POST GRADUATE DEGREE PROGRAMME (CBCS)

M.SC. IN MATHEMATICS

SEMESTER I

SELF LEARNING MATERIAL

PAPER: COR 1.1
(Pure and Applied Streams)

Real Analysis I

Complex Analysis I

Functional Analysis I

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

Course Preparation Team	
Dr. Abhijit Benarjee, Professor, Department of Mathematics, University of Kalyani	Dr. Indrajit Lahiri, Professor, Department of Mathematics, University of Kalyani
Dr. Pulak Sahoo, Professor, Department of Mathematics, University of Kalyani	Dr. Animesh Biswas, Professor, Department of Mathematics, University of Kalyani
Dr. Biswajit Mallick, Assistant Professor (Cont), DODL, University of Kalyani	Ms. Audrija Choudhury, Assistant Professor (Cont), DODL, University of Kalyani

Dec 2021

Directorate of Open and Distance Learning, University of Kalyani

Published by the Directorate of Open and Distance Learning

University of Kalyani, 741235, West Bengal

SYLLABUS

COR 1.1

Marks: 100; Credits: 6

Unit	Topic	Counselling Duration
Block I: Real Analysis I; Marks 32 (SEE: 25; IA: 07)		
1	Cardinal number **:** Definition, Schröder-Bernstein theorem, Order relation of cardinal numbers, Arithmetic of cardinal numbers, Continuum hypothesis	54 Mins
2	Cantor's set **:** Construction and its presentation as an uncountable set of measure zero	54 Mins
3	Functions of bounded variation **:** Definition and basic properties, Lipschitz condition, Jordan decomposition,	54 Mins
4	Nature of points of discontinuity, Nature of points of non-differentiability, Convergence in variation (Helly's First theorem)	54 Mins
5	Absolutely continuous functions **:** Definition and basic properties, Deduction of the class of all absolutely continuous functions as a proper subclass of all functions of bounded variation,	54 Mins
6	Characterization of an absolutely continuous function in terms of its derivative vanishing almost everywhere	54 Mins
7	Riemann-Stieltjes integral **:** Existence and basic properties, Integration by parts, Integration of a continuous function with respect to a step function,	54 Mins
8	Convergence theorems in respect of integrand, convergence theorem in respect of integrator (Helly's Second theorem)	54 Mins
9	Gauge partition **:** Definition of a delta-fine tagged partition and its existence, Lebesgue's criterion for Riemann integrability,	54 Mins
10	Delta-fine free tagged partition and an equivalent definition of the Riemann integral	54 Mins
Block II: Complex Analysis I; Marks 36 (SEE: 30; IA: 06)		
11	Riemann's sphere, point at infinity and the extended complex plane	54 Mins
12	Functions of a complex variable, limit and continuity. Analytic functions, Cauchy-Riemann equations	54 Mins
13	Complex integration. Cauchy's fundamental theorem (statement only) and its consequences. Cauchy's integral formula. Derivative of an analytic function	54 Mins
14	Morera's theorem, Cauchy's inequality, Liouville's theorem, Fundamental theorem of classical algebra	54 Mins
15	Uniformly convergent series of analytic functions. Power series. Taylor's theorem. Laurent's theorem	54 Mins
Block III: Functional Analysis I; Marks 32 (SEE: 25; IA: 07)		

16	Metric spaces. Brief discussions of continuity, completeness, compactness. Hölder's and Minkowski's inequalities (statement only)	54 Mins
17	Baire's (category) theorem. The spaces and. Banach's fixed point theorem	54 Mins
18	Applications to solutions of certain systems of linear algebraic equations, Fredholm's integral equation of the second kind, implicit function theorem. Kannan's fixed point theorem	54 Mins
19	Real and Complex linear spaces. Normed induced metric. Banach spaces, Riesz's lemma	54 Mins
20	Finite dimensional normed linear spaces and subspaces, completeness, compactness criterion, equivalent norms	54 Mins
Total		18 Hours

Block I
Real Analysis I

Unit 1

Course Structure

1. Cardinal Number
2. Schröder Berstein Theorem
3. Order Relation of cardinal numbers
4. Arithmetic of cardinal numbers
5. Continuum hypothesis

1 Introduction

Cardinal numbers, or cardinals for short, are a generalization of the natural numbers used to measure the cardinality (size) of sets. The cardinality of a finite set is a natural number: the number of elements in the set.

The notion of cardinality, as now understood, was formulated by Georg Cantor, the originator of set theory, in 1874–1884. Cardinality can be used to compare an aspect of finite sets; e.g. the sets $\{1, 2, 3\}$ and $\{4, 5, 6\}$ are not equal, but have the same cardinality. We will now move on to the formal definitions and preliminaries.

1.1 Cardinal Numbers

Definition 1.1. *A set A is said to be equipotent to a set B written as $A \approx B$, there exists a bijective mapping from A to B.*

For arbitrary sets A, B, C we have
(i) $A \approx A$ (identity map)
(ii) $A \approx B \Rightarrow B \approx A$ (inverse map)
(iii) $A \approx B$ and $B \approx C \Rightarrow A \approx C$ (composite map).

It, therefore, follows that in any given family of sets, the relation of equipotence is an equivalence relation which therefore partitions the given family into pairwise disjoint equivalence classes, the members of of the same class being mutually equipotent. Intuitively, we think that every two equipotent sets have the same number of elements. This leads to the following postulates:

With each set A there is associated a unique object, to be called the cardinal number (or, power) of the set A, denoted by $\overline{\overline{A}}$, such that two sets have the same cardinal number if and only if they are equipotent. That is $\overline{\overline{A}} = \overline{\overline{B}} \iff A \approx B$.

By saying that m is the cardinal number of the set A, we shall mean that $m = \overline{\overline{A}}$. However the cardinal number of a finite set is taken to be the number of elements in the set. Thus $\overline{\overline{\Phi}} = 0$ and $\{\overline{\overline{1}}\} = 1$, $\left\{\overline{\overline{1, 2, \ldots, n}}\right\} = n$.

Two finite sets are equipotent iff they have the same number of elements.

If $A = \{1, 2, 3, \ldots\}$, $B = \{2, 4, 6, \ldots\}$ and $C = \{-1, -2, -3, \ldots\}$, we can easily see that $A \approx B \approx C$.

Theorem 1.1. *Given any index set $\{m_i\}_{i\in\Gamma}$ of cardinal numbers, there exists pairwise disjoint set such that $m_i = \overline{\overline{A_i}}$ for each $i \in \Gamma$.*

Proof. By definition there exists sets such that $m_i = \overline{\overline{B_i}}$ for each $i \in \Gamma$. We evidently have $A_i \cap A_j = \phi$ for $i \neq j$. Thus $\{A_i\}_{i\in\Gamma}$ is a family of pairwise disjoint set. Now for each $i \in \Gamma$, the prescription $b \to (b, i)$ for all $b \in B_i$ obviously defines a bijective mapping from B_i to A_i. So we have $m_i = \overline{\overline{A_i}}$ for each $i \in \Gamma$. This proves the theorem. □

1.2 Addition of Cardinal numbers:

For any two cardinal numbers m and n, the sum $m + n$ is defined to be the cardinal number of the set $A \cup B$, where A and B are disjoint sets with $m = \overline{\overline{A}}$ and $n = \overline{\overline{B}}$.

Justification:

Given the cardinal numbers m and n, we know that there exists disjoint sets A and B such that $m = \overline{\overline{A}}$ and $n = \overline{\overline{B}}$. Then by definition $m + n = \overline{\overline{A \cup B}}$. Suppose A_1 and B_1 are also disjoint sets with $m = \overline{\overline{A_1}}$ and $n = \overline{\overline{B_1}}$. Then to show that $m+n$ is unambiguously defined, we are to show that $\overline{\overline{A \cup B}} = \overline{\overline{A_1 \cup B_1}}$. Now, since $\overline{\overline{A}} = m = \overline{\overline{A_1}}$ and $\overline{\overline{B}} = n = \overline{\overline{B_1}}$, there exist bijections $f : A \longrightarrow A_1$ and $g : B \longrightarrow B_1$.

Let $h : A \cup B \longrightarrow A_1 \cup B_1$ be defined by

$$h(x) = \begin{cases} f(x), \ if \ x \in A \\ g(x), \ if \ x \in B. \end{cases}$$

Since $A \cap B = \phi$, h is a well defined function. Also since, f and g are surjective, we have $h[a \cup B] = h[A] \cup h[B] = A_1 \cup B_1$. Thus h is surjective.

We now show that h is injective. Let x_1, x_2 be any two distinct points of $A \cup B$. If both x_1, $x_2 \in A$, then $h(x_1) = f(x_1) \neq f(x_2) = h(x_2)$, since f is injective.

If both $x_1, x_2 \in B$, then $h(x_1) = g(x_1) \neq g(x_2) = h(x_2)$, since g is injective.

The remaining possibility is that one of x_1, x_2 be in A and the other is in B. But then one of $h(x_1)$, $h(x_2)$ belongs to $f[A] = A_1$ and the other to $g[B] = B_1$. Since $A \cap B = \phi$, we can not have $h(x_1) = h(x_2)$. Thus h is bijective map from $A \cup B$ to $A_1 \cup B_1$. Therefore, $\overline{\overline{A \cup B}} = \overline{\overline{A_1 \cup B_1}}$.

Theorem 1.2. *For any two cardinals m and n, $m+n = n+m$* $\left[\overline{\overline{A \cup B}} = \overline{\overline{B \cup A}}\right]$.

Theorem 1.3. *For any three cardinals m, n, p, we have $(m+n)+p = m+(n+p)$.*

1.3 Multiplications of Cardinal numbers:

Definition 1.2. *The product of two cardinal numbers m and n denoted by mn is defined by $mn = \overline{\overline{A \times B}}$, where $m = \overline{\overline{A}}$ and $n = \overline{\overline{B}}$.*

To show that the product is uniquely defined, we show that if $m = \overline{\overline{A}} = \overline{\overline{A_1}}$ and $n = \overline{\overline{B}} = \overline{\overline{B_1}}$, then $\overline{\overline{A \times B}} = \overline{\overline{A_1 \times B_1}}$.
Now there exists bijections $f : A \longrightarrow A_1$ and $g : B \longrightarrow B_1$.
Let us define $h : A \times B \longrightarrow A_1 \times B_1$ by $h(a, b) = (f(a), f(b))$ for all $(a, b) \in A \times B$.
Then h is a function from $A \times B \longrightarrow A_1 \times B_1$. If $h(a, b) = h(a_1, b_1)$, then

$$\begin{aligned}(f(a), g(b)) = (f(a_1, g(b_1)) &\implies f(a) = f(a_1) \text{ and } g(b) = g(b_1)\\ &\implies a = a_1 \text{ and } b = b_1\\ &\implies (a, b) = (a_1, b_1).\end{aligned}$$

So h is injective. Now we shall show that h is surjective.
Again, if $(a_1, b_1) \in A_1 \times B_1$, then $a_1 \in A_1$ and $b_1 \in B_1$, but then we have $a \in A$ such that $f(a) = a_1$ and $b \in B$ such that $g(b) = b_1$. It follows that $h(a, b) = (f(a), g(b)) = (a_1, b_1)$. So, h is surjective. Therefore, h is bijective.

Theorem 1.4. *For any three cardinals, we have (i) $mn = nm$, (ii) $(mn)p = m(np)$, (iii) $m(n + p) = mn + mp$.*

Proof. (ii) Suppose $m = \overline{\overline{A}}$, $n = \overline{\overline{B}}$ and $p = \overline{\overline{C}}$. Then $np = \overline{\overline{B \times C}}$.
Also, $(mn)p = \overline{\overline{(A \times B) \times C}}$ and $m(np) = \overline{\overline{A \times (B \times C)}}$.
clearly the map $h : (A \times B) \times C \longrightarrow A \times (B \times C)$ defined by

$$h((a, b), c) = (a, (b, c))$$

is a bijection.

Hence, $(mn)p = m(np)$.

(iii) We assume that $B \cap C = \Phi$. Then $(n + p) = \overline{\overline{B \cup C}}$ and so $m(n + p) = \overline{\overline{A \times (B \cup C)}}$

Also $mn = \overline{\overline{A \times B}}$ and $mp = \overline{\overline{A \times C}}$.

Since $B \cap C = \Phi$, we have $(A \times B) \cap (A \times C) = \Phi$.

So, $mn + mp = \overline{\overline{(A \times B)(A \times C)}}$. But we know that $A \times (B \cup C) = (A \times B) \cup (A \times C)$.

It follows that

$$m(n + p) = mn + mp.$$

□

Example 1.1. *Show that for any two cardinal $m.0 = 0$ and $m.1 = m$.*

Proof. Let $m = \overline{\overline{A}}$. We have $\overline{\overline{\Phi}} = 0$ and $\{\overline{\overline{a}}\} = 1$. Then $m.0 = \overline{\overline{A \times \Phi}} = \overline{\overline{\Phi}} = 0$. Also $m.1 = \overline{\overline{A \times \{1\}}} \longrightarrow A$. But the prescription $a \longrightarrow (a, 1)$, $a \in A$ obviously defines a bijection from $A \longrightarrow A \times \{1\}$. *i.e.*,

$$\overline{\overline{A \times \{1\}}} = \overline{\overline{A}} = m = m.1.$$

□

Theorem 1.5. *Let A be the union of an indexed family $\{A_i\}$ of pairwise disjoint sets such that $\overline{\overline{A_i}} = m$ (a fixed cardinal) for all $i \in \Gamma$. Then $\overline{\overline{A_i}} = mn$, where $n = \overline{\overline{\Gamma}}$.*

Proof. Let $m = \overline{\overline{B}}$. Then for each $i \in \Gamma$, there is a bijection $f_i :\to B$. WE define a mapping $f : A \to B \times \Gamma$ by stipulating that $f(x) = (f_i(x), i)$ if $x \in A_i$, $i \in \Gamma$.

Since A is the sum of the sets A_i and since the sets A_i are pairwise disjoint to each x, there corresponds a unique $i \in \Gamma$ such that $x \in A_i$. From this observation it follows that f is a well defined function from A to $B \times \Gamma$.
Now if $f(x) = f(y)$ for some $x, y \in A$, then $(f_i(x), i) = (f_j(y), j)$, where i is the unique index for which $x \in A_i$ and j is the unique index for which $y \in A_j$.

Then $f_i(x) = f_j(y)$ and $i = j$. So $f_i(x) = f_i(y)$, $x, y \in A_i$.
m Since f_i is injective, we have $x = y$. Thus f is injective.
Again given any $(b, i) \in B \times \Gamma$, we note that $f_i : A_i \to B$ is surjective and hence there is an $x \in A_i$ for which $f_i(x) = b$. Then $f(x) = (f_i(x), i) = (b, i)$, since $x \in A_i$.
Thus f is surjective and hence bijective. Consequently, $\overline{\overline{A}} = \overline{\overline{B \times \Gamma}} = \overline{\overline{B\Gamma}} = mn$. □

1.4 Exponentiation of the Cardinals:

If m and n are two cardinals, then we explain the meaning of m^n.

Definition 1.3. *A set of all functions from a given set X to a given set Y is denoted by Y^X. i.e., $f \in Y^X \Longrightarrow f$ is a function from $X \longrightarrow Y$.*

A set of two elements is usually denoted by 2. We specially take $2 = \{0, 1\}$. Thus for any set X, 2^X denotes the set of all functions from X to $2 = \{0, 1\}$. The set of Y^X is clearly a subset of power set of $X \times Y$ *i.e.*, $P(X \times Y)$.
If $X = \phi$, then Y^X consists of only one member the empty subset of $X \times Y$. This is the only subset of $X \times Y$, since when X is empty so is $X \times Y$. If $Y = \phi$ and $X \neq \phi$, then Y^X is empty. When $X = \phi$, then $Y^\phi = \{\phi\}$. $\phi^X = \phi$ if $X = \phi$.

Definition 1.4. *For any two cardinal numbers m and n, m^n is defined to be the cardinal number of the set A^B of all functions from B to A, where $m = \overline{\overline{A}}$ and $n = \overline{\overline{B}}$, $\overline{\overline{A^B}} = \overline{\overline{A}}^B$, $A^B = A_1^{B_1}$, $m = \overline{\overline{A}} = \overline{\overline{A_1}}$ and $n = \overline{\overline{B}} = \overline{\overline{B_1}}$.*

Definition 1.5. *Given a function $f : X \to Y$ and a subset $A \subset X$, the function $f|_A : A \to Y$ defined by $(f|_A)(x) = f(x)$ for all $x \in A$, is called the restriction of f to the subset A.*

Theorem 1.6. *For any three cardinals m, n, p, we have*
(i) $m^{n+p} = m^n m^p$
(ii) $(mn)^p = m^n m^p$
(iii) $(m^n)^p = m^{np}$.

Proof. Let A, B, C be pairwise disjoint sets with $m = \overline{\overline{A}}$, $n = \overline{\overline{B}}$ and $p = \overline{\overline{C}}$.

(i) We have $n + p = \overline{\overline{B \cup C}}$, $m^n = \overline{\overline{A^B}}$, $m^p = \overline{\overline{A^C}}$. $m^{n+p} = \overline{\overline{A^{B\cup C}}}$, $m^n n^p = \overline{\overline{A^B \times A^C}}$.

We are to show the existence of a bijection $\phi : \overline{\overline{A^{B\cup C}}} \to A^B \times A^C$.

We define ϕ as follows:

$\phi(f) = (f|_B, f|_C)$ for all $f \in A^{B\cup C}$.

Then ϕ is a well defined function from $A^{B\cup C}$ to $A^B \times A^C$.

Suppose $\phi(f) = \phi(g)$ for some $f, g \in A^{B\in C}$. Then $(f|_B, f|_C) = (g|_B, g|_C)$ whence $f|_B = g|_C$ and $f|_C = g|_C$ which imply that $f = g$. Thus ϕ is injective.

Consider any element $(f_1, f_2) \in A^B \times A^C$. Then $f_1 : B \to A$, $f_2 : C \to A$. Since B and C are disjoint, obviously $f(x) = f_1(x)$, if $x \in B$ and $f(x) = f_2(x)$, if $x \in C$ defines a function from $B \cup C$ to A, i.e., $f \in A^{B\cup C}$. Then we have $\phi(f) = (f|_B, f|_C) = (f_1, f_2)$. So ϕ is surjective. Thus ϕ is bijective.
Hence $m^{n+p} = m^n m^p$.

(ii) We have $(mn)^p = \overline{\overline{(A \times B)^C}}$, i.e., $m^p n^p = \overline{\overline{(A \times B)^C}}$. We note that if $f \in (A \times B)^C$, then for every $c \in C$ we have $f(c) \in A \times B$, so that $f(c) = (f_A(c), f_B(c))$, where $f_A(c)$ and $f_B(c)$ are elements of A and B respectively uniquely determined by f and C. Thus f determines a unique pair of functions $f_A : C \to A$ and $f_B : C \to B$ such that $f(c) = (f_A(c), f_B(c))$ for all $c \in C$. We define $\phi(f) = (f_A, f_B)$. Then ϕ is a function from $(A \times B)^C$ to $A^C \times B^C$.

To show that ϕ is injective, we suppose $\phi(f) = \phi(g)$ for some $f, g \in (A\times B)^C$. Then $(f_A, f_B) = (g_A, g_B)$ whence $f_A = g_A$ and $f_B = g_B$. Then for all $c \in C$ we have

$$f(c) = (f_A(c), f_B(c)) = (g_A(c), g_B(c)) = g(c).$$

So $f = g$. Thus, ϕ is injective.

To show ϕ is surjective, consider any element $(f_1, f_2) \in A^C \times B^C$. Then $f_1 : C \to A$ and $f_2 : C \to B$. We define $f : C \to A \times B$ by $f(c) = (f_1(c), f_2(c))$ for all $c \in C$.

Then $f \in (A \times B)^C$ and obviously $f_A = f_1$ and $f_B = f_2$ for this f. Thus $\phi(f) = (f_A, f_B) = (f_1, f_2)$. Thus ϕ is surjective and hence bijective. So, $(mn)^p = m^p n^p$.

(iii) We have $(m^n)^p = \overline{\overline{(A^B)^C}}$ and $m^{np} = \overline{\overline{(A^{B\times C})}}$. We note that for every $f \in (A^B)^C$, the value f_c of f at a point $c \in C$ is a member of A^B. i.e., $f_c : B \to A$

for every $c \in C$ and so $f_c(b) \in A$ for all $b \in B$. Thus, every $f \in (A^B)^C$ includes a unique function $f^* : B \times C \to A$ defined by $f^*(b, c) = f_c(b)$ for all $(b, c) \in B \times C$.

Thus we get a function $\phi : (A^B)^C \to A^{B\times C}$ defined by $\phi(f) = f^*$ for all $f \in (A^B)^C$.

To show ϕ is injective suppose $\phi(f) = \phi(g)$ for some $f, g \in (A^B)^C$. Then $f^* = g^*$ so that $f^{(b,c)} = g^*(b, c)$ for all $(b, c) \in B \times C$. i.e., $f_c(b) = g_c(b)$ for all $b \in B$ and all $c \in C$. i.e., $f_c = g_c$ for all $c \in C$. So, $f = g$. Hence, ϕ is injective.

To show ϕ is surjective, we consider any $h \in A^{B\times C}$ for all $(b, c) \in B \times C$, we have $h(b, c) \in A$.

Let us define $f : C \to A^B$ by stipulating that for every $c \in C$, f_c is that function from $B \to A$ for which $f_c(b) = h(b, c)$ for all $b \in B$. For this f and for all $(b, c) \in B \times C$, we have

$$f^*(b, c) = f_c(b) = h(b, c).$$

In other words, $\phi(f) = f^* = h$. Thus ϕ is surjective and hence bijective. So, $(m^n)^p = m^{np}$. □

Theorem 1.7. *For any set A, power set* $\overline{\overline{P(A)}} = \overline{\overline{2^A}} = 2^{\overline{\overline{A}}}$.

Proof. For each $B \in P(A)$, let 1_B denote the characteristic function of Bin A. *i.e.*,

$$1_B(x) = \begin{cases} 1, \ if \ x \in B \\ 0, \ if \ x \in A - B. \end{cases}$$

1_B is a function from $A \longrightarrow 2$. So $1_B \in 2^A$.

Now we define $\Phi : P(A) \longrightarrow 2^A$ by $\Phi(B) = 1_B$ for all $B \in P(A)$.
We show that Φ is injective. Suppose $\Phi(B) = \Phi(C)$ for some $B, C \in P(A)$. Then

$$1_B = 1_C.$$

i.e., $1_B(x) = 1_C(x)$ for all $x \in A$. In-particular, $x \in B \iff 1_B(x) = 1 = 1_C(x) \iff x \in C$. So, $B = C$. Hence, Φ is injective.

Next we show that Φ is surjective.
Consider any $g \in 2^A$. Put $B = g^{-1}[\{1\}]$. Since, the set of the functions maps from A to $\{0, 1\}$ belongs to 2^A, clearly $B \in P(A)$, we have $1_B = g$. *i.e.*, $\Phi(B) = g$. Thus Φ is surjective and hence bijective. Thus

$$\overline{\overline{P(A)}} = \overline{\overline{2^A}} = 2^{\overline{\overline{A}}}.$$

□

Note: The result of the above theorem is an extension of the fact that for a finite set A of n elements, $P(A)$ consists of exactly 2^n elements.

Example 1.2. *Let m be any cardinal number. Then*

$$(i)\ m^1 = m\ (ii)\ 1^m = 1\ (iii)\ m^0 = 1\ (iv)\ 0^m = 0,\ if\ m \neq 0.$$

Soln: (i) Suppose $m = \overline{\overline{A}}$, $m^1 = A^{\overline{\overline{\{1\}}}}$. Let $\Phi : A^{\{1\}} \longrightarrow A$ given by $\Phi(f) = f(1)$.
Clearly, for every $a \in A$, $\Phi(f_a) = a$, where $f_a(1) = a$.
Clearly,

$$\Phi(f_a(1)) = \Phi(f_b(1))$$

$$\Rightarrow a = b$$

$$\Rightarrow f_a(1) = f_b(1)$$

$$\Rightarrow f_a = f_b.$$

Thus for every $a \in A$, there exists $f_a \in A^{\{1\}}$ such that $f_a(1) = a$. So, Φ is surjective.

(ii) $1^m = \{1\}^{\overline{\overline{A}}}$. But $\{1\}^A$ consists of exactly one function which is $f : A \longleftarrow \{1\}$; *i.e.*,

$$1^m = 1.$$

Example 1.3. *Let m, n denote cardinal numbers. Show that*
(i) $mn = 1$ if and only if $m = n = 1$
(ii) $mn = 0$ if and only if atleast one of m and n is zero.

Soln: Suppose $m = \overline{\overline{A}}$, $n = \overline{\overline{B}}$. Then $mn = \overline{\overline{A \times B}}$.

(i) $A \times B$ consists of one element if and only if A and B consists of only one element.i.e., $n = \overline{\overline{B}} = 1$ if and only if $n = \overline{\overline{A}} = 1$ and $n = \overline{\overline{B}} = 1$. Therefore, $mn = 1$ if and only if $m = 1$ and $n = 1$.

(ii) $n = \overline{\overline{A \times B}} = 0$ if and only if either $n = \overline{\overline{A}} = 0$ or $n = \overline{\overline{B}} = 0$. Therefore, $mn = 0$ if and only if either $m = 0$ or $n = 0$.

1.5 Ordering of Cardinal Numbers:

A cardinal number m is said to be less than or equal to a cardinal number n, written as $m \leq n$, if there exists sets A and B with $m = \overline{\overline{A}}$ and $n = \overline{\overline{B}}$ such that there is an injective map from A to B (i.e., m is the cardinal number of a subset of B). If $m \leq n$ but $m \neq n$, then we write $m < n$.

Theorem 1.8. *If $m \leq n$, then for every two sets A, B with $m = \overline{\overline{A}}$ and $n = \overline{\overline{B}}$, there is an injection from A to B.*

Proof. Since $m \leq n$, by definition, there is a pair of sets A_0 and B_0 with $m = \overline{\overline{A_0}}$ and $n = \overline{\overline{B_0}}$ for which there is an injection f from A_0 to B_0. Now, since $m = \overline{\overline{A_0}} = \overline{\overline{A}}$ and $n = \overline{\overline{B_0}} = \overline{\overline{B}}$, there exists bijections $g : A \longrightarrow A_0$ and $h : B_0 \longrightarrow B$.

Thus $h \circ f \circ g$ is an injection from $A \longrightarrow B$. □

Example 1.4. *If $A \subset B$, then $\overline{\overline{A}} \leq \overline{\overline{B}}$.*

Soln: *$I : A \longrightarrow B$ is the identity mapping from A to B.*

Example 1.5. *If $\overline{\overline{A}} = m$, then $0 \leq m$.*

Soln: $\Phi \subset A \Rightarrow \overline{\overline{\Phi}} \leq \overline{\overline{A}} \Rightarrow 0 \leq m$.

Example 1.6. *If $m \neq 0$, then $1 \leq m$.*

Soln: *Let $\overline{\overline{A}} = m \neq 0$ and $a \in A$. Then*

$$\{a\} \subset A$$

$$\Rightarrow \overline{\overline{\{1\}}} \leq \overline{\overline{A}}$$

$$\Rightarrow 1 \leq m.$$

Theorem 1.9. *If $m \leq n$ and $n \leq p$, then $m \leq p$.*

Proof. Let $\overline{\overline{A}} = m$, $\overline{\overline{B}} = n$ and $\overline{\overline{C}} = p$. Suppose $f : A \longrightarrow B$ is an injection and $g : B \longrightarrow C$ is an injection. So, $g \circ f : A \longrightarrow C$ is an injection. So, $m \leq p$. □

Theorem 1.10. (Schroder-Bernstein theorem) [S. B. theorem] *If $m \leq n$ and $n \leq m$, then $m = n$. (Antisymmetry)*

Proof. We first suppose that $m \leq n$. Then if $m = \overline{\overline{A}}$ and $n = \overline{\overline{B}}$, there is an injective map $f : A \longrightarrow B$. Let us put $C = B - f[A]$ and $P = \overline{\overline{C}}$.

Since $f : A \longrightarrow f[A]$ is a bijection, we have $\overline{\overline{P[A]}} = \overline{\overline{A}} = m$. Since $f[A]$ and C are disjoint sets, we have $m + p = \overline{\overline{f[A] \cup [B - f[A]]}} = n$. i.e., $m + p = n$.

Conversely, suppose that $m + p = n$ for some p. Then there exists disjoint sets A and P with $\overline{\overline{A}} = m$ and $\overline{\overline{P}} = p$ and $n = \overline{\overline{A \cup P}}$.
Since $A \subset A \cup P$, we have $\overline{\overline{A}} \leq \overline{\overline{A \cup P}}$. i.e., $m \leq n$. □

Theorem 1.11. *If m, n, p be three cardinals with $m \leq n$, then (i) $m+p \leq n+p$, (ii) $mp \leq np$, (iii) $m^p \leq n^p$.*

Proof. (i) Since $m \leq n$, there is a cardinal r such that $n = m + r$. Then $n+p = (m+r)+p = m+(r+p) = m+(p+r) = (m+p)+r$. i.e., $m+p \leq n+p$.

(ii) Also, $np = (m + r)p = mp + rp$. Therefore, $mp \leq np$.

(iii) To prove (iii), let $m = \overline{\overline{A}}$, $n = \overline{\overline{B}}$, where $A \cap B = \Phi$. Then $n = \overline{\overline{A \cup B}}$. If $p = \overline{\overline{C}}$, then since $A^c \subset (A \cup B)^c$, it follows that $m^p \leq n^p$. □

Theorem 1.12. *Given the cardinal numbers m and n, we have $m \leq n$ if and only if there is a cardinal p such that $n = m + p$.*

Proof. We first suppose that $m \leq n$. Then if $m = \overline{\overline{A}}$ and $n = \overline{\overline{B}}$. There is an injective map $f : A \to B$. Let us put $C = B - f[A]$ and $p = \overline{\overline{C}}$.

Since $f : A \to f[A]$ is a bijection, we have $\overline{\overline{f[A]}} = \overline{\overline{A}} = m$.

Since $f[A]$ and C are disjoint sets, we have

$$m + p = \overline{\overline{f[A] \cup (B - f[A])}} = n.$$

Conversely, suppose $m + p = n$ for some p. Then there exists disjoint sets A and P with $\overline{\overline{A}} = m$ and $\overline{\overline{P}} = p$ and $n = \overline{\overline{A \cup P}}$. Since, $A \subset A \cup P$, we have $\overline{\overline{A}} \leq \overline{\overline{A \cup P}}$. i.e., $m \leq n$. □

Theorem 1.13. *For any cardinal m, $m < 2^m$.*

Proof. Let $m = \overline{\overline{A}}$. Then we know that $2^m = \overline{\overline{P(A)}}$. Now the prescription $x \to \{x\}$, $x \in A$ clearly defines a injective map from A to $P(A)$. So, we have $m \leq 2^m$.

We show that $m \neq 2^m$. Consider any map $f : A \to P(A)$. We note that for every $x \in A$, $f(x) \in P(A)$. i.e., $f(x)$ is a subset of A. Let then $B = \{x \in A : x \notin f(x)\}$. We have $B \in P(A)$. Consider any $x \in A$. If $x \in B$, then by definition of B, we have $x \notin f(x)$ and so $f(x) \neq B$. If $x \notin B$, then by definition of B, we have $x \in f(x)$ and hence $f(x) \neq B$. Thus no function from A to $P(A)$ is surjective. Hence we can not have $\overline{\overline{A}} = \overline{\overline{P(A)}}$, so that $m \neq 2^m$. Consequently, $m < 2^m$. □

Some Particular Cardinals:

The letters a and c will be used to denote some particular cardinals.

Definition 1.6. *$a = \overline{\overline{\mathbb{N}}}$, where $\mathbb{N} = \{1, 2, 3, \ldots\}$. Here a is called the denumerable cardinal. $c = \overline{\overline{\mathbb{R}}}$, where $\mathbb{R}$ is the set of all real numbers. Here, c is called the power of the continuum, $\mathbb{R}$ is called the arithmetic continuum.*

Example 1.7. $\overline{\overline{\{0, \pm 1, \pm 2, \ldots\}}} = a$.

Definition 1.7. *A set X is said to be denumarable (or, enumarable, or countably infinite) if $\overline{\overline{X}} = a$. A set Y is called countable if it is either finite or denumarable. A set is said to be uncountable if it is not countable. Thus a set X is denumarable iff the elements of X can be arranged in a sequence (infinite) $x_1, x_2, \ldots$ of distinct terms.*

Cardinal numbers of finite sets are called finite cardinals and those of infinite sets are called infinite or transfinite cardinals. Thus a and c are transfinite cardinals, while $0, 1, 2, \ldots$ are finite cardinals. Now, we require the knowledge of radix fraction.

Definition 1.8. *Let r be any integer ≥ 2. A number of the form $\frac{p}{r^m}$ where p is any integer and m is any integer ≥ 0 is called a radix fraction with radix r.*

Every real number x can be represented uniquely by a series of radix fraction with any given radix $r(\geq 2)$ as follows:

$$x = x_{0+\frac{x_1}{r}} + \frac{x_2}{r^2} + \ldots + \frac{x_n}{r^n} + \ldots.$$

The representation is unique if x is not a number of the form $\frac{m}{2^n}$ $(m = 1, 3, \ldots, 2^n - 1)$.

$$\frac{3}{8} = \begin{cases} 0.011000\ldots \\ 0.010111\ldots \end{cases}$$

If however we do not use representation where $x_k = 1$ for some steps occurred. Then for every real number $x \in (0, 1)$, the representation is unique.

Here (i) x_0 is the greatest integer not greater than x.

(ii) Each x_n is an integer on $0 \leq x_n \leq (r-1)$ for all $n = 1, 2, 3, \ldots$.

(iii) $x_n \leq (r-2)$ for infinitely many n.

The above representation is uniquely written as

$$x = x_0 x_1 \ldots (\text{radix } r).$$

For $r = 2$, the representation is called binary rdyadic.
For $2 = 3$, the representation is called ternary.

For $r = 10$, the representation is called decimary.

For $r = n$, the representation is called n-ary.

Theorem 1.14. *The open interval* $(0, 1)$ *of real numbers is uncountable.*

Proof. The set $(0, 1)$ is evidently infinite. Any real number in $(0, 1)$ can be written as an infinite decimal of the form $d_1 d_2 d_3 \ldots$, where $d_1, d_2, \ldots$ are digital numbers $0, 1, 2, \ldots, 9$.

Suppose for a contradiction that $(0, 1)$ is denumerable. So we can enumerate the set of all real numbers in $(0, 1)$ as $\{a_1, a_2, a_3, \ldots\}$, where

$$a_1 = .d_{11} d_{12} d_{13} \ldots$$

$$a_2 = .d_{21} d_{22} d_{23} \ldots$$

$$\ldots\ldots\ldots\ldots$$

$$a_n = .d_{n1} d_{n2} d_{n3} \ldots$$

Every d_{ij} being a digital number. Now, we choose a real number a in the following manner:

$$a = .d_1 d_2 d_3 \ldots,$$

where $d_i = 1$ if $d_{ii} \neq 1$ and $d_i = 2$ if $d_{ii} = 1$ for $i = 1, 2, \ldots$. Clearly, $a \in (0, 1)$ and a is different from $a_1, a_2, \ldots, a_n, \ldots$, which is a contradiction. Hence, the interval $(0, 1)$ is uncountable. □

Note:- In choosing a, we should avoid 0 or 9 for d_i's since many rational numbers in $(0,1)$may have two decimal expansions, one of them having 0 recurring and the other 9 recurring.

i.e.,

$$\frac{1}{2} = .5000\ldots$$

$$\frac{1}{4} = .2500\ldots = .2499.$$

Verification:-

$$\begin{aligned} & \frac{4}{10} + \frac{9}{10^2} + \frac{9}{10^3} + \ldots \\ = \ & \frac{4}{10} + 9(-1 - \frac{1}{10} + 1 + \frac{1}{10} + \frac{1}{10^2} + \frac{1}{10^3} + \ldots) \\ = \ & \frac{4}{10} + 9(\frac{1}{1-\frac{1}{10}} - 1 - \frac{1}{10}) \\ = \ & \frac{4}{10} + 9(\frac{10}{9} - \frac{11}{10}) \\ = \ & \frac{4}{10} + 9\frac{1}{10} = \frac{5}{10} = \frac{1}{2}. \end{aligned}$$

Theorem 1.15. *Every infinite subset of a denumerable set is denumerable.*

Proof. Let B be an infinite subset of a denumerable set A. The elements of A can be arranged in a sequence of distinct terms as

$$a_1, a_2, a_3, \ldots, a_n, \ldots.$$

Let n_1 be the smallest index such that $a_{n_1} \in B$. In general, having defined $n_1, n_2, \ldots, n_k$, let n_{k+1} denotes the smallest index such that $a_{n_{k+1}} \in B - \{a_{n_1}, a_{n_2}, \ldots, a_{n_k}\}$.

Since $B \subset A$ and since B is infinite, we get a well defined sequence of positive integers $n_1 < n_2 < \ldots$

Since all the elements of B occur in, it is clear that every $x \in B$ is some a_{n_k}.

Thus $B = \{a_{n_1}, a_{n_2}, \ldots, a_{n_k}\}$ which shows that B is denumerable. □

Corollary 1.1. $\mathbb{R}$ *is non-denumerable. For if* $\mathbb{R}$ *is denumerable and since* $(0,1) \subset \mathbb{R}$*, then* $(0,1)$ *is denumerable, which is a contradiction. So* $\mathbb{R}$ *is non-denumerable.*

Corollary 1.2. $a < c$.

Proof. We have $\mathbb{N} \subset \mathbb{R}$. This implies $\overline{\overline{\mathbb{N}}} \leq \overline{\overline{\mathbb{R}}}$. Therefore, $a \leq c$.
It is clear that $(0,1)$ is not equivalent to any subset of $\mathbb{N}$. But a subset of $(0,1)$ can be found out such that $\mathbb{N}$ is equivalent to that subset. Hence $a \neq c$. Therefore, $a < c$. □

Theorem 1.16. *The union of a countable family of countable sets is countable.*

Proof. Let B denotes the union of a countable family $\mathcal{B}$ of countable sets. If $\mathcal{B} = \phi$, then $B = \phi$ and B is of course finite and hence countable. If $\mathcal{B} \neq \phi$, then being countable, its non-empty members can be arranged in a finite or infinite sequence $B_1, B_2, B_3, \ldots$ so that $B = \cup_i B_i$. Since each B_i is countable and non-empty, the members of B_i can be arranged in a finite or infinite sequence of distinct terms as $b_{i1}, b_{i2}, b_{i3} \ldots$. Now each $x \in B$, let $m(x)$ denote the smallest positive integer such that $x \in b_m(x)n(x)$. We define a mapping $f : B \to \mathbb{N}$ by saying that $f(x) = 2^{m(x)}3^{n(x)}$.

If $f(x) = f(y)$, then $2^{m(x)}3^{n(x)} = 2^{m(y)}3^{n(y)}$. Since 2 and 3 are distinct primes, it follows that $m(x) = m(y)$ and $n(x) = n(y)$.

Then $x = b_{m(x)n(x)} = b_{m(y)n(y)} = y$. Thus f is injective. Hence, $\overline{\overline{B}} \leq \overline{\overline{\mathbb{N}}}$ and B is countable. □

Corollary 1.3. *The cartesian product of two denumerable sets is denumerable.*

Proof. Let A and B be denumerable sets. Then $A \times B = \cup_{b \in B}\{(x, b) : x \in A\}$ shows that $A \times B$ is the denumerable union of the denumerable sets $\{(x, b) : x \in A\}$. Hence, $A \times B$ is denumerable. It follows from the definition of the product of two cardinals and the above corollary that $aa = a$. □

Theorem 1.17. *(i) $a + a = a$, (ii) $n + a = a$ if n is finite. (iii) $c + a = c$.*

Proof. Let λ denote a or a finite cardinal. Let $a = \overline{\overline{A}}$ and $\lambda = \overline{\overline{B}}$, where $A \cap B = \phi$. Then $\lambda + a = \overline{\overline{B \cup A}}$. Since, $B \cup A$ is the union of two countable sets of which atleast one is denumerable, so $B \cup A$ is denumerable. i.e., $\overline{\overline{B \cup A}} = a$. Thus $\lambda + a = a$. This proves (i) and (ii).

Now since $a < c$, we have $c = a + m$ for some cardinal m. So, $c + a = (a + m) + a = (a = a) + m = a + m = c$. □

Theorem 1.18. *The set of all rational numbers is denumerable.*

Proof. Let $\mathbb{Q} = \{\frac{p}{n} : n \in \{1, 2, 3, \ldots\}, p \in \{0, \pm 1, \pm 2, \ldots\}\}$. Let $\mathbb{Q}_n = \{\frac{p}{n} : p \in \{0, \pm 1, \pm 2, \ldots\}\}$, $n = 1, 2, 3, \ldots$. Then each $\mathbb{Q}_n$ is obviously denumerable and we have $\mathbb{Q} = \cup_{n=1}^{\infty}\mathbb{Q}_n$. Hence, $\mathbb{Q}$ is denumerable. □

A real number is called an algebraic number if it is a root of a polynomial equation of the form

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0 = 0,$$

where $a_n \neq 0$ and all a_k's are integers. A real number which is not algebraic is called transcendental number.
A rational number is algebraic since $x = \frac{p}{q}$ is a root of the equation $qx - p = 0$. But all algebraic numbers are not rational. For example, $x = \sqrt{2}$ is not rational but is a root of the equation $x^2 - 2 = 0$ and hence, it is algebraic.

Theorem 1.19. *The set of all algebraic numbers is denumerable.*

Proof. Let $f(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0$, where $a_n \neq 0$ and all a_k's be integers. We assume that $a_n > 0$. We associated with every polynomial $f(x)$, its height h defined by $h = n + |a_n| + |a_{n-1}| + |a_{n-2}| + \ldots + |a_1| + |a_0|$. Clearly, h is a positive integer≥ 1, there is only a finite number of polynomials with h as height, since, $n \leq h$ and every $|a_k| \leq h$. So, corresponding to any height h, there exists only a finite number of algebraic numbers [omitting complex roots]. If h runs through the set of all positive integers $\{1, 2, 3, \ldots\}$, then writing down the roots in succession omitting those which have already occurred, we get a sequence of distinct algebraic numbers. Since every polynomial has a height, all algebraic numbers appear in the sequence. This shows that the set of all algebraic numbers is denumerable. □

Theorem 1.20. *Every open interval has the power of the continuum.*

Proof. Given any open interval $(a, b) \subset \mathbb{R}$, we define $f : (a, b) \longrightarrow \mathbb{R}$ by

$$f(x) = \frac{1}{x-a} + \frac{1}{x-b}.$$

Clearly, $f'(x) = -\dfrac{1}{(x-a)^2} - \dfrac{1}{(x-b)^2}$. Therefore, f is continuous and strictly decreasing in (a, b).

Since further, $\lim\limits_{x \to a^+} f(x) = +\infty$ and $\lim\limits_{x \to b^-} f(x) = -\infty$, it follows that f is one-one and onto in $\mathbb{R}$. i.e., f is a bijective mapping. Hence $\overline{\overline{(a, b)}} = \overline{\overline{\mathbb{R}}} = c$. □

Corollary 1.4. *For any set E of real numbers containing an interval, $\overline{\overline{E}} = c$.*

Proof. Since E contains an interval, we can find an open interval $I \subset E \subset \mathbb{R}$. Then $c = \overline{\overline{I}} \leq \overline{\overline{E}} \leq \overline{\overline{\mathbb{R}}} = c$. So, by S.B. theorem, we have

$$\overline{\overline{E}} = c.$$

□

Note:- cardinal numbers of finite sets are called finite cardinals and those of infinite sets are called infinite or transfinite cardinals. So, a and c are transfinite cardinals where as $0, 1, 2, \ldots$ are finite cardinals.

Theorem 1.21. $a < c$.

Theorem 1.22. *(i) $a + a = a$, (ii) $n + a = a$, if n is finite, (iii) $c + a = c$.*

Proof. Suppose λ denotes a of a finite cardinal. Let $a = \overline{\overline{A}}$ and $\lambda = \overline{\overline{B}}$, where $A \cap B = \phi$. Then $\lambda + a = \overline{\overline{A \cup B}}$.

Since $B \cup A$ is the union of two countable sets of which at-least one is denumerable (countable infinite). So, $B \cup A$ is denumerable. i.e., $\overline{\overline{B \cup A}} = a$ $[\overline{\overline{B}} + \overline{\overline{A}} = a]$.

Thus $\lambda + a = a$. This proves (i) or (ii).

Since $a < c$, we have $c = a + m$. Therefore,

$$c + a = (a + m) + a = (m + a) + a = m(a + a) = a + m = c.$$

□

Theorem 1.23. *Suppose T be a countable subset of a set S having the power of the continuum. Then $S - T$ also has the power of the continuum.*

Proof. We have $\overline{\overline{S}} = \overline{\overline{\mathbb{R}}}$, where $\mathbb{R}$ is a set of real numbers. We may suppose that $S = \mathbb{R}$ and $T \subset \mathbb{R} = S$. Then if T is finite, we can evidently find an open interval contained in $\mathbb{R} - T$. In which case we know that $\overline{\overline{\mathbb{R} - T}} = c$. So, we assume that T is countably infinite. i.e., $\overline{\overline{T}} = a$. Now let, $A = \{x + y : x \in T, y \in T\}$. Since $A = \cup_{y \in T}\{x + y : x \in T\}$, so A is denumerable. But $\mathbb{R}$ is uncountable so $\mathbb{R} - A$ and set $B = \{\xi - x : x \in T\}$. Then clearly $\overline{\overline{B}} = \overline{\overline{T}} = a$. we also observe that $B \cap T = \phi$.

For otherwise, $\xi - x = y$ for some $x, y \in T$. or, $\xi = x + y$ for some $x, y \in T$ which contradicts that $\xi \notin A$. Thus, $B \subset \mathbb{R} - T$, so $\overline{\overline{B}} \leq \overline{\overline{\mathbb{R} - T}}$. or, $a \leq c^{'}$, where $c^{'} = \overline{\overline{\mathbb{R} - T}}$.

So, $c^{'} = a + m$ for some cardinal m. Then $c = \overline{\overline{T \cup (\mathbb{R} - T)}} = \overline{\overline{T}} + \overline{\overline{\mathbb{R} - T}} = a + c^{'} = a + a + m = a + m = c^{'}$. □

Corollary 1.5. *The set of all irrational numbers in any interval has the power of the continuum.*

Theorem 1.24. $2^a = c$.

Proof. Let $A = \{0, 1\}$, $\mathbb{N} = \{1, 2, 3, \ldots\}$, $I = [0, 1)$. Then $a = \overline{\overline{\mathbb{N}}}$, $\overline{\overline{A^{\mathbb{N}}}} = 2^a$, $\overline{\overline{I}} = c$. We note that for every $f \in A^{\mathbb{N}}$ and every $n \in \mathbb{N}$, $f(n)$ is either 0 or 1. We define $\phi : A^{\mathbb{N}} \to I$ by stipulating that for all $f \in A^{\mathbb{N}}$,

$$\phi(f) = \frac{f(1)}{3} + \frac{f(2)}{3^2} + \ldots + \frac{f(n)}{3^n} + \ldots\ldots$$

Since $f(n) = 0$ or 1 for each n, so

$$0 \leq \phi(f) \leq \frac{1}{3} + \frac{1}{3^2} + \ldots\ldots = \frac{1}{2}(< 1).$$

Thus ϕ is a well defined function from $A^{\mathbb{N}}$ into I.

To show that ϕ is injective, we suppose $\phi(f) = \phi(g)$ for some $f, g \in A^{\mathbb{N}}$.

Then $\sum_{n=1}^{\infty} \frac{f(n)}{3^n} = \sum_{n=1}^{\infty} \frac{g(n)}{3^n}$, where every $f(n)$ and $g(n)$ is 0 or 1. Hence, by the uniqueness of ternary representation of real numbers, we have $f(n) = g(n)$ for all n. Then $f = g$. Thus ϕ is injective.

Hence, $\overline{\overline{A^{\mathbb{N}}}} \leq \overline{\overline{I}}$. i.e., $2^a \leq c$.

Again, each $x \in [0,1) = I$ can be represented uniquely as

$$x = \frac{x_1}{2} + \frac{x_2}{2^2} + \ldots + \frac{x_n}{2^n} + \ldots,$$

where each x_n is 0 or 1 and $x_n = 0$ for infinitely many n.

We define $\psi : I \to A^{\mathbb{N}}$ by stipulating that for each $x \in I$, $\psi(x)$ is the function $f_x : \mathbb{N} \to A$ defined by $f_x(x) = x_n$ for all $n \in \mathbb{N}$.

We now show that ψ is injective.

Suppose for some $x, y \in I$. Then $f_x = f_y$. So, $f_x(n) = f_y(n)$ for all n. So, $x_n = y_n$ for all n. i.e., $x = y$. Thus ψ is injective.

So, $\overline{\overline{I}} \leq \overline{\overline{A^{\mathbb{N}}}}$. i.e., $c \leq 2^a$. Hence, by S.B. theorem, $2^a = c$. □

Theorem 1.25. *The family of the finite subsets of a denumerable set is denumerable and the family of the infinite subset has the power of the continuum.*

Proof. Let $\mathcal{F}$ denote the family of the finite subsets of a denumerable set A. Let the elements of A be arranged in a sequence $a_1, a_2, \ldots, a_n, \ldots$ of distinct terms. For each $E \in \mathcal{F}$ and for each positive integer n. Let us define

$$||E_n|| = \begin{cases} 1 & if\ a_n \in E \\ 0 & if a_n \notin E \end{cases} .$$

Let $\mathbb{Q}$ denote the set of all rational numbers and let $\phi : \mathcal{F} \to \mathbb{Q}$ be defined by

$$\phi(E) = \frac{||E||}{3} + \frac{||E||_2}{3^2} + \ldots + \frac{||E||_n}{3^n} + \ldots.$$

Since every E is a finite set, so all but finitely many $||E||_n$ are zero. So, ϕ is a well defined function from $\mathcal{F}$ to $\mathbb{Q}$. So if $\phi(E) = \phi(f)$, then by the uniqueness of ternary representation of real numbers, we have $||E||_n = ||F||_n$ for all n. This means that $a_n \in E$ iff $a_n \in F$ whence $E = F$. Thus ϕ is injective. So, $\overline{\overline{\mathcal{F}}} \leq \overline{\overline{\mathbb{Q}}} = a$.

On the other hand, $\mathcal{F}$ contains the denumerable family $\{\{a_1\}, \{a_2\}, \ldots, \{a_n\}, \ldots\}$ which implies that $a \leq \overline{\overline{\mathcal{F}}}$. Hence, by S.B.'s theorem $\overline{\overline{\mathcal{F}}} = a$.

Now, we know that $\overline{\overline{P(A)}} = \overline{\overline{2^A}} = 2^a = c$. Since, $\overline{\overline{\mathcal{F}}} \subset P(A)$ and $\mathcal{F}$ is denumerable, it follows that (from previous theorem) $\overline{\overline{P(A)}}/\mathcal{F} = c$. In other words, the family of the infinite subset has the power of the continuum. □

Theorem 1.26. $c^a = c$, $c^c = c$, $ca = c$, $c + c = c$, $a^a = c$, $2^c = a^c = c^c$.

Proof. We know that $a^a = a = a + a$ and $2^a = c$. So, $c^a = (2^a)^a = 2^{aa} = 2^a = c$ and $c^c = c^a c^a = c^{a+a} = c^a = c$.

Since, $1 < a < c$, so $c.1 \leq ca \leq cc$. i.e., $c \leq ca \leq c$. Here, $ca = c$.

$c + c = ca + ca = c(a + a) = ca = c$. Since, $2 < a < c$, so $2^a \leq a^a \leq c^a \leq c$. Hence, $a^a = c$. Finally, $2 < a < c \implies 2^c \leq a^c \leq c^c$. But $c^c = (2^a)^c = 2^{ac} = 2^{ca} = 2^c$. Hence, $2^c = a^c = c^c$. □

Theorem 1.27. *The set $\mathcal{F}$ of the real valued continuous functions from one interval to another interval J has the power of the continuum.*

Proof. For each $y \in J$, let f_y denote the constant function $f_y : I \to J$, defined by $f_y(x) = y$ for each $x \in I$. Then $f_y \in \mathcal{F}$ and the mapping $\phi : J \to \mathcal{F}$ defined by $\phi(y) = f_y$ is clearly injective. So, $\overline{\overline{J}} \leq \overline{\overline{\mathcal{F}}}$. i.e., $c \leq \overline{\overline{\mathcal{F}}}$. Again, let $\mathbb{Q}$ denotes the set of rationals in I. For each $f \in \overline{\overline{\mathcal{F}}}$, let $f_{\mathbb{Q}}$ denote the restriction of f to $\mathbb{Q}$. Then $f_{\mathbb{Q}} \in J^{\mathbb{Q}}$. Consider the map $\psi : \mathcal{F} \to J^{\mathbb{Q}}$, defined by $\psi(f) = f_{\mathbb{Q}}$, then ψ is injective. Suppose, $\psi(f) = \psi(g)$; $f, g \in \mathcal{F}$. Then $f_{\mathbb{Q}} = g_{\mathbb{Q}}$. i.e., $f(x) = g(x)$ for all $x \in \mathbb{Q}$.

Since, the rational numbers in I are dense in I and since the functions f and g are continuous on I, it follows that $f(x) = g(x)$ for all $x \in I$. Then $f = g$ and ψ is injective. Therefore, $\overline{\overline{\mathcal{F}}} \leq \overline{\overline{J^{\mathbb{Q}}}} = \overline{\overline{J}}^{\overline{\overline{\mathbb{Q}}}} c^a = c$. Hence, $\overline{\overline{F}} = c$.
$f(x) = g(x)$, $x \in I$, $\phi(x) = f(x) - g(x)$, $\phi(x) = 0$ at rational point. If $\phi(x) \neq 0$ at some point α by continuity of ϕ then $\phi(\alpha) > 0$.

$(\phi(\alpha) - \epsilon < \phi(x) < \phi(\alpha) + \epsilon \implies \phi(x) > \phi(\alpha) - \epsilon > 0 \implies \epsilon < \phi(\alpha).)$

This implies $\phi(x) > c - \epsilon > 0$ for all nbd of α which contains $\phi(x) = 0$ at rational point. □

Summary

In this unit, we have studied about cardinal numbers and their relevant properties and their application.

Unit 2

Course Structure

1. Cantor Set:its construction

2. And its presentation as an uncountable set of measure zero.

2 Introduction

The Cantor set, denoted by C is a subset of the interval $[0,1]$ which is left after the removal of certain specified countable (infinite) collection of open intervals from $[0,1]$. To construct the set C, we proceed as follows:

Let C_0 denote the interval $[0,1]$. Remove the following from C_0 in succession (one-by-one):

1. The open interval $(a_1,b_1)=\left(\frac{1}{3},\frac{2}{3}\right)$, the middle third of the interval C_0 leaving behind the set $C_1=[0,\frac{1}{3}]\cup[\frac{2}{3},1]$.

2. The open intervals $(a_2,b_2)=(\frac{1}{9},\frac{2}{9})$ and $(a_3,b_3)=(\frac{7}{9},\frac{8}{9})$, the middle third of the two closed intervals $[0,\frac{1}{3}]$ and $[\frac{2}{3},1]$ in C_1, leaving behind the set

$$C_2=\left[0,\frac{1}{9}\right]\cup\left[\frac{2}{9},\frac{1}{3}\right]\cup\left[\frac{2}{3},\frac{7}{9}\right]\cup\left[\frac{8}{9},1\right]$$

3. The open intervals $(a_4,b_4)=(\frac{1}{27},\frac{2}{27})$, $(a_5,b_5)=(\frac{7}{27},\frac{8}{27})$, $(a_6,b_6)=(\frac{19}{27},\frac{20}{27})$ and $(a_7,b_7)=(\frac{25}{27},\frac{26}{27})$, the middle thirds of the four closed intervals in C_2, leaving behind the set

$$\left[0,\frac{1}{27}\right]\cup\left[\frac{2}{27},\frac{1}{9}\right]\cup\left[\frac{2}{9},\frac{7}{27}\right]\cup\left[\frac{8}{27},\frac{1}{3}\right]\cup\left[\frac{2}{3},\frac{19}{27}\right]\cup\left[\frac{20,}{27},\frac{7}{9}\right]\cup\left[\frac{8}{9},\frac{25}{27}\right]\cup\left[\frac{26}{27},1\right].$$

and

4. Continue this process generating a sequence $\{C_n\}$ of sets, where C_{n+1} is obtained from C_n by removing the middle thirds of the 2^n disjoint closed intervals of which C_n is composed of. The points of the interval $[0,1]$ which are never removed in the process constitute the cantor set C. More precisely, the points common to all the sets C_n, i.e., $C=\cap_{n=1}^{\infty}C_n$.

Each of the set C_n is nonempty, closed and bounded. Also $C_{n+1}\subset C_n$ for all n. Hence the set C is nonempty closed and bounded.

Let E_n denote the set composed of all open intervals removed at the n-th stage, for instance $E_1=(a_1,b_1)=(\frac{1}{3},\frac{2}{3})$, $E_2=(a_2,b_2)\cup(a_3,b_3)=(\frac{1}{9},\frac{2}{9})\cup(\frac{7}{9},\frac{8}{9})$,

$$\begin{aligned}E_3 &= (a_4,b_4)\cup(a_5,b_5)\cup(a_6,b_6)\cup(a_7,b_7)\\ &= \left(\frac{1}{27},\frac{2}{27}\right)\cup\left(\frac{7}{27},\frac{8}{27}\right)\cup\left(\frac{19}{27},\frac{20}{27}\right)\cup\left(\frac{25}{27},\frac{26}{27}\right).\end{aligned}$$

Thus, it follows that the Cantor set C can also be expressed as the complement of the union $\cup_{n=1}^{\infty} E_n$ with respect to the set $[0,1]$. i.e., $C = [0,1] - \cup_{n=1}^{\infty} E_n$. Hence C is closed.

Note: In the following, the Cantor set C, the length of the open intervals removed at different stages are given by

$$l(E_1) = \frac{1}{3},\ l(E_2) = \frac{2}{3^2},\ l(E_3) = \frac{4}{3^3}\ \textit{and in general},\ l(E_n) = \frac{1}{3}\left(\frac{2}{3}\right)^{n-1}.$$

Thus the sum of the lengths of all open intervals removed upon the n-th stage is given by,

$$S_n = \sum_{n=1}^{n} l(E_i) = 1 - \left(\frac{2}{3}\right)^n$$

which implies $\lim_{n\to\infty} S_n = \lim_{n\to\infty}\left[1 - \left(\frac{2}{3}\right)^n\right] = 1$.

Hence, the sum of lengths of all intervals removed is the length of the original C_0 which is $[0,1]$. As such the set remaining on $[0,1]$ which in fact is the cantor set may seem so sparse as to be insignificant. Intuitively it may appear that the only points left in the Cantor set are the end points $0, 1, \frac{1}{3}, \frac{2}{3}, \frac{1}{9}, \frac{2}{9}, \frac{7}{9}, \frac{8}{9}, \ldots$ which are denumerable in number but this is wrong.

2.1 Some important results

Theorem 2.1. *The Cantor set C has power c.*

Proof. We express all the real numbers in $[0,1]$ in ternary decimals. Then every point in this interval is of the form

$$x = 0.\alpha_1\alpha_2\alpha_3\ldots\ (\alpha_k = 0\ or\ 1\ or\ 2)$$

That is

$$x = \frac{\alpha_1}{3} + \frac{\alpha_2}{3^2} + \frac{\alpha_3}{3^3} + \ldots.$$

Then each of the end intervals renamed in the construction of C admits of two such representations.

$$\frac{1}{3} = \begin{cases} 0.1222\ldots \\ 0.200\ldots \end{cases}, \frac{2}{3} = \begin{cases} 0.100\ldots \\ 0.0222\ldots \end{cases}, \frac{1}{3^2} = \begin{cases} 0.0100\ldots \\ 0.0022\ldots \end{cases}, \frac{2}{3^2} = \begin{cases} 0.01222\ldots \\ 0.02000\ldots \end{cases}$$

In general, the point of E_n has the representation

$$x = \frac{1}{3^n} + \sum_{i=1}^{n-1} \frac{b^i}{3^i} + \sum_{i=n+1}^{\infty} \frac{a_i}{3^i},$$

where $b_i = 0\ or\ 2$ and $a_i = 0, 1, 2$ but a_i's are neither all zeros nor all 2's.

So the construction of the set C at the finite steps when we remove the interval $(\frac{1}{3}, \frac{2}{3})$, actually we remove those points which lie in $(1, 2)$ i.e., those points whose ternary expansion have $\alpha_1 = 1$.

At the second step, removal of the intervals $(\frac{1}{3^2}, \frac{2}{3^2})$ and $(\frac{7}{3^2}, \frac{8}{9})$ means removal of the points which lie in $(.01, .02)$ and $(.21, .22)$ respectively i.e., those points for we have $\alpha_2 = 1$ in the ternary expansion.

We note that

$$\frac{7}{9} = .21000 = .202222,$$

$$\frac{8}{9} = .220000 = .21222.$$

Proceeding in this way, we see that C contains those points x only which can be expressed as $x = 0.\alpha_1\alpha_2\alpha_3\ldots$, where $\alpha_k = 0,\ or\ 2$.

Thus, $P_0 = C = \{0.\alpha_1\alpha_2\alpha_3\ldots,\}$ $(\alpha_k = 0\ or\ 2)$.

Now, if we express the points of $U = [0, 1]$ in binary expansions like $0.\beta_1\beta_2\beta_3\ldots$ $(\beta_k = 0,\ or\ 1)$ and establish a correspondence between P_0 and $[0, 1]$, we see that for every point of $[0, 1]$, there correspondence a point of P_0 and conversely. (Here the correspondence is an interchange between 1 and 2 in $0.\alpha_1\alpha_2\alpha_3\ldots$ and $0.\beta_1\beta_2\beta_3\ldots$). So, $P_0 = C$ has power c. □

Theorem 2.2. *The Cantor set C is perfect.*

Proof. In the construction of C, when the interval $(\frac{1}{3}, \frac{2}{3})$ is removed, we denote by $C_1 = \left[0, \frac{1}{3}\right] \cup \left[\frac{2}{3}, 1\right]$.

Similarly, let $C_2 = \left[0, \frac{1}{3^2}\right] \cup \left[\frac{2}{3^2}, \frac{3}{3^2}\right] \cup \left[\frac{6}{3^2}, , \frac{7}{3^2}\right] \cup \left[\frac{8}{3^2}, 1\right]$ and so on.
Therefore, continuing in this way, we get a sequence $C_1 \supset C_2 \supset C_3 \supset \ldots$ and $C = \cap_{i=1}^{\infty} C_n$ and clearly C is closed.

Now let $x_0 \in C$ and (α, β) be any neighbourhood of x_o. Also let I_n be that interval of C_n which contains x_0. Then for sufficiently large n, $I_n \subset [\alpha, beta]$ and x_n is an end point of I_n such that $x_n \neq x_0$.Obviously $x_n \in C$. Thus every neighbourhood (α, β) of x_0 contains a point of C which is distinct from x_0. Hence, x_0 is a limit point of C and hence C is dense in itself. C being closed and dense in itself is perfect. □

Theorem 2.3. *The Cantor set is uncountable.*

Proof. Let, if possible, the Cantor set C be countable. Then we may write $C = P_0$ as $C = \{x_1, x_2, x_3, \ldots, x_n, \ldots\}$.

Write the elements in C in ternary expansion as

$$x_1 = .{}_3a_{11}a_{12}a_{13}\ldots a_{1n}\ldots,$$

$$x_2 = .{}_3a_{21}a_{22}a_{23}\ldots a_{2n}\ldots,$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$x_n = {}_{.3}a_{n1}a_{n2}a_{n3}\ldots a_{nn}\ldots,$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots,$$

where $a_{ij} = 0$ *or* 2. Consider a sequence $\{a_n\}$, where

$$a_n = \begin{cases} 0, \ if \ a_{nn} = 2 \\ 2, \ if \ a_{nn} = 0 \end{cases} .$$

Clearly, the element

$$x = {}_{.3}a_1a_2a_3\ldots a_n\ldots$$

in C. But $x \neq x_1$ since it differs from x_n atleast in the n-th place. This is true for each n and as much x should not be in C. Hence, the result is proved by contradiction. □

Summary

In this section, we have studied about Cantor set, its properties, applications and examples.

Units 3 & 4

Course Structure

1. Functions of bounded variation : Definition and basic properties
2. Lipschitz condition, Jordan decomposition
3. Nature of points of discontinuity, Nature of points of non-differentiability
4. Convergence in variation (Helly's First theorem)

3 Introduction

A function of bounded variation, also known as BV function, is a real-valued function whose total variation is bounded (finite): the graph of a function having this property is well behaved in a precise sense. For a continuous function of a single variable, being of bounded variation means that the distance along the direction of the y-axis, neglecting the contribution of motion along x-axis, travelled by a point moving along the graph has a finite value. We now move on to the various definitions and preliminary ideas related to the chapter.

3.1 Functions of Bounded Variation

Definition 3.1. *Let S be a non-empty subset of $\mathbb{R}$. A real number u is said to be an upper bound of S if $x \in S \Rightarrow x \leq u$. A real number l is said to be a lower bound of S if $x \in S \Rightarrow x \geq l$.*

A non-empty set $S \subseteq \mathbb{R}$ is said to be bounded above if there exists a real number u such that $x \leq u$, for all $x \in S$. A non-empty S is said to be bounded below if there exists a real number l such that $x \geq l$, for all $x \in S$.

A non-empty S is said to be a bounded set if S is bounded above as well as bounded below.

Definition 3.2. *A real number M is said to be a least upper bound(or supremum) of a non-empty S (lub S or Sup S) if it has the following two properties.*
(i) M is an upper bound of S i.e., $x \leq M$, for all $x \in S$.
(ii) for each $\epsilon > 0$, there exists an element $y(\epsilon)$ in S such that $M - \epsilon < y \leq M$.

A real number m is said to be a greatest lower bound(or infimum) of a non-empty S (glb S or inf S) if it has the following two properties.

(i) m is a lower bound of S i.e., $x \geq m$, for all $x \in S$.

(ii) for each $\epsilon > 0$, there exists an element $y(\epsilon)$ in S such that $m \leq y < m + \epsilon$

We now state some fundamental properties of the set $\mathbb{R}$.

(1) Algebraic properties of $\mathbb{R}$.

(2) Order properties of $\mathbb{R}$.

(3) Completeness property of $\mathbb{R}$.

Every non-empty subset of $\mathbb{R}$ that is bounded above has a least upper bound (or a supremum) or every non-empty subset of $\mathbb{R}$ that is bounded below has a greatest lower bound (or an infimum).

(4) Archimedean property of $\mathbb{R}$.

If $x,\ y \in \mathbb{R}$ and $y > 0$, then there exists a natural number n such that $ny > x$.

(5) Density property of $\mathbb{R}$.

If $x,\ y$ are real numbers with $x < y$, then there exists a rational number r such that $x < r < y$. If $x,\ y$ are real numbers with $x < y$, then there exists an irrational number s such that $x < s < y$.

Definition 3.3. *Let $[a, b]$ be a closed and bounded interval. A partition P of $[a, b]$ is a finite ordered set $\{x_0, x_1, \ldots, x_n\}$ of points of $[a, b]$ such that $a = x_0 < x_1 < \ldots < x_n = b$.*

The family of all partitions of $[a, b]$ is denoted by $P[a, b]$ and the partition $P = \{x_0, x_1, \ldots, x_n\}$ is a member of $P[a, b]$.

For example $P = \{0, \frac{1}{2}, 1\}$ is a partition of $[0, 1]$, $Q = \{0, \frac{1}{8}, \frac{1}{2}, \frac{7}{8}, 1\}$ is another partition of $[0, 1]$.

The partition $P = \{x_0, x_1, \ldots, x_n\}$ of $[a, b]$ divides the interval $[a, b]$ into non-overlapping subintervals $[a, x_1], [x_1, x_2], \ldots, [x_{n-1}, b]$.

Definition 3.4. *Let $[a, b]$ be a closed and bounded interval and $f : [a, b] \to \mathbb{R}$ be a function. Let $P = \{x_0, x_1, \ldots, x_n\}$ be a partition of $[a, b]$. Let us consider the sum*

$$V(P, f) = \sum_{i=1}^{n} |f(x_i) - f(x_{i-1})|.$$

For different partitions $P \in P[a, b]$, $V(P, f)$ gives a set of non-negative real numbers. If the set

$$\{V(P, f) : P \in P[a, b]\}$$

is bounded above, then f is said to be a function of bounded variation (or a BV-function) on $[a, b]$.
The supremum of the set $\{V(P, f) : P \in P[a, b]\}$ is said to be the total variation of f on $[a, b]$ and is denoted by $V(a, b; f)$ i.e.,

$$V(a, b; f) = \sup\{V(P, f) : P \in P[a, b]\}.$$

Thus the function f is said to be of bounded variation on $[a, b]$ if total variation is finite i.e., $V(a, b; f) < +\infty$.

Note 3.1. *Since each sum $V(P,f) \geq 0$, it follows that $V(a,b;f) = 0$ iff f is a constant function on $[a,b]$.*

Note 3.2. *$V(P,f) \leq V(a,b;f)$, for all $P \in P[a,b]$.*

Example 3.1. *Let $k \in \mathbb{R}$ and $f(x) = k$, for all $x \in [a,b]$.*
Let $P = \{x_0, x_1, \ldots, x_n\}$, where $a = x_0 < x_1 < \ldots < x_n = b$ be a partition of $[a,b]$. Then

$$\begin{aligned} V(P,f) &= \sum_{i=1}^{n} |f(x_i) - f(x_{i-1})| \\ &= \sum_{i=1}^{n} |k - k| = 0. \end{aligned}$$

Consequently $V(a,b;f) = \sup\{V(P,f) : P \in P[a,b]\} = 0$. Therefore f is a function of bounded variation on $[a,b]$.

Example 3.2. *Let $f(x) = x$, $x \in [a,b]$.*
Let $P = \{x_0, x_1, \ldots, x_n\}$, where $a = x_0 < x_1 < \ldots < x_n = b$ be a partition of $[a,b]$. Then

$$\begin{aligned} V(P,f) &= \sum_{i=1}^{n} |f(x_i) - f(x_{i-1})| \\ &= \sum_{i=1}^{n} |x_i - x_{i-1}| \\ &= \sum_{i=1}^{n} (x_i - x_{i-1}) \\ &= (x_1 - x_0) + (x_2 - x_1) + \ldots + (x_n - x_{n-1}) \\ &= x_n - x_0 \\ &= b - a. \end{aligned}$$

Consequently $V(a,b;f) = \sup\{V(P,f) : P \in P[a,b]\} = b - a$. Therefore f is a function of bounded variation on $[a,b]$.

Example 3.3. *Let $f(x) = \sin x$, $x \in [a,b]$.*
Let $P = \{x_0, x_1, \ldots, x_n\}$, where $a = x_0 < x_1 < \ldots < x_n = b$ be a partition of $[a,b]$. Then

$$\begin{aligned} V(P,f) &= \sum_{i=1}^{n} |f(x_i) - f(x_{i-1})| \\ &= |\sin x_1 - \sin x_0| + |\sin x_2 - \sin x_1| + \ldots + |\sin x_n - \sin x_{n-1}|. \end{aligned}$$

By Mean Value Theorem we have

$$|f(x_r) - f(x_{r-1})| = |x_r - x_{r-1}||\cos \xi_r|,$$

for some ξ_r *satisfying* $x_{r-1} < \xi_r < x_r$. *This holds for* $r = 1, 2, \ldots, n$. *Therefore* $|f(x_r) - f(x_{r-1})| \leq |x_r - x_{r-1}|$, *since* $|\cos \xi_r| \leq 1$. *Now*

$$\begin{aligned} V(P, f) &\leq |x_1 - x_0| + |x_2 - x_1| + \ldots + |x_n - x_{n-1}| \\ &= (x_1 - x_0) + (x_2 - x_1) + \ldots + (x_n - x_{n-1}) \\ &= x_n - x_0 \\ &= b - a. \end{aligned}$$

Consequently $V(a, b; f) = \sup\{V(P, f) : P \in P[a, b]\} \leq b - a$. *i.e.,* $V(a, b; f) < +\infty$. *Therefore* f *is a function of bounded variation on* $[a, b]$.

Theorem 3.1. *Let* $[a, b] \subset \mathbb{R}$ *and* $f : [a, b] \to \mathbb{R}$ *be a function of bounded variation on* $[a, b]$. *Then* f *is bounded on* $[a, b]$.

Proof. Let $f(x)$ is of bounded variation on $[a, b]$ on all divisions of $[a, b]$. Then there is some positive real number M such that $V_f[a, b] \leq M$. If $x \in [a, b]$, we have $a \leq x \leq b$ and

$$|f(x) - f(a)| + |f(b) - f(x)| \leq V_f[a, b] \leq M.$$

i.e.,

$$|f(x)| - |f(a)| \leq |f(x) - f(a)| \leq V_f[a, b] \leq M.$$

i.e.,

$$|f(x)| - |f(a)| \leq M.$$

i.e.,

$$|f(x)| \leq M + |f(a)|$$

showing that f is bounded on $[a, b]$. □

Following example shows that the converse of the above theorem is not always true.

Example 3.4. *Let* $f : [0, 1] \to \mathbb{R}$ *be defined by*

$$\begin{aligned} f(x) &= 1, \ if \ x \ is \ rational, \\ &= 0, \ if \ x \ is \ irrational. \end{aligned}$$

Let $P = \{x_0, x_1, \ldots, x_{2n}\}$ *be a partition of* $[0, 1]$ *such that* $x_0, x_2, \ldots, x_{2n}$ *are all rational and* $x_1, x_3, \ldots, x_{2n-1}$ *are all irrational. Then*

$$\begin{aligned} V(P, f) &= \sum_{i=1}^{2n} |f(x_i) - f(x_{i-1})| \\ &= |f(x_1) - f(x_0)| + |f(x_2) - f(x_1)| + \ldots + |f(x_{2n}) - f(x_{2n-1})| \\ &= |1 - 0| + |1 - 0| + \ldots + |1 - 0| \ \ (2n \ - times) \\ &= 2n. \end{aligned}$$

Clearly the set $\{V(P, f) : P \in P[0, 1]\}$ *is not bounded above and therefore* f *is not a function of bounded variation on* $[0, 1]$.

Note 3.3. *The function f is bounded but not a function of bounded variation.*

Theorem 3.2. *If a function $f(x)$ be monotonic on $[a,b]$, then it is of function of bounded variation $[a,b]$.*

Proof. Let f be monotonic increasing on $[a,b]$.

Let $P = \{a = x_0, x_1, \ldots, x_n = b\}$ be any partition of $[a,b]$. Then

$$\begin{aligned} V(P,f) &= \sum_{i=1}^{n} |f(x_i) - f(x_{i-1})| \\ &= \sum_{i=1}^{n} [f(x_i) - f(x_{i-1})] \\ &= [f(x_1) - f(x_0)] + [f(x_2) - f(x_1)] + \ldots + [f(x_n) - f(x_{n-1})] \\ &= f(x_n) - f(x_0) \\ &= f(b) - f(a). \end{aligned}$$

Therefore $V(a,b;f) = \sup\{V(P,f) : P \in P[a,b]\} = f(b) - f(a)$, a finite number.

Thus a monotonic increasing bounded function is of bounded variation on $[a,b]$.

Similarly, it may be shown that a monotonic decreasing bounded function is of bounded variation with $V(a,b;f) = f(a) - f(b)$. □

Note 3.4. *The function $f(x) = [x]$, where $[x]$ denotes the greatest integer not greater than x is a function of bounded variation on $[0,2]$.*

Theorem 3.3. *A function of bounded variation is necessarily bounded.*

Proof. Let $f : [a,b] \to \mathbb{R}$ be a function of bounded variation on $[a,b]$. Since f is a function of bounded variation on $[a,b]$, it follows that $V(a,b;f)$ is finite. Let $V(a,b;f) = M$, where M is a non-negative real number.

Let $x \in (a,b)$. Let $P_0 = \{a, x, b\}$ be a partition of $[a,b]$. Then

$$V(P_0,f) \leq V(a,b;f) = M$$

$$\Rightarrow |f(x) - f(a)| + |f(b) - f(x)| \leq M$$

$$\Rightarrow |f(x) - f(a)| \leq M.$$

Therefore

$$\begin{aligned} |f(x)| &= |f(x) - f(a) + f(a)| \\ &\leq |f(x) - f(a)| + |f(a)| \\ &\leq |f(a)| + M. \end{aligned}$$

If however $x = a$, then $|f(x)| = |f(a)| \leq |f(a)| + M$ and also if $x = b$, then

$$V(P_0,f) = |f(x) - f(a)| + |f(b) - f(b)| = |f(x) - f(a)|$$

and so $|f(x) - f(a)| \leq M$. This implies that $|f(x)| \leq |f(a)| + M$.

Thus $|f(x)| \leq |f(a)| + M$, $\forall$ x$\in [a,b]$ and so f is bounded on $[a,b]$. □

Remark 3.1. *The converse of the theorem is not true. A function f bounded on $[a,b]$ may not be a function of bounded variation on $[a,b]$.*

For example, let $f:[0,1]\to\mathbb{R}$ be defined by

$$\begin{aligned} f(x) &= x\sin\frac{\pi}{x}, \quad x\neq 0 \\ &= 0, \quad x=0. \end{aligned}$$

Then f is bounded on $[0,1]$, since $|f(x)|\leq 1,\ \forall\ x\in[0,1]$.

Let us choose the partition $P=\{0,\frac{2}{2n+1},\frac{2}{2n-1},\ldots,\frac{2}{5},\frac{2}{3},1\}$. Therefore

$$\begin{aligned} V(P,f) &= \sum_{i=1}^{n}|f(x_i)-f(x_{i-1})| \\ &= \Big|f\Big(\frac{2}{2n+1}\Big)-f(0)\Big|+\Big|f\Big(\frac{2}{2n-1}\Big)-f\Big(\frac{2}{2n+1}\Big)\Big|+\ldots+\Big|f\Big(\frac{2}{3}\Big)-f\Big(\frac{2}{5}\Big)\Big| \\ &\quad +\Big|f(1)-f\Big(\frac{2}{3}\Big)\Big| \\ &= \Big|f(1)-f\Big(\frac{2}{3}\Big)\Big|+\Big|f\Big(\frac{2}{3}\Big)-f\Big(\frac{2}{5}\Big)\Big|+\ldots+\Big|f\Big(\frac{2}{2n-1}\Big)-f\Big(\frac{2}{2n+1}\Big)\Big| \\ &\quad +\Big|f\Big(\frac{2}{2n+1}\Big)-f(0)\Big| \\ &= \frac{2}{3}+\Big(\frac{2}{3}+\frac{2}{5}\Big)+\Big(\frac{2}{5}+\frac{2}{7}\Big)+\ldots+\Big(\frac{2}{2n-1}+\frac{2}{2n+1}\Big)+\frac{2}{2n+1} \\ &= 4\Big[\frac{1}{3}+\frac{1}{5}+\frac{1}{7}+\ldots+\frac{1}{2n+1}\Big]. \end{aligned}$$

Since the infinite series $\frac{1}{3}+\frac{1}{5}+\frac{1}{7}+\ldots$ is not convergent, its partial sums sequence $\{S_n\}$, where $S_n=\frac{1}{3}+\frac{1}{5}+\frac{1}{7}+\ldots+\frac{1}{2n+1}$ is not bounded above.

Thus $V(P,f)$ can be made arbitrarily large by taking n sufficiently large. Consequently $V(0,1;f)\to\infty$ and so f is not of bounded variation on $[0,1]$.

Remark 3.2. *A continuous function f defined on a closed and bounded interval $[a,b]$ may not be a function of bounded variation on $[a,b]$.*

For example let,

$$\begin{aligned} f(x) &= x\sin\frac{\pi}{x}, \quad x\neq 0 \\ &= 0, \qquad\quad x=0. \end{aligned}$$

Then f is continuous on $[0,1]$. But f is not a function of bounded variation on $[0,1]$.

Theorem 3.4. *If the derivative f' exists and is bounded on $[a,b]$, then the function f is of bounded variation on $[a,b]$.*

Proof. Since f' is bounded on $[a,b]$, therefore there exists $k > 0$ such that $|f'(x)| \leq k$, $\forall\ x \in [a,b]$.

Let $P = \{a = x_0, x_1, \ldots, x_n = b\}$ be any partition of $[a,b]$. Then

$$V(P,f) = \sum_{i=1}^{n} |f(x_i) - f(x_{i-1})|.$$

By Mean Value Theorem we have

$$f(x_i) - f(x_{i-1}) = (x_i - x_{i-1}) f'(\xi_i)$$

for some ξ_i satisfying $x_{i-1} < \xi < x_i$. Therefore

$$|f(x_i) - f(x_{i-1})| = |x_i - x_{i-1}||f'(\xi_i)| \leq k(x_i - x_{i-1}),$$

for all $i = 1, 2, \ldots, n$. This implies that

$$V(P,f) \leq k \sum_{i=1}^{n} (x_i - x_{i-1}) = k(b-a).$$

Consequently $V(a,b;f) \leq k(b-a)$ and so f is a function of bounded variation on $[a,b]$. □

Remark 3.3. *Boundedness of f' is sufficient condition.*

Boundedness of f' is not necessary for the function f to be of bounded variation on $[a,b]$. For example let $f(x) = \sqrt{x}$, $x \in [0,1]$. Then f is a monotonic increasing function on $[0,1]$ and therefore it is a function of bounded variation on $[0,1]$. But f' is not bounded on $[0,1]$.

Example 3.5. *A function $f : [0,1] \to \mathbb{R}$ is defined by,*

$$\begin{aligned} f(x) &= x^2 \cos\frac{1}{x}, \quad if \ \ x \neq 0 \\ &= 0, \qquad\qquad if \ \ x = 0. \end{aligned}$$

we have,

$$\begin{aligned} f'(x) &= \sin\frac{1}{x} + 2x\cos\frac{1}{x}, \quad if \ \ x \neq 0 \\ &= 0, \qquad\qquad\qquad\quad if \ \ x = 0. \end{aligned}$$

so that $|f'(x)| \leq 3$, $\forall x \in [0,1]$. Hence $f(x)$ is of bounded variation on $[0,1]$.

Definition 3.5. *A function $f : [a,b] \to \mathbb{R}$ is said to satisfy a Lipschitz condition on $[a,b]$ if there exists a positive real number M such that*

$$|f(x_1) - f(x_2)| \leq M|x_1 - x_2|$$

for any two points $x_1,\ x_2$ in $[a,b]$. In this case f is also said to be a Lipschitz function on $[a,b]$.

Note 3.5. *M must be independent upon the choice of x_1 and x_2.*

Theorem 3.5. *Let $f : [a, b] \to \mathbb{R}$ be a Lipschitz function on $[a, b]$. Then f is a function of bounded variation on $[a, b]$.*

Proof. Let $P = \{a = x_0, x_1, \ldots, x_n = b\}$ be any partition of $[a, b]$. Since f is a Lipschitz function on $[a, b]$, there is a positive real number M such that

$$|f(x_r) - f(x_{r-1})| \leq M|x_r - x_{r-1}|, \quad for \;\; r = 1, 2, \ldots, n.$$

Therefore

$$\begin{aligned} V(P, f) = \sum_{r=1}^{n} |f(x_r) - f(x_{r-1})| &\leq M \sum_{r=1}^{n} |x_r - x_{r-1}| \\ &= M \sum_{r=1}^{n} (x_r - x_{r-1}) = M(b - a). \end{aligned}$$

Consequently $V(a, b; f) \leq M(b - a)$ and so f is a function of bounded variation on $[a, b]$. □

Remark 3.4. *The converse of the theorem is not true. A function f of bounded variation on $[a, b]$ may not be a Lipschitz function on $[a, b]$. For example let $f : [0, 1] \to \mathbb{R}$ be defined by $f(x) = \sqrt{x}$, $x \in [0, 1]$.*

Then f being a monotonic increasing function on $[0, 1]$ is a function of bounded variation on $[0, 1]$. But f is not a Lipschitz function on $[0, 1]$, because if $x_1 = 0$, no positive real number M can be found to satisfy the condition

$$|f(x_2) - f(x_1)| \leq M|x_2 - x_1|, \forall x_2 \in (0, 1].$$

3.2 Some properties of functions of bounded variation

Theorem 3.6. *The sum(difference) of two functions of bounded variation is also of bounded variation.*

Proof. Let f and g be two functions of bounded variation on [a,b].

For any partition $P = \{a = x_0, x_1, \ldots, x_n = b\}$ of $[a, b]$ we have

$$\begin{aligned} V(P, f + g) &= \sum_{i=1}^{n} |(f + g)(x_i) - (f + g)(x_{i-1})| \\ &= \sum_{i=1}^{n} |\{f(x_i) - f(x_{i-1})\} + \{g(x_i) - g(x_{i-1})\}| \\ &\leq \sum_{i=1}^{n} |f(x_i) - f(x_{i-1})| + \sum_{i=1}^{n} |g(x_i) - g(x_{i-1})| \\ &= V(P, f) + V(P, g) \\ &\leq V(a, b; f) + V(a, b; g), \end{aligned}$$

i.e.,

$$V(P, f+g) \leq V(a,b;f) + V(a,b;g), \quad \forall P \in P[a,b].$$

Therefore,

$$\begin{aligned} V(a,b;f+g) &= \sup\{V(P,f+g) : P \in P[a,b]\} \\ &\leq V(a,b;f) + V(a,b;g) \end{aligned}$$

and so $f+g$ is of bounded variation.

Similarly it may be shown that $f-g$ is of bounded variation. □

Theorem 3.7. *The product of two functions of bounded variation is also of bounded variation.*

Proof. Let f and g be two functions of bounded variation on $[a,b]$. Clearly f and g are bounded and therefore a positive number K exists such that $|f(x)| \leq K$, $|g(x)| \leq K$, $\forall\ x \in [a,b]$.
For any partition $P = \{a = x_0, x_1, \ldots, x_n = b\}$, we have

$$\begin{aligned} V(P,fg) &= \sum_{i=1}^{n} |(fg)(x_i) - (fg)(x_{i-1})| \\ &= \sum_{i=1}^{n} |f(x_i)g(x_i) - f(x_{i-1})g(x_{i-1})| \\ &= \sum_{i=1}^{n} |f(x_i)\{g(x_i) - g(x_{i-1})\} + g(x_{i-1})\{f(x_i) - f(x_{i-1})\}| \\ &\leq \sum_{i=1}^{n} |f(x_i)||g(x_i) - g(x_{i-1})| + \sum_{i=1}^{n} |g(x_{i-1})||f(x_i) - f(x_{i-1})| \\ &\leq K\sum_{i=1}^{n} |g(x_i) - g(x_{i-1})| + K\sum_{i=1}^{n} |f(x_i) - f(x_{i-1})| \\ &= KV(P,g) + KV(P,f) \\ &\leq K[V(a,b;f) + V(a,b;g)], \end{aligned}$$

i.e.,

$$V(P,fg) \leq K[V(a,b;f) + V(a,b;g)], \forall\ \ P \in P[a,b].$$

Therefore

$$V(a,b;fg) = \sup\{V(P,fg) : P \in P[a,b]\} \leq K[V(a,b;f) + V(a,b;f)]$$

and so fg is a function of bounded variation on $[a,b]$. □

Theorem 3.8. *If f is a function of bounded variation on $[a,b]$ and if there exists a positive number K such that $|f(x)| \geq K$, $\forall\ x \in [a,b]$, then $\frac{1}{f}$ is also of bounded variation on $[a,b]$.*

Proof. For any partition $P = \{a = x_0, x_1, \ldots, x_n = b\}$, we have

$$\begin{aligned} V\Big(P, \frac{1}{f}\Big) &= \sum_{i=1}^{n} \Big|\frac{1}{f}(x_i) - \frac{1}{f}(x_{i-1})\Big| \\ &= \sum_{i=1}^{n} \Big|\frac{f(x_{i-1}) - f(x_i)}{f(x_i)f(x_{i-1})}\Big| \\ &\leq \sum_{i=1}^{n} \frac{1}{|f(x_i)|}\frac{1}{|f(x_{i-1})|}|f(x_i) - f(x_{i-1})| \\ &\leq \frac{1}{K^2}\sum_{i=1}^{n} |f(x_i) - f(x_{i-1})| \\ &= \frac{1}{K^2}V(P,f) \leq \frac{1}{K^2}V(a,b;f). \end{aligned}$$

Therefore,

$$V\left(P, \frac{1}{f}\right) \leq \frac{1}{K^2}V(a,b;f), \forall\ P \in P[a,b].$$

Hence

$$V\left(a,b;\frac{1}{f}\right) = \sup\{V\left(P,\frac{1}{f}\right) : P \in P[a,b]\} \leq \frac{1}{K^2}V(a,b;f)$$

and so $\dfrac{1}{f}$ is of bounded variation on $[a,b]$. □

Theorem 3.9. *If f is a function of bounded variation on $[a,b]$, then it is also a of bounded variation on $[a,c]$ and $[c,b]$, where c is a point of $[a,b]$ and conversely. Also $V(a,b;f) = V(a,c;f) + V(c,b;f)$.*

Theorem 3.10. *Let $f : [a,b] \to \mathbb{R}$ be a function of bounded variation on $[a,b]$. Then $|f|$ is a function of bounded variation on $[a,b]$.*

Proof. For any partition $P = \{a = x_0, x_1, \ldots, x_n\}$ we have

$$V(P,f) = \sum_{i=1}^{n} |f(x_i) - f(x_{i-1})|$$

and

$$V(P,|f|) = \sum_{i=1}^{n} \Big||f(x_i)| - |f(x_{i-1})|\Big|.$$

But

$$\Big||f(x_i)| - |f(x_{i-1})|\Big| \leq |f(x_i) - f(x_{i-1})|$$

and so

$$\begin{aligned} V(P,|f|) &= \sum_{i=1}^{n} \Big|f(x_i)| - |f(x_{i-1})|\Big| \\ &\leq \sum_{i=1}^{n} |f(x_i) - f(x_{i-1})| \\ &= V(P,f) \leq V(a,b;f), \end{aligned}$$

i.e.,

$$V(P,|f|) \leq V(a,b;f), \ \ \forall\ P \in P[a,b].$$

Therefore,

$$V(a,b;|f|) = \sup\{V(P,|f|) : P \in P[a,b]\} \leq V(a,b;f)$$

and so $|f|$ is a function of bounded variation on $[a,b]$. □

Theorem 3.11. *Let f be defined on $[a,b]$. If $f \in V[a,c]$ for any $0 < c < b$ and if there exists a number M such that $V_f[a,c] \leq M$ for any $a < c < b$, then $f \in V[a,b]$.*

Proof. Let $P = \{x_0, x_1, \ldots, x_n\}$ be any partition of $[a,b]$. Set $P' = \{x_0, x_1, \ldots, x_{n-1}\}$. Then P' is a partition of $[a, x_{n-1}]$ and hence $V_f[a, x_{n-1}] \leq M$, we have

$$\begin{aligned} V(P,f) &= V(P',f) + |f(a) - f(x_{n-1})| \\ &\leq V_f[a,x_{n-1}] + |f(a) - f(x_{n-1})| \\ &\leq M + |f(b) - f(c)| + |f(c) - f(x_{n-1})| \\ &\leq M + |f(b) - f(c)| + V_f[a,x_{n-1}] \\ &\leq 2M + |f(b) - f(c)|. \end{aligned}$$

Thus $V(P,f)$ is bounded by $2M + |f(b) - f(c)|$ and therefore $f \in V[a,b]$. □

3.3 Variation Function

Let f be a function of bounded variation on $[a,b]$ and x is a point of $[a,b]$. Then the total variation of f, $V(a,x;f)$ on $[a,x]$, which clearly is a function of x, is called the total variation function or simply the variation function of f and is denoted by $V_f(x)$. Thus,

$$V_f(x) = V(a,x;f), x \in [a,b].$$

If x_1, x_2 are two points of $[a, b]$ such that $x_2 > x_1$, then

$$\begin{aligned} 0 \leq |f(x_2) - f(x_1)| &\leq V(x_2, x_1; f) \\ &= V(a, x_2; f) - V(a, x_1; f) \\ &= V_f(x_2) - V_f(x_1). \end{aligned}$$

Therefore $V_f(x_2) \geq V_f(x_1)$, i.e., $V_f(x)$ is a monotonic increasing function on $[a, b]$.

Note 3.6. *If f is of bounded variation on $[a, b]$, then $V_f \pm f$ is a monotonic increasing function on $[a, b]$.*

Note 3.7. *The variation function of a function f of bounded variation is continuous iff f is a continuous function.*

Theorem 3.12. *A function f is of bounded variation on $[a, b]$ iff f is the difference of two monotone functions on $[a, b]$.*

Theorem 3.13. *If f is of bounded variation on $[a, b]$, then $f'(x)$ exists for at most all x in $[a, b]$.*

Theorem 3.14. *Let $f : [a, b] \to \mathbb{R}$ be a function of bounded variation on $[a, b]$. Then f can have only discontinuities of the first kind and the points of discontinuity of f form a countable set.*

Proof. Since f is a function of bounded variation on $[a, b]$, f can be expressed as $f(x) = g(x) - h(x)$, where g and h are monotone functions on $[a, b]$.

A monotone function can have only discontinuities of the first kind and the set of points of discontinuities is a countable set.

Let $c \in (a, b)$. Then each of $g(c + 0), g(c - 0), h(c + 0), h(c - 0)$ exists and therefore each of $f(c + 0), f(c - 0)$ exists. Also each of $f(a + 0), f(b - 0)$ exists. If follows that f can have only discontinuities of the first kind on $[a, b]$.

Let E_1, E_2 be respectively the sets of points of discontinuities of g and h. Then $E_1 \cup E_2$ is the set of points of discontinuity of f. But both E_1 and E_2 are countable sets. Therefore the set $E_1 \cup E_2$ is countable. □

Theorem 3.15. *The set of points of discontinuity of a function which is monotonic in an interval $[a, b]$ is at most denumerable.*

Proof. We shall prove the theorem for a monotonic increasing function $f(x)$ defined on $[a, b]$. If $c \in [a, b]$ then it follows that $f(c - 0) \leq f(c) \leq f(c + 0)$.

Thus a monotonic function is discontinuous at a point c iff. $f(c+0)-f(c-0) > 0$, where we agree that $f(a - 0) = f(a)$ and $f(b + 0) = f(b)$. We suppose that

$$a = x_0 < x_1 < x_2 < \ldots < x_n < x_{n+1} = b$$

and choose points $y_0, y_1, \ldots, y_n$ such that $x_k < y_k < x_{k+1}, \quad k = 0, 1, \ldots, n$.

Then since

$$f(x_k+0) \le f(y_k) \quad (for \ \ a \ \ point \ \ k=0)$$
$$f(x_k-0) \ge f(y_{k-1}) \quad (for \ \ b \ \ point \ \ k=n+1),$$

we have $f(x_k+0)-f(x_k-0) \le f(y_k)-f(y_{k-1}) \quad$ for $k=1,2,\ldots,n$

and for

$$\begin{cases} f(a+0)-f(a) \le f(y_0)-f(a) \ \ (k=0) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ f(b)-f(b-0) \le f(b)-f(y_n) \ \ (k=n+1) \end{cases}$$

Now adding the above inequalities we obtain

$$[f(a+c)-f(0)]+[f(b)-f(b-0)]+\sum_{k=1}^{n}[f(x_k+0)-f(x_k-0)]$$
$$\le \quad f(b)-f(a)$$

Therefore it follows that if the interior points $x_1, x_2, \ldots, x_n$ are points of discontinuity with

$$f(x_k+0)-f(x_k-0) \ge \sigma = \frac{1}{p} \ \ (say),$$

Then $n\frac{1}{p} \le f(b)-f(a)$

i.e.,$n \le p\{f(b)-f(a)\}$

Thus n can not be arbitrary large.So number of points of discontinuity such that

$f(x+0)-f(x-0) \ge \frac{1}{p}$ is finite.

If N be the set of points of discontinuity of $f(x)$ on $[a,b]$ and $N_p(p=1,2,\ldots)$ be the set of points of discontinuity where $f(x+0)-f(x-0) \ge \frac{1}{p}$ then obviously

$N = N_1 \cup N_2 \cup N_3 \cup \ldots$

where each N_p is finite . So N is denumerable.

This prove the theorem. □

Theorem 3.16. *A function of b.v. has at most a denumerable number of points of discontinuity.*

Proof. From the penultimate theorem we have $f(x)=\phi(x)-\psi(x)$ where $\phi(x)$ and $\psi(x)$ are increasing. Also since $\phi(x)$ and $\psi(x)$ has at most denumerable number of points of discontinuity. The function $f(x)$ has at most denumerable no. of points of discontinuity. □

Theorem 3.17. *Let $f(x)$ be a function of b.v. on $[a,b]$. If $f(x)$ is continuous at the point $x_0 \in [a,b]$. Thus the function $F(x)=V_f[a,x]$ is also continuous at x_0. Conversely if $F(x)$ is continuous at x_0 the function $f(x)$ is continuous at x_0.*

Proof. Let $f(x)$ be continuous and $x_0 < b$.Let $\epsilon(> 0)$ be arbitrary. Then there exists a $\delta(> 0)$ such that

$|f(x) - f(x_0)| < \frac{\epsilon}{2}$ whenever $|x - x_0| < \delta$.

We consider a division of $[x_0, b]$ given by $x_0 < x_1 < x_2 < \ldots < x_n = b$ such that $\sum_{p=0}^{n-1} |f(x_{p+1}) - f(x_p)| > V_f[x_0, b] - \frac{\epsilon}{2}$.

Since the sum in the above inequality increases on introduction of new points of divisions, we may assume that

$|f(x_1) - f(x_0)| < \frac{\epsilon}{2}$ whenever $(x - x_0) < \delta$. Therefore,

$$\begin{aligned} V_f[x_0, b] = \frac{\epsilon}{2} &< |f(x_1) - f(x_0)| + \sum_{p=1}^{n-1} |f(x_{p+1}) - f(x_p)| \\ &< \frac{\epsilon}{2} + \sum_{p=1}^{n-1} |f(x_{p+1}) - f(x_p)| \\ &< \frac{\epsilon}{2} + V_f[x_1, b] \end{aligned}$$

So, $V_f[x_0, b] - V_f[x_1, b] < \epsilon$ and we have $V_f[x_0, x_1] < \epsilon$ $\quad \Big[\; x_0 < x_1 < b \; , \;\; V_f[x_0, b] = V_f[x_0, x_1] + V_f[x_1, b] \; \Big]$

Consequently, for $0 < (x_1 - x_0) < \delta$ we have

$|F(x_1) - F(x_0)| < \epsilon$ $\quad \Big[\; a < x_0 < x_1 \; , \;\; V_f[a, x_1] = V_f[a, x_0] + V_f[x_0, x_1] \; \Big]$.

This implies that $\lim\limits_{x \to x_0^+} F(x) = F(x_0)$.

Similarly, we can show that if $x_0 > a$, then $\lim\limits_{x \to x_0^-} F(x) = F(x_0)$.

This proves the continuity of $F(x)$ at x_0.

Conversely, let $F(x)$ be continuous at x_0.Then for any $\epsilon(> 0)$ $\exists$ a $\delta > 0$ such that

$|F(x) - F(x_0)| < \epsilon$ for $|x - x_0| < \delta$.

Also $|f(x) - f(x_0)| \leq V_f[x_0, x] = F(x) - F(x_0)$ if $x > x_0$

$|f(x) - f(x_0)| \leq V_f[x, x_0] = F(x_0) - F(x)$ if $x < x_0$

Hence, it follows that

$|f(x) - f(x_0)| \leq |F(x) - F(x_0)| < \epsilon$ whenever $|x - x_0| < \delta$.

Hence, $f(x)$ is continuous at x_0. □

Corollary 3.1. *A continuous function $f(x)$ of b.v. can be expressed as the difference of two continuous increasing functions.*

Proof. In fact $f(x)$ can be expressed as

$$f(x) = F(x) - G(x),$$

where both $F(x)$ and $G(x)$ are increasing. Since $f(x)$ is continuous, by the above theorem $F(x)$ is continuous and $G(x) = F(x) - f(x)$ being the difference of two continuous function is continuous. □

Corollary 3.2. *If a function $f(x)$ defined on the interval $[a,b]$ is of b.v. on $[a,b]$. Then $f(x)$ is measurable on $[a,b]$.*

Proof. For $f(x)$ has at most denumerable number of points of discontinuity that is $f(x)$ is continuous almost everywhere. But a function which is continuous a.e. is measurable. □

3.4 Some Examples on Bounded variation

Example 3.6. *Show that the following function given by*

$$f(x) = \begin{cases} x^{\alpha} \sin \frac{1}{x^{\beta}} & if\ 0 < x \leq 1 \\ 0 & if\ x = 0, \end{cases}$$

where α and β are positive numbers is of bounded variation on $[0,1]$ if $\alpha > \beta + 1$.

Solution: Since for $x \neq 0$,

$$\begin{aligned} f'(x) &= \alpha x^{\alpha-1} \sin \frac{1}{x^{\beta}} - \beta x^{\alpha} \frac{1}{x^{\beta+1}} \cos \frac{1}{x^{\beta}} \\ &= \alpha x^{\alpha-1} \sin \frac{1}{x^{\beta}} - \beta x^{\alpha-\beta-1} \cos \frac{1}{x^{\beta}} \\ &= x^{\alpha-\beta-1} \Big[\alpha x^{\beta} \sin \frac{1}{x^{\beta}} - \beta \cos \frac{1}{x^{\beta}} \Big]. \end{aligned}$$

Therefore, when $\alpha > \beta + 1$, we have $|f'(x)| \leq \alpha + \beta$, $x \in (0,1]$.

Thus, f is a function of bounded variation on $(0,1]$ when $\alpha > \beta + 1$.

Example 3.7. *Verify either the following functions is of B.V.*

$$f(x) = \begin{cases} \sqrt{x} \sin \frac{1}{x} & if\ x \neq 0 \\ 0 & if\ x = 0 \end{cases}$$

Example 3.8. *Show that a polynomial f is of bounded variation on every closed interval $[a,b]$.*

Example 3.9. *Let f be a function of bounded variation on $[a,b]$. Then*
(i) f is of bounded variation on every closed subinterval of $[a,b]$
(ii) kf is of bounded variation on $[a,b]$, k be a constant.

Example 3.10. *If $f : [a,b] \to \mathbb{R}$ is of bounded variation on every closed subinterval of (a,b) it may yet found to be of bounded variation on $[a,b]$.*

Example 3.11. *Verify whether the following functions is of B.V. on $[0,1]$.*
(i)

$$f(x) = \begin{cases} \sqrt{x} \sin \frac{1}{x} & if\ x \neq 0 \\ 0 & if\ x = 0 \end{cases}$$

(ii)

$$f(x) = \begin{cases} x^{\frac{1}{3}} \sin \frac{\pi}{x} \; if \; x \neq 0 \\ 0 \quad\quad if \; x = 0 \end{cases}$$

Summary

In this unit, we have been acquainted with the Functions of Bounded variations, its various relevant properties; about variation function, and their applications.

Units 5 & 6

Course Structure

1. Absolutely continuous functions : Definition and basic properties
2. Deduction of the class of all absolutely continuous functions as a proper subclass of all functions of bounded variation
3. Characterization of an absolutely continuous function in terms of its derivative vanishing almost everywhere

4 Introduction

Absolute continuity is a smoothness property of functions that is stronger than continuity and uniform continuity. The notion of absolute continuity allows one to obtain generalizations of the relationship between the two central operations of calculus— differentiation and integration. This relationship is commonly characterized (by the fundamental theorem of calculus) in the framework of Riemann integration, but with absolute continuity it may be formulated in terms of Lebesgue integration. For real-valued functions on the real line two interrelated notions appear: absolute continuity of functions and absolute continuity of measures.

4.1 Absolute Continuity

A real-valued function f defined on $[a, b]$ is said to be absolutely continuous on $[a, b]$ if given $\epsilon > 0$ there exists a $\delta > 0$ such that

$$\sum_{i=1}^{n} |f(x_i^{'}) - f(x_i)| < \epsilon$$

for every finite collection $\{(x_i, x_i^{'})\}$ of non-overlapping intervals with

$$\sum_{i=1}^{n} |x_i^{'} - x_i| < \delta.$$

If in the above definition we consider only one interval ($n = 1$), then it is seen that an absolutely continuous function is continuous. Therefore, f is continuous on $[a, b]$ and so f is uniformly continuous on $[a, b]$.

Definition 4.1. (**Uniform Continuity**) *Let $f(x)$ be defined on the interval $[a, b]$. Then $f(x)$ is said to be uniformly continuos on $[a, b]$ if corresponding to $\epsilon(> 0)$ arbitrary, there exists a $\delta(> 0)$ depending only on (ϵ) such that $|f(x^{'}) - f(x^{''})| < \epsilon$ whenever $|x^{'} - x^{''}| < \delta$ for arbitrary pair of points $x^{'}$, $x^{''}$ in $[a, b]$.*

Note: Any uniformly continuous function is obviously continuous but the converse is not true.

Example 4.1. *$f(x) = x^2$ and $x \in (-\infty, \infty)$. But $|f(x^{'}) - f(x^{''})| < 1$ whenever $|x^{'} - x^{''}| < \delta$ implies $|f(x + \frac{\delta}{2}) - f(x)| < 1 = \epsilon$. [since $|(x + \frac{\delta}{2}) - x| < \delta$]*
or, $\frac{\delta}{2}|2x + \frac{\delta}{2}| < 1$ for all $x \in (-\infty, \infty)$. But this inequality does not hold for all sufficiently large values of x. So, $f(x)$ is not uniformly continuous.

Definition 4.2. (**AbsoluteContinuity**) *Let a function $f(x)$ be defined and finite in the interval $[a, b]$. Also, let (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) be non-overlapping intervals in $[a, b]$ such that $y_i \leq x_{i+1}$ $(i = 1, 2, \ldots, n-1)$. Then, given $\epsilon(> 0)$, if there exists a $\delta(> 0)$ such that $\sum_{i=1}^{n} |f(y_i) - f(x_i)| < \epsilon$ for $\sum_{i=1}^{n} |y_i - x_i| < \epsilon$, then the function $f(x)$ is said to be absolutely continuous or AC in $[a, b]$.*

If in the above definition, we consider only one interval ($n = 1$), then it is seen that an absolutely continuous function is continuous. In fact any absolutely continuous function is uniformly continuous and so continuous but the converse is not true.

Example 4.2. *Let $f(x) = \sqrt{x}$, $0 \leq x \leq \frac{1}{2}$. Let $f(1) = 0$ and define f to be linear on $[\frac{1}{2}, 1]$. Let $f(x + k) = f(x)$ for each $k \in \mathbb{Z}$ and each x. Show that f is continuous on $\mathbb{R}$ but not absolutely continuous.*

Solution: From the definition, f is continuous on $[0, 1]$. Given δ, $0 < \delta < \frac{1}{2}$, let $x_i = i$, $y_i = i + \frac{\delta}{i^2}$. Then for each n, $\sum_{i=1}^{n} |x_i - y_i| < 2\delta$ but $\sum_{i=1}^{n} |x_i - y_i| = \sum_{i=1}^{n} \frac{\sqrt{\delta}}{i}$ which tends to infinity with n. So, f is not absolutely continuous.

4.2 Some basic results

Theorem 4.1. *An absolutely continuous function is of bounded variation*

Proof. Let an absolutely continuous function f be defined in the interval $[a, b]$. Then corresponding to any positive number K there exists $\delta > 0$ such that for every finite collection $\{(x_i, x_i^{'})\}$ of non-overlapping intervals we have,

$$\sum_{i=1}^{n} |f(x_i^{'}) - f(x_i)| < K \quad whenever \quad \sum_{i=1}^{n} |x_i - x_i^{'}| < \delta \tag{4.1}$$

Now we divide the interval $[a, b]$ by the points $a = z_0 < z_1 < \ldots < z_p = b$. in such a way that $z_i - z_{i-1} < \delta$, for $i = 1, 2, \ldots, p$.
We consider the subinterval $[z_{i-1}, z_i]$ and let $P = \{z_{i-1} = t_0, t_1, t_2, \ldots, t_q = z_i\}$ be any partition of $[z_{i-1}, z_i]$. Note that

$$\sum_{k=1}^{q} |t_k - t_{k-1}| = |z_i - z_{i-1}| < \delta,$$

i.e.,

$$\sum_{k=1}^{q} |t_k - t_{k-1}| < \delta. \tag{4.2}$$

Then from (4.1) and (4.2) we get,

$$V(P, f) = \sum_{k=1}^{q} |f(t_k) - f(t_{k-1})| < K$$

and so

$$V(z_{i-1}, z_i; f) \leq K.$$

Consequently,

$$V(a, b; f) \leq Kp \leq \infty,$$

which shows that f is a function of bounded variation on $[a, b]$. □

Remark 4.1. *Converse of the theorem is not true. For example let $f : [0, 2] \to \mathbb{R}$ be defined by $f(x) = [x]$. Clearly f is a function of bounded variation on $[0, 2]$. Since every absolutely continuous function is continuous, it follows that f is not absolutely continuous on $[0, 2]$.*

Theorem 4.2. *If $f(x)$ has bounded derivative in $[a, b]$, then $f(x)$ is absolutely continuous in $[a, b]$.*

Proof. Let $|f'(x)| \leq K$, $\forall x \in [a, b]$. By mean value theorem for any pair of points x_1, x_2 in $[a, b]$, we have

$$\begin{aligned} |f(x_2) - f(x_1)| &= |x_2 - x_1||f'(\xi)|, \ \ where \ \ \xi \in (x_1, x_2) \\ &\leq K|x_2 - x_1| \ \ (\ since \ \ |f'(\xi)| \leq K) \end{aligned}$$

$$\Rightarrow |f(x_2) - f(x_1)| \leq K|x_2 - x_1|$$

So $f(x)$ satisfies Lipschitz condition in $[a, b]$. Choose $\epsilon > 0$ arbitrarily. If $\delta = \frac{\epsilon}{k}$ and $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be any system of non-overlapping intervals in $[a, b]$ with

$$\sum_{i=1}^{n} |y_i - x_i| < \delta.$$

Then,

$$\begin{aligned} \sum_{i=1}^{n} |f(y_i) - f(x_i)| &\leq \sum_{i=1}^{n} K|y_i - x_i| \\ &= K \sum_{i=1}^{n} |y_i - x_i| \\ &\leq K\delta = \epsilon \end{aligned}$$

Hence $f(x)$ is absolutely continuous in $[a, b]$. □

Note 4.1. *If $f(x)$ satisfies the Lipschitz condition on $[a,b]$, then $f(x)$ is absolutely continuous in $[a,b]$.*

Example 4.3. *Give an example of a function f which is continuous but not absolutely continuous.*

Solution:- Let $f:[0,1]\to\mathbb{R}$ be defined by,

$$\begin{aligned} f(x) &= x\sin\frac{\pi}{x}, \quad if \quad x\neq 0 \\ &= 0 \qquad\quad if \quad x=0. \end{aligned}$$

Then f is continuous on $[0,1]$. But f is not a function of bounded variation on $[0,1]$. Since every absolutely continuous function is a function of bounded variation , it follows that f is not absolutely continuous on $[0,1]$.

Theorem 4.3. *If $f(x)$ and $g(x)$ are two absolutely continuous functions then*

(i) $f(x)+g(x)$

(ii) $f(x)-g(x)$

(iii) $f(x)g(x)$ *and*

(iv) $\frac{f(x)}{g(x)}$ $(g(x)\neq 0)$ *are absolutely continuous.*

Proof. (i) Since $f(x)$ and $g(x)$ are both absolutely continuous, given $\epsilon>0$ there exists a $\delta>0$ such that for any system of non-overlapping intervals $\{(x_i,y_i)\}$ $(i=1,2,\ldots,n)$ in $[a,b]$,

$$\sum_{i=1}^{n}|f(y_i)-f(x_i)|<\frac{\epsilon}{2}$$

and

$$\sum_{i=1}^{n}|g(y_i)-g(x_i)|<\frac{\epsilon}{2},$$

whenever

$$\sum_{i=1}^{n}|y_i-x_i|<\delta.$$

Now we have,

$$\Big|[f(y_i)+g(y_i)]-[f(x_i)+g(x_i)]\Big| \;\leq\; |f(y_i)-f(x_i)|+|g(y_i)-g(x_i)|,$$

for $i = 1, 2, \ldots, n$. Therefore,

$$\begin{aligned}\sum_{i=1}^{n} \Big|[f(y_i) + g(y_i)] - [f(x_i) + g(x_i)]\Big| &\leq \sum_{i=1}^{n} |f(y_i) - f(x_i)| + \sum_{i=1}^{n} |g(y_i) - g(x_i)| \\ &\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,\end{aligned}$$

whenever

$$\sum_{i=1}^{n} |y_i - x_i| < \delta.$$

Hence $f + g$ is absolutely continuous.
(ii) Do yourself.
(iii) Since f and g are absolutely continuous on $[a, b]$, it follows that f and g are continuous on $[a, b]$ and hence f and g are bounded. Let $|f(x)| \leq B$ and $|g(x)| \leq B$, $\forall x \in [a, b]$.
Since f and g are both absolutely continuous on $[a, b]$, given $\epsilon > 0$ there exists a $\delta > 0$ such that for every system of non-overlapping intervals $\{(x_i, y_i)\}$ $(i = 1, 2, \ldots, n)$ of $[a, b]$ we have,

$$\sum_{i=1}^{n} |f(y_i) - f(x_i)| < \frac{\epsilon}{2B} \text{ and } \sum_{i=1}^{n} |g(y_i) - g(x_i)| < \frac{\epsilon}{2B}, \tag{4.3}$$

whenever $\sum_{i=1}^{n} |y_i - x_i| < \delta$.
Note that,

$$\begin{aligned}& |f(y_i)g(y_i) - f(x_i)g(x_i)| \\ = \; & |g(y_i)\{f(y_i) - f(x_i)\} + f(x_i)\{g(y_i) - g(x_i)\}| \\ \leq \; & |g(y_i)||f(y_i) - f(x_i)| + |f(x_i)||g(y_i) - g(x_i)|\end{aligned}$$

Therefore,

$$\begin{aligned}\sum_{i=1}^{n} |f(y_i)g(y_i) - f(x_i)g(x_i)| &\leq \sum_{i=1}^{n} |g(y_i)||f(y_i) - f(x_i)| + \sum_{i=1}^{n} |f(x_i)||g(y_i) - g(x_i)| \\ &\leq \sum_{i=1}^{n} B|f(y_i) - f(x_i)| + \sum_{i=1}^{n} B|g(y_i) - g(x_i)| \\ &= B\sum_{i=1}^{n} |f(y_i) - f(x_i)| + B\sum_{i=1}^{n} |g(y_i) - g(x_i)|. \end{aligned} \tag{4.4}$$

Now from (4.3) and (4.4) we have,

$$\sum_{i=1}^{n} |f(y_i)g(y_i) - f(x_i)g(x_i)| < B\frac{\epsilon}{2B} + B.\frac{\epsilon}{2B} = \epsilon,$$

whenever

$$\sum_{i=1}^{n} |y_i - x_i| < \delta.$$

Thus fg is absolutely continuous.
(iv) Since g is continuous and $g(x) \neq 0$, it follows that $\frac{1}{g(x)}$ is continuous on $[a, b]$ and hence $\frac{1}{g(x)}$ is bounded on $[a, b]$. Then there exists $M > 0$ such that $\left|\frac{1}{g(x)}\right| \leq \frac{1}{M}$. Since g is absolutely continuous on $[a, b]$, given $\epsilon > 0$, there exists a $\delta > 0$ such that for any system of non-overlapping intervals $\{(x_i, y_i)\}$ $(i = 1, 2, \ldots, n)$ of $[a, b]$, we have

$$\sum_{i=1}^{n} |g(y_i) - g(x_i)| < M^2\epsilon, \ \ whenever \ \ \sum_{i=1}^{n} |y_i - x_i| < \delta.$$

Note that,

$$\begin{aligned}
\sum_{i=1}^{n} \Big|\frac{1}{g(y_i)} - \frac{1}{g(x_i)}\Big| &= \sum_{i=1}^{n} \frac{|g(x_i) - g(y_i)|}{|g(x_i)||g(y_i)|} \\
&\leq \sum_{i=1}^{n} \frac{|g(y_i) - g(x_i)|}{M^2} \\
&= \frac{1}{M^2} \sum_{i=1}^{n} |g(y_i) - g(x_i)| \\
&< \frac{1}{M^2}\epsilon M^2 = \epsilon, \ \ whenever \sum_{i=1}^{n} |y_i - x_i| < \delta.
\end{aligned}$$

This shows that $\frac{1}{g(x)}$ is absolutely continuous on $[a, b]$.
Now by applying (iii) $\frac{f}{g}$ is absolutely continuous on $[a, b]$. □

Example 4.4. *Prove that the function $f : [0, 1] \to \mathbb{R}$ be defined by,*

$$\begin{aligned}
f(x) &= x^2 \cos\frac{1}{x}, \ \ if \ \ 0 < x \leq 1 \\
&= 0, \qquad\qquad if \ \ x = 0.
\end{aligned}$$

is absolutely continuous on $[0, 1]$.

Solution: Note that

$$\begin{aligned}
f'(x) &= 2x \cos\frac{1}{x} + \sin\frac{1}{x}, \ \ if \ \ 0 < x \leq 1 \\
&= 0, \qquad\qquad\qquad\quad if \ \ x = 0.
\end{aligned}$$

and so $|f'(x)| \leq 3$ on $[0, 1]$. Then f has bounded derivatives on $[0, 1]$ and so f is absolutely continuous on $[0, 1]$.

Example 4.5. *Give an example to show that every monotone function is not necessarily absolutely continuous.*

Solution: Let $f(x) = [x], \forall\ x \in [0, 2]$. Clearly f is monotone increasing function on $[0, 2]$. But f is not absolutely continuous, because every absolutely continuous function is continuous.

Example 4.6. *The function $f(x) = \sqrt{x}$ is absolutely continuous on $[0, 1]$.*

Theorem 4.4. *If f is absolutely continuous on $[a, b]$, then f has derivative almost everywhere.*

Theorem 4.5. *If f is absolutely continuous on $[a, b]$ and $f'(x) = 0$ almost everywhere, then f is constant.*

Theorem 4.6. *Let a function f defined in $[a, b]$ be absolutely continuous. If $c \leq f(x) \leq d, \forall\ x \in [a, b]$ and $F(y)$ satisfies Lipschitz condition in $[c, d]$, then $F(f(x))$is absolutely continuous in $[a, b]$.*

Proof. Since $F(y)$ satisfies Lipchitz condition, there exists constant $K > 0$ such that

$$|F(y_1) - F(y_2)| \leq K|y_1 - y_2| \tag{4.5}$$

for any two points $y_1$and y_2 in $[c, d]$.

Now if $\{(x_i, y_i)\}\ \ (i = 1, 2, \ldots, n)$ be any system of non-overlapping intervals in $[a, b]$.

Then given any $\epsilon > 0$ there exists a $\delta > 0$ such that,

$$\sum_{i=1}^{n} |f(y_i) - f(x_i)| < \frac{\epsilon}{K}, \tag{4.6}$$

whenever

$$\sum_{i=1}^{n} |y_i - x_i| < \delta.$$

Therefore from (4.5) and (4.6) we get ,

$$\sum_{i=1}^{n} |F(f(y_i)) - F(f(x_i))| \ \ \leq \ \ K \sum_{i=1}^{n} |f(y_i) - f(x_i)| < K.\frac{\epsilon}{K} = \epsilon,$$

whenever $\sum_{i=1}^{n} |y_i - x_i| < \delta$. Hence $F(f(x))$ is absolutely continuous on $[a, b]$. $\square$

Theorem 4.7. *Let f be a function of bounded variation on $[a, b]$. Then f is absolutely continuous on $[a, b]$ iff the variation function $F(x) = V_f[a, x]$ is absolutely continuous on $[a, b]$.*

Proof. The absolute continuity of $F(x)$ clearly implies that of f since $|f(x_i) - f(x_{i-1})| \leq V_f[x_{i-1}, x_i] = V - f[a, x_i] - V_f[a, x_{i-1}] = F(x_i) - F(x_{i-1})$ for $a \leq x_{i-1} < x_i \leq b$.

Thus for a finite collection of disjoint intervals $(x_i, y_i) \subset [a, b]$, we have due to the absolute continuity of $F(x)$.

Then for arbitrary $\epsilon > 0$, there corresponds a $\delta > 0$ such that $\sum_{i=1}^{n} |F(x_i) - F(y_i)| < \epsilon$ whenever $\sum_{i=1}^{n} |x_i - y_i| < \delta$.
Clearly, $\sum_{i=1}^{n} |f(x_i) - f(y_i)| < \epsilon$ whenever $\sum_{i=1}^{n} |x_i - y_i| < \delta$. Hence, f is absolutely continuous.

On the other hand, assume f is absolutely continuous. Then given an $\epsilon > 0$, there exists a $\delta > 0$ such that for any finite collection $C = \{(x_i, x_i^{'}) : i = 1, 2, \ldots, n\}$ of pairwise disjoint intervals in $[a, b]$ with $\sum_{i=1}^{n} |x_i^{'} - x_i| < \delta$, we have

$$\sum_{i=1}^{n} |f(x_i^{'}) - f(x_i)| < \epsilon.$$

For each i, let $P_i = |\{x_i = a_0^i < a_1^i < a_2^i < \ldots < a_{m_i^i} = x_i^{'}\}$ be a partition of $[x_i, x_i^{'}]$. Since

$$\sum_{i=1}^{n} \sum_{j=1}^{m_i} |a_j^i - a_{j-1}^i| = \sum_{i=1}^{n} |x_i^{'} - x_i| < \delta,$$

we have

$$\sum_{i=1}^{n} \sum_{j=1}^{m_i} |f(a_j^i) - f(a_{j-1}^i)| < \epsilon.$$

This implies

$$\sum_{j=1}^{m_i} |f(a_j^{'}) - f(a_{j-1}^{'})| + \sum_{j=1}^{m_i} |f(a_j^2) - f(a_{j-1}^2)| + \ldots + \sum_{j=1}^{m_i} |f(a_j^n) - f(a_{j-1}^n)| < \epsilon.$$

Now fixing the collection C but varying the partition P_i of each $[x_i, x_i^{'}]$, we have, upon taking the supremum over all such partitions P_i,

$$V_f[x_1, x_1^{'}] + V_f[x_2, x_2^{'}] + \cdots + V_f[x_n, x_n^{'}] < \epsilon$$

$$\implies \sum_{i=1}^{n} \{V_f[a, x_i^{'}] - V_f[a, x_i]\} < \epsilon$$

$$\implies \sum_{i=1}^{n} \{F(x_i^{'}) - F(x_i)\} < \epsilon.$$

So, $F(x)$ is absolutely continuous. □

Set of measure zero:

Definition 4.3. *A property P is said to hold good almost everywhere (abbreviated a.e.) on a set S if the set of points of S where P fails to hold has measure zero.*

Theorem 4.8. *If f is absolutely continuous on $[a,b]$ and $f' = 0$ a.e., then f is a constant function.*

Theorem 4.9. *Let a function $f(x)$ defined on $[a,b]$ be absolutely continuous. If for all x, $c \leq f(x) \leq d$ and $F(y)$ satisfies Lipschitz condition in $[c,d]$, then $F(f(x))$ is absolutely continuous in $[a,b]$.*

Proof. Since $F(y)$ satisfies Lipschitz condition, there exists a constant k such that $|F(y_1) - F(y_2)| \leq k|y_1 - y_2|$ for any two points y_1 and y_2 in $[c,d]$.
Now if $\{(x_i, y_i) : i = 1, 2, \ldots, n\}$ be any system of non-overlapping intervals in $[a,b]$, then given $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\sum_{i=1}^{n} |f(y_i) - f(x_i)| < \frac{\epsilon}{k}$$

when $\sum_{i=1}^{n}(y_i - x_i) < \delta$. So,

$$\sum_{i=1}^{n} |F(f(y_i)) - F(f(x_i))| \leq k \sum_{i=1}^{n} |f(y_i) - f(x_i)| < k.\frac{\epsilon}{k} = \epsilon$$

when $\sum_{i=1}^{n}(y_i - x_i) < \delta$. So, $F(f(x))$ is absolutely continuous. □

Outer measure:

Let us consider the family $\mathcal{F}$ of all countable collections of open intervals.For any arbitrary $\mathcal{I} \in \mathcal{F}$, the sum $\sum_{I \in \mathcal{I}} l(I)$ is a non-negative extended real number. Let E be an arbitrary set. Consider the sub family C of $\mathcal{F}$ containing of countable collections $\mathcal{I}$ of open intervals $\{I_i\}$ such that $E \subset \cup I_i$ i.e.,

$$C = \{\mathcal{I} : \mathcal{I} \in \mathcal{F} \text{ and } \mathcal{I} \text{ covers } E\}.$$

The subfamily C is obviously non-empty. Thus we obtain a well defined number $m^*(E)$ in the set of all non negative extended real numbers given by

$$m^*(E) = inf\{\sum_{I \in \mathcal{I}} l(I) : I \in C\}.$$

Definition 4.4. *The Lebesgue outer measure $m^*(E)$ of an arbitrary set is given by $m^*(E) = inf \sum_i l(I_i)$, where the infimum is taken over all countable collections $\{I_i\}$ of open intervals such that $E \subset \cup_i I_i$.*

Vitali cover:

Definition 4.5. *Let $E \subset \mathbb{R}$. A collection γ of intervals is said to be a Vitali cover of the set E if for each $x \in E$ and each $\epsilon > 0$, there exists an interval $I \in \mathcal{I}$ with $x \in \mathcal{I}$ and $l(I) < \epsilon$.*

Example 4.7. *Let r_n be an enumeration of the rationals in $[a, b]$. Then the collection $I_{n,i}$, where $I_{n,i} = [r_n - \frac{1}{i}, r_n + \frac{1}{i}]$, $n, i \in \mathbb{N}$ forms a Vitali cover of $[a, b]$.*

4.3 Vitali's covering Theorem:

Let E be a set of finite outer measure and $\mathcal{I}$ be a collection of intervals which covers E in the sense of Vitali. Then given $\epsilon > 0$, there is a finite disjoint collection $\{I_1, I_2, \ldots, I_n\}$ of intervals in $\mathcal{I}$ such that $m^*(E - \cup_{i=1}^N) < \epsilon$, $m^*(E - \cup_{i=1}^N) < \epsilon$.

Theorem 4.10. *Let $\{E_n\}$ be a countable collection of sets. Then $m^*(\cup_n) \leq \sum_n m^*(E_n)$.*

Theorem 4.11. *If $f(x)$ is absolutely continuous on $[a, b]$ and $f' = 0$ a.e., then $f(x)$ is constant.*

Proof. It is sufficient to prove that $f(a) = f(b)$. For then, for any x in $a < x \leq b$, we have in $[a, x]$ $f(x) = f(a)$ which implies $f(x)$ is constant. Let $E = \{x \in [a, b] : f'(x) = 0\}$ so that $m^*(E) = b - a$. Then for every $x \in E$ and arbitrary $\epsilon(> 0)$, we get a sufficiently small $h(> 0)$ such that $\frac{|f(x+h)-f(x)|}{h} < \epsilon$.

i.e.,

$$|f(x + h) - f(x)| < \epsilon h \tag{4.7}$$

The closed interval $[x, x + h]$ covers the set E in Vitali's sense. So we can pick out a finite set of non-overlapping intervals out of these, say $\delta_r = [x_r, x_r + h_r]$, $r = 1, 2, \ldots, n$ such that $m^*(E - \cup_{i=1}^n) < \delta$, where $\delta(> 0)$ is a given number. Let $x_r < x_{r+1}$, $r = 1, 2, \ldots, n$.

Now

$$\begin{aligned} b - a = m^*(E) &\leq \sum_{r=1}^{n} m\delta_r + m^*(E - \cup_{r=1}^n \delta_r) \\ &< \sum_{r=1}^{n} m\delta_r + \delta. \end{aligned}$$

i.e.,

$$(b - a) - \sum_{r=1}^{n} m\delta_r < \delta. \tag{4.8}$$

From (4.8), it is clear that the lenght of the intervals (a, x_1), $(x_1 + h_1, x_2)$,..., $(x_n + h_n, b)$ is less than δ.
We have

$$\begin{aligned} & |f(b) - f(a)| \\ = \; & |f(x_1) - f(a) + f(x_1 + h_1) - f(x_1) + f(x_2) + \ldots + f(b) - f(x_n + h_n)| \\ \leq \; & \{|f(x_1) - f(a)| + |f(x_2) - f(x_1 + h_1)| + \ldots + |f(b) - f(x_n + h_n)|\} \\ & + \{|f(x_1 + h_1) - f(x_1)| + |f(x_2 + h_2) - f(x_2)| \\ & + \ldots + |f(x_n + h_n) - f(x_n)|\} \end{aligned} \tag{4.9}$$

Since $f(x)$ is a.c., the first sum on the right hand side of (4.9) is less than ϵ. Also since $\sum_{r=1}^{n} h_r = \sum_{r=1}^{n} m\delta_r \leq (b-a)$, by (4.7), the second sum on the right hand side of (4.9) is $< \epsilon(b-a)$. Consequently, from (4.9), it follows that

$$|f(b) - f(a)| < \epsilon + \epsilon(b-a) = \epsilon(1 + b - a).$$

Since $\epsilon(> 0)$ is arbitrary, $f(b) = f(a)$. □

Inner Measure:

Let the set E is in the bounded interval $[a, b]$. Then the inner measure of the set E is defined to be $m_*(E) = b - a = m^*(E^c)$. A set is said to be measurable if $m_*(E) = m^*(E)$.

Corollary 4.1. *If $f(x)$ and $g(x)$ are two a.c. functions and $f'(x) = g'(x)$ a.e., then $f(x) - g(x)$ is constant.*

Summary

In this unit, we have learnt about absolute continuity and its properties, the Vitali's Covering Theorem.

Units 7 & 8

Course Structure

1. Riemann-Stieltjes integral : Existence and basic properties
2. Integration by parts, Integration of a continuous function with respect to a step function
3. Convergence theorems in respect of integrand
4. Convergence theorem in respect of integrator (Helly's Second theorem)

5 Introduction

The Riemann–Stieltjes integral is a generalization of the Riemann integral, named after Bernhard Riemann and Thomas Joannes Stieltjes. The definition of this integral was first published in 1894 by Stieltjes. It serves as an instructive and useful precursor of the Lebesgue integral, and an invaluable tool in unifying equivalent forms of statistical theorems that apply to discrete and continuous probability.

5.1 Riemann Stieltjes Integral

Definition 5.1. *Let $[a, b]$ be a finite closed interval. Let $P = \{a = x_0, x_1, \ldots, x_n = b\}$ be a partition of $[a, b]$ and $I_r = [x_{r-1}, x_r]$, $r = 1, 2, \ldots, n$. The length of the r-th sub-interval I_r is denoted by δ_r, i.e., $\delta_r = x_r - x_{r-1}$. The set of all possible partitions of $[a, b]$ is denoted by $\mathcal{P}[a, b]$.*
The length of the greatest of all the intervals I_r of the partition P will be called its norm and denoted by $||P||$ or $\mu(P)$. Thus

$$||P|| = \mu(P) = \max\{\delta_r : r = 1, 2, \ldots, n\}.$$

If P_1 and P_2 be two partitions of $[a, b]$ such that $P_2 \supseteq P_1$, then we say that P_2 is finer than P_1 or P_2 is a refinement of P_1.
If P_1 and P_2 are two partitions of $[a, b]$, then $P_1 \cup P_2$ is called a common refinement of P_1 and P_2.

Let f and α be bounded functions on $[a, b]$ and α be monotonic increasing on $[a, b]$.
Corresponding to any partition $P = \{a = x_0, x_1, \ldots, x_n = b\}$ of $[a, b]$ we write

$$\delta\alpha_r = \alpha(x_r) - \alpha(x_{r-1}), \quad r = 1, 2, \ldots, n.$$

Note that

$$\sum_{r=1}^{n} \delta\alpha_r = \alpha(b) - \alpha(a).$$

Since α is monotonic increasing on $[a, b]$, we have

$$\delta\alpha_r \geq 0, \forall\ r = 1, 2, \ldots, n.$$

The function f which is bounded on $[a, b]$ is also necessarily bounded in each sub-interval $I_r = [x_{r-1}, x_r]$. Let M_r, m_r be the supremum and the infimum of f on I_r. We define two sums,

$$U(P, f, \alpha) = \sum_{r=1}^{n} M_r \delta\alpha_r$$

$$L(P, f, \alpha) = \sum_{r=1}^{n} m_r \delta\alpha_r.$$

$U(P, f, \alpha)$ and $L(P, f, \alpha)$ are called upper and lower Riemann-Stieltjes sums respectively of f with respect to α and corresponding to the partition P. If M, m are respectively the upper and lower bounds of f on $[a, b]$, we have

$$m \leq m_r \leq M_r \leq M, \quad r = 1, 2, \ldots, n$$

$$\Rightarrow m\delta\alpha_r \leq m_r\delta\alpha_r \leq M_r\delta\alpha_r \leq M\delta\alpha_r, \ \ r = 1, 2, \ldots, n.$$

Adding all inequalities we get

$$m\{\alpha(b) - \alpha(a)\} \leq L(P, f, \alpha) \leq U(P, f, \alpha) \leq M\{\alpha(b) - \alpha(a)\}. \tag{5.1}$$

From (5.1) we see that both the sets $\{U(P, f, \alpha) : P \in \mathcal{P}[a, b]\}$ and $\{L(P, f, \alpha) : P \in \mathcal{P}[a, b]\}$ are bounded. We now define two integrals, which always exist as follows:

$$\int_a^{\bar{b}} f\mathrm{d}\alpha = \inf\{U(P, f, \alpha) : P \in \mathcal{P}[a, b]\} \tag{5.2}$$

and

$$\int_{\underline{a}}^{b} f\mathrm{d}\alpha = \sup\{L(P, f, \alpha) : P \in \mathcal{P}[a, b]\}. \tag{5.3}$$

These are respectively called the upper and the lower integrals of f with respect to α on $[a, b]$.
These two integrals may or may not be equal. In case these two integrals are equal, i.e.,

$$\int_a^{\bar{b}} f\mathrm{d}\alpha = \int_{\underline{a}}^{b} f\mathrm{d}\alpha$$

we say that f is integrable with respect to α in the Riemann sense and write $f \in \mathrm{R}(\alpha)$. Their common value is denoted by

$$\int_a^b f \mathrm{d}\alpha.$$

and is called the Riemann-Stieltjes integral of f with respect to α on $[a, b]$.

Theorem 5.1. *The lower Riemann-Stieltjes integral cannot exceed the upper Riemann-Stieltjes integral, i.e.,*

$$\int_{\underline{a}}^b f \mathrm{d}\alpha \leq \int_a^{\overline{b}} f \mathrm{d}\alpha.$$

Remark 5.1. *From (5.2) and (5.3), it follows that*

$$\int_a^{\overline{b}} f \ \mathrm{d}\alpha \leq U(P, f, \alpha) \quad \textit{and} \quad L(P, f, \alpha) \leq \int_{\underline{a}}^b f \ \mathrm{d}\alpha,$$

$\forall \ P \in \mathcal{P}[a, b]$ *and so*

$$L(P, f, \alpha) \leq \int_{\underline{a}}^b f \ \mathrm{d}\alpha \leq \int_a^{\overline{b}} f \ \mathrm{d}\alpha \leq U(P, f, \alpha),$$

$\forall \ P \in \mathcal{P}[a, b]$. *If* $f \in \mathrm{R}(\alpha)$, *then* $\int_a^b f \mathrm{d}\alpha$ *lie between* $U(P, f, \alpha)$ *and* $L(P, f, \alpha)$.

Remark 5.2. *As a particular case, if* $\alpha(x) = x$, *then the so called Riemann-Stieltjes sums reduce to the Riemann sums. So by taking* $\alpha(x) = x$, *the Riemann integral can be easily seen to be a special case of Riemann-Stieltjes integral.*

Remark 5.3. *In this chapter, unless otherwise stated, all functions will be taken as bounded and the function* α *will be always monotonic increasing.*

Theorem 5.2. *If* P^* *is a refinement of* P, *then*

$$(i) \ \ L(P, f, \alpha) \leq L(P^*, f, \alpha) \quad (ii) \ \ U(P^*, f, \alpha) \leq U(P, f, \alpha).$$

Theorem 5.3. *If* P_1 *and* P_2 *be any two partitions of* $[a, b]$, *then* $U(P_2, f, \alpha) \geq L(P_1, f, \alpha)$.

Theorem 5.4. *If* $a \in \mathbb{R}$ *and* $0 \leq a < \varepsilon$ *holds for every positive* ε, *then* $a = 0$.

5.2 A Condition Of Integrability

Theorem 5.5. *A function f is integrable with respect to α on $[a,b]$ if and only if for every $\varepsilon > 0$ there exists a partition P of $[a,b]$ such that*

$$U(P,f,\alpha) - L(P,f,\alpha) < \varepsilon.$$

Proof. **The condition is necessary.**
Let $f \in \mathrm{R}(\alpha)$ over $[a,b]$. Then

$$\underline{\int_a^b} f\mathrm{d}\alpha = \overline{\int_a^b} f\mathrm{d}\alpha = \int_a^b f\mathrm{d}\alpha.$$

Let $\varepsilon > 0$ be any real number. Since $\overline{\int_a^b} f\mathrm{d}\alpha = \inf\{U(P,f,\alpha) : P \in \mathcal{P}[a,b]\}$ and $\underline{\int_a^b} f\mathrm{d}\alpha = \sup\{L(P,f,\alpha) : P \in \mathcal{P}[a,b]\}$, it follows that there exist partitions P_1 and P_2 of $[a,b]$ such that

$$U(P_1,f,\alpha) < \overline{\int_a^b} f\mathrm{d}\alpha + \frac{1}{2}\varepsilon \quad or \quad U(P_1,f,\alpha) < \int_a^b f\mathrm{d}\alpha + \frac{1}{2}\varepsilon \tag{5.4}$$

and

$$L(P_2,f,\alpha) > \underline{\int_a^b} f\mathrm{d}\alpha - \frac{1}{2}\varepsilon \quad or \quad \int_a^b f\mathrm{d}\alpha < L(P_2,f,\alpha) + \frac{1}{2}\varepsilon. \tag{5.5}$$

Let $P = P_1 \cup P_2$. Then P is the common refinement of P_1 and P_2. Now from (5.4) and (5.5) we have

$$U(P,f,\alpha) \le U(P_1,f,\alpha) < \int_a^b f\mathrm{d}\alpha + \frac{1}{2}\varepsilon < L(P_2,f,\alpha) + \varepsilon < L(P,f,\alpha) + \varepsilon.$$

Thus $U(P,f,\alpha) - L(P,f,\alpha) < \varepsilon$, showing that the condition is necessary.
The condition is sufficient.
Let for every $\varepsilon > 0$, there exists a partition P such that

$$U(P,f,\alpha) - L(P,f,\alpha) < \varepsilon. \tag{5.6}$$

We know that

$$\overline{\int_a^b} f\ \mathrm{d}\alpha \le U(P,f,\alpha) \quad and \quad L(P,f,\alpha) \le \underline{\int_a^b} f\ \mathrm{d}\alpha$$

$$\Rightarrow \overline{\int_a^b} f \, \mathrm{d}\alpha \leq U(P, f, \alpha) \ \ and \ \ -\underline{\int_a^b} f \, \mathrm{d}\alpha \leq -L(P, f, \alpha). \tag{5.7}$$

Now from (5.7) we get

$$\overline{\int_a^b} f \, \mathrm{d}\alpha - \underline{\int_a^b} f \, \mathrm{d}\alpha \leq U(P, f, \alpha) - L(P, f, \alpha). \tag{5.8}$$

Therefore from (5.6) and (5.8) we get

$$0 \leq \overline{\int_a^b} f \, \mathrm{d}\alpha - \underline{\int_a^b} f \, \mathrm{d}\alpha < \varepsilon,$$

for every $\varepsilon > 0$. This shows that

$$\overline{\int_a^b} f \, \mathrm{d}\alpha - \underline{\int_a^b} f \, \mathrm{d}\alpha = 0, \ \ i.e., \ \ \overline{\int_a^b} f \, \mathrm{d}\alpha = \underline{\int_a^b} f \, \mathrm{d}\alpha,$$

showing that the function $f \in \mathrm{R}(\alpha)$ over $[a, b]$ and so the condition is sufficient. □

Example 5.1. *Let $f : [0, 1] \longrightarrow \mathbb{R}$ be defined by*

$$f(x) = \begin{cases} 1, & if\, x \ \ is\, rational \\ -1, & if\, x \ \ is \ \ irrational \end{cases}$$

and α is a bounded monotonic increasing function on $[0, 1]$ such that $\alpha(0) < \alpha(1)$. Then prove that $f \notin \mathrm{R}(\alpha)$ on $[0, 1]$.

Proof. Clearly f is bounded on $[0, 1]$. Let $P = \{0 = x_0, x_1, \ldots, x_{r-1}, x_r, \ldots, x_n = 1\}$ be any partition of $[0, 1]$. Let m_r and M_r be the infimum and the supremum of f on $[x_{r-1}, x_r]$. Then $m_r = -1$ and $M_r = 1$ for $r = 1, 2, \ldots, n$. Note that

$$U(P, f, \alpha) = \sum_{r=1}^{n} M_r \delta\alpha_r = \sum_{r=1}^{n} \delta\alpha_r = \alpha(1) - \alpha(0)$$

and

$$L(P, f, \alpha) = \sum_{r=1}^{n} m_r \delta\alpha_r = -\sum_{r=1}^{n} \delta\alpha_r = -[\alpha(1) - \alpha(0)].$$

Therefore

$$\overline{\int_0^1} f \mathrm{d}\alpha = \inf\{U(P, f, \alpha) : P \in \mathcal{P}[0, 1]\} = \alpha(1) - \alpha(0)$$

and

$$\underline{\int_0^1} f\mathrm{d}\alpha = \sup\{L(P,f,\alpha) : P \in \mathcal{P}[0,1]\} = -[\alpha(1)-\alpha(0)].$$

Since $\overline{\int_0^1} f\mathrm{d}\alpha \neq \underline{\int_0^1} f\mathrm{d}\alpha$, it follows that $f \notin \mathrm{R}(\alpha)$ on $[0,1]$. □

Example 5.2. *Let $f(x) = k$ be a constant function defined on $[a,b]$ and α a monotonic increasing function on $[a,b]$. Prove that $f \in \mathrm{R}(\alpha)$ on $[a,b]$ and $\int_a^b f\ \mathrm{d}\alpha = k[\alpha(b)-\alpha(a)]$.*

Proof. Hints: $m_r = M_r = k$. □

5.3 Integral as a limit sum

Let f be a bounded function and α be a monotonic increasing function on $[a,b]$. Let $P = \{a = x_0, x_1, \ldots, x_{r-1}, x_r, \ldots, x_n = b\}$ be any partition of $[a,b]$. Let $\xi_r \in [x_{r-1}, x_r]$, for $r = 1, 2, \ldots, n$. Then the sum

$$S(P,f,\alpha) = \sum_{r=1}^{n} f(\xi_r)\delta\alpha_r$$

is called the Riemann-Stieltjes sum of f relative to α on $[a,b]$ corresponding to the partition P. We say that $S(P,f,\alpha)$ converges to A as $\mu(P) \to 0$, i.e.,

$$\lim_{\mu(P)\to 0} S(P,f,\alpha) = A$$

if for every $\varepsilon > 0$ there exists $\delta > 0$ such that $|S(P,f,\alpha) - A| < \varepsilon$ for every partition $P = \{a = x_0, x_1, \ldots, x_{r-1}, x_r, \ldots, x_n = b\}$ of $[a,b]$ with $\mu(P) < \delta$ and every choice of ξ_r in $[x_{r-1}, x_r]$.

Remark 5.4. *Let $M_r = \sup_{x\in I_r} f(x)$ and $m_r = \inf_{x\in I_r} f(x)$ for $r = 1, 2, \ldots, n$. Then*

$$m_r \leq f(\xi_r) \leq M_r \quad for \ \ r = 1, 2, \ldots, n$$

$$\Rightarrow m_r\delta\alpha_r \leq f(\xi_r)\delta\alpha_r \leq M_r\delta\alpha_r \quad for \ \ r = 1, 2, \ldots, n$$

$$\Rightarrow \sum_{r=1}^{n} m_r\delta\alpha_r \leq \sum_{r=1}^{n} f(\xi_r)\delta\alpha_r \leq \sum_{r=1}^{n} M_r\delta\alpha_r$$

$$L(P,f,\alpha) \leq S(P,f,\alpha) \leq U(P,f,\alpha).$$

Theorem 5.6. *If* $\lim_{\mu(P)\to 0} S(P, f, \alpha)$ *exists, then* $f \in \mathrm{R}(\alpha)$ *and*

$$\lim_{\mu(P)\to 0} S(P, f, \alpha) = \int_a^b f \, \mathrm{d}\alpha.$$

Remark 5.5. *The theorem asserts that the existence of the limit of* $S(P, f, \alpha)$ *implies that* $f \in \mathrm{R}(\alpha)$. *The existence of the limit is a sufficient condition for* $f \in \mathrm{R}(\alpha)$ *but it is not a necessary condition, i.e., functions exist which are integrable but for which limit of* $S(P, f, \alpha)$ *does not exist.*

Example 5.3. *Let* $f, \alpha : [-1, 1] \longrightarrow \mathbb{R}$ *be defined by*

$$f(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1, & \text{if } x \geq 0 \end{cases}$$

and

$$\alpha(x) = \begin{cases} 0, & \text{if } x \leq 0 \\ 1, & \text{if } x > 0 \end{cases}$$

Then $f \in \mathrm{R}(\alpha)$ *but* $\lim_{\mu(P)\to 0} S(P, f, \alpha)$ *does not exist.*

Theorem 5.7. *If* f *is continuous on* $[a, b]$, *then* $f \in \mathrm{R}(\alpha)$ *and* $\lim_{\mu(P)\to 0} S(P, f, \alpha) = \int_a^b f \, \mathrm{d}\alpha$.

Proof. Let $\varepsilon > 0$ be given. Let us choose $\eta > 0$ such that

$$\eta[\alpha(b) - \alpha(a)] < \varepsilon. \tag{5.9}$$

Since f is continuous on $[a, b]$, so f is uniformly continuous on $[a, b]$. Hence for $\eta > 0$, there exists $\delta > 0$ such that

$$|f(x^{'}) - f(x^{''})| < \eta \tag{5.10}$$

whenever $|x^{'} - x^{''}| < \delta$. Let $P = \{a = x_0, x_1, \ldots, x_{r-1}, x_r, \ldots, x_n = b\}$ be any partition of $[a, b]$ with $\mu(P) < \delta$.
Let m_r, M_r be the infimum and supremum of f in $[x_{r-1}, x_r]$. Since f is continuous on $[a, b]$, so it is continuous on each sub-interval $[x_{r-1}, x_r]$ and hence f will attain its bounds m_r, M_r on $[x_{r-1}, x_r]$. Therefore there exist $y^{'}, y^{''} \in [x_{r-1}, x_r]$ such that $f(y^{'}) = M_r$ and $f(y^{''}) = m_r$. Since $|y^{'} - y^{''}| \leq |x_r - x_{r-1}| < \delta$, from (5.10)

$$|f(y^{'}) - f(y^{''})| < \eta$$

$$\Rightarrow M_r - m_r < \eta, \quad r = 1, 2, \ldots, n. \tag{5.11}$$

Now from (5.9) and (5.11),

$$\begin{aligned} U(P,f,\alpha) - L(P,f,\alpha) &= \sum_{r=1}^{n}(M_r - m_r)\delta\alpha_r < \eta\sum_{r=1}^{n}\delta\alpha_r \\ &= \eta\sum_{r=1}^{n}[\alpha(x_r) - \alpha(x_{r-1})] \\ &= \eta[\alpha(b) - \alpha(a)] < \varepsilon. \end{aligned}$$

Hence $f \in \mathrm{R}(\alpha)$ over $[a,b]$. We know that $S(P,f,\alpha)$ and $\int_a^b f\mathrm{d}\alpha$ lie between $U(P,f,\alpha)$ and $L(P,f,\alpha)$ for all partition P with $\mu(P) < \delta$ and for every choice of ξ_r in $[x_{r-1}, x_r]$. Hence we have

$$|S(P,f,\alpha) - \int_a^b f\mathrm{d}\alpha| \leq U(P,f,\alpha) - L(P,f,\alpha) < \varepsilon.$$

Therefore $\lim_{\mu(P)\to 0} S(P,f,\alpha) = \int_a^b f\ \mathrm{d}\alpha$. □

Remark 5.6. *From the above theorem, it follows that continuity is sufficient condition for integrability of a function. But it is not necessary condition, i.e., there exist functions which are integrable but function is not continuous. See Example 2.3.*

Theorem 5.8. *Suppose f is bounded on $[a,b]$, f has only finitely many points of discontinuity on $[a,b]$ and α is continuous at every point at which f is discontinuous. Then $f \in \mathrm{R}(\alpha)$ on $[a,b]$.*

Theorem 5.9. *Suppose f is continuous on $[a,b]$ and α is of bounded variation on $[a,b]$. Then $f \in \mathrm{R}(\alpha)$ on $[a,b]$.*

Theorem 5.10. *Suppose $f \in \mathrm{R}(\alpha)$ on $[a,b]$, $m \leq f(x) \leq M$, $\forall\ x \in [a,b]$, ϕ is continuous on $[m,M]$ and $h(x) = \phi(f(x))$ on $[a,b]$. Then $h \in \mathrm{R}(\alpha)$ on $[a,b]$.*

Theorem 5.11. *If f is monotonic on $[a,b]$ and if α is continuous and monotonic increasing on $[a,b]$, then $f \in \mathrm{R}(\alpha)$.*

Proof. Let $\varepsilon > 0$ be given. Since α is monotonic increasing on $[a,b]$, it follows that $\sup_{x\in[a,b]} \alpha(x) = \alpha(b)$ and $\inf_{x\in[a,b]} \alpha(x) = \alpha(a)$. Again since α is continuous on $[a,b]$, so α attains every value between its bounds $\alpha(a)$ and $\alpha(b)$. Consequently for any positive integer n, choose a partition $P = \{a = x_0, x_1, \ldots, x_{r-1}, x_r, \ldots, x_n = b\}$ of $[a,b]$ such that

$$\delta\alpha_r = \frac{\alpha(b) - \alpha(a)}{n}, \quad for\ \ r = 1, 2, \ldots, n.$$

Let f be monotonic increasing on $[a, b]$.
Let $M_r = \sup_{x\in[x_{r-1},x_r]} f(x)$ and $m_r = \inf_{x\in[x_{r-1},x_r]} f(x)$. Since f is monotonic increasing on $[a, b]$, it follows that $m_r = f(x_{r-1})$ and $M_r = f(x_r)$ for $r = 1, 2, \ldots, n$.
Now

$$\begin{aligned} U(P, f, \alpha) - L(P, f, \alpha) &= \sum_{r=1}^{n} (M_r - m_r)\delta\alpha_r \\ &= \frac{\alpha(b) - \alpha(a)}{n} \sum_{r=1}^{n} [f(x_r) - f(x_{r-1})] \\ &= \frac{\alpha(b) - \alpha(a)}{n} [f(b) - f(a)] \\ &< \varepsilon, \end{aligned}$$

for large n. This shows that $f \in \mathrm{R}(\alpha)$ over $[a, b]$.
Similarly we can prove that $f \in \mathrm{R}(\alpha)$ over $[a, b]$, when f is monotonic decreasing.
□

5.4 First mean value theorem

Theorem 5.12. *If f is continuous function on $[a, b]$ and α is monotonic increasing on $[a, b]$, then there exists a number ξ in $[a, b]$ such that*

$$\int_a^b f \, \mathrm{d}\alpha = f(\xi)\{\alpha(b) - \alpha(a)\}.$$

Proof. Let $M = \sup_{x\in[a,b]} f(x)$ and $m = \inf_{x\in[a,b]} f(x)$. Let $P = \{a = x_0, x_1, \ldots, x_{r-1}, x_r, \ldots, x_n = b\}$ be any partition of $[a, b]$. Let $M_r = \sup_{x\in I_r} f(x)$ and $m_r = \inf_{x\in I_r} f(x)$ for $r = 1, 2, \ldots, n$.
Then

$$m \le m_r \le M_r \le M \quad for \;\; r = 1, 2, \ldots, n$$

$$\Rightarrow m\delta\alpha_r \le m_r\delta\alpha_r \le M_r\delta\alpha_r \le M\delta\alpha_r \quad for \;\; r = 1, 2, \ldots, n$$

$$\Rightarrow \sum_{r=1}^{n} m\delta\alpha_r \le \sum_{r=1}^{n} m_r\delta\alpha_r \le \sum_{r=1}^{n} M_r\delta\alpha_r \le \sum_{r=1}^{n} M\delta\alpha_r$$

$$m\{\alpha(b) - \alpha(a)\} \le L(P, f, \alpha) \le U(P, f, \alpha) \le M\{\alpha(b) - \alpha(a)\}.$$

Since f is continuous and α is monotonic increasing, therefore $f \in \mathrm{R}(\alpha)$ on $[a, b]$ and so

$$L(P, f, \alpha) \leq \int_a^b f \, \mathrm{d}\alpha \leq U(P, f, \alpha).$$

Consequently

$$m\{\alpha(b) - \alpha(a)\} \leq \int_a^b f \, \mathrm{d}\alpha \leq M\{\alpha(b) - \alpha(a)\}.$$

Hence there exists a number μ, with $m \leq \mu \leq M$ such that

$$\int_a^b f \, \mathrm{d}\alpha = \mu\{\alpha(b) - \alpha(a)\}.$$

Again since f is continuous, there exists $\xi \in [a, b]$ such that $f(\xi) = \mu$ and so

$$\int_a^b f \, \mathrm{d}\alpha = f(\xi)\{\alpha(b) - \alpha(a)\}.$$

□

Corollary 5.1. *If $f \in \mathrm{R}(\alpha)$ over $[a, b]$ and $f(x) \geq 0$, for all $x \in [a, b]$, then $\int_a^b f \, \mathrm{d}\alpha \geq 0$.*

Proof. Let $m = \inf_{x \in [a,b]} f(x)$. Since $f(x) \geq 0$, for all $x \in [a, b]$, it follows that $m \geq 0$. We know that $\int_a^b f \, \mathrm{d}\alpha \geq m\{\alpha(b) - \alpha(a)\}$ and so $\int_a^b f \, \mathrm{d}\alpha \geq 0$. □

Corollary 5.2. *Let $f_1, f_2 \in \mathrm{R}(\alpha)$ on $[a, b]$ and $f_1(x) \leq f_2(x)$ for all $x \in [a, b]$, then $\int_a^b f_1 \, \mathrm{d}\alpha \leq \int_a^b f_2 \, \mathrm{d}\alpha$.*

Lemma 5.1. *Let $f : [a, b] \longrightarrow \mathbb{R}$ be bounded on $[a, b]$ and $M = \sup_{x \in [a,b]} f(x)$, $m = \inf_{x \in [a,b]} f(x)$. Then*

$$M - m = \sup\{|f(\alpha) - f(\beta)| : \alpha, \beta \in [a, b]\}.$$

Remark 5.7. *$M_r - m_r = \sup\{|f(\alpha) - f(\beta)| : \alpha, \beta \in [x_{r-1}, x_r]\}$, where $M_r = \sup_{x \in [x_{r-1}, x_r]} f(x)$, $m_r = \inf_{x \in [x_{r-1}, x_r]} f(x)$.*

Theorem 5.13. *If $f \in \mathrm{R}(\alpha)$ on $[a,b]$, then $|f| \in \mathrm{R}(\alpha)$ on $[a,b]$ and $|\int_a^b f \, \mathrm{d}\alpha| \leq \int_a^b |f| \, \mathrm{d}\alpha$.*

Proof. Since $f \in \mathrm{R}(\alpha)$, therefore for a given $\varepsilon > 0$, there exists a partition $P = \{a = x_0, x_1, \ldots, x_n = b\}$ of $[a,b]$ such that

$$U(P, f, \alpha) - L(P, f, \alpha) < \varepsilon. \tag{5.12}$$

Let M_r, m_r and $M_r^{'}, m_r^{'}$ be respectively the supremum and infimum of f and $|f|$ respectively in the sub-interval $[x_{r-1}, x_r]$. Now if ξ_1 and ξ_2 be any two points in $[x_{r-1}, x_r]$, then we have

$$\Big||f(\xi_1)| - |f(\xi_2)|\Big| \leq |f(\xi_1) - f(\xi_2)|. \tag{5.13}$$

From (5.13) we see that

$$\sup\{\Big||f(\xi_1)| - |f(\xi_2)|\Big| : \xi_1, \xi_2 \in [x_{r-1}, x_r]\} \leq \sup\{|f(\xi_1) - f(\xi_2)| : \xi_1, \xi_2 \in [x_{r-1}, x_r]\}$$

$$\Rightarrow M_r^{'} - m_r^{'} \leq M_r - m_r$$

$$\Rightarrow \sum_{r=1}^{n} (M_r^{'} - m_r^{'}) \delta\alpha_r \leq \sum_{r=1}^{n} (M_r - m_r) \delta\alpha_r$$

$$\Rightarrow U(P, |f|, \alpha) - L(P, |f|, \alpha) \leq U(P, f, \alpha) - L(P, f, \alpha).$$

Then from (5.12) we have $U(P, |f|, \alpha) - L(P, |g|, \alpha) < \varepsilon$ and so $|f| \in \mathrm{R}(\alpha)$ on $[a,b]$. We know that

$$-|f(x)| \leq f(x) \leq |f(x)|, \quad \forall\, x \in [a,b].$$

Since $f, |f|, -|f| \in \mathrm{R}(\alpha)$ on $[a,b]$, it follows that

$$-\int_a^b |f| \mathrm{d}\alpha \leq \int_a^b f \mathrm{d}\alpha \leq \int_a^b |f| \mathrm{d}\alpha$$

and so $|\int_a^b f \, \mathrm{d}\alpha| \leq \int_a^b |f| \, \mathrm{d}\alpha$. □

Remark 5.8. *The converse of the theorem is not true. For example, let $f : [0,1] \longrightarrow \mathbb{R}$ be defined by*

$$f(x) = \begin{cases} 1, & \text{if } x \text{ is rational} \\ -1, & \text{if } x \text{ is irrational} \end{cases}$$

and $\alpha(x) = x$. Clearly $f \notin \mathrm{R}(\alpha)$ on $[0,1]$. Note that $|f(x)| = 1$ for all $x \in [0,1]$ and so $|f| \in \mathrm{R}(\alpha)$ on $[0,1]$.

Theorem 5.14. *If $f \in \mathrm{R}(\alpha)$ on $[a,b]$ and $g \in \mathrm{R}(\alpha)$ on $[a,b]$, then $fg \in \mathrm{R}(\alpha)$ on $[a,b]$.*

Proof. Since f and g are bounded on $[a,b]$, there exists $M > 0$, such that $|f(x)| < M$ and $|g(x)| < M$, for all $x \in [a,b]$.
Since $f \in \mathrm{R}(\alpha)$ and $g \in \mathrm{R}(\alpha)$, for a given $\varepsilon > 0$, there exists a partition $P = \{a = x_0, x_1, \ldots, x_n = b\}$ of $[a,b]$ such that

$$U(P,f,\alpha) - L(P,f,\alpha) < \frac{\varepsilon}{2M} \text{ and } U(P,g,\alpha) - L(P,g,\alpha) < \frac{\varepsilon}{2M}. \quad (5.14)$$

Let M_r, m_r, $M_r^{'}, m_r^{'}$ and $M_r^{''}, m_r^{''}$ be respectively the supremum and infimum of f, g and fg on $[x_{r-1}, x_r]$. Then for any two points $\xi_1, \xi_2 \in [x_{r-1}, x_r]$, we have

$$\begin{aligned} |f(\xi_1)g(\xi_1) - f(\xi_2)g(\xi_2)| &= |f(\xi_1)\{g(\xi_1) - g(\xi_2)\} + g(\xi_2)\{f(\xi_1) - f(\xi_2)\}| \\ &\leq |f(\xi_1)||g(\xi_1) - g(\xi_2)| + |g(\xi_2)||f(\xi_1) - f(\xi_2)| \\ &< M\Big[|f(\xi_1) - f(\xi_2)| + |g(\xi_1) - g(\xi_2)|\Big]. \end{aligned}$$

This shows that

$$\begin{aligned} &\sup\{|f(\xi_1)g(\xi_1) - f(\xi_2)g(\xi_2)| : \xi_1, \xi_2 \in [x_{r-1}, x_r]\} \\ \leq\ & M \sup\{|f(\xi_1) - f(\xi_2)| : \xi_1, \xi_2 \in [x_{r-1}, x_r]\} + M \sup\{|g(\xi_1) - g(\xi_2)| : \xi_1, \xi_2 \in [x_{r-1}, x_r]\}, \end{aligned}$$

i.e.,

$$M_r^{''} - m_r^{''} \leq M[M_r - m_r] + M[M_r^{'} - m_r^{'}]$$

$$\Rightarrow \sum_{r=1}^{n}[M_r^{''} - m_r^{''}]\delta\alpha_r \leq M\sum_{r=1}^{n}[M_r - m_r]\delta\alpha_r + M\sum_{r=1}^{n}[M_r^{'} - m_r^{'}]\delta\alpha_r$$

$$U(P,fg,\alpha) - L(P,fg,\alpha) \leq M[U(P,f,\alpha) - L(P,f,\alpha)] + M[U(P,g,\alpha) - L(P,g,\alpha)].$$

Then from (5.14) we have $U(P,fg,\alpha) - L(P,fg,\alpha) < \varepsilon$. This shows that $fg \in \mathrm{R}(\alpha)$ on $[a,b]$. □

Remark 5.9. *The converse of the theorem is not true. For example, let $f, g : [0,1] \longrightarrow \mathbb{R}$ be defined by*

$$f(x) = \begin{cases} 1, & \text{if } x \text{ is rational} \\ -1, & \text{if } x \text{ is irrational} \end{cases}$$

and

$$g(x) = \begin{cases} -1, & \text{if } x \text{ is rational} \\ 1, & \text{if } x \text{ is irrational} \end{cases}$$

where $\alpha(x) = x$. Clearly $f, g \notin \mathrm{R}(\alpha)$ on $[0,1]$. Note that $f(x)g(x) = -1$ for all $x \in [0,1]$ and so $fg \in \mathrm{R}(\alpha)$ on $[0,1]$.

Theorem 5.15. *If $f \in \mathrm{R}(\alpha)$ on $[a,b]$, then $f^2 \in \mathrm{R}(\alpha)$ on $[a,b]$.*

Proof. Hints:

$$\begin{aligned} |f^2(\xi_1) - f^2(\xi_2)| &= |f(\xi_1) - f(\xi_2)||f(\xi_1) + f(\xi_2)| \\ &\leq |f(\xi_1) - f(\xi_2)|[|f(\xi_1)| + |f(\xi_2)|] \\ &< 2M|f(\xi_1) - f(\xi_2)|. \end{aligned}$$

□

Remark 5.10. *The converse of the theorem is not true. For example, let $f : [0,1] \longrightarrow \mathbb{R}$ be defined by*

$$f(x) = \begin{cases} 1, & \text{if } x \text{ is rational} \\ -1, & \text{if } x \text{ is irrational} \end{cases}$$

and $\alpha(x) = x$. Clearly $f \notin \mathrm{R}(\alpha)$ on $[0,1]$. Note that $f^2(x) = 1$ for all $x \in [0,1]$ and so $f^2 \in \mathrm{R}(\alpha)$ on $[0,1]$.

Theorem 5.16. *If $f \in \mathrm{R}(\alpha)$ on $[a,b]$ and c is a constant, then $cf \in \mathrm{R}(\alpha)$ on $[a,b]$ and*

$$\int_a^b cf \, \mathrm{d}\alpha = c \int_a^b f \, \mathrm{d}\alpha.$$

Theorem 5.17. *If $f_1, f_2 \in \mathrm{R}(\alpha)$ on $[a,b]$, then $f_1 + f_2 \in \mathrm{R}(\alpha)$ on $[a,b]$ and*

$$\int_a^b (f_1 + f_2) \, \mathrm{d}\alpha = \int_a^b f_1 \, \mathrm{d}\alpha + \int_a^b f_2 \, \mathrm{d}\alpha.$$

Proof. Since $f_1, f_2 \in \mathrm{R}(\alpha)$ on $[a,b]$, therefore for a given $\varepsilon > 0$, there exist partitions P_1 and P_2 of $[a,b]$ such that

$$U(P_1, f_1, \alpha) - L(P_1, f_1, \alpha) < \frac{\varepsilon}{2}, \;\; U(P_2, f_2, \alpha) - L(P_2, f_2, \alpha) < \frac{\varepsilon}{2}. \tag{5.15}$$

Let $P = P_1 \cup P_2$. Then P is a refinement of both P_1 and P_2. Now from (5.15) (see Theorem 1) we have

$$U(P, f_1, \alpha) - L(P, f_1, \alpha) < \frac{\varepsilon}{2}, \;\; U(P, f_2, \alpha) - L(P, f_2, \alpha) < \frac{\varepsilon}{2}. \tag{5.16}$$

Let $f = f_1 + f_2$. Corresponding to the partition $P = \{a = x_0, x_1, \ldots, x_n = b\}$, let $m_r^{'}, M_r^{'}$, $m_r^{''}, M_r^{''}$ and m_r, M_r be the infimum and supremum of f_1, f_2 and f respectively on $[x_{r-1}, x_r]$. Then for any two points $\xi_1, \xi_2 \in [x_{r-1}, x_r]$, we have

$$\begin{aligned} |f(\xi_1) - f(\xi_2)| &= |\{f_1(\xi_1) - f_1(\xi_2)\} + \{f_2(\xi_1) - f_2(\xi_2)\}| \\ &\leq |f_1(\xi_1) - f_1(\xi_2)| + |f_2(\xi_1) - f_2(\xi_2)|. \end{aligned}$$

This shows that

$$\begin{aligned} & \sup\{|f(\xi_1) - f(\xi_2)| : \xi_1, \xi_2 \in [x_{r-1}, x_r]\} \\ \leq\ & \sup\{|f_1(\xi_1) - f_1(\xi_2)| : \xi_1, \xi_2 \in [x_{r-1}, x_r]\} + \sup\{|f_2(\xi_1) - f_2(\xi_2)| : \xi_1, \xi_2 \in [x_{r-1}, x_r]\}, \end{aligned}$$

i.e.,

$$M_r - m_r \leq M_r^{'} - m_r^{'} + M_r^{''} - m_r^{''}$$

$$\Rightarrow \sum_{r=1}^{n} [M_r - m_r]\delta\alpha_r \leq \sum_{r=1}^{n} [M_r^{'} - m_r^{'}]\delta\alpha_r + \sum_{r=1}^{n} [M_r^{''} - m_r^{''}]\delta\alpha_r,$$

i.e.,

$$U(P, f, \alpha) - L(P, f, \alpha) \leq U(P, f_1, \alpha) - L(P, f_1, \alpha) + U(P, f_2, \alpha) - L(P, f_2, \alpha).$$

Then from (5.16) we have $U(P, f, \alpha) - L(P, f, \alpha) < \varepsilon$. This shows that $f \in \mathrm{R}(\alpha)$ on $[a, b]$.
Let us now proceed to prove the second part.
Since the upper integral is the infimum of the upper sums, therefore there exists partitions P_1 and P_2 such that

$$U(P_1, f_1, \alpha) < \int_a^b f_1 \mathrm{d}\alpha + \frac{\varepsilon}{2}, \quad U(P_2, f_2, \alpha) < \int_a^b f_2 \mathrm{d}\alpha + \frac{\varepsilon}{2}.$$

If $P = P_1 \cup P_2$, we have

$$U(P, f_1, \alpha) < \int_a^b f_1 \mathrm{d}\alpha + \frac{\varepsilon}{2}, \quad U(P, f_2, \alpha) < \int_a^b f_2 \mathrm{d}\alpha + \frac{\varepsilon}{2}.$$

For such a partition P, we have

$$\int_a^b f \mathrm{d}\alpha \leq U(P, f, \alpha) \leq U(P, f_1, \alpha) + U(P, f_2, \alpha) \leq \int_a^b f_1 \mathrm{d}\alpha + \int_a^b f_2 \mathrm{d}\alpha + \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, we get

$$\int_a^b f \mathrm{d}\alpha \leq \int_a^b f_1 \mathrm{d}\alpha + \int_a^b f_2 \mathrm{d}\alpha. \tag{5.17}$$

Proceeding with $(-f_1)$ and $(-f_2)$ instead of f_1 and f_2 we get

$$\int_a^b f \mathrm{d}\alpha \geq \int_a^b f_1 \mathrm{d}\alpha + \int_a^b f_2 \mathrm{d}\alpha. \tag{5.18}$$

From (5.17) and (5.18) we get

$$\int_a^b f\mathrm{d}\alpha = \int_a^b f_1\mathrm{d}\alpha + \int_a^b f_2\mathrm{d}\alpha.$$

□

Theorem 5.18. *If $f \in \mathrm{R}(\alpha_1)$ and $f \in \mathrm{R}(\alpha_2)$ on $[a,b]$, then $f \in \mathrm{R}(\alpha_1+\alpha_2)$ on $[a,b]$ and*

$$\int_a^b f\ \mathrm{d}(\alpha_1+\alpha_2) = \int_a^b f\ \mathrm{d}\alpha_1 + \int_a^b f\ \mathrm{d}\alpha_2.$$

Proof. Since $f \in \mathrm{R}(\alpha_1)$ and $f \in \mathrm{R}(\alpha_2)$ on $[a,b]$, therefore for a given $\varepsilon > 0$, there exist partitions P_1 and P_2 of $[a,b]$ such that

$$U(P_1,f,\alpha_1) - L(P_1,f,\alpha_1) < \frac{\varepsilon}{2},\ \ U(P_2,f,\alpha_2) - L(P_2,f,\alpha_2) < \frac{\varepsilon}{2}. \tag{5.19}$$

Let $P = P_1 \cup P_2$. Then P is a refinement of both P_1 and P_2. Now from (5.19) (see Theorem 1) we have

$$U(P,f,\alpha_1) - L(P,f,\alpha_1) < \frac{\varepsilon}{2},\ \ U(P,f,\alpha_2) - L(P,f,\alpha_2) < \frac{\varepsilon}{2}. \tag{5.20}$$

Let $\alpha = \alpha_1 + \alpha_2$. Corresponding to the partition $P = \{a = x_0, x_1, \ldots, x_n = b\}$, let m_r and M_r be the infimum and supremum of f on $[x_{r-1}, x_r]$. Note that $\delta\alpha_r = \delta\alpha_{1r} + \delta\alpha_{2r}$ and so from (5.20) we get

$$\begin{aligned} U(P,f,\alpha) - L(P,f,\alpha) &= \sum_{r=1}^n (M_r - m_r)\delta\alpha_r \\ &= \sum_{r=1}^n (M_r - m_r)\delta\alpha_{1r} + \sum_{r=1}^n (M_r - m_r)\delta\alpha_{2r} \\ &= U(P,f,\alpha_1) - L(P,f,\alpha_1) + U(P,f,\alpha_2) - L(P,f,\alpha_2) < \varepsilon. \end{aligned}$$

This shows that $f \in \mathrm{R}(\alpha_1+\alpha_2)$ on $[a,b]$.
Now to prove the second part, we see that

$$\begin{aligned} \int_a^b f\mathrm{d}\alpha &= \inf\{U(P,f,\alpha) : P \in \mathcal{P}[a,b]\} \qquad (5.21) \\ &= \inf\{U(P,f,\alpha_1) + U(P,f,\alpha_2) : P \in \mathcal{P}[a,b]\} \\ &\geq \inf\{U(P,f,\alpha_1) : P \in \mathcal{P}[a,b]\} + \inf\{U(P,f,\alpha_2) : P \in \mathcal{P}[a,b]\} \\ &= \int_a^b f\mathrm{d}\alpha_1 + \int_a^b f\mathrm{d}\alpha_2. \end{aligned}$$

Similarly

$$\int_a^b f\mathrm{d}\alpha = \sup\{L(P,f,\alpha) : P \in \mathcal{P}[a,b]\} \leq \int_a^b f\mathrm{d}\alpha_1 + \int_a^b f\mathrm{d}\alpha_2. \tag{5.22}$$

From (5.21) and (5.22) we have

$$\int_a^b f\mathrm{d}\alpha = \int_a^b f\mathrm{d}\alpha_1 + \int_a^b f\mathrm{d}\alpha_2.$$

□

Theorem 5.19. *If $f \in \mathrm{R}(\alpha)$ on $[a,b]$ and c be a point such that $a < c < b$, then $f \in \mathrm{R}(\alpha)$ on $[a,c]$ as well as on $[c,b]$ and*

$$\int_a^b f\ \mathrm{d}\alpha = \int_a^c f\ \mathrm{d}\alpha + \int_c^b f\ \mathrm{d}\alpha.$$

Remark 5.11. *The converse of the above theorem is not always true. For example, let f and α be defined on $[0,2]$ as follows*

$$f(x) = \begin{cases} 0, & \text{if } 0 \leq x < 1 \\ 1, & \text{if } 1 \leq x \leq 2 \end{cases}$$

and

$$\alpha(x) = \begin{cases} 0, & \text{if } 0 \leq x < 1 \\ 1, & \text{if } 1 \leq x \leq 2 \end{cases}$$

Clearly $\int_0^1 f\ \mathrm{d}\alpha$ and $\int_1^2 f\ \mathrm{d}\alpha$ both exist, but the $\int_0^2 f\ \mathrm{d}\alpha$ does not exist.

5.5 Reduction of Riemann-Stieltjes integral into Riemann Integral

Theorem 5.20. *If $f \in \mathrm{R}[a,b]$ and α be a monotonic increasing function on $[a,b]$ such that $\alpha' \in \mathrm{R}[a,b]$, then $f \in \mathrm{R}(\alpha)$ on $[a,b]$ and $\int_a^b f\ \mathrm{d}\alpha = \int_a^b f(x)\alpha'(x)\mathrm{d}x$.*

Proof. Since $f \in \mathrm{R}[a,b]$, it follows that f is bounded on $[a,b]$ and so there exists a positive real number k such that

$$|f(x)| \leq k, \ \ \forall\ x \in [a,b]. \tag{5.23}$$

Let $\varepsilon > 0$ be given. Since $f, \alpha^{'} \in \mathrm{R}[a,b]$, it follows that $f\alpha^{'} \in \mathrm{R}[a,b]$. Then there exists $\delta_1 > 0$ such that

$$\Big|\sum_{r=1}^{n} f(\xi_r)\alpha^{'}(\xi_r)\delta_r - \int_a^b f\alpha^{'}\mathrm{d}x\Big| < \frac{\varepsilon}{2}, \tag{5.24}$$

for $\mu(P) < \delta_1$ and for all $\xi_r \in I_r$, where $I_r = [x_{r-1}, x_r]$ is the r-th sub-interval of the partition $P = \{a = x_0, x_1, \ldots, x_n = b\}$ of $[a,b]$. Again since $\alpha^{'} \in \mathrm{R}[a,b]$, so there exists $\delta_2 > 0$ such that

$$U(P,\alpha^{'}) - L(P,\alpha^{'}) = \sum_{r=1}^{n}(M_r - m_r)\delta_r < \frac{\varepsilon}{2k}, \tag{5.25}$$

for $\mu(P) < \delta_2$ and $M_r = \sup\limits_{x\in I_r} \alpha^{'}(x)$, $m_r = \inf\limits_{x\in I_r} \alpha^{'}(x)$.
Now for $\mu(P) < \delta_2$ and for all ξ_r, $\eta_r \in I_r$, we have (see Remark 7)

$$M_r - m_r = \sup\{|\alpha^{'}(\xi_r) - \alpha^{'}(\eta_r)| : \xi_r, \eta_r \in I_r\}$$

and so we have

$$|\alpha^{'}(\xi_r) - \alpha^{'}(\eta_r)| \leq M_r - m_r,$$

for all ξ_r, $\eta_r \in I_r$.

$$\Rightarrow \sum_{r=1}^{n} |\alpha^{'}(\xi_r) - \alpha^{'}(\eta_r)|\delta_r \leq \sum_{r=1}^{n}(M_r - m_r)\delta_r, \tag{5.26}$$

for all ξ_r, $\eta_r \in I_r$. Now from (5.25) and (5.26) we have

$$\sum_{r=1}^{n} \Big|\alpha^{'}(\xi_r) - \alpha^{'}(\eta_r)\Big|\delta_r < \frac{\varepsilon}{2k}, \tag{5.27}$$

for $\mu(P) < \delta_2$ and for all $\xi_r, \eta_r \in I_r$.
Let $\delta = \min\{\delta_1, \delta_2\}$. Let P be a partition with $\mu(P) < \delta$ and $\xi_r \in [x_{r-1}, x_r]$. Then by Lagrange's Mean Value theorem, there exists $\eta_r \in (x_{r-1}, x_r)$ such that

$$\alpha^{'}(\eta_r) = \frac{\alpha(x_r) - \alpha(x_{r-1})}{x_r - x_{r-1}} = \frac{\delta\alpha_r}{\delta_r},$$

i.e.,

$$\delta\alpha_r = \alpha^{'}(\eta_r)\delta_r. \tag{5.28}$$

Now from (5.23), (5.24), (5.27) and (5.28) we have

$$\begin{aligned}
&\Big|\sum_{r=1}^{n} f(\xi_r)\delta\alpha_r - \int_a^b f\alpha'\ \mathrm{d}x\Big| \\
=\ &\Big|\sum_{r=1}^{n} f(\xi_r)\alpha'(\eta_r)\delta_r - \int_a^b f\alpha'\ \mathrm{d}x\Big| \\
=\ &\Big|\sum_{r=1}^{n} f(\xi_r)\alpha'(\xi_r)\delta_r - \int_a^b f\alpha'\ \mathrm{d}x + \sum_{r=1}^{n} f(\xi_r)\{\alpha'(\eta_r) - \alpha'(\xi_r)\}\delta_r\Big| \\
\leq\ &\Big|\sum_{r=1}^{n} f(\xi_r)\alpha'(\xi_r)\delta_r - \int_a^b f\alpha'\ \mathrm{d}x\Big| + \sum_{r=1}^{n}\Big|f(\xi_r)\Big|\Big|\alpha'(\eta_r) - \alpha'(\xi_r)\Big|\delta_r \\
<\ &\frac{\varepsilon}{2} + k\frac{\varepsilon}{2k} = \varepsilon.
\end{aligned}$$

Hence for any $\varepsilon > 0$, there exists $\delta > 0$ such that for all partitions with $\mu(P) < \delta$

$$\Big|\sum_{r=1}^{n} f(\xi_r)\delta\alpha_r - \int_a^b f\alpha'\ \mathrm{d}x\Big| < \varepsilon$$

$$\Rightarrow \lim_{\mu(P)\to 0}\sum_{r=1}^{n} f(\xi_r)\delta\alpha_r \quad \textit{exists and equals to} \quad \int_a^b f\alpha'\ \mathrm{d}x$$

$$\Rightarrow f \in \mathrm{R}(\alpha) \ \textit{ on } \ [a,b] \ \textit{ and } \ \int_a^b f\ \mathrm{d}\alpha = \int_a^b f(x)\alpha'(x)\mathrm{d}x.$$

□

Theorem 5.21. *If f is continuous on $[a,b]$ and α has a continuous derivatives on $[a,b]$, then $f \in \mathrm{R}(\alpha)$ on $[a,b]$ and $\int_a^b f\ \mathrm{d}\alpha = \int_a^b f(x)\alpha'(x)\mathrm{d}x$.*

Example 5.4. *Evaluate (i) $\int_0^1 x\ \mathrm{d}(e^{2x})$ and (ii) $\int_0^2 [x]\ \mathrm{d}(x^2)$.*

Proof. (i) $\int_0^1 x\ \mathrm{d}(e^{2x}) = \int_0^1 x.2e^{2x}\ \mathrm{d}x = [x.e^{2x}]_0^1 - \int_0^1 1.e^{2x}\ \mathrm{d}x = \frac{1+e^2}{2}$.
(ii) $\int_0^2 [x]\ \mathrm{d}(x^2) = \int_0^2 [x].2x\ \mathrm{d}x = \int_0^1 [x].2x\ \mathrm{d}x + \int_1^2 [x].2x\ \mathrm{d}x = 3.$ □

Integration by Parts

Theorem 5.22. *If $f \in \mathrm{R}(\alpha)$ on $[a,b]$, then $\alpha \in \mathrm{R}(f)$ on $[a,b]$ and we have*

$$\int_a^b f(x)\,\mathrm{d}\alpha(x) = [f(b)\alpha(b) - f(a)\alpha(a)] - \int_a^b \alpha(x)\,\mathrm{d}f(x).$$

Second Mean Value Theorem

Theorem 5.23. *If f is monotonic increasing and α is continuous on $[a,b]$, then there exists a point $\xi \in [a,b]$ such that*

$$\int_a^b f(x)\,\mathrm{d}\alpha(x) = f(a)\int_a^\xi \mathrm{d}\alpha(x) + f(b)\int_\xi^b \mathrm{d}\alpha(x).$$

Example 5.5. *Evaluate* $\int_0^3 x^2\,\mathrm{d}([x]-x)$.

Proof.

$$\begin{aligned}
\int_0^3 x^2\,\mathrm{d}([x]-x) &= \Big[x^2([x]-x)\Big]_0^3 - \int_0^3 2x([x]-x)\,\mathrm{d}x \\
&= 0 - 2\int_0^3 x[x]\,\mathrm{d}x + 2\int_0^3 x^2\,\mathrm{d}x \\
&= -2\Big[\int_0^1 x[x]\,\mathrm{d}x + \int_1^2 x[x]\,\mathrm{d}x + \int_2^3 x[x]\,\mathrm{d}x\Big] + 18 \\
&= \cdots\cdots \\
&= 5.
\end{aligned}$$

□

Definition 5.2. *A function α defined on $[a,b]$ is called a step function if there is a partition $P = \{a = x_0, x_1, \ldots, x_n = b\}$ such that α is constant on each open subinterval (x_{r-1}, x_r). The number $\alpha(x_r+) - \alpha(x_r-)$ is called the jump at x_r, if $1 \le r \le n-1$. The jump at x_0 is $\alpha(x_0+) - \alpha(x_0)$ and the jump at x_n is $\alpha(x_n) - \alpha(x_n-)$.*

Theorem 5.24. *If f is continuous on $[a,b]$ and α is a step function such that α is constant on each subinterval $(x_{r-1}, x_r)(r = 0, 1, \ldots, n)$ where $a = x_0 < x_1 < \ldots < x_n = b$,*

$$\int_a^b f\,\mathrm{d}\alpha = \sum_{r=0}^{n} f(x_r)\Big[\alpha(x_r+) - \alpha(x_r-)\Big],$$

provided $\alpha(x_0-) = \alpha(x_0) = \alpha(a)$ and $\alpha(x_n+) = \alpha(x_n) = \alpha(b)$.

Example 5.6. *Evaluate* $\int_0^5 (x^2+1)\ \mathrm{d}[x]$.

Proof. Hints. Here $f(x) = x^2 + 1$ and $\alpha(x) = [x]$, for all $x \in [0, 5]$.
Note that $\alpha(x) = 0$, if $0 \leq x < 1$, $= 1$, if $1 \leq x < 2$, $= 2$, if $2 \leq x < 3$, $= 3$, if $3 \leq x < 4$, $= 4$, if $4 \leq x < 5$, $= 5$, if $x = 5$.
Clearly $\int_0^5 f(x)\ \mathrm{d}\alpha(x) = \sum_{r=0}^{5} f(r)\Big[\alpha(r+) - \alpha(r-)\Big]$, where $\alpha(0-) = \alpha(0) = 0$ and $\alpha(5+) = \alpha(5) = 5$.
Ans.=60. □

Summary

In this chapter we have learnt about the Riemann-Stieltjes Integral and its various properties; reduction into Riemann Integrals and have also done relevant examples.

Units 9 & 10

Course Structure

1. Definition of a delta-fine tagged partition and its existence
2. Lebesgue's criterion for Riemann integrability
3. Delta-fine free tagged partition and an equivalent definition of the Riemann integral

6 Introduction

Let an interval $[a,b] \subset \mathbb{R}$, $-\infty < a < +\infty$ be given. A division of $[a,b]$ is a finite collection of non overlapping closed intervals whose union is $[a,b]$. A pair (c,J) of point $c \in \mathbb{R}$ and an associated closed interval $J \in \mathbb{R}$ is called a tagged point interval pair where the point c is called the tag of J. A finite collection $\Delta = \{(c_i, J_i) : 1 \leq i \leq n\}$ of tagged point interval pairs is called a tagged system in $[a,b]$ if $c_i \in J_i \subset [a,b]$ for each $i = 1,2,\ldots,n$ and the intervals are non overlapping. A tagged system $\Delta = \{(c_i, j_i) : 1 \leq i \leq n\}$ is called a tagged partition of $[a,b]$ if $\cup_{i=1}^{n} J_i = [a,b]$.

Given a positive function δ defined on $[a,b]$ $\delta : [a,b] \longrightarrow (0,\infty)$ is called a Guage function on $[a,b]$. A tagged interval (c,J) with $c \in [a,b]$ and $J \subset [a,b]$ is said to be δ-fine if $c \in J \subset (c - \delta(c), c + \delta(c))$.

A tagged partition $\Delta = \{(c_i, J_i) : 1 \leq i \leq n\}$ is δ-fine if the point interval pair (c_i, J_i) is δ-fine, for each $i = 1,2,\ldots,n$ and is denoted by $\{c_i, [x_{i-1}, x_i] : 1 \leq i \leq n\}$, where $a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b$ and $c_i \in [x_{i-1}, x_i] \subset (c_i - \delta(c_i), c_i + \delta(c_i))$ for each index i.

Guage determines the size of the interval associated with a given tag. In a δ-fine tagged partion, the tags must be choosen first, then intervals of the rigiud size are choosen for each tag.

If the infimum of the set $\{\delta(x) : x \in [a,b]\}$ is positive, then it is clear that δ-fine tagged partition of $[a,b]$ exists, but if the infimum is zero, then the proof of the existence of δ-fine tagged partition is required.

Note: If δ' is a gauge finer than δ, i.e., $0 < \delta'(x) \leq \delta(x)$ for all x, then every δ' fine tagged partition is also δ-fine. In fact if $(x,[u,v])$ is a point interval pair in a δ'-fine partition, then

$$x \in [u,v] \subset (x - \delta', x + \delta') \subset (x - \delta(x), x + \delta(x))$$

shows that $(x,[u,v])$ may also be a point interval pair in some δ-fine partition.

Example 6.1. *Suppose δ is a positive function defined on $[0,1]$ by*

$$\delta(x) = \begin{cases} \frac{x}{2}, \ if \ 0 < x \leq 1 \\ 1, if \ x = 0. \end{cases}$$

Then $inf\{\delta(x) : 0 \leq x \leq 1\} = 0$.

We note that $(x - \delta(x), x + \delta(x))$ *does not contain zero unless* $x = 0$. *Consequently, any* δ*-fine tagged partition of* $[0, 1]$ *must have* 0 *as atag. Here infimum of* $\{\delta : x \in [a, b]\}$ *is not a value assumed by* $\delta(x)$.

Causin's Lemma:

If δ is a positive function defined on $[a, b]$, then there exists a δ-fine tagged partition on $[a, b]$.

Note: If δ is a Gauge finer than δ, i.e., $0 < \delta^{'}(x) \leq \delta(x)$ for all x, then every $\delta^{'}$ fine tagged partition is also δ-fine. In fact if $(x, [u, v])$ is a point interval pair in a $\delta^{'}$ fine partition, then $x \in [u, v] \subset [x - \delta^{'}(x), x + \delta^{'}(x)] \subset (x - \delta(x), x + \delta(x))$ shows that $(x, [u, v])$ may also be a point interval pair in some δ-fine partition.

6.1 Application of Tagged Gauge partition:

We know that a funcytion f to be $\mathbb{R}$-integrable (Reimann) on $[a, b]$ iff for each $\epsilon(> 0)$, there exists a partition $\{[x_{i-1}, x_i] : 1 \leq i \leq n\}$ on $[a, b]$ such that $\sum_{i=1}^{n} \omega(f, [x_{i-1}, x_i])(x_i - x_{i-1}) < \epsilon$, where the oscillation $\omega(f, [c, d])$ of the function f on the interval $[c, d]$ is defined by $\omega(f, [c, d]) = sup\{|f(t) - f(s)| : t, s \in [c, d]\}$.

Theorem 6.1. *If* f *is bounded and continuous function on* $[a, b]$, *then* f *is R-integrable on* $[a, b]$.

Proof. Let $\epsilon(> 0)$ be arbitrary. Since f is continuous on $[a, b]$ for each $x \in [a, b]$, there exists $\delta(x) > 0$ such that $|f(t) - f(x)| < \epsilon$ for all $t \in [a, b]$ satisfying $|t - x| < \delta(x)$.

This denotes a positive function δ on $[a, b]$.

By Caucin's lemma, let $\{(\xi_i, [x_{i-1}, x_i]) : 1 \leq i \leq n\}$ be a δ-fine tagged partition on $[a, b]$.

For $s, t \in [x_{i-1}, x_i]$, $i = 1, 2, \ldots, n$, we have

$$\begin{aligned} |f(t) - f(s)| &\leq |f(t) - f(\xi_i)| + |f(\xi_i) - f(s)| \\ &< 2\epsilon, \end{aligned}$$

for each index $s, t \in [x_{i-1}, x_i]$, $i = 1, 2, \ldots, n$.
It follows that

$$\sum_{i=1}^{n} \omega(f, [x_{i-1}, x_i])(x_{[}i] - x_{i-1}) \leq 2\epsilon \sum_{i=1}^{n} (x_i - x_{i-1}) = 2\epsilon(b - a).$$

Hence, f is Riemann integrable. □

We recall that a set E has measure zero if for each $\epsilon > 0$, there exists a sequence $\{I_k\}$ of open intervals such that $E = \bigcup_{k=1}^{\infty} I_k$ and $\sum_{k=1}^{\infty} |I_k| < \epsilon$, where $|I_k|$ denotes the length of the interval I_k. A property is said to hold almost everywhere if it fails to hold only at points in a set of measure zero.

6.2 Lebesgue criteria for Riemann integrability:

Theorem 6.2. *If f is bounded and continuous everywhere on $[a,b]$, then f is $\mathcal{R}$-integrable on $[a,b]$.*

Proof. Let $M(>0)$ be an upper bound for the function $|f|$ on $[a,b]$ and let D be the set of all points $x \in [a,b]$ such that f is not continuous at x. Let $\epsilon > 0$ be arbitrary. Since D has measure zero, there exists a sequence $\{I_k\}$ of open intervals such that $D \subset \bigcup_{k=1}^{\infty} I_k$ and $\sum_{k=1}^{\infty} |I_k| < \frac{\epsilon}{2M}$.

If $x \in D$, choose $\delta(x) > 0$ so that $(x-\delta(x), x+\delta(x)) \subset I_k$ some index k. If $x \notin D$, then by continuity of f at x, we can choose $\delta(x) > 0$ so that $|f(t)-f(x)| < \frac{\epsilon}{2}$ for all $t \in [a,b]$ that satisfy $|t-x| < \delta(x)$.
This defines a positive function δ on $[a,b]$. So, by Causin's lemma let $\{(\xi_i, [x_{i-1}, x_i]) : 1 \le i \le n\}$ be a δ-fine tagged partition of $[a,b]$. Define $S_0 = \{i : \xi_i \notin D\}$ and $S_D = \{i : \xi_i \in D\}$.

Since the intervals are non-overlapping, we see that

$$\begin{aligned}
\sum_{i=1}^{n} \omega(f,[x_{i-1},x_i])(x_i - x_{i-1}) &= \sum_{i\in S_0} \omega(f,[x_{i-1},x_i])(x_i - x_{i-1}) \\
&\quad + \sum_{i\in S_D} \omega(f,[x_{i-1},x_i])(x_i - x_{i-1}) \\
&\le \epsilon \sum_{i\in S_0} (x_i - x_{i-1}) + 2M \sum_{i\in S_D} (x_i - x_{i-1}) \\
&\le \epsilon(b-a) + 2M \sum_{k=1}^{\infty} |I_k| \\
&< \epsilon(b-a+1).
\end{aligned}$$

Hence, the function f is $\mathcal{R}$-integrable on $[a,b]$. □

Definition of HK integral:

Let f be a real valued function defined on a bounded closed interval $[a,b]$, where $a < b$. For any gauge δ, we define

$$(\delta)\int_a^{\bar{b}} = sup\{\sum_i f(x_i)(b_i - a_i) : \{x_i, [a_i, b_i]\} \; is \; a \; \delta \; fine \; tagged \; partition \; of \; [a,b]\},$$

$$(\delta)\int_{\underline{a}}^{b} = inf\{\sum_i f(x_i)(b_i - a_i) : \{x_i, [a_i, b_i]\} \; is \; a \; \delta \; fine \; tagged \; partition \; of \; [a,b]\}.$$

Considering all gauges δ, we define

$$(HK)\int_a^{\bar{b}} f = inf_\delta(\delta)\int_a^{\bar{b}} f$$

and

$$(HK)\int_{\underline{a}}^{b} f = sup_{\delta}(\delta)\int_{\underline{a}}^{b} f.$$

We call $(HK)\int_a^{\overline{b}} f$, the upper HK integral of f on $[a,b]$ and $(HK)\int_{\underline{a}}^{b} f$, the lower HK integral of f on $[a,b]$.

If $(HK)\int_a^{\overline{b}} f = (HK)\int_{\underline{a}}^{b} f = \alpha(say) \neq \pm\infty$, then the function f is said to be HK-integrable on $[a,b]$ and the common finite value α is called the definite HK integral of f on $[a,b]$ written as $(HK)\int_a^b f = \alpha$.

Definition 6.1. *A function $f : [a,b] \longrightarrow \mathbb{R}$ is said to be Henstock integrable to the real number I on $[a,b]$ if for every $\epsilon(>0)$, there exists a positive function δ on $[a,b]$, $\delta : [a,b] \longrightarrow (0,\infty)$ such that for every δ-fine tagged partition $P = \{(\xi_i, [x_{i-1}, x_i]) : 1 \leq i \leq n\}$ of $[a,b]$, the inequality $\left|\sum_{i=1}^{n} f(\xi_i)(x_i - x_{i-1}) - I\right| < \epsilon$ holds.*

If the integral exists, we write $f \in H[a,b]$ and the real number I is called the Henstock integral of f on $[a,b]$ and we write $I = (H)\int_a^b f(x)dx$ or simply $I = (H)\int_a^b f$.

For simplicity, we shall often write $\sum_{i=1}^{n} f(\xi_i)(x_i - x_{i-1}) = S(P,f)$ corresponding to the δ fine tagged partition $P = \{(\xi_i, [x_{i-1}, x_i]) : 1 \leq i \leq n\}$ of $[a,b]$ and the integrals agree.

Theorem 6.3. *If f is R-integrable on $[a,b]$, then f is H-integrable on $[a,b]$ and the two integrals agree.*

Proof. Let $\epsilon(>0)$ be arbitrary. Then there exists a real number I and a positive constant δ_1 such that for every Riemann partition $P = \{[y_{i-1}, y_i], \eta_i : i = 1, 2, \ldots, m\}$ of $[a,b]$ with $||P|| < \delta_1$, we have

$$\Big|\sum_{i=1}^{n} f(\eta_i)(y_i - y_{i-1}) - I\Big| < \epsilon.$$

□

Define a positive function δ on $[a,b]$ $\delta : [a,b] \longrightarrow (0,\infty)$ by $\delta(x) = \delta_1$ for all $x \in [a,b]$. Clearly, every δ-fine tagged partition of $[a,b]$ is a Riemann δ_1-tagged partition of $[a,b]$. Consequently, every Henstock sum $S(P,f)$ is a Riemann sum corresponding to Riemanian $\delta = \delta_1$ partition P_1 and we have

$$|S(P_1, f) - I| < \epsilon.$$

This shows that $f \in H[a,b]$ and

$$(H)\int_a^b f(x)dx = I = (R)\int_a^b f(x)dx.$$

Example 6.2. *Consider the function f defined on $[0,1]$ by*

$$f(x) = \begin{cases} 1, if\ x\ is\ rational \\ 0, if,\ x\ is\ irrational. \end{cases}$$

Clearly, $(H)\int_0^1 f(x)dx = 0$ and $(R)\int_0^1 f(x)dx = 1$. Therefore, f does not belong to $R[0,1]$.
Let $\epsilon(>0)$ be given and let $\{r_i : i = 1, 2, \ldots, n, \ldots\}$ be an enumarable of rationals in $[0,1]$.

Define $\delta(r_i) = \frac{\epsilon}{2^{i+1}}$ and $\delta(x) = 1$, $x \neq r_i$.
Let $P = \{(\xi_i, [u_i, v_i]) : i = 1, 2, \ldots, n\}$ be an arbitrary δ-fine partition of $[0,1]$. Therefore,

$$\begin{aligned} 0 \leq S(P,f) &= \sum_{\xi_i \in \mathbb{Q}\cap[0,1]} f(\xi_i)(v_i - u_i) + \sum_{\xi_i \notin \mathbb{Q}\cap[0,1]} f(\xi_i)(v_i - u_i) \\ &= \sum_{\xi_i \in \mathbb{Q}\cap[0,1]} f(\xi_i)(v_i - u_i) + 0 \\ &= \sum_{\xi_i \in \mathbb{Q}\cap[0,1]} (v_i - u_i) \\ &< \sum_i 2\delta(\xi_i) = \sum_i 2\frac{\epsilon}{2^{i+1}} \\ &= \sum_i \frac{\epsilon}{2^i} \\ &= \epsilon[\frac{1}{2} + \frac{1}{2^2} + \ldots] \\ &= \frac{\epsilon}{2}\frac{1}{1-\frac{1}{2}} = \epsilon. \end{aligned}$$

Therefore, $f \in H[a,b]$. The value of the integral is zero.

Though the function is Henstock integrable, but it is not Riemann integrable.

Summary

In this chapter, we have learnt about delta-fine tagged partition and its properties, its applications, Lebesgue's criterion for Riemann integrability and various related theorems and examples.

References

[1] T. M. Apostol : Mathematical Analysis.

[2] S. C. Malik, S. Arora : Mathematical Analysis.

[3] W.F. Trench : Introduction to Real Analysis.

[4] W. Rudin : Principles of Mathematical Analysis.

Block II
Complex Analysis I

Unit 11

Course Structure

1. Point at infinity and the extended complex plane
2. Riemann's sphere

1 Point at infinity

By means of the transformation $w = \frac{1}{z}$, the point $z = 0$, that is, the origin is mapped into $w = \infty$, called the point at infinity in the w-plane. Similarly we denote by $z = \infty$, the point at infinity in the z-plane.

To consider the behaviour of $f(z)$ at $z = \infty$, it suffices to let $z = \frac{1}{w}$ and examine the behaviour of $f\left(\frac{1}{w}\right)$ at $w = 0$.

2 The Extended Complex Plane

By the extended complex number system, we shall mean the complex plane $\mathbb{C}$ along with ∞, the point at infinity, which satisfies the following properties:

1. If $z \in \mathbb{C}$, then we have,

$$z + \infty = z - \infty = \infty$$

and

$$\frac{z}{\infty} = 0$$

2. If $z \in \mathbb{C}$, and $z \neq 0$, then $z.\infty = \infty$ and $\frac{z}{0} = \infty$.
3. $\infty + \infty = \infty$ and $\infty.\infty = \infty$.
4. $\frac{\infty}{z} = \infty$, $(z \neq \infty)$.

Then the set $\mathbb{C} \cup \infty$ is called the extended complex plane.

3 Stereographic Projection

Let $\mathbb{C}$ be the complex plane and consider a unit sphere $\overline{S}$ of radius 1 tangent to $\mathbb{C}$ at $z = 0$. The diameter NS is perpendicular to $\mathbb{C}$ and we call the points n and S, the north and south poles respectively. For any point A on $\mathbb{C}$, we can construct a line NA joining N and A, which intersects $\overline{S}$ at the point A'. Thus, to each point A of the complex plane, there corresponds a unique point A' on the sphere $\overline{S}$, and we can represent any complex number by a point on the sphere. For completeness, we say that the point N itself corresponds to

the point at infinity of the extended complex plane. The set of all points of the complex plane including the point at infinity is called the entire complex plane or the entire z-plane. This method of mapping the plane onto the sphere is called stereographic projection. The sphere is sometimes called the Riemann sphere.

Unit 12

Course Structure

1. Functions of a complex variable
2. Limit and Continuity
3. Analytic functions
4. Cauchy-Riemann Equations

4 Functions, Limit and Continuity

Definition 1. Any collection of point in the complex plane is called a point set and each point is called an element of the set.

Definition 2. A neighbourhood of a point $z_0 \in \mathbb{C}$ is the set of all points z such that $|z - z_0| < r(r > 0)$, that is, the set of all points lying in the disc with centre z_0 and radius r. Take deleted neighbourhood of a point $z_0 \in \mathbb{C}$ is a neighbourhood of z_0 in which z_0 is omitted. The set of all points z such that $|z| > k$, where k is any positive real number, is called the neighbourhood of the point at infinity.

Definition 3. A point z_0 is called the limit point of a set S in the complex plane if every deleted neighbourhood of z_0 contains at least one point of S. A limit point may or may not belong to the set.
We consider the set of points defined by $|z| < r$. Evidently, all points on the circle $|z| = r$ are the limit points of the set, but they do not belong to the set. Again, all the points within the circle $|z| = r$ are also limit points of the set defined $|z| < r$ and they belong to the set.

Definition 4. A point $z_0 \in S$ is called an interior point of the set S if there exists a neighbourhood of z_0 contained entirely within S.

Definition 5. A point $z_0 \in S$ is called an exterior point of the set S if there exists a neighbourhood of z_0 which contains no point of S.

Definition 6. A point $z_0 \in S$ is called the boundary point of the set S if every neighbourhood of z_0 contains atleast one point of S and at least one point outside S. The collection of all boundary points of S is called the boundary of S.

Definition 7. A set S in the complex plane is said to be open if it consists only of its interior points. For example, the open disc $|z - z_0| < r$ is an open set.

Definition 8. A set is called closed if its complement is open. Equivalently, a set S is said to be closed if every limit point of S belongs to S, or if S has no limit point.

There are sets which are neither open nor closed. For example, the set $\{z \in \mathbb{C} : |z| < 1\} \cup \{1\}$ is neither open nor closed.

Definition 9. A set of points S is said to be bounded if there exists a positive number M such that $|z| < M \ \forall z \in S$. If there exists no such M, the set S is said to be unbounded.

Definition 10. A set which is bounded and closed is called compact set.

Definition 11. The set of all limit points of a set S is called the derived set of S and is denoted by S'.

The union of a set S and its derived set S' is called the closure of S and is denoted by $\overline{S}$ or $cl(S)$.

Then, $\overline{S} = S \cup S'$.

Definition 12. (**Connected Set**) A set is said to be connected if any two of its points can be joined by a polygon which completely lies inside the set.

4.1 Jordan Curve

Definition 13. (**Open and Closed Regions**) An open connected set is called a domain or an open region. If however the boundary points are also included, then it is called a closed region or a closed domain.

Definition 14. The equation $z = z(t) = x(t) + iy(t)$ where $x(t)$ and $y(t)$ are real continuous functions of a real variable t, defined in the interval $a \leq t \leq b$, determines a set of points in the complex plane which we call a continuous arc.

The equation $z = z(t) = x(t) + iy(t)$ determines a simple arc if $t_1 \neq t_2$ implies $z(t_1) \neq z(t_2)$.

The equation $z = z(t) = x(t) + iy(t)$ determines a simple closed curve if $t_1 < t_2$ and $z(t_1) = z(t_2)$ implies $t_1 = a$ and $t_2 = b$.

Simple arcs and simple closed arcs are often called Jordan arcs and Jordan curves respectively. A simple example of a Jordan arc is the polygonal arc which consists of a finite number of line segments.

Theorem 1. (Jordan Curve Theorem) A Jordan curve divides the complex plane into two regions having the curve as a common boundary. The region which is bounded is called the interior of the curve and the other region is called the exterior of the curve.

Theorem 2. (Bolzano Weierstrass Theorem) If a set is bounded and contains infinitely many points then it possesses at least one limit point.

Definition 15. (**Variables and Functions**)A symbol z which can stand for any one of a set of complex numbers is called a complex variable.

If for each value of z there corresponds one or more values of a complex variable w, we say that w is a function of z and we write $w = f(z)$.

The variable z is called an independent variable, while w is called a dependent variable.

Definition 16. (**Single and Multi – valued Functions**) If any one value of w corresponds to each value of z, we say that w is a single-valued function of z or that $f(z)$ is single-valued.

If more than one value of w corresponds to each value of z, we say that w is a multi-valued function or many valued function of z.

A multiple valued function can be considered as a collection of single valued functions, each member of which is called a branch of the function.

Examples 1.

1. If $w = z^2$, then to each value of z there is only one value of w. Hence $w == f(z) = z^2$ is a single valued fucntion of z.
2. If $w = \sqrt{z}$, then to each value of z, there are two values of w. Hence, $w = f(z) = \sqrt{z}$, is a multiple valued function of z.
3. If $w = z^{1/p}$, where p is any natural number, then to each value of z, there are p values of w. Hence, $w = f(z) = z^{1/p}$, is a multiple valued function of z.

Definition 17. (**Rational Function**) Rational functions are defined by

$$w = \frac{P(z)}{Q(z)}$$

where $P(z)$ and $Q(z)$ are polynomials in z and $Q(z) \neq 0$.

Definition 18. (**Limit of a function**) Let $w = f(z)$ be defined in a domain D except perhaps of the point z_0 of D. A complex number l is said to be a limit of f as $z \to z_0$, or symbolically,

$$l = \lim_{z \to z_0} f(z)$$

if for given $\epsilon > 0$, there exists a $\delta > 0$ such that

$$|f(z) - l| < \epsilon \; whenever \; 0 < |z - z_0| < \delta$$

If no such l exists we say that $\lim_{z \to z_0} f(z)$ does not exist. Note that, z is allowed to approach z_0 in an arbitrary manner, not just from some particular direction. The limit is clearly independent of the path by which z approaches z_0.

Geometrically, if z_0 is a point in the complex plane, $\lim_{z \to z_0} f(z) = l$ if the difference in absolute value between $f(z)$ and l can be made as small as we wish. By choosing points z sufficiently close to z_0(excluding $z = z_0$ itself).

Theorem 3. A necessary and sufficient condition that the function $f(z) = u+iv$ may tend to $l = \alpha + i\beta$ as $z = x + iy$ tends to $z_0 = a + ib$ is that

$$u(x, y) \to \alpha \; and \; v(x, y) \to \beta \; as \; (x, y) \to (a, b)$$

Proof. We first suppose that $\lim_{z\to z_0} f(z) = l$. Then, for given $\epsilon > 0$, there exists a $\delta > 0$ such that

$$|f(z) - l| < \epsilon \text{ whenever } 0 < |z - z_0| < \delta$$
$$\text{or, } |u(x,y) + iv(x,y) - \alpha - i\beta| < \epsilon \text{ whenever } 0 < |x + iy - a - ib| < \delta$$
$$\text{or, } |(u(x,y) - \alpha) + i(v(x,y) - \beta)| < \epsilon \text{ whenever } 0 < |(x-a) + i(y-b)| < \delta$$
$$\text{or, } |u(x,y) - \alpha| < \epsilon \text{ and } |v(x,y) - \beta| < \epsilon \text{ whenever } 0 < \sqrt{(x-a)^2 + (y-b)^2} < \delta$$

as $|Re(z)| \leq |z|$ and $|Im(z)| \leq |z|$. This implies that

$$\lim_{(x,y)\to(a,b)} u(x,y) = \alpha \text{ and } \lim_{(x,y)\to(a,b)} v(x,y) = \beta$$

This proves the necessary part.

Conversely, let

$$\lim_{(x,y)\to(a,b)} u(x,y) = \alpha \text{ and } \lim_{(x,y)\to(a,b)} v(x,y) = \beta$$

Then, for $\epsilon > 0$ we can find a $\delta > 0$ such that

$$|u(x,y)-\alpha| < \epsilon/2 \text{ and } |v(x,y)-\beta| < \epsilon/2 \text{ whenever } 0 < \sqrt{(x-a)^2 + (y-b)^2} < \delta.$$

So, for $0 < \sqrt{(x-a)^2 + (y-b)^2} < \delta$, that is for $0 < |z - z_0| < \delta$, we get,

$$\begin{aligned} |f(z) - l| &= |(u(x,y) - \alpha) + i(v(x,y) - \beta)| \\ &= |u(x,y) - \alpha| + |v(x,y) - \beta| \\ &= < \epsilon/2 + \epsilon/2 \\ &= \epsilon \end{aligned} \tag{1}$$

This implies that $\lim_{z\to z_0} f(z) = l$. Hence proved. □

Theorem 4. Suppose that $\lim_{z\to z_0} f(z) = l$ and $\lim_{z\to z_0} g(z) = m$. Then,

1. $\lim_{z\to z_0}[f(z) \pm g(z)] = l \pm m$.
2. $\lim_{z\to z_0} cf(z) = cl$ for som complex number c.
3. $\lim_{z\to z_0} f(z)g(z) = lm$.
4. $\lim_{z\to z_0} \frac{f(z)}{g(z)} = \frac{l}{m}$, provided $g(z) \neq 0$ and $m \neq 0$.

Proof. Proof of this theorem is left as an exercise. □

Definition 19. 1. Let $f(z)$ be defined in a domain D except perhaps at the point z_0. The function f is said to tend to infinity as $z \to z_0$ if for any real number $k(> 0)$, however large, there exists a $\delta(> 0)$ such that $|f(z)| > k$ whenever $0 < |z - z_0| < \delta$. Symbolically, we write,

$$\lim_{z\to z_0} f(z) = \infty$$

2. Let f be defined for $|z| > k > 0$. Then the function is said to tend to a finite limit l as $z \to \infty$, if for any $\epsilon > 0$ there exists a number $k_0 > 0$ such that $|f(z) - l| < \epsilon$ whenever $|z| > k_0$.

3. we say that $\lim_{z \to z_0} f(z) = \infty$ if for each number $k > 0$, there exists a number $k_0 > 0$ such that $|f(z)| > k$ whenever $|z| > k_0$.

Exercise 1. Using definition of limit, show that $\lim_{z \to z_0}(az^2 + bz + c) = az_0^2 + bz_0 + c$, where a, b, c are complex constants.

Exercise 2. If $\lim_{z \to z_0} f(z) = l$, then show that the limit is unique.

4.2 Continuity

Let $f(z)$ be a complex function defined in some neighbourhood of z_0(including the point). The function is said to be continuous at z_0 if for any $\epsilon > 0$ there corresponds a $\delta > 0$ such that

$$|f(z) - f(z_0)| < \epsilon \; whenever \; |z - z_0| < \delta$$

Symbolically we write

$$\lim_{z \to z_0} f(z) = f(z_0)$$

This means that for continuity at a point, the limiting value and the functional value at the point have the same value.

A function $f(z)$ is continuous on a set S if it is continuous at every point of S.

If a function is not continuous at z_0, then we say that the function is discontinuous at z_0 or z_0 is a point of discontinuity.

Note 1. To examine the continuity of a function $f(z)$ at $z = \infty$, replace z by $\frac{1}{\zeta}$ and examine the continuity of $g(\zeta) = f\left(\frac{1}{\zeta}\right)$ at $\zeta = 0$.

Remark 1. Suppose that f and g are continuous functions at z_0. Then the following functions are continuous at z_0:

1. the sum $f(z) + g(z)$,

2. the difference $f(z) - g(z)$,

3. their product $f(z)g(z)$,

4. quotient $f(z)/g(z)$, provided $g(z) \neq 0$

Remark 2. If the function $f(z)$ is continuous, so are the functions $|f(z)|$, $f(\overline{z})$ and $\overline{f(z)}$.

Example 1. Test the continuity of the function

$$f(z) = \frac{z^3 + (1+i)z^2 + (2+i)z + 2}{z - i}$$

at $z = i$.

Clearly the function is undefined at the point $z = i$. Now,

$$\begin{aligned} \lim_{z \to i} f(z) &= \lim_{z \to i} \frac{z^3 + (1+i)z^2 + (2+i)z + 2}{z - i} \\ &= \lim_{z \to i} \frac{(z-i)(z^2 + 2iz + z + 2i)}{z - i} \\ &= -3 + 3i \end{aligned}$$

Hence, if we define

$$\begin{aligned} f(z) &= \frac{z^3 + (1+i)z^2 + (2+i)z + 2}{z - i},\ z \neq i \\ &= -3 + 3i,\ z = i \end{aligned}$$

Then the function $f(z)$ is continuous at $z = i$.

Exercise 3. Is the function

$$f(z) = \frac{z^2 + (2-i)z - 2i}{z - i}$$

continuous at $z = i$? If not, can it be made continuous by redefining at $z = i$?

Theorem 5. The composition of two continuous functions is continuous.

Proof. Let $f(z)$ be continuous at some point z_0. Then $f(z)$ is defined in some neighbourhood of z_0. Suppose that $g(w)$ is a function which is defined on the image of this neighbourhood. Given that $g(w)$ is continuous at $w_0 = f(z_0)$. Then, for given $\epsilon > 0$, there exists an $r > 0$ such that

$$\begin{aligned} |g(w) - g(w_0)| &< \epsilon \ whenever\ |w - w_0| < r \\ or,\ |g(f(z)) - g(f(z_0))| &< \epsilon \ whenever\ |f(z) - f(z_0)| < r \end{aligned} \tag{2}$$

Now, $f(z)$ is continuous at z. Hence, for this $r(> 0)$, there exists a $\delta > 0$ such that

$$|f(z) - f(z_0)| < r \ whenever\ |z - z_0| < \delta \tag{3}$$

Combining (2) and (3), we have

$$|gf(z) - gf(z_0)| < \epsilon \ whenever\ |z - z_0| < \delta$$

Hence the composition function gf is continuous at z_0. □

Theorem 6. If $f(z)$ is continuous in a region, then its real and imaginary parts are also continuous in that region.

Theorem 7. If $f(z)$ is continuous in a closed region, then it is bounded in that region.

Theorem 8. If a function $f(z)$ is continuous on a bounded and closed set $S \subset \mathbb{C}$, then the minimum and maximum of $|f(z)|$ exist on S.

Proof. Given that $f(z) = u(x, y) + iv(x, y)$ is continuous on S. This implies that the component functions $u(x, y)$ and $v(x, y)$ are continuous on S. Hence,

$$|f(z)| = \sqrt{u^2(x, y) + v^2(x, y)}$$

is a real valued continuous function on the closed and bounded set S. Hence by real calculus, $|f(z)|$ attains its maximum and minimum on S. This completes the proof. □

4.3 Uniform Continuity

A function $f(z)$ is said to be uniformly continuous on a set S if for given $\epsilon > 0$, there exists a $\delta > 0$ such that

$$|f(z_1) - f(z_2)| < \epsilon \; whenever \; |z_1 - z_2| < \delta \; \forall z_1, z_2 \in S$$

Here, the choice of δ is independent of z_1 and z_2 in S.

Theorem 9. Let $f(z)$ be a continuous function on a closed and bounded set S in the complex plane. Then it is uniformly continuous on S.

Proof. Let F be a compact set and $z \in F$. Then F can be covered by a finite number of neighbourhoods $N(z)$, that is, $F \subset N(z_1) \cup N(z_2) \cup N(z_3) \cup N(z_k)$. We now consider several points $\zeta_1, \zeta_2, \ldots, \zeta_n$ in S. Then using continuity of f at this point, we get

$$|f(z) - f(\zeta_k)| < \epsilon \; whenever \; |z - \zeta_k| < \delta(\epsilon, \zeta_k), \; k = 1, 2, \ldots, n$$

We consider the neighbourhoods of $\zeta_k, \; k = 1, 2, \ldots, n$ as

$$N(\zeta_k) = \{z : |z - \zeta_k| < \frac{1}{\delta}(\epsilon, \zeta_k)\}$$

Since S is compact, by Heine-Borel theorem, there exists a finite number of points $\zeta_1, \zeta_2, \ldots, \zeta_n$ such that, $S \subset N(\zeta_1) \cup N(\zeta_2) \cup \ldots \cup N(\zeta_n)$. Let $d = \min\{\delta(\epsilon, \zeta_k) : k = 1, 2, \ldots, n\}$ and $z_1, \; z_2$ be any two points in S such that $|z_1 - z_2| < \frac{d}{2}$ and $z_1 \in N(\zeta_i)$(say). Then,

$$|z_1 - z_2| < \frac{1}{2}\delta(\epsilon, \zeta_i) < \delta(\epsilon, \zeta_i)$$

Now,

$$\begin{aligned} |z_2 - \zeta_i| &\leq |z_1 - \zeta_i| + |z_1 - z_2| \\ &< \frac{1}{2}\delta + \frac{d}{2} \\ &< \frac{1}{2}\delta(\epsilon, \zeta_i) + \frac{1}{2}\delta(\epsilon, \zeta_i) \\ &= \delta(\epsilon, \zeta_i) \end{aligned}$$

Applying continuity of $f(z)$ in $|z-\zeta_i|<\delta$, we find

$$|f(z_1)-f(z_2)| \leq |f(z_1)-f(\zeta_i)|+|f(z_2)-f(\zeta_i)|<2\epsilon$$

Hence, f is continuous on S. □

Example 2. Show that $f(z)=z^2$ is uniformly continuous in the region $|z|<1$, but the function $g(z)=\frac{1}{z}$ is not uniformly continuous in the region.

We take any two points z_1 and z_2 in $|z|<1$ such that $|z_1-z_2|<\delta$. Then,

$$|f(z_1)-f(z_2)|=|z_1^2-z_2^2| \leq |z_1-z_2|(|z_1|+|z_2|)<2\delta$$

Thus if we choose $\delta=\frac{\epsilon}{2}$, we obtain $|f(z_1)-f(z_2)|<\epsilon$. This shows that f is uniformly continuous in the region $|z|<1$.

For the second case, if possible let us assume that $g(z)=\frac{1}{z}$ is uniformly continuous in the region $|z|<1$. Then for given $\epsilon>0$, we can find a $\delta>0$, say, between 0 and 1 such that,

$$|g(z_1)-g(z_2)|<\epsilon \; whenever \; |z_1-z_2|<\delta$$

for all z_1, z_2 in the region.
Fix $z_1=\delta$ and $z_2=\frac{\delta}{1+\epsilon}$. Clearly, z_1 and z_2 are in $|z|<1$ and

$$|z_1-z_2|=\left|\delta-\frac{\delta}{1+\epsilon}\right|=\frac{\epsilon\delta}{1+\epsilon}<\delta$$

Hence,

$$\begin{aligned} |g(z_1)-g(z_2)| &= \left|\frac{1}{z_1}-\frac{1}{z_2}\right| \\ &= \left|\frac{1}{\delta}-\frac{1+\epsilon}{\delta}\right| \\ &= \frac{\epsilon}{\delta} \\ &> \epsilon(0<\delta<1) \end{aligned}$$

Thus, we reach at a contradiction and therefore the function $g(z)=\frac{1}{z}$ can not be uniformly continuous in the given region.

Example 3. Prove that the function $f(z)=\frac{1}{z^2}$ is not uniformly continuous in the region $|z|\leq 1$ but it is uniformly continuous in the region $\frac{1}{2}\leq|z|\leq 1$.

For $|z|\leq 1$,

$$\begin{aligned} |f(z_1)-f(z_2)| &= |\frac{1}{z_1^2}-\frac{1}{z_2^2}| \\ &= \frac{|z_1-z_2||z_1+z_2|}{z_1^2.z_2^2} \\ &\leq \frac{2\delta}{z_1^2.z_2^2} \to \infty \; as \; z_1, z_2 \to 0 \; and \; |z_1-z_2|<\delta \end{aligned}$$

Hence f is not uniformly continuous in $|z| \leq 1$. But when $\frac{1}{2} \leq |z| \leq 1$, we have

$$\begin{aligned} |f(z_1) - f(z_2)| &= |\frac{1}{z_1^2} - \frac{1}{z_2^2}| \\ &= \frac{|z_1 - z_2||z_1 + z_2|}{z_1^2.z_2^2} \\ &\leq \frac{2\delta}{z_1^2.z_2^2} \\ &\leq \frac{2\delta}{16} < \epsilon \; when \; |z_1 - z_2| < \delta \end{aligned}$$

If we take $\delta = 7\epsilon$ then we will have,

$$|f(z_1) - f(z_2)| < \epsilon \; whenever \; |z_1 - z_2| < \delta$$

Hence f is uniformly continuous in the region $\frac{1}{2} \leq |z| \leq 1$.

4.4 Differentiation

Definition 20. Let $f(z)$ be a single-valued function defined in a domain D of the complex plane $\mathbb{C}$. If $z_0 \in D$ and if

$$\lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0} \tag{4}$$

exists, we denote this limit by $f'(z_0)$ and call it the derivative of $f(z)$ at the point z_0.

If $f'(z_0)$ exists, then f is said to be differentiable at the point z_0. Equivalently, we can write

$$\begin{aligned} f'(z_0) &= \lim_{h \to 0} \frac{f(z_0 + h) - f(z_0)}{h} \\ &= \lim_{\Delta z \to 0} \frac{f(z_0 + \Delta z) - f(z_0)}{\Delta z} \end{aligned}$$

If $f(z)$ is differentiable at each point of D, then we say that f is differentiable in D. We state once again that the limit 4 exists means that the value of the limit is same along any path in which z approaches z_0.

Theorem 10. If $f(z)$ is differentiable at z_0 then it is continuous there.

Proof. Since f is differentiable at z_0, we have

$$f'(z_0) = \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}$$

Now,

$$\begin{aligned}\lim_{z\to z_0}\{f(z)-f(z_0)\} &= \lim_{z\to z_0}\frac{f(z)-f(z_0)}{z-z_0}(z-z_0)\\ &= f'(z_0).0\\ &= 0\\ \lim_{z\to z_0} f(z) &= f(z_0)\end{aligned}$$

This proves that $f(z)$ is continuous at z_0. □

The following example shows that the converse of the above theorem is not necessarily true.

Example 4. Show that the function $f(z)=\overline{z}$ is continuous at z_0 is continuous at the point $z=z_0$, but the derivative does not exist.

$$\begin{aligned}f'(z_0) &= \lim_{\Delta z\to 0}\frac{f(z_0+\Delta z)-f(z_0)}{\Delta z}\\ &= \lim_{\Delta z\to 0}\frac{\overline{z_0}+\Delta\overline{z}-\overline{z_0}}{\Delta z}\\ &= \lim_{\Delta z\to 0}\frac{\Delta\overline{z}}{\Delta z}\\ &= 1,\ \textit{along real axis}\\ &= -1,\ \textit{along imaginary axis}\end{aligned}$$

Therefore the derivative does not exist since the value of the limit depends on the path along which Δz approaches 0.

The function $f(z)=\overline{z}$ is clearly continuous at $z=z_0$(Go by definition).

4.5 Geometrical Interpretation of Complex Derivative

Let $P(z_0)$ be a point in the z plane and $P'(w_0)$ be its image in the w plane under the transformation $w=f(z)$. If we give z_0 an increment Δz we obtain the point, say $Q(z_0+\Delta z)$. This point has image Q' in the w plane. Thus we see that $P'Q'$ represents the complex number

$$\Delta w = f(z_0+\Delta z)-f(z_0)$$

It follows that the derivative at z_0, if it exists, is given by

$$\begin{aligned}f'(z_0) &= \lim_{\Delta z\to 0}\frac{f(z_0+\Delta z)-f(z_0)}{\Delta z}\\ &= \lim_{\Delta z\to 0}\frac{\Delta w}{\Delta z}\\ &= \lim_{Q\to P}\frac{P'Q'}{PQ}\end{aligned}$$

that is, the limit of the ratio $Q'P'$ to QP at the point Q approaches the point P.

4.6 Analytic Functions

A function $f(z)$ defined in a domain D is said to be an analytic function in D, if $f(z)$ has a derivative at each point of D. The terms *regular* and *holomorphic* are also used instead of analytic.

The function $f(z)$ is said to be analytic at a point z_0 of D if it is analytic in a neighbourhood of z_0, that is, if there exists a neighbourhood of z_0 at all points of which $f'(z)$ exists.

If $f(z)$ is not analytic at a point z_0, then z_0 is called a singular point or a singularity of $f(z)$.

Example 5. Show that the function $f(z) = z\overline{z}$ is differentiable only at origin.

We have

$$f(z) = z\overline{z} = |z|^2$$

So, we have

$$\begin{aligned} f'(z_0) &= \lim_{\Delta z \to 0} \frac{f(z_0 + \Delta z) - f(z_0)}{\Delta z} \\ &= \lim_{\Delta z \to 0} \frac{(z_0 + \Delta z)\overline{(z_0 + \Delta z)} - z_0\overline{z_0}}{\Delta z} \\ &= \lim_{\Delta z \to 0} \left(z_0 \frac{\Delta \overline{z}}{\Delta z} + \overline{z_0} + \overline{\Delta z} \right) \end{aligned}$$

Since

$$\begin{aligned} \lim_{\Delta z \to 0} \frac{\Delta \overline{z}}{\Delta z} &= 1, \textit{ along real axis} \\ &= -1, \textit{ along imaginary axis.} \end{aligned} \tag{5}$$

Hence,

$$\lim_{\Delta z \to 0} \frac{\Delta \overline{z}}{\Delta z}$$

does not exist. Thus the given function is differentiable only at the origin.

Example 6. If

$$\begin{aligned} f(z) &= \frac{x^2 y(y - ix)}{x^4 + y^2},\ z \neq 0 \\ &= 0,\ z = 0 \end{aligned}$$

Prove that

$$\frac{f(z) - f(0)}{z} \to 0$$

along radius vector but $f'(0)$ does not exist.

Suppose $z \to 0$ along any radius vector $y = mx$. Then,

$$\begin{aligned}\lim_{z\to 0}\frac{f(z)-f(0)}{z-0} &= \lim_{(x,y)\to(0,0)}\frac{x^2y(y-ix)}{(x^4+y^2)(x+iy)}\\ &= \lim_{(x,y)\to(0,0)}\frac{-iyx^2}{x^4+y^2}\\ &= \lim_{x\to 0}\frac{-imx^3}{x^4+m^2x^2}\\ &= 0\end{aligned}$$

But for $y = x^2$,

$$\begin{aligned}\lim_{z\to 0}\frac{f(z)-f(0)}{z-0} &= \lim_{(x,y)\to(0,0)}\frac{-iyx^2}{x^4+y^2}\\ &= \lim_{x\to 0}\frac{-ix^4}{x^4+x^4}\\ &= -\frac{i}{2}\end{aligned}$$

This shows that $f'(0)$ does not exist since the two limits are different.

Try the next example for yourself.

Example 7. If

$$\begin{aligned}f(z) &= \frac{xy^2(x+iy)}{x^2+y^4},\ z\neq 0\\ &= 0,\ z=0\end{aligned}$$

Prove that

$$\frac{f(z)-f(0)}{z} \to 0\ as\ z\to 0$$

along any straight line but $f'(0)$ does not exist.

4.7 Cauchy-Riemann Equations

Theorem 11. A necessary condition for $w = f(z) = u(x,y) + iv(x,y)$ to be differentiable at a point $z_0 = x_0 + iy_0$ is that

$$\begin{aligned}u_x(x_0,y_0) &= v_y(x_0,y_0)\\ u_y(x_0,y_0) &= -v_x(x_0,y_0)\end{aligned}$$

Proof. Suppose that $f'(z_0)$ exists. Then

$$\begin{aligned} f'(z_0) &= \lim_{z\to z_0} \frac{f(z)-f(z_0)}{z-z_0} \\ &= \lim_{(x,y)\to(x_0,y_0)} \frac{(u(x,y)+iv(x,y))-(u(x_0,y_0)+iv(x_0,y_0))}{(x+iy)-(x_0+iy_0)} \\ &= \lim_{(x,y)\to(x_0,y_0)} \frac{(u(x,y)-u(x_0,y_0))+i(v(x,y)-v(x_0,y_0))}{(x-x_0)+i(y-y_0)} \end{aligned} \tag{6}$$

Since $f'(z_0)$ exists, (6) must exist for all modes of approach of (x, y) to (x_0, y_0) and all the limiting values must be same.

Let $z \to z_0$ along a line parallel to the real axis. Then $y = y_0$ and $x \to x_0$. So from (6), we obtain

$$\begin{aligned} f'(z_0) &= \lim_{x\to x_0} \frac{u(x,y)-u(x_0,y_0)}{x-x_0} + i \lim_{x\to x_0} \frac{v(x,y)-v(x_0,y_0)}{x-x_0} \\ &= u_x(x_0,y_0) + iv_x(x_0,y_0) \end{aligned} \tag{7}$$

Now letting $z \to z_0$ along a line parallel to the imaginary axis, we have $x = x_0$ and $y \to y_0$. So from (6), we have,

$$\begin{aligned} f'(z_0) &= \lim_{y\to y_0} \frac{u(x,y)-u(x_0,y_0)}{i(y-y_0)} + \lim_{y\to y_0} \frac{v(x,y)-v(x_0,y_0)}{y-y_0} \\ &= -iu_y(x_0,y_0) + v_y(x_0,y_0) \end{aligned} \tag{8}$$

Comparing (7) and (8) and equating the real and imaginary parts we get

$$u_x(x_0,y_0) = v_y(x_0,y_0) \text{ and } u_y(x_0,y_0) = -v_x(x_0,y_0)$$

This proves the theorem. □

Note 2. The differential equations

$$u_x = v_y \text{ and } u_y = -v_x$$

are known as Cauchy-Riemann (CR) equations.

Example 8. Find the nature of CR equations for the function $f(z) = |z|^2$.

We have,

$$f(z) = |z|^2 = x^2 + y^2$$

Hence,

$$u(x,y) = x^2 + y^2,\ v(x,y) = 0$$

Therefore,

$$u_x = 2x,\ u_y = 2y,\ v_x = 0,\ v_y = 0$$

Thus the CR equations for the given function are not satisfied unless $x = y = 0$, that is, at the origin and hence $f'(z)$ does not exist at any point $z \neq 0$.

Example 9. Let $f(z) = |z|^4$. Show that $f(z)$ is differentiable but not analytic at the origin.

Try for yourself.

Example 10. Let

$$\begin{aligned} f(z) &= \frac{x^3 - y^3}{x^2 + y^2} + i\frac{x^3 + y^3}{x^2 + y^2},\ z \neq 0 \\ &= 0,\ z = 0 \end{aligned}$$

Show that though CR equations are satisfied at $(0, 0)$, $f'(0)$ does not exist.

Here,

$$\begin{aligned} u(x, y) &= \frac{x^3 - y^3}{x^2 + y^2},\ z \neq 0 \\ &= 0,\ z = 0 \end{aligned}$$

and

$$\begin{aligned} u(x, y) &= \frac{x^3 + y^3}{x^2 + y^2},\ z \neq 0 \\ &= 0,\ z = 0 \end{aligned}$$

Now,

$$\begin{aligned} u_x(0, 0) &= \lim_{x \to 0} \frac{u(x, 0) - u(0, 0)}{x - 0} = 1 \\ u_y(0, 0) &= \lim_{y \to 0} \frac{u(0, y) - u(0, 0)}{y - 0} = -1 \\ v_x(0, 0) &= \lim_{x \to 0} \frac{v(x, 0) - v(0, 0)}{x - 0} = 1 \\ v_y(0, 0) &= \lim_{y \to 0} \frac{v(0, y) - v(0, 0)}{y - 0} = 1 \end{aligned}$$

Since $u_x = v_y$ and $u_y = -v_x$ at $(0, 0)$, CR equations are satisfied at the origin. Now,

$$\begin{aligned} f'(0) &= \lim_{z \to} \frac{f(z) - f(0)}{z - 0} \\ &= \lim_{(x,y) \to (0,0)} \frac{(x^3 - y^3) + i(x^3 + y^3)}{(x^2 + y^2)(x + iy)} \end{aligned}$$

We put $y = mx$. Then

$$\begin{aligned} f'(0) &= \lim_{x \to 0} \frac{(1 - m^3) + i(1 + m^3)}{(1 + m^2)(1 + im)} \\ &= \frac{(1 - m^3) + i(1 + m^3)}{(1 + m^2)(1 + im)} \end{aligned}$$

Since the value of the limit depends on m, $f'(0)$ does not exist.

Example 11. Let

$$\begin{aligned} f(z) &= e^{-z^{-4}},\ z \neq 0 \\ &= 0,\ z = 0 \end{aligned}$$

Show that though CR equations are satisfied at $(0,0)$, $f'(0)$ does not exist.

Try for yourself.

The above example shows that the validity of CR equations is not sufficient to ensure the analyticity.

Theorem 12. A single-valued continuous function $w = f(z) = u(x,y)+iv(x,y)$ is differentiable in a domain D if the partial derivatives $u_x,\ u_y,\ v_x,\ v_y$ exist and are continuous and they satisfy CR equations.

Proof. We are to show that

$$f'(z) = \lim_{\Delta z \to 0} \frac{\Delta w}{\Delta z}$$

exist at each point of D. Let $z = x + iy$ be any arbitrary point of D. Since $u_x,\ u_y,\ v_x,\ v_y$ exist and are continuous at (x,y), $u(x,y)$ and $v(x,y)$ are differentiable at the point (x,y). Therefore,

$$\begin{aligned} \Delta u &= u(x+\Delta x, y+\Delta y) - u(x,y) \\ &= u_x \Delta x + u_y \Delta y + \epsilon_1 \Delta x + \epsilon_2 \Delta y \\ &= v_y \Delta x - v_x \Delta y + \epsilon_1 \Delta x + \epsilon_2 \Delta y \end{aligned}$$

where $\epsilon_1, \epsilon_2 \to 0$ as $(\Delta x, \Delta y) \to (0,0)$. Also,

$$\begin{aligned} \Delta v &= v(x+\Delta x, y+\Delta y) - v(x,y) \\ &= v_x \Delta x + v_y \Delta y + \eta_1 \Delta x + \eta_2 \Delta y \\ &= v_y \Delta x + u_x \Delta y + \eta_1 \Delta x + \eta_2 \Delta y \end{aligned}$$

where $\eta_1, \eta_2 \to 0$ as $(\Delta x, \Delta y) \to (0,0)$. Now,

$$\begin{aligned} \Delta w &= \Delta u + i\Delta v \\ &= u_x(\Delta x + i\Delta y) + v_x(i\Delta x - \Delta y) + (\epsilon_1 + i\eta_1)\Delta x + (\epsilon_2 + i\eta_2)\Delta y \\ &= (u_x + iv_x)(\Delta x + i\Delta y) + (\epsilon_1 + i\eta_1)\Delta x + (\epsilon_2 + i\eta_2)\Delta y \\ or,\ \frac{\Delta w}{\Delta z} &= u_x + iv_x + (\epsilon_1 + i\eta_1)\frac{\Delta x}{\Delta z} + (\epsilon_2 + i\eta_2)\Delta y \Delta z \end{aligned} \tag{9}$$

Now,

$$\left|(\epsilon_1 + i\eta_1)\frac{\Delta x}{\Delta z}\right| = |\epsilon_1 + i\eta_1| \left|\frac{\Delta x}{\Delta z}\right| \leq (|\epsilon_1| + |\eta_1|) \to 0 \ as\ (\Delta x, \Delta y) \to (0,0)$$

Similarly,

$$|(\epsilon_2 + i\eta_2)\Delta y \Delta z| \to 0 \; as \; (\Delta x, \Delta y) \to (0,0)$$

Hence taking limit as $\Delta z \to 0$, we get from (9)

$$\lim_{\Delta z \to 0} \frac{\Delta w}{\Delta z} = u_x + iv_x$$

that is, $f'(z)$ exists and is equal to $u_x + iv_x$. Since z is any point of D, we thus conclude that f is differentiable in D. This proves the theorem. □

Example 12. Let $f(z) = u + iv$ be analytic in a domain D and $|f(z)|$ is equal to a constant in D. Then show that $f(z)$ is constant in D.

We have,

$$|f(z)| = constant = c(say) or, \; u^2 + v^2 = c^2$$

Differentiating with respect to x and y, we obtain

$$uu_x + vv_x = 0 \tag{10}$$

$$uu_y + vv_y = 0 \tag{11}$$

Since $u_x = v_y$ and $u_y = -v_x$ (11) gives

$$-uv_x + vu_x = 0 \tag{12}$$

From (10) and (12) we get

$$(u^2 + v^2)u_x = 0$$

If $u^2 + v^2 = 0$, then $u = v = 0$. Then $f(z) = 0$, a constant function. Hence $u_x = 0$. Similarly from (10) and (12), we obtain $v_x = 0$. Hence $u_x = v_x = u_y = v_y = 0$. Hence,

$$\begin{aligned} du &= u_x dx + u_y dy \\ &= 0 \\ or, \; u &\equiv constant \end{aligned}$$

Similarly, $v \equiv constant$ and so f is constant.

Note 3. In polar form the CR equations can be written as

$$\frac{\partial u}{\partial r} = \frac{1}{r}\frac{\partial v}{\partial \theta}, \; \frac{\partial v}{\partial r} = -\frac{1}{r}\frac{\partial u}{\partial \theta}$$

Example 13. Show that the function $f(z) = xy + iy$ is everywhere continuous but not analytic.

Example 14. Show that the function $f(z) = \overline{z}$ is non-analytic everywhere.

Example 15. Let $f = u + iv$ be analytic in a domain D. Show that f is constant in D if any one of the following conditions hold:

1. $f'(z) \equiv 0$ in D.
2. $Re\{f(z)\} = constant$ in D.
3. $Im\{f(z)\} = constant$ in D.
4. $arg\{f(z)\} = constant$ in D.

Unit 13

Course Structure

1. Complex Integration
2. Cauchy's Fundamental Theorem and its consequences
3. Cauchy's Integral formula
4. Derivative of an analytic function.

Introduction

In this unit, we shall learn about curves, contours, Jordan curves, Complex Integration and a few examples to grasp it completely. Then we have Cauchy's Fundamental Theorem and its applications; Cauchy's Integral formula and derivatives of an analytic function.

Definition :- A curve Γ in the Complex plane is a continuous Complex valued function

Γ:- ƶ (t) =x (t)+ iy(t) defined on a real interval a≤ t ≤b. The point Ƶ(a) is called the initial point of Γ and the point Ƶ (b) is called the terminal point of Γ. If Ƶ(b) = Ƶ(a) i.e., the terminal and initial point coincide, the curve Γ is called a closed curve.

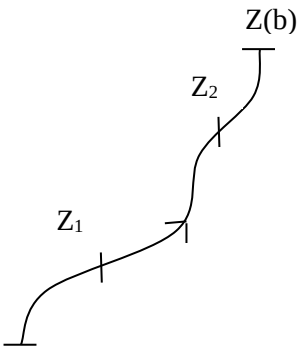


Fig -1

In calling Ƶ (a) the initial point and Ƶ (b) the terminal point of the curve Γ we describe an orientation of Γ. This means that a

þoint $Ƶ_1 = Ƶ(t_1) \in Γ$ is regarded as preceding a point $Ƶ_2 = Ƶ(t_2) \in Γ$ if $Ƶ_1 \neq Ƶ_2$ and $t_1 < t_2$.

[Fig - `1']. From this it follows that the Curve Γ may be thought of having two orientations according as t varies from a to b or from b to a. The Curve differing from Γ only by the direction is denoted by (-Γ If for more than one value of t, we get the same point Ƶ = Ƶ(t),then Ƶ is called a multiple point of the curve (Fig - `2'). If a curve has no multiple point, then it is called a `simple curve'.

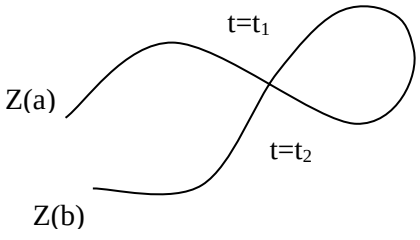


Fig -2

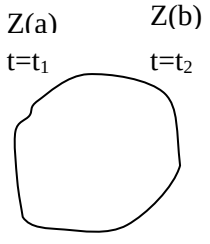
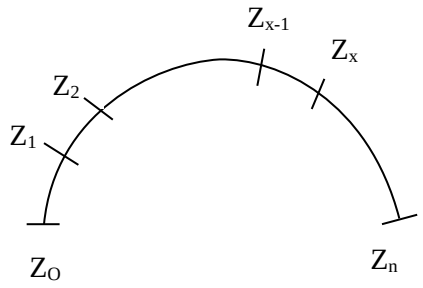


Fig -3

Rectifiable curve:- Let Γ : Ƶ= Ƶ(t), a ≤ t ≤ b, be a continuous curve. By a partition of [a,b], we mean a set of points P={a=t_o, t_1, t_2 …, t_n -1, t_n=b} satisfying a=to < t_1<t_2< … < t_n =b. We denote by P [a,b] the collection of all possible partitions of [a,b]. We put Ƶ$_k$ = Ƶ (t_k) for k=0,1,2, …, n. Then corresponding to the partition P we get a set of points Ƶ$_o$, Ƶ$_1$, Ƶ$_2$,…, Ƶ$_n$ or Γ dividing it into smaller arcs.



We now construct the sum,

Lp = |Ƶ$_1$ –Ƶ$_0$| + |Ƶ$_2$ – Ƶ$_1$| + … + |Ƶ$_k$ – Ƶ$_{k-1}$| + …+ |Ƶ$_n$ –Ƶ$_{n-1}$|

Clearly, Lp denotes the length of the polygon inscribed, which is obtained by drawing straight lines from Ƶ$_0$ to Ƶ$_1$ from Ƶ$_1$ to Ƶ$_2$ and so on.

We now consider the aggregate {Lp: PϵP [a, b]}. If the aggregate {Lp :PϵP [a,b]} is bounded above, then the curve Γ is said to be a rectifiable curve.

Definition : - A simple curve Γ:Ƶ =Ƶ(t)= ϰ(t) +iy(t) a≤+≤b is called regular curve if the derivatives of ϰ(t) and y (t) exists and are continuous and they do not vanish simultaneously over the whole interval [a,b].

For example, the unit circle,

Ƶ = e^{it} = cost + i sint , 0≤ t ≤ 2π is a regular curve.

Theorem: - Every regular curve is rectifiable.

Complex Integration

Let Γ:Ƶ =Ƶ(t), a ≤ t ≤ b be a rectifiable curve. Let. F be a complex function defined on Γ. Suppose that P= { a=t_0, t_1, t_2, …, t_{k-1}, t_k, …, t_n=b} be a partition of [a,b]. We put Ƶ$_k$ = Ƶ (t_k)k=0, 1,2, …, n. Then we get a set of point. Ƶ$_0$, Ƶ$_1$, Ƶ$_2$, …, Ƶ$_{k-1}$, Ƶ$_k$, …, Ƶ$_n$ which divide the curve Γ, into smaller arcs $\overset{\frown}{z_{k-1}z_k}$, where k= 1,2, …,n.

We choose $\varepsilon_{k-1} \in \overset{\frown}{z_{k-1}z_k}$ for k= 1,2,3,…,n and form the sum

$$Sp = \sum_{k-1}^{n} \mathrm{f}(\varepsilon_{k-1})(\mathrm{Z}_k - \mathrm{Z}_{k-1})$$

If $\lim_{||P||\to 0} S\rho$ exists and is equal I, which is

Independent of the partition P and the points ε_{k-1}, we say that f is integrable on Γ, and we write,

$$I = \int_{\Gamma} f(Z)dZ,$$

where ||P|| denotes the norm of the partition P

equivalently, we can write

$I = \int_{\Gamma} f(Z)dZ$= lim Sp

Max|Zk-Zk-1|→0

1≤k≤ n

As the case of real definite integral, it can be prove that if f is continuous on Γ, then f is integrable on Γ.

Some Elementary properties:-

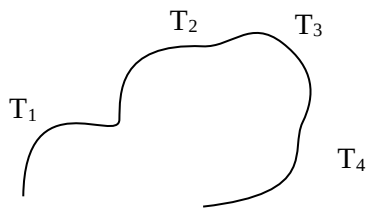
If f and g are integrable on a rectifiable curve Γ, then

(i) $\int_{\Gamma} (f(Zl \pm g(Z))dZ = \int_{\Gamma} f(Z)dZ \pm \int_{\Gamma} g(Z)dZ$

(ii) $\int_{\Gamma} k.f(Z)dZ = k\int_{\Gamma} f(Z)dZ$, where k is a constant.

(iii) $\int_{-\Gamma} f(Z)dZ = -\int_{\Gamma} f(Z)dZ$,

(iv) $\int_{\Gamma} f(Z)dZ = \sum_{i=0}^{n} \int_{\Gamma j} f(Z)dZ$, where



$\Gamma=\Gamma_1+\Gamma_2+\text{---}+\Gamma_n$

An inequality for complex Integral (ML-formula)

Theorem: If f is integrable on a rectifiable Curve Γ of length L and if there exists a positive number M such that

$|f(Z)| \leq M \; \forall \; Z \in \Gamma$, then

$\left|\int_\Gamma f(Z)dZ\right| \leq ML.$

Proof :- We divide the curve Γ into smaller arcs by the point $Z_0, Z_1, \text{---}, Z_{n-1}, Z_n, \text{--}$ where Z_0 is initial point and Z_n is the terminal point of Γ. We choose

$\varphi_k \in Z_{k-1}\, Z_k$ for k=1, 2, ---, n and form the sum

$$S = \sum_{k=1}^{n} f(\varphi_k)(Z_k - Z_{k-1}).$$

By the given condition, We have,

$|S| = \left|\sum_{k=1}^{n} f(\varphi_k)(Z_k - Z_{k-1})\right|.$

$\leq \sum_{k=1}^{n} |f(\varphi_k)||Z_k - Z_{k-1}|.$

$\leq M \sum_{k=1}^{n} |Z_k - Z_{k-1}| \leq ML.$

Since, f is integrable on Γ, we get,

$$\left|\int_\Gamma f(Z)dZ\right| = \left|\lim_{\substack{\max|Z_k - Z_{k-1}|\to 0 \\ 1\leq k\leq n}} S\right| = \lim_{\substack{\max|Z_k - Z_{k-1}|\to 0 \\ 1\leq k\leq n}} |S| \leq ML.$$

This proves the theorem.

Cauchy's Fundamental Theorem

Theorem: - If f is analytic with in a simple closed rectifiable curve Γ and continuous on Γ, then $\int_\Gamma f(Z)dZ = 0$

For a closed curve Γ, we use the notation $\oint$ to denote the integration along Γ in the counter-clockwise sense and $\oint$ denotes the same integration in the clock wise sense clearly.

$$\oint_{\Gamma} = -\oint_{\Gamma}$$

Definition :- A region D is called simply connected if every closed curve drawn in the region encloses only the points of the region.

A region which is not simply connected is called multiply connected region.

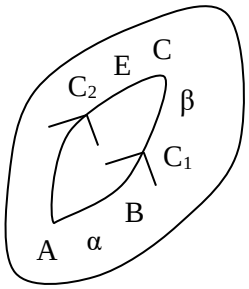
Consequences of Cauchy's fundamental theorem :-

(i) Let f be analytic in a simply connected domain D and let $\propto$ and β be any two points in D. Then

$\int_{\propto}^{\beta} f(Z)dZ$ is independent of the curve in D joining $\propto$ and β.

$0 = \int_{ABCDEFA} f(Z)dZ = \int_{AB} f(Z)dZ + \int_{BCD} f(Z)dZ + \int_{DE} f(Z)dZ + \int_{EFA} f(Z)dZ$ ------ (i)

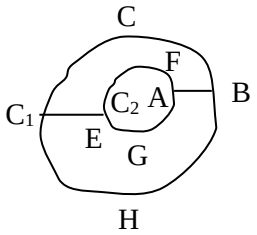
$0 = \int_{DHBAGED} f(Z)dZ = \int_{DHB} f(Z)dZ + \int_{BA} f(Z)dZ + \int_{AGE} f(Z)dZ + \int_{ED} f(Z)dZ$ ------ (ii)



Adding (i) and (ii) we get,

$$0 = \oint_{C_1} f(Z)dZ - \oint_{C_2} f(Z)dZ$$

$$= \text{»} \oint_{C_1} f(Z)dZ - \oint_{C_2} f(Z)dZ$$



II. Let, C1 and C2 be two simple closed rectifiable curves, C2 lying wholly with in C1. If f is analytic in the closed annulus determined by C1 and C2, then

$$\oint_{C_1} f(\mathrm{Z})d\mathrm{Z} = \oint_{C_2} f(\mathrm{Z})d\mathrm{Z}$$

III. $0 = \int_{ABCDEFGHIA} f(Z)dZ = \int_{AB} f(Z)dZ + \int_{BCD} f(Z)dZ$

$$+ \int_{DE} f(Z)dZ + \int_{EFG} f(Z)dZ + \int_{GH}^{\Gamma} f(Z)dZ +$$

$\int_{HIA}^{L} f(Z)dZ$ - - (i)

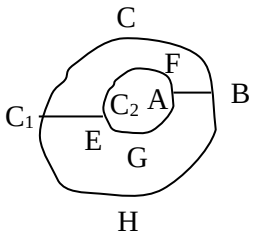
$$0 = \int_{DLBAJHGKED} f(Z)dZ = \int_{DLB} f(Z)dZ + \int_{BA} f(Z)dZ + \int_{AJH} f(Z)dZ$$

$$+ \int_{HG} f(Z)dZ + \int_{GKE} f(Z)dZ + \int_{ED} f(Z)dZ \quad \text{- - - (ii)}$$

Now we add (i) and (ii)

$0 = \oint_{\Gamma} f(\mathrm{Z})\, d\mathrm{Z} + \oint_{\Gamma_1} f(\mathrm{Z})\, d\mathrm{Z} + \oint_{\Gamma_2} f(\mathrm{Z})\, d\mathrm{Z}$

$= \oint_{\Gamma} f(\mathrm{Z})\, d\mathrm{Z} - \oint_{\Gamma_1} f(\mathrm{Z})\, d\mathrm{Z} - \oint_{\Gamma_2} f(\mathrm{Z})\, d\mathrm{Z}$



Therefore,

$\oint_{\Gamma} f(\mathrm{Z})\, d\mathrm{Z} = \oint_{\Gamma_1} f(\mathrm{Z})\, d\mathrm{Z} - \oint_{\Gamma_2} f(\mathrm{Z})\, d\mathrm{Z}$

(III) If Γ_1, Γ_2, ---, Γ_n are simple closed rectifiable curves, no two of which have common point, and if Γ is a simple closed rectifiable curve, which contains Γ_1, Γ_2, - - - , Γ_n in its interior, then .

$\oint_{\Gamma} f(\mathrm{Z})\, d\mathrm{Z} = \sum_{i=0}^{n} \oint_{\Gamma_i} f(\mathrm{Z})\, d\mathrm{Z}$

Provided f is analytic in the closed region bounded by this curve.

Example :- Evaluate

$\oint_{|Z|=1} \frac{dZ}{Z+2}$ and deduce that $\int_0^{\Pi} \frac{1+2COS\theta}{5+4\,cos\theta} d\theta = 0$.

Solution:- Since $\frac{1}{Z+2}$ is analytic with in and on lZl =1, by Cauchy's fundamental theorem, weget

$\oint_{lZl=1} \frac{dZ}{Z+2} = 0$

Putting, Z= $e^{i\theta}$, 0≤θ≤2Л we get, dZ = $i.e^{i\theta}$dθ

$$0 = \oint_{lZl=1} \frac{dZ}{Z+2} = \int_0^{2\Pi} \frac{e^{i\theta}d\theta}{(e^{i\theta}+2)} = i\int_0^{2\Pi} \frac{e^{i\theta}(e^{-i\theta}+2)}{(e^{i\theta}+2)\,(e^{-i\theta}+2)}d\theta$$

$$= i\int_0^{2\Pi} \frac{1+2\,e^{i\theta}}{5+2(e^{-i\theta}+2)}d\theta$$

$$= i\int_0^{2\Pi} \frac{1+2\,Cos\theta+2isin\theta}{5+2\,\mathrm{Cos}\theta}d\theta$$

Equating the imaginary parts of both sides, we get,

$$i\int_0^{2\Pi} \frac{1+2\,Cos\theta}{5+4\,\mathrm{Cos}\theta}d\theta=0$$

$$i.e,2\int_0^{\Pi} \frac{1+2\,Cos\theta}{5+4\,\mathrm{Cos}\theta}d\theta=0$$

$$= ie\int_0^{\Pi} \frac{1+2\,Cos\theta}{5+4\,\mathrm{Cos}\theta}d\theta=0$$

$$[\because \int_0^{2a} \quad f(x)dx = 2\int_0^{a} \quad f(x)dx$$

$$if \int \quad f(2a-x) = f(x)]$$

2) Evaluate , $\oint \frac{dZ}{Z-\propto}$where C denotes a simple closed rectifiable curve and ∝ is an interior point.

Solution:- Let Γ be a circle lying with in C with ∞ as centre and radius ϒ. since, $\frac{1}{Z-\propto}$ is analytic in the closed region bounded by C and Γ, putting

$Z-\propto = \gamma.e^{i\theta}$, 0 ≤ θ∝ 2Л ,

We get $\oint_c \frac{dZ}{Z-\infty} = \oint_r \frac{dZ}{Z-\infty} = \int_0^{2\Pi} \frac{i.\gamma\, e^{i\theta}d\theta}{\gamma e \quad i\theta}$

$$= i\int_0^{2\Pi} \quad d\theta = 2\Pi i.$$

Cauchy's Integral Formula :-

Theorem: Let f be analytic within and on a simple closed rectifiable curve C and let, $\propto$ be an interior point of C. then

$f(\propto) = \frac{1}{2\pi i}\oint_c \frac{f(Z)}{Z-\propto} dZ$ ---- (1)

Theorem: - let, f be analytic within and on a simple closed rectifiable curve C. If ∞ is a point interior to C, then

$f(\propto) = \frac{1}{2\pi i}\oint_c \frac{f(Z)}{(Z-\propto)^2} dZ$ ---- (1)

Proof:- Let d be the lower bound of the distances of the point $\propto$ from the points on C.

If h denotes a complex number such that $|h| < d$, $|(\propto+h)- \propto| = |h| < d$

Then the point $(\propto+h)$ also lies with in C.

Therefore, we use Cauchy' integral formula.

$$f(\propto) = \frac{1}{2\pi i}\oint_c \frac{d(Z)}{Z-\propto} dZ$$

$$f(\propto +h) = \frac{1}{2\pi i}\oint_c \frac{f(Z)}{(Z-\propto -h)} dZ$$

$$f(\propto +h) - f(\propto) = \frac{1}{2\pi i}\oint_c (Z)\{\frac{f(Z)}{(Z-\propto -h)} - \frac{1}{Z-\propto}\} dZ$$

$$= \frac{1}{2\pi i}\oint_c (Z)\frac{h}{(Z-\propto -h)(Z-\propto)} . dZ$$

$$\text{i. e,} \quad \frac{f(\propto +h) - f(\propto)}{h} = \frac{1}{2\pi i}\oint_c \frac{f(Z)dZ}{(Z-\propto -h)(Z-\propto)} ----(2)$$

from (2) we get

$$\left|\frac{f(\propto+h)-f(\propto)}{h} - \frac{1}{2\pi i}\oint_c \frac{f(Z)}{(Z-\propto)2} dZ\right| = \left|\frac{1}{2\pi i}\oint_c f(Z)\{\frac{1}{(Z-\infty)(Z-\infty-h)} - \frac{1}{(Z-\infty)2}\} dZ\right|$$

$$= \frac{1}{2\pi i}\left|\oint_c f(Z)\{\frac{(Z-\propto) - (Z-\propto -h)}{(Z-\propto)2(Z-\propto -h)} dZ\right|$$

$$= \frac{|h|}{2\pi} \left| \oint_c \frac{f(Z)}{(Z-\propto)^2(Z-\propto -h)} dZ \right| - - - (3)$$

Since f is analytic on C, it is continuous there, so, there exists a positive number M such that $|f(Z)| \leq M$ $Z \epsilon C$

Also for $Z \epsilon C$, we have,

$|Z-\propto| \geq d$ and $|Z-\propto-h| \geq |Z-\propto-|h| \geq d-|h|$.

Therefore, for all $Z \epsilon C$, we have,

$\left| \frac{f(Z)}{(Z-d)^2(Z-d-h)} \right| \leq \frac{M}{d^2(d-lhl)}$.

So, by M-L formula, we get from (3)

$\left| \frac{f(\propto+h)-f(\propto)}{h} - \frac{1}{2\pi i} \oint_c \frac{f(Z)}{(Z-\propto)^2} dZ \right| \leq \frac{|h|}{2\pi} \cdot \frac{ML}{d^2(d-lhl)} \to 0$ as $h \to 0$

i.e.

$\lim_{h-0} \left(\frac{f(\propto+h)-f(\propto)}{h} \right) = \frac{1}{2\pi i} \oint_c \frac{f(Z)}{(Z-\propto)^2} dZ$

i.e. $f(\propto) = \frac{1}{2\pi i} \oint_c \frac{f(Z)}{(Z-\propto)^2} dZ$

This proof's the theorem

Theorem : Let f be analytic within and on a simple closed rectifiable curve C. Then for any point $\propto$ interior to C.

$$f^{(n)}(\propto) = \frac{n!}{2\pi i} \oint_c \frac{f(Z)}{(Z-\propto)^{n+1}} dZ$$

for n=0, 1, 2, ---

Proof: We prove the theorem by Mathematical Induction. First we note that, the formula (1) is valid for n=0,1.
We suppose that, (1) is true for n=m and prove that it true for n=m+1.
We choose a positive number R sufficiently large, such that the curve C is Contained in $|Z| < R$, d be the lower bound of the distances of $\propto$ from the points on C.

If a complex number h is such that $|h| < d$, then, the point $\propto+h$ lies with in C.

Therefore, from (1), we get for n=m,

$$f^{(m)}(\propto) = \frac{m!}{2\pi i}\oint_c \frac{f(Z)}{(Z-\propto)^{m+1}}dZ$$

$$f^{(m)}(\propto +h) = \frac{m!}{2\pi i}\oint_c \frac{f(Z)}{(Z-\propto)^{m+1}}dZ$$

Then

$$f^{(m)}(\propto +h) - f^{(m)}(\propto) = \frac{m!}{2\pi i}\oint_c f(Z)\{\frac{1}{(Z-\propto -h)^{m+1}} - \frac{1}{(Z-\propto)^{m+1}}\}\,dZ$$

$$= \frac{m!}{2\pi i}\oint_c f(Z)\frac{(Z-\propto)^{m+1}-(Z-\propto-h)^{m+1}}{(Z-\propto)^{m+1}(Z-\propto-h)^{m+1}}dZ \text{ ---- (2)}$$

Let us denote Z -∝by t,

Then, $(Z-\propto)^{m+1}$- $(Z-\propto -h)^{m+1}$

$= t^{m+1} - (t-h)^{m+1}$

$=n\{t^m + t^{m-1}(t-h) + \cdots + (t-h)^m\}$

So, from (2)

$$\frac{f^{(m)}(\propto+h)-f^{(m)}(\propto)}{h} = \frac{m!}{2\pi i}\oint_c f(Z)\frac{t^m+t^{m-1}(t-h)+\cdots+(t-h)^m}{t^{m+1}(t-n)^{m+1}}.dZ$$

Now

$$f^{(m)}(\propto +h) - f^{(m)}(\propto) = \frac{(m+l)!}{2\Pi i}\oint_c \frac{l}{(Z-\propto)^{m+2}}dZ$$

$$= \frac{(m+l)!}{2\pi i}\oint_c f(Z)\{\frac{t^m+t^{m-1}(t-h)+\cdots+(t-h)^m}{t^{m+1}(t-n)^{m+1}} - \frac{m+l}{t^{m+2}}\,dZ$$

$$= \frac{m!}{2\pi i}\oint_c f(Z)\{\frac{t^m+t^{m-1}(t-h)+\cdots+(t-h)^m-(m+l)(t-h)^{m+1}}{t^{m+1}(t-n)^{m+1}}\,dZ \quad \text{------ (3)}$$

Also ,

$$t^{m+1} + t^m(t-h) + \cdots + t(t-h)^m - (m+l)(t-h)^{m+1}$$

$$= \{t^{m+l} + (t-h)^{m-1}\} + (t-h)\{t^m - (t-h)^m\} + \cdots + (t-h)^m\{t-(t+h)\}$$

$$= h[\{t^m + t^{m-!}(t-h) + \cdots (t-h)^m\}+\} + \cdots + (t-h)^m\{t^{m-1} + t^{m-2} + (t-2) + \cdots + (t-h)^{m-1}\}\} + \cdots + (t-h)^m]$$

So, from (3) we get,

$$\frac{f^{(m)}(\propto+h)-f^{(m)}(\propto)}{h} - \frac{(m-1)!}{2\pi i}.\oint_c \frac{f(Z)}{(Z-\propto)^{m+2}}dZ$$

$$= \frac{hm!}{2\pi i}\oint_c f(Z)\frac{t^m+t^{m-1}(t+n)+\cdots+(t-n)^m+t^{m-1}(t+n)+\cdots+(t+h)^m}{t^{(m+2)}(t-h)^{m-1}}.dZ \text{--- (4)}$$

For Ƶ∈C, we get,

l+l=|Ƶ-∝|≥d and |t – h| =|Ƶ- ∝-h|≥|Ƶ-∝|-|h|≥d-|h|

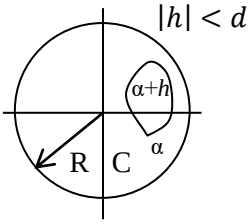
And
|t|=|Ƶ - ∞|≤2R and |t-h| =|Ƶ- (∞+h)|≤2R
so, for Ƶ∞C, we have.

$$\left|\frac{t^{m}+t^{m-1}(t-h)+---+(t-n)^{m}}{t^{(m+2)}(t-h)^{m+1}}\right| \leq \frac{(2R)^{m}N}{d^{m+2}(d-lhl)^{m+1}},$$

where N is the number of terms in the numerator.

Since f is continuous on C, there exists a positive number M such that, lf(Ƶ)l ≤ M. for all Ƶ∈C suppose that L is the length of C, then by M-L formula, we get from (4).

$$\left|\frac{f^{(m)}(\propto+h)-f^{(m)}(\propto)}{h}-\frac{<m+1}{2\pi i}\oint_{c}\frac{f(Ƶ)}{(Ƶ-\propto)^{m+2}}dƵ\right|$$



$$\leq \frac{lhlm!}{2\pi}\frac{(2R)^{m}MLN}{d^{m+2}(d-lhl)^{m+1}} \rightarrow \text{as } h \rightarrow 0$$

So,

$$\lim_{n\rightarrow 0}\frac{f^{(m)}(\propto+h)-f^{(m)}(\propto)}{h}=\frac{(m+1)!}{2\Pi i}\oint_{c}\frac{f(Ƶ)}{(Ƶ-\propto)^{m+2}}dƵ.$$

i.e,

$$f^{(m)}(h)=\frac{(m+1)!}{2\pi i}\oint_{c}\frac{f(Ƶ)}{(Ƶ-\propto)^{m+2}}dƵ$$

Therefore, By Mathematical induction, the theorem is proved.

Some Problems on Cauchy's integral formula:

1) Dictation of the proof of the Cauchy's Integral Formula.

2) Cauchy's Integral For multiple connected domain.

3) Evaluate $\oint_{lƵl=2}\frac{ƵdƵ}{(9-Ƶ^{2})(z-i)}$

(*) Cauchy's Integral Formula for derivatives : -

(i) Cauchy's Integral Formula for the first derivative.

Cauchy's integral formula:

Let, f be analytic within and on a simple closed rectifiable curve C and $\propto$ be any

interior point of C then

$$f(\propto) = \frac{1}{2\pi i}\oint_c \frac{f(Z)}{(Z-\propto)}dZ$$

⇨ Choose a circle C_0, with centre $\propto$ and radius r_0 such that C_0 lies in the interior of C
Now, $\propto$ is the only point inside C at which the function $\frac{f(Z)}{(Z-\propto)}$ is not analytic and analytic in the region D consisting of all point inside and on C except the points interior to C_0
Hence $\oint_c \frac{f(Z)dZ}{Z-\propto} = \oint_{c_0} \frac{f(Z)}{Z-\propto}dZ$

$$= \oint_{c_0} \frac{(f(Z)-f(\propto)+f(\propto))}{Z-\propto}dZ$$

$$= \oint_{c_0} \frac{f(Z)-f(\propto)}{Z-\propto}dZ + \oint_{c_0} \frac{f(\propto)}{Z-\propto}dZ$$

$$= \oint_{c_0} \frac{f(Z)-f(\propto)}{Z-\propto}.dZ + f(\propto)\oint_{c_0} \frac{dZ}{Z-\propto}$$

$$= \oint_{c_0} \frac{f(Z)-f(\propto)}{Z-\propto}.dZ + f(\propto)\times 2\pi i$$

Thus,

$$\oint_{c_0} f(Z)\frac{dZ}{Z-\propto} = \oint_{c_0} (\frac{f(Z)-f(\propto)}{Z-\propto}).dZ + 2\pi i\,.f(\propto)$$

We now claim that

$$\oint_{c_0} \frac{f(Z) - f(\propto)}{Z - \propto} dZ = 0$$

Sine, f(Z) is analytic inside and on C it is continuous at $\propto$.

Given $\varepsilon(>0)$ there exists $\delta(>0)$ such that,

$$|Z - \propto| < \delta \Rightarrow |f(Z) - f(\propto)| < \varepsilon.$$

We choose $\gamma_0 < \delta$ then $|Z - \propto| = \gamma_0 \Rightarrow |f(Z) - f(\propto)| < \varepsilon$

Hence,

$$\left| \oint_{c_0} \frac{f(Z) - f(\propto)}{Z - \infty} dZ \right| < \left(\frac{\varepsilon}{\gamma 0}\right). 2\pi\gamma_0 [By\ M - L\ formula]$$

$$= 2\pi\varepsilon$$

Thus,

$$\left| \oint_{c_0} \frac{f(Z) - f(\propto)}{Z - \propto} dZ \right| < 2\pi\varepsilon .$$

Since ε is arbitrary, we have

$$\oint_{c_0} \frac{f(Z) - f(\propto)}{Z - \propto} dZ = 0.$$

From (1) we have

$$\oint_{c} \frac{f(Z)}{Z - \propto} dZ = 2\pi i\, f(\propto)$$

and so

$$f(\propto) = \frac{1}{2\pi i} \oint_c \frac{f(Z)}{Z - \propto} dZ$$

(2) Solve:- $\oint_{|Z|=2} \frac{Z dZ}{(9 - Z^2)(Z + i)}$

Let, $f(Z) = \frac{Z}{9 - Z^2}$. Clearly f(Z), is analytic C within and on C, where C is the Circle $|Z| = 2$

By Cauchy's integral formula, $\oint_c \frac{Z}{(9-Z^2)(Z+i)} dZ = \oint_c \frac{f(Z)}{(Z+i)} dZ$

$= 2\pi i\, f(-i)$

$= 2\pi i \times \frac{-i}{9-(i)^2}$

$= -\frac{2\pi i^2}{10} = \frac{2\pi}{10} = \frac{\pi}{5}$

Theorem:- Let, f be analytic in a domain D then all derivatives of F exists and are analytic in D.

Proof: - let, $Z_0 \epsilon D$ and C be circle with centre at Z_0 and contained in D. If $\propto$ is an interior point of C, then

$$f^{(n)}(\propto) = \frac{n!}{2\pi i} \oint_c \frac{f(Z)}{(Z-\propto)^{n+1}} dZ$$

for n=0,1,2, …
Thus, f has derivatives of all orders in a neighborhood of Z_0. Since Z_0 is any point of D, the theorem is proved.

Summary

This unit dealt mainly with the preliminaries of complex integration and related properties of which the Cauchy's Fundamental Theorem and the Cauchy's Integral formula are very useful in solving relevant problems.

Unit 14

Course Structure

1. Morera's Theorem
2. Cauchy's Inequality
3. Liouville's Theorem
4. Fundamental Theorem of classical algebra.

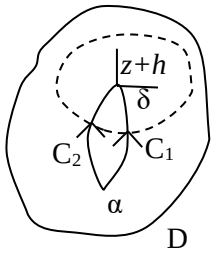
Introduction

This unit deals in Morera's Theorem, which can be viewed as a converse of Cauchy's Fundamental Theorem; Cauchy's inequality; Liouville's theorem and hence, derive the Fundamental Theorem of classical algebra using these results. Let us first start with the Morera's Theorem.

Morera's Theorem:-

If f is continuous in a simply connected domain D and if $\oint_c f(\mathrm{Z})\, d\mathrm{Z} = 0$ for every closed rectifiable curve C in D, then f is analytic D.

Proof:- Let $\propto$ be a fixed and Z be a variable point in D. let, C_1 and C_2 be any two rectifiable curves in D joining $\propto$ and Z. Then the curve consisting of C_1 and C_2 is a closed rectifiable curve in D. So, by the given condition, we have



$$\int_{c_1} f(\mathrm{Z})d\mathrm{Z} + \int_{c_2} f(\mathrm{Z})d\mathrm{Z} = 0$$

$$\Rightarrow \int_{c_1} f(\mathrm{Z})d\mathrm{Z} - \int_{c_2} f(\mathrm{Z})d\mathrm{Z} = 0$$

$$\Rightarrow \int_{c_1} f(\mathrm{Z})d\mathrm{Z} = \int_{c_2} f(\mathrm{Z})d\mathrm{Z} = 0$$

This shows that, the integral of f is independent of the path, so long as the path lies in D. we now define a function ϕ in D as

$$\phi(\mathrm{Z}) = \int_{\propto}^{\mathrm{Z}} f(t)dt ----(1)$$

The definition of ϕ is justified because integral (1) depends only on the upper limit Z and not on the path joining $\propto$ and Z, where $\propto$ is fixed.

Then, ϕ (Z+h) = $\int_{\propto}^{\mathrm{Z}+h} f(t)dt$

And so,

ϕ (Z+h) - ϕ (Z) = $\int_{\infty}^{\mathrm{Z}+h} f(t)dt$ - $\int_{\infty}^{\mathrm{Z}} f(t)dt$ = $\int_{\infty}^{\mathrm{Z}+h} f(t)dt$ - - - (2)

the integral (2) being independent of the path of integration, we may take the same along the line segment joining Z and Z+h.

Now,

$$\frac{\phi(\mathrm{Z}+h) - \phi(\mathrm{Z})}{h} - f(\mathrm{Z}) = \frac{l}{h}\int_{\mathrm{Z}}^{\mathrm{Z}+h} f(t)dt - f(\mathrm{Z}) = = \frac{l}{h}\int_{\mathrm{Z}}^{\mathrm{Z}+h} \{f(t) - f(\mathrm{Z})\}dt ---(3)$$

Since, by the given condition, f is continuous at Z, for given Ɛ (>0) there exists a Ɛ (>0) such that

|f(t) – f(Z)|<Ɛ whenever |t-Z| <δ

We choose, Z +h in D such that |h| <δ.

Then, for every point t, on the straight line joining Z to Z+h, we have.

|f (t) – f(Z) | <Ɛ.

Hence, from (3), we get by ML formula,

$$|\frac{\phi(\mathrm{Z}+\mathrm{h}) - \phi(\mathrm{Z})}{h} - \mathrm{f}(\mathrm{Z})| = \frac{\mathrm{l}}{|h|}|\int_{\mathrm{Z}}^{\mathrm{Z}+h} \{f(t) - f(\mathrm{Z})\}.d\mathrm{Z}| \leq \frac{\mathrm{l}}{|h|}.\mathrm{Ɛ}.|\mathrm{h}| = \mathrm{Ɛ}$$

for 0<|h|<δ.

This gives,

$$\lim_{h\to 0}\frac{\phi(Z+h)-\ \phi\ (Z)}{h}=f\ (Z)$$

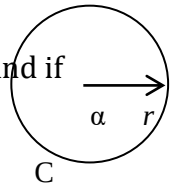
i.e. $\phi'(Z) = f(Z)$.

Thus, $\phi'(Z)$ exists at every point of D and so, ϕ is analytic in D. Since the derivative of an analytic function is analytic, we see that $f(Z) = \phi'(Z)$ is analytic in D.

This proves the theorem.

Cauchy's Inequality :-

Theorem :- If f is analytic within and on a circle C with centre $\propto$ and radius γ and if



$|f(Z)|\leq M\ \forall Z\epsilon$ C, where M is a positive number, then,

$|f^{(n)}(\propto)|\ \leq\ \frac{M<n!}{\gamma^n}$ for n=0,1,2,3,....

Proof :- By Cauchy's integral formula for general order derivative, we have,

$$f^{(n)}(\propto)\ =\frac{n!}{2\pi i}\oint_c\frac{f(Z)}{(Z-d)^{n+1}}dZ---(1)$$

for n=0,1,2, ----

For $Z\epsilon C$, we get,

$$\left|\frac{f(Z)}{(Z-\propto)^{n+1}}\right|\ \leq\frac{M}{\gamma^{n+1}}$$

therefore, from (1), we get, by M-L formula ,

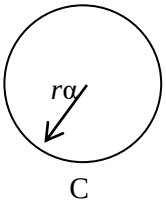
$$f^{(n)}(\propto)=\frac{n!}{2\pi}\left|\oint_c\frac{f(Z)}{(Z-\propto)^{n+1}}dZ\right|\ \leq\frac{n!}{2\pi}\cdot\frac{M}{\gamma^{n+1}}2\pi\gamma=\ \frac{Mn!}{\gamma^{n}}\ \ for, n=0,1,2,\ldots$$

This proves the theorem.

Definition : - A function of a complex variable, which is analytic throughout the complex plane, is called an Entire or an integral function.

For example :- A polynomial, e^{Z}, sinZ, CosZ etc. are entire function.

Theorem:- **Liouville's theorem** :- Every bounded entire function is constant .



Proof: - Let, $\propto$ be any point in the complex plane, and let C: |Z -$\propto$|=γ .

Then there exists a constant M such that $|f(Z)| \leq M \; \forall Z \epsilon C$

and so $|f(Z)| \leq M \; \forall \; Z \epsilon C$, no matter how large the radius γ is.

The function f is analytic within and on the circle C and $\propto$ is a point interior to C.

So, by Cauchy's integral formula for derivatives, we get

$$f'^{(\propto)} = \frac{1}{2\pi i} \oint_c \frac{f(Z)}{(Z-\propto)^2} dZ.$$

Also, on C, we get,

$$\left| \frac{f(Z)}{(Z-\propto)^2} \right| \leq \frac{M}{\gamma^2}$$

Therefore, By ML, formula, We get,

$$|f'^{(\propto)}| = \left| \frac{1}{2\pi i} \oint_c \frac{f(Z)}{(Z-\propto)^2} dZ \right| \leq \frac{1}{2\pi} \cdot \frac{M}{\gamma^2} \cdot 2\pi\gamma = \frac{M}{\gamma} \to 0 \; as \; \gamma \to \propto$$

And So, $f'(\propto) = 0$

Since $\propto$ is an arbitrary point of the complex plane $\mathbb{C}$, we have,

$$f'(\propto) = 0 \; \forall \; Z \epsilon \mathbb{C}$$

We now choose any two points Z_1 and Z_2 of $\mathbb{C}$

Then, substituting w = f(Z), we get

$$0 = \int_{Z1}^{Z2} f'(Z) dZ = \int_{f(Z1)}^{f(Z2)} dw \; = \; f(Z_2) - \; f(Z_1)$$

i.e. $f(Z_1) = f(Z_2)$.

Since, Z_1 and Z_2 are any two points of $\mathbb{C}$, it follows that, f is a constant function.

This proves the theorem.

Fundamental Theorem of Classical Algebra :-

Theorem:- If f is a polynomial of degree n with real or complex coefficients, then the equation,

$f(Z) = 0$ has at least one root.

Proof :- Let, $f(Z) = a_0 + a_1Z + - - - a_nZ^n$ ($a_n \neq 0$) be a polynomial of degree n. If possible, suppose that no value of Z exists for which $f(Z) = 0$.

We shall show that this leads to a contradiction. Since, f is a polynomial, it is an entire function. Also since $f(Z) \neq 0$ for any Z, it follows that, $\phi(Z) = \frac{1}{f(Z)}$ is an entire function.

Now, for $Z \neq 0$ we have

$|f(Z)| = |a_0 + a_1Z + \text{----} + anZ^n|$

$= |Z|^n|a_n + \frac{an-1}{Z} + - - - + \frac{a0}{Zn}|$

$\geq |Z|^n \{ |a_n| - \frac{|a_{n-1}|}{|Z|} - - - - \frac{|a_n|}{|Z|^n}$

This shows that, $\lim_{Z \to \infty} lf(Z) = \infty$

and so, $\lim_{Z \to \infty} \phi(Z) = 0$

Therefore, ϕ is a bounded entire function, and so by Liouville's theorem is a constant, Hence, $f(Z)$ is also a constant which is impossible. So, $f(Z) = 0$ has at least one root in the complex plane $\mathbb{C}$.

Unit 15

Course Structure

1. Uniformly convergent series of analytic functions
2. Power series
3. Taylor's Theorem
4. Laurent's Theorem

Introduction

Just as series of real functions, we will be reading about the series of complex valued functions, their region of convergence, radius of convergence and uniform convergence. As in real analysis, here we will see that any analytic function can be expressed as a Taylor series in its region of convergence. Furthermore, we will learn about Laurent's series expansion of a complex valued function analytic in an annular region.

Uniformly Convergent Series:-

Let, $\sum_{n=1}^{\infty} f_n(Z)$ be an infinite series whose terms are functions of a complex variable defined on a set E in the Complex plane. Further let, $S_n(Z) = f_1(Z) + f_2(Z) + \text{----} + f_n(Z)$ be the nth partial sum of the series.

The series $\sum_{n=1}^{\infty} f_n(Z)$ is said to be uniformly convergent on E for E, for given $\varepsilon(>0)$ there exists a positive integer $N=N(\varepsilon)$, depending only on ε such that

$$|f_{n+p}(Z) - f_n(Z)| < \varepsilon \ for\ all\ n > N,$$

p=0,1,2, ….

and all $Z \epsilon E$.

It follows, from Cauchy's general principle of convergence that every series which is uniformly convergent on a set E is also convergent on E. Hence, there exists a function f(Z) on E, called the sum of the series, such that

$$F(\mathrm{Z}) \;=\; \sum_{n=1}^{\infty} \mathrm{f}_n(\mathrm{Z}) \;\forall\; \mathrm{Z}\epsilon E$$

Example:- The geometric series $\sum_{n=1}^{\infty} \mathrm{Z}^n$ converges in E={Z: lZl <1}, because

$$Sn(\mathrm{Z}) \;+\; \mathrm{Z} + \mathrm{Z}2 + \;-\;-\;-\;-\; + \mathrm{Z}n$$

$$= \;\frac{\mathrm{Z}(l - \mathrm{Z}^{\,n})}{l - \mathrm{Z}} \rightarrow \frac{\mathrm{Z}}{l - \mathrm{Z}}\; for\; \mathrm{Z}\epsilon E$$

Now

$\mathrm{S_{n+p}}(\mathrm{Z})- \mathrm{S_n}(\mathrm{Z})$

$= (\mathrm{Z}+\mathrm{Z}^2+ \text{- - -} + \mathrm{Z}^{\mathrm{n}} +\mathrm{Z}^{\mathrm{n\text{-}1}} + \text{- - -} +\mathrm{Z}^{\,\mathrm{n+p}}) - (\,\mathrm{Z}+\mathrm{Z}^2+\mathrm{Z}^3+ \text{- - -} +\mathrm{Z}^{\mathrm{n}})$

$= \mathrm{Z}^{\mathrm{n+l}} + \mathrm{Z}^{\mathrm{n+2}} + \mathrm{Z}^{\mathrm{n+p}}$

$=\mathrm{Z}^{\mathrm{n+l}}\,(1+2+ \text{- - -} + \mathrm{Z}^{\mathrm{p-l}})$

$=\mathrm{Z}^{\mathrm{n+l}}\,\frac{1-\mathrm{Z}^{\mathrm{p}}}{\mathrm{l-Z}}$

So

$$|S_{n+p}(\mathrm{Z}) - S_n(\mathrm{Z})| \;=\; |\mathrm{Z}|^{n+l}|\frac{1 - |\mathrm{Z}|^p}{l - \mathrm{Z}}|$$

$$\geq\; |\mathrm{Z}|^{n+p}\,\frac{1 - |\mathrm{Z}|^p}{|\,l - \mathrm{Z}|}$$

We chose, p=n and $\mathrm{Z_n} = 1 - \frac{1}{n}\,\epsilon\mathrm{E}$

Then

$$|S_{2n}(\mathrm{Z}_n) - S_n(\mathrm{Z}_n)| \geq\; (1 - \frac{l}{n}\,)^{n+1}\,(\frac{1 - (1 - \frac{l}{n}\,)^n}{|\,1 - 1 + \frac{l}{n}|}\,)$$

$$=\; (1 - \frac{l}{n}\,)^n\,(1 + \frac{l}{n}\,)^n\,\{1 - (1 - \frac{l}{n}\,)^n\}\; \rightarrow \infty\; as\; n \rightarrow \infty$$

Thus, for sufficiently large n and suitable p and ~~z~~ in E, we can make the difference $|\mathrm{S_{n+p}}(\mathrm{Z}) - \mathrm{S_n}(\mathrm{Z})|$ as large as we please. So, the series $\sum_{n=1}^{\infty} \mathrm{Z}^{\mathrm{p}}$ is not uniformly convergent in E ={Z : |Z|< 1}

Theorem: A convergent series $\sum_{n=1}^{\infty} f_n(\text{Ƶ}) = f(\text{Ƶ}), (\text{Ƶ}\epsilon E)$is uniformly convergent on E iff given Ɛ(>0) there exists a positive integer N=N(Ɛ) such that $|S_n(\text{Ƶ})$ –f(Ƶ) <Ɛ for n>N and for all ƵϵE.

Proof : - If the series $\sum_{n=1}^{\infty} f_n(\text{Ƶ})$ is uniformly convergent on E, then for given Ɛ(>0) there exists a positive integer N=N(Ɛ) such that $|S_{n+p}\ (\text{Ƶ}) - S_n(\text{Ƶ})\ | < \frac{\varepsilon}{2}$ whenever n>N, p=0,1,2, … and for all ƵϵE.

Letting p→∞, we get

$$|S_n(\text{Ƶ}) - f(\text{Ƶ})\ | \ \leq \frac{\varepsilon}{2} < \varepsilon \ whenever\ n > N\ and\ for\ all\ \text{Ƶ}\epsilon E.$$

Conversely,

If there exists a positive integer N=N(Ɛ) such that ,

$$|S_n(\text{Ƶ}) - f(\text{Ƶ})\ | \ < \frac{\varepsilon}{2}\ for\ n > N\ and\ for\ all\ \text{Ƶ}\epsilon E,$$

then for n>N and for all ƵϵE we get for p=0,1,2, …

$$|S_{n+p}(\text{Ƶ}) - S_n(\text{Ƶ})\ | \ \leq \ | S_{n+p}(\text{Ƶ}) - f(\text{Ƶ})\ | \ + \ | S_n\ (\text{Ƶ}) - f(\text{Ƶ})|$$

$$< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

Therefore the series $\sum_{n=1}^{\infty} f_n(\text{Ƶ})$ converges uniformly on E.

This proves the theorem.

Weierstrass M-Test :-

Theorem:- Given a convergent series $\sum_{n=1}^{\infty} M_n$ of positive constants and a series $\sum_{n=1}^{\infty} f_n(\text{Ƶ})$ of functions defined on a set E. If there exists a positive integer N_1 such that $|f_n(\text{Ƶ})\ | \leq M_n$ for n≥N, and forrall ƵϵE,

then $\sum_{n=1}^{\infty} f_n(\text{Ƶ})$ is uniformly and absolutely convergent on E.

Proof: Since, $\sum_{n=1}^{\infty} M_n$ Converges, for given Ɛ(>0) there exists a positive integer $N_2 = N_2(\varepsilon)$ such that , $M_{n+1}+M_{n+2}+ - - - + M_{n+p}<\varepsilon$ for $n>N_2$ and p=1,2, ---

Let, N=Max { N_1, N_2} . Then, N depends only on Ɛ> Now for n>N and for all ƵϵE, p=1,2, --- we get.

$| S_{n+p}(\text{Ƶ}) - S_n(\text{Ƶ})\ | \ = \ | f_{n+1}(\text{Ƶ}) \ + \ f_{n+2}\ (\text{Ƶ}) \ + \ - \ - \ - \ + \ f_{n+p}(\text{Ƶ})\ |$

$\leq |f_{n+1}(Z) + | + |f_{n+2}(Z) + - - - + |f_{n+p}(Z)|$

$\leq M_{n+1} + M_{n+2} + - - - + M_{n+p}$

$< \varepsilon$.

So, $\sum_{n=1}^{\infty} f_n(Z)$ is uniformly convergent on E.

We further note that, $\sum_{n=1}^{\infty} f_n(Z)$ is absolutely convergent. This follows from comparison test.

This proves the theorem.

Definition:- If every point of a set E is a limit point of E, then E is called dense in itself.

Example:- A domain or a continuous curve is in itself.

Theorem : Given a uniformly convergent series $f(Z) = \sum_{n=1}^{\infty} f_n(Z)$ defined on a dense in itself set E such that each term $f_n(Z)$ is continuous on E. Then the sum f(Z) is also continuous on E.

- ⇨ If Z and Z_0 are any points of E. Then $|f(Z) - f(Z_0)| = |f(Z) - S_n(Z) + S_n(Z) - S_n(Z_0) + S_n(Z_0) - f(Z_0)$
- ⇨ $|S_n(Z) - f(Z)| + |S_n(Z) - S_n(Z_0)| + |S_n(Z_0) - f(Z_0)|$ ------ (1)
- ⇨ Since, $f(Z) = \sum_{n=1}^{\infty} f_n(Z)$ converges uniformly in E, for given $\varepsilon>0$ there exists a positive integer $N=N(\varepsilon)$, such that,
- ⇨ $|S_n(Z) - f(Z)| < \frac{\varepsilon}{3}$ whenever n>N and forall $Z \epsilon E$.

Let, $n_0>N$ be a fixed positive integer. Then from (i), we get,

$|f(Z) - f(Z_0)| \leq |S_{n0}(Z) - f(Z)| + |S_{n0}(Z) - S_{n0}(Z_0) + |S_{n0}(Z_0) - f(Z_0)| < \frac{\varepsilon}{3} + \frac{\varepsilon}{3}|S_{n0}(Z) - S_{n0}(Z_0)|$ - - - - (2)

Since $S_{n0}(Z)$ is continuous at Z_0, $\exists$ a $\delta>0$ such that,

$|S_{n0}(Z) - S_{no}(Z_0)| < \frac{\varepsilon}{3}$ whenever $|Z - Z_0| < \delta$.

Hence, from (2), we get,

$|f(Z) - f(Z_0)| < \frac{2\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon$ whenever $|Z - Z_0| < \delta$.

So, f is continuous at Z_0 and since Z_0 is arbitrary, it follows that f is continuous on E. This proves the theorem.

Theorem:- Given a rectifiable curve C, suppose that the series f(Ƶ) = , $\sum_{n=1}^{\infty} f_n(Ƶ)$ is uniformly convergent on C and every term $f_n(Ƶ)$ is continuous on C. Then the series can be integrable term by term on C, i.e,

$$\oint_c f(Ƶ)dƵ = \sum_{n=1}^{\infty} \oint_c f(Ƶ)dƵ$$

Proof : Since, each term of $\sum_{n=1}^{\infty} f_n(Ƶ)$ is continuous on C, f(Ƶ) is also continuous on C and so it is integral on C. Since f(Ƶ) = $\sum_{n=1}^{\infty} f_n(Ƶ)$ converges uniformly on C, for given Ɛ(>0) ∃ converges uniformly on C, for given Ɛ(>0) ∃a positive integer N=N(Ɛ) such that.

$|S_n(Ƶ) - f(Ƶ)| < \frac{Ɛ}{l}$ wherever n>N for all ƵϵC whereℓ is the length of C.

Now by ML formula.

$|\sum_{j=1}^{n} \int_c f_j(Ƶ)dƵ - \int_c f(Ƶ)dƵ|$

$= |\sum_{j=1}^{n} \int_c f_j(Ƶ)d(Ƶ) - \int_c f(Ƶ)dƵ|$

$=| \int_c S_n(Ƶ)dƵ - \int_c f(Ƶ)dƵ|$

$=\left|\int_c \{S_n(Ƶ) - f(Ƶ)\right\} dƵ|$

$< \frac{Ɛ}{\ell}.\ell = Ɛ$ whenever n>N.

Therefore $\lim_{n-\infty} \sum_{j=1}^{n} \int_c f_j(Ƶ)dƵ = \int_c f(Ƶ)dƵ$

i.e. , $\sum_{n=1}^{\infty} \int_c f_n(Ƶ)dƵ = \int_c f(Ƶ)dƵ$

Theorem:- A series f(Ƶ) = $\sum_{n=1}^{\infty} \int_c f_n(Ƶ)$, which is convergent on a domain G, is uniformly convergent on every compact subset of G if and only if every point $Ƶ_0$ϵG has a nbd N($Ƶ_0$) ⊂G in which the series is uniformly convergent.

The condition is necessary. We suppose that the given series converges uniformly on every compact subset of G.

Let, $Ƶ_0$ ϵ G be an arbitrary point. Further suppose that, $\bar{N}(Ƶ_0)$ = {Ƶ : |Ƶ- $Ƶ_0$| < Ɛ}

and $\bar{N}(Ƶ_0) = \{Ƶ : |Ƶ - Ƶ_0| \leq Ɛ\}$,

where Ɛ(>0) is so small that $\bar{N}(Ƶ_0) \subset G$.

Then the series converges uniformly on $\bar{N}(Ƶ_0)$, because $\bar{N}(Ƶ_0)$ is a compact subset of G. Since, $N(Ƶ_0), \subset \bar{N}(Ƶ_0)$, it follows that the given series converges uniformly on $N(Ƶ_0)$.

The condition is sufficient. We suppose that every point Ƶ of G has an open neighborhood N(Ƶ) on which the given series converges uniformly. Let A be a compact subset of G. Then {N(Ƶ) : ƵϵA} is an open cover of A. Since A is compact, there exists $Ƶ_1, Ƶ_2$, - - -, ƵpϵA such that $A \subset \bigcup_{J=1}^{P} \bar{N}(Ƶ_j)$.

Since the given series converges uniformly on each N(Ƶj) (j=1,2, - - -, P), it converges uniformly on $U_{j=1}^{p} f(Ƶ_j)$and so, converges uniformly on A, because $A \subset \bigcup_{n=1}^{P} N(Ƶ_j)$. Since A is any compact subset of G, it follows that the given series converges uniformly on every compact subset of G.

This proves the theorem.

Weierstrass Theorem on Uniformly Convergent series of Analytic function: -

Theorem: - If the series $f(Ƶ) = \sum_{n=1}^{\infty} f_n(Ƶ)$ is uniformly convergent on every compact subset of a domain G and if every term $f_n(Ƶ)$ is analytic on G, then the sum $f(Ƶ)$ of the series is also analytic on G. Moreover, the series can be differentiated as$f^k(Ƶ) = \sum_{n=1}^{\infty} f_n^{(k)}(Ƶ)$ (k=0,1,2, - - -) for all ƵϵG.

Also, each differentiated series is uniformly convergent on every compact subset of G.

Proof : - Let, $Ƶ_0$ be an arbitrary point of G.

We choose p>0 such that G contains the circle $ϒ_p$: $|Ƶ - Ƶ_0| = p$ and its interior. Since

by the hypothesis the series $f(Ƶ) = \sum_{n=1}^{\infty} f_n(Ƶ)$, converges uniformly on ϒp, each of the series.

$$\frac{\lfloor k}{2\pi i}\sum_{n=1}^{\infty}\frac{f_n(Ƶ)}{(Ƶ-\propto)^{k+1}} = \frac{\lfloor k}{2\pi i}\frac{f(Ƶ)}{(Ƶ-\propto)^{k+1}} \text{ - - - - - - - (1)(k=0,1,2, - - - -)}$$

converges uniformly on $ϒ_p$, where ∝ is an interior point of $ϒ_p$. *For*

$$\left|\frac{\lfloor k}{2\pi i}\frac{1}{(Ƶ-\propto)^{k+1}}\right| \leq \frac{\lfloor k}{2\pi p_0^{k+1}} \text{ for every } Ƶ \epsilon ϒp.$$

where p^0 is the minimum distance of ∝ from ϒp. Since, each $f_n(Ƶ)$ is analytic and so, it is continuous in G and $f(Ƶ) = \sum_{n=1}^{\infty} f_n(Ƶ)$ converges uniformly on ϒp we see that $f(Ƶ)$is continuous on ϒp. Therefore, we can integrate (1) term-by-term to obtained,

$\sum_{n=1}^{\infty} \frac{\lfloor k}{2\pi i} \oint_{\gamma_p} \frac{f_n(Z)}{(Z-\alpha)^{k+1}} = \frac{\lfloor k}{2\pi i} \oint_{\gamma_p} \frac{f_n(Z)}{(Z-\alpha)^{k+1}} dZ$. - - - - -(2)

for k= 0,1,2, - - -

For, k= 0, we get from (2)

$$\sum_{n=1}^{\infty} \frac{1}{2\pi i} \oint_{\gamma_p} \frac{f_n(Z)}{(Z-\alpha)} dZ = \frac{1}{2\pi i} \oint_{\gamma_p} \frac{f(Z)}{(Z-\alpha)} dZ$$

i.e, $f(\alpha) = \frac{l}{2\pi i} \oint_{\gamma p} \frac{f(Z)}{(Z-\alpha)} dZ.$ $- - - - (3)$

Since f is continuous on γp, it is bounded there. So, proceeding as the proof of Cauchy's integral formula for the first derivative, we can deduce from (3), that ,

$f'(Z_0) = \frac{l}{2\pi i} \oint_{\gamma p} \frac{f(Z)}{(Z-Zo)^2} dZ.$

Since, Z_0 is an arbitrary point of G, it follows that f is analytic on G.

Again, from (2), we get for k=1,2, - - - and for $\alpha = Z_0$,nsing Cauchy's integral formula,

for general order derivation, that,$\sum_{n=1}^{\infty} f\frac{k}{n}(Zo) = f^{k}(Zo)$.

Let $N(Z_0) = \{ Z : |Z - Z_0| \leq \frac{p}{2} \}$ be a nbd of

$Z_0 \epsilon$ G. Since $f(Z) = \sum_{n=1}^{\infty} f_n(Z)$ converges uniformly on γp, for given ε (>0) there

exists some (+positive) integer $N(\varepsilon)$ depending only on ε such that

$|S_n(Z) - f(Z)| < \varepsilon$ for n >N (ε) and for all $Z \epsilon \gamma p$

Now, for all $Z \epsilon N(Z_0)$ and n>N(ε), we get by the ML formula ,

$|\sum_{j=1}^{n} f\frac{K}{n}(Z) - f^{K}(Z)| = |\sum_{j=1}^{n} \frac{\lfloor k}{2\pi i} \oint_{\gamma p} \frac{fn(\xi)}{(\xi-Z)^{k+1}} d\xi - \frac{\lfloor k}{2\pi i} \oint_{\gamma p} \frac{f(\xi)}{(\xi-Z)^{k+1}} d\xi|$

$= |\frac{\lfloor k}{2\pi i} \oint_{\gamma p} \frac{Sn(\xi)-f(\xi)}{(\xi-Z)^{k+1}} d\xi$

$\leq \frac{\lfloor k}{2\pi} \cdot \frac{\xi}{(\frac{p}{2})^{k+1}}$

Because, $|\xi - Z| \geq \frac{p}{2}$ for $\xi \epsilon \gamma p$.

This shows that the series $f^{k}(Z) = \sum_{n=1}^{\infty} f_n^k(Z)$ converges uniformly on $N(Z_0)$. Since Z_0 is on arbitrary point of G, we see that every point of G has a nbd in which the differentiated series converges uniformly. Therefore,

$$f^{k}(Z) = \sum_{n=1}^{\infty} f_n^k(Z)$$

converges uniformly on every compact subset of G.

This proves the theorem.

Definition:- A series of the form $\sum_{n=0}^{\infty} a_n(Z - Z_0)^n$ is called a power series where Z_0, a_0, a_1, a_2, -- - are given complex numbers.

Cauchy – Hadamard Theorem:-

For a power series $\sum_{n=0}^{\infty} a_n(Z - Z_0)^n$, let, $R=\frac{1}{\Lambda}$, where $\Lambda = \limsup_{n\to\infty} \sqrt[n]{|a_n|}$, and let Υ be the circle given by Υ: $|Z - Z_0| = R$, with interior $I(\Upsilon)$ and exterior $E(\Upsilon)$. Then, there are there

Possibilities:

(1) If R=0, then $\sum_{n=0}^{\infty} a_n(Z - Z_0)^n$ converges only for $Z=Z_0$
(2) If, $0<R<\infty$, then $\sum_{n=0}^{\infty} a_n(Z - Z_0)^n$ converges absolutely for $Z \in I(\Upsilon)$ and does not converge for any $Z \in E(\Upsilon)$
(3) If $R=\infty$, then $\sum_{n=0}^{\infty} a_n(Z - Z_0)^n$ converges absolutely for all finite Z.

Proof:
We examine each of the three possibilities one-by-one. We also note that the given power series converges absolutely for $Z=Z_0$

Case- I : Let R=0.Then $\Lambda = \infty$ and so

$\frac{l}{|Z-Z0|} < \Lambda$ for any $Z \neq Z_0$
$Hence\ \frac{l}{|Z-Z0|} < \limsup_{n\to\infty} \sqrt[n]{|a_n|}$ and so,

$\frac{l}{|Z-Z0|} < \sqrt[n_k]{a_{n_k}}$ | for k=1,2, - - -, where $\{a_{n_k}\}$ is a subsequence of $\{a_n\}$
$Now, raising\ power\ to\ n_k$ and then, by cross multiplication, we get,
$|a_{n_k}(Z - Z_0)a_n| > 1$ for k=1,2, - - -

This implies{ $a_n(Z - Z)^2$} does not converge to zero for any $Z \neq Z_0$.
Therefore the given power series does not converge for any $Z \neq Z_0$, when R=0.

Case 2 :-Let, 0<R<∞. So, 0 <Λ < ∞.Let, Zϵ I (Y) and $Z \neq Z_0$, Then $|Z - Z_0|$ <R = $\frac{l}{\Lambda}$ and so, we can put,

$|Z - Z_0| = \frac{\theta^2}{\Lambda}$ where 0<θ<1.

We note that

$\Lambda < \frac{\Lambda}{\theta} = \frac{\theta}{|Z - Z_0|}$

i.e, $\sqrt[n]{|a_n|} < \frac{\theta}{|Z - Z_0|}$ for all large values of n

i.e, $|a_n(Z - Z_0)^2| < \theta^n$ for all large values of n Since, $\sum_{n=0}^{\infty} \theta^n$ is a convergent geometric series, by comparison test, we see that the given power series converges absolutely for ZϵI (Y), because the series obviously converges absolutely for $Z \neq Z_0$.

One the other hand, if Zϵ E (Y), then

$|Z - Z_0| > R = \frac{l}{\Lambda}$

i.e, , $\frac{l}{|Z - Z_0|} < \Lambda$

Therefore, there exists a subsequence { $\sqrt[n_k]{|a_{n_k}|}$ } of { $\sqrt[n]{|a_n|}$ } for which $\frac{l}{|Z - Z_0|} < \sqrt[n_k]{|a_{n_k}|}$ for k=1,2, - - -

Now, raising both sides to the power nk and then by cross multiplication, we get,

$\left||a_{n_k}(Z - Z_0)^{n_k}\right| > 1$ for k=1,2,3, - -

This shows that the general term of the given power series does not converge to zero. Therefore, the given power series does not converge for any Zϵ E (Y) .

Case – 3 : Let, R= ∞. Then $\Lambda = 0$ and so for any $Z \neq Z_0$ and any θ, 0< 0< 1, we have

$\Lambda < \frac{\theta}{|Z - Z_0|}$ So, for all large values of n.

we get $\sqrt[n]{|a_n|} < \frac{\theta}{|Z - Z_0|}$

Now raising both sides to the power n and then by cross multiplication, we get,

$|a_n(Z - Z_0)^n| < \theta^n$ for all large values of n. Since the geometric series $\sum_{n=0}^{\infty} \theta^n$ is convergent, by the comparison test, the given power series, converges absolutely for any finite Z (because the given power series automatically converges absolutely for $Z=Z_0$).

This proves the theorem.

Note: The circle $\Upsilon: |Z - Z_0| + R$ is called the circle of convergence and R is called the radius of convergence of the power series $\sum_{n=0}^{\infty} a_n (Z - Z_0)^n$

Theorem : let, $\Upsilon: |Z - Z_0| + R$ be be the circle of convergence of the power series $\sum_{n=0}^{\infty} a_n (Z - Z_0)^n$, then the series is uniformly convergent on every compact subset of $I(\Upsilon)$

Proof: First, we verify that the given power series

converges uniformly in the closed circular

disc $|Z - Z_0| \leq \gamma$, where $0 < \gamma < R$. We take a number p such that $\gamma < p < R$

And ξ be a number such that $p = |\xi - Z_0|$

Then clearly $\xi \in I(\Upsilon)$ and so by Cauchy Hadamard theorem the given power series Converges absolutely for $Z = \xi$

i,e. the series

$\sum_{n=0}^{\infty} |a_n (Z - Z_0)^n| = \sum_{n=0}^{\infty} |a_n| p^n$ is convergent.

Since, for $|Z - Z_0| \leq \gamma$ *we have*

$|a_n (Z - Z_0)^n| \leq |a_n| |\xi - Z_0|^n = |a_n| p^n$

By Weierstrass M- test we see that, the given power series converges uniformly in $|Z - Z_0| \leq \gamma$.

Since, every point of $I(\Upsilon)$ has a sufficiently small neighborhood that is contained in $|Z - Z_0| \leq \gamma$ for γ sufficiently closed to R, from the above discussion we see that every point of $I(\Upsilon)$ has a neighborhood in which the given power series converges uniformly.

Therefore the given power series converges uniformly on every compact subset of $I(\Upsilon)$.

This proves the theorem.

Note :- The power series $\sum_{n=0}^{\infty} |a_n (Z - Z_0)^n$ need not converge uniformly on $I(\Upsilon)$ itself. For the geometric series $\sum_{n=0}^{\infty} Z^n$ does not converge uniformly in $|Z| < 1$, having 1 as its radius of convergence.

But by the above theorem, the series converges uniformly on every compact subset of $|Z| < 1$.

Taylor's Theorem :

Theorem:- Let f be analytic in the interior of $\propto$ circle C with centre $\propto$ and radius γ. Then at each point Z interior to C,

$F(Z) = \sum_{n=0}^{\infty} a_n (Z-\propto)^n$, where $a_n = \frac{f^{(n)}(\propto)}{\lfloor n}$

Let, Z_0 be an arbitrary, but

fixed point with in C and let, $|Z_0 - \propto| = R| < \gamma$ We now choose a positive number p such that R <p< γ let C_1 denote the circle $|Z - \propto| = p$

Then C_1 lies entirely within C and Z_0 is an interior point of C_1. Clearly f is analytic within and on C_1. Hence by Cauchy's integral formula,

We get,

$f(Z_0) = \frac{1}{2\pi i}\oint_{C_1} \frac{f(Z)}{Z-\propto} dZ$

$= \frac{1}{2\pi i}\oint_{C_1} \frac{f(Z)}{(Z-\propto)(1-\frac{Z_0-\propto}{Z-\propto})} dZ$

$= \frac{1}{2\pi i}\oint_{C_1} \frac{f(Z)}{Z-\propto} \frac{1-t^n+t^n}{1-t}\, dZ$, where $t = \frac{Z_0-\propto}{Z-\propto}$

$= \frac{1}{2\pi i}\oint_{C_1} \frac{f(Z)}{Z-\propto} .\left(1 + t + t^n + - - - + t^{n-1}\frac{t^n}{1-t}\right) dZ$

$= \sum_{n=0}^{\infty} \frac{1}{2\pi i}\oint_{C_1} \frac{f(Z)}{(Z-\propto)} t^n . dZ + \frac{1}{2\pi i}\oint_{C_1} \frac{f(Z)}{Z-\propto} . \frac{t^n}{1-t} . dZ$

$= \sum_{k=0}^{n-1} \frac{1}{\lfloor k} \{ \frac{1}{2\pi i}\oint_{C_1} \frac{f(Z)}{(Z-\propto)^{k+1}} dZ\} (Z-\propto)^k + \frac{1}{2\pi i}\oint_{C_1} \frac{f(Z)(Z_0-\propto)^n}{(Z-\propto)^n(Z-Z_0)} dZ$

$= \sum_{k=0}^{n-1} a_k (Z-\propto)^k + R_n$ ----- (1)

where, $a_k = \frac{f^k(\propto)}{\lfloor k} = \frac{1}{\lfloor k} . \frac{\lfloor k}{2\pi i}\oint_{C_1} \frac{f(Z)}{(Z-\propto)^{k+1}} dZ$

And $R_n = \frac{1}{2\pi i}\oint_{C_1} \frac{f(Z)(Z_0-\propto)^n}{(Z-\propto)^n(Z-Z_0)} dZ$

Since f is analytic within and on C_1, it is bounded on C_1. So, there exists a positive number M such that If (Ƶ) | ≤ M ∀ Ƶϵ C_1.

Also, for all Ƶϵ C_1 we have

$|\frac{Ƶ_0-\propto}{Ƶ-\propto}| = \frac{R}{p}|$

And $|Ƶ - Ƶ_0| = |(Ƶ-\propto)-(Ƶ_0-\propto)| \geq |Ƶ-\propto| - |Ƶ_0-\propto| = p-R$

Therefore for all ƵϵC_1, we obtained

$|\frac{f(Ƶ)(Ƶ_0-\propto)^n}{(Ƶ-\propto)^n(Ƶ-Ƶ_0)}| \leq \frac{M}{p-R}.\left(\frac{R}{p}\right)^n$

Now by ML formula we get

$|R_n| = |\frac{1}{2\pi i}\oint_{C_1}\frac{f(Ƶ)(Ƶ_0-\propto)^n}{(Ƶ-\propto)^n(Ƶ-Ƶ_0)}dƵ| \leq \frac{l}{2\pi}\frac{M}{p-R}\left(\frac{R}{p}\right)^n 2\pi p$

$[\because \frac{R}{p} < 1\ ] = \frac{MP}{(P-R)}\left(\frac{R}{P}\right)^n \to 0$ as n→0

So, $\lim_{n\to\infty} R_n = 0$.

Considering limit as n→∞, we get from (1)

$f(Ƶ_0) = \sum_{n=0}^{\propto} a_n (Ƶ-\propto)^n$ where $a_n = \frac{f^n(\propto)}{\llcorner n}$

Since, $Ƶ_0$ is an arbitrary point with in C_1 for every Ƶ with in C we get,

$f(Ƶ) = \sum_{n=0}^{\propto} a_n (Ƶ-\propto)^n$ where $a_n = \frac{f^k(\propto)}{\llcorner n}$.

This proves the theorem.

Note -1: - The power series representing f is called the Taylor series of f about the point ∝. The Taylor's theorem shows that if f is analytic in a nbd of ∝, then f can be represented in that nbd by a power series in (Ƶ-∝) with a positive redius of converegence.

Note-2 :- Let, f be analytic at $\propto$. Then there exists a circle C with centre at $\propto$ such that f is analytic with in C. Then for each point Z within C, we have $f(Z) = \sum_{n=0}^{\infty} a_n(Z-\propto)^n$. The radius of the greatest circle with the power series $\sum_{n=0}^{\infty} a_n(Z-\propto)^n$ converges to f(Z) is the distance of the point $\propto$ $from\ the\ sigular\ point\ of\ f\ which\ is\ nearest\ to\ \propto$.

Note – 3:- If f is an entire function, then it has a Taylor expansion about the origin of the form $f(Z) = \sum_{n=0}^{\infty} a_n Z^n$, which is valid for all Z. This series is called an entire series.

Laurent's Theorem :-

Theorem :- Let, f be an analytic function in a_n annular region D: $r_2 < |Z - \propto| < r_1$. Then at each point $Z \epsilon D$, f can represented by a series of the form $f(Z) = \sum_{n=0}^{\infty} a_n(Z-\propto)^n + \sum_{n=1}^{\infty} b_n(Z-\propto)^n$, where the coefficients a_n and b_n depend only on the function f and the point - $\propto$.

Proof :- Let, Z_0 be an arbitrary point of D. and let $|Z_0 - \propto| = p$ We new choose

Two positive numbers p_1 and p_2
Such that $r_2 < p_2 < p < p_1 < r_1$.
Let C_1: $|Z - \propto| = p_1$ and C_2: $|Z - \propto| = p_2$. Then f is analytic in the closed annular region bounded by C_1 and C_2.Also the point Z_0 lies inside the closed annular region $p_2 \leq |Z - \propto| \leq p_i$.
Hence, by Cauchy's integral formula for annular region we get,

$f(Z_0) = \frac{1}{2\pi i}\oint_{C_1} \frac{f(Z)}{Z-Z_0} dZ - \frac{1}{2\pi i}\oint_{C_2} \frac{f(Z)}{Z-Z_0} dZ$ - - - (1)

Now

$\frac{1}{2\pi i}\oint_{C_1} \frac{f(Z)}{Z-\propto} dZ = \frac{1}{2\pi i}\oint_{C_1} \frac{f(Z)}{(Z-\propto)\ (Z_0-\propto)} dZ$

$= \frac{1}{2\pi i}\oint_{C_1} \frac{f(Z)}{(Z-\propto)\ (1-\frac{Z_0-\propto}{Z-\propto})} dZ$

$= \frac{1}{2\pi i}\oint_{C_1} \frac{f(Z)(1-t^n+t^n)}{(Z-\propto)\ (1-t)} dZ$, putting $t = \frac{Z_0-\propto}{Z-\propto}$

$= \frac{1}{2\pi i}\oint_{C_1} \frac{f(Z)}{(Z-\propto)} .(1 + t^2 + - - - + t^{n-1}) dZ + \frac{1}{2\pi i}\oint_{C1} \frac{f(Z)}{(Z-\propto)} . \frac{t^n}{1-t} . dZ$

$= \sum_{k=0}^{n-1} \frac{1}{2\pi i}\oint_{C_1} \frac{f(Z)}{Z-\propto} . t^k\, dZ + \frac{1}{2\pi i}\oint_{C_1} \frac{f(Z)}{Z-\propto} . \frac{t^n}{1-t} . dZ$

$= \sum_{k=0}^{n-1} (\frac{1}{2\pi i}\oint_{C_1} \frac{f(Z)}{(Z-\propto)^{k+1}} . (Z_0-\propto)^k + \frac{1}{2\pi i}\oint_{C_1} \frac{f(Z)}{(Z-\propto)^n (Z-Z_0)} . dZ$

$= \sum_{k=0}^{n-1} a_k \, (Z_0 - \propto)^k + \; R_n$ ----- (2)

Where, $a_k = \frac{1}{2\pi i} = \oint_{C_1} \frac{f(Z)}{(Z-\propto)^{k+1}} dZ$

and $R_n = \frac{1}{2\pi i} \oint_{C_1} \frac{f(Z).(Z_0-\propto)^n}{(Z-\propto)^n (Z-Z_0)} dZ$

Since, f is analytic on C_1, there exists a `+ve' number M such that $|f(Z)| \leq M \; \forall \; Z \epsilon \; C_1$.

Also$\forall \; Z \epsilon \; C_1$ we get

$|Z - Z_0| = |(Z-\propto)+(\propto - Z_0)|$

$\geq |Z-\propto| - |Z_0 - \propto|$

$= p_1 - p.$

Hence, on C_1 we have

$|f(Z) \frac{(Z_0-\propto)^n}{(Z-\propto)^n (Z-Z_0)}| \leq \frac{M}{p_1 - p} . \left(\frac{p}{p_1}\right)^n$

Therefore by ML formula, we get

$|R_n| \leq \frac{1}{2\pi} \frac{M}{p_1 - P} \left(\frac{p}{p_1}\right)^n x 2\pi p_1 \rightarrow 0$

as $n \rightarrow \propto$ $[\because p < p_1]$

$$\therefore \lim_{n \to \propto} R_n = 0$$

Therefore considering limit as $n \rightarrow \infty$, we get from (2)

$\frac{1}{2\pi i} \oint_{C_1} \frac{f(Z)}{(Z-Z_0)} dZ = \sum_{n=0}^{\propto} a_n (Z_0 - \propto)^n.$

Next

- $\frac{1}{2\pi i} \oint_{C_2} \frac{f(Z)}{Z-Z_0} dZ = \frac{1}{2\pi i} \oint_{C_2} \frac{f(Z)}{Z_0 - Z} dZ$

$= \frac{1}{2\pi i} \oint_{C_2} \frac{f(Z)}{(Z_0-\propto)-(Z-\propto)} dZ$

$= \frac{1}{2\pi i} \oint_{C_2} \frac{f(Z)}{(Z_0-\propto)-(1-s)} dZ$ where $s = \frac{Z-\propto}{(Z_0-\propto)}$

$= \frac{1}{2\pi i}\oint_{C_2} \frac{f(Z)}{Z_0-\propto}.\frac{1-s^n+s^n}{1-s}dZ$

$=\frac{1}{2\pi i}\oint_{C_2} (\frac{f(Z)}{Z_0-\propto}\sum_{k=0}^{n-1} s^k)dZ + \frac{1}{2\pi i}\oint \frac{f(Z)}{Z_0-\propto}.\frac{s^n}{1-s}dZ$

$=\sum_{k=0}^{n-1} (\frac{1}{2\pi i}\oint_{C_2} \frac{f(Z)}{(Z-\propto)^{-k}}dZ)\,(Z-\propto)^{-(k+1)} + \frac{1}{2\pi i}\oint_{C_2} \frac{f(Z)(Z-\propto)^n}{(Z-\propto)^n(Z_0-Z)}dZ$

$=\sum_{m=1}^{n} (\frac{1}{2\pi i}\oint_{C_2} \frac{f(Z)}{(Z-\propto)^{-m+1}}dZ)\,(Z_0-\propto)^{-m} + R'_n$ - - - (3),

where k+1 we put k+1 =m

and $R'_n = \frac{l}{2\pi i}\oint_{C_2} \frac{f(Z)(Z-\propto)^n}{(Z_0-\propto)^n(Z_0-Z)}dZ$

Since f is analytic on C_2, it is bounded there. So there exists a positive number N such that,

$|f(Z)| \leq N \;\forall\; Z \epsilon\, C_2$.

Also, for, $Z \epsilon C_2$ we have

$|Z - \propto| = p_2$

Since $|Z - \propto| = p$ we get for all $Z \epsilon C_2$

$|Z_0 - Z| \geq |Z_0 - \propto| - |Z - \propto| = p - p_2$

Therefore for all $Z \epsilon C_2$ we have

$|\frac{f(Z)(Z-\propto)^n}{(Z_0-\propto)^n(Z_0-Z)}| \leq \frac{N}{p-p_2}.\left(\frac{p_2}{p}\right)^n$

Hence by ML formula, we obtained

$|R'_n| \leq \frac{l}{2\pi}.\frac{N}{p-p_2}.\left(\frac{p_2}{p}\right)^n.2\pi p_2$

$\rightarrow 0$ as $n \rightarrow \propto$ [$\because p_2 < p$]

i, e, $\lim_{n-\propto} R'_n = 0$

Therefore, considering limit as $n \rightarrow \infty$, we get from (3)

- $\frac{l}{2\pi i}\oint_{C_2} \frac{f(Z)}{(Z-Z_0)}dZ = \sum_{n=1}^{\infty} b_n\,(Z_0-\infty)^n$, where $b_n = \frac{1}{2\pi i}\oint_{C_2} \frac{f(Z)}{(Z_0-\propto)^{-n+1}}dZ$

Now from (1), we get

$f(Z_0) = \sum_{n=0}^{\infty} a_n\,(Z_0-\propto)^n + \sum_{n=1}^{\propto} b_n\,(Z_0-\propto)^{-n}$.

Since, Z_0 is a_n arbitrary point of D, we have

$f(Z) = \sum_{n=0}^{\infty} a_n (Z-\propto)^n + \sum_{n=1}^{\propto} b_n (Z-\propto)^{-n}$

where, $a_n = \frac{l}{2\pi i} \oint_{C_1} \frac{f(Z)}{(Z-\propto)^{n+1}} dZ$,

$b_n = \frac{1}{2\pi i} \oint_{C_2} \frac{f(Z)}{(Z-\propto)^{-n+1}} dZ$

and C_1: $|Z - \propto| = p_1$, C_2 : $|Z - \propto| = p_2$,

$r_2 < p_2 < |Z - \propto| = p_1 < r_1$.

We note that, the functions

$\frac{f(Z)}{(Z-\propto)^{n+1}}$ and $\frac{f(Z)}{(Z-\propto)^{-n+1}}$ are

analytic in D. So, the values of a_n and b_n do not depend on the circles C_1 and C_2. Infact, if we consider a circle C lying in D with $\propto$ as its centre, then

$a_n = \frac{l}{2\pi i} \oint_{C} \frac{f(Z)}{(Z-\propto)^{n+1}} dZ$,

and $b_n = \frac{l}{2\pi i} \oint_{C} \frac{f(Z)}{(Z-\propto)^{-n+1}} dZ$.

So, $b_n = a_{-n}$ for n=1,2, - - -

and the above series can also be written as

$f(Z) = \sum_{n=0}^{\infty} a_n (Z_0-\propto)^n$.

This proves the theorem.

Note :- The series $f(Z) = \sum_{n=0}^{\infty} a_n (Z-\propto)^n + \sum_{n=1}^{\propto} b_n (Z-\propto)^{-n}$

is called the Laurent's series for the function f(Z).

Example :- Expand $f(Z) = \frac{1}{(Z+1)(Z+3)}$ in a Laurent's series.

valid for (i) $|Z| < 1$, (ii) $1 < |Z| < 3$, (iii) $|Z| > 3$ (iv) $0 < |Z+1| < 2$.

Sol^n – (i) when $|Z| < 1$ we get

$f(Z) = \frac{1}{(Z+1)(Z+3)}$

$= \frac{1}{2}\left(\frac{1}{Z+1} - \frac{1}{Z+3}\right)$

$= \frac{1}{2}(1+Z)^{-1} - \frac{1}{6}\left(1+\frac{Z}{3}\right)$

$= \frac{1}{2}(1+Z+Z^2 - \ldots.) - \frac{1}{6}\left(1-\frac{Z}{3}+\frac{Z^2}{9} - - - -\right)$

$= \frac{1}{3} - \frac{4}{9}Z + \frac{13}{27}Z^2 - - - - -,$

which is the Laurent's expansion of f in $|Z| < 1$.

(ii) When $1<|Z| < 3$, we get

$f(Z) = \frac{1}{(Z+1)(Z+3)} = \frac{1}{2}\left(\frac{1}{Z+1} - \frac{1}{Z+3}\right)$

$= \frac{1}{2Z}\left(1+\frac{1}{Z}\right)^{-1} - \frac{1}{6}\left(1+\frac{Z}{3}\right)^{-1}$

$= \frac{1}{2Z}\left(1-\frac{1}{Z}+\frac{1}{Z^2} - - -\right) - \frac{1}{6}\left(1-\frac{Z}{3}+\frac{Z^2}{9} - - -\right),$

which is Laurent's expansion of f in $1<|Z| < 3$.

(iii) When $|Z| > 3$, we get ,

$f(Z) = \frac{1}{(Z+1)(Z+3)} = \frac{1}{2}\left(\frac{1}{Z+1} - \frac{1}{Z+3}\right)$

$= \frac{1}{2Z}\left(1+\frac{1}{Z}\right)^{-1} - \frac{1}{2Z}\left(1+\frac{3}{Z}\right)^{-1}$

$= \frac{1}{2Z}\left(1-\frac{1}{Z}+\frac{1}{Z^2} - - -\right) - \frac{1}{2Z}\left(1-\frac{3}{Z}+\frac{9}{Z^2} - - -\right),$

which is Laurent's expansion of f in the region $|Z| > 3$.

(iv) When $0<|Z+1| <2$, we get,

$f(Z) = \frac{1}{(Z+1)(Z+3)} = \frac{1}{2}\left(\frac{1}{Z+1} - \frac{1}{Z+3}\right)$

$= \frac{1}{2}\cdot\frac{1}{Z+1} - \frac{1}{2(Z+1+2)}$

$= \frac{1}{2}\cdot\frac{1}{Z+1} - \frac{1}{4}\left(1+\frac{Z+1}{2}\right)^{-1}$

$= \frac{1}{2}\cdot\frac{1}{Z+1} - \frac{1}{4}\left(1-\frac{Z+1}{2}+\frac{(Z+1)^2}{4} - - -\right),$

which is Laurent's expansion of f in the region $0<|Z+1| <2$.

Example: Show that $\cosh\left(Z+\frac{1}{Z}\right) = a_0 + \sum_{n=1}^{\infty} a_n \left(Z^n + \frac{1}{Z^n}\right)$, where

$a_n = \frac{l}{2\pi}\int_0^{2\pi} cosn\theta \cosh(2\,cos\theta)d\theta$.

Sol^{n}:$\rightarrow$ The function f(Z) = cosh $(Z + \frac{1}{Z})$ is analytic every where except at the origin. Theorefore we can expand f in a Laurent series around the arigin and get.

f(Z) = cosh $(Z + \frac{1}{Z})$ = $\sum_{n=0}^{\infty} a_n Z^n + \sum_{n=1}^{\infty} b_n \cdot \frac{1}{Z^n}$, where

$a_n = \frac{1}{2\pi i} \oint_{|Z|=1} \frac{f(Z)}{Z^{n+1}} dZ$, for n=0,1,2, ----

and b_n = a_{-n} for n= 1,2,3, - - -

Now, $a_n = \frac{1}{2\pi i} \int_0^{2\pi} \frac{Cosh(e^{i\theta}+e^{-i\theta})}{e^{(n+1)i\theta}} i.e^{i\theta}.d\theta$ [putting $Z = e^{i\theta}$ $0 \leq \theta \leq 2\pi$]

= $\frac{1}{2\pi} \int_0^{2\pi} \cosh(2\cos\theta)\,(cosn\,\theta - isinn\theta)d\theta$

= $\frac{1}{2\pi} \int_0^{2\pi} \cosh(2\cos\theta)\, cosn\,\theta.d\theta$ - $\frac{l}{2\pi} \int_0^{2\pi} \cosh(2\cos\theta)\, sin\, n\theta.d\theta$ - - - - (1)

Putting $\theta = 2\pi - \phi$, we get,

$\int_0^{2\pi} \cosh(2\cos\theta)\, sin\, n\theta.d\theta = \int_{2\pi}^{0} \cosh(2\cos\phi)\, sin\, n\phi.d\phi$

$= -\int_0^{2\pi} \cosh(2\cos\theta)\, sin\, n\theta.d\theta$

and so, $\int_0^{2\pi} \cosh(2\cos\theta)\, sin\, n\theta.d\theta = 0$

Therefore, from (1) we get,

$a_n = \frac{1}{2\pi} \int_0^{2\pi} \cosh(2\cos\theta)\, cos\, n\,\theta.d\theta$

Also,

$b_n = a_{-n}$

= $\frac{l}{2\pi} \int_0^{2\pi} \cosh(2\cos\theta)\, cos\,(-n)\,\theta.d\theta$

= $\frac{1}{2\pi} \int_0^{2\pi} \cosh(2\cos\theta)\, cos\, n\,\theta.d\theta$

Hence, cosh $(Z + \frac{1}{Z})$ = $a_0 + \sum_{n=1}^{\infty} a_n (Z^n + \frac{1}{Z^n})$ in $0<|Z|<\infty$.

Block III
Functional Analysis I

Unit 16

Course Structure

1. Metric Spaces
2. Continuity, Completeness, Compactness (in brief)
3. Holder's and Minkowski's Inequalities

1.1 Some basic concepts

Definition 1.1.Let X be any non empty set and let d be any real valued function defined on $X \times X$ such that for all $x, y, z \in X$, we have

i) $0 \leq d(x, y) < +\infty$ and $d(x, y) = 0$ if and only if $x = y$

ii) $d(x, y) = d(y, x)$ (Symmetry)

iii) $d(x, y) \leq d(x, z) + d(z, y)$ (triangular inequality)

Then d is called a metric or distance function on X, and the set X together with the metric d, written (X, d), is called a metric space.

Example 1.1.Given any set X, let us define

$$d_1(x, y) = \begin{cases} 0 & if\ x = y \text{ where } x, y \in X \\ 1 & if\ x \neq y \end{cases}$$

Then d_1 is a metric on X and it is called the trivial metric or discrete metric.

Example1.2.$d(x, y) = |x - y|\ \forall x, y \in \Phi = \mathbb{R}$ or $\mathbb{C}$ is called usual metric.

Definition1.2.A metric d on a set X induces a unique metric d_0 on a subset $X_0 \subset X$ defined by

$$d_0(x, y) = d(x, y)\ \forall x, y \in X_0.$$

The metric space (X_0, d_0) is called the subspace of the metric space (X, d).

Lemma1.1.In any metric space (x, d), $d(x, y) \geq |d(x, z) - d(y, z)|$.

Proof. Since

$$d(x, z) \leq d(x, y) + d(y, z),$$

$$\text{so} d(x, y) \geq d(x, z) - d(y, z) \qquad \ldots (1)$$

Also $d(x, y) = d(y, x)$ (by symmetry)

$$\geq d(y, z) - d(x, z) \qquad \ldots (2)$$

Combining (1) and (2), we get

$d(x, y) \geq |d(x, z) - d(y, z)|$.

Definition1.3. A sequence $\{x_n\}$ in a metric space (X, d) is said to be Cauchy or fundamental if $d(x_m, x_n) \to 0$as $m, n \to \infty$, i.e., for every $\varepsilon > 0$, there is a positive integer N such that $d(x_m, x_n) < \epsilon$ for all $m, n \geq N$.

The metric space (X, d) is said to be complete or d-complete if and only if every Cauchy Sequence in X is convergent in X.

Definition1.4.Let E be a subset of a metric space X. A family of open sets $\{G_\alpha\}$ in X is said to be an open cover of the set Eif $E \subset \bigcup_\alpha G_\alpha$.

Definition1.5. The subset E is said to be compact if every open cover of E has a finite sub cover. i.e., whenever $\{G_\alpha\}$ is a family of open sets in X with $E \subset \bigcup_\alpha G_\alpha$, then there is a finite subfamily $\{G_{\alpha_1}, G_{\alpha_2}, \ldots, G_{\alpha_n}\} \subset \{G_\alpha\}$ such that $E \subset G_{\alpha_1} \cup G_{\alpha_2} \cup \ldots \cup G_{\alpha_n}$.

Definition1.6. A subset E of a metric space X is said to be sequentially compact if every sequence in E has a convergent subsequence whose limit belongs to E.

Definition1.7.A subset E of a metric space X is said to be countably compact if every infinite subset of E has a limit point in E.

Definition1.8.Let (X, d) and (Y, ρ) be any two metric spaces. A function $f : (X, d) \to (Y, \rho)$ is said to be continuous at a point $c \in X$ if for every $\varepsilon > 0$ there is a $\delta > 0$ such that for all $x \in X$ with $d(c, x) < \delta$, we have $\rho(f(c), f(x)) < \varepsilon$.

The function f is said to be continuous (or continuous on X) if it is continuous at every point of X.

1.2 Cantor's Intersection Theorem

Let (X, d) be a complete metric space and let $\{F_n\}$ be a decreasing sequence of non empty closed subsets of X such that $d(F_n) \to 0$. Then$\bigcap_{n=1}^{\infty} F_n$ consists of exactly one point.

Proof. For each n we select a point x_n from the non empty set F_n.

Since $d(F_n) \to 0$, given $\varepsilon > 0$, there is a positive integer N such that $d(F_n) < \varepsilon$.

Now for all $m, n \geq N$, we have $x_m \in F_m \subset F_N$ and $x_n \in F_n \subset F_N$. So $d(x_m, x_n) \leq F_N < \varepsilon$.

Thus $\{x_n\}$ is a Cauchy Sequence in X.

Since X is complete, there is a point $x \in X$ such that $x_n \to x$.

We shall now show that $\bigcap_{n=1}^{\infty} F_n = \{x\}$.

For any positive integer n and for all k = 1, 2, …, we have $x_{n+k} \in F_{n+k} \subset F_N$. So, $\{x_{n+k}\}_{k=1}^{\infty}$ is a sequence in F_n, and we have $\lim_{k\to\infty} x_{n+k} = \lim_{p\to\infty} x_p = x$. Since the set F_n is closed, it follows that $x \in F_n$ for n = 1, 2, … . Thus $x \in \bigcap_{n=1}^{\infty} F_n$.

On the other hand if $\acute{x} \in \bigcap_{n=1}^{\infty} F_n$, then $0 \le d(x, \acute{x}) \le d(F_N)$ (since $x, \acute{x} \in F_n$)

$\to 0$ as $n \to \infty$.

So, $d(x,\acute{x}) = 0$. Hence by definition of a metric, $\acute{x} = x$.

Hence $\bigcap_{n=1}^{\infty} F_n = \{x\}$.

1.3. Some Standard Inequalities

Let $p > 1$ and $q > 1$ be any two real numbers such that $\frac{1}{p} + \frac{1}{q} = 1$, then p and q are called conjugate exponents.

1.3.1 Holder's Inequality: Let $\{x_1, x_2, \dots, x_k\}$ and $\{y_1, y_2, \dots, y_k\}$ be any two sets of k complex numbers. Then for any two conjugate exponents p and q, we have

$\sum_{j=1}^{k} |x_j y_j| \le \left(\sum_{j=1}^{k} |x_j|^p\right)^{\frac{1}{p}} \left(\sum_{j=1}^{k} |y_j|^q\right)^{\frac{1}{q}}$.

Note 1. Since 2, 2 are conjugate exponents, so in particular

$\sum_{j=1}^{k} |x_j y_j| \le \left(\sum_{j=1}^{k} |x_j|^2\right)^{\frac{1}{2}} \left(\sum_{j=1}^{k} |y_j|^2\right)^{\frac{1}{2}}$. This is Cauchy Schwarz's Inequality.

Note 2. The inequality extends to sequences of complex numbers $\{x_n\}_{n=1}^{\infty}$ and $\{y_n\}_{n=1}^{\infty}$ also.

i.e., $\sum_{j=1}^{\infty} |x_j y_j| \le \left(\sum_{j=1}^{\infty} |x_j|^2\right)^{\frac{1}{2}} \left(\sum_{j=1}^{\infty} |y_j|^2\right)^{\frac{1}{2}}$.

1.3.2 Minkowski's Inequality: For any two sets of complex numbers $\{x_1, x_2, \dots, x_k\}$ and $\{y_1, y_2, \dots, y_k\}$ and for any real number $p \ge 1$, we have

$\left(\sum_{j=1}^{k} |x_j + y_j|^p\right)^{\frac{1}{p}} \le \left(\sum_{j=1}^{k} |x_j|^p\right)^{\frac{1}{p}} + \left(\sum_{j=1}^{k} |y_j|^p\right)^{\frac{1}{p}}$.

Note 3. As like Holder's Inequality, Minkowski's Inequality can also extends to two sequences $\{x_n\}_{n=1}^{\infty}$ and $\{y_n\}_{n=1}^{\infty}$ of complex numbers. i.e.,

$\left(\sum_{j=1}^{\infty} |x_j + y_j|^p\right)^{p} \le \left(\sum_{j=1}^{\infty} |x_j|^p\right)^{\frac{1}{p}} + \left(\sum_{j=1}^{\infty} |y_j|^p\right)^{\frac{1}{p}}$.

Unit 17

Course Structure

1. Baire's Category Theorem
2. Banach Fixed Point Theorem

2.1 Definition.

A subset E of a metric space X is said to be

1) dense in a subset $A \subset X$ if $A \subset \bar{E}$, i.e., if each point of $A \backslash E$ is a limit point of E.

2) nowhere dense (or non dense) in X if E is not dense in any open ball in X.

3) of the first category (or meager (very little)) in X if we can write $E = \bigcup_{n=1}^{\infty} E_n$, where each E_n is nowhere dense in X.

4) of the second category in X if E is not of the first category in X.

2.2 Lemma.

Let B_0 be an open ball in a metric space (X, d). Let E be a subset of X which is not dense in B_0. Then for every $\varepsilon > 0$ there is an open ball B in X such that $\bar{B} \subset B_0$, $\bar{B} \cap E = \emptyset$ and $d(\bar{B}) < \epsilon$.

Proof. Since E is not dense in B_0, there is a point $x_0 \in B_0 \backslash \bar{E}$. So, x_0 is not a limit point of E. Then there is an open ball $B(x_0; r_1)$ such that $B(x_0; r_1) \cap E = \emptyset$.

Since $x_0 \in B_0$ and B_0 is open, there is an open ball $B(x_0; r_2) \subset B_0$.

Let $r = \min\left\{\frac{\epsilon}{3}, \frac{r_1}{3}, \frac{r_2}{3}\right\}$. We take the open ball $B = B(x_0; r)$. Clearly, $\bar{B} \subset B(x_0; r_1) \cap B(x_0; r_2)$.

Therefore, $\bar{B} \subset B_0$, $\bar{B} \cap E = \emptyset$ [since $B(x_0; r_1) \cap B(x_0; r_2) \subset B(x_0; r_2) \subset B_0$]

and $d(\bar{B}) \leq d(B(x_0, r]) \leq 2r \leq \frac{2}{3}\varepsilon < \varepsilon$.

2.3 Baire's Category Theorem.

A non empty complete metric space is of second category in itself.

Proof. Suppose for a contradiction that (X, d) is a complete metric space which is of first category in itself. Then X can be written as $X = \bigcup_{n=1}^{\infty} E_n$, where each E_n is nowhere dense in X.

Now, since $X \neq \emptyset$, so there is an open ball B_0 in X.

Since E_1 is not dense in B_0, so by the previous lemma, there is an open ball B_1 such that

$\bar{B}_1 \subset B_0, \bar{B}_1 \cap E_1 = \emptyset$ and $d(\bar{B}_1) < \frac{1}{1}$.

Since E_2 is not dense in B_1, again there is an open ball B_2 such that $\bar{B}_2 \subset B_1, \bar{B}_2 \cap E_2 = \emptyset$ and $d(\bar{B}_2) < \frac{1}{2}$.

Proceeding thus, we obtain a sequence of open balls $\{B_n\}_{n=1}^{\infty}$ such that

(i) $\bar{B}_1 \supset \bar{B}_2 \supset \bar{B}_3 \supset \cdots$

(ii) $\bar{B}_n \cap E_n = \emptyset$ for all n

(iii) $d(\bar{B}_n) < \frac{1}{n}$ for all n.

Since the metric space (X, d) is complete, so the conditions (i) and (iii) imply by Cantor's intersection theorem that there is a unique point $x \in \bigcap_{n=1}^{\infty} \bar{B}_n$.

Now by (ii), we have then $x \notin E_n$ for all n.

So, $x \notin \bigcup_{n=1}^{\infty} E_n = X$.

This is a contradiction.

Thus X is not of first category. Hence X is of second category in itself.

2.2 $\mathbb{R}^k, \mathbb{C}^k$ Spaces

Let $\Phi = \mathbb{R}$ or $\Phi = \mathbb{C}$. Then Φ^k denotes for each positive integer k the set of all ordered k-tuples $x = (x_1, x_2, \ldots, x_k)$ where $x_j \in \Phi$.

For $x, y \in \Phi^k$, we define $d(x, y) = \left(\sum_{j=1}^{k} |x_j - y_j|^2\right)^{\frac{1}{2}}$.

Clearly, $0 \leq d(x, y) < +\infty$ and $d(x, y) = 0$ if and only if $x_j = y_j$ for $j = 1, 2, \ldots, k$.

i.e., if and only if $x = y$.

It is also clear that $d(x, y) = d(y, x)$.

Finally, by Minkowski's Inequality, for all $x, y, z \in \Phi^k$, we have

$$d(x, z) + d(z, y) = \left(\sum_{j=1}^{k} |x_j - z_j|^2\right)^{\frac{1}{2}} + \left(\sum_{j=1}^{k} |z_j - y_j|^2\right)^{\frac{1}{2}} \geq \left(\sum_{j=1}^{k} |(x_j - z_j) + (z_j - y_j)|^2\right)^{\frac{1}{2}}$$
$$= d(x, y)$$

Thus d is a metric on Φ^k. This is called the Euclidean Metric on Φ^k.

If $\Phi = \mathbb{R}$, then $(\mathbb{R}^k, d)$ is called the k – dimensional Euclidean Space or the Euclidean k – Space.

If $\Phi = \mathbb{C}$, then $(\mathbb{C}^k, d)$ is called the k – dimensional Complex Euclidean Space or the k – dimensional unitary space.

Sometimes another metric for Φ^k becomes useful. This metric is given by

$$\rho(x,y) = \max_{1\le j\le k} |x_j - y_j|.$$

This metric ρ is equivalent to the above metric *d*.

2.3. The Function Space C[a, b]

Let $a, b \in \mathbb{R}$, $a < b$. Then $C[a, b]$ denotes the set of all real valued continuous functions on $[a, b]$.

For $f, g \in C[a, b]$, we define

$$d(f, g) = \sup_{a\le t\le b} |f(t) - g(t)|.$$

Since $f - g$ is continuous in the closed interval $[a, b]$, so it is bounded on $[a, b]$.

Therefore, $0 \le d(f, g) \le +\infty$.

Also $d(f, g) = 0$ if and only if $f(t) = g(t)$ for all $t \in [a, b]$. i.e., if and only if $f = g$.

It is also clear that $d(g, f) = d(f, g)$.

Finally, for all $f, g, h \in C[a, b]$, we have for all $t \in [a, b]$

$|f(t) - g(t)| \le |f(t) - h(t)| + |h(t) - g(t)| \le d(f,h) + d(h,g)$.

Thus $|f(t) - g(t)| \le d(f,h) + d(h,g)$ for $a \le t \le b$.

$\therefore d(f, g) = \sup_{a\le t\le b} |f(t) - g(t)| \le d(f,h) + d(h,g)$.

Hence d is a metric on $C[a, b]$.

This metric is called sup metric or Tchebycheff's Metric.

2.4 Theorem

The space $C[a, b]$ with the sup metric is complete; and convergence in $C[a, b]$ is equivalent to uniformly convergence of functions on $[a, b]$.

Note. Because of this result, the sup metric on $C[a, b]$ is also called the uniform metric.

Proof. Let $\{f_n\}$ be a Cauchy Sequence in the metric space $C[a, b]$. Then given $\varepsilon > 0$ there is a positive integer N such that

$d(f_m, f_n) = \sup_{a\le t\le b} |f_m(t) - f_n(t)| \le \varepsilon$ for all $m, n \ge N$.

Consequently, we have

(1) $|f_m(t) - f_n(t)| \le \varepsilon$ for all $m, n \ge N$; $a \le t \le b$.

This is the Cauchy's Criterion for uniform convergence of the sequence $\{f_n\}_{n=1}^{\infty}$ on $[a, b]$.

Since each $f_n \in C[a, b]$, so f_n is continuous on $[a, b]$.

Therefore, the sequence $\{f_n\}_{n=1}^{\infty}$converges uniformly to a continuous function, f, say, on $[a , b]$.

Then $f \in C[a, b]$.

Thus $lim_{n\to\infty} f_n(t) = f(t)$ for all $t \in [a, b]$ where $f \in C[a, b]$.

Now keeping n and t fixed in (1) and taking the $limit m \to \infty$ we get

(2) $|f(t) - f_n(t)| \leq \varepsilon$ for $n \geq N$ and $a \leq t \leq b$.

Consequently, we have

(3) $d(f, f_n) = \sup_{a\leq t\leq b} |f(t) - f_n(t)| \leq \varepsilon$ for all $n \geq N$ and $a \leq t \leq b$.

Hence $d(f, f_n) \to 0$.

$\therefore$ by definition of convergence in a metric space, we have $f_n \to f$ in $C[a, b]$, where $f \in C[a, b]$.

Hence the space $C[a, b]$ is complete.

Let $\{f_n\}$ be any sequence in $C[a, b]$ which converges to $f \in C[a, b]$.

Then $d(f, f_n) \to 0$.

Hence for every $\varepsilon > 0$ there is a positive integer N for which (3) is true. But (3) implies (2). Then (2) implies by definition that the sequence $\{f_n\}$ converges uniformly to the function f on $[a, b]$.

Conversely, let $\{f_n\}$ be any sequence of continuous functions on $[a, b]$ that converges uniformly to the function f on $[a, b]$. Then f is continuous on $[a, b]$. Then $\{f_n\}$ is a sequence in $C[a, b]$ and $f \in C[a, b]$.

By uniform convergence, for every $\varepsilon > 0$ there is a positive integer N for which (2) is true.

But (2) implies (3). Then by definition, (3) gives that $d(f, f_n) \to 0$.

Hence $\{f_n\}$ converges to f in $C[a, b]$.

Therefore, we conclude that convergence in $C[a, b]$ is equivalent to uniformly convergence of sequence of continuous functions on $[a, b]$.

Note. It can be shown that the space $C[a, b]$ is separable also.

2.5 The Sequence Space $\ell_p (1 \leq p \leq \infty)$

Let $\Phi = \mathbb{R}$ or $\Phi = \mathbb{C}$. Let ℓ_p denotes the set of all sequences $x = \{x^{(k)}\}_{k=1}^{\infty}$ with all $x^{(k)} \in \Phi$ such that $\sum_{k=1}^{\infty} |x^{(k)}|^p < +\infty$.

In ℓ_p we define $d(x,y) = (\sum_{k=1}^{\infty} |x^{(k)} - y^{(k)}|^p)^{\frac{1}{p}}$, $x, y \in \ell_p$.

By Minkowski's Inequality, we have

$\left(\sum_{k=1}^{\infty}\left|x^{(k)} - y^{(k)}\right|^p\right)^{\frac{1}{p}} \leq \left(\sum_{k=1}^{\infty}\left|x^{(k)}\right|^p\right)^{\frac{1}{p}} + \left(\sum_{k=1}^{\infty}\left|y^{(k)}\right|^p\right)^{\frac{1}{p}} < +\infty$, since $x, y \in \ell_p$.

$\therefore 0 \leq d(x,y) < +\infty$.

Clearly, $d(x, y) = 0$ if and only if $x^{(k)} = y^{(k)}$ for all $k = 1, 2, \ldots$.

i.e., if and only if $x = y$.

It is also clear that $d(y, x) = d(x, y)$.

Finally, by Minkowski's Inequality, again we have for all $x, y, z \in \ell_p$,

$$d(x,z) + d(z,y) = \left(\sum_{k=1}^{\infty}\left|x^{(k)} - z^{(k)}\right|^p\right)^{\frac{1}{p}} + \left(\sum_{k=1}^{\infty}\left|z^{(k)} - y^{(k)}\right|^p\right)^{\frac{1}{p}}$$

$$\geq \left(\sum_{k=1}^{\infty}\left|x^{(k)} - z^{(k)} + z^{(k)} - y^{(k)}\right|^p\right)^{\frac{1}{p}}$$

$$= \left(\sum_{k=1}^{\infty}\left|x^{(k)} - y^{(k)}\right|^p\right)^{\frac{1}{p}} = d(x,y).$$

Hence d is a metric on ℓ_p.

By the metric space ℓ_p we mean the space (ℓ_p, d).

2.6 The Space ℓ_p is Complete

Let $\{x_n\}$ be any Cauchy Sequence in ℓ_p, $x_n = \left\{x_n^{(k)}\right\}_{k=1}^{\infty}$. Then for every $\varepsilon > 0$ there is a positive integer N such that

(1) $d(x_m, x_n) = \left(\sum_{k=1}^{\infty}\left|x_m^{(k)} - x_n^{(k)}\right|^p\right)^{\frac{1}{p}} < \varepsilon$ for all $m, n \geq N$.

Then for each k, we have

$\left(\left|x_m^{(k)} - x_n^{(k)}\right|^p\right)^{\frac{1}{p}} < \varepsilon$ for all $m, n \geq N$.

So the sequence $\left\{x_n^{(k)}\right\}_{n=1}^{\infty}$ is a Cauchy Sequence of numbers. Hence this sequence is convergent.

We put

(2) $\lim_{n\to\infty} x_n^{(k)} = \xi^{(k)}$, $k = 1, 2, \ldots$

Now by (1), for each positive integer j, we have

$\left(\sum_{k=1}^{j}\left|x_m^{(k)} - x_n^{(k)}\right|^p\right)^{\frac{1}{p}} < \varepsilon$ for all *m, n*≥*N*.

Letting $m \to \infty$ in this relation and using (2), we get

$\left(\sum_{k=1}^{j}\left|\xi^{(k)} - x_n^{(k)}\right|^p\right)^{\frac{1}{p}} < \varepsilon$ for all *n*≥*N*.

Letting $j \to \infty$ in this relation, we have

(3) $\left(\sum_{k=1}^{\infty}\left|\xi^{(k)} - x_n^{(k)}\right|^p\right)^{\frac{1}{p}} < \varepsilon$ for all *n*≥*N*.

In particular, for $n = N$, from (3) we get

$\sum_{k=1}^{\infty}\left|\xi^{(k)} - x_N^{(k)}\right|^p \leq \varepsilon^p$.

Hence by Minkowski's Inequality, we have

$$\left(\sum_{k=1}^{\infty}\left|\xi^{(k)}\right|^p\right)^{\frac{1}{p}} = \left(\sum_{k=1}^{\infty}\left|\left(\xi^{(k)} - x_N^{(k)}\right) + x_N^{(k)}\right|^p\right)^{\frac{1}{p}}$$

$$\leq \left(\sum_{k=1}^{\infty}\left|\xi^{(k)} - x_N^{(k)}\right|^p\right)^{\frac{1}{p}} + \left(\sum_{k=1}^{\infty}\left|x_N^{(k)}\right|^p\right)^{\frac{1}{p}} < +\infty.$$

$\therefore \xi = \left\{\xi^{(k)}\right\}_{k=1}^{\infty} \in \ell_p$.

Then (3) shows that $d(\xi, x_n) \leq \varepsilon$ for all $n \geq N$.

Hence by definition $x_n \to \xi$ in ℓ_p.

Hence the space ℓ_p is complete.

Note. It can be shown that the space ℓ_p is separable.

2.7 Definition

A mapping *T: X* →*X* is called a self mapping and a point $x_0 \in X$ is called a fixed point of *T* if $Tx_0 = x_0$.

A mapping *T:* (*X, d*) →(*X, d*) is said to satisfy Lipschitz Condition with Lipschitz Constant α if $d(Tx, Ty) \leq \alpha\, d(x, y)$ for all *x, y*∈*X*.

If the constant be such that 0 ≤α< 1 and *T* satisfy Lipschitz Condition, then *T* is called a contraction on (*X, d*).

2.8Theorem

Let (X, d) be a metric space and let $T: X \to X$ be a contraction. Then T is uniformly continuous on X.

Proof. Since T is a contraction on (X, d), there is a number α, $0 \leq \alpha < 1$, such that

$d(Tx, Ty) \leq \alpha\, d(x, y)$ for all $x, y \in X$.

Now given $\varepsilon > 0$ let us take $\delta = \frac{\varepsilon}{1+\alpha}$.

Then for all $x, y \in X$ with $d(x, y) < \delta$, we have

$d(Tx, Ty) \leq \alpha\, d(x, y) < \alpha\delta = \alpha \frac{\varepsilon}{1+\alpha} < \varepsilon$.

Hence T is uniformly continuous on X.

2.9 Banach's Fixed Point Theorem (*or Contraction Principle*)

Let (X, d) be a complete metric space, where $X \neq \phi$. Then every contraction mapping $T: (X, d) \to (X, d)$ has a unique fixed point.

Proof. The proof is by iteration.

Since T is a contraction on (X, d), there is a constant α, $0 \leq \alpha < 1$, such that

(1) $d(Tx, Ty) \leq \alpha\, d(x, y)$ for all $x, y \in X$.

Starting with an arbitrary point $x_0 \in X$, by a recursion on n, we define the sequence $\{x_n\}_{n=1}^{\infty}$ in X as follows:

(2) $x_1 = Tx_0,\ x_2 = Tx_1 = T^2x_0,\ x_3 = Tx_2 = T^3x_0,\ \ldots$

We show that $\{x_n\}_{n=1}^{\infty}$ is a Cauchy Sequence.

By (1) and (2) for each n, we have

$$
\begin{aligned}
d(x_n, x_{n+1}) &= d(Tx_{n-1}, Tx_n) \\
&\leq \alpha\, d(x_{n-1}, x_n) = \alpha\, d(Tx_{n-2}, Tx_{n-1}) \\
&\leq \alpha.\alpha\, d(x_{n-2}, x_{n-1}) \\
&\vdots \\
&\leq \alpha^{n-1}\, d(x_1, x_2) = \alpha^{n-1}\, d(Tx_0, Tx_1) \\
&\leq \alpha^n\, d(x_0, x_1)
\end{aligned}
$$

$\therefore$ for any two positive integers m, n $(> m)$, we have

$$
\begin{aligned}
d(x_m, x_n) &\leq d(x_m, x_{m+1}) + d(x_{m+1}, x_n) \\
&\leq d(x_m, x_{m+1}) + d(x_{m+1}, x_{m+2}) + d(x_{m+2}, x_n) \\
&\vdots
\end{aligned}
$$

$$\leq d(x_m, x_{m+1}) + d(x_{m+1}, x_{m+2}) + \cdots + d(x_{n-1}, x_n)$$

$$\therefore d(x_m, x_n) \leq \alpha^m d(x_0, x_1) + \alpha^{m+1} d(x_0, x_1) + \cdots + \alpha^{n-1} d(x_0, x_1)$$

$$= (\alpha^m + \alpha^{m+1} + \cdots + \alpha^{n-1}) d(x_0, x_1)$$

$$< \frac{\alpha^m}{1-\alpha} d(x_0, x_1).$$

Since $0 \leq \alpha < 1$, so $\alpha^m \to 0$.

Hence it follows that $d(x_m, x_n) \to 0$ as $m, n \to \infty$.

So $\{x_n\}_{n=1}^{\infty}$ is a Cauchy Sequence in (X, d).

Since (X, d) is complete, so there is a point $x \in X$ such that $x_n \to x$ in (X, d).

We now show that T has the unique fixed point x.

We have

$$d(x, Tx) \leq d(x, Tx_n) + d(Tx_n, Tx)$$

$$\leq d(x, x_{n+1}) + \alpha d(x_n, x)$$

$$\to 0 \text{ as } n \to \infty.$$

Hence $d(x, Tx) = 0$.

$\therefore Tx = x$.

Thus x is fixed point of T.

Suppose $\acute{x} \in X$ is another fixed point of T.

$\therefore T\acute{x} = \acute{x}$.

So, $d(x, \acute{x}) = d(Tx, T\acute{x}) \leq \alpha d(x, \acute{x})$.

Hence $0 \leq (1-\alpha) d(x, \acute{x}) \leq 0 \Rightarrow (1-\alpha) d(x, \acute{x}) = 0$.

But $1 - \alpha \neq 0$. Therefore $d(x, \acute{x}) = 0$. Thus $\acute{x} = x$.

Hence T has the unique fixed point x.

Unit 18

Course Structure

1. Applications to solutions of certain systems of linear algebraic equations
2. Implicit Function theorem
3. Kannan's Fixed Point Theorem

3.1 Theorem

A system of *n* linear equations in *n* unknowns

$x_i = \sum_{j=1}^{n} a_{ij}x_j + b_i$ i = 1,2, .., n

where all a_{ij} , b_i are given real numbers such that $\sum_{j=1}^{n} |a_{ij}| < 1$, i = 1,2, .., n

has a unique solution for $(x_1, x_2, \dots, x_n)$.

Proof. Clearly, the problem is to find a fixed point of the mapping $T: \mathbb{R}^n \to \mathbb{R}^n$ defined for $x = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ by $Tx = \acute{x}$,where $\acute{x} = (\acute{x_1}, \acute{x_2}, \dots, \acute{x_n})$ is given by

$\acute{x_\iota} = \sum_{j=1}^{n} a_{ij}x_j + b_i$ i = 1,2, .., n.

We know that $\mathbb{R}^n$ is a complete metric space with the metric ρ defined by

$\rho(x, y) = \max_{1 \le i \le n} |x_j - y_j|$.

Now for $Tx = \acute{x}$ and $Ty = \acute{y}$, we have

$\rho(Tx, Ty) = \rho(\acute{x}, \acute{y}) = \max_{1 \le i \le n} |\acute{x}_i - \acute{y}_i|$.

$= \max_{1 \le i \le n} | \sum_{j=1}^{n} a_{ij}x_j + b_i - (\sum_{j=1}^{n} a_{ij}y_j + b_i) |$

$= \max_{i} | \sum_{j=1}^{n} a_{ij}(x_j - y_j) |$

$\le \max_{i} \sum_{j=1}^{n} |a_{ij}| | x_j - y_j|$

$\le \left(\max_{i} \sum_{j=1}^{n} |a_{ij}| \right) \rho(x, y)$

$= \alpha. \rho(x, y)$, where $\alpha = \max_{i} \sum_{j=1}^{n} |a_{ij}| < 1$.

Thus T is a contraction on the complete metric space $\mathbb{R}^n$. Hence by Banach's Fixed Point Theorem, T has a unique fixed point $x = (x_1, x_2, \dots, x_n)$ which is the required solution.

3.2 Theorem

Let $v: [a,b] \to \mathbb{R}$ and $k: [a,b] \times [a,b] \to \mathbb{R}$ be two continuous functions and let μ be a real number such that $|\mu| < \frac{1}{K(b-a)}, K > 0$, where K is an upper bound of the functions $|k|$. Then the Fredholm Integral Equation of second kind with kernel k,

$$x(t) = \mu \int_a^b k(t,s).x(s)\, ds + v(t)$$

Has a unique continuous solution x: $[a, b] \to \mathbb{R}$.

Proof. We know that the class of continuous functions $C[a, b]$ is a complete metric space under the *sup* metric defined by

$d(x,y) = \sup_{a \le t \le b} |x(t) - y(t)|, x, y \in C[a,b]$.

Clearly, then the problem is to find a fixed point of the mapping $T: C[a,b] \to C[a,b]$ defined by $Tx = \bar{x}$, where $\bar{x}(t) = \mu \int_a^b k(t,s).x(s)\, ds + v(t); a \le t \le b$.

Since the functions v, k, x are continuous, so $Tx = \bar{x} \in C[a,b]$.

Thus T is a well defined mapping.

Now for all $x, y \in C[a, b]$, we have

$$\begin{aligned} d(Tx, Ty) = d(\bar{x}, \bar{y}) &= \sup_{a \le t \le b} |\bar{x}(t) - \bar{y}(t)| \\ &= \sup_{a \le t \le b} |\mu \int_a^b k(t,s).\big(x(s) - y(s)\big) ds\,| \\ &\le |\mu| \sup_{a \le t \le b} \int_a^b |k(t,s)|.|x(s) - y(s)| ds \\ &\le |\mu| \sup_{a \le t \le b} \int_a^b K.d(x,y)\, ds \\ &= |\mu|.K.d(x,y)(b-a) \\ &= \alpha.d(x,y) \end{aligned}$$

where $\alpha = |\mu|.K(b-a) < 1$.

Therefore, T is a contraction on $C[a, b]$.

Hence by Banach's Fixed Point Theorem, T has a unique fixed point, which is the required solution.

3.3 Theorem (*Implicit Function Theorem*)

Let $f:[a,b]\times\mathbb{R}\to\mathbb{R}$ be a continuous function such that $\frac{\partial f}{\partial y}$ exists and satisfies the bounding condition $0<m\le\frac{\partial f}{\partial y}(x,y)\le M$, for all $(x,\ y)\ \in[a,b]\times\mathbb{R}$, then there is a unique continuous function $y=\varphi x$ on $[a,\ b]$ such that $f(x,\ \varphi(x))=0$, for all $x\in[a,b]$.

Proof. Clearly, the problem is to find a fixed point of the mapping $T:C[a,b]\to C[a,b]$ defined by $T\varphi=\bar{\varphi}$, where $\bar{\varphi}(x)=\varphi(x)-\frac{1}{M}f(x,\varphi(x))$, $a\le x\le b$; $\varphi\in C[a,b]$.

Since $\varphi\in C[a,b]$ is continuous on $[a,\ b]$ and f is continuous on $[a,\ b]$, so $\bar{\varphi}$ is well defined continuous function in $C[a,\ b]$.

Thus the mapping T is well defined.

Now for all $\varphi,\psi\ \in C[a,b]$, we have

$$d(T\varphi,T\psi)=d(\bar{\varphi},\bar{\psi})=\sup_{a\le x\le b}|\bar{\varphi}(x)-\bar{\psi}(x)|$$

$$=\sup_{a\le x\le b}|\varphi(x)-\psi(x)-\frac{1}{M}(f(x,\varphi(x)-f(x,\psi(x))|$$

$$=\sup_{a\le x\le b}|\varphi(x)-\psi(x)-\frac{1}{M}(\varphi(x)-\psi(x))\frac{\partial f}{\partial y}(x,\xi(x))|$$

[by MVT, where $\xi(x)\in(\varphi(x),\psi(x))$]

$$=\sup_{a\le x\le b}|\varphi(x)-\psi(x)|\,.\,|\frac{M-\frac{\partial f}{\partial y}(x,\xi(x))}{M}|$$

$\le d(\varphi,\psi).\frac{M-m}{M}$ (by hypothesis)

$$=\alpha.d(\varphi,\psi)$$

where $\alpha=\frac{M-m}{M}=1-\frac{m}{M}<1$.

$\therefore$ T is a contraction. Hence by Banach's Fixed Point Theorem, T has a unique fixed point, which is the required solution.

3.4 Theorem (*Kannan's Fixed Point Theorem*)

Let T: $(X,\ d)\to(X,\ d)$ be a self mapping and X be a complete metric space where T satisfies the condition $d(Tx,Ty)\le\beta[d(x,Tx)+d(y,Ty)]$ with $0<\beta<\frac{1}{2}$ and $x,\ y\in X$. Then T has a unique fixed point.

Proof. Let $x_0\in X$ be an arbitrary point and we define

$$x_1=Tx_0,\qquad x_2=Tx_1,\qquad x_3=Tx_2,\dots,x_n=Tx_{n-1},\dots$$

Then by the given condition,

$$d(x_1, x_2) = d(Tx_0, Tx_1)$$
$$\leq \beta\,[d(x_0, Tx_0) + d(x_1, Tx_1)]$$
$$= \beta\,[d(x_0, Tx_0) + d(x_1, x_2)]$$

$\therefore d(x_1, x_2) \leq \frac{\beta}{1-\beta} d(x_0, Tx_0)$.

And so $d(x_2, x_3) \leq \frac{\beta}{1-\beta} d(x_1, Tx_1)$

$$= \frac{\beta}{1-\beta} d(x_1, x_2)$$

$\leq \left(\frac{\beta}{1-\beta}\right)^2 d(x_0, Tx_0)$.

Similarly, $d(x_3, x_4) \leq \left(\frac{\beta}{1-\beta}\right)^3 d(x_0, Tx_0)$.

In general, for any positive integer *n*, we have

$$d(x_n, x_{n+1}) \leq \left(\frac{\beta}{1-\beta}\right)^n d(x_0, Tx_0)$$

Hence for any two positive integers *m*, *n* (> *m*), we have

$$d(x_m, x_n) \leq d(x_m, x_{m+1}) + d(x_{m+1}, x_n)$$
$$\leq d(x_m, x_{m+1}) + d(x_{m+1}, x_{m+2}) + d(x_{m+2}, x_n)$$
$$\vdots$$
$$\leq d(x_m, x_{m+1}) + d(x_{m+1}, x_{m+2}) + \cdots + d(x_{n-1}, x_n)$$
$$\therefore d(x_m, x_n) \leq \alpha^m d(x_0, Tx_0) + \alpha^{m+1} d(x_0, Tx_0) + \cdots + \alpha^{n-1} d(x_0, Tx_0)$$
$$= (\alpha^m + \alpha^{m+1} + \cdots + \alpha^{n-1})\, d(x_0, Tx_0)$$

Where $\alpha = \frac{\beta}{1-\beta}$. Since $0 < \beta < \frac{1}{2}$, $0 < \alpha < 1$.

Hence $d(x_m, x_n) < \frac{\alpha^m}{1-\alpha} d(x_0, Tx_0)$.

Letting $m \to \infty$, we get $d(x_m, x_n) \to 0$.

So the sequence$\{x_n\}_{n=1}^{\infty}$ is a Cauchy Sequence in (*X*, *d*).

Since (*X*, *d*) is complete, so there is a point $x \in X$ such that $x_n \to x$ in (*X*, *d*).

We now show that *T* has the unique fixed point *x*.

We have

$$d(x, Tx) \leq d(x, x_n) + d(x_n, Tx)$$

$\leq d(x, x_n) + d(Tx_{n-1}, Tx)$

$\leq d(x, x_n) + \beta[d(x_{n-1}, Tx_{n-1}) + d(x, Tx)]$

$\therefore d(x, Tx) \leq \frac{1}{1-\beta} d(x, x_n) + \frac{\beta}{1-\beta} d(x_{n-1}, x_n) \to 0$ as $n \to \infty$.

Hence $Tx = x$, which implies that x is fixed point of T.

Suppose $\acute{x} \in X$ is another fixed point of T.

$\therefore T\acute{x} = \acute{x}$.

So, $(x, \acute{x}) = d(Tx, T\acute{x}) \leq \beta[d(x, Tx) + d(T\acute{x}, \acute{x})] = 0$.

$\Longrightarrow \acute{x} = x$.

Hence T has the unique fixed point x.

Unit 19

Course Structure

1. Linear Spaces
2. Norm Induced Metric
3. Banach Spaces
4. Riesz's Lemma

4.1 Linear Space or Vector Space

Recall that any additive abelian group is any abstract set X equipped with a binary operation called addition which assigns to each pair $(x, y) \in X \times X$ a unique element $(x + y) \in X$ and satisfies the following conditions for all $x, y, z \in X$:

i) $x + y = y + x$

ii) $(x + y) + z = x + (y + z)$

iii) there is a unique element $0 \in X$, called the zero or null element or origin of X such that $x + 0 = 0 + x = x$

iv) for each $x \in X$ there is a unique element, denoted by $(-x) \in X$, called the negative of x such that $x + (-x) = (-x) + x = 0$.

Let now Φ denotes either the field of real numbers or the field of complex numbers.

A linear space or a vector space over the scalar field Φ is any abstract abelian group X equipped with a scalar multiplication $\Phi \times X \to X$, $(\alpha, x) \to \alpha x$, which satisfies the following conditions for all scalars $\alpha, \beta, 1 \in \Phi$ and all vectors $x, y \in X$:

i) $\alpha(x + y) = \alpha x + \alpha y$

ii) $(\alpha + \beta)x = \alpha x + \beta x$

iii) $(\alpha\beta)x = \alpha(\beta x)$

iv) $1x = x$.

The linear space X is called real or complex according as Φ is $\mathbb{R}$ or $\mathbb{C}$.

A mapping from a vector space X to its associated scalar field Φ is called a functional.

4.2 Theorem. Every linear space has the following properties:

(i) $0x = 0$ and $\alpha 0 = 0$

(ii) $-1x = -x$

(iii) $(\alpha - \beta)x = \alpha x - \beta x$

(iv) If $\alpha x = \alpha y$ and $\alpha \neq 0$, then $x = y$.

(v) If $\alpha x = \beta x$ and $x \neq 0$, then $\alpha = \beta$.

Let x be any linear space. A finite set of vectors $\{x_1, x_2, \dots, x_k\} \subset X$ is said to be linearly dependent if there exist scalar $\alpha_1, \alpha_2, \dots, \alpha_k$, not all zero such that

$\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_k x_k = 0.$

Otherwise the set $\{x_1, x_2, \dots, x_k\}$ is called linearly independent.

Thus the set $\{x_1, x_2, \dots, x_k\}$ is linearly independent if and only if the relation

$\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k = 0$ holds if and only if $\beta_1 = \beta_2 = \cdots = \beta_k = 0.$

An arbitrary subset $M \subset X$ is said to be linearly independent if every non empty finite subset of M is linearly independent.

The linear space X is said to have finite dimension k (a positive integer) if there is a linearly independent set of k vectors $\{e_1, e_2, \dots, e_k\} \subset X$ such that $x \in X$ can be expressed as a linear combination of the form $x = \alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_k e_k$ where $\alpha_1, \alpha_2, \dots, \alpha_k$ are suitable scalars depending on the vector x.

Such a set $\{e_1, e_2, \dots, e_k\}$ is then called a basis of the vector space X.

If $X = \{0\}$, then X has no basis but it is said to have finite dimension zero.

A subset $Y \subset X$ is called a linear subspace or a linear manifold of X if Y itself is a linear space under the addition and scalar multiplication induced by those in X.

Then a subset $Y \subset X$ is a linear subspace of X if and only if $0 \in Y$ and $\alpha x + \beta y \in Y$ whenever x, $y \in Y$ and $\alpha, \beta \in \Phi$.

A subspace $Y \subset X$ is called proper subspace if $Y \neq X$.

4.3 Theorem. If X is a linear space of finite dimension n and Y is a proper subspace of X then Y has some finite dimension $m < n$.

4.4 Example. $\mathbb{R}^k$ is a real linear space of dimension *k*. $\mathbb{C}^k$ is a complex linear space of dimension *k*.

Ex. What is the dimension of $\mathbb{C}$ considering it as a real vector space?

4.5 Definition. Let *X* be a linear space over the field Φ ($\mathbb{R}$ or $\mathbb{C}$). A real valued function defined on *X*, $\| \cdot \|$, $x \to \|x\|$, is called a norm on *X*, if for all $x, y \in X$ and all scalars $\alpha \in \Phi$, the following conditions are satisfied:

i) $0 \le \|x\| < +\infty$ and $\|x\| = 0$ if and only if $x = 0$.

ii) $\|\alpha x\| = |\alpha| \, \|x\|$ (absolute homogeneity)

iii) $\|x + y\| \le \|x\| + \|y\|$ (triangular inequality)

The linear space *X* equipped with a norm $\| \cdot \|$, written $(X, \| \cdot \|)$, is called a normed linear space.

Let $(X, \| \cdot \|)$ is a normed linear space. We define

$d(x, y) = \|x - y\|$ for all $x, y \in X$.

Then $0 \le d(x, y) < +\infty$ and $d(x, y) = 0$ if and only if $x - y = 0$ i.e., if and only if $x = y$.

Also $d(y, x) = \|y - x\| = \|-1(x - y)\| = |-1| \, \|x - y\| = 1.d(x, y) = d(x, y)$.

Finally, $d(x, y) = \|x - y\| = \|x - z + z - y\|$

$\le \|x - z\| + \|z - y\| = d(x, z) + d(z, y)$.

Hence *d* is a metric on *X*. This is called the metric induced by the norm on *X*.

Whenever we shall consider a normed linear space $(X, \| \cdot \|)$, we shall always suppose that *X* is a metric space under the metric *d* induced by the norm as defined above.

Thus the notions of convergence, open set, closed set, etc. in *X* are always with respect to the induced metric. In particular, $x_n \to x$ in *X* is equivalent to $d(x, x_n) = \|x - x_n\| \to 0$.

4.6 Definition. A normed linear space $(X, \| \cdot \|)$ which is complete with respect to the induced metric defined by $d(x, y) = \|x - y\|$ is called a Banach Space.

4.7 Theorem. Every normed linear space *X* has the following properties:

i) $\|-x\| = \|x\|$

ii) $\|x \pm y\| \ge | \, \|x\| - \|y\| \, |$

iii) Addition in *X* is continuous, i.e., $x_n \to x$ and $y_n \to y$ in *X* then $x_n + y_n \to x + y$.

iv) Scalar multiplication in X is continuous, i.e., if $x_n \to x$ in X and $\alpha_n \to \alpha$ in Φ, then $\alpha_n x_n \to \alpha x$ in X.

v) If $\{x_n\}$ and $\{y_n\}$ are Cauchy Sequences in X and $\{\alpha_n\}$ is a Cauchy Sequences in Φ, then both $\{x_n + y_n\}$ and $\{\alpha_n x_n\}$ are Cauchy sequences in X.

vi) The norm function in X is continuous, i.e., if $x_n \to x$ in X, then $||x_n|| \to ||x||$ in $\mathbb{R}$.

Proof.

(i) We have $||-x|| = ||(-1)x|| = |-1|\,||x|| = 1.\,||x|| = ||x||$.

(ii) We have $||x|| = ||(x \pm y) \mp y||$

$\leq ||(x \pm y)|| + ||\mp y||$

$= ||(x \pm y)|| + ||y||$

$\therefore\ ||x|| - ||y|| \leq ||(x \pm y)||$ …(1)

Interchanging x and y, we get

$||y|| - ||x|| \leq ||(y \pm x)|| = ||\pm(x \pm y)|| = ||(x \pm y)||$ …(2)

Since $|\,||x|| - ||y||\,| = ||x|| - ||y||$ or $-(||x|| - ||y||) = ||y|| - ||x||$, it follows from (1) and (2) that

$||x \pm y|| \geq |\,||x|| - ||y||\,|$

(iii) We have $0 \leq ||x_n + y_n - (x + y)|| = ||x_n - x + y_n - y||$

$\leq ||x_n - x|| + ||y_n - y|| \to 0.$

Hence $||x_n + y_n - (x + y)|| \to 0$.

So $x_n + y_n \to x + y$

(iv) We have $||\alpha_n x_n - \alpha x||$

$= ||\alpha_n(x_n - x) + (\alpha_n - \alpha)x||$

$\leq |\alpha_n|\,.\,||x_n - x|| + |\alpha_n - \alpha|\,.\,||x||$

$\to |\alpha|.0 + 0.||x|| = 0$

Hence $\alpha_n x_n \to \alpha x$.

(v) We have $||x_m + y_m - (x_n + y_n)|| = ||(x_m - x_n) + (y_m - y_n)||$

$\leq ||x_m - x_n|| + ||y_m - y_n||$

$\to 0$ as $m, n \to \infty$.

Hence $\{x_n + y_n\}$ is a Cauchy sequence.

Again $||\alpha_m x_m - \alpha_n x_n||$

$= ||\alpha_m(x_m - x_n) + (\alpha_m - \alpha_n)x_n||$

$\leq |\alpha_m|. \|x_m - x_n\| + |\alpha_m - \alpha_n| \, . \|x_n\|$

$\to 0$ as $m, n \to \infty$. [since the Cauchy Sequences $\{\alpha_n\}$ and $\{x_n\}$ are bounded]

Hence $\{\alpha_n x_n\}$ is a Cauchy sequence.

(vi) We have $| \|x_n\| - \|x\| | \leq \|x_n - x\| \to 0$.

Hence $\|x_n\| \to \|x\|$.

4.8. Theorem. Let $\Phi = \mathbb{R}$ or $\Phi = \mathbb{C}$ and let k be a positive integer. Then Φ^k is a Banach Space of finite dimension k under the Euclidean Norm.

Proof. The elements of Φ^k are ordered k – tuples $x = (x_1, x_2, \ldots, x_k)$, $x_i \in \Phi$.

In Φ^k we define addition and scalar multiplication by

$x + y = (x_1 + y_1, x_2 + y_2, \ldots, x_k + y_k)$ and $\alpha x = (\alpha x_1, \alpha x_2, \ldots, \alpha x_k)$, $\alpha \in \Phi$.

It is easy to verify that with these operations of addition and scalar multiplication, Φ^k is a vector space of dimension k over the field Φ.

We define

$\|x\| = (|x_1|^2 + |x_2|^2 + \ldots + |x_k|^2)^{\frac{1}{2}}$.

Then $0 \leq \|x\| < +\infty$. Also $\|x\| = 0$ if and only if $x_1 = x_2 = \ldots x_k = 0$.

i.e., if and only if $x = (0, 0, \ldots, 0) = 0 \in \Phi^k$.

Again for all $\alpha \in \Phi$ and $x \in \Phi^k$,

$\|\alpha x\| = \|(\alpha x_1, \alpha x_2, \ldots, \alpha x_k)\|$

$= (|\alpha x_1|^2 + |\alpha x_2|^2 + \cdots + |\alpha x_k|^2)^{\frac{1}{2}}$

$= (|\alpha|^2 |x_1|^2 + |\alpha|^2 |x_2|^2 + \cdots + |\alpha|^2 |x_k|^2)^{\frac{1}{2}}$

$= |\alpha| (|x_1|^2 + |x_2|^2 + \cdots + |x_k|^2)^{\frac{1}{2}}$

$= |\alpha| \, \|x\|$

Finally,

$\|x+y\| = \|(x_1 + y_1, x_2 + y_2, \ldots, x_k + y_k)\|$

$= (\sum_{j=1}^{k} |x_j + y_j|^2)^{\frac{1}{2}}$

$\leq (\sum_{j=1}^{k} |x_j|^2)^{\frac{1}{2}} + (\sum_{j=1}^{k} |y_j|^2)^{\frac{1}{2}}$ (by Minkowski's Inequality)

$= \|x\| + \|y\|$.

Hence || . || is a norm on Φ^k.

This norm is called the Euclidean norm on Φ^k.

The metric induced by this norm is given by

d(x, y) = ||x – y||

$= \|(x_1 - y_1, x_2 - y_2, \dots, x_k - y_k)\|$

$= (\sum_{j=1}^{k} |x_j - y_j|^2)^{\frac{1}{2}}$

This is Euclidean metric on Φ^k and we know that Φ^k is complete with respect to this metric induced by the norm. Hence with this definition, Φ^k becomes a Banach Space.

4.9. Theorem. The space $C[a, b]$ is an infinite dimensional Banach Space under the sup norm.

Proof. The elements of $C[a, b]$ are the real valued continuous functions on $[a, b]$. For $f, g \in C[a,b]$ and $\alpha \in \mathbb{R}$, wedefine $f + g$ and αf by

$(f + g)(x) = f(x) + g(x)$, $(\alpha f)(x) = \alpha f(x)$, $x \in [a, b]$, then $f + g \in C[a, b]$ and $\alpha f \in C[a, b]$.

It is easy to verify that $C[a, b]$ is a real vector space under these operations, addition and scalar multiplication. Also $C[a, b]$ has infinite dimension.

We define

$$||f|| = \sup_{a \le x \le b} |f(x)|, f \in C[a, b].$$

Since $f \in C[a, b]$ is continuous on the closed interval $[a, b]$, so f is bounded on $[a, b]$.

Hence $0 \le ||f|| \le +\infty$. Also $||f|| = 0$ iff $f(x) = 0$ for all $x \in [a, b]$. i.e., iff f is the constant function zero.

For each $\alpha \in \mathbb{R}$ and $f \in C[a, b]$, we have

$$|\alpha f(x)| = |\alpha| \ |f(x)| \le |\alpha| \ ||f||, x \in [a, b]$$

Hence

$\sup_{a \le x \le b} |\alpha f(x)| \le |\alpha| \ ||f||$; i.e., $||\alpha f|| \le |\alpha| \ ||f||$.

If $\alpha = 0$, the equality holds.

If $\alpha \neq 0$ then

$$||f|| = ||\alpha^{-1} \alpha f|| \le |\alpha^{-1}| \ ||\alpha f|| = \frac{1}{|\alpha|} ||\alpha f||$$

$$\therefore \ |\alpha| \ ||f|| \ \le ||\alpha f||$$

Combining both the inequalities we have $||\alpha f|| = |\alpha| \ ||f||$.

Finally,

$$|(f+g)(x)| = |f(x)+g(x)| \le |f(x)| + |g(x)| \le \big||f|\big| + \big||g|\big|, x \in [a,b]$$

$$\therefore \sup_{a\le x\le b} |(f+g)(x)| \le \big||f|\big| + \big||g|\big|$$

i.e., $||f+g|| \le \big||f|\big| + \big||g|\big|$.

Hence $||.||$ defines a norm in $C[a,b]$.

This is called the supnorm in $C[a,b]$.

The metric induced by this norm is given by

$$d(f,g) = \big||f-g|\big| = \sup_{a\le x\le b} |f(x)-g(x)|$$

This is the supmetric for $C[a,b]$ and we know that $C[a,b]$ is complete with respect to this metric. Hence $C[a,b]$ is an infinite dimensional Banach space under the supnorm.

4.10. Theorem. The space ℓ_p is a Banach Space.

Proof. The elements of ℓ_p are the sequences $x = \{x_k\}_{k=1}^{\infty},\ x_k \in \Phi$ ($\mathbb{R}$ *or* $\mathbb{C}$) with

$\sum_{k=1}^{\infty} |x_k|^p < +\infty$.

For $x, y \in \ell_p$ and $\alpha \in \Phi$, we define $x+y = \{x_k + y_k\}$ and $\alpha x = \{\alpha x_k\}$.

It is clear that $\alpha x \in \ell_p$. Also by Minkowski's inequality, we have

(1) $\left(\sum_{k=1}^{\infty} \left|x^{(k)} + y^{(k)}\right|^p\right)^{\frac{1}{p}} \le \left(\sum_{k=1}^{\infty} \left|x^{(k)}\right|^p\right)^{\frac{1}{p}} + \left(\sum_{k=1}^{\infty} \left|y^{(k)}\right|^p\right)^{\frac{1}{p}} < +\infty$, so, $x + y \in \ell_p$.

It is now easy to verify that under these operations of addition and scalar multiplication, ℓ_p is a vector space of infinite dimension over Φ.

We define

$\big||x|\big| = \left(\sum_{k=1}^{\infty} \left|x^{(k)}\right|^p\right)^{\frac{1}{p}}, x \in \ell_p$.

Then $0 \le ||x|| < +\infty$ and $||x|| = 0$ if and only if $x_k = 0$ for all k. i.e., $x = \{0, 0, \dots, 0\} = 0 \in \ell_p$.

For $x \in \ell_p$ and $\alpha \in \Phi$, we have $||\alpha x|| = \left(\sum_{k=1}^{\infty} \left|\alpha x^{(k)}\right|^p\right)^{\frac{1}{p}} = \left(\sum_{k=1}^{\infty} |\alpha|^p \left|x^{(k)}\right|^p\right)^{\frac{1}{p}} = |\alpha|\big||x|\big|$.

Finally, from (1) we have $\big||x+y|\big| \le \big||x|\big| + ||y||$.

Hence $||.||$ defines a norm in ℓ_p.

The metric induced by this norm is given by $d(x,y) = \big||x-y|\big| = \left(\sum_{k=1}^{\infty} \left|x^{(k)} - y^{(k)}\right|^p\right)^{\frac{1}{p}}$

and we know that this is a metric in ℓ_p with respect to which ℓ_p is complete. Hence ℓ_p is an infinite dimensional Banach space under the above norm.

4.11Example. Not every normed linear space is a Banach space.

We know that $C[a,b]$ is a Banach space under the supnorm.

Let $P[a,b] = \{f \in C[a,b] f \text{ is a polynomial}\}$

Then $P[a,b]$ is a normed linear subspace of $C[a,b]$.Now, let us define

$$P_n(t) = 1 + \frac{t}{1!} + \frac{t^2}{2!} + \cdots + \frac{t^n}{n!}, n = 1, 2, \ldots \text{ and } a \leq t \leq b$$

Then $\{P_n\}$ is a sequence in $P[a,b]$. So, $\{P_n\}$ is also a sequence in $C[a,b]$, and we know that this sequence converges uniformly to the function $f(t) = e^t$.

Since uniform convergence of continuous functions on $[a,b]$ is equivalent to convergence in the metric space $C[a,b]$, it follows that $P_n \to f$ in $C[a,b]$.

Therefore, $\{P_n\}$ is a sequence in $P[a,b]$ which convergence to $f \in C[a,b]$, but $f \notin P[a,b]$. Hence the normed linear space $P[a,b]$ is not complete. Thus $P[a,b]$ is a normed linear space but it is not a Banach space.

4.12Example. Not every complete linear space is normable.

Proof. Let us consider any linear space $X \neq \{0\}$ with the discrete metric

$$d\ (x,y) = \begin{cases} 0 & if\ x = y \\ 1 & if\ x \neq y \end{cases}$$

Let $\{x_n\}$ be any Cauchy sequence in X. Then for $\epsilon = \frac{1}{2}$ there is a positive integer N such that $d(x_m, x_n) < \frac{1}{2}$ for all $m, n \geq N$.

By definition of d we see that $x_m = x_n$ for all $m, n \geq N$.

In particular, $x_n = x_N$ for all $n \geq N$.

So, $d(x_n, x_N) = 0$ for all $n \geq N$.

Thus $d(x_n, x_N) \to 0$ as $n \to \infty$. i.e., $x_n \to x_N$ in X.

Hence the metric space (x,d) is complete.

We shall show however that the space (x,d) is not normable. i.e., the metric d does not come from any norm defined on X.

Suppose for a contradiction that $||.||$ is a norm on$X$ which induces the metric d. Then

$$d(x,y) = ||x-y|| = \begin{cases} 0 & if\ x = y \\ 1 & if\ x \neq y \end{cases}$$

Now since $X \neq \{0\}$, $x_0 \in$ X such that $x_0 \neq 0$.

Then $d(x_0, 0) = 1$, i.e., $\big|\,||x_0 - 0||\,\big| = 1$, i.e., $\big|\,||x_0||\,\big| = 1$.

Also, $\big|\,||2x_0||\,\big| = 2\big|\,||x_0||\,\big| = 2.1 = 2$

i.e., $\big|\,||2x_0 - 0||\,\big| = 2$, i.e., $d(2x_0, 0) = 2$, i.e., $1 = 2$.

This contradiction proves our assertion.

4.13Theorem. If $Y$is a linear subspace of a normed linear space X, then $\bar{Y}$ is also a linear subspace of X.

Proof. Since Y is a linear subspace of a linear space X, so $0 \in Y$, then $0 \in \bar{Y}$.

Let now $x, y \in \bar{Y}$ and α, β be scalars. Then there are sequences $\{x_n\}$ and $\{y_n\}$ in Y such that$x_n \to x$ and $y_n \to y$.

Consequently, $\{\alpha x_n + \beta y_n\}$ is a sequence in Y and $\alpha x_n + \beta y_n \to \alpha x + \beta y$ in Y.

Hence $\alpha x_n + \beta y_n \in \bar{Y}$.

$\therefore \bar{Y}$ is a linear subspace of X.

4.14Theorem.A normed linear space $X \neq \{0\}$ is a Banach space iff the unit sphere

$$S(0;1] = \{x \in X : \big|\,||x||\,\big| = 1\}$$

(called the surface of the closed unit ball) is complete.

Proof. First suppose that the normed linear space X is a Banach space, i.e., it is complete under the metric induced by the norm.

Now, $S(0;1] = B(0;1]\backslash B(0;1)$, where the closed ball $B(0;\ 1]$ is a closed set and the open ball $B(0;1)$ is an open set. Therefore, $S(0;1]$ is a closed set. Since, X is complete so its closed subset $S(0;1]$ is alsocomplete.

Conversely, assume that $S(0;1]$ is complete. Let $\{x_n\}$ be any Cauchy sequence in X. Then

$$\Big|\,||x_m|| - ||x_n||\,\Big| \leq \big|\,||x_m - x_n||\,\big| \to 0 \text{ as } m, n \to \infty.$$

$\therefore \{x_n\}$ is a Cauchy sequence of non-negative real numbers. Hence, there is a number $\alpha \geq 0$ such that $\big|\,||x_n||\,\big| \to \alpha$.

If $\alpha = 0$ then $\big|\,||x_n - 0||\,\big| = \big|\,||x_n||\,\big| \to \alpha = 0$.

Hence in this case $x_n \to 0$ in X.

Let now $\alpha > 0$. Then taking $\epsilon = \frac{1}{2}\alpha > 0$, there is a positive integer N such that

$$0 < \alpha - \epsilon = \tfrac{1}{2}\alpha < \big||x_n|\big| \text{ for all } n \geq N.$$

[$\Big|\big||x_n|\big| - \alpha\Big| < \varepsilon \Rightarrow \alpha - \epsilon < \big||x_n|\big| < \alpha + \epsilon$]

Then for all $m, n \geq N$ we have

$$\left|\left|\frac{1}{\big||x_m|\big|}.x_m - \frac{1}{\big||x_n|\big|}.x_n\right|\right| = \left|\left|\frac{1}{\big||x_m|\big|}(x_m - x_n) + \left(\frac{1}{\big||x_m|\big|} - \frac{1}{\big||x_n|\big|}\right).x_n\right|\right|$$

$$\leq \frac{1}{\big||x_m|\big|}\big||x_m - x_n|\big| + \frac{\Big|\big||x_n|\big| - \big||x_m|\big|\Big|}{\big||x_m|\big|.\big||x_n|\big|}\big||x_n|\big| \leq \frac{2}{\big||x_m|\big|}\big||x_m - x_n|\big|$$

$\leq \frac{4}{\alpha}\big||x_m - x_n|\big| \to 0$ as $m, n \to \infty$.

Hence $\left\{\frac{1}{\big||x_n|\big|}.x_n\right\}$ is a Cauchy sequence in $S(0;1]$, since $\left|\left|\frac{1}{\big||x_n|\big|}x_n\right|\right| = \frac{\big||x_n|\big|}{\big||x_n|\big|} = 1$.

Since $S(0;1]$ is assumed to be complete, i.e., there is $x \in S(0;1]$, i.e., $\big||x|\big| = 1$ such that $\frac{1}{\big||x_n|\big|}.x_n \to x$. i.e., $x_n = \big||x_n|\big|\left(\frac{x_n}{\big||x_n|\big|}\right) \to \alpha x \in X$.

Hence the normed linear space X is complete, i.e., it is a Banach space.

4.15Lemma. For every linearly independent set of vectors $\{e_1, e_2, \dots, e_k\}$in a normed linear space X, there exist two constants $\lambda > 0$ and $\mu > 0$ such that for all sets of scalars $\{\alpha_1, \alpha_2, \dots, \alpha_k\}$ we have $\lambda(|\alpha_1| + |\alpha_2| + \cdots + |\alpha_k|) \leq \big||\alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_k e_k|\big| \leq \mu(|\alpha_1| + |\alpha_2| + \cdots + |\alpha_k|)$.

*Proof.*Let Φ ($\mathbb{R}$ *or* $\mathbb{C}$) denote the scalar field of X. Then the set

$$E = \{(\beta_1, \beta_2, \dots, \beta_k) \in \Phi^k : |\beta_1| + |\beta_2| + \cdots + |\beta_k| = 1\}$$

is compact since it is closed and bounded.

We consider the function $f : E \to \mathbb{R}$ defined by

$$f(\beta_1, \beta_2, \dots, \beta_k) = \big||\beta_1 e_1 + \beta_2 e_2 + \cdots + \beta_k e_k|\big|, \qquad (\beta_1, \beta_2, \dots, \beta_k) \in E.$$

Since $|\beta_1| + |\beta_2| + \cdots + |\beta_k| = 1$, so $\beta_1, \beta_2, \dots, \beta_k$ cannot be all zero.

Again, since the set of vectors $\{e_1, e_2, \dots, e_k\}$is linearly independent, it follows that

$\beta_1 e_1 + \beta_2 e_2 + \cdots + \beta_k e_k \neq 0$. Therefore, $\big||\beta_1 e_1 + \beta_2 e_2 + \cdots + \beta_k e_k|\big| > 0$.

i.e., $f(\beta_1, \beta_2, \dots, \beta_k) > 0$ for all $(\beta_1, \beta_2, \dots, \beta_k) \in E$.

Since, scalar multiplication, addition of vectors and the norm function are continuous, we have f is also continuous.

Therefore, the continuous image $f(E)$ of the compact set E is compact in $\mathbb{R}$. So, $f(E)$ is closed and bounded.

Since $f > 0$ on E, it follows that

$$0 < \lambda = \inf f(E) \leq \sup f(E) = \mu < +\infty$$

Consider now any set of scalars $\{\alpha_1, \alpha_2, \dots, \alpha_k\} \subset \Phi$.

First, let $|\alpha_1| + |\alpha_2| + \cdots + |\alpha_k| = t \neq 0$. Then

$$\left|\frac{\alpha_1}{t}\right| + \left|\frac{\alpha_2}{t}\right| + \cdots + \left|\frac{\alpha_k}{t}\right| = \frac{|\alpha_1| + |\alpha_2| + \cdots + |\alpha_k|}{t} = \frac{t}{t} = 1.$$

So, $\left(\frac{\alpha_1}{t}, \frac{\alpha_2}{t}, \dots, \frac{\alpha_k}{t}\right) \in E$

Therefore, by definition of λ and μ, we have

$\lambda \leq f\left(\frac{\alpha_1}{t}, \frac{\alpha_2}{t}, \dots, \frac{\alpha_k}{t}\right) \leq \mu$, i.e., $\lambda \leq \left|\left|\frac{\alpha_1}{t}e_1 + \frac{\alpha_2}{t}e_2 + \cdots + \frac{\alpha_k}{t}e_k\right|\right| \leq \mu$

i.e., $\lambda \leq \frac{1}{t}\left||\alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_k e_k|\right| \leq \mu$

i.e., $\lambda(|\alpha_1| + |\alpha_2| + \cdots + |\alpha_k|) \leq \left||\alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_k e_k|\right| \leq \mu(|\alpha_1| + |\alpha_2| + \cdots + |\alpha_k|)$.

Since this is trivially true even when $t = |\alpha_1| + |\alpha_2| + \cdots + |\alpha_k| = 0$, the proof is complete.

4.16Theorem. Every finite dimensional subspace Y of a normed linear space X is complete and Y is closed in X. In particular, every finite dimensional normed linear space is a Banach space.

Proof. If $\dim Y = 0$, then $Y = \{0\}$, which is obviously complete and closed.

Let now $\dim Y = k \geq 1$. Then Y has a basis of $k$vectors $\{e_1, e_2, \dots, e_k\}$, say. Since $\{e_1, e_2, \dots, e_k\}$ is linearly independent, then by the previous Lemma there are two constants $\lambda > 0$ and $\mu > 0$ such that

(1) ... $\lambda(|\alpha_1| + |\alpha_2| + \cdots + |\alpha_k|) \leq \left||\alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_k e_k|\right| \leq \mu(|\alpha_1| + |\alpha_2| + \cdots + |\alpha_k|)$

for all $(\alpha_1, \alpha_2, \dots, \alpha_k) \in \Phi^k$, where Φ ($= \mathbb{R}$ or $\mathbb{C}$)is the associated scalar field of X.

Now let $\{y_n\}$ be any Cauchy sequence in Y. Then each y_n has a unique representation

$y_n = \alpha_1^{(n)} e_1 + \alpha_2^{(n)} e_2 + \cdots + \alpha_k^{(n)} e_k$ where $\left(\alpha_1^{(n)}, \alpha_2^{(n)}, \dots, \alpha_k^{(n)}\right) \in \Phi^k$.

For any two positive integers m and n by (1) we have

$$\sqrt{|\alpha_1^{(m)} - \alpha_1^{(n)}|^2 + |\alpha_2^{(m)} - \alpha_2^{(n)}|^2 + \cdots + |\alpha_k^{(m)} - \alpha_k^{(n)}|^2}$$

$$\leq |\alpha_1^{(m)} - \alpha_1^{(n)}| \quad + |\alpha_2^{(m)} - \alpha_2^{(n)}| \quad + \cdots + |\alpha_k^{(m)} - \alpha_k^{(n)}|$$

$$\leq \frac{1}{\lambda}\left|\left|\left(\alpha_1^{(m)} - \alpha_1^{(n)}\right)e_1 + \left(\alpha_2^{(m)} - \alpha_2^{(n)}\right)e_2 + \cdots + \left(\alpha_k^{(m)} - \alpha_k^{(n)}\right)e_k\right|\right|$$

$$= \frac{1}{\lambda}\left||y_m - y_n|\right| \to 0 \text{ as } m, n \to \infty$$

Hence $\left\{\left(\alpha_1^{(n)}, \alpha_2^{(n)}, \ldots, \alpha_k^{(n)}\right)\right\}_{n=1}^{\infty}$ is a Cauchy sequence in Φ^k. Since, Φ^k is complete, so there is a point $(\alpha_1, \alpha_2, \ldots, \alpha_k) \in \Phi^k$ such that $\left(\alpha_1^{(n)}, \alpha_2^{(n)}, \ldots, \alpha_k^{(n)}\right) \to (\alpha_1, \alpha_2, \ldots, \alpha_k)$as $n \to \infty$.

We now put $y = \alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_k e_k$, then $y \in Y$, and we have

$$\left||y - y_n|\right| = |\left|\alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_k e_k - \left(\alpha_1^{(n)} e_1 + \alpha_2^{(n)} e_2 + \cdots + \alpha_k^{(n)} e_k\right)\right||$$

$$= |\left|\left(\alpha_1 - \alpha_1^{(n)}\right)e_1 + \left(\alpha_2 - \alpha_2^{(n)}\right)e_2 + \cdots + \left(\alpha_k - \alpha_k^{(n)}\right)e_k\right||$$

$$\leq \mu\left(\left|\alpha_1 - \alpha_1^{(n)}\right| + \left|\alpha_2 - \alpha_2^{(n)}\right| + \cdots + \left|\alpha_k - \alpha_k^{(n)}\right|\right), \text{ by (1)}$$

$$\to 0 \text{ as } n \to \infty$$

Thus $\left||y - y_n|\right| \to 0$. Hence $y_n \to y$in Y.

Hence the subspace Y is complete, and so further Y is closed in X. (complete subspace of a metric space is closed)

Finally, if X itself is a finite dimensional normed linear space, then $X \subset X$ is a finite dimensional subspace of X. Hence by the above X is complete. i.e., X is a Banachspace.

4.17Theorem. Every closed and bounded subset of a normed linear space X of finite dimension is compact. The result may fail if X is not of finite dimension.

Proof. Let Y be a closed and bounded subset of a normed linear space with $\dim X$ finite.

If $\dim X = 0$, then $X = \{0\}$, so that the result is obvious.

Let now $\dim X = k \geq 1$. Then X has a basis of k vectors $\{e_1, e_2, \ldots, e_k\}$, say.

Since $\{e_1, e_2, \ldots, e_k\}$ is linearly independent, then by the previous Lemma there are two constants $\lambda > 0$ and $\mu > 0$ such that

(1) ... $\lambda(|\alpha_1| + |\alpha_2| + \cdots + |\alpha_k|) \leq \left||\alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_k e_k|\right| \leq \mu(|\alpha_1| + |\alpha_2| + \cdots + |\alpha_k|)$

for all $(\alpha_1, \alpha_2, \ldots, \alpha_k) \in \Phi^k$, where Φ $(= \mathbb{R}$ or $\mathbb{C})$ is the associated scalar field of X.

Since Y is bounded, there is a constant $M > 0$ such that

(2) ... $||y|| < M$ for all $y \in Y$.

Let us define a function $f: \Phi^k \to X$ by $f(\alpha_1, \alpha_2, \dots, \alpha_k) = \alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_k e_k$ for $(\alpha_1, \alpha_2, \dots, \alpha_k) \in \Phi^k$.

Since scalar multiplication and vector addition are both continuous, clearly, the function f is continuous.

Therefore, since Y is closed in X, so $f^{-1}(Y)$ is closed in Φ^k.

Also, if $(\alpha_1, \alpha_2, \dots, \alpha_k) \in f^{-1}(Y)$, then $f(\alpha_1, \alpha_2, \dots, \alpha_k) = \alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_k e_k \in Y$, so that by (2) we have $||\alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_k e_k|| < M$.

Then by (1) we have

$$\sqrt{|\alpha_1|^2 + |\alpha_2|^2 + \cdots + |\alpha_k|^2} \leq |\alpha_1| + |\alpha_2| + \cdots + |\alpha_k| \leq \frac{1}{\lambda}||\alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_k e_k|| < \frac{M}{\lambda}.$$

Thus $f^{-1}(Y)$ is bounded besides being closed in Φ^k. Hence $f^{-1}(Y)$ is compact in Φ^k.

Now, for each $y \in Y$, there are scalars $\alpha_1, \alpha_2, \dots, \alpha_k$ such that

$$y = \alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_k e_k = f(\alpha_1, \alpha_2, \dots, \alpha_k)$$

So f is onto Y. Hence $f\left(f^{-1}(Y)\right) = Y$.

Since $f^{-1}(Y)$ is compact in Φ^k, so its continuous image $f\left(f^{-1}(Y)\right)$ is continuous on X. i.e., Y is compact.

To show that the result may fail if $\dim X$ is not finite, we can consider the Banachspace l_2.

Let $e_1 = (1, 0, 0, \dots)$, $e_2 = (0, 1, 0, \dots)$, $e_3 = (0\ 0, 1, \dots)$, …

Then $\{e_n\}_{n=1}^{\infty}$ is a sequence in l_2.

We have $||e_n|| = \sqrt{|1|^2} = 1 \ \forall n$.

So the subset $Y = \{e_1, e_2, \dots\} \subset l_2$ is bounded.

Clearly, the set of vectors Y is linearly independent.

For $m \neq n$, $||e_m - e_n|| = \left(\sum_{k=1}^{\infty} |e_m{}^{(k)} - e_n{}^{(k)}|^2\right)^{\frac{1}{2}} = \sqrt{1^2 + (-1)^2} = \sqrt{2} > 0$.

Hence Y has no limit point in l_2, and so Y isclosed.

Thus Y is bounded and closed in l_2 which is of infinite dimension.

But Y is a countable subset of l_2 having no limit point. So, Y is not countably compact. Hence the metric space Y is not compact.

4.18Riesz's Lemma. Let Y be proper, closed subspace of a normed linear space X. Then for each real number α with $0 < \alpha < 1$, there is a vector $x_\alpha \in X\backslash Y$ such that $\big||x_\alpha|\big| = 1$ and $\big||x_\alpha - y|\big| > \alpha$ for all $y \in Y$.

Proof. Since Y is a proper subset of X, there is a vector $x_0 \in X\backslash Y$.

Since Y is closed in X and $x_0 \notin Y$, so $\delta = dist\,(x_0, Y) = \inf_{y\in Y}\big||x_0 - y|\big| > 0$.

Now, since $0 < \alpha < 1$, so $\delta < \frac{\delta}{\alpha}$.

Hence there is $y_0 \in Y$ such that $\delta < ||x_0 - y_0|| < \frac{\delta}{\alpha}$. We take

$$x_\alpha = \frac{1}{\big||x_0 - y_0|\big|}(x_0 - y_0)$$

Clearly, $x_\alpha \in X\backslash Y$ and $\big||x_\alpha|\big| = 1$.

Also, since Y is a subspace of X, so for all $y \in Y$, we have $y_0 + ||x_0 - y_0||y \in Y$, and hence

$$\delta \le \Big||x_0 - \big(y_0 + ||x_0 - y_0||y\big)|\Big| = \Big||(x_0 - y_0) - |x_0 - y_0||y|\Big|$$

$$= \Big|\big||x_0 - y_0|\big|x_\alpha - |x_0 - y_0||y\ \big|\Big| = \big||x_0 - y_0|\big|\big||x_\alpha - y|\big|$$

Hence $\big||x_\alpha - y|\big| \ge \frac{\delta}{||x_0 - y_0||} > \alpha$.

Thus $\big||x_\alpha - y|\big| > \alpha$ for all $y \in Y$.

This completes the proof.

Unit 20

Course Structure

1. Finite dimensional normed linear spaces
2. Equivalent norms.

4.19Theorem. If the unit sphere $S(0;1] = \{x \in X: ||x|| = 1\}$ in a normed linear space X is compact, then X is finite dimensional.

Proof. If $S = \emptyset$ then $X = \{0\}$, which is of finite dimension zero.

Suppose, $S \neq \emptyset$.Then the family of open balls $\left\{B\left(x;\frac{1}{2}\right)\right\}$ in X with centers on S is an open cover of the compact set S. So there is a finite number of open balls

$$B\left(x_1;\frac{1}{2}\right), B\left(x_2;\frac{1}{2}\right), \dots, B\left(x_n;\frac{1}{2}\right)$$

which together covers S. i.e., $S \subset \bigcup_{i=1}^{n} B\left(x_i;\frac{1}{2}\right)$.

Now, let Y be the linear subspace of X generated by the set of vectors $\{x_1, x_2, \dots, x_n\}$.

Then Y has some finite dimension $k \leq n$.

Hence by a known result Y is closed in X.

We assert that $Y = X$. If not, then by Riesz'z lemma, there exists vector $x_{\frac{1}{2}} \in X \backslash Y$ such that $\left\|\left\|x_{\frac{1}{2}}\right\|\right\| = 1$ and $\left\|\left\|x_{\frac{1}{2}} - y\right\|\right\| > \frac{1}{2}$ for all $y \in Y$.

Now since, $\left\|\left\|x_{\frac{1}{2}}\right\|\right\| = 1$, so $x_{\frac{1}{2}} \in S$.

Therefore, we have $x_{\frac{1}{2}} \in B\left(x_i;\frac{1}{2}\right)$ for some $i = 1, 2, \dots, n$.

Hence, $\left\|\left\|x_{\frac{1}{2}} - x_i\right\|\right\| < \frac{1}{2}$.

But since $x_i \in Y$, so by our choice of $x_{\frac{1}{2}}$, $\left\|\left\|x_{\frac{1}{2}} - x_i\right\|\right\| > \frac{1}{2}$.

Thus we arrive at a contradiction.

Hence we must have $Y = X$ and so X is of finite dimension k.

4.20Theorem. Every finite dimensional linear space X can be made to a Banach space.

Proof. If $\dim X = 0$, then $X = \{0\}$, and clearly, , $\big||0|\big| = 0$ defines a norm on X and so $(X, ||.||)$ is a Banach space.

Let now $\dim X = k \geq 1$. We select a basis $\{e_1, e_2, \dots, e_k\}$of X.

Then each $x \in X$ has a unique representation $x = \alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_k e_k$

We define $\big||x|\big| = |\alpha_1| + |\alpha_2| + \cdots + |\alpha_k|$.

Then $0 \leq ||x|| < +\infty$. Also $||x|| = 0$ if and only if $\alpha_1 = \alpha_2 = \cdots = \alpha_k$. i.e., if and only if x = 0.

Again for any scalar α, we have

$$\alpha x = \alpha(\alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_k e_k) = (\alpha\alpha_1)e_1 + (\alpha\alpha_2)e_2 + \cdots + (\alpha\alpha_k)e_k$$

and $\big||\alpha x|\big| = |\alpha\alpha_1| + |\alpha\alpha_2| + \cdots + |\alpha\alpha_k| = |\alpha|(|\alpha_1| + |\alpha_2| + \cdots + |\alpha_k|) = |\alpha|\ ||x||$.

Finally, if $y = \beta_1 e_1 + \beta_2 e_2 + \cdots + \beta_k e_k \in X$ then

$$||x + y|| = ||(\alpha_1 + \beta_1)e_1 + (\alpha_2 + \beta_2)e_1 + \cdots + (\alpha_k + \beta_k)e_k||$$

$$= |\alpha_1 + \beta_1| + |\alpha_2 + \beta_2| + \cdots + |\alpha_k + \beta_k|$$

$$\leq |\alpha_1| + |\beta_1| + |\alpha_2| + |\beta_2| + \cdots + |\alpha_k| + |\beta_k| = ||x|| + ||y||$$

Hence $||.||$ is in fact a norm on X.

Since, every finite dimensional normed linear space is a Banachspace, it follows that $(X, ||.||)$ is a Banach space.

4.21Theorem. If a sequence $\{x_n\}$ in a Banach space X is such that $\sum_{n=1}^{\infty}\big||x_n|\big| < +\infty$, then $\sum_{n=1}^{\infty} x_n$ converges in X.

Proof. Let $s_n = x_1 + x_2 + \dots + x_n \in X$

Then $\Big|\big|s_{n+p} - s_n\big|\Big| = \Big|\big|x_{n+1} + x_{n+2} + \cdots + x_{n+p}\big|\Big| \leq \big||x_{n+1}|\big| + \big||x_{n+2}|\big| + \cdots + + \Big|\big|x_{n+p}\big|\Big|$.

Since $\sum_{n=1}^{\infty}\big||x_n|\big| < +\infty$, so $\big||x_{n+1}|\big| + \big||x_{n+2}|\big| + \cdots + + \Big|\big|x_{n+p}\big|\Big| \to 0$ as $n, p \to \infty$ (by Cauchy's criterion of convergence)

Therefore, $\Big|\big|s_{n+p} - s_n\big|\Big| \to 0$as $n, p \to \infty$.

Hence $\{s_n\}$ is a Cauchy sequence in the Banach space X. Therefore, the sequence converges in X. i.e., the series $\sum_{n=1}^{\infty} x_n$ converges in X.

4.22Equivalent Norms

Two norms $||.||_1$and $||.||_2$ on a linear space X are said to be equivalent if there exists two constants $a > 0$ and $b > 0$ such that $a||x||_1 \leq ||x||_2 \leq b||x||_1$ for all $x \in X$.

4.23Theorem. Any two norms on a finite dimensional linear space $X$are equivalent.

Proof. If the linear space$X = \{0\}$ then there is only one norm on X given by ,$\big||0|\big| = 0$. So the result is trivially true in this case.

Let now $\dim X = k \geq 1$.

Let $||.||_1$and $||.||_2$ be two norms on X. We choose a basis $\{e_1, e_2, \dots, e_k\}$of X. Then we know that there are constants $\lambda_1,\ \mu_1,\ \lambda_2,\ \mu_2 > 0$ such that

$\lambda_1(|\alpha_1| + |\alpha_2| + \cdots + |\alpha_k|) \leq \big||\alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_k e_k|\big|_1 \leq \mu_1(|\alpha_1| + |\alpha_2| + \cdots + |\alpha_k|)$and

$\lambda_2(|\alpha_1| + |\alpha_2| + \cdots + |\alpha_k|) \leq \big||\alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_k e_k|\big|_2 \leq \mu_2(|\alpha_1| + |\alpha_2| + \cdots + |\alpha_k|)$

Now each $x \in X$ has a unique representation $x = \alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_k e_k$, where the scalars $\alpha_1, \alpha_2, \dots, \alpha_k$depends on x.

Then by the above

$\lambda_1(|\alpha_1| + |\alpha_2| + \cdots + |\alpha_k|) \leq \big||x|\big|_1 \leq \mu_1(|\alpha_1| + |\alpha_2| + \cdots + |\alpha_k|)$and

$$\lambda_2(|\alpha_1| + |\alpha_2| + \cdots + |\alpha_k|) \leq \big||x|\big|_2 \leq \mu_2(|\alpha_1| + |\alpha_2| + \cdots + |\alpha_k|)$$

$$\therefore \frac{\lambda_2}{\mu_1}\big||x|\big|_1 \leq \lambda_2(|\alpha_1| + |\alpha_2| + \cdots + |\alpha_k|) \leq \big||x|\big|_2 \leq \mu_2(|\alpha_1| + |\alpha_2| + \cdots + |\alpha_k|)$$

$$= \frac{\mu_2}{\lambda_1}\lambda_1(|\alpha_1| + |\alpha_2| + \cdots + |\alpha_k|) \leq \frac{\mu_2}{\lambda_1}\big||x|\big|_1$$

$\therefore a||x||_1 \leq ||x||_2 \leq b||x||_1$, where $a = \frac{\lambda_2}{\mu_1} > 0$and $b = \frac{\mu_2}{\lambda_1} > 0$.

Hence $||.||_1$and $||.||_2$ are equivalent.

4.24Definition. A subset $E$of a linear space X is said to be convex if for all $x, y \in E$ we have $tx + (1-t)y \in E$ for all $t \in [0,1]$.

4.25Theorem. The intersection of an arbitrary family of convex sets $\{E_i\}$in a linear space X is convex.

Proof. Let $x, y \in \bigcap_i E_i$ and let $t \in [0,1]$.

For each i, $E_i$is convex and $x, y \in E_i$.

So $tx + (1-t)y \in E_i$for each i.

Hence $\bigcap_i E_i$ is convex.

4.26Theorem. In any normed linear space X, we have

i) The closure $\bar{E}$ of a convex set $E \subset X$ is convex.

ii) All balls in X are convex.

Proof.

i) Let $x, y \in \bar{\bar{E}}$and let $t \in [0,1]$.Then there are sequences $\{x_n\}$ and $\{y_n\}$ in E such that $x_n \to x$and $y_n \to y$.

Since E is convex, so $tx_n + (1-t)y_n \in E$ for all n; and we have

$$tx_n + (1-t)y_n \to tx + (1-t)y$$

Hence $tx + (1-t)y \in \bar{\bar{E}}$.

Thus the set $\bar{\bar{E}}$ is convex.

ii) Consider any open ball $B(x_0; r)$in X.

Let $x, y \in B(x_0; r)$ and let $0 \leq t \leq 1$.

Then $\big||x - x_0|\big| < r$and $\big||y - x_0|\big| < r$.

So,

$$\big||tx + (1-t)y - x_0|\big| = \big||t(x - x_0) + (1-t)(y - x_0)|\big| \leq |t|\big||x - x_0|\big| + |1-t|\big||y - x_0|\big|$$
$$< t.r + (1-t)r = r$$

Hence $tx + (1-t)y \in B(x_0; r)$.

Hence the open ball $B(x_0; r)$ is convex.

Similarly, we can show that the closed balls $B(x_0; r]$ is also convex.

Hence the theorem.

References :

1. E. Kreyszig : Introductory Functional Analysis with Applications.

2. B. V. Limaye : Functional Analysis.

3. W. Rudin : Functional Analysis.

4. A. E. Taylor : Introduction to Functional Analysis.

5. P. K. Jain, Khalil Ahmad, Om P. Ahuja: Functional Analysis

POST GRADUATE DEGREE PROGRAMME (CBCS)

M.SC. IN MATHEMATICS

SEMESTER I

SELF LEARNING MATERIAL

PAPER: COR 1.2
(Pure and Applied Streams)

Ordinary Differential Equations

Partial Differential Equations

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

Course Preparation Team

Dr. Pulak Sahoo, Professor, Department of Mathematics, University of Kalyani

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

Dec 2021

Directorate of Open and Distance Learning, University of Kalyani

Published by the Directorate of Open and Distance Learning

University of Kalyani, 741235, West Bengal

COR 1.2

Marks: 100; Credits: 6

Unit	Topic	Counselling Duration
Block I: Ordinary Differential Equations; Marks 50 (SEE: 40; IA: 10)		
1	Existence of solutions**:** Picard's Existence theorem for equation dy / dx = f(x,y), Gronwall's lemma, Picard-Lindelöf method of successive approximations.	54 Mins
2	Solutions of linear differential equations of nth order. Wronskian, Abel's identity.	54 Mins
3	Linear dependence and independence of the solution set, Fundamental set of solutions.	54 Mins
4	Green's function for boundary value problem and solution of non-homogenous linear equations.	54 Mins
5	Adjoint and self-adjoint equations. Lagrange's identity.	54 Mins
6	Sturm's separation and comparison theorems for second order linear equations. Regular Sturm-Liouville problems for second order linear equations.	54 Mins
7	Eigen values and eigen functions, expansion in eigen functions.	54 Mins
8	Solution of linear ordinary differential equations of second order in complex domain.	54 Mins
9	Existence of solutions near an ordinary point and a regular singular point.	54 Mins
10	Solutions of Hyper geometric equation and Hermite equation, Introduction to special functions.	54 Mins
Block II: Partial Differential Equations; Marks 50 (SEE: 40; IA: 10)		
11	Introduction and pre-requisite, Genesis and types of solutions of Partial Differential Equations.	54 Mins

12	First order Partial Differential Equations, Classifications of First Order Partial Differential Equations. Charpit's Method for the solution of First Order non-linear Partial Differential Equation.	54 Mins
13	Linear Partial Differential Equations of second and higher order, Linear Partial Differential Equation with constant coefficient, Solution of homogeneous irreducible Partial Differential Equations	54 Mins
14	Method of separation of variables, Particular integral for irreducible non-homogeneous equations	54 Mins
15	Linear partial Differential equation with variable coefficients, Cannonical forms, Classificatin of second order partial differential equations, Canonical transformation of linear second order partial differential equations	54 Mins
16	Parabolic equation, Initial and boundary conditions, Heat equation under Dirichlet's Condition, Solution of Heat equation under Dirichlet's Condition ,	54 Mins
17	Solution of Heat equation under Neuman Condition, Solution of Parabolic equation under non-homogeneous boundary condition	54 Mins
18	Hyperbolic equation, occurrence of wave equations, in Mathematical Physics, Initial and boundary conditions, Initial value problem	54 Mins
19	D' Alembert's solutions, vibration of a sting of finite length, Initial value problem for a non-homogeneous wave equation	54 Mins
20	Elliptic equations, Gauss Divergence Theorem, Green's identities, Harmonic functions, Laplace equation in cylindrical and spherical polar coordinates, Dirichlet's Problem, Neumann Problem	54 Mins
Total		18 Hours

Block I

Ordinary Differential Equations

Units 1 & 2

General Theory of Linear Differential Equation

Introduction:

In mathematics, a linear differential equation is a differential equation that is defined by a linear polynomial in the unknown function and its derivatives. In this unit, we gave the definitions of linear ordinary differential equations of n-th order and some associated theorems.

Definition: An ordinary differential equation is an equation which involves only ordinary differential coefficients of a single independent variable.

Linear equations with variable coefficient:

A linear differential equation of order n with variable coefficients is an equation of the form

$P_0(x)\frac{d^n y}{dx^n} + P_1(x)\frac{d^{n-1} y}{dx^{n-1}} + \ldots + P_n(x)y = R(x)$

or more simply

$P_0(x)y^{(n)} + P_1(x)\, y^{(n-1)} + \ldots + P_n(x)y = R(x)$

where $P_0(x)$, $P_1(x)$, . . . , $P_n(x)$, R(X) are complex valued functions on some real interval I = [a,b]. Points where $P_0(x) = 0$ are called singular points and often the equation requires special consideration at such points. We assume that $P_0(x) \neq 0$ on I. Dividing by $P_0(x)$ we can obtain an equation of the same form, but with P_0 replaced by the constant 1. Thus we consider the equation

$$y^{(n)} + P_1(x)y^{(n-1)} + \ldots + P_n(x)y = R(x) \text{ ----------(1)}$$

We designate the left side of (1) by L(y) . Thus

$$L(y) = y^{(n)} + P_1(x)\, y^{(n-1)} + \ldots + P_n(x)y$$

and (1) becomes

L(y) = R(x) ---------(2)

If R(x) = 0 for all x in I, we say

L(y) = 0 --------(3)

is a homogeneous differential equation, whereas if $R(x) \neq 0$ for some x in I, the equation

L(y) = R(x)

is called a non-homogeneous equation. To study the general solution (2) it is necessary to consider the homogeneous equation (3). The reason of this is easily seen.

Suppose that in some way we know that $y_g(x, c_1, c_2, \ldots, c_n)$ is the general solution of (3), we expect it to contain n arbitrary constants since the equation is of n^{th} order and that $y_p(x)$ is a particular solution of (2). If y(x) is any solution of (2), we have

$$L(y-y_p) = L(y) - L(y_p) = R(x) - R(x) = 0$$

This shows that $y(x) - y_p(x)$ is a solution of (3).

Since $y_g(x, c_1, c_2, \ldots, c_n)$ is the general solution of (3), it follows that

$$y(x) - y_p(x) = y_g(x, c_1, c_2, \ldots, c_n)$$

i.e., $y(x) = y_p(x) + y_g(x, c_1, c_2, \ldots, c_n)$ for a suitable choice of constants $c_1, c_2, \ldots, c_n$. Thus we obtain the following theorem.

Theorem 1.1: If y_g is the general solution of (3) and y_p is any particular solution of (2), then $y_g + y_p$ is the general solution of (2).

We shall later show that if y_g is known, then a formal procedure is available for finding y_p. This shows that the central problem in the theory of linear differential equation is that of solving the corresponding homogeneous equation. Accordingly most of our attention will be devoted to study the structure of y_g. The following existence and uniqueness theorem assure ourselves that the homogeneous equation (3) really has a solution.

Theorem 1.2: Let $P_1(x), P_2(x), \ldots, P_n(x)$ be continuous functions on an interval containing the point x_0. If $a_0, a_1, \ldots, a_{n-1}$ are any n constants, then the homogeneous equation $L(y)=y^{(n)} + P_1(x)\, y^{(n-1)}+\text{-------}+P_n(x)y = 0$ has one and only one solution y(x) on I satisfying $y(x_0) = a_0$, $y'(x_0) = a_1$, $y''(x_0) = a_2$, --- $y^{(n-1)}(x_0) = a_{n-1}$. Clearly if $y(x) \equiv 0$, then $y^{(r)}(x)=0$, r =1,2, ---, n-1 and is therefore a unique solution of (3).

Definition: A set of m functions, $f_1(x), f_2(x), \text{---}, f_m(x)$ defined on some interval I are said to be linearly dependent on I if there exist constants $C_1, C_2, \text{---}, C_m$, not all zero, but

$c_1f_1(x) + c_2f_2(x) + \text{---} + c_mf_m(x) = 0$ for all x in I.

If $c_1 f_1(x) + c_2f_2(x) + \text{---} + c_mf_m(x) = 0$ implies $c_1=c_2 = \text{---} = c_m = 0$, then the functions will be linearly independent.

It is worth to noting that if one of $f_i(x)$ (i = 1, 2, --- , m) say $f_p(x)$ is identically zero on I, then the set of functions are linearly dependent, as it is sufficient to chose $c_p =1$ and $c_i = 0$ for all other i.

If $f_1, f_2, \text{---}, f_m$ are any m solutions of L(y) = 0 on an interval I and $c_1, c_2, \text{---}, c_m$ are any m constants, then

$L(c_1f_1 + c_2f_2 + \text{---} + c_mf_m)$

$= c_1L(f_1) + c_2L(f_2) + \text{---} + c_mL(f_m)$

$= 0,$

which implies that $c_1f_1+c_2f_2+ \text{---}+c_mf_m$ is also a solution of L(y) = 0. Therefore any linear combination of solutions is again a solution. The trivial solution is the function which is identically zero on I.

Theorem 1.3: There exist n linearly independent solutions of L(y) = 0 on an interval I where L(y) is given by

$L(y)= y^{(n)}+p_1(x)y^{(n-1)} + \text{---} + p_n(x)y.$

Proof: Let x_0 be a point on I. In view of Theorem 1.2, there is a solution y of L(y) = 0 satisfying

$y_1(x_0) = 1, y_1'(x_0) = 0, y_1''(x_0) = 0, \text{---}, y_1^{(n-1)}(x_0) = 0.$

In general, for each i =1, 2, ---, n, there is a solution y_i satisfying

$y_i^{(i-1)}(x_0) = 1, y_i^{(j-1)}(x_0) = 0, \ j \neq i$ —— (1)

We have to show that $y_1, y_2, \text{---}, y_n$ are linearly independent. Suppose there are constants $c_1, c_2, \text{---}, c_n$ such that

$c_1y_1(x)+c_2y_2(x) + \text{---} + c_ny_n(x) = 0$ —— (2) for all $x \varepsilon I$.

Differentiating successively n-1 times we obtain

$$\left.\begin{array}{l} c_1\, y_1'(x) + c_2\, y_2'(x) + \text{---} + c_n\, y_n'(x) = 0 \\ \text{----------} \\ \text{----------} \\ c_1y_1^{(n-1)}(x) + c_2y_2^{(n-1)}(x) + \text{---} c_ny_n^{(n-1)}(x) = 0 \end{array}\right\} \text{—— (3)}$$

for all x in I. In particular, (2) and (3) must hold at x_0.

Putting $x= x_0$ in (2) and using (1), we obtain,

$c_1 . 1 + c_2 . 0 + - - - + c_n . 0 = 0$

Putting $x = x_0$ in the equations (3) and using (1), we obtain

$c_2 = c_3 = --- = c_n = 0,$

and thus the solutions $y_1 , y_2 --- , y_n$ are linearly independent. This proves the theorem.

Definition: let $f_1(x), - - - f_n(x)$ be n functions defined on an interval I = [a, b] and each possesing (n-1)th derivatives. The Wronskian of these n functions is defined by the determinant

$$W(f_1 f_2, \dots f_n)(x) = \begin{vmatrix} f_1(x) & f_2(x) \dots & f_n(x) \\ f_1'(x) & f_2'(x) \dots & f_n'(x) \\ \vdots & \vdots \ \vdots \ \vdots & \vdots \\ f_1^{(n-1)}(x) & f_2^{(n-1)}(x) & \dots f_n^{(n-1)}(x) \end{vmatrix}$$

Theorem 1.4: If $y_1, y_2, \dots, y_n$ are n solutions of L(y) = 0 on some interval I, they are linearly independent there if and only if $W(y_1, y_2, \dots, y_n)(x) \neq 0 \quad \forall\, x \in I.$

Proof: First we assume that $W(y_1, y_2, \dots, y_n)(x) \neq 0 \quad \forall\, x \in I.$

If there are constants $y_1, y_2, \dots, y_n$ such that

$c_1y_1(x) + c_2y_2(x) + \dots + c_ny_n(x) = 0$ —— (1) $\forall x \in I$

then clearly

$$\left.\begin{array}{l} c_1 y_1'(x) + c_2 y_2'(x) + \dots + c_n y_n'(x) = 0 \\ \vdots \qquad\qquad \vdots \quad \vdots \qquad \vdots \\ c_1 y_1^{(n-1)}(x) + c_2 y_2^{(n-1)}(x) + \dots + c_n y_n^{(n-1)}(x) = 0 \end{array}\right\} \text{—— (2) } \forall x \in I$$

For a fixed $x \in I$, the equations (1) and (2) are linear homogeneous equations satisfied by $c_1, c_2, \dots, c_n$.

The determinant of the coefficients is just $W(y_1, y_2, \dots, y_n)(x)$, which is not zero. Hence there is only one solution of the system viz. $c_1 = c_2 = c_3 = \dots = c_n = 0$. Therefore $y_1, y_2, \dots, y_n$ are linearly independent in I.

Conversely, suppose that $y_1, y_2, \dots, y_n$ are linearly independent on I. Suppose that there is an $x_0 \in I$ such that

$$W(y_1, y_2, \dots, y_n)(x_0) = 0$$

This implies that the system of n linear equations

$$\left.\begin{array}{l} c_1y_1(x_0) + c_2y_2(x_0) + \dots + c_ny_n(x_0) = 0 \\ c_1 y_1'(x_0) + c_2 y_2'(x_0) + \dots + c_n y_n'(x_0) = 0 \\ \vdots \qquad\qquad \vdots \quad \vdots \\ c_1 y_1^{(n-1)}(x_0) + c_2 y_2^{(n-1)}(x_0) + \dots + c_n y_n^{(n-1)}(x_0) = 0 \end{array}\right\} \text{—— (3)}$$

has a solution $c_1, c_2, \dots, c_n$ not all zero.

We consider the function $\Phi = c_1y_1 + c_2y_2 + \dots + c_ny_n$

Now, $L(\Phi)=0$ and from (3), we see that

$\Phi(x_0) = 0, \Phi'(x_0) = 0, \dots , \Phi^{(n-1)}(x_0) = 0$

From the uniqueness theorem (Theorem 2), $\Phi(x) \equiv 0 \quad \forall x \in I$, and thus

$c_1y_1(x) + c_2y_2(x) + \dots + c_ny_n(x) = 0 \ \forall x \in I.$

But this contradicts the fact that $y_1, y_2, \ldots\ldots, y_n$ are linearly independent in I. Thus our supposition that there is a point x_0 in I such that $W(y_1, y_2, \ldots\ldots, y_n)(x_0) = 0$ must be false.

Hence $W(y_1, y_2, \ldots\ldots, y_n)(x_0) \neq 0$ for all $x \in I$. This proves the theorem.

Theorem 1.5: Let $y_1, y_2, \ldots\ldots, y_n$ be n linearly independent solutions of $L(y) = 0$ on an interval I. If y(x) is any solution of $L(y) = 0$ on I, it can be expressed in the form $y = c_1y_1 + c_2y_2 + \ldots\ldots + c_ny_n$, where $c_1, c_2, \ldots\ldots, c_n$ are constants.

Proof: Let x_0 be a point in I, and suppose that

$y(x_0) = a, y'(x_0) = a_2, \ldots\ldots, y^{(n-1)}(x_0) = a_n.$

We show that there exist unique constants $c_1, c_2, \ldots\ldots, c_n$ such that $\Phi = c_1y_1 + c_2y_2 + \ldots\ldots + c_ny_n$ is a solution of $L(y) = 0$ satisfying $\Phi(x_0) = a_1, \Phi^1(x_0) = a_2, \ldots\ldots, \Phi^{(n-1)}(x_0) = a_n.$

By the uniqueness result of Theorem 1.2, we then have $y = \Phi$ i.e. $y = c_1y_1 + c_2y_2 + \ldots\ldots + c_ny_n.$ The initial conditions for Φ are equivalent to the following equation for $c_1, c_2, \ldots\ldots, c_n$:

$$\left.\begin{aligned} &c_1y_1(x_0) + c_2y_2(x_0) + \ldots\ldots\ldots\ldots + c_ny_n(x_0) = a_1 \\ &c_1y_1'(x_0) + c_2y_2'(x_0) + \ldots\ldots\ldots\ldots + c_ny_n'(x_0) = a_2 \\ &\vdots \qquad\quad \vdots \qquad\qquad\quad \vdots \quad \vdots \\ &c_1y_1^{(n-1)}(x_0) + c_2y_2^{(n-1)}(x_0) + \ldots\ldots\ldots\ldots + c_ny_n^{(n-1)}(x_0) = a_n \end{aligned}\right\} \text{—— (1)}$$

This is a system of n non-homogeneous equations for $c_1, c_2, \ldots\ldots, c_n$. The determinant of the coefficients is just $W(y_1, y_2, \ldots\ldots, y_n)(x_0)$, which is not zero, since $y_1, y_2, \ldots\ldots, y_n$ are linearly independent. Therefore there is a unique solution $c_1, c_2, \ldots\ldots, c_n$ of the system (1) and this completes the proof.

Theorem 1.6: Let $y_1, y_2, \ldots\ldots, y_n$ be n solutions of $L(y) = 0$ on an interval I and x_0 be any point in I. Then

$$W(y_1, y_2, \ldots\ldots, y_n)(x) = \exp\left[-\int_{x_0}^{x} p_1(t)dt\right] w(y_1, y_2, \ldots, y_n)(x_0) \ldots (1)$$

Proof: We first prove the theorem for n = 2.

In this case y_1, y_2 are solutions of the second order linear homogeneous differential equation

$$y'' + p_1(x)y' + p_2(x)y = 0 \text{ —(1)}$$

Therefore, the Wronskian

$$W(y_1, y_2) = \begin{vmatrix} y_1 & y_2 \\ y_1' & y_2' \end{vmatrix} = y_1y_2' - y_1'y_2.$$

$\therefore W'(y_1, y_2) = y_1y_2'' - y_1''y_2.$

Since y_1, y_2 are solutions of (1), we have

$$y_1'' = -p_1(x)y_1' + p_2(x)y_1$$

$$y_2'' = -p_1(x)y_2' + p_2(x)y_2$$

Hence $W'(y_1, y_2) = y_1(-p_1(x)y_2' + p_2(x)y_2) - (-p_1(x)y_1' + p_2(x)y_1)y_2$

$$= -p_1(x)[y_1y_2' - y_1'y_2]$$

$$= -p_1(x)w(y_1y_2)$$

From this we see that $W(y_1\ y_2)$ is a solution of the first order linear differential equation

$$z' = -p_1(x)z.$$

Solving we get

$$W(y_1, y_2)(x) = c \exp\left[-\int_{x_0}^{x} p_1(t)dt\right]$$

Putting $x=x_0$, we see that $c = W(y_1, y_2)(x_0)$ and therefore

$$W(y_1, y_2)(x) = \exp\left[-\int_{x_0}^{x} p_1(t)dt\right] w(y_1, y_2)(x_0).$$

This proves the result for n = 2.

Now we consider the theorem for general n.

For brevity, we write $W(x) = W(y_1, y_2, \ldots\ldots, y_n)(x)$

Then $W(x) = \begin{vmatrix} y_1 & y_2 \ldots & y_n \\ y_1' & y_2' \ldots & y_n' \\ \vdots & \vdots & \vdots \\ y_1^{(n-1)} & y_2^{(n-1)} \ldots & y_n^{(n-1)} \end{vmatrix}$

Differentiating w.r.t. x, we get

$$W'(x) = \begin{vmatrix} y_1 & y_2 \ldots & y_n \\ y_1' & y_2' \ldots & y_n' \\ \vdots & \vdots & \vdots \\ y_1^{(n)} & y_2^{(n)} \ldots & y_n^{(n)} \end{vmatrix}$$

Since $y_1, y_2, \ldots, y_n$ are solutions of

$L(y) = y^{(n)} + p_1(x)\, y^{(n-1)} + P_2(x)\, y^{(n-2)} + \ldots + p_{n-1}(x)y' + p_n(x)\, y = 0,$

we have

$y_i^{(n)} = -\, p_1(x)y_i^{(n-1)} - p_2(x)y_i^{(n-2)} - \ldots - P_{n-1}(x)y_i' - p_n(x)y_i,\ \ i = 1, 2, \ldots, n,$ and therefore

$$\therefore\ W'(x) = \begin{vmatrix} y_1 & y_2 \ldots & y_n \\ y_1' & y_2' \ldots & y_n' \\ \vdots & \vdots & \vdots \\ y_1^{(n-2)} & y_2^{(n-2)} \ldots & y_n^{(n-2)} \\ -\sum_{i=1}^{n} p_i(x)y_1^{(n-i)} & -\sum_{i=1}^{n} p_i(x)y_2^{(n-i)} \ldots & -\sum_{i=1}^{n} p_i(x)y_n^{(n-i)} \end{vmatrix}$$

Multiplying 1^{st} row by p_n, 2^{nd} row by p_{n-1}, ...,$(n-1)^{th}$ row by p_2 and then adding with last row, we get

$$W'(x) = \begin{vmatrix} y_1 & y_2 \ldots & y_n \\ y_1' & y_2' \ldots & y_n' \\ \vdots & \vdots & \vdots \\ y_1^{(n-2)} & y_2^{(n-2)} \ldots & y_n^{(n-2)} \\ -p_i(x)y_1^{(n-1)} & -p_i(x)y_2^{(n-1)} \ldots & -p_i(x)y_n^{(n-1)} \end{vmatrix}$$

$$= -p_1(x)\, W(x).$$

Solving we get,

$W(x) = c \exp\left[-\int_{x_0}^{x} p_1(t)dt\right]$

Putting $x = x_0$, we get $c = W(x_0)$

$\therefore W(x) = \exp\left[-\int_{x_0}^{x} p_1(t)dt\right] W(x_0)$

This proves the theorem.

Note: The identity (1) is known as Abel's identity.

Summary:

- Linear equations with variable coefficient
- Theorem on general and particular solution
- Linear independent solution of linear ODE
- Wronskian

Unit 3

Second Order Linear Differential Equation

Introduction:

In this unit, we will study linear ordinary differential equations of the second order. In this section we introduce the method of variation of parameters to find particular solutions to non-homogeneous differential equation. We give a detailed examination of the method as well as derive a formula that can be used to find particular solutions. Then we will discuss about the fundamental set of solutions and second order linear ODE in its normal form.

Method of Variation of parameters:

Consider the non-homogeneous equation

$L(y) = R(x)$ ─ (1)

where L(y) is given by

$L(y) = y^{(n)} + p_1(x)y^{(n-1)} + \text{---} + p_n(x)y.$

If $P_i(x)$, i =1, 2,...,n are constants and R(x) has a particular simpler form viz., an exponential, sine, cosine, or a polynomial, we can obtain a particular solution of (1) and $y = y_g + y_p$ is the general solution.

We now develop a powerful method that always works regardless of the nature of $P_i(x)$ and R(x), provided that the general solution of the homogeneous equation L(y) = 0 is known.

Let $y(x) = c_1y_1(x) + c_2y_2(x) + ... + c_ny_n(x)$ - - - - (2)

be the general solution of the homogeneous equation L(y) = 0, - - - - (3)

where $y_1, y_2, ..., y_n$ are n linearly independent solutions of (3). We replace the constants in (2) by unknown functions $v_1(x), v_2(x), ..., v_n(x)$ in such a manner that

$y(x) = v_1(x)y_1(x) + v_2(x)y_2(x) + ... + v_n(x)y_n(x)$ will be a solution of (1).

To find n unknown function it will be necessary to have n equations relating these functions.

We consider successive (n-1)th derivative of y(x) and write

$y'(x) = v_1y_1' + v_2y_2' + \ldots + v_ny_n'$

so that $v_1'y_1 + v_2'y_2 + \ldots + v_n'y_n = 0$

$y''(x) = v_1y_1'' + v_2y_2'' + \ldots + v_ny_n''$

so that $v_1'y_1' + v_2'y_2' + \ldots + v_n'y_n' = 0$

: : : :

$y^{(n-1)}(x) = v_1y_1^{(n-1)} + v_2y_2^{(n-1)} + \ldots + v_ny_n^{(n-1)}$

so that $v_1'y_1^{(n-2)} + v_2'y_2^{(n-2)} + \ldots + v_n'y_n^{(n-2)} = 0$

Differentiating once w.r.t. x, we get

$y^{(n)}(x) = v_1y_1^{(n)} + v_2y_2^{(n)} + ... + v_ny_n^{(n)} + v_1'y_1^{(n)} + v_2'y_2^{(n-1)} + ... + v_n'y_n^{(n-1)}$

Substituting the values of $y'(x), y''(x), \ldots, y^{(n)}(x)$ in (1), we get

$R(x) = v_1[y_1^{(n)} + p_1y_1^{(n-1)} + \ldots + p_{n-1}y_1' + p_ny_1]$

$+ v_2[y_2^{(n)} + p_1y_2^{(n-1)} + \ldots + p_{n-1}y_2' + p_ny_2]$

$+ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$

$+ v_n[y_n^{(n)}+p_1y_n^{(n-1)}+ \ldots +p_{n-1}y_n' +p_ny_n]$

$+ v_1'y_1^{(n-1)}+ v_2'y_2^{(n-1)}+ \ldots + v_n'y_n^{(n-1)}$

Since $y_1, y_2, \ldots, y_n$ are solutions of (3), we obtain from above

$$v_1'y_1^{(n-1)}+ v_2'y_2^{(n-1)}+ \ldots + v_n'y_n^{(n-1)} = R(x)$$

Thus we obtain the non-homogeneous system

$$\left.\begin{aligned} v_1'y_1^{(i)}+ v_2'y_2^{(i)}+ \ldots + v_n'y_n^{(i)} &= 0 \\ i &= 0, 1, 2, \ldots, n-2 \\ v_1'y_1^{(n-1)}+ v_2'y_2^{(n-1)}+ \ldots + v_n'y_n^{(n-1)} &= R(x) \end{aligned}\right\} \text{—— (4)}$$

Since the Wronskian $w(y_1, y_2, \ldots, y_n)(x) \neq 0$, the system (4) has a non-trivial solution $v_1', v_2', \ldots, v_n'$ so that

$v_i = \frac{W_i(x)}{W(x)} R(x)$, $i = 1, 2, \ldots, n$

where $W_i(x)$ is obtained from $W(x)$ by replacing the i-th column of $W(x)$ by $(0,0,0, \ldots, 1)$

$\therefore\ v_i(x) = \int_{x_0}^{x} \frac{W_i(t)}{W(t)} R(t)dt$, $i = 1, 2, \ldots\ldots, n$

With these values of $v_i(x)$, $i = 1, 2, \ldots\ldots, n$, we obtain a particular solution of (1) as

$y(x) = (c_1 + v_1)\, y_1(x) + (c_2 + v_2)y_2(x) + \ldots + (c_n + v_n)y_n(x)$

where $c_1, c_2, \ldots\ldots, c_n$ are arbitrary constants.

Example 1: Find a particular solution of

$$y'' + 2y' + y = e^{-x}\log x$$

Solution: The corresponding homogeneous equation is

$$y'' + 2y' + y = 0$$

The complementary function is $y(x) = (A+Bx)e^{-x}$

We assume $y(x) = v_1(x)e^{-x} + V_2(x)xe^{-x}$ to be a solution of the given equation subject to the conditions

$v_1'(x)e^{-x} + v_2'(x)\, xe^{-x} = 0$

and $-v_1'(x)e^{-x} + (e^{-x} - xe^{-x})\, v_2'(x) = e^{-x}\log x$

Here,

$$w(x) = \begin{vmatrix} e^{-x} & xe^{-x} \\ -e^{-x} & e^{-x} - xe^{-x} \end{vmatrix}$$

$$= e^{-2x} \begin{vmatrix} 1 & x \\ -x & 1 - x \end{vmatrix}$$

$= e^{-2x} \neq 0$

$$\therefore v_1'(x) = \frac{\begin{vmatrix} 0 & xe^{-x} \\ 1 & e^{-x} - xe^{-x} \end{vmatrix}}{e^{-2x}} e^{-x}\log x$$

$= -x\log x$

$$\text{and } v_2'(x) = \frac{\begin{vmatrix} e^{-x} & 0 \\ -e^{-x} & 1 \end{vmatrix}}{e^{-2x}} e^{-x}\log x$$

$= \log x$

$\therefore v_1(x) = -\int_0^x t \log t \, dt$

$= -\frac{x^2}{2} \log x + \frac{x^2}{4}$

$v_2(x) = \int_0^x \log t \, dt = x \log x - x$

So a particular solution of the given equation is

$y(x) = (\frac{x^2}{4} - \frac{x^2}{2} \log x) e^{-x} + (x \log x - x) x e^{-x}$

Example 2: Find a particular solution of the equation

$y'' - 2y' - 3y = 64xe^{-x}$

Solution: The corresponding homogeneous equation is

$y'' - 2y' - 3y = 0$

The complementary function is $y(x) = Ae^{3x} + Be^{-x}$

We assume $y(x) = v_1(x)e^{3x} + v_2(x)e^{-x}$ to be a solution of the given equation subject to the conditions

$v_1'(x)e^{3x} + v_2'(x)e^{-x} = 0$

and $3v_1'(x)e^{3x} - v_2'(x)e^{-x} = 64xe^{-x}$

Here, $w(x) = \begin{vmatrix} e^{3x} & e^{-x} \\ 3e^{3x} & -e^{-x} \end{vmatrix}$

$= e^{2x} \begin{vmatrix} 1 & 1 \\ 3 & -1 \end{vmatrix}$

$= -4e^{2x} \neq 0$

$\therefore v_1(x) = \frac{\begin{vmatrix} 0 & e^{-x} \\ 1 & -e^{-x} \end{vmatrix}}{-4e^{2x}} 64xe^{-x}$

$= \frac{-e^{-x}}{-4e^{2x}} 64xe^{-x}$

$= 16xe^{-4x}$

and $v_2'(x) = \frac{\begin{vmatrix} e^{3x} & 0 \\ 3e^{3x} & 1 \end{vmatrix}}{-4e^{2x}} 64xe^{-x}$

$= \frac{e^{3x}}{-4e^{2x}} 64xe^{-x}$

$= -16x$

$\therefore v_1(x) = \int 16xe^{-4x} \, dx$

$= 16[-\frac{e^{-4x}}{4} x - \frac{e^{-4x}}{16}]$

$= -4xe^{-4x} - e^{-4x}$

and $v_2(x) = -\int 16x \, dx$

$= -8x^2$

Hence a particular solution of the given equation is

$y(x) = -(4x+1)e^{-x} - 8x^2e^{-x}$

Exercise: Find a particular solution of each of the following:

i) $y'' + y = \text{cosec } x$
ii) $y'' + 4y = \tan 2x$
iii) $y'' + 2y' + 5y = e^{-x}\sec 2x$

Fundamental set of solutions:

A set of functions which has the property that if y_1, y_2 belong to the set and c_1, c_2 are any two constants, then $c_1y_1 + c_2y_2$ belongs to the set also, is called a linear space of functions. We have seen that the set of all solutions of $L(y) = 0$ on an interval I is a linear space of functions.

If a linear space of functions contains n functions $y_1, y_2, ..., y_n$ which are linearly independent and such that every function in the space can be represented as a linear combination of these functions, then $y_1, y_2, ..., y_n$ is called a basis of the linear space, and the dimension of the linear space is the integer n. A basis is sometimes called a fundamental set of solutions.

Theorem 2.1: If ϕ_1 is a solution of

$$L(y) = y'' + p_1(x)y' + p_2(x)y = 0 \quad — (1)$$

on an interval I, and $\phi_1(x) \neq 0 \ \forall \ x \in I$, a second solution $\phi_2(x)$ of (1) on I is given by

$$\phi_2(x) = \phi_1(x)\int_{x_0}^{x} \frac{1}{[\phi_1(s)]^2} \exp\left[-\int_{x_0}^{s} p_1(t)dt\right]ds$$

The functions ϕ_1 and ϕ_2 form a basis for the solutions of (1).

Proof: Here $L(y) = y'' + p_1(x)y' + p_2(x)y = 0$

Let $\phi_2 = u\phi_1$ be a solution on I.

Then $L(u\phi_1) = (u\phi_1)'' + P_1(x)(u\phi_1)' + P_2(x)\, u\, \phi_1 = 0$

i.e., $u''\phi_1 + 2u'\phi_1' + u\phi_1'' + P_1(x)\, u'\phi_1 + P_1(x)\, u\, \phi_1' + P_2(x)\, u\, \phi_1 = 0$

i.e., $u''\phi_1 + u'[2\phi_1' + P_1(x)\phi_1] + u[\phi_1'' + P_1(x)\phi_1' + P_2(x)\phi_1] = 0$

i.e., $u''\phi_1 + (2\phi_1' + P_1(x)\phi_1)u' = 0$

Let $v = u'$.

Then we have

$$\phi_1 v' + 2\phi_1' v + P_1\phi_1 v = 0 \ldots\ldots (2)$$

i.e., $\phi_1^2 v' + 2\phi_1\phi_1' v + P_1\phi_1^2 v = 0 \ \ldots\ldots \ (3)$

i.e., $(\phi_1^2 v)' + P_1(\phi_1^2 v) = 0$

This implies that

$$\phi_1^2 v(x) = c \exp\left[-\int_{x_0}^{x} p_1(t)dt\right]$$

where x_0 is a point in I and c is any constant.

Since any constant multiple of a solution of (3) is again a solution, we see that

$$v(x) = \frac{1}{[\phi_1(x)]^2} \exp\left[-\int_{x_0}^{x} p_1(t)dt\right]$$

is a solution of (3), and also of (2).

Therefore two independent solutions of the equation (1) on I are ϕ_1 and ϕ_2 where

$$\phi_2(x) = \phi_1(x)\int_{x_0}^{x} \frac{1}{[\phi_1(s)]^2} \exp\left[-\int_{x_0}^{s} p_1(t)dt\right]ds$$

This completes the proof.

Example 3: Find the general solution of $x^2y'' - 2y = 0$, $0 < x < \infty$.

Solution: It is easy to verify that $\phi_1 = x^2$ is a solution of the given homogeneous equation in $0 < x < \infty$. Since this function does not vanish on 0, $0 < x < \infty$, there is another independent solution ϕ_2 of the form $\phi_2 = u\,\phi_1 = ux^2$.
So we obtain $x^2(ux^2)'' - 2ux^2 = 0$

i.e., $x^2(u''x^2+4xu' +2u) - 2ux^2 = 0$

i.e., $u'' x^2 + 4xu' = 0$

i.e., $v'x^2 + 4xv = 0 \quad [u' = v]$

i.e., $v'x + 4v = 0$

i.e., $\frac{v'}{v} = -\frac{4}{x}$

i.e., $v = ce^{-\int \frac{4}{x} dx}$

i.e., $v = x^{-4} \quad$ [taking c = 1]

$\therefore u = -\frac{1}{3} x^{-3}$

This gives $\phi_2(x) = -\frac{1}{3x}$

Since any constant multiple of a solution is also a solution, we may choose second solution as $\phi_2(x) = \frac{1}{x}$

Thus x^2, x^{-1} form a basis of the given equation on 0, $0 < x < \infty$.

Example 4: Find the general solution of $(x^2-1)y'' - 2xy' + 2y = (x^2-1)^2$, $0 < x < \infty$, given that $y = x$ is a solution of the corresponding homogeneous equation.

Solution. The given equation is

$$= (x^2-1)\, y'' - 2xy' + 2xy' - 2y = (x^2-1)^2 \quad \cdots\cdots (1)$$

The corresponding homogeneous equation is

$$(x^2-1)\, y'' - 2xy' + 2y = 0 \quad \ldots\ldots (2)$$

Given that $\phi_1 = x$ is a solution of the homogeneous equation (2). Let $\phi_2 = u\,\phi_1 = ux$ be another solution of (2).

So we obtain from (2),

$$(x^2-1)\,(ux)'' - 2x(ux)' + 2ux = 0$$

or, $(x^2-1)\,(u''x+2u') - 2x(u'x+u) + 2ux = 0$

or, $u''x\,(x^2-1) + 2u'\,(x^2-1) - 2x^2u' = 0$

or, $u''x\,(x^2-1) - 2u' = 0$

or, $v'x\,(x^2-1) - 2v = 0 \quad [u' = v]$

or, $\frac{v'}{v} = \frac{2}{x(x^2-1)}$

or, $\log v = -2\int \frac{dx}{x} + \int \frac{dx}{x+1} + \int \frac{dx}{x-1}$

or, $\log v = -2\log x + \log(x+1) + \log(x-1) + \log c$

or, $v = \frac{(x+1)(x-1)c}{x^2}$

$\therefore\ v = \frac{x^2-1}{x^2} \quad$ [taking c = 1]

$\therefore\ u = \int (1 - \frac{1}{x^2})dx$

or, $u = (x + \frac{1}{x})$

This gives $\phi_2(x) = (x + \frac{1}{x})\, x$

i.e., $\phi_2(x) = x^2 + 1$

Thus x and $x^2 + 1$ form a basis of the given equation on 0, $0<x<\infty$.

Example 5: Find the general solution of $x^2y'' - 7xy' + 15y = 0$ given that $\phi_1(x) = x^3$ is a solution.

Solution: The given homogeneous equation is

$$x^2y'' - 7xy' + 15y = 0$$

Given that $\phi_1 = x^3$ is a solution of the homogeneous equation on 0, $0<x<\infty$.

Since this function does not vanish on 0, $0<x<\infty$, there is another independent solution ϕ_2 of the form $\phi_2 = u\,\phi_1 = ux^3$

So we obtain $x^2(ux^3)'' - 7x(ux^3)' + 15\,ux^3 = 0$

or, $x^2(u''x^3 + 2u'.\,3x^2 + u.6x) - 7x(u\,3x^2 + u'x^3) + 15ux^3 = 0$

or, $u''x^5 + 6u'x^4 + 6ux^3 - 21ux^3 - 7u'x^4 + 15ux^3 = 0$

or, $u''x^5 - u'x^4 = 0$

or, $\log v = \log cx$

$\therefore v = x$ [taking c = 1]

$\therefore u = \int x\,dx$

i.e., $u = \frac{x^2}{2}$

This gives $\phi_2 = \frac{x^2}{2} . x^3 = \frac{x^5}{2}$

Since any constant multiple of a solution is also a solution, we may choose second solution $\phi_2(x) = x^5$

Thus x^3, x^5 form a basis of the given equation on 0, $0 < x < \infty$.

Second order linear differential equation – Normal form:

We have already seen that a linear second order equation as

$$\frac{d^2y}{dx^2} + p\frac{dy}{dx} + Qy = R \quad \ldots\ldots \quad (1)$$

can be solved when one integral solution of its corresponding homogeneous equation is known. But when it is not possible to get such an integral solution, then we can find a method by which (1) can be solved.

Let $y = uv$ be a solution of (1), where u and v are functions of x and none of them is an integral solution of the corresponding homogeneous equation of (1).

Now, $\frac{dy}{dx} = u\frac{dv}{dx} + v\frac{du}{dx}$

and $\frac{d^2y}{dx^2} = u\frac{d^2v}{dx^2} + 2\frac{du}{dx}\frac{dv}{dx} + v\frac{d^2u}{dx^2}$

Then from (1) we got,

$$u\frac{d^2v}{dx^2} + 2\frac{du}{dx}\frac{dv}{dx} + v\frac{d^2u}{dx^2} + p\left(u\frac{dv}{dx} + v\frac{du}{dx}\right) + Quv = R$$

or, $u\frac{d^2v}{dx^2} + \left(2\frac{du}{dx} + pu\right)\frac{dv}{dx} + v\left(\frac{d^2u}{dx^2} + p\frac{du}{dx} + Qu\right) = R \quad \ldots\ldots \quad (2)$

Now, the term containing $\frac{dv}{dx}$ may be removed from (2), if

we put $2\frac{du}{dx} + pu = 0 \quad \ldots\ldots \quad (3)$

i.e., $u = e^{-\frac{1}{2}\int p\,dx}$ . . . . . (4)

Using (3) in (2), we obtain,

$u\frac{d^2v}{dx^2} + v(\frac{d^2u}{dx^2} + p\frac{du}{dx} + Qu) = R$

i.e., $\frac{d^2v}{dx^2} + \frac{1}{u}(\frac{d^2u}{dx^2} + p\frac{du}{dx} + Qu)v = \frac{R}{u}$ . . . . . (5)

From (3) we get,

$\frac{du}{dx} = -\frac{1}{2}pu$

$\therefore \frac{d^2u}{dx^2} = -\frac{1}{2}p\frac{du}{dx} - \frac{1}{2}u\frac{dp}{dx}$

or, $\frac{d^2u}{dx^2} = \frac{1}{4}p^2u - \frac{1}{2}u\frac{dp}{dx}$

So from (5), we get,

$\frac{d^2v}{dx^2} + \frac{1}{u}(\frac{1}{4}p^2u - \frac{1}{2}u\frac{dp}{dx} + p\frac{du}{dx} + Qu)v = \frac{R}{u}$

or, $\frac{d^2v}{dx^2} + \frac{1}{u}(\frac{1}{4}p^2u - \frac{1}{2}u\frac{dp}{dx} - \frac{1}{2}p^2u + Qu)v = \frac{R}{u}$

or, $\frac{d^2v}{dx^2} + (Q - \frac{1}{2}\frac{dp}{dx} - \frac{1}{4}p^2)v = \frac{R}{u}$

i.e., $\frac{d^2v}{dx^2} + Lv = M$ . . . . . (6)

where $L = Q - \frac{1}{2}\frac{dp}{dx} - \frac{1}{4}p^2$ and $M = \frac{R}{u}$

The equation (6), which does not contain the term $\frac{dv}{dx}$, is known as normal form of (1).

Example 6: Solve $\frac{d^2v}{dx^2} - 4x\frac{dy}{dx} + (4x^2-1)y = -3e^{x2}\sin 2x$ . . . . . (1) by reducing it in its normal form.

Solution: Here $P = -4x$

$Q = 4x^2 - 1$

$R = -3e^{x2}\sin 2x$

In order to remove the first derivative, we choose

$u = e^{-\frac{1}{2}\int p\,dx}$

i.e., $u = e^{-\frac{1}{2}\int -4x\,dx}$

i.e., $u = e^{x2}$

Now, putting $y = uv = ve^{x2}$ in (1), we get

$\frac{d^2v}{dx^2} + Lv = M$ — (2)

where $L = Q - \frac{1}{2}\frac{dp}{dx} - \frac{1}{4}p^2$

$= (4x^2 - 1) - \frac{1}{2}(-4) - \frac{1}{4}.16\,x^2$

$= 4x^2 - 1 + 2 - 4x^2$

$= 1$

and M = $\frac{R}{u}$ = − 3sin 2x

Now (2) reduces to

$\frac{d^2v}{dx^2}$ + v = −3sin 2x . . . . . (3)

Now the complementary function is

v = $c_1$cos x + c_2 sin x, where c_1, c_2 are arbitrary constants.

P.I = $\frac{1}{D^2+1}$ (-3 sin 2x)

= sin 2 x

So the general solution of (3) is

v = $c_1$cos x + c_2 sin x + sin 2x

∴ The general solution of (1) is

y = uv

i.e., y = e^{x2}($c_1$cos x + c_2 sin x + sin 2x)

Example 7: Solve $\frac{d^2y}{dx^2}$ −2 tan x $\frac{dy}{dx}$ −(a^2+1)y = e^x sec x by reducing to normal form.

Solution: The given equation is

$\frac{d^2y}{dx^2}$ −2 tan x $\frac{dy}{dx}$ −(a^2+1)y = e^x sec x . . . . . (1)

Here, P = −2 tan x

Q = − (a^2+1)

R = e^x sec x

In order to remove the first derivative, we choose

u = $e^{-\frac{1}{2}\int p\,dx}$

i.e., u = $e^{-\frac{1}{2}\int -2\,tan\,x\,dx}$

i.e., u = sec x

Now putting y = uv = v sec x in (1) we have

$\frac{d^2y}{dx^2}$ + Lv = M — (2)

where L = Q − $\frac{1}{2}\frac{dp}{dx}$ − $\frac{1}{4}$ p^2

= −(a^2+1) − $\frac{1}{2}$ (−2 sec^2x) − $\frac{1}{4}$. 4 tan^2 x

= − a^2 – 1 + sec^2x – tan^2x = −a^2

and M = $\frac{R}{u}$ = e^x

Now, (2) reduces to

$\frac{d^2v}{dx^2}$ – a^2 v = e^x . . . . . (3)

The complementary function is

$v = c_1e^{ax} + c_2e^{-ax}$, where c_1 and c_2 are arbitrary constants

P. I. $= \frac{1}{D^2-a^2}e^x = \frac{e^x}{1-a^2}$

So, general solution of (3) is

$$v = c_1e^{ax} + c_2e^{-ax} + \frac{e^x}{1-a^2}$$

∴ The general solution of (1) is

$$y = uv$$

i.e., $y = \sec x\ (c_1e^{ax} + c_2e^{-ax} + \frac{e^x}{1-a^2})$

Example 8: Solve $\frac{d^2y}{dx^2} - \frac{2}{x}\frac{dy}{dx} + (a^2+\frac{2}{x^2})y = 0$ by reducing to normal form.

Solution: The given equation is

$$\frac{d^2y}{dx^2} - \frac{2}{x}\frac{dy}{dx} + (a^2+\frac{2}{x^2})y = 0 \quad \ldots\ldots \quad (1)$$

Here, $P = -\frac{2}{x}$

$$Q = a^2+\frac{2}{x^2}$$

$$R = 0$$

In order to remove the first order derivative, we choose

$$u = e^{-\frac{1}{2}\int p dx}$$

i.e., $u = e^{-\frac{1}{2}\int -\frac{2}{x}dx}$

i.e., $u = x$

Putting $y = uv = vx$ in (1), we have

$$\frac{d^2v}{dx^2} + Lv = M \quad \ldots\ldots \quad (2)$$

where $L = Q - \frac{1}{2}\frac{dp}{dx} - \frac{1}{4}p^2 = a^2 + \frac{2}{x^2} - \frac{1}{2}(\frac{2}{x^2}) - \frac{1}{4}.\frac{4}{x^2}$

$$= a^2 + \frac{2}{x^2} - \frac{1}{x^2} - \frac{1}{x^2}$$

$$= a^2$$

and $M = \frac{R}{u} = 0$

Therefore, (2) reduces to

$$\frac{d^2v}{dx^2} + a^2v = 0 \quad -(3)$$

Therefore, the general solution of (3) is

$V = c_1 \cos ax + c_2 \sin ax$, where c_1 and c_2 are arbitrary constants

Hence the general solution of the given equation is

$$y = uv$$

i.e., $y = x\ (c_1 \cos ax + c_2 \sin ax)$

Summary:

- Non-homogeneous second order linear ODE
- Method of variation of parameters
- Fundamental set of solutions
- Second order linear ODE – Normal forms

Unit 4

Initial Value Problem and Boundary Value Problem

Introduction:

Initial value problem does not require to specifying the value at boundaries, instead it needs the value during initial condition. These usually apply for dynamic system that is changing over time as in Physics. An example, to solve a particle position under differential equation, we need the initial position and also initial velocity. Without these initial values, we cannot determine the final position from the equation. In contrast, boundary value problems not necessarily used for dynamic system. Instead, it is very useful for a system that has space boundary.

Initial Value Problem (IVP):

An initial value problem is a differential equation together with subsidiary conditions to be satisfied by the solution function and its derivative, all given at the same value of the independent variable.

For example, $\frac{d^2y}{dx^2} + 4y = 0$, $y(0) = 0$, $y'(0) = 2$ is an IVP.

Boundary value problem (BVP):

A boundary value problem is a differential equation together with subsidiary conditions to be satisfied by the solution function and its derivatives, where the conditions are given for more than one value of the independent variable. As for example,

$y'' - 9y = 0$, $y(0) = 0$, $y'(1) = 1$ is a BVP.

The general form of BVP is

Solve: $\frac{d^2y}{dx^2} + P\frac{dy}{dx} + Qy = R$ with the boundary conditions

$A_1y(a) + B_1y'(a) = c_1$

$A_2y(b) + B_2y'(b) = c_2$,

where P, Q, R are functions of x and A_1, B_1, C_1, A_2, B_2, C_2 are all real constants. Also $a \neq b$; A_1, B_1 are not together zero and also A_2, B_2 are not together zero.

The method of successive approximation:

We now face up to the general problem of finding solution of the equation

$\frac{dy}{dx} = f(x, y)$

where f is any real-valued continuous function defined on some rectangle

$R = \{(x,y) : |x-x_0| \leq a, |y - y_0| \leq b, a > 0, b > 0\}$

Our aim is to show that on some interval I containing x_0, there is a real-valued differentiable function y_1 such that the points $(x, y_1(x)) \in R$ for all $x \in I$ and $y_1'(x) = f(x, y_1(x))$, $y_1(x_0) = y_0$. Such a function y_1 is called a solution to the initial value problem

$$\left.\begin{array}{l} y' = f(x, y) \\ y(x_0) = y_0 \end{array}\right\} \quad \ldots\ldots \quad (1)$$

Our first step will be to show that the initial value problem (1) is equal to an integral equation

$$y = y_0 + \int_{x_0}^{x} f(t, y)dt \quad \ldots\ldots \quad (2)$$

on I.

By a solution of (2) on I is meant a real valued continuous function y_1 on I such that $(x, y_1(x)) \in R$ for all $x \in I$ and $y_1(x) = y_0 + \int_{x_0}^{x} f(t, y_1(t))dt \ldots\ldots$ (3)

for all x∈I.

Theorem 3.1 : A function y_1 is a solution of the IVP (1) on an interval I if and only if it is a solution of the integral equation (2) on I.

Proof: Suppose y_1 is a solution of the IVP (1) on I. Then

$$y_1'(x) = f(x, y_1(x)) \quad -(3)$$

Since y_1 is continuous on I, and f is continuous on R, the function F defined by $F(x) = f(x, y_1(x))$ is continuous on I .

Integrating (3) from x_0 to x, we get

$$y_1(x) = y_1(x_0) + \int_{x_0}^{x} f(t, y_1(t))dt \qquad \forall\, x \in I$$

and since $y_1(x_0) = y_0$ we see that y_1 is a solution of (2).

Conversely, suppose that y_1 satisfies (3) on I. Differentiating we find that

$$y_1'(x) = f(x, y_1(x)) \quad \forall\, x \in I.$$

Moreover, from the above we see that $y_1(x_0) = y_0$, and thus y_1 is a solution of the initial value problem (1). This proves the theorem.

Picard's Method of Successive Approximation:

We know that the initial value problem

$y' = f(x, y)$, $y(x_0) = y_0$

is equivalent to the integral equation

$$y(x) = y_0 + \int_{x_0}^{x} f(t, y)\, dt.$$

Since the information concerning the expression of y in term of x is absent the integral on right hand side of the above integral equation cannot be evaluated. Here the exact value of y cannot be obtained. Therefore we determine a sequence of approximations to the solution of the integral equation as follows:

First we put $y = y_0$ and obtain $y_1(x) = y_0 + \int_{x_0}^{x} f(x, y_0)\, dx$

where $y_1(x)$ is the corresponding value of y(x) and is called first approximation of y(x) at any x. To determine still better approximation we replace y by y_1 in the above integral and obtain a second approximation y_2 as

$$y_2(x) = y_0 + \int_{x_0}^{x} f(x, y_1(x))\, dx.$$

Proceeding in this way, we obtain the n-th approximation y_n as

$$y_n(x) = y_0 + \int_{x_0}^{x} f(x, y_{n-1}(x))\, dx.$$

Thus we arrive at a sequence of approximate solutions $y_1(x), y_2(x), \ldots, y_n(x)$

Example 1: Using Picard's method of successive approximation, find the third approximation of the solution of the equation

$\frac{dy}{dx} = x + y^2$, where $y = 0$ when $x = 0$.

Solution : The given problem is

$$y' = x + y^2,\ y(0) = 0$$

We know that n-th approximation y_n of the initial value problem $y' = f(x,y)$, $y(x_0) = y_0$ is given by

$y_n(x) = y_0 + \int_{x_0}^{x} f(x, y_{n-1}(x))\, dx \quad -(1)$

Here $f(x,y) = x + y^2$, $x_0 = 0$ and $y_0 = 0$.

So from (1) we get, $y_n = \int_0^x (x + {y_{n-1}}^2)\, dx \quad -(2)$

Putting n = 1 in (2), we obtain the first approximation as

$$y_1 = \int_0^x (x + y_0^2)\, dx$$

$$= \int_0^x x\, dx = \frac{x^2}{2}$$

Putting n = 2 in (2), we obtain the second approximation as

$$y_2 = \int_0^x (x + y_1^2)\, dx$$

$$= \int_0^x (x + \frac{x^4}{4}) dx$$

$$= \frac{x^2}{2} + \frac{x^5}{20}$$

Putting n = 3 in (2), we obtain the third approximation as

$$y_3 = \int_0^x (x + y_2^2)\, dx$$

$$= \int_0^x \left(x + \frac{x^4}{4} + \frac{x^7}{20} + \frac{x^{10}}{400}\right) dx$$

$$= \frac{x^2}{2} + \frac{x^5}{20} + \frac{x^8}{160} + \frac{x^{11}}{4400}$$

Hence the third approximation of the solution of the given equation is

$$y_3 = \frac{x^2}{2} + \frac{x^5}{20} + \frac{x^8}{160} + \frac{x^{11}}{4400}.$$

Example 2: Using Picard's method of successive approximation, obtain the third approximation of the solution of the initial value problem

$$y' = 2 - \frac{y}{x}, \quad y(1) = 2.$$

Solution: Given problem is

$$y' = 2 - \frac{y}{x}, \quad y(1) = 2.$$

We know that the n-th approximation y_n of the initial value problem $y' = f(x, y)$, $y(x_0)$ is given by

$$y_n(x) = y_0 + \int_{x_0}^{x} f(x, y_{n-1}(x))\, dx \quad -(1)$$

Here $f(x,y) = 2 - \frac{y}{x}$, $x_0 = 1$ and $y_0 = 2$.

So from (1) we get,

$$y_n = 2 + \int_1^x \left(2 - \frac{y_{n-1}}{x}\right) dx \quad -(2)$$

Putting n = 1 in (2), we obtain the first approximation as

$$y_1 = 2 + \int_1^x \left(2 - \frac{y_0}{x}\right) dx$$

$$= 2 + \int_1^x \left(2 - \frac{2}{x}\right) dx$$

$= 2 + [2x-2\log x]_1^x = 2 + 2x - 2\log x - 2 + 0 = 2x - 2\log x$

Putting n = 2 in (2), we obtain the second approximation as

$$y_2 = 2 + \int_1^x \left(2 - \frac{y_1}{x}\right) dx$$

$$= 2 + \int_1^x \left(2 - \frac{2x - 2\log x}{x}\right) dx$$

$$= 2 + \int_1^x 2\frac{\log x}{x} dx = 2 + (\log x)^2$$

Putting n = 3 in (2), we obtain the third approximation as

$$y_3 = 2 + \int_1^x \left(2 - \frac{y_2}{x}\right) dx$$

$$= 2 + \int_1^x \left(2 - \frac{2 + (\log x)^2}{x}\right) dx$$

$$= 2 + [2x - 2\log x - \frac{1}{3}(\log x)^3]_1^x = 2x - 2\log x - \frac{1}{3}(\log x)^3$$

Hence the third approximation of the solution of the given equation is

$$y_3 = 2x - 2\log x - \frac{1}{3}(\log x)^3$$

The Lipschitz's Condition:

Let f be a function defined for (x,y) in a set S. We say f satisfy Lipschitz's Condition on S if there exists a positive constant k such that

$$|f(x,y_1) - f(x,y_2)| \leq k\,|y_1 - y_2| \quad \text{for all } (x, y_1), (x, y_2) \in S.$$

The constant k is called Lipschitz's constant.

Note: If f is continuous and satisfies Lipschitz's Condition on the rectangle R, then the successive approximation converge to a solution of the IVP $\frac{dy}{dx} = f(x, y)$, $y(x_0) = y_0$ in $|x-x_0| \leq \alpha$.

Theorem 3.2:

Suppose that S is either a rectangle $|x-x_0| \leq a$, $|y-y_0| \leq b$, a, b > 0 or a strip $|x-x_0| \leq a$, $|y| < \propto$, a > 0, and that f is a real-valued function defind on S such that $\frac{\partial f}{\partial y}$ exists, is continuous on S, and $|\frac{\partial f}{\partial y}(x,y)| \leq k$ for all $(x,y) \in s$, k > 0 is a constant. Then f satisfies Lipschitz's Condition on S with Lipschitz's constant k.

Proof:

By the M.V.T. for two distinct points (x, y_1) and (x, y_2), we have

$$|f(x,y_1) - f(x,y_2)| = |\frac{\partial f}{\partial y}(x, y^*)|\,|y_1 - y_2|$$

where y^* lies in between y_1 and y_2. Since $|\frac{\partial f}{\partial y}(x,y)| \leq k$ for all $(x, y) \in S$, from above we obtain

$$|f(x,y_1) - f(x,y_2)| \leq k\,|y_1 - y_2| \ .$$

Hence f satisfies Lipschitz's Condition on S with Lipschitz's constant k.

Example 3: Show that the function $f(x,y) = xy^2$ satisfies Lipschitz's Condition on a rectangle R given by R: $|x| \leq 1$, $|y| \leq 1$. What happen if R is given by $|x| \leq 1$, $|y| < \propto$.

Solution: Here $f(x,y) = xy^2$

Therefore $|\frac{\partial f}{\partial y}(x,y)| = |2xy| \leq 2$ for $(x, y) \in R$

Hence f satisfies Lipschitz's Condition on R with Lipschitz's constant 2.

Since on the strip S : | x | ≤ 1, | y |< ∞ we have

$|\frac{f(x,y_1)-f(x,0)}{y_1-0}|$ = | x ||y_1| → ∞ as |y_1| → ∞, if |x| ≠ 0, the above function does not satisfy Lipschitz's Condition on S.

Theorem 3.3: Picard's Existence and Uniqueness Theorem:

Let $(X_0,Y_0) \in R^2$ and let f be a real-valued continuous function defined on the closed rectangle Q = {((x, y) : |x-x_0| ≤ a, |y-y_0| ≤ b, a, b > 0} in R^2. Further assume that f(x,y) satisfies Lipschitz's Condition with respect to y in Q, i.e., there exists a positive constant k such that |f(x,y_1) – f(x,y_2) | ≤ k |y_1 – y_2 | for all (x,y_i) ϵ Q, i = 1, 2. Let h = min{a, $\frac{b}{M}$}, where |f(x,y) | ≤ M for all (x,y) ϵ Q. Then the initial value problem

$\frac{dy}{dx}$ = f (x, y), y(x_0) = y_0

has a unique solution in |x-x_0| ≤ h .

Proof:

Since f(x, y′) is continuous in Q, the initial value problem

$\frac{dy}{dx}$ = f (x, y), y(x_0) = y_0 ..... (1)

is equivalent to the integral equation

y(x) = y_0 + $\int_{x_0}^{x} f(t,\ y)$ dt ..... (2)

We have to find out a unique continuous solution of (2). We define a sequence of function {y_n(x)} as follows :

y_0(x) = y_0

y_1(x) = y_0 + $\int_{x_0}^{x} f(t,\ y_0(t))$ dt

y_2(x) = y_0 + $\int_{x_0}^{x} f(t,\ y_1(t))$ dt

: : : :

y_n(x) = y_0 + $\int_{x_0}^{x} f(t,\ y_{n-1}(t))$ dt

Now, for all |x-x_0| ≤ h, we have

| y_1(x) - y_0 | = | $\int_{x_0}^{x} f(t,\ y_0(t))$ dt

≤ $\int_{x_0}^{x}$ | $f(t,\ y_0(t))$ dt |

≤ M$\int_{x_0}^{x}$| dt |

= M |x-x_0| ≤ Mh ≤ b

Hence (x, y_1) ϵQ. By induction we assume that (x, y_{n-1}) ϵQ.

Then | y_n(x) - y_0 | ≤ $\int_{x_0}^{x}$ |$f(t,\ y_{n-1}(t))$ dt |

≤ M$\int_{x_0}^{x}$| dt |

= M |x-x_0| ≤ Mh ≤ b

Hence (x, y_n) ϵ Q for all n.

Now, for all |x-x_0| ≤ h, we get

| y_n(x) – y_{n-1}(x) | ≤ $\int_{x_0}^{x}$ |$f(t,\ y_{n-1}(t))$ - $f(t,\ y_{n-2}(t))$ | | dt |

$$\leq k \int_{x_0}^{x} | y_{n-1}(t) - y_{n-2}(t) | | dt | \quad \ldots\ldots \quad (3)$$

where k is a Lipschitz's constant.

Taking n = 1, 2,..., we obtain

$$|y_1(x) - y_0(x)| \leq \int_{x_0}^{x} |f(t, y_0(t))\, dt|$$

$$\leq M \int_{x_0}^{x} | dt |$$

$$= M |x-x_0|$$

$$|y_2(x) - y_1(x)| \leq k \int_{x_0}^{x} |y_1(t) - y_0(t)| |dt|$$

$$\leq k M \int_{x_0}^{x} | t - x_0(t) | | dt |$$

$$= k M \frac{|x-x_0|^2}{2!}$$

By induction, we assume that

$$|y_n(x) - y_{n-1}(x)| \leq \frac{Mk^{n-2}}{(n-1)!} | x-x_0|^{n-1}$$

Then $|y_n(x) - y_{n-1}(x)| \leq k \frac{Mk^{n-2}}{(n-1)!} \int_{x_0}^{x} | t - x_0|^{n-1} | dt |$

$$= \frac{Mk^{n-2}}{(n-1)!} \frac{|x-x_0|^n}{n}$$

$$= \frac{M}{k} \frac{k^n |x-x_0|^n}{n!}$$

Since the series $\sum_{n=1}^{\infty} \frac{M}{k} \frac{k^n |x-x_0|^n}{n!}$ converges uniformly to $\frac{M}{k} [e^{k|x-x_0|}-1]$, the series

$$y_0(x) + \sum_{n=1}^{\infty} (y_n(x) - y_{n-1}(x)) \quad \ldots\ldots \quad (4)$$

converges absolutely in $|x-x_0| \leq h$.

Now, $y_n(x) = y_0(x) + [y_1(x) - y_0(x)] + [y_2(x) - y_1(x)] + \ldots + [y_n(x) - y_{n-1}(x)]$

= a partial sum of the series (4)

and so the sequence $\{y_n(x)\}$ also converges uniformly in $|x-x_0| \leq h$.

Let $\lim_{n\to\infty} y_n(x) = y(x), |x - x_0| \leq h$

Since each $y_n(x)$ is continuous is $|x-x_0| \leq h$, by construction, and since the convergence is uniform, the limit function y(x) is also continuous in $|x-x_0| \leq h$.

We shall show that y(x) is the desired continuous solution of the equation (2).

Since (x,y(x)), (x, $y_{n-1}(x)$) ϵQ, we have

$$\left|\int_{x_0}^{x} [f(t, y_{n-1}(t)) - f(t, y(t))] dt\right|$$

$$\leq \int_{x_0}^{x} |f(t, y_{n-1}(t)) - f(t, y(t))] dt|$$

$$\leq k \int_{x_0}^{x} |y_{n-1}(t) - y(t)| |dt|$$

$\to 0$ as $n \to \infty$

$$\therefore \int_{x_0}^{x} f(t, y_{n-1}(t)) dt \to \int_{x_0}^{x} f(t, y(t))] dt| \text{ as } n \to \infty$$

Taking limit as $n \to \infty$ in $y_n(x) = y_0 + \int_{x_0}^{x} f(t, y_{n-1}(t)) dt$, we obtain $y(x) = y_0 + \int_{x_0}^{x} f(t, y(t)) dt$

This proves that y(x) is a continuous solution of the integral equation (2) and that of initial value problem (1).

We now show that this continuous solution y(x) is unique.

If possible, let y(x) and z(x) be the two continuous solution of (1) satisfying the initial conditions $y(x_0) = y_0$ and $z(x_0) = y_0$ respectively.

Then we have $y(x) = y_0 + \int_{x_0}^{x} f(t, y(t))dt$

$z(x) = y_0 + \int_{x_0}^{x} f(t, z(t))dt$

Now, $|y(x) - z(x)| \leq \int_{x_0}^{x} |f(t, y(t)) - f(t, z(t))||dt|$

$\leq k \int_{x_0}^{x} |y(t) - z(t)||dt| \quad \ldots\ldots \quad (5)$

as $(x, y(x)) \in Q$ and $(x, z(x)) \in Q$.

Since $y(t) - z(t)$ is continuous in $|x-x_0| \leq h$, there exists a positive constant M_1, say, such that

$|y(t) - z(t)| \leq M_1$, for all t in $|x-x_0| \leq h$

Hence from (5) we get,

$|y(x) - z(x)| \leq M_1 k \int_{x_0}^{x} |dt|$

$= M_1 k |x-x_0|$

Substituting this in (5), we obtain

$|y(x) - z(x)| \leq M_1 k^2 \int_{x_0}^{x} |t - x_0||dt|$

$= M_1 k^2 \frac{|x-x_0|^2}{2!}$

Continuing this process we obtain at the n-th stage

$|y(x) - z(x)| \leq M_1 \frac{k^n |x-x_0|^n}{n!} \quad \ldots\ldots \quad (6)$

Since the series

$\sum_{n=0}^{\infty} \frac{k^n |x-x_0|^n}{n!}$ is convergent, we have $\frac{k^n |x-x_0|^n}{n!} \to 0$ as $n \to \infty$.

Therefore from (6), we obtain

$|y(x) - z(x)| = 0 \ \forall$ x in $|x-x_0| \leq h$

i.e., $y(x) = z(x) \ \forall$ x in $|x-x_0| \leq h$

Thus the solution is unique. This proves the theorem.

Example 4: If R is defined by the rectangle $|x| \leq a$, $|y| \leq b$, show that the function $f(x,y) = x \sin y + y \cos x$ satisfies the Lipschitz's Condition in R with Lipschitz's constant a + 1.

Solution: $f(x,y) = x \sin y + y \cos x$

and $\left|\frac{\partial f}{\partial y}\right| = |x \cos y + \cos x|$

$\leq |x||\cos y| + |\cos x|$

$\leq |x| + 1$

$\leq a + 1$

Hence f satisfies Lipschitz's Conditions in R with Lipschitz's constant a+1.

Gronwall's Lemma:

Let u and v be two real-valued positive continuous functions defined in the interval $[t_0, t_0 + h]$. Also suppose that

$u(t) \le c + \int_{t_0}^{t} u(s)v(s)ds \ \forall\ t \in [t_0, t_0 + h], c \ge 0$ is a constant.

Then $u(t) \le c \exp\left[\int_{t_0}^{t} v(s)ds\right]$

Proof:

We consider the following two cases separately.

Case -1: Let $c > 0$

Let $U(t) = c + \int_{t_0}^{t} u(s)v(s)ds \ \forall t \in [t_0, t_0 + h]$

Cleary $U(t) > 0$ and $u(t) \le U(t) \quad \forall t \in [t_0, t_0 + h]$

Since $\frac{dU}{dt} = u(t)v(t)$, it follows that

$\frac{d\,U(t)}{U(t)} = \frac{u(t)v(t)}{U(t)}\ dt \le v(t)dt$

i.e., $\frac{d}{dt}\ [log\ U(t)] \le v(t),\ t_0 \le t \le t_0 + h$

Integrating over $[t_0, t]$, we get

$$\int_{t_0}^{t} \frac{d}{ds}\ [\log U(s)ds \le \int_{t_0}^{t} v(s)ds$$

i.e., $log \frac{U(t)}{U(t_0)} \le \int_{t_0}^{t} v(s)ds$

i.e, $U(t) \le U(t_0) \exp[\int_{t_o}^{t} v(s)ds]$

Since $U(t_0) = c$ and $u(t) \le U(t)$ for all $t \in [t_0, t_0 + h]$,

we get $u(t) \le c \exp[\int_{t_0}^{t} v(s)ds]$

Case -2: Let $c = 0$

We define $c_p = \frac{1}{p}, p > 0$

Then $c > c_p$

Now by the hypothesis, $(t) \le c_p + \int_{t_0}^{t} v(s)ds$, where $c_p > 0$

So by Case 1, we obtain

$$u(t) \le c_p \exp\ [\int_{t_0}^{t} v(s)ds]$$

$$= \frac{1}{p}\ exp[\int_{t_0}^{t} v(s)ds]$$

Taking limit $p \to \infty$, we get

$$u(t) = 0 = cexp\ [\int_{t_0}^{t} v(s)ds]$$

This completes the proof of the theorem.

Application of Gronwall's Lemma :

Suppose that y(x) and z(x) are two solutions of the initial value

$\frac{dy}{dx} = f(x,y), y(x_0) = y_0 \; on \; [x_0, x_0 + h]$.

Then (x, y(x)), (x, z(x)) ϵ Q and also

$\frac{dy(x)}{dx} = f\big(x, y(x)\big) \quad , y(x_0) = y_0$

$\frac{dz(x)}{dx} = f\big(x, z(x)\big) \quad , z(x_0) = y_0$

Integrating from x_0 to x we get

$$y(x) = y(x_0) + \int_{x_0}^{x} f\big(t, y(t)\big)dt$$

and $z(x) = z(x_0) + \int_{x_0}^{x} f\big(t, z(t)\big)dt$

This give $|y(x) - z(x)| = \left| \int_{x_0}^{x} f\big(t, y(t)\big)dt - \int_{x_0}^{x} f\big(t, z(t)\big)dt \right|$

$$\leq \int_{x_0}^{x} \left| f\big(t, y(t)\big) - f\big(t, z(t)\big) \right| dt$$

$$\leq k \int_{x_0}^{x} |y(t) - z(t)| dt$$

where k is a Lipchitz's constant.

We define $u(x) = |y(x) - z(x)|$, $v(x) = k$ and $c = 0$, for $x_0 \leq x \leq x_0 + h$

Then $u(x) \leq c + \int_{x_0}^{x} u(t)v(t)dt$

So by Gronwall's lemma, we obtain

$$u(x) \leq c \; exp \left[\int_{x_0}^{x} v(t)dt\right]$$

$$= c \; exp \left[\int_{x_0}^{x} k \; dt\right]$$

$$= c \exp[k|x - x_0|]$$

Hence $u(x) = |y(x) - x(x)| = 0 \quad \forall \; x \in [x_0, x_0 + h]$

Hence $u(x) \equiv z(x) \forall \; x \in [x_0, x_0 + h]$ and the solution is unique.

<u>Fundamental Inequality</u>

Let y(x) and z(x) be two solutions of the differential equation $\frac{dy}{dx} = f(x)$.

Then $|y(x) - z(x)| \leq |y(x_0) - z(x_0)| \exp\left[\int_{x_0}^{x} k dt\right]$ in $[x_0, x_0 + h]$,

assuming that f satisfied Lipschitz's Condition in Q with Lipschitz's constant K. This is known as fundamental inequality.

Proof: We have,

$$\frac{dy(x)}{dx} = f(x, y(x))$$

$$and\ \frac{dz(x)}{dx} = f(x, z(x))$$

Integrating from x_0 to x, we obtain

$$y(x) = y(x_0) + \int_{x_0}^{x} f(t, y(t))dt$$

$$and\ z(x) = z(x_0) + \int_{x_0}^{x} f(t, z(t))dt$$

This gives $|y(x) - z(x) \leq |y(x_0) - z(x_0)| + \int_{x_0}^{x} |f(t, y(t)) - (t, z(t))|dt$

$$\leq |y(x_0) - z(x_0)| + k \int_{x_0}^{x} |y(t) - (t)|dt$$

We define,

$u(x) = |y(x) - z(x)|$

$\vee(x) = k$

$c = |y(x_0) - z(x_0)| \quad in\ [x_0, x_0 + h]$

Now , applying Gronwall's lemma, we obtain

$$|y(x) - z(x)| \leq |y(x_0) - z(x_0)| \exp[\int_{x_0}^{x} kdt]$$

Example 5: Show that $\frac{dy}{dx} = f(x, y) = y^{2/3},\ y(0) = 0$ has solutions but not unique.

Solution: The given equation is $\frac{dy}{dx} = y^{2/3}$

$$i.e., \quad \frac{dy}{y^{2/3}} = dx$$

On integration, $3y^{1/3} = x + c$, where c is a constant. Now, $y(0) = 0,\ \therefore c = 0$

Thus we get

$x = 3y^{1/3}$

Therefore $x = 3y^{1/3}$ is a solution of the given equation.

We note that y = 0 is also a solution of the given equation. We show that this happens due to the lack of Lipschitz's Condition. We choose

$y_1 = \frac{1}{n}$, $y_2 = 0$

Then $|f(x_1, y_1) - f(x_2, y_2)| = |y_1^{2/3} - y_2^{2/3}| = n^{1/3}|\frac{1}{n^{2/3}} - 0| = n^{1/3}|\frac{1}{n} - 0| = n^{1/3}|y_1 - y_2|$

For sufficiently large n, $n^{1/3}$ is greater than any finite number. Hence there is not positive k such that

$$| f(x, \frac{1}{n}) - f(x, 0) | \leq k | \frac{1}{n} - 0|$$

and so Lipschitz's Condition is not satisfied.

Note: The above example shows that the continuity of f(x,y) is not enough to ensure the uniqueness of the solution of the initial value problem

$\frac{dy}{dx} = y^{2/3}$, y(0) = 0.

Example 6: For the initial value problem $\frac{dy}{dx} = y^2 + cos^2 x,\ y(0) = 0$, determine the interval of existence of its solution given that R is the rectangle containing origin, i.e., R = {(x,y) : 0 ≤ x ≤ a, | y | ≤ b, a > $\frac{1}{2}$, b > 0}.

Solution: Here, f(x,y) = $y^2 + cos^2 x$

Now, | f(x,y) | = $|y^2 + cos^2 x|$ ≤ 1 + b^2 = M, say

Since $|\frac{\partial f}{\partial y}(x, y)| = |2y| \leq 2b$, we see that f(x,y) satisfies Lipschitz's Condition with Lipschitz's constant 2b.

∴ By Picard's existence and uniqueness theorem, the solution y(x) exists in the interval 0 ≤ x ≤ h, where

h = min{a, $\frac{b}{M}$}

= min {a, $\frac{b}{b^2+1}$}

Now, $\frac{b}{b^2+1} = \frac{1}{b+\frac{1}{b}} = \frac{1}{(\sqrt{b}-\frac{1}{\sqrt{b}})^2+2} \leq \frac{1}{2}$

Hence h = min{a, $\frac{b}{b^2+1}$} ≤ $\frac{1}{2}$

Hence y(x) exists in the interval 0 ≤ x ≤ $\frac{1}{2}$.

Example 7: Consider the initial value problem

$\frac{dy}{dx} = y^2$, y(0) = 2.

Let R be the rectangle given by

R = {(x,y) : | x | ≤ a, | y – 2 | ≤ b, a > 0, b >0}

Find the largest interval of existence of its solution.

Solution: Here, the given initial value problem is

$\frac{dy}{dx} = y^2$, y(0) = 2.

Let f(x,y) = y^2

So, , | f(x,y) | = $|y^2|$ ≤ $(2 + b)^2$ = M, say

Also, $\frac{\partial f}{\partial y}$ = 2 y

So, $|\frac{\partial f}{\partial y}|$ =2 | y | ≤ 2(2+b) [∵ b > 0]

We see that f(x,y) satisfies Lipschitz's Condition with Lipschitz's constant 2 (2+b).

Therefore by Picard's existence and uniqueness theorem, the solution y(x) exists in the interval 0 ≤ x ≤ h, where

h = min{a, $\frac{b}{M}$} = min {a, $\frac{b}{(b+1)^2}$}

Now, $\frac{b}{(b+2)^2} = \frac{b}{b^2+4b+4} = \frac{b}{b(b+4+\frac{4}{b})} = \frac{1}{b+4+\frac{4}{b}} = \frac{1}{(\sqrt{b}-\frac{2}{\sqrt{b}})^2+8} \leq \frac{1}{8}$

Hence y(x) exists in the interval of $0 \leq x \leq \frac{1}{8}$.

Unit 5 & Unit 6

Sturm Liouville Problem

Adjoint Equation:

Let L be a differential operation of the n-th order defined by

$L[u] = P_0(x)\frac{d^n u}{dx^n} + P_1(x)\frac{d^{n-1}u}{dx^{n-1}} + \quad ... + p_{n-1}(x)\frac{du}{dx} + p_n(x)u.$

Then the differential operation $\bar{L}$ or L^+ defined by

$\bar{L}[v] = (-1)^n \frac{d^n}{dx^n}(p_0 v) + (-1)^{n-1}\frac{d^{n-1}}{dx^{n-1}}(p_1 v) + \ldots + (-1)^n \frac{d}{dx}(p_{n-1}v) + p_n v$

is called the adjoint of L and the equation $\bar{L}[v] = 0$ is called the adjoint equation of $L[u] = 0$.

Particular case:

The adjoint equation of the second order homogeneous linear differential equation

$L[u] = p_0\frac{d^2u}{dx^2} + p_1\frac{du}{dx} + p_2 u = 0 \quad \ldots\ldots \quad (1)$

is the differential equation

$\bar{L}[v] = (-1)^2\frac{d^2}{dx^2}(p_0 v) + (-1)\frac{d}{dx}(p_1 v) + p_2 v = 0$

i.e., $p_0 v'' + (2p_0' - P_1)v' + (p_0'' - p_1' + p_2)v = 0 \quad \ldots\ldots \quad (2)$

The differential operator $\bar{L}$, where

$\bar{L} = (-1)^2\frac{d^2}{dx^2}(P_0) - \frac{d}{dx}(P_1) + P_2$

is called the adjoint of the linear operator

$L \equiv P_0\frac{d^2}{dx^2} + P_1\frac{d}{dx} + P_2$

Note: We note that the adjoint equation of (2) is the differential equation (1) for if we write (2) in the form $q_0 v'' + q_1 v' + q_2 v = 0$, then the adjoint of (2) is differential equation

$(q_0 w)'' - (q_1 w)' + q_2 w = 0$

i.e., $q_0 w'' + (2q_0' - q_1)w' + (q_0'' - q_1' + q_2)w = 0$

i.e., $p_0 w'' + (2p_0' - 2p_0' + p_1)w' + (p_0'' - 2p_0'' + p_1' + p_0'' - p_1' + p_2)w = 0$

i.e., $p_0 w'' + p_1 w' + p_2 w = 0.$

Self Adjoint Equation:

Homogeneous linear differential equations which coincide with their adjoint are called self adjoint.

From (1) and (2), we see that for the differential equation (1) to be self-adjoint, it is necessary that

$2p_0' - p_1 = p_1$

i.e., $p_0' = p_1$

The condition is also sufficient for then

$p_0'' - p_1' + p_2 = p_1' - p_1' + p_2 = p_2$

Theorem 4.1: The second order linear homogeneous differential equation

$p_0u'' - p_1u' + p_2u = 0$

is self-adjoint if and only if it has the form

$\frac{d}{dx}(p(x)\frac{du}{dx}) + q(x)\,u = 0.$

Proof:

It is known that the given differential equation is self-adjoint if and only if $p_0' = p_1$.

On substitution it follows that

$p_0u'' + p_0'u' + p_2u = 0$

i.e., $\frac{d}{dx}(p_0\frac{du}{dx}) + p_2u = 0,$

which is the given form.

Theorem 4.2: The second order homogeneous differential equation $p_0u'' + p_1u' + p_2u = 0$ can be reduced to self-adjoint form by multiplying throughout by factor $h(x) = \frac{1}{p_0}\exp[\int \frac{p_1}{p_0}\,dx]$

Proof:

Multiplying by $H(x)/p_0$ $[H(x)\neq 0]$ throughout we get from given differential equation

$H(x)u'' + H(x)\frac{p_1}{p_0}u' + H(x)\frac{p_2}{p_0}u = 0$

This will be self- adjoint if and only if

$H'(x) = H(x)\frac{p_1}{p_0}$

i.e., $\frac{H'(x)}{H(x)} = \frac{p_1}{p_0},$

which on integrating gives

$H(x) = \exp[\int \frac{p_1}{p_0}\,dx]$

$\therefore h(x) = \frac{H(x)}{p_0} = \frac{1}{p_0}\exp[\int \frac{p_1}{p_0}\,dx]$

Example 1: Obtain the adjoint equation of the differential equation

$(1+x^2)\frac{d^2u}{dx^2} + x\frac{du}{dx} - 4u = 0.$

Solution: The given differential equation is

$(1+x^2)\frac{d^2u}{dx^2} + x\frac{du}{dx} - 4u = 0$

Therefore its adjoint equation will be

$(-1)^2\frac{d^2}{dx^2}((1+x^2)v) + (-1)\frac{d}{dx}(xv) - 4v = 0$

i.e., $(1+x^2)v'' + 4xv' + 2v - xv' - v - 4v = 0$

i.e.,$(1+x^2)v'' + 3xv' - 3v = 0.$

Example 2: Reduce the following equation to self-adjoint form.

$(1-x^2)\,u''-xu'+\lambda u = 0.$

Solution: Comparing the given equation with the standard form

$p_0u''+ p_1u'+ p_2u = 0,$

we get

$p_0 = 1-x^2,\ p_1 = -x,\ p_2 = \lambda$

So, $h(x) = \frac{1}{p_0} \exp\left[\int \frac{p_1}{p_0}\, dx\right]$

$= \frac{1}{1-x^2} \exp\left[\int \frac{-x}{1-x^2}\, dx\right]$

$= \frac{1}{1-x^2} \exp\left[\frac{1}{2} log|1-x^2|\right]$

$= \frac{1}{\sqrt{1-x^2}}$

Multiplying the given equation throughout by h(x), we get

$\sqrt{1-x^2}\ u'' - \frac{x}{\sqrt{1-x^2}}\ u' + \frac{\lambda}{\sqrt{1-x^2}}\ u = 0,$

which is obviously a self-adjoint equation.

Regular Sturm –Liouville Problem:

A second order Sturm –Liouville problem is a homogeneous boundary value problem of the form

$$\frac{d}{dx}\left(p(x)\frac{dy}{dx}\right) + \left(q(x) + \lambda r(x)\right)y = 0 \quad \text{——} (1)$$

together with boundary conditions

$$\left.\begin{aligned} A_1\, y(a) + B_1 y'(a) = 0 \\ A_2\, y(b) + B_2 y'(b) = 0 \end{aligned}\right\} \quad \text{——} (2)$$

Where $p(x), q(x), r(x)\ and\ p'(x)$ are real valued continuous functions on [a, b] and λ is a parameter.

Also A_1, B_1, A_2, B_2 are any constants such that A_1 and B_1 are not both zero and A_2 and B_2 are not both zero.

If $p(x)$ and $r(x)$are positive for all x in [a. b], then equation (1) together with with boundary condition (2) is called regular Sturm-Liouville problem. If the conditions "$p(x)$ and $r(x)$ are positive" are not satisfied, then the problem is known as Singular S-L problem.

Remark:

Let $p(x) = 1 = r(x)\ and\ q(x) = 0$ in (1).
Also let $A_1 = A_2 = 1,\ B_1 = B_2 = 0$ in (2). Then (1) and (2) reduces to $y'' + \lambda y = 0, y(a) = 0 = y(b)$.

This is the simplest form of regular S-L problem.

Definition:

A non-trivial solution of a regular Sturm - Liouville problem is called an eigen function and the corresponding λ is called its eigen value.

Definition:

Two integrable functions $f(x)$and $g(x)$ are said to be orthogonal with weight function $\pi(x)$ on an interval I = [a, b], if $\int_a^b \pi(x)f(x)g(x) = 0$.

Conversion of a 2nd order liner differential equation to S-L form

Let us consider a second order linear differential equation of the form

$$p_0(x)\frac{d^2y}{dx^2} + p_1(x)\frac{dy}{dx} + p_2(x)y + \lambda p_3(x)y = 0$$

Where $p_0(x)(\neq 0)\ and\ p_3(x)$ are positive in the interval where the problem is considered.

Multiplying the above equation by $(x) = \exp[\int \frac{p_1(x)}{p_0(x)}\,dx]$, we get

$$p_0(x)\left[I(x)\frac{d^2y}{dx^2} + I(x)\frac{p_1(x)}{p_0(x)}\frac{dy}{dx}\right] + I(x)p_2(x)y + I(x)\lambda p_3(x)y = 0$$

i.e., $p_0(x)\frac{d}{dx}\left[I(x)\frac{dy}{dx}\right] + I(x)p_2(x)y + I(x)\lambda p_3(x)y = 0$

$$i.e.,\quad \frac{d}{dx}\left[I(x)\frac{dy}{dx}\right] + I(x)\frac{p_2(x)}{p_0(x)}y + \lambda I(x)\frac{p_3(x)}{p_0(x)}y = 0.$$

Now, putting $(x) = I(x),\ q(x) = I(x)\frac{p_2(x)}{p_0(x)}\ and\ r(x) = I(x)\frac{p_3(x)}{p_0(x)}$, we get

$$\frac{d}{dx}\left[p(x)\frac{dy}{dx}\right] + [q(x) + \lambda r(x)]y = 0,$$

which is of Sturm-Liouville form.

Example 3: Find the eigen values and eigen functions of the differential equation

$$\frac{d^2y}{dx^2} + \lambda y = 0\ , y(0) = 0\ and\ y(\pi) = 0,\qquad where\ \lambda\ is\ a\ parameter.$$

Solution: If $\lambda = 0$, then $y'' = 0$

Therefore, $y(x) = Ax + B$.

Since $y(0) = 0\ and\ y(\pi) = 0$, we obtain A = B = 0.

Therefore $y(x) = 0$, a trivial solution.

If λ < 0, then the general solution of the given equation is

$y(x) = Ae^{\sqrt{-\lambda}x} + Be^{\sqrt{-\lambda}x}$

Using boundary conditions, we obtain A= B = 0.

Hence, we again obtain a trivial solution.

Finally, if λ > 0, then the general solution of the given equation become

$y(x) = A\ cos\sqrt{\lambda}x$ + B $\sin\sqrt{\lambda}x$

Now, $y(0) = 0 \Rightarrow A = 0$

and $y(\pi) = 0 \Rightarrow B\ Sin\ \sqrt{\lambda}\pi = 0$

Now, B ≠ 0; for, in that case solution becomes trivial.

So $\sin\sqrt{\lambda}\pi = 0$

i.e., $\sqrt{\lambda} =$ n, an integer

i.e., $\lambda = n^2$, n = 1, 2, 3, ...

Thus the solution of the given equation is given by y(x) = Sin nx, n = 1, 2, 3, ... The values of λ namely 1, 4, 9, 16 ... are called the eigen values and the corresponding solutions $sinx,\ sin2x,\ sin3x$ ... are called the eigen functions.

The general solution of the given equation becomes

$$y = \sum_{n=1}^{\infty} a_n \sin nx \qquad \text{—— (1)}$$

where a_n are arbitrary constants.

Since $\int_0^{\pi} \sin mx \, sin \, nx \, dx = 0$ if $m \neq n$

$$= \frac{\pi}{2} \text{ if } m{=}n$$

It follows that

$$\int_0^{\pi} y(x) \sin mx \, dx = an \int_0^{\pi} \sin mx \, \text{Sin} \, dx$$

$$\text{i. e.}, a_n = \frac{2}{\pi} \int_0^{\pi} y(x) \, sin \, nx \, dx, n = 1, 2, 3, .. [\because m = n] \ldots\ldots \text{ (2)}$$

Hence (1) is the general solution of the given equation where the coefficient a_n are calculated from (2).

Theorem 4.3: Sturm Separation Theorem:

Statement: Let $f(x)$and $g(x)$ be two linearly independent solutions of

$u''(x) + p(x)u'(x) + q(x)\,u(x) = 0.$

Then the zeros of $f(x)$ and $g(x)$ occur alternatively.

Proof:

Let $g(x)$ vanishes at x = x_i. Since f(x) and g(x) are linearly independent solutions of the given equation, the Wronskian

$$W(f, g; x_i) = \begin{vmatrix} f(x_i) & g(x_i) \\ f'(x_i) & g'(x_i) \end{vmatrix} = f(x_i)g'(x_i) \neq 0 \; if \; g(x_i) = 0$$

This shows that $f(x_i) \neq 0 \; and \; g'(x_i) \neq 0.$

If x_1 and x_2 are two consecutive zeros of g(x), then $g'(x_1),\ g'(x_2),\ f(x_1) \; and \; f(x_2)$ are all non-zero. Moreover, $g'(x_1) \; and \; g'(x_2)$ cannot have the some sign, because if the function $g(x)$ is decreasing at x = x_1, then it must be increasing at $x = x_2$ and vice –versa. Since $w(f, g;\ x_i)$ has constant sign, it follows that $f(x)$ must vanish somewhere between x_1 and x_2. This proves the theorem.

Theorem 4.4: Sturm Comparison theorem:

Statement: Let $f(x)$ and $g(x)$ be two non-trivial solution of the differential equations

$u''(x) + p(x)\,u(x) = 0$ and $v''(x) + q(x)\,v(x) = 0$

respectively where $p(x) > q(x)$. Then $f(x)$ vanishes at least once in between two consecutive zeros of g(x).

Proof:

Let x_1 and x_2 be two consecutive zeros of g(x), i. e, $g(x_1) = g(x_2) = 0$ and $f(x) \neq 0$ at any point in the open interval (x_1, x_2). Without any loss of generality, we may assume that both $f(x)$ and $g(x)$ are positive in (x_1, x_2), for either of the functions can be replaced by its negative. Since g(x) is a non-trivial solution, it follows that $g'(x_1) \neq 0$ and also $g'(x_2) \neq 0$.

Again, since g(x) is positive in (x_1, x_2), it follows $g'(x_1) > 0$ and $g'(x_2) < 0$.

So, $W(f, g;\ x_1) = \begin{vmatrix} f(x_1) & g(x_1) \\ f'(x_1) & g'(x_1) \end{vmatrix} = f(x_1)g'(x_1) > 0$

and $W(f, g;\ x_2) = \begin{vmatrix} f(x_2) & g(x_2) \\ f'(x_2) & g'(x_2) \end{vmatrix} = f(x_2)g'(x_2) < 0.$

Now, $\frac{d}{dx} W(f,\ g;\ x) = \frac{d}{dx}\left(f(x)\, g'(x) - f'(x)g(x)\right)$

$= f(x)$ g''(x)-f''$(x)g(x)$

$= f(x)g(x)[p(x) - q(x)] > 0$ in (x_1, x_2).

Hence, W(f, g; x) is strictly increasing in (x_1, x_2) and so W(f, g; x_2) > W(f, g; x_1), a contradiction. Thus f(x) has at least one zero in between two consecutive zeros of g(x).

Theorem 4.5:

The eigen functions of the Regular Sturm-Liouville problem

$$\frac{d}{dx}\left[p(x)\frac{dy}{dx}\right] + \{q(x) + \lambda r(x)\}y = 0,$$

with boundary conditions

$$A_1 y(a) + B_1 y'(a) = 0$$

and $A_2 y(b) + B_2 y'(b) = 0$

are orthogonal in [a, b] with weight function $r(x)$.

Proof: Let u(x) and v(x) are eigen functions of the given problem with eigen values λ and μ respectively.

Then $\frac{d}{dx}\left(p(x)u'(x)\right) + \left(q(x) + \lambda r(x)\right)u(x) = 0$ —— (1)

and $\frac{d}{dx}\left(p(x)v'(x)\right) + \left(q(x) + \mu r(x)\right)v(x) = 0$ —— (2)

Now, (2) x u(x) – (1) x v(x) gives

$$u(x)\frac{d}{dx}\left(p(x)v'(x)\right) - v(x)\frac{d}{dx}(p(x)u'(x)) = (\lambda - \mu)\, r(x)u(x)\, v(x)$$

i.e, $\frac{d}{dx}[u(x)\left(p(x)v'(x)\right) - v(x)\left(p(x)\, u'(x)\right)] = (\lambda - \mu)\, r(x)u(x)\, v(x)$

Integrating from a to b, we get

$$(\lambda - \mu)\int_b^a r(x)u(x)\, v(x)dx = [u(x)p(x)v'(x) - v(x)p(x)u'(x)]_a^b$$

$$= [u(b)p(b)v'(b) - v(b)p(b)u'(b)]$$

$$- [u(a)p(a)\, v'(a) - v(a)p(a)u'(a)]$$

$$= \left[-\frac{B_2}{A_2}\, u'(b)p(b)\, v'(b) + \frac{B_2}{A_2}\, v'(b)p(b)u'(b)\right]$$

$$-[-\frac{B_1}{A_1}\, u'(a)p(a)\, v'(a) + \frac{B_1}{A_1}\, v'(a)\, p(a)\, u'(a)]$$

$= 0$ (using boundary condions)

Since $(\lambda - \mu) \neq o$ it follows that

$$\int_a^b r(x)u(x)v(x)dx = 0$$

This completes the proof.

Note:

(1) The eigen values of a Sturm –Liouville problem are all real and non-negative.
(2) The eigen values of a Sturm-Liouville problem can be arranged to from a strictly increasing infinite sequence and $\lambda_n \to \infty$ an $n \to \infty$.
(3) For each eigen value of a Sturm-Liouville problem, there exists one and only one linearly independent eigen function.

Sturm-Liouville Expansion:

Let $y_1(x)$, $y_2(x)$, ...,$y_n(x)$, ... are eigen functions of the Regular Sturm-Liouville problem. Suppose that a function f(x) is given in the interval $a < x < b$ and that we wish to express f(x) in terms of the eigen functions $y_n(x)$, i.e., we wish to have

$$f(x) = \sum_{n=1}^{\infty} a_n y_n\,(x) \text{——} (1)$$

$= a_1y_1(x) + a_2y_2(x) +. \ . \ .+ a_ny_n(x) + \ldots$

This gives

$$\int_a^b r(x)y_n\,(x)f(x)dx = a_n \int_a^b r(x)[y_n(x)]^2 dx$$

[$\because \int_a^b r(x)y_m\,(x)y_n(x)dx = 0$ if $m \neq n$]

If we write $\propto_n = \int_a^b r(x)[y_n(x)]^2 dx$, we obtain

$$a_n = \frac{1}{\propto_n}\int_a^b r(x)y_n(x)\,f(x)dx \quad \text{——} (2)$$

Expansion of the form (1) with their coefficients given by (2) often called Sturm-Liouville expansion or expansion in eigen functions.

Example 4: For the equation $y''+\lambda y = 0$ ($\lambda>0$), find the eigen values and eigen functions which satisfy the end point conditions $y(0) = 0$ and $y(2\pi)= 0$. Verify also that these eigen functions corresponding to different eigen values are orthogonal in $[0,2\pi]$.

Solution: The general solution of the given equation is

$y(x) = A \cos \sqrt{\lambda}\, x + B \sin \sqrt{\lambda}\, x$

Putting the initial conditions, we have

$A = 0$

and $B \sin 2\sqrt{\lambda}\, \pi = 0$

If $B = 0$, then we get the trivial solution.

So, to get the non-trivial solution, we consider $B \neq 0$.

Then $\sin 2\sqrt{\lambda}\,\pi = 0$

i.e., $\sin 2\sqrt{\lambda}\,\pi = \sin n\pi$, n = 1, 2, 3, ...

i.e., $\lambda = \frac{n^2}{4}$

Thus the solution of the given equation is given by $y = \sin\frac{nx}{2}$, n = 1, 2, 3,..... The values of λ viz. $\frac{1}{4}$, $\frac{4}{4}$, $\frac{9}{4}$, ... are called eigen values and the corresponding solutions $\sin\frac{x}{2}$, $\sin\frac{2x}{2}$,... are called the eigen functions.

The general solution of the given equation becomes

$$y = \sum_{n=1}^{\infty} a_n \sin\frac{nx}{2} \quad \text{——} (1)$$

where a_n are arbitrary constants.

Now, the weight function is r(x) = 1 and $y_n(x) = \sin\frac{nx}{2}$, $u(x) = \sin\frac{n_1 x}{2}$, $v(x) = \sin\frac{n_2 x}{2}$

Then

$$\int_0^{2\pi} r(x)\,u(x)\,v(x)\,dx$$

$$= \int_0^{2\pi} \sin\frac{n_1 x}{2}\sin\frac{n_2 x}{2}\,dx$$

$$= \frac{1}{2}\int_0^{2\pi} \left[\cos\frac{(n_2 - n_1)x}{2} - \cos\frac{(n_2 + n_1)x}{2}\right]dx$$

$$= 0$$

$$\therefore \int_0^{2\pi} r(x)\,u(x)\,v(x)\,dx = 0$$

This proves that the eigen functions corresponding to the different eigen values are orthogonal.

Unit 7

Green's Function

Green's Function:

Suppose that we want to solve a non-homogeneous equation

$$L[u(x)] = f(x), \quad a \le x \le b \quad \text{——} (1)$$

where L is a differential operator defined by

$$L \equiv \frac{d}{dx}[p(x)\frac{d}{dx}] + q(x).$$

Here p(x), p′(x) and q(x) are given real valued continuous functions defined on [a,b].

Equation (1) is considered with boundary conditions

$$m_1 u(a) + m_2 u'(a) = 0 \text{——} (2)$$

and $$m_3 u(b) + m_4 u'(b) = 0 \text{——} (3)$$

with the usual assumption that at least one of m_1 and m_2 and one of m_3 and m_4 are non-zero.

If we can find two linearly independent solutions of the homogeneous equation L[u(x)] = 0, the solution of (1) can be obtained in the form

$$u(x) = \int_a^b G(x,z)f(z)dz.$$

The function G(x, z) is called the Green's function of the problem.

Suppose we have to solve the boundary value problem

$$\frac{d^2u}{dx^2} + k(x)\frac{du}{dx} + p(x)u = f(x), \quad a \le x \le b \quad \text{——} (4)$$

$$m_1u(a) + m_2u'(a) = 0 \text{——} (5)$$

$$m_3u(b) + m_4u'(b) = 0 \text{——} (6)$$

Let $u_1(x)$ and $u_2(x)$ be two linearly independent solutions of the corresponding homogeneous equation of (4). For simplicity, we may assume that u_1 satisfy the boundary condition at x = a and u_2 satisfy the boundary condition at x= b.

i.e., $$m_1u_1(a) + m_2u_1'(a) = 0$$

and $$m_3u_2(b) + m_4u_2'(b) = 0$$

Then the general solution of (4) can be written as

$$u(x) = c_1u_1(x) + c_2u_2(x) + \int_a^x \frac{u_1(z)u_2(x) - u_2(z)u_1(x)}{W(z)} f(z)\,dz \text{——} (7)$$

where W(z) is the Wronskian of u_1 and u_2 and $W \not\equiv 0$ because u_1 and u_2 are linearly independent.

Differentiating (7) with respect to x, we get

$$u'(x) = c_1u_1'(x) + c_2u_2'(x) + \int_a^x [u_1(z)u_2'(x) - u_2(z)u_1'(x)]\frac{f(z)}{w(z)}\,dz$$

[by Leibnitz's rule $\frac{d}{dx}\int_a^b f(x,z)dz = \int_a^b \frac{\partial f(x,z)}{\partial x}\,dz$]

Now, we apply the boundary condition (5) to the general solution u(x).

At x = a, we have $m_1u(a) + m_2u'(a) = c_1(m_1u_1(a) + m_2u_1'(a)) + c_2(m_1u_2(a)+m_2u_2'(a)) = 0$.

Since u_1 satisfy the boundary condition at x = a, we have

$c_2(m_1u_2(a)+m_2u_2'(a)) = 0$ i.e., $c_2 = 0$

At x = b, we have $m_3u(b) + m_4 u'(b) = 0$

i.e., $c_1(m_3u_1(b) + m_4u_1'(b)) + \int_a^b [u_1(z)[m_3u_2(b) + m_4u_2'(b)] - u_2(z)[m_3u_1(b) + m_4u_1'(b)]] \frac{f(z)}{w(z)} = 0$

i.e., $c_1(m_3u_1(b) + m_4u_1'(b)) - (m_3u_1(b) + m_4u_1'(b)) \int_a^b \frac{u_2(z)f(z)}{w(z)}\, dz = 0$

i.e., $c_1 = \int_a^b \frac{u_2(z)f(z)}{w(z)}\, dz$

Then from (7) we get

$u(x) = u_1(x)\int_a^b u_2(z) \frac{f(z)}{w(z)} dz + \int_a^x [u_1(z)u_2(x) - u_2(z)u_1(x)] \frac{f(z)}{w(z)} dz$

$= \int_a^x u_1(z)u_2(x) \frac{f(z)}{w(z)} dz + \int_x^b u_1(x)\, u_2(z) \frac{f(z)}{w(z)} dz$ ..... (8)

These two integrals can be combined into one. We first define Green's function for the problem (4), (5), (6) as

$G(x,z) = \frac{u_1(Z)u_2(x)}{w(z)}$, $a \le z \le x$

$= \frac{u_2(Z)u_1(x)}{w(z)}$, $x \le z \le b$

Then the formula given in (8) simplifies to

$u(x) = \int_a^b G(x,z)f(z)\, dz$.

Example 1: Solve by constructing Green's function

$\frac{d^2u}{dx^2} - u = -1$, $0 < x < 1$

$u(0) = 0$, $u(1) = 0$

Solution: First we have to find out two linearly independent solutions of the homogeneous differential equation

$u'' - u = 0$

that satisfies the boundary conditions as required.

The general solution of the homogeneous differential equation is

$u(x) = c_1 \cos hx + c_2 \sin hx$.

Since $u_1(x)$ is required to satisfy the condition at the left, $u_1(0) = 0$, we take $c_1 = 0$, $c_2 = 1$ and conclude $u_1(x) = \sin hx$.

The second solution is to satisfy $u_2(1) = 0$. We may take

$u_2(x) = \sin h1 \cos hx - \cosh 1 \sin hx = \sin h(1-x)$

The Wronskian of u_1 and u_2 is

$$W(x) = \begin{vmatrix} \sinh x & \sinh(1-x) \\ \cosh x & -\cosh(1-x) \end{vmatrix} = -\sinh 1$$

Therefore, the Green's function for this problem is

$G(x,z) = -\frac{\sinh z \ \sinh(1-x)}{\sinh 1}, \quad 0 \le z \le x$

$= -\frac{\sinh x \ \sinh(1-z)}{\sinh 1}, \quad x \le z \le 1$

Furthermore, since $f(x) = -1$, the solution is the integral

$u(x) = \int_0^1 -G(x,z)dz$

$= \int_0^x \frac{\sinh z \ \sinh(1-x)}{\sinh 1} dz + \int_x^1 \frac{\sinh x \ \sinh(1-z)}{\sinh 1} dz$

$= \frac{\sinh(1-x)}{\sinh 1}(\cosh x - 1) - \frac{\sinh x}{\sinh 1}(1 - \cosh(1-x)$

$= \frac{1}{\sinh 1}[\sinh(1-x)\cosh x + \cosh(1-x)\sinh x] - \frac{\sinh(1-x)+\sinh x}{\sinh 1}$

$= 1 - \frac{\sinh(1-x)+\sinh x}{\sinh 1}$

Example 2: Solve the following boundary value problem, by constructing Green's function,

$\frac{1}{x}\frac{d}{dx}\left(x\frac{du}{dx}\right) = f(x)$, $0 < x < 1$,

given that $u(0)$ is bounded and $u(1) = 0$.

Solution: The corresponding homogeneous equation is $\frac{1}{x}(xu'' + u') = 0$

$$\text{or, } u'' + \frac{u'}{x} = 0$$

$$\text{or, } \frac{d}{dx}\left(\frac{du}{dx}\right) + \frac{1}{x}\left(\frac{du}{dx}\right) = 0$$

$$\text{or, } \frac{d\left(\frac{du}{dx}\right)}{\frac{du}{dx}} + \frac{dx}{x} = 0$$

Integrating, $\log\left(\frac{du}{dx}\right) + \log x = \log c_1$, c_1 = constant

$$\text{or, } x\frac{du}{dx} = c_1$$

$$\text{or, } du = c_1 \frac{du}{dx}$$

$\therefore$ Integrating, $u = c_1 \log x + c_2$, c_2 = constant

Since u_1 is required to satisfy the condition at the left, $u_1(0)$ is bounded we take $c_1 = 0$ and $c_2 = 1$ and conclude $u_1(x) = 1$.

The second solution is to satisfy $u_2(1) = 0$. We take take $c_1 = 1$ and $c_2 = 0$ and conclude $u_2(x) = \log x$.

$\therefore W(z) = \begin{vmatrix} 1 & \log z \\ 0 & 1/z \end{vmatrix} = \frac{1}{z}$

Therefore, the Green's function for this problem is

$\therefore G(x,z) = z\log x, \quad 0 < z \le x$

$= z \log z, \quad x \le z < 1$

Therefore the solution of the given differential equation is

$u(x) = \int_0^1 G(x,z)\, f(z)\, dz$.

Another definition of Green's function:

A function G(x, z) defined on [a,b] ×[a,b] is called a Green's function for the boundary value problem L[u] = 0 with boundary conditions

$m_1u(a) + m_2u'(a) = 0$

$m_3u(b) + m_4\, u'(b) = 0,$

where $L \equiv \frac{d}{dx}[\, p(x) \frac{d}{dx}] + q(x),$

if for a given z,

$G(x, z) = G_1(x, z)$ if $x < z$

$= G_2\,(x, z)$ if $x > z$

where G_1 and G_2 are such that

(1) G_1 satisfies the boundary condition at x = a and $L(G_1) = 0$ for $x < z$;
(2) G_2 satisfies the boundary condition at x = b and $L(G_2) = 0$ for $x > z$;
(3) The function G(x, z) is continuous at x = z;
(4) The derivative of G with respect to x has a jump discontinuity at x = z and $[\, \frac{\partial G_2}{\partial x} - \frac{\partial G_1}{\partial x}\,]_{x=z} = -\frac{1}{p(z)}$.

With this definition, the Green's function for the above boundary value problem is constructed.

Example 3: Find solution of the following B.V.P. by constructing Green's function $u'' = f(x)$, $u(0) = 0 = u\,(1)$

Solution: The general solution of the corresponding homogeneous equation is $u(x) = ax+b$.
Let $G_1(x, z) = c_1x + c_2$, $0< x < z$
and $G_2(x, z) = c_3x + c_4$, $z< x < 1$
Then $G_1(0, z) = 0 \Longrightarrow c_2 = 0$
$G_2(1, z) = 0 \Longrightarrow c_3 + c_4 = 0$
Now, $[\, \frac{\partial G_2}{\partial x} - \frac{\partial G_1}{\partial x}\,]_{x=z} = -1 \Longrightarrow c_3 + c_1 = -1 \Longrightarrow c_3 = c_1 - 1$

G is continuous at $x = z \Longrightarrow c_1z = c_3z + c_4$
i.e., $z(c_1 - c_3) = c_4$
i.e., $c_4 = z$

$\therefore c_3 = -z$ and $c_1 = -z + 1$
Thus $G(x, z) = x(1-z)$, $0 < x < z$
$= z(1-x)$, $z < x < 1$.

∴ The solution of the problem is

$$u(x) = \int_0^1 G\,(x,z) f(z)dz$$

where $G(x,z) = z(1-x), 0 < z < x$
$= x(1-z), x < z < 1.$

Exercise:
1. Find the Green's function of the B.V.P.

$$u'' = 0, u(0) = 0 = u(l).$$

Exercise:
2. Solve the following B.V.P. by constructing Green's function

$$u'' = f(x) = 0,\ u(0) = 0,\ u(i) = u'(1) = 0$$

Unit 8 & Unit 9

Second Order Linear Differential Equation in Complex Domain

Solution of 2nd order linear differential equation in complex domain:

Let $\frac{d^2w}{dz^2} + p(z)\frac{dw}{dz} + q(z)w = 0$ —— (1)

be a given differential equation in the complex variable z, where p(z) and q(z) are functions of z. A point z = z_0 is called an ordinary point of (1) if both p(z) and q(z) are analytic at z_0, i.e., if p(z) and q(z) have the following expressions.

$p(z) = \sum_{n=0}^{\infty} p_n (z - z_0)^n$ and $q(z) = \sum_{n=0}^{\infty} q_n (z - z_0)^n$.

A point $z = z_0$is called a regular singularity of the equation (1) if at least one of p(z) and q(z) are not analytic at z_0, but $(z - z_0)$ p(z) and $(z - z_0)^2 q(z)$ are analytic at z_0, i.e., if z_0 is at best a simple pole of p(z) and a double pole of q(z). If z_0 is neither an ordinary point nor a regular singularity of (1), then z_0 is called an irregular singularity of (1).

We are interested to obtain solutions of the differential equation (1) in the neighborhood of an ordinary point and regular singularity. The result in the neighbourhood of an ordinary point is due to Fuch.

Theorem 6.1: Fuch's theorem:

Statement: Let z_0 be an ordinary point of the differential equation

$$\frac{d^2w}{dz^2} + p(z)\frac{dw}{dz} + q(z)w = 0 \text{ —— (1)}$$

and let a_0, a_1 be arbitrary constants. Then there exists a unique function w(z), which is analytic at z_0 , is a solution of (1) in certain nbd of z_0 and satisfies the initial conditions $w(z_0) = a_0$ and $w'(z_0) = a_1$. Furthermore, if the power series expansion of p(z) and q(z) are valid in $|z-z_0| < R$, (R>0), then the power series expansion of w(z) is also valid in $|z-z_0| < R$.

Proof:

Without loss of generality, we may assume that $z_0 = 0$. Since p(z) and q(z) are both analytic at z_0, we have

$$p(z) = \sum_{n=0}^{\infty} p_n z^n \text{ and } q(z) = \sum_{n=0}^{\infty} q_n z^n \text{ in } |z| < R, R > 0.$$

Let $w(z) = \sum_{n=0}^{\infty} a_n z^n$ be a solution of (1).

Then $\frac{dw}{dz} = \sum_{n=1}^{\infty} n a_n z^{n-1} = \sum_{n=0}^{\infty} (n+1)\, a_{n+1} z^n$

and $\frac{d^2w}{dz^2} = \sum_{n=2}^{\infty} n(n-1) a_n z^{n-2} = \sum_{n=0}^{\infty} (n+1)(n+2)\, a_{n+2}\, z^n$

So from (1) we obtain

$$\sum_{n=0}^{\infty} (n+2)(n+1) a_{n+2} z^n + \left(\sum_{n=0}^{\infty} p_n z^n\right) \sum_{n=0}^{\infty} (n+1) a_{n+1}\, z^n + \left(\sum_{n=0}^{\infty} q_n z^n\right) \sum_{n=0}^{\infty} a_n\, z^n = 0$$

i. e., $\sum_{n=0}^{\infty}\left[(n+2)(n+1)a_{n+2}+\sum_{k=0}^{n} p_{n-k}\,(k+1)a_{k+1}+\sum_{k=0}^{n} q_{n-k}\,a_k\right]\right]z^n = 0$

Thus a_n must satisfy the relation

$$-(n+2)(n+1)a_{n+2}=\sum_{k=0}^{n}[(k+1)p_{n-k}\,a_{k+1}+a_k q_{n-k}] \quad \text{——} \quad (2)$$

Putting n=0, 1, 2, ... we get

-2.1.$a_2 = a_1p_0 + q_0a_0$

-3.2.$a_3 = a_1p_1 + 2a_2p_0 + a_0q_1 + a_1q_0$

-4.3.$a_4 = a_1p_2 + 2a_2p_1 + 3a_3\,p_0 + a_0q_2 + a_1q_1 + a_2q_0$

.

.

.

The above recurrence relations exhibit that the coefficients a_n, n = 2, 3, are obtained uniquely in terms of a_0 and a_1. With these coefficients a_n, $w(z)=\sum_{n=0}^{\infty} a_n z^n$ satisfies the given equation (1).

We now show that the expansion of w(z) is valid in $|z| < R$ so that the solution is analytic. Let r be a positive real number less than R. Since p(z) and q(z) are analytic in $|z| < R$, the series $\sum_{n=0}^{\infty} p_n z^n$ and $\sum_{n=0}^{\infty} q_n z^n$ are both converges in $|z| < R$. By Cauchy's inequality, we have

$|p_n| \le \frac{M}{r^n}$ and $|q_n| \le \frac{M}{r^n}$ $\forall$ n,

where M = max{| p(z) |, | q (z) | } in $|z| < R$.

Then from (2), we get

$|(n+1)(n+2)a_{n+2}| \le \frac{M}{r^n}\sum_{k=0}^{n}[(k+1)|a_{k+1}| + |a_k|]r^k$

i.e., $(n+1)(n+2)|a_{n+2}| \le \frac{M}{r^n}\sum_{k=0}^{n}[(k+1)|a_{k+1}| + |a_k|]r^k + M|a_{n+1}|r$ —— (3)

We define $b_0 = |a_0|$, $b_1 = |a_1|$ and b_n by

$$(n+1)(n+2)\,b_{n+2} \le \frac{M}{r^n}\sum_{k=0}^{n}[(k+1)\,b_{k+1}+b_k]r^k + M\,b_{n+1}\,r \quad \text{——} \quad (4)$$

n = 0, 1, 2, ...

Comparing (4) with (3) we see that an induction yields $|a_n| \le b_n$ $\forall$n, $b_n \ge 0$

n = 0, 1, 2, ...

We now show that the series $\sum_{n=0}^{\infty} b_n z^n$ converges in $|z| < r$.

In fact,

$n(n+1)b_{n+1} = \frac{M}{r^{n-1}}\sum_{k=0}^{n-1}[(k+1)b_{k+1}+b_k]r^k + Mb_n r$ —— (5)

and $(n-1)n\,b_n = \frac{M}{r^{n-2}}\sum_{k=0}^{n-2}[(k+1)b_{k+1}+b_k]r^k + Mb_{n-1}r$ —— (6)

Now, r ×(6) – (7) gives.

$n(n+1)\,b_{n+1}r - (n-1)nb_n = M(nb_n + b_{n-1})\,r + Mb_nr^2 - Mb_{n-1}r$

i.e., $\frac{b_{n+1}}{b_n} = \frac{Mnr+Mr^2+n(n-1)}{n(n+1)r}$

Thus $\lim_{n\to\infty}\frac{b_{n+1}}{b_n} = \frac{1}{r}$

i.e., $\lim_{n\to\infty}\left|\frac{b_{n+1}z^{n+1}}{b_n z^n}\right| = \frac{|z|}{r}$

Thus the series $\sum_{n=0}^{\infty} b_n z^n$ converges in | z | < r. Since | a_n| ≤ b_n for all n, by comparison test, the series $\sum_{n=0}^{\infty} a_n z^n$ converges in | z | < r. Since r <R is arbitrary, it follows that the series $\sum_{n=0}^{\infty} a_n z^n$ converges in| z | < R.
Therefore w(z) = $\sum_{n=0}^{\infty} a_n z^n$ is analytic in | z | < R which satisfies the given differential equation (1) with w(0) = a_0 and w′(0) = a_1. Since for a given set of values a_0, a_1, we obtain the coefficients uniquely in terms of a_0 and a_1, the solution w(z) = $\sum_{n=0}^{\infty} a_n z^n$ is unique. This proves the theorem.

Example 1: Obtain the solution of the differential equation

$$(1-z^2)\frac{d^2w}{dz^2} - 2z\frac{dw}{dz} + n(n+1)\,w = 0 \qquad \text{——} \qquad (1)$$

in the neighbourhood of z = 0.

Solution: Comparing the given equation to the standard form

$$w''(z) + p(z)w'(z) + q(z)\,w = 0,$$

we obtain $p(z) = -\frac{2z}{1-z^2}$ and $q(z) = \frac{n(n+1)}{1-z^2}$, both of which are analytic at z = 0.
Hence z = 0 is an ordinary point of the given equation. So according to Fuch's theorem, there exists a unique solution w(z) which is analytic in certain neighbourhood of z = 0 which satisfies the initial conditions w(0) = a_0 and w′(0) = a_1, where a_0 and a_1 are arbitrary constants. We write

$$w(z) = \sum_{k=0}^{\infty} a_k z^k$$

Then

$$\frac{dw}{dz} = \sum_{k=1}^{\infty} k a_k z^{k-1}$$

$$\frac{d^2w}{dz^2} = \sum_{k=2}^{\infty} k(k-1) a_k z^{k-2}$$

Substituting in (1), we get

$$(1 - z^2 \sum_{k=2}^{\infty} k(k-1) a_k z^{k-2} - 2z \sum_{k=1}^{\infty} k a_k z^{k-1} + n(n+1) \sum_{k=0}^{\infty} a_k z^k = 0$$

Equating the coefficient of z^k to zero, we obtain

$(k+2)(k+1)a_{k+2} - k(k-1)a_k - 2ka_k + n(n+1)\,a_k = 0$

i.e., $(k+2)(k+1)a_{k+2} = [k(k-1) + 2k - n(n+1)]a_k$

i.e., $a_{k+2} = \frac{k(k+1)-n(n+1)}{(k+2)(k+1)}\,a_k$

$$= -\frac{(n-k)(n+k+1)}{(k+2)(k+1)}\,a_k$$

Putting k = 0, 1, 2, ... in succession, we get

$$a_2 = -\frac{n(n+1)}{1.2}\,a_0$$

$$a_3 = -\frac{(n-1)(n+2)}{2.3}\,a_1$$

$$a_4 = -\frac{(n-2)(n+3)}{3.4}\,a_2 = \frac{(n-2)n(n+1)(n+3)}{4!}\,a_0$$

$$a_5 = \frac{-(n-3)(n+4)}{4.5} a_3 = \frac{(n-3)(n-1)(n+2)(n+4)}{5!} a_1$$

and so on.

Therefore the solution of (1) becomes

$$w(z) = \sum_{k=0}^{\infty} a_k z^k$$

$$= a_0[1 - \frac{n(n+1)}{2!} z^2 + \frac{(n-2)n(n+1)(n+3)}{4!} z^4 - \ldots]$$

$$+ a_1[z - \frac{(n-1)(n+2)}{3!} z^3 + \frac{(n-3)(n-1)(n+2)(n+4)}{5!} z^5 - \ldots]$$

where a_0 and a_1 are arbitrary constants.

Note: The differential equation (1) is known as Legendre differential equation.

Theorem 6.2: Frobenius Theorem

Statement: Let z = 0 be a regular singularity of the differential equation

$$\frac{d^2w}{dz^2} + p(z)\frac{dw}{dz} + q(z)w = 0$$

and let the power series expansion of zp(z) and $z^2q(z)$ are valid in $|z| < R$, $R > 0$. Then there exists at least one solution of the given equation of the form

$$w(z) = z^m \sum_{n=0}^{\infty} a_n z^n$$

where m is a scalar and the series $\sum_{n=0}^{\infty} a_n z^n$ is convergent at least in $|z| < R$.

Example 2: Obtain the solution of the differential equation

$$z^2 \frac{d^2w}{dz^2} + z\frac{dw}{dz} + (z^2 - \nu^2)w = 0 \quad \text{——} \quad (1)$$

in the nbd of z = 0 , ν is neither zero nor an integer.

Solution: Comparing (1) with the equation $w''(z) + p(z)w'(z) + q(z)w(z) = 0$

we get $p(z) = \frac{1}{z}$ and $q(z) = \frac{z^2 - \nu^2}{z^2}$

Since both zp(z) = 1 and $z^2q(z) = z^2 - \nu^2$ are analytic at z = 0, the point z =0 is a regular singularity of (1).

Let

$$w(z) = \sum_{n=0}^{\infty} a_n z^{n+m}$$

be a solution of (1). Then

$$w'(z) = \sum_{n=0}^{\infty} (n+m) a_n z^{n+m-1}$$

$$w''(z) = \sum_{n=0}^{\infty}(n+m)(n+m-1)a_n z^{n+m-2}$$

So from (1) we obtain

$$z^2\sum_{n=0}^{\infty}(n+m)(n+m-1)a_n z^{n+m-2} + z\sum_{n=0}^{\infty}(n+m)a_n z^{n+m-1} + (z^2-\nu^2)\sum_{n=0}^{\infty}a_n z^{n+m} = 0$$

Equation the coefficient of z^m to zero, we get

$[m(m-1) + m - \nu^2]a_0 = 0$

If $a_0 \neq 0$ is arbitrary, then the indicial equation is

$m^2 - \nu^2 = 0.$

So, the exponents are $m = \pm\nu$

Equating to zero, the coefficient of z^{n+m}, we get

$(n+m)(n+m-1)\,a_n + (n+m)\,a_n - \nu^2\,a_n = 0$

i.e., $(n+m)^2\,a_n - \nu^2 a_n = -\,a_{n-2}$

i.e., $a_n\,(n+m+\nu)\,(n+m-\nu) = -\,a_{n-2}$

i.e., $a_n = -\frac{1}{(n+m+\nu)\,(n+m-\nu)}\;a_{n-2}$ —— (2)

The relation (2) will determine the even coefficients in terms of a_0 and the odd coefficients in terms of a_1.

Since $a_{-1} = 0$, we get $a_1 = 0$.

Consequently $a_{2n+1} = 0$ for all n.

Now, from (2) we get for $m = \nu$,

$$a_n = \frac{-1}{n(n+2\nu)}\;a_{n-2}$$

Putting n = 2, 4, 6, ..., we obtain

$$a_2 = -\frac{1}{2(2+2\nu)}\;a_0 = -\frac{1}{2.2(\nu+1)}\,a_0$$

$$a_4 = -\frac{1}{4(4+2\nu)}\;a_2 = (-1)^2\frac{1}{2^{2.2}1.2(\nu+1)(\nu+2)}\,a_0$$

$$a_6 = -\frac{1}{6(6+2\nu)}\;a_4 = (-1)^3\frac{1}{2^{2.3}1.2.3\,(\nu+1)(\nu+2)(\nu+3)}\,a_0$$

In general,

$$a_{2n} = (-1)^n\frac{1}{2^{2n}1.2\ldots n(\nu+1)(\nu+2)\ldots(\nu+n)}\,a_0$$

$$= (-1)^n\frac{\Gamma(\nu+1)}{2^{2n}n!\Gamma(\nu+1)(\nu+1)(\nu+2)\ldots(\nu+n)}\,a_0$$

$$= (-1)^n\frac{\Gamma(\nu+1)}{2^{2n}\Gamma(n+1)\Gamma(n+\nu+1)}\,a_0$$

Hence the general solution of (1) near z = 0 corresponding to the exponent ν is given by

$$w_1(z) = z^{\nu}\sum_{n=0}^{\infty}a_{2n}z^{2n}$$

$$= z^{\nu} \sum_{n=0}^{\infty} (-1)^n \frac{\Gamma(\nu+1)\mathrm{a}_0}{2^{2\mathrm{n}}\Gamma(n+1)\Gamma(\mathrm{n}+\nu+1)} z^{2n}$$

Taking $\mathrm{a}_0 = \frac{1}{2^{\nu}\Gamma(\nu+1)}$, we obtain a particular solution denoted by

$$J_{\nu}(z) = (\frac{z}{2})^{\nu} \sum_{n=0}^{\infty} (-1)^n \frac{(\frac{\mathrm{z}}{2})^{2\mathrm{n}}}{\Gamma(n+1)\Gamma(\mathrm{n}+\nu+1)}$$

Solution corresponding to the exponent $-\nu$ will be obtained symmetrically as

$$J_{-\nu}(z) = \sum_{n=0}^{\infty} (-1)^n \frac{(\frac{\mathrm{z}}{2})^{2\mathrm{n}-\nu}}{\Gamma(n+1)\Gamma(\mathrm{n}-\nu+1)}$$

Therefore the general solution of (1) is given by

$$\mathrm{w(z)} = \mathrm{A}\, J_{\nu}(\mathrm{z}) + \mathrm{B} J_{-\nu}(\mathrm{z})$$

where A and B are arbitrary constants.

Note: If ν is an integer, then

$$J_{-\nu}(z) = \sum_{n=0}^{\infty} (-1)^n \frac{(\frac{\mathrm{z}}{2})^{2\mathrm{n}-\nu}}{\Gamma(n+1)\Gamma(\mathrm{n}-\nu+1)}$$

$$= \sum_{n=\nu}^{\infty} (-1)^n \frac{(\frac{\mathrm{z}}{2})^{2\mathrm{n}-\nu}}{\Gamma(n+1)\Gamma(\mathrm{n}-\nu+1)} \qquad \because \frac{1}{\Gamma(p)} = 0 \; if \; p = 0 \; or \; negative \; integer$$

$$= \sum_{k=0}^{\infty} (-1)^{k+\nu} \frac{(\frac{\mathrm{z}}{2})^{2\mathrm{k}+\nu}}{\Gamma(k+\nu+1)\Gamma(\mathrm{k}+1)} \qquad [\text{Putting } \mathrm{n} = \mathrm{k} + \nu]$$

$$= (-1)^{\nu} J_{\nu}(z)$$

In this case the solutions are linearly dependent.

Note: The differential equation (1) is known as the Bessel's differential equation.

Exercise 1: Obtain a solution of the Gauss hypergeometric differential equation

$$\mathrm{z}(1-\mathrm{z})\frac{d^2w}{dz^2} + [c - (a+b+1)z]\frac{dw}{dz} - abw = 0 \quad \text{——} \quad (1)$$

near z = 0, c is neither an integer nor zero.

Exercise 2: Obtain a solution of the Hermite differential equation

$$\frac{d^2w}{dz^2} - 2z\frac{dw}{dz} + 2n\,w = 0$$

near z = 0, n being a constant.

Unit 10

10 Special Functions

10.1 Introduction

The classification theorem, concerning the type of O.D.E.'s that arise in mathematical physics, implies that once the solution to the standard form differential equations, are known, the solutions appropriate to most physical problems may be obtained as special cases. The present chapter is therefore devoted to applying the techniques to solve these two master equations. The solutions so generated are, of course, in series form and in general cannot be expressed in closed form in terms of elementary functions. As a result much of the chapter is devoted to constructing analytic continuations of the solutions to the general complex plane and using these to derive the properties that are useful in manipulating these functions in physical problems. The general properties are then written explicitly for the special cases of Bessel's and Legendre's equations since these are the most commonly encountered and furnish classic examples of the techniques described.

10.2 Hypergeometric Functions

The aim is to explicitly generate a series solution to the Hypergeometric equation

$$z(1-z)y''(z) + [c - (1+a+b)z]y'(z) - aby(z) = 0. \tag{10.2.1}$$

Since the coefficient functions have a simple form in powers of z it is convenient to expand about the regular singular point at $z = 0$. From Fuchs's theorem it follows that the solution must have the following form:

$$y(z) = \sum_{n=0}^{\infty} y_n z^{n+s} \tag{10.2.2}$$

which, when substituted into the differential equation gives the following indicial equation:

$$I(s) = s(s-1+c) = 0 \tag{10.2.3}$$

and recursion relation:

$$y_{n+1} = \left[\frac{(n+s+a)(n+s+b)}{(n+s+1)(n+s+c)}\right] y_n \qquad n \geq 0. \tag{10.2.4}$$

The roots of the indicial equation are $s_1 = 0$ and $s_2 = 1-c$. From the general discussion in section 8.5 the series ansatz may fail to generate both solutions if c should be an integer. Recalling the property $\Gamma(z+1) = z\Gamma(z)$ satisfied by Euler's gamma function allows the general solution to the recursion relation (10.2.4) to be written down immediately:

$$y_n(s) = C\left[\frac{\Gamma(n+s+a)\Gamma(n+s+b)}{\Gamma(n+s+1)\Gamma(n+s+c)}\right]. \tag{10.2.5}$$

The constant C is determined to be $y_0[\Gamma(s+1)\Gamma(s+c)]/[\Gamma(s+a)\Gamma(s+b)]$ by the initial condition that eq. (10.2.5) reduce to the free parameter y_0 when $n = 0$. Eq. (10.2.5) then reduces to:

$$y_n(s) = \left[\frac{\Gamma(c+s)\Gamma(s+1)}{\Gamma(a+s)\Gamma(b+s)}\right]\left[\frac{\Gamma(n+s+a)\Gamma(n+s+b)}{\Gamma(n+s+c)\Gamma(n+s+1)}\right] y_0. \tag{10.2.6}$$

Inspection of the limit

$$\lim_{n\to\infty} \frac{y_{n+1}(s)z^{s+n+1}}{y_n(s)z^{s+n}} = \lim_{n\to\infty}\left[\frac{(n+s+a)(n+s+b)}{(n+s+1)(n+s+c)}\right] z = z, \tag{10.2.7}$$

together with the ratio test, implies that for any a, b, c and s the series in eq. (10.2.2) converges for $|z| < 1$.

The solution corresponding to $s = 0$ is now easily written. With the conventional choice that $y_0 = 1$ the series solution using $s = 0$ in eq. (10.2.6) is:

$$\begin{aligned} y_1(z) &= \frac{\Gamma(c)}{\Gamma(a)\Gamma(b)}\sum_{n=0}^{\infty}\frac{\Gamma(n+a)\Gamma(n+b)}{\Gamma(n+c)n!}z^n \\ &= 1 + \frac{ab}{c}z + \frac{a(a+1)b(b+1)}{c(c+1)}\frac{z^2}{2} + ... \\ &\equiv F(a,b;c;z) \\ &\equiv {}_2F_1(a,b;c;z). \end{aligned} \tag{10.2.8}$$

The function $F(a,b;c;z)$ defined by this series is called the *Hypergeometric function* and its definition makes sense provided that $c \neq 0, -1, -2, ...$ and $|z| < 1$.

The second solution corresponds to the choice $s = 1 - c$. The corresponding solution is therefore given by:

$$\begin{aligned} y_2(z) &= \frac{\Gamma(2-c)}{\Gamma(a+1-c)\Gamma(b+1-c)}\sum_{n=0}^{\infty}\frac{\Gamma(n+a+1-c)\Gamma(n+b+1-c)}{\Gamma(n+2-c)n!}z^{n+1-c} \\ &= z^{1-c}\left[1 + \frac{(a+1-c)(b+1-c)}{2-c}z + ...\right] \\ &= z^{1-c}F(a+1-c, b+1-c; 2-c; z) \end{aligned} \tag{10.2.9}$$

This second solution is well defined only if $c \neq 2, 3, ...$ and $|z| < 1$. If $c = 1$ then the solutions (10.2.8) and (10.2.9) are not distinct. For non-integral c the two solutions $y_1(z)$ and $y_2(z)$ are linearly independent since they behave differently as $z \to 0$ and so cannot be proportional to one another.

10.3 Confluent Hypergeometric Functions

The series solution to the Confluent Hypergeometric equation

$$zy''(z) + (c - z)y'(z) - ay(z) = 0 \tag{10.3.1}$$

can be obtained in a similar fashion. It is more instructive, however, to obtain it directly from the previously constructed Hypergeometric function. To do so take $u = z/\lambda$ and $b = 1/\lambda$ in eq. (10.2.6). The solution to the Confluent equation is obtained by taking the limit as $\lambda \to 0$ with all other quantities fixed:

$$\begin{aligned} y_1(u) &\equiv M(a, c; u) \\ &\equiv {}_1F_1(a; c; u) \\ &= \lim_{\lambda \to 0} F(a, 1/\lambda; c; \lambda u) \end{aligned} \tag{10.3.2}$$

for $c \neq 0, -1, ...$ The required limit for the $n+1$'th term of the series is:

$$\begin{aligned} X &\equiv \lim_{\lambda \to 0} \left[\lambda^n \frac{\Gamma(n + 1/\lambda)}{\Gamma(1/\lambda)} \right] \\ &= \lim_{\lambda \to 0} \left[\lambda^n \left(\frac{1}{\lambda} \right) \left(\frac{1}{\lambda} + 1 \right) ... \left(\frac{1}{\lambda} + n - 1 \right) \right] \\ &= \lim_{\lambda \to 0} \left[1(1 + \lambda)(1 + 2\lambda)...(1 + (n-1)\lambda) \right] \\ &= 1. \end{aligned} \tag{10.3.3}$$

This leaves the following series solution to eq. (10.3.1):

$$\begin{aligned} M(a, c; z) &= \left[\frac{\Gamma(c)}{\Gamma(a)} \right] \sum_{n=0}^{\infty} \left[\frac{\Gamma(a + n)}{\Gamma(c + n)} \right] \frac{z^n}{n!} \\ &= 1 + \left[\frac{a}{c} \right] z + \left[\frac{a(a+1)}{c(c+1)} \right] \frac{z^2}{2} + ... \end{aligned} \tag{10.3.4}$$

for $c \neq 0, -1, ...$ This function is called the *Confluent Hypergeometric function.*

Using the limit:

$$\lim_{n \to \infty} \frac{y_{n+1} z^{n+1}}{y_n z^n} = \lim_{n \to \infty} \left[\frac{(a + n)}{(c + n)(n + 1)} \right] z = 0 \tag{10.3.5}$$

in the ratio test implies that the series (10.3.4) has an infinite radius of convergence.

The second solution is similarly found to be given by:

$$y_2(z) = z^{1-c} M(a + 1 - c, 2 - c; z) \tag{10.3.6}$$

provided $c \neq 2, 3, 4....$ For $c = 1$ solutions (10.3.2) and (10.3.6) are not distinct. For noninteger c these solutions are well-defined for all finite z and linearly independent.

EXAMPLES:

we know that the solutions to Bessel's equation and the Associated Legendre equation can be directly expressed in terms of the Hypergeometric and Confluent Hypergeometric series. relates the solutions for the Associated Legendre equation:

$$y_1(z) = C_m \left(\frac{1 - z}{1 + z} \right)^{m/2} F \left[\frac{1}{2}(1 + \sqrt{1 - 4t}), \frac{1}{2}(1 - \sqrt{1 - 4t}); 1 + m; \frac{1}{2}(1 - z) \right]. \tag{10.3.7}$$

with $m \neq -1, -2, ...$, to the Hypergeometric function. C_m denotes a constant. If $m < 0$ the solution can be taken as eq. (10.3.7) with m replaced everywhere by $-m$ since the Associated Legendre equation is invariant under this substitution. This is equivalent to the expression (10.2.9) for $y_2(z)$. Since the two roots to the indicial equation differ by $c = 1 + m$, the second linearly independent solution does not have the simple series form and must be constructed

Similarly, the solutions to Bessel's equation are given in terms of the confluent hypergeometric series by:

$$y_1(z) = C_\nu z^\nu e^{-iz} M\left[\frac{1}{2} + \nu, 1 + 2\nu; 2iz\right] \tag{10.3.8}$$

for $\nu \neq -\frac{1}{2}, -1, -\frac{3}{2}, ...$ C_ν again denotes a constant. The second solution is simply found by taking $\nu \to -\nu$ in this equation:

$$y_2(z) = C_{-\nu} z^{-\nu} e^{-iz} M\left[\frac{1}{2} - \nu, 1 - 2\nu; 2iz\right] \tag{10.3.9}$$

provided $\nu \neq \frac{1}{2}, 1, \frac{3}{2}, ...$

This construction is not limited to the Legendre or Bessel equations. To illustrate this point and for convenience of reference, the solution to the most frequently occuring O.D.E.'s of mathematical physics are briefly listed here in terms of Hypergeometric or Confluent Hypergeometric series:

1. ULTRASPHERICAL (GEGENBAUER) EQUATION:

$$(1 - z^2)y''(z) - (2\beta + 1)zy'(z) + n(n + 2\beta)y(z) = 0. \tag{10.3.10}$$

The regular solutions to this equation are denoted ${T_n}^\beta(z)$ and are called *Ultraspherical functions*. They are related to the Hypergeometric series by:

$${T_n}^\beta(z) = \frac{\Gamma(n + 2\beta + 1)}{2^\beta n!\Gamma(\beta + 1)} F\left[-n, n + 2\beta + 1; 1 + \beta; \frac{1}{2}(1 - z)\right]. \tag{10.3.11}$$

2. ASSOCIATED LEGENDRE FUNCTIONS:

$$y''(z) - \frac{2z}{1 - z^2}y'(z) + \left[\frac{\ell(\ell + 1)}{1 - z^2} - \frac{m^2}{(1 - z^2)^2}\right] y(z) = 0. \tag{10.3.12}$$

The regular solutions are the *Associated Legendre functions*, denoted ${P_\ell}^m(z)$:

$${P_\ell}^m(z) = \frac{1}{2^m m!}\frac{(\ell + m)!}{(\ell - m)!}(1 - z^2)^{m/2} F\left[m - \ell, m + \ell + 1; m + 1; \frac{1}{2}(1 - z)\right]. \tag{10.3.13}$$

3. CHEBYSHEV POLYNOMIALS:

$$y''(z) - \frac{z}{1-z^2}y'(z) + \frac{n^2}{1-z^2}y(z) = 0. \tag{10.3.14}$$

has regular solutions, $T_n(z)$, called *Chebyshev polynomials*:

$$T_n(z) = F\left[-n, n; \frac{1}{2}; \frac{1}{2}(1-z)\right]. \tag{10.3.15}$$

These are again special cases of the ultraspherical functions.

4. BESSEL FUNCTIONS:

$$y''(z) + \frac{1}{z}y'(z) + \left(1 - \frac{\nu^2}{z^2}\right)y(z) = 0. \tag{10.3.16}$$

has as regular solutions the *Bessel functions* $J_\nu(z)$. As shown these are given by:

$$J_\nu(z) = \frac{1}{\Gamma(\nu+1)}\left(\frac{z}{2}\right)^\nu e^{-iz} M\left[\nu + \frac{1}{2}, 2\nu + 1; 2iz\right]. \tag{10.3.17}$$

5. HERMITE FUNCTIONS:

$$y''(z) - 2zy'(z) + 2ny(z) = 0. \tag{10.3.18}$$

has *Hermite functions*, $H_n(z)$, as regular solutions:

$$H_{2n}(z) = (-)^n \frac{(2n)!}{n!} M\left[-n, \frac{1}{2}; z^2\right] \tag{10.3.19}$$

$$H_{2n+1}(z) = (-)^n \frac{2(2n+1)!}{n!} z M\left[-n, \frac{3}{2}; z^2\right]. \tag{10.3.20}$$

6. ASSOCIATED LAGUERRE FUNCTIONS:

$$y''(z) + \frac{k+1-z}{z}y'(z) + \frac{n}{z}y(z) = 0. \tag{10.3.21}$$

The solutions are the *Associated Laguerre functions*, ${L_n}^k(z)$, given by:

$${L_n}^k(z) = \frac{(n+k)!}{n!k!} M[-n, k+1; z]. \tag{10.3.22}$$

Block II

Partial Differential Equations

Unit 11 & Unit 12

FIRST-ORDER PARTIAL DIFFERENTIAL EQUATIONS

First-Order Partial Differential Equations

A first order PDE in two independent variables x, y and the dependent variable z can be written in the form

$$f(x, y, z, \frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}) = 0. \tag{1}$$

For convenience, we set

$$p = \frac{\partial z}{\partial x}, \quad q = \frac{\partial z}{\partial y}.$$

Equation (1) then takes the form

$$f(x, y, z, p, q) = 0. \tag{2}$$

The equations of the type (2) arise in many applications in geometry and physics. For instance, consider the following geometrical problem.

Example 1. *Find all functions $z(x, y)$ such that the tangent plane to the graph $z = z(x, y)$ at any arbitrary point $(x_0, y_0, z(x_0, y_0))$ passes through the origin characterized by the PDE $xz_x + yz_y - z = 0$.*

The equation of the tangent plane to the graph at $(x_0, y_0, z(x_0, y_0))$ is

$$z_x(x_0, y_0)(x - x_0) + z_y(x_0, y_0)(y - y_0) - (z - z(x_0, y_0)) = 0.$$

This plane passes through the origin $(0, 0, 0)$ and hence, we must have

$$-z_x(x_0, y_0)x_0 - z_y(x_0, y_0)y_0 + z(x_0, y_0) = 0. \tag{3}$$

For the equation (3) to hold for all (x_0, y_0) in the domain of z, z must satisfy

$$xz_x + yz_y - z = 0,$$

which is a first-order PDE.

Example 2. *The set of all spheres with centers on the z-axis is characterized by the first-order PDE $yp - xq = 0$.*

The equation

$$x^2 + y^2 + (z - c)^2 = r^2, \tag{4}$$

where r and c are arbitrary constants, represents the set of all spheres whose centers lie on the z-axis. Differentiating (4) with respect to x, we obtain

$$2\left(x + (z - c)\frac{\partial z}{\partial x}\right) = 2\left(x + (z - c)p\right) = 0. \tag{5}$$

Differentiate (4) with respect to y to have

$$y + (z - c)q = 0. \tag{6}$$

Eliminating the arbitrary constant c from (5) and (6), we obtain the first-order PDE

$$yp - xq = 0. \tag{7}$$

Equation (4) in some sense characterized the first-order PDE (7).

Example 3. *Consider all surfaces described by an equation of the form*

$$z = f(x^2 + y^2), \tag{8}$$

where f is an arbitrary function, described by the first-order PDE.

Writing $u = x^2 + y^2$ and differentiating (8) with respect to x and y, it follows that

$$p = 2xf'(u); \qquad q = 2yf'(u),$$

where $f'(u) = \frac{df}{du}$. Eliminating $f'(u)$ from the above two equations, we obtain the same first-order PDE as in (7).

Remark 4. *The function z described by each of the equations (4) and (8), in some sense, a solution to the PDE (7). Observe that, in Example 2, PDE (7) is formulated by eliminating arbitrary constants from (4) whereas in Example 3, PDE (7) is formed by eliminating an arbitrary function.*

1 Formation of first-order PDEs

The applications of conservation principles often yield a first-order PDEs. We have seen in the previous two examples that a first-order PDE can be formed either by eliminating arbitrary constants or an arbitrary function involved. Below, we now generalize the arguments of Example 2 and Example 3 to show that how a first-order PDE can be formed.

Method I (*Eliminating arbitrary constants*): Consider two parameters family of surfaces described by the equation

$$F(x, y, z, a, b) = 0, \tag{9}$$

where a and b are arbitrary constants. Equation (9) may be thought of as a generalization of the relation (4).

Differentiating (9) with respect to x and y, we obtain

$$\frac{\partial F}{\partial x} + p\frac{\partial F}{\partial z} = 0 \tag{10}$$

$$\frac{\partial F}{\partial y} + q\frac{\partial F}{\partial z} = 0. \tag{11}$$

Eliminate the constants a, b from equations (9), (10) and (11) to obtain a first-order PDE of the form

$$f(x, y, z, p, q) = 0. \tag{12}$$

This shows that a family of surfaces described by the relation (9) gives rise to a first-order PDE (12).

Method II (*Eliminating arbitrary function*): Now consider the generalization of Example 3. Let $u(x, y, z) = c_1$ and $v(x, y, z) = c_2$ be two known functions of x, y and z satisfying a relation of the form

$$F(u, v) = 0, \tag{13}$$

where F is an arbitrary function of u and v. Differentiating (13) with respect to x and y lead to the equations

$$F_u(u_x + u_z p) + F_v(v_x + v_z p) = 0$$

$$F_u(u_y + u_z q) + F_v(v_y + v_z q) = 0.$$

Eliminating F_u and F_v from the above two equations, we obtain

$$p\frac{\partial(u, v)}{\partial(y, z)} + q\frac{\partial(u, v)}{\partial(z, x)} = \frac{\partial(u, v)}{\partial(x, y)}, \tag{14}$$

which is a first-order PDE of the form $f(x, y, z, p, q) = 0$. Here, $\frac{\partial(u,v)}{\partial(x,y)} = u_x v_y - u_y v_x$.

2 Classification of first-order PDEs

We classify the equation (1) depending on the special forms of the function f. If (1) is of the form

$$\boxed{a(x, y)\frac{\partial z}{\partial x} + b(x, y)\frac{\partial z}{\partial y} + c(x, y)z = d(x, y)}$$

then it is called **linear** first-order PDE. Note that the function f is linear in $\frac{\partial z}{\partial x}$, $\frac{\partial z}{\partial y}$ and z with all coefficients depending on the independent variables x and y only.

If (1) has the form

$$\boxed{a(x, y)\frac{\partial z}{\partial x} + b(x, y)\frac{\partial z}{\partial y} = c(x, y, z)}$$

then it is called **semilinear** because it is linear in the leading (highest-order) terms $\frac{\partial z}{\partial x}$ and $\frac{\partial z}{\partial y}$. However, it need not be linear in z. Note that the coefficients of $\frac{\partial z}{\partial x}$ and $\frac{\partial z}{\partial y}$ are functions of the independent variables only.

If (1) has the form

$$\boxed{a(x,y,z)\frac{\partial z}{\partial x} + b(x,y,z)\frac{\partial z}{\partial y} = c(x,y,z)}$$

then it is called **quasi-linear** PDE. Here the function f is linear in the derivatives $\frac{\partial z}{\partial x}$ and $\frac{\partial z}{\partial y}$ with the coefficients a, b and c depending on the independent variables x and y as well as on the unknown z. Note that linear and semilinear equations are special cases of quasi-linear equations.

Any equation that does not fit into one of these forms is called **nonlinear**.

Example 5.

1. $xz_x + yz_y = z$ (linear)
2. $xz_x + yz_y = z^2$ (semilinear)
3. $z_x + (x+y)z_y = xy$ (linear)
4. $zz_x + z_y = 0$ (quasilinear)
5. $xz_x^2 + yz_y^2 = 2$ (nonlinear)

3 Cauchy's problem or IVP for first-order PDEs

Recall the initial value problem for a first-order ODE which ask for a solution of the equation that takes a given value at a given point of $\mathbb{R}$. The IVP for first-order PDE ask for a solution of (2) which has given values on a curve in $\mathbb{R}^2$. The conditions to be satisfied in the case of IVP for first-order PDE are formulated in the classic problem of Cauchy which may be stated as follows:

Let C be a given curve in $\mathbb{R}^2$ described parametrically by the equations

$$x = x_0(s), \qquad y = y_0(s); \quad s \in I, \tag{15}$$

where $x_0(s)$, $y_0(s)$ are in $C^1(I)$. Let $z_0(s)$ be a given function in $C^1(I)$. The IVP or Cauchy's problem for first-order PDE

$$f(x,y,z,p,q) = 0 \tag{16}$$

is to find a function $u = u(x,y)$ with the following properties:

- $u(x, y)$ and its partial derivatives with respect to x and y are continuous in a region Ω of $\mathbb{R}^2$ containing the curve C.
- $u = u(x, y)$ is a solution of (16) in Ω, i.e.,

$$f(x, y, u(x, y), u_x(x, y), u_y(x, y)) = 0 \quad \text{in } \Omega.$$

- On the curve C

$$u(x_0(s), y_0(s)) = z_0(s), \quad s \in I. \tag{17}$$

The curve C is called the initial curve of the problem and the function $z_0(s)$ is called the initial data. Equation (17) is called the initial condition of the problem.

NOTE: Geometrically, Cauchy's problem may be interpreted as follows: To find a solution surface $u = u(x, y)$ of (16) which passes through the curve C whose parametric equations are

$$x = x_0(s), \quad y = y_0(s) \quad z = z_0(s). \tag{18}$$

Further, at every point of which the direction $(p, q, -1)$ of the normal is such that

$$f(x, y, z, p, q) = 0.$$

The proof of existence of a solution of (16) passing through a curve with equations (18) requires some more assumptions on the function f and the nature of the curve C. We now state the classic theorem due to Kowalewski in the following theorem (cf. [10]).

Theorem 6. *(Kowalewski) If $g(y)$ and all its derivatives are continuous for $|y - y_0| < \delta$, if x_0 is a given number and $z_0 = g(y_0)$, $q_0 = g'(y_0)$, and if $f(x, y, z, q)$ and all its partial derivatives are continuous in a region S defined by*

$$|x - x_0| < \delta, \quad |y - y_0| < \delta, \quad |q - q_0| < \delta,$$

then there exists a unique function $\phi(x, y)$ such that:
(a) $\phi(x, y)$ and all its partial derivatives are continuous in a region
$\Omega : |x - x_0| < \delta_1, \ |y - y_0| < \delta_2$;
(b) For all (x, y) in Ω, $z = \phi(x, y)$ is a solution of the equation

$$\frac{\partial z}{\partial x} = f(x, y, z, \frac{\partial z}{\partial y})$$

(c) For all values of y in the interval $|y - y_0| < \delta_1, \ \phi(x_0, y) = g(y)$.

We conclude this lecture by introducing different kinds of solutions of first-order PDE.

Definition 7. *(**A complete solution or a complete integral**) Any relation of the form*

$$\boxed{F(x, y, z, a, b) = 0} \tag{19}$$

*which contains two arbitrary constants a and b and is a solution of a first-order PDE is called **a complete solution or a complete integral** of that first-order PDE.*

Definition 8. *(**A general solution or a general integral**) Any relation of the form*

$$\boxed{F(u, v) = 0}$$

involving an arbitrary function F connecting two known functions $u(x, y, z)$ and $v(x, y, z)$ and providing a solution of a first-order PDE is called a general solution or a general integral of that first-order PDE.

It is possible to derive a general integral of the PDE once a complete integral is known.

With $b = \phi(a)$, if we take any one-parameter subsystem

$$f(x, y, z, a, \phi(a)) = 0$$

of the system (19) and form its envelope, we obtain a solution of equation (16). When $\phi(a)$ is arbitrary, the solution obtained is called the general integral of (16) corresponding to the complete integral (19).

When a definite $\phi(a)$ is used, we obtain a particular solution.

Definition 9. *(**A singular integral**) The envelope of the two-parameter system (19) is also a solution of the equation (16). It is called the singular integral or singular solution of the equation.*

NOTE: The general solution of an equation of type (1) can be obtained by solving systems of ODEs. This is not true for higher-order equations or for systems of first-order equations.

Practice Problems

1. Classify whether the following PDE is linear, quasi-linear or nonlinear:

   $$(a)\ zz_x - 2xyz_y = 0;\ \ (b)\ z_x^2 + zz_y = 2;\ \ (c)\ z_x + 2z_y = 5z;\ \ (d)\ xz_x + yz_y = z^2.$$

2. Eliminate the arbitrary constants a and b from the following equations to form the PDE:

   $$(a)\ ax^2 + by^2 + z^2 = 1;\ \ (b)\ z = (x^2 + a)(y^2 + b).$$

Linear First-Order PDEs

The most general first-order linear PDE has the form

$$a(x,y)z_x + b(x,y)z_y + c(x,y)z = d(x,y), \tag{1}$$

where a, b, c, and d are given functions of x and y. These functions are assumed to be continuously differentiable. Rewriting (1) as

$$a(x,y)z_x + b(x,y)z_y = -c(x,y)z + d(x,y), \tag{2}$$

we observe that the left hand side of (2), i.e.,

$$a(x,y)z_x + b(x,y)z_y = \nabla z \cdot (a,b)$$

is (essentially) a directional derivative of $z(x,y)$ in the direction of the vector (a,b), where (a,b) is defined and nonzero. When a and b are constants, the vector (a,b) had a fixed direction and magnitude, but now the vector can change as its base point (x,y) varies. Thus, (a,b) is a vector field on the plane.

The equations

$$\boxed{\frac{dx}{dt} = a(x,y), \qquad \frac{dy}{dt} = b(x,y),} \tag{3}$$

determine a family of curves $x = x(t)$, $y = y(t)$ whose tangent vector $(\frac{dx}{dt}, \frac{dy}{dt})$ coincides with the direction of the vector (a,b). Therefore, the derivative of $z(x,y)$ along these curves becomes

$$\begin{aligned}
\frac{dz}{dt} = \frac{d}{dt} z\{(x(t),y(t))\} &= \frac{\partial z}{\partial x}\frac{dx}{dt} + \frac{\partial z}{\partial y}\frac{dy}{dt} \\
&= z_x(x(t),y(t))a(x(t),y(t)) + z_y(x(t),y(t))b(x(t),y(t)) \\
&= -c(x(t),y(t))z(x(t),y(t)) + d(x(t),y(t)) \\
&= -c(t)z(t) + d(t),
\end{aligned}$$

where we have used the chain rule and (1). Thus, along these curves, $z(t) = z(x(t),y(t))$ satisfies the ODE

$$z'(t) + c(t)z(t) = d(t). \tag{4}$$

Let $\mu(t) = \exp\left[\int_0^t c(\tau)d\tau\right]$ be an integrating factor for (4). Then, the solution is given by

$$z(t) = \frac{1}{\mu(t)}\left[\int_0^t \mu(\tau)d(\tau)d\tau + z(0)\right]. \tag{5}$$

The approach described above to solve (1) by using the solutions of (3)-(4) is called **the method of characteristics**. It is based on the geometric interpretation of the partial differential equation (1).

NOTE: (*i*) The ODEs (3) is known as the characteristics equation for the PDE (1). The solution curves of the characteristic equation are the characteristics curves for (1).

(*ii*) Observe that $\mu(t)$ and $d(t)$ depend only on the values of $c(x, y)$ and $d(x, y)$ along the characteristics curve $x = x(t)$, $y = y(t)$. Thus, equation (5) shows that the values $z(t)$ of the solution z along the entire characteristics curve are completely determined, once the value $z(0) = z(x(0), y(0))$ is prescribed.

(*iii*) Assuming certain smoothness conditions on the functions a, b, c, and d, the existence and uniqueness theory for ODEs guarantees a unique solution curve $(x(t), y(t), z(t))$ of (3)-(4) (i.e., a characteristic curve) passes through a given point (x_0, y_0, z_0) in (x, y, z)-space.

1 The method of characteristics for solving linear first-order IVP

In practice we are not interested in determining a general solution of the partial differential equation (1) but rather a specific solution $z = z(x, y)$ that passes through or contains a given curve C. This problem is known as the initial value problem for (1). The method of characteristics for solving the initial value problem for (1) proceeds as follows.

Let the initial curve C be given parametrically as:

$$x = x(s), \quad y = y(s), \quad z = z(s). \tag{6}$$

for a given range of values of the parameter s. The curve may be of finite or infinite extent and is required to have a continuous tangent vector at each point.

Every value of s fixes a point on C through which a unique characteristic curve passes (see, Fig. 2.1). The family of characteristic curves determined by the points of C may be parameterized as

$$x = x(s, t), \quad y = y(s, t), \quad z = z(s, t)$$

with $t = 0$ corresponding to the initial curve C. That is, we have

$$x(s, 0) = x(s), \quad y(s, 0) = y(s), \quad z(s, 0) = z(s).$$

In other words, we have the following:

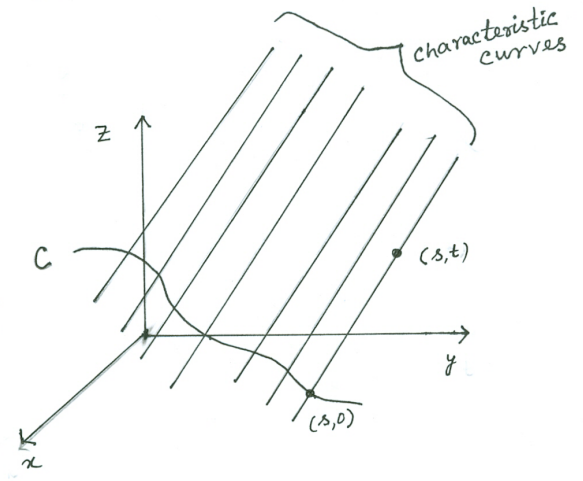


Figure 2.1: Characteristic curves and construction of the integral surface

The functions $x(s,t)$ and $y(s,t)$ are the solutions of the characteristics system (for each fixed s)

$$\frac{d}{dt}x(s,t) = a(x(s,t), y(s,t)), \quad \frac{d}{dt}y(s,t) = b(x(s,t), y(s,t)) \tag{7}$$

with given initial values $x(s,0)$ and $y(s,0)$.

Suppose that

$$z(x(s,0), y(s,0)) = g(s), \tag{8}$$

where $g(s)$ is a given function. We obtain $z(x(s,t), y(s,t))$ as follows: Let

$$z(s,t) = z(x(s,t), y(s,t)),\ c(s,t) = c(x(s,t), y(s,t)),\ d(s,t) = d(x(s,t), y(s,t)) \tag{9}$$

and

$$\mu(s,t) = \exp\left[\int_0^t c(s,t)dt\right]. \tag{10}$$

Analogous to formula (5), for each fixed s, we obtain

$$z(s,t) = \frac{1}{\mu(s,t)}\left[\int_0^t \mu(s,t)d(s,t)dt + g(s)\right]. \tag{11}$$

$z(s,t)$ is the value of z at the point $(x(s,t), y(s,t))$. Thus, as s and t vary, the point (x,y,z), in xyz-space, given by

$$x = x(s,t), \quad y = y(s,t), \quad z = z(s,t), \tag{12}$$

traces out the surface of the graph of the solution z of the PDE (1) which meets the initial curve (8). The equations (12) constitute the parametric form of the solution of (1) satisfying the initial condition (8) [i.e., a surface in (x, y, z)-space that contains the initial curve]

NOTE: If the Jacobian $J(s,t) = x_s y_t - x_t y_s \neq 0$, then the equations $x = x(s,t)$ and $y = y(s,t)$ can be inverted to give s and t as (smooth) functions of x and y i.e., $s = s(x,y)$ and $t = t(x,y)$. The resulting function $z = z(x,y) = z(s(x,y), t(x,y))$ satisfies the PDE (1) in a neighborhood of the curve C (in view of (4) and the initial condition (6)) and is the unique solution of the IVP.

Example 1. *Determine the solution the following IVP:*

$$\frac{\partial z}{\partial y} + c\frac{\partial z}{\partial x} = 0, \quad z(x,0) = f(x),$$

where $f(x)$ is a given function and c is a constant.

Solution. A step by step procedure for the finding solution is given below.

Step 1.(Finding characteristic curves)

To apply the method of characteristics, parameterize the initial curve C as follows: as follows:

$$x = s, \quad y = 0, \quad z = f(s). \tag{13}$$

The family of characteristics curves $x((s,t), y(s,t))$ are determined by solving the ODEs

$$\frac{d}{dt}x(s,t) = c, \quad \frac{d}{dt}y(s,t) = 1$$

The solution of the system is

$$x(s,t) = ct + c_1(s) \quad \text{and} \quad y(s,t) = t + c_2(s).$$

Step 2. (Applying IC)
Using the initial conditions

$$x(s,0) = s, \quad y(s,0) = 0.$$

we find that

$$c_1(s) = s, \quad c_2(s) = 0,$$

and hence

$$x(s,t) = ct + s \quad \text{and} \quad y(s,t) = t.$$

Step 3. (Writing the parametric form of the solution)
Comparing with (1), we have $c(x, y) = 0$ and $d(x, y) = 0$. Therefore, using (10) and (11), we find that

$$d(s, t) = 0, \quad \mu(s, t) = 1.$$

Since $z(x(s, 0), y(s, 0)) = z(s, 0) = g(s) = f(s)$, we obtain $z(s, t) = f(s)$. Thus, the parametric form of the solution of the problem is given by

$$x(s, t) = ct + s, \quad y(s, t) = t, \quad z(s, t) = f(s).$$

Step 4. (Expressing $z(s, t)$ in terms of $z(x, y)$) Expressing s and t as $s = s(x, y)$ and $t = t(x, y)$, we have

$$s = x - cy, \quad t = y.$$

We now write the solution in the explicit form as

$$z(x, y) = z(s(x, y), y(x, y)) = f(x - cy).$$

Clearly, if $f(x)$ is differentiable, the solution $z(x, y) = f(x - cy)$ satisfies given PDE as well as the initial condition.

NOTE: Example 1 characterizes unidirectional wave motion with velocity c. If we consider the initial function $z(x, 0) = f(x)$ to represent a waveform, the solution $z(x, y) = f(x - cy)$ shows that a point x for which $x - cy =$ constant, will always occupy the same position on the wave form. If $c > 0$, the entire initial wave form $f(x)$ moves to the right without changing its shape with speed c (if $c < 0$, the direction of motion is reversed).

Example 2. *Find the parametric form of the solution of the problem*

$$-yz_x + xz_y = 0$$

with the condition given by

$$z(s, s^2) = s^3, \quad (s > 0).$$

Solution. To find the solution, let's proceed as follows.

Step 1. (Finding characteristic curves)

The family of characteristics curves $(x(s, t), y(s, t))$ are determined by solving

$$\frac{d}{dt}x(s, t) = -y(s, t), \quad \frac{d}{dt}y(s, t) = x(s, t)$$

with initial conditions

$$x(s, 0) = s, \quad y(s, 0) = s^2.$$

The general solution of the system is

$$x(s,t) = c_1(s)\cos(t) + c_2(s)\sin(t) \quad \text{and} \quad y(s,t) = c_1(s)\sin(t) - c_2(s)\cos(t).$$

Step 2. (Applying IC)

Using ICs, we find that

$$c_1(s) = s, \quad c_2(s) = -s^2,$$

and hence

$$x(s,t) = s\cos(t) - s^2\sin(t) \quad \text{and} \quad y(s,t) = s\sin(t) + s^2\cos(t).$$

Step 3. (Writing the parametric form of the solution)

Comparing with (1), we note that $c(x,y) = 0$ and $d(x,y) = 0$. Therefore, using (10) and (11), it follows that

$$d(s,t) = 0, \quad \mu(s,t) = 1.$$

In view of the given condition curve and $z = z(s,t)$, we obtain

$$z(x(s,0), y(s,0)) = z(s,s^2) = g(s) = s^3, \quad z(s,t) = s^3.$$

Thus, the parametric form of the solution of the problem is given by

$$x(s,t) = s\cos(t) - s^2\sin(t), \quad y(s,t) = s\sin(t) + s^2\cos(t), \quad z(s,t) = s^3.$$

Step 4. (Expressing $z(s,t)$ in terms of $z(x,y)$)

Writing s and t as a function of x and y, it is an easy exercise to show that

$$z(x,y) = \frac{1}{\sqrt{8}}\left[-1 + \sqrt{1 + 4(x^2+y^2)}\right]^{3/2}.$$

Practice Problems

1. Find the general solution of the following PDE in the indicated domain.

   (A) $xz_x + 2yz_y = 0$, for $x > 0, y > 0$

   (B) $yz_x - 4xz_y = 2xy$, for all (x,y)

   (C) $xz_x - xyz_y = z$, for all (x,y)

2. Find a particular solution of the following PDEs satisfying the given side conditions.

Quasilinear First-Order PDEs

A first order quasilinear PDE is of the form

$$a(x,y,z)\frac{\partial z}{\partial x} + b(x,y,z)\frac{\partial z}{\partial y} = c(x,y,z). \tag{1}$$

Such equations occur in a variety of nonlinear wave propagation problems. Let us assume that an integral surface $z = z(x,y)$ of (1) can be found. Writing this integral surface in implicit form as

$$F(x,y,z) = z(x,y) - z = 0.$$

Note that the gradient vector $\nabla F = (z_x, z_y, -1)$ is normal to the integral surface $F(x,y,z) = 0$. The equation (1) may be written as

$$az_x + bz_y - c = (a,b,c)\cdot(z_x, z_y, -1) = 0. \tag{2}$$

This shows that the vector (a,b,c) and the gradient vector ∇F are orthogonal. In other words, the vector (a,b,c) lies in the tangent plane of the integral surface $z = z(x,y)$ at each point in the (x,y,z)-space where $\nabla F \neq 0$.

At each point (x,y,z), the vector (a,b,c) determines a direction in (x,y,z)-space is called the characteristic direction. We can construct a family of curves that have the characteristic direction at each point. If the parametric form of these curves is

$$x = x(t), \quad y = y(t), \quad \text{and} \quad z = z(t), \tag{3}$$

then we must have

$$\frac{dx}{dt} = a(x(t),y(t),z(t)), \quad \frac{dy}{dt} = b(x(t),y(t),z(t)), \quad \frac{dz}{dt} = c(x(t),y(t),z(t)), \tag{4}$$

because $(dx/dt, dy/dt, dz/dt)$ is the tangent vector along the curves. The solutions of (4) are called the characteristic curves of the quasilinear equation (1).

We assume that $a(x,y,z)$, $b(x,y,z)$, and $c(x,y,z)$ are sufficiently smooth and do not all vanish at the same point. Then, the theory of ordinary differential equations ensures that a unique characteristic curve passes through each point (x_0,y_0,z_0). The IVP for (1) requires that $z(x,y)$ be specified on a given curve in (x,y)-space which determines a curve C in (x,y,z)-space referred to as the initial curve. To solve this IVP, we pass a characteristic curve through each point of the initial curve C. If these curves generate a surface known as integral surface. This integral surface is the solution of the IVP.

Remark 1. (*i*) *The characteristic equations (4) for x and y are not, in general, uncoupled from the equation for z and hence differ from those in the linear case (cf. Eq. (3) of Lecture 2).*

(*ii*) *The characteristics equations (4) can be expressed in the nonparametric form as*

$$\frac{dx}{a} = \frac{dy}{b} = \frac{dz}{c}. \tag{5}$$

Below, we shall describe a method for finding the general solution of (1). This method is due to Lagrange hence it is usually referred to as *the method of characteristics* or *the method of Lagrange*.

1 The method of characteristics

It is a method of solution of quasi-linear PDE which is stated in the following result.

Theorem 2. *The general solution of the quasi-linear PDE (1) is*

$$F(u, v) = 0, \tag{6}$$

where F is an arbitrary function and $u(x, y, z) = c_1$ and $v(x, y, z) = c_2$ form a solution of the equations

$$\frac{dx}{a} = \frac{dy}{b} = \frac{dz}{c}. \tag{7}$$

Proof. If $u(x, y, z) = c_1$ and $v(x, y, z) = c_2$ satisfy the equations (1) then the equations

$$u_x dx + u_y dy + u_z dz = 0,$$
$$v_x dx + v_y dy + v_z dz = 0$$

are compatible with (7). Thus, we must have

$$au_x + bu_y + cu_z = 0,$$
$$av_x + bv_y + cv_z = 0.$$

Solving these equations for a, b and c, we obtain

$$\frac{a}{\frac{\partial(u,v)}{\partial(y,z)}} = \frac{b}{\frac{\partial(u,v)}{\partial(z,x)}} = \frac{c}{\frac{\partial(u,v)}{\partial(x,y)}}. \tag{8}$$

Differentiate $F(u, v) = 0$ with respect to x and y, respectively, to have

$$\frac{\partial F}{\partial u}\left\{\frac{\partial u}{\partial x} + \frac{\partial u}{\partial z}\frac{\partial z}{\partial x}\right\} + \frac{\partial F}{\partial v}\left\{\frac{\partial v}{\partial x} + \frac{\partial v}{\partial z}\frac{\partial z}{\partial x}\right\} = 0$$
$$\frac{\partial F}{\partial u}\left\{\frac{\partial u}{\partial y} + \frac{\partial u}{\partial z}\frac{\partial z}{\partial y}\right\} + \frac{\partial F}{\partial v}\left\{\frac{\partial v}{\partial y} + \frac{\partial v}{\partial z}\frac{\partial z}{\partial y}\right\} = 0.$$

Eliminating $\frac{\partial F}{\partial u}$ and $\frac{\partial F}{\partial v}$ from these equations, we obtain

$$\frac{\partial z}{\partial x}\frac{\partial(u,v)}{\partial(y,z)} + \frac{\partial z}{\partial y}\frac{\partial(u,v)}{\partial(z,x)} = \frac{\partial(u,v)}{\partial(x,y)} \tag{9}$$

In view of (8), the equation (9) yields

$$a\frac{\partial z}{\partial x} + b\frac{\partial z}{\partial y} = c.$$

Thus, we find that $F(u,v) = 0$ is a solution of the equation (1). This completes the proof. □.

Remark 3.
- *All integral surfaces of the equation (1) are generated by the integral curves of the equations (4).*
- *All surfaces generated by integral curves of the equations (4) are integral surfaces of the equation (1).*

Example 4. *Find the general integral of* $xz_x + yz_y = z$.

Solution. The associated system of equations are

$$\frac{dx}{x} = \frac{dy}{y} = \frac{dz}{z}.$$

From the first two relation we have

$$\frac{dx}{x} = \frac{dy}{y} \Longrightarrow \ln x = \ln y + \ln c_1 \Longrightarrow \frac{x}{y} = c_1.$$

Similarly,

$$\frac{dz}{z} = \frac{dy}{y} \Longrightarrow \frac{z}{y} = c_2.$$

Take $u_1 = \frac{x}{y}$ and $u_2 = \frac{z}{y}$. The general integral is given by

$$F(\frac{x}{y}, \frac{z}{y}) = 0.$$

Example 5. *Find the general integral of the equation*

$$z(x+y)z_x + z(x-y)z_y = x^2 + y^2.$$

Solution. The characteristic equations are

$$\frac{dx}{z(x+y)} = \frac{dy}{z(x-y)} = \frac{dz}{x^2+y^2}.$$

Each of these ratio is equivalent to

$$\frac{ydx + xdy - zdz}{0} = \frac{xdx - ydy - zdz}{0}.$$

Consequently, we have

$$d\{xy - \frac{z^2}{2}\} = 0 \quad \text{and} \quad d\{\frac{1}{2}(x^2 - y^2 - z^2)\} = 0.$$

Integrating we obtain two integrals

$$2xy - z^2 = c_1 \quad \text{and} \quad x^2 - y^2 - z^2 = c_2,$$

where c_1 and c_2 are arbitrary constants. Thus, the general solution is

$$F(2xy - z^2, x^2 - y^2 - z^2) = 0,$$

where F is an arbitrary function.

Next, we shall discuss a method for solving a Cauchy problem for the first-order quasi-linear PDE (1). The following theorem gives conditions under which a unique solution of the initial value problem for (1) can be obtained.

Theorem 6. *Let $a(x, y, z)$, $b(x, y, z)$ and $c(x, y, z)$ in (1) have continuous partial derivatives with respect to x, y and z variables. Let the initial curve C be described parametrically as*

$$x = x(s), \quad y = y(s), \quad \text{and} \quad z = z(x(s), y(s)).$$

The initial curve C has a continuous tangent vector and

$$J(s) = \frac{dy}{ds}a[x(s), y(s), z(s)] - \frac{dx}{ds}b(x(s), y(s), z(s)] \neq 0 \tag{10}$$

on C. Then, there exists a unique solution $z = z(x, y)$, defined in some neighborhood of the initial curve C, satisfies (1) and the initial condition $z(x(s), y(s)) = z(s)$.

Proof. The characteristic system (4) with initial conditions at $t = 0$ given as $x = x(s), y = y(s)$, and $z = z(s)$ has a unique solution of the form

$$x = x(s, t), \quad y = y(s, t), \quad z = z(s, t),$$

with continuous derivatives in s and t, and

$$x(s, 0) = x(s), \quad y(s, 0) = y(s), \quad z(s, 0) = z(s).$$

This follows from the existence and uniqueness theory for ODEs. The Jacobian of the transformation $x = x(s, t), y = y(s, t)$ at $t = 0$ is

$$J(s) = J(s, t)|_{t=0} = \left| \begin{array}{cc} \frac{\partial x}{\partial s} & \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial s} & \frac{\partial y}{\partial t} \end{array} \right|_{t=0} = \left[\frac{\partial y}{\partial t}a - \frac{\partial x}{\partial t}b\right]_{t=0} \neq 0. \tag{11}$$

in view of (10). By the continuity assumption, the Jacobian $J \neq 0$ in a neighborhood of the initial curve. Thus, by the implicit function theorem, we can solve for s and t as functions of x and y near the initial curve. Then

$$z(s,t) = z(s(x,y), t(x,y)) = Z(x,y).$$

a solution of (1), which can be easily seen as

$$\begin{aligned} c = \frac{dz}{dt} &= \frac{\partial z}{\partial x}\frac{dx}{dt} + \frac{\partial z}{\partial y}\frac{dy}{dt} \\ &= a\frac{\partial z}{\partial x} + b\frac{\partial z}{\partial y}, \end{aligned}$$

where we have used (4). The uniqueness of the solution follows from the fact that any two integral surfaces that contain the same initial curve must coincide along all the characteristic curves passing through the initial curve. This is a consequence of the uniqueness theorem for the IVP for (4). This completes our proof. □.

Example 7. *Consider the IVP:*

$$\begin{aligned} &\frac{\partial z}{\partial y} + z\frac{\partial z}{\partial x} = 0 \\ &z(x,0) = f(x), \end{aligned}$$

where f(x) is a given smooth function.

Solution. We solve this problem using the following steps.

Step 1. (Finding characteristic curves)
To solve the IVP, we parameterize the initial curve as

$$x = s, \quad y = 0, \quad z = f(s).$$

The characteristic equations are

$$\frac{dx}{dt} = z, \quad \frac{dy}{dt} = 1, \quad \frac{dz}{dt} = 0.$$

Let the solutions be denoted as $x(s,t)$, $t(s,t)$, and $z(s,t)$. We immediately find that

$$x(s,t) = zt + c_1(s), \quad y(s,t) = t + c_2(s), \quad z(s,t) = c_3(s),$$

where c_i, $i = 1, 2, 3$ are constants to be determined using IC.

Step 2. (Applying IC) The initial conditions at $s = 0$ are given by

$$x(s,0) = s, \quad y(s,0) = 0, \quad z(s,0) = f(s).$$

Using these condition, we obtain

$$x(s,t) = zt + s, \quad y(s,t) = t, \quad z(s,t) = f(s).$$

Step 3. (Writing the parametric form of the solution)

The solutions are thus given by

$$x(s,t) = zt + s = f(s)t + s, \quad y(s,t) = t, \quad z(s,t) = f(s).$$

Step 4. (Expressing $z(s,t)$ in terms of $z(x,y)$) Applying the condition (10), we find that $J(s) = -1 \neq 0$, along the entire initial curve. We can immediately solve for $s(x,y)$ and $t(x,y)$ to obtain

$$s(x,y) = x - tf(s), \quad t(x,y) = y.$$

Since $t = y$ and $s = x - tf(s) = x - yz$, the solution can also be given in implicit form as

$$z = f(x - yz).$$

Example 8. *Solve the following quasi-linear PDE:*

$$zz_x + yz_y = x, \quad (x,y) \in \mathbf{R}^2$$

subject to the initial condition

$$z(x,1) = 2x, \quad x \in \mathbf{R}.$$

Solution. Here $a(x,y,z) = z, \quad b(x,y,z) = y, \quad c(x,y,z) = x$. The characteristics equations are

$$\begin{aligned}
\frac{dx}{dt} &= z, \quad x(s,0) = s, \\
\frac{dy}{dt} &= y, \quad y(s,0) = 1, \\
\frac{dz}{dt} &= x, \quad z(s,0) = 2s.
\end{aligned}$$

On solving the above ODEs, we obtain

$$x(s,t) = \frac{s}{2}(3e^t - e^{-t}), \;\; y(s,t) = e^t, \;\; z(s,t) = \frac{s}{2}(3e^t + e^{-t}).$$

Solving for (s,t) in terms of (x,y), we obtain

$$\begin{aligned}
s(x,y) &= \frac{2xy}{3y^2 - 1}, \quad t(x,y) = ln(y), \\
z(x,y) &= z(s(x,y), t(x,y)) = \frac{(3y^2 + 1)x}{(3y^2 - 1)}.
\end{aligned}$$

Note that the characteristics variables imply that y must be positive ($y = e^t$). In fact, the solution z is valid only for $3y^2 - 1 > 0$, i.e., for $y > \frac{1}{\sqrt{3}} > 0$. Observe that the change of variables is valid only where

$$\begin{vmatrix} x_s(s,t) & x_t(s,t) \\ y_s(s,t) & y_t(s,t) \end{vmatrix} \neq 0.$$

It is easy to verify that this condition leads to $y \neq 1/\sqrt{3}$.

Practice Problems

1. Find a solution of the PDE $z_x + zz_y = 6x$ satisfying the condition $z(0, y) = 3y$.

2. Find the general integral of the PDE

$$(2xy - 1)z_x + (z - 2x^2)z_y = 2(x - yz)$$

   and also the particular integral which passes through the line $x = 1$, $y = 0$.

3. Solve $z_x + zz_y = 2x, \quad z(0, y) = f(y)$.

4. Find the solution of the equation $z_x + zz_y = 1$ with the data

$$x(s, 0) = 2s, \quad y(s, 0) = s^2, \quad z(0, s^2) = s.$$

5. Find the characteristics of the equation $z_x z_y = z$, and determine the integral surface which passes through the parabola $x = 0$, $y^2 = z$.

Nonlinear First-Order PDEs

The general nonlinear first-order PDE is written in the form

$$F(x, y, z, z_x, z_y) = 0, \tag{1}$$

where F is not linear in z_x and z_y. Setting $z_x = p$ and $z_y = q$, rewrite (1) as

$$F(x, y, z, p, q) = 0. \tag{2}$$

1 The method of characteristics for nonlinear PDEs

Recall the method of characteristics for solving first-order linear PDE:

$$F(x, y, z, p, q) = a(x, y)p + b(x, y)q + c(x, y)z - d(x, y) = 0.$$

In this method, the PDE becomes an ODEs along the characteristics curves which may be regarded as the solutions of the system

$$x'(t) = a(x(t), y(t)) \quad \text{and} \quad y'(t) = b(x(t), y(t)). \tag{3}$$

Note that $F_p = a(x, y)$ and $F_q = b(x, y)$. Hence, (3) may be written as

$$x'(t) = F_p \quad \text{and} \quad y'(t) = F_q. \tag{4}$$

For solving first-order nonlinear PDE (1), the relation (4) motivates us to define characteristics curves as solutions of the system

$$x'(t) = F_p(x(t), y(t), z(t), p(t), q(t)) \quad \text{and} \quad y'(t) = F_q(x(t), y(t), z(t), p(t), q(t)), \tag{5}$$

where $z(t) = z(x(t), y(t))$, $p(t) = z_x(x(t), y(t))$, $q(t) = z_y(x(t), y(t))$. However, unlike the linear case, the right sides of (5) depend not only on $x(t)$ and $y(t)$, but also on $z(t)$, $p(t)$ and $q(t)$. Thus, we can expect a large system of five ODEs for the five unknown $x(t)$, $y(t)$, $z(t)$, $p(t)$ and $q(t)$. For the remaining three equations, notice that

$$\begin{aligned} z'(t) &= \frac{d}{dt}\{z(x(t), y(t))\} \\ &= z_x x'(t) + z_y y'(t) \\ &= p(t)x'(t) + q(t)y'(t) \\ &= p(t)F_p(x(t), y(t), z(t), p(t), q(t)) + q(t)F_q(x(t), y(t), z(t), p(t), q(t)). \end{aligned} \tag{6}$$

Along a characteristics p is a function of t. The equation for $p'(t)$ is obtained as follows:

$$\begin{aligned} p'(t) &= \frac{d}{dt}\{z_x(x(t), y(t))\} \\ &= z_{xx}x'(t) + z_{xy}y'(t) \\ &= z_{xx}F_p(x(t), y(t), z(t), p(t), q(t)) + z_{xy}F_q(x(t), y(t), z(t), p(t), q(t)). \end{aligned} \tag{7}$$

Using the fact that $z(x, y)$ should solve the PDE (1), we obtain

$$\begin{aligned} 0 &= \frac{d}{dx}\{F(x, y, z(x, y), z_x(x, y), z_y(x, y))\} \\ &= F_x + F_z z_x + F_p z_{xx} + F_q z_{yx}. \end{aligned}$$

Therefore,

$$p'(t) = z_{xx}F_q + z_{xy}F_q = -(F_x + pF_z). \tag{8}$$

Similarly,

$$q'(t) = -[F_y + qF_z]. \tag{9}$$

Thus, we have the following system of five ODEs

$$\boxed{\begin{aligned} x'(t) &= F_p(x(t), y(t), z(t), p(t), q(t)) \\ y'(t) &= F_q(x(t), y(t), z(t), p(t), q(t)) \\ z'(t) &= p(t)F_p(x(t), y(t), z(t), p(t), q(t)) + q(t)F_q(x(t), y(t), z(t), p(t), q(t)) \\ p'(t) &= -\{F_x(x(t), y(t), z(t), p(t), q(t)) + p(t)F_z(x(t), y(t), z(t), p(t), q(t))\} \\ q'(t) &= -\{F_y(x(t), y(t), z(t), p(t), q(t)) + q(t)F_z(x(t), y(t), z(t), p(t), q(t))\} \end{aligned}} \tag{10}$$

These equations constitute the characteristics system of the PDE (1) and are known as the characteristics equations associated with PDE (1).

NOTE: If the functions which appear in equations (10) satisfy a Lipschitz condition, there is a unique solution of the equations for each prescribed set of initial values of the variables. Therefore the characteristics strip is uniquely determined by any initial element $(x(t_0), y(t_0), z(t_0), p(t_0), q_(t_0))$ at any initial point t_0 of t.

An important result about characteristic strips is given below.

Theorem 1. *The function $F(x, y, z, p, q)$ is a constant along every characteristics strip of the equation $F(x, y, z, p, q) = 0$.*

Proof. Along a characteristic strip, we have

$$\begin{aligned}\frac{d}{dt}\{F(x(t),y(t),z(t),p(t),q(t))\} &= F_x x'(t) + F_y y'(t) + F_z z'(t) + F_p p'(t) + F_q q'(t) \\ &= F_x F_p + F_y F_q + F_z(pF_p + qF_q) - F_p(F_x + pF_z) - F_q(F_y + qF_z) \\ &= 0.\end{aligned}$$

This implies $F(x,y,z,p,q) = k$, a constant along the strip.

2 Solving Cauchy's problem for nonlinear PDEs

The objective of this section to solve PDE

$$F(x,y,z,z_x,z_y) = 0$$

subject to an appropriate initial condition (i.e., z assume prescribed values on some curve).

Let $(f(s),g(s))$ traces out a regular curve in the xy-plane as s varies. We regard this curve as being an initial curve. We seek a solution $u(x,y)$ of the following problem (known as Cauchy's problem).

$$F(x,y,z,z_x,z_y) = 0, \quad u(f(s),g(s)) = G(s), \tag{11}$$

where $G(s)$ is a continuously differentiable function. Such a problem may have no solution (e.g., the PDE $z_x^2 + z_y^2 + 1 = 0$). However, if a solution exists in some neighborhood of the initial curve, then such a solution can often be determined using the following steps (cf. [1]).

Step 1: Find functions $h(s)$ and $k(s)$ (if possible) such that

$$\begin{aligned}&F(f(s),g(s),G(s),h(s),k(s)) = 0, \;\; G'(s) = h(s)f'(s) + k(s)g'(s) \;\text{ and} \\ &F_p(f(s),g(s),G(s),h(s),k(s))g'(s) - F_q(f(s),g(s),G(s),h(s),k(s))f'(s) \neq 0.\end{aligned} \tag{12}$$

Note that if $h(s)$ and $k(s)$ do not exist, then (11) has no solution. If there are several choices for $(h(s),k(s))$, then a solution of (11) exists for each such choice.

Step 2: For each fixed s, solve the following charateristics system for $x(s,t)$, $y(s,t)$, $z(s,t)$, $p(s,t)$,$q(s,t)$ with the given initial conditions $p(s,0) = h(s)$, $q(s,0) = k(s)$, where $h(s)$ and $k(s)$ are the functions found in Step 1.

$$\begin{aligned}
\frac{d}{dt}x(s,t) &= F_p(x(s,t),y(s,t),z(s,t),p(s,t),q(s,t)) \\
\frac{d}{dt}y(s,t) &= F_q(x(s,t),y(s,t),z(s,t),p(s,t),q(s,t)) \\
\frac{d}{dt}z(s,t) &= p(s,t)F_p(x(s,t),y(s,t),z(s,t),p(s,t),q(s,t)) \\
&\quad +q(s,t)F_q(x(s,t),y(s,t),z(s,t),p(s,t),q(s,t)) \\
\frac{d}{dt}p(s,t) &= -[F_x(x(s,t),y(s,t),z(s,t),p(s,t),q(s,t)) \\
&\quad +p(s,t)F_z(x(s,t),y(s,t),z(s,t),p(s,t),q(s,t))] \\
\frac{d}{dt}q(s,t) &= -[F_y(x(s,t),y(s,t),z(s,t),p(s,t),q(s,t)) \\
&\quad +q(s,t)F_z(x(s,t),y(s,t),z(s,t),p(s,t),q(s,t))]
\end{aligned} \tag{13}$$

Step 3: As s and t vary, the point (x,y,z), defined by

$$x = x(s,t),\ y = y(s,t),\ z = z(s,t) \tag{14}$$

traces out the graph of a solution z of (11) in the xyz-space, in a neighborhood of the curve traced out by $(f(s),g(s),G(s))$. In some cases, one can use the first two equations in (14) to solve for s and t in terms of x and y (say, $s = s(x,y)$ and $t = t(x,y)$) to obtain a solution $z(x,y) = z(s(x,y),t(x,y))$, for (x,y) in a neighborhood of the curve $(f(s),g(s))$.

To illustrate the above steps, let us consider the following example.

Example 2. *Solve the PDE $z_x z_y - z = 0$ subject to the condition $z(s,-s) = 1$.*

Solution. Here, we have

$$F(x,y,z,p,q) = pq - z.$$

The characteristics system (13) takes the form

$$\begin{aligned}
&\frac{dx}{dt} = F_p = q(t), \quad \frac{dy}{dt} = F_q = p(t), \quad \frac{dz}{dt} = pF_p + qF_q = 2p(t)q(t), \\
&\frac{dp}{dt} = -[F_x + p(t)F_z] = p(t), \quad \frac{dq}{dt} = -[F_y + q(t)F_z] = q(t).
\end{aligned}$$

Note that

$$\frac{dp}{dt} = p(t) \Longrightarrow p(t) = ce^t \ \text{ and } \ \frac{dq}{dt} = q(t) \Longrightarrow q(t) = de^t,$$

where c and d are arbitrary constants. Since we are looking for a characteristics strip (i.e., $F(x,y,z,p,q) = 0$), we set $z(t) = p(t)q(t) = cde^{2t}$. The equations for the characteristic strip are:

$$x(t) = de^t + d_1, \quad y(t) = ce^t + c_1, \quad z(t) = cde^{2t}, \quad p(t) = ce^t, \quad q(t) = de^t,$$

where c_1 and d_1 are constants.

The initial condition $z(s,-s)=1$ is given on the line $y=-x$ traced out by $(s,-s)$, in (11), we have $f(s)=s$ and $g(s)=-s$. We must find $h(s)$ and $k(s)$ such that

$$1=G(s)=h(s)k(s) \quad 0=G'(s)=h(s)-k(s),$$
$$0\neq F_p(\ldots)(-1)-F_q(\ldots)(1)=-k(s)-h(s).$$

Thus, we have two choices $h(s)=1$ and $k(s)=1$, or $h(s)=-1$ and $k(s)=-1$. For the choice $h(s)=1$ and $k(s)=1$, we obtain

$$x(s,t)=e^t-1+s, \quad y(s,t)=e^t-1-s, \quad z(s,t)=e^{2t}, \quad p(s,t)=e^t, \quad q(s,t)=e^t.$$

From the first two equations, we obtain

$$e^t=(x+y+2)/2.$$

Then the solution is

$$z(x,y)=e^{2t}=\frac{(x+y+2)^2}{4}.$$

If we choose $h(s)=-1$ and $k(s)=-1$, the solution is given by

$$z(x,y)=\frac{(x+y-2)^2}{4}.$$

Practice Problems

Solve the following Cauchy's problem:

1. $pq-z=0, \quad z(x,-x)=x$
2. $p+zq=2x, \quad z(0,y)=f(y)$
3. Find the solution of the equation $p+zq=1$ with the data

$$x(s,0)=2s, \quad y(s,0)=s^2, \quad z(0,s^2)=s.$$

4. Find the characteristics of the equation $pq=z$, and determine the integral surface which passes through the parabola $x=0$, $y^2=z$.

Compatible Systems and Charpit's Method

In this lecture, we shall study compatible systems of first-order PDEs and the Charpit's method for solving nonlinear PDEs. Let's begin with the following definition.

Definition 1. ***(Compatible systems of first-order PDEs)*** *A system of two first-order PDEs*

$$f(x, y, z, p, q) = 0 \tag{1}$$

and

$$g(x, y, z, p, q) = 0 \tag{2}$$

are said to be compatible if they have a common solution.

Theorem 2. *The equations $f(x, y, z, p, q) = 0$ and $g(x, y, z, p, q) = 0$ are compatible on a domain D if*

(i) $J = \frac{\partial(f,g)}{\partial(p,q)} = \begin{vmatrix} f_p & f_q \\ g_p & g_q \end{vmatrix} \neq 0$ *on D.*

(ii) p and q can be explicitly solved from (1) and (2) as $p = \phi(x, y, z)$ and $q = \psi(x, y, z)$. Further, the equation

$$dz = \phi(x, y, z)dx + \psi(x, y, z)dy$$

is integrable.

Theorem 3. *A necessary and sufficient condition for the integrability of the equation $dz = \phi(x, y, z)dx + \psi(x, y, z)dy$ is*

$$\boxed{[f, g] \equiv \frac{\partial(f, g)}{\partial(x, p)} + \frac{\partial(f, g)}{\partial(y, q)} + p\frac{\partial(f, g)}{\partial(z, p)} + q\frac{\partial(f, g)}{\partial(z, q)} = 0.} \tag{3}$$

In other words, the equations (1) and (2) are compatible iff (3) holds.

Example 4. *Show that the equations*

$$xp - yq = 0, \qquad z(xp + yq) = 2xy$$

are compatible and solve them.

Solution. Take $f \equiv xp - yq = 0, \qquad g \equiv z(xp + yq) - 2xy = 0$. Note that

$$f_x = p, \;\; f_y = -q, \;\; f_z = 0, \;\; f_p = x, \;\; f_q = -y.$$

and

$$g_x = zp - 2y, \;\; g_y = zq - 2x, \;\; g_z = xp + yq, \;\; g_p = zx, \;\; g_q = zy.$$

Compute

$$J \equiv \frac{\partial(f,g)}{\partial(p,q)} = \begin{vmatrix} f_p & f_q \\ g_p & g_q \end{vmatrix} = \begin{vmatrix} x & -y \\ zx & zy \end{vmatrix} = zxy + zxy = 2zxy \neq 0$$

for $x \neq 0$, $y \neq 0$, $z \neq 0$. Further,

$$\begin{aligned}
\frac{\partial(f,g)}{\partial(x,p)} &= \begin{vmatrix} f_x & f_p \\ g_x & g_p \end{vmatrix} = \begin{vmatrix} p & x \\ zp-2y & zx \end{vmatrix} = zxp - x(zp-2y) = 2xy \\
\frac{\partial(f,g)}{\partial(z,p)} &= \begin{vmatrix} f_z & f_p \\ g_z & g_p \end{vmatrix} = \begin{vmatrix} 0 & x \\ xp+yq & zx \end{vmatrix} = 0 - x(xp+yq) = -x^2p - xyq \\
\frac{\partial(f,g)}{\partial(y,q)} &= \begin{vmatrix} f_y & f_q \\ g_y & g_q \end{vmatrix} = \begin{vmatrix} -q & -y \\ zq-2x & zy \end{vmatrix} = -qzy + y(zq-2x) = -2xy \\
\frac{\partial(f,g)}{\partial(z,q)} &= \begin{vmatrix} f_z & f_q \\ g_z & g_q \end{vmatrix} = \begin{vmatrix} 0 & -y \\ xp+yq & zy \end{vmatrix} = y(xp+yq) = y^2q + xyp.
\end{aligned}$$

It is an easy exercise to verify that

$$\begin{aligned}
[f,g] &\equiv \frac{\partial(f,g)}{\partial(x,p)} + \frac{\partial(f,g)}{\partial(y,q)} + p\frac{\partial(f,g)}{\partial(z,p)} + q\frac{\partial(f,g)}{\partial(z,q)} \\
&= 2xy - x^2p^2 - xypq - 2xy + y^2q^2 + xypq \\
&= y^2q^2 - x^2p^2 \\
&= 0.
\end{aligned}$$

So the equations are compatible.

Next step to determine p and q from the two equations $xp - yq = 0$, $z(xp+yq) = 2xy$. Using these two equations, we have

$$\begin{aligned}
zxp + zyq - 2xy = 0 &\implies xp + yq = \frac{2xy}{z} \\
&\implies 2xp = \frac{2xy}{z} \implies p = \frac{y}{z} = \phi(x,y,z).
\end{aligned}$$

and

$$\begin{aligned}
xp - yq = 0 &\implies q = \frac{xp}{y} = \frac{xy}{yz} = \frac{x}{z} \\
&\implies q = \frac{x}{z} = \psi(x,y,z).
\end{aligned}$$

Substituting p and q in $dz = pdx + qdy$, we get

$$zdz = ydx + xdy = d(xy),$$

and hence integrating, we obtain

$$z^2 = 2xy + k,$$

where k is a constant.

NOTE: For the compatibility of $f(x, y, z, p, q) = 0$ and $g(x, y, z, p, q) = 0$ it is not necessary that every solution of $f(x, y, z, p, q) = 0$ be a solution of $g(x, y, z, p, q) = 0$ or vice-versa as is generally believed. For instance, the equations

$$f \equiv xp - yq - x = 0 \tag{4}$$

$$g \equiv x^2p + q - xz = 0 \tag{5}$$

are compatible. They have common solutions $z = x + c(1 + xy)$, where c is an arbitrary constant. Note that $z = x(y + 1)$ is a solution of (4) but not of (5).

Charpit's Method: It is a general method for finding the complete integral of a nonlinear PDE of first-order of the form

$$f(x, y, z, p, q) = 0. \tag{6}$$

Basic Idea: The basic idea of this method is to introduce another partial differential equation of the first order

$$g(x, y, z, p, q, a) = 0 \tag{7}$$

which contains an arbitrary constant a and is such that

(i) Equations (6) and (7) can be solved for p and q to obtain

$$p = p(x, y, z, a), \qquad q = q(x, y, z, a).$$

(ii) The equation

$$dz = p(x, y, z, a)dx + q(x, y, z, a)dy \tag{8}$$

is integrable.

When such a function g is found, the solution

$$F(x, y, z, a, b) = 0$$

of (8) containing two arbitrary constants a, b will be the solution of (6).

Note: Notice that another PDE g is introduced so that the equations f and g are compatible and then common solutions of f and g are determined in the Charpit's method.

The equations (6) and (7) are compatible if

$$[f, g] \equiv \frac{\partial(f, g)}{\partial(x, p)} + \frac{\partial(f, g)}{\partial(y, q)} + p\frac{\partial(f, g)}{\partial(z, p)} + q\frac{\partial(f, g)}{\partial(z, q)} = 0.$$

Expanding it, we are led to the linear PDE

$$f_p\frac{\partial g}{\partial x}+f_q\frac{\partial g}{\partial y}+(pf_p+qf_q)\frac{\partial g}{\partial z}-(f_x+pf_z)\frac{\partial g}{\partial p}-(f_y+qf_z)\frac{\partial g}{\partial q}=0. \tag{9}$$

Now solve (9) to determine g by finding the integrals of the following auxiliary equations:

$$\boxed{\frac{dx}{f_p}=\frac{dy}{f_q}=\frac{dz}{pf_p+qf_q}=\frac{dp}{-(f_x+pf_z)}=\frac{dq}{-(f_y+qf_z)}} \tag{10}$$

These equations are known as Charpit's equations which are equivalent to the characteristics equations (10) of the previous Lecture 4.

Once an integral $g(x,y,z,p,q,a)$ of this kind has been found, the problem reduces to solving for p and q, and then integrating equation (8).

Remark 5. *1. For finding integrals, all of Charpit's equations (10) need not to be used.*
2. p or q must occur in the solution obtained from (10).

Example 6. *Find a complete integral of*

$$p^2x+q^2y=z. \tag{11}$$

Solution. To find a complete integral, we proceed as follows.

Step 1: (Computing f_x, f_y, f_z, f_p, f_q).
Set $f\equiv p^2x+q^2y-z=0$. Then

$$f_x=p^2,\quad f_y=q^2,\quad f_z=-1,\quad f_p=2px,\quad f_q=2qy.$$

$$\Longrightarrow pf_p+qf_q=2p^2x+2q^2y,\ \ -(f_x+pf_z)=-p^2+p,\ \ -(f_y+qf_z)=-q^2+q.$$

Step 2: (Writing Charpit's equations and finding a solution $g(x,y,z,p,q,a)$).

The Charpit's equations (or auxiliary) equations are:

$$\begin{aligned}&\frac{dx}{f_p}=\frac{dy}{f_q}=\frac{dz}{pf_p+qf_q}=\frac{dp}{-(f_x+pf_z)}=\frac{dq}{-(f_y+qf_z)}\\ \Longrightarrow\quad &\frac{dx}{2px}=\frac{dy}{2qy}=\frac{dz}{2(p^2x+q^2y)}=\frac{dp}{-p^2+p}=\frac{dq}{-q^2+q}\end{aligned}$$

From which it follows that

$$\begin{aligned}&\frac{p^2dx+2pxdp}{2p^3x+2p^2x-2p^3x}=\frac{q^2dy+2qydq}{2q^3y+2q^2y-2q^3y}\\ \Longrightarrow\quad &\frac{p^2dx+2pxdp}{p^2x}=\frac{q^2dy+2qydq}{q^2y}\end{aligned}$$

On integrating, we obtain

$$\begin{aligned} & \log(p^2x) = \log(q^2y) + \log a \\ \Longrightarrow \quad & p^2x = aq^2y, \end{aligned} \tag{12}$$

where a is an arbitrary constant.

Step 3: (Solving for p and q).

Using (11) and (12), we find that

$$\begin{aligned} & p^2x + q^2y = z, \quad p^2x = aq^2y \\ \Longrightarrow \quad & (aq^2y) + q^2y = z \Longrightarrow q^2y(1+a) = z \\ \Longrightarrow \quad & q^2 = \frac{z}{(1+a)y} \Longrightarrow q = \left[\frac{z}{(1+a)y}\right]^{1/2}. \end{aligned}$$

and

$$\begin{aligned} & p^2 = aq^2\frac{y}{x} = a\frac{z}{(1+a)y}\frac{y}{x} = \frac{az}{(1+a)x} \\ \Longrightarrow \quad & p = \left[\frac{az}{(1+a)x}\right]^{1/2}. \end{aligned}$$

Step 4: (Writing $dz = p(x,y,z,a)dx + q(x,y,z,a)dy$ and finding its solution).

Writing

$$\begin{aligned} & dz = \left[\frac{az}{(1+a)x}\right]^{1/2}dx + \left[\frac{z}{(1+a)y}\right]^{1/2}dy \\ \Longrightarrow \quad & \left(\frac{1+a}{z}\right)^{1/2}dz = \left(\frac{a}{x}\right)^{1/2}dx + \left(\frac{1}{y}\right)^{1/2}dy. \end{aligned}$$

Integrate to have

$$[(1+a)z]^{1/2} = (ax)^{1/2} + (y)^{1/2} + b$$

the complete integral of the equation (11).

Practice Problems

1. Show that the equations $xp - yq = x$ and $x^2p + q = xz$ are compatible and solve them.

2. Show that the equations $f(x,y,p,q) = 0$ and $g(x,y,p,q) = 0$ are compatible if
$$\frac{\partial(f,g)}{\partial(x,p)} + \frac{\partial(f,g)}{\partial(y,p)} = 0.$$

3. Find complete integrals of the equations:
(i) $p = (z+qy)^2$; (ii) $(p^2+q^2)y = qz$

Some Special Types of First-Order PDEs

We shall consider some special types of first-order partial differential equations whose solutions may be obtained easily by Charpit's method.

Type (a): (Equations involving only p and q)
If the equation is of the form

$$f(p, q) = 0 \tag{1}$$

then Charpit's equations take the form

$$\frac{dx}{f_p} = \frac{dy}{f_q} = \frac{dz}{pf_p + qf_q} = \frac{dp}{0} = \frac{dq}{0}$$

An immediate solution is given by $p = a$, where a is an arbitrary constant. Substituting $p = a$ in (1), we obtain a relation

$$q = Q(a).$$

Then, integrating the expression

$$dz = adx + Q(a)dy$$

we obtain

$$z = ax + Q(a)y + b, \tag{2}$$

where b is a constant. Thus, (2) is a complete integral of (1).

Note: Instead of taking $dp = 0$, we can take $dq = 0 \Rightarrow q = a$. In some problems, taking $dq = 0$ the amount of computation involved may be reduced considerably.

Example 1. *Find a complete integral of the equation* $pq = 1$.

Solution. If $p = a$ then $pq = 1 \Rightarrow q = \frac{1}{a}$. In this case, $Q(a) = 1/a$. From (2), we obtain a complete integral as

$$\begin{aligned} & z = ax + \frac{y}{a} + b \\ \Longrightarrow \quad & a^2x + y - az = c, \end{aligned}$$

where a and c are arbitrary constants.

Type (b) (Equations not involving the independent variables):
For the equation of the type

$$f(z, p, q) = 0, \tag{3}$$

Charpit's equation becomes

$$\frac{dx}{f_p} = \frac{dy}{f_q} = \frac{dz}{pf_p + qf_q} = \frac{dp}{-pf_z} = \frac{dq}{-qf_z}.$$

From the last two relation, we have

$$\begin{aligned} & \frac{dp}{-pf_z} = \frac{dq}{-qf_z} \implies \frac{dp}{p} = \frac{dq}{q} \\ \implies \quad & p = aq, \end{aligned} \tag{4}$$

where a is an arbitrary constant. Solving (3) and (4) for p and q, we obtain

$$q = Q(a, z) \Longrightarrow p = aQ(a, z).$$

Now

$$\begin{aligned} & dz = pdx + qdy \\ \implies \quad & dz = aQ(a, z)dx + Q(a, z)dy \\ \implies \quad & dz = Q(a, z)\left[adx + dy\right]. \end{aligned}$$

It gives complete integral as

$$\int \frac{dz}{Q(a, z)} = ax + y + b, \tag{5}$$

where b is an arbitrary constant.

Example 2. *Find a complete integral of the PDE* $p^2z^2 + q^2 = 1$.

Solution. Putting $p = aq$ in the given PDE, we obtain

$$\begin{aligned} & a^2q^2z^2 + q^2 = 1 \\ \implies \quad & q^2(1 + a^2z^2) = 1 \\ \implies \quad & q = (1 + a^2z^2)^{-1/2}. \end{aligned}$$

Now,

$$\begin{aligned} & p^2 = (1 - q^2)/z^2 = \left(1 - \frac{1}{(1 + a^2z^2)}\right)\left(\frac{1}{z^2}\right) \\ \implies \quad & p^2 = \frac{a^2}{1 + a^2z^2} \\ \implies \quad & p = a(1 + a^2z^2)^{-1/2}. \end{aligned}$$

Substituting p and q in $dz = pdx + qdy$, we obtain

$$\begin{aligned} & dz = a(1 + a^2z^2)^{-1/2}dx + (1 + a^2z^2)^{-1/2}dy \\ \implies \quad & (1 + a^2z^2)^{1/2}dz = adx + dy \\ \implies \quad & \frac{1}{2a}\left\{az(1 + a^2z^2)^{1/2} - \log[az + (1 + a^2z^2)^{1/2}]\right\} = ax + y + b, \end{aligned}$$

which is the complete integral of the given PDE.

Type (c): (Separable equations)

A first-order PDE is separable if it can be written in the form

$$f(x,p) = g(y,q). \tag{6}$$

That is, a PDE in which z is absent and the terms containing x and p can be separated from those containing y and q. For this type of equation, Charpit's equations become

$$\frac{dx}{f_p} = \frac{dy}{-g_q} = \frac{dz}{pf_p - qg_q} = \frac{dp}{-f_x} = \frac{dq}{-g_y}.$$

From the last two relation, we obtain an ODE

$$\frac{dp}{-f_x} = \frac{dx}{f_p} \implies \frac{dp}{dx} + \frac{f_x}{f_p} = 0 \tag{7}$$

which may be solved to yield p as a function of x and an arbitrary constant a. Writing (7) in the form $f_p dp + f_z dx = 0$, we see that its solution is $f(x,p) = a$. Similarly, we get $g(y,q) = a$. Determine p and q from the equation

$$f(x,p) = a, \quad g(y,q) = a$$

and then use the relation $dz = pdx + qdy$ to determine a complete integral.

Example 3. *Find a complete integral of* $p^2y(1+x^2) = qx^2$.

Solution. First we write the given PDE in the form

$$\frac{p^2(1+x^2)}{x^2} = \frac{q}{y} \quad \text{(separable equation)}$$

It follows that

$$\frac{p^2(1+x^2)}{x^2} = a^2 \implies p = \frac{ax}{\sqrt{1+x^2}},$$

where a is an arbitrary constant. Similarly,

$$\frac{q}{y} = a^2 \implies q = a^2 y.$$

Now, the relation $dz = pdx + qdy$ yields

$$dz = \frac{ax}{\sqrt{1+x^2}}dx + a^2 y dy \implies z = a\sqrt{1+x^2} + \frac{a^2y^2}{2} + b,$$

where a and b are arbitrary constant, a complete integral for the given PDE.

Type (d): (Clairaut's equation)

A first-order PDE is said to be in Clairaut form if it can be written as

$$z = px + qy + f(p, q). \tag{8}$$

Charpit's equations take the form

$$\frac{dx}{x + f_p} = \frac{dy}{y + f_q} = \frac{dz}{px + qy + pf_p + qf_q} = \frac{dp}{0} = \frac{dq}{0}.$$

Now, $dp = 0 \Longrightarrow p = a$, where a is an arbitrary constant.
$dq = 0 \Longrightarrow q = b$, where b is an arbitrary constant.
Substituting the values of p and q in (8), we obtain the required complete integral

$$z = ax + by + f(a, b).$$

Example 4. *Find a complete integral of* $(p + q)(z - xp - yq) = 1$.

Solution. The given PDE can be put in the form

$$z = xp + yq + \frac{1}{p + q}, \tag{9}$$

which is of Clairaut's type. Putting $p = a$ and $q = b$ in (9), a complete integral is given by

$$z = ax + by + \frac{1}{a + b},$$

where a and b are arbitrary constants.

Practice Problems

Find complete integrals of the following PDEs.

1. $p + q = pq$
2. $\sqrt{p} + \sqrt{q} = 1$
3. $z = p^2 - q^2$
4. $p(1 + q) = qz$
5. $p^2 + q^2 = x + y$
6. $z = px + qy + \sqrt{1 + p^2 + p^2}$
7. $zpq = p^2(xq + p^2) + q^2(yp + q^2)$

Unit 13

SECOND-ORDER PARTIAL DIFFERENTIAL EQUATIONS

Classification of Second-Order PDEs

Classification of PDEs is an important concept because the general theory and methods of solution usually apply only to a given class of equations. Let us first discuss the classification of PDEs involving two independent variables.

1 Classification with two independent variables

Consider the following general second order linear PDE in two independent variables:

$$A\frac{\partial^2 u}{\partial x^2} + B\frac{\partial^2 u}{\partial x \partial y} + C\frac{\partial^2 u}{\partial y^2} + D\frac{\partial u}{\partial x} + E\frac{\partial u}{\partial y} + Fu + G = 0, \tag{1}$$

where A, B, C, D, E, F and G are functions of the independent variables x and y. The equation (1) may be written in the form

$$Au_{xx} + Bu_{xy} + Cu_{yy} + f(x, y, u_x, u_y, u) = 0, \tag{2}$$

where

$$u_x = \frac{\partial u}{\partial x},\ \ u_y = \frac{\partial u}{\partial y},\ \ u_{xx} = \frac{\partial^2 u}{\partial x^2},\ \ u_{xy} = \frac{\partial^2 u}{\partial x \partial y},\ \ u_{yy} = \frac{\partial^2 u}{\partial y^2}.$$

Assume that A, B and C are continuous functions of x and y possessing continuous partial derivatives of as high order as necessary.

The classification of PDE is motivated by the classification of second order algebraic equations in two-variables

$$ax^2 + bxy + cy^2 + dx + ey + f = 0. \tag{3}$$

We know that the nature of the curves will be decided by the principal part $ax^2+bxy+cy^2$ i.e., the term containing highest degree. Depending on the sign of the discriminant b^2-4ac, we classify the curve as follows:

If $b^2 - 4ac > 0$ then the curve traces hyperbola.
If $b^2 - 4ac = 0$ then the curve traces parabola.
If $b^2 - 4ac < 0$ then the curve traces ellipse.

With suitable transformation, we can transform (3) into the following normal form

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1 \quad \text{(hyperbola)}.$$
$$x^2 = y \quad \text{(parabola)}.$$
$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \quad \text{(ellipse)}.$$

SECOND-ORDER PARTIAL DIFFERENTIAL EQUATIONS

Linear PDE with constant coefficients. Let us first consider the following general linear second order PDE in two independent variables x and y with constant coefficients:

$$Au_{xx} + Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu + G = 0, \tag{4}$$

where the coefficients A, B, C, D, E, F and G are constants. The nature of the equation (4) is determined by the principal part containing highest partial derivatives i.e.,

$$Lu \equiv Au_{xx} + Bu_{xy} + Cu_{yy}. \tag{5}$$

For classification, we attach a symbol to (5) as $P(x, y) = Ax^2 + Bxy + Cy^2$ (as if we have replaced x by $\frac{\partial}{\partial x}$ and y by $\frac{\partial}{\partial y}$). Now depending on the sign of the discriminant (B^2-4AC), the classification of (4) is done as follows:

$$\boxed{\begin{aligned} B^2 - 4AC > 0 &\implies \text{Eq. (4) is hyperbolic} \quad (6)\\ B^2 - 4AC = 0 &\implies \text{Eq. (4) is parabolic} \quad (7)\\ B^2 - 4AC < 0 &\implies \text{Eq. (4) is elliptic} \quad (8)\end{aligned}}$$

Linear PDE with variable coefficients. The above classification of (4) is still valid if the coefficients A, B, C, D, E and F depend on x, y. In this case, the conditions (6), (7) and (8) should be satisfied at each point (x, y) in the region where we want to describe its nature e.g., for elliptic we need to verify

$$B^2(x, y) - 4A(x, y)C(x, y) < 0$$

for each (x, y) in the region of interest. Thus, we classify linear PDE with variable coefficients as follows:

$$\boxed{\begin{aligned} B^2(x, y) - 4A(x, y)C(x, y) > 0 \text{ at } (x, y) &\implies \text{Eq. (4) is hyperbolic at } (x, y) \quad (9)\\ B^2(x, y) - 4A(x, y)C(x, y) = 0 \text{ at } (x, y) &\implies \text{Eq. (4) is parabolic at } (x, y) \quad (10)\\ B^2(x, y) - 4A(x, y)C(x, y) < 0 \text{ at } (x, y) &\implies \text{Eq. (4) is elliptic at } (x, y) \quad (11)\end{aligned}}$$

Note: Eq. (4) is hyperbolic, parabolic, or elliptic depends only on the coefficients of the second derivatives. It has nothing to do with the first-derivative terms, the term in u, or the nonhomogeneous term.

Example 1.

1. $u_{xx} + u_{yy} = 0$ (Laplace equation). Here, $A = 1$, $B = 0$, $C = 1$ and $B^2 - 4AC = -4 < 0$. Therefore, it is an elliptic type.

2. $u_t = u_{xx}$ (Heat equation). Here, $A = -1$, $B = 0$, $C = 0$. $B^2 - 4AC = 0$. Thus, it is of parabolic type.

3. $u_{tt} - u_{xx} = 0$ (Wave equation). In this case, $A = -1$, $B = 0$, $C = 1$ and $B^2 - 4AC = 4 > 0$. Hence, it is of hyperbolic type.

4. $u_{xx} + xu_{yy} = 0, \quad x \neq 0$ (Tricomi equation). $B^2 - 4AC = -4x$. Given PDE is hyperbolic for $x < 0$ and elliptic for $x > 0$. This example shows that equations with variable coefficients can change form in the different regions of the domain.

2 Classification with more than two variables

Consider the second-order PDE in general form:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{n} b_i \frac{\partial u}{\partial x_i} + cu + d = 0, \tag{12}$$

where the coefficients a_{ij}, b_i, c and d are functions of $x = (x_1, x_2, \cdots, x_n)$ alone and $u = u(x_1, x_2, \cdots, x_n)$.

Its principal part is

$$L \equiv \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\frac{\partial^2}{\partial x_i \partial x_j}. \tag{13}$$

It is enough to assume that $A = [a_{ij}]$ is symmetric if not, let $\bar{a}_{ij} = \frac{1}{2}(a_{ij} + a_{ji})$ and rewrite

$$L \equiv \sum_{i=1}^{n}\sum_{j=1}^{n} \bar{a}_{ij}\frac{\partial^2}{\partial x_i \partial x_j}. \tag{14}$$

Note that $\frac{\partial^2 u}{\partial x_i \partial x_j} = \frac{\partial^2 u}{\partial x_j \partial x_i}$. As in two-space dimension, let us attach a quadratic form P with (14) (i.e., replacing $\frac{\partial u}{\partial x_i}$ by x_i).

$$P(x_1, x_2, \cdots, x_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} x_i x_j. \tag{15}$$

Since A is a real valued symmetric ($a_{ij} = a_{ji}$) matrix, it is diagonalizable with real eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ (counted with their multiplicities). In other words, there exists a corresponding set of orthonormal set of n eigenvectors, say $\sigma_1, \sigma_2, \cdots, \sigma_n$ with $R =$

$[\sigma_1, \sigma_2, \cdots, \sigma_n]$ as column vectors such that

$$R^T A R = \begin{bmatrix} \lambda_1 & & & & \\ & \lambda_2 & & \bigcirc & \\ & & \cdot & & \\ & & & \cdot & \\ & \bigcirc & & \cdot & \\ & & & & \lambda_n \end{bmatrix} = D \tag{16}$$

We now classify (12) depending on sign of eigenvalues of A:

(a) If $\lambda_i > 0 \ \forall i$ **or** $\lambda_i < 0 \ \forall i$ then (12) is elliptic type.
(b) If one **or** more of the $\lambda_i = 0$ then (12) is parabolic type.
(c) If one of the $\lambda_i < 0$ **or** $\lambda_i > 0$, and all the remaining have opposite sign then (12) is said to be of hyperbolic type.

Example 2.

1. $\nabla^2 u = u_{xx} + u_{yy} + u_{zz} = 0$. In this case, $\lambda_i = 1 > 0$ for all $i = 1, 2, 3$. It is an elliptic PDE since all eigenvalues are of one sign.
2. It is an easy exercise to check that $u_t - \nabla^2 u = 0$ is of parabolic type.
3. The equation $u_{tt} - \nabla^2 u = 0$ is of hyperbolic type.

Example 3. *Classify* $u_{x_1 x_1} + 2(1 + cx_2) u_{x_2 x_3} = 0$.

To symmetrize, write it as

$$u_{x_1 x_1} + (1 + cx_2) u_{x_2 x_3} + (1 + cx_2) u_{x_3 x_2} = 0$$

i.e., $\partial_x^T A \partial_x - c\partial_{x_2} = 0$, where

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 + cx_2 \\ 0 & 1 + cx_2 & 0 \end{bmatrix} \quad \partial_x = \begin{bmatrix} \partial_{x_1} \\ \partial_{x_2} \\ \partial_{x_3} \end{bmatrix}$$

Eigenvalues are $\lambda_1 = 1$, $\lambda_2 = 1 + cx_2$, $\lambda_3 = -(1 + cx_2)$ and normalized eigenvectors

$$\sigma_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad \sigma_2 = \begin{bmatrix} 0 \\ 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} \quad \sigma_3 = \begin{bmatrix} 0 \\ 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$$

So

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$$

Note that $R = R^T = R^{-1}$.

$$R^T AR = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 + cx_2 & 0 \\ 0 & 0 & -(1 + cx_2) \end{bmatrix} = D$$

Equation is parabolic if $x_2 = -\frac{1}{c}$ $(c \neq 0)$, hyperbolic if $x_2 > -\frac{1}{c}$ and $x_2 < -\frac{1}{c}$. For $c = 0$, $\lambda_1 = \lambda_2 = 1$ and $\lambda_3 = -1$, it is hyperbolic type.

Practice Problems

1. Classify the following equations into hyperbolic, elliptic or parabolic type.

   (A) $5u_{xx} - 3u_{yy} + (\cos x)u_x + e^y u_y + u = 0.$

   (B) $e^x u_{xx} + e^y u_{yy} - u = 0.$

   (C) $xu_{xx} + u_{yy} = 0.$

   (D) $8u_{xx} + u_{yy} - u_x + [\log(2 + x^2)]u = 0.$

   (E) $\sin^2 xu_{xx} + \sin 2xu_{xy} + \cos^2 xu_{yy} = x.$

2. Classify the following equations into elliptic, parabolic, or hyperbolic type.

   (A) $u_{xx} + 2u_{yz} + (\cos x)u_z - e^{y^2} u = \cosh z.$

   (B) $u_{xx} + 2u_{xy} + u_{yy} + 2u_{zz} - (1 + xy)u = 0.$

   (C) $e^z u_{xy} - u_{xx} = \log[x^2 + y^2 + z^2 + 1].$

3. Determine the regions where $u_{xx} - 2x^2 u_{xz} + u_{yy} + u_{zz} = 0$ is of hyperbolic, elliptic and parabolic.

Unit 14

SECOND-ORDER PARTIAL DIFFERENTIAL EQUATIONS

Canonical Forms or Normal Forms

By a suitable change of the independent variables we shall show that any equation of the form

$$Au_{xx} + Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu + G = 0, \tag{1}$$

where A, B, C, D, E, F and G are functions of the variables x and y, can be reduced to a canonical form or normal form. The transformed equation assumes a simple form so that the subsequent analysis of solving the equation will be become easy.

Consider the transformation of the indpendent variables from (x, y) to (ξ, η) given by

$$\xi = \xi(x, y), \quad \eta = \eta(x, y). \tag{2}$$

Here, the functions ξ and η are continuously differentiable and the Jacobian

$$J = \frac{\partial(\xi, \eta)}{\partial(x, y)} = \begin{vmatrix} \xi_x & \xi_y \\ \eta_x & \eta_y \end{vmatrix} = (\xi_x\eta_y - \xi_y\eta_x) \neq 0 \tag{3}$$

in the domain where (1) holds.

Using chain rule, we notice that

$$\begin{aligned}
u_x &= u_\xi\xi_x + u_\eta\eta_x \\
u_y &= u_\xi\xi_y + u_\eta\eta_y \\
u_{xx} &= u_{\xi\xi}\xi_x^2 + 2u_{\xi\eta}\xi_x\eta_x + u_{\eta\eta}\eta_x^2 + u_\xi\xi_{xx} + u_\eta\eta_{xx} \\
u_{xy} &= u_{\xi\xi}\xi_x\xi_y + u_{\xi\eta}(\xi_x\eta_y + \xi_y\eta_x) + u_{\eta\eta}\eta_x\eta_y + u_\xi\xi_{xy} + u_\eta\eta_{xy} \\
u_{yy} &= u_{\xi\xi}\xi_y^2 + 2u_{\xi\eta}\xi_y\eta_y + u_{\eta\eta}\eta_y^2 + u_\xi\xi_{yy} + u_\eta\eta_{yy}
\end{aligned}$$

Substituting these expression into (1), we obtain

$$\bar{A}(\xi_x, \xi_y)u_{\xi\xi} + \bar{B}(\xi_x, \xi_y; \eta_x, \eta_y)u_{\xi\eta} + \bar{C}(\eta_x, \eta_y)u_{\eta\eta} = F(\xi, \eta, u(\xi, \eta), u_\xi(\xi, \eta), u_\eta(\xi, \eta)), \tag{4}$$

where

$$\begin{aligned}
\bar{A}(\xi_x, \xi_y) &= A\xi_x^2 + B\xi_x\xi_y + C\xi_y^2 \\
\bar{B}(\xi_x, \xi_y; \eta_x, \eta_y) &= 2A\xi_x\eta_x + B(\xi_x\eta_y + \xi_y\eta_x) + 2C\xi_y\eta_y \\
\bar{C}(\eta_x, \eta_y) &= A\eta_x^2 + B\eta_x\eta_y + C\eta_y^2.
\end{aligned}$$

An easy calculation shows that

$$\bar{B}^2 - 4\bar{A}\bar{C} = (\xi_x\eta_y - \xi_y\eta_x)^2(B^2 - 4AC). \tag{5}$$

The equation (5) shows that the transformation of the independent variables does not modify the type of PDE.

We shall determine ξ and η so that (4) takes the simplest possible form. We now consider the following cases:

Case I: $B^2 - 4AC > 0$ (Hyperbolic type)
Case II: $B^2 - 4AC = 0$ (Parabolic type)
Case III: $B^2 - 4AC < 0$ (Elliptic type)

Case I: Note that $B^2 - 4AC > 0$ implies the equation $A\alpha^2 + B\alpha + C = 0$ has two real and distinct roots, say λ_1 and λ_2. Now, choose ξ and η such that

$$\frac{\partial \xi}{\partial x} = \lambda_1 \frac{\partial \xi}{\partial y} \text{ and } \frac{\partial \eta}{\partial x} = \lambda_2 \frac{\partial \eta}{\partial y}. \tag{6}$$

Then the coefficients of $u_{\xi\xi}$ and $u_{\eta\eta}$ will be zero because

$$\bar{A} = A\xi_x^2 + B\xi_x\xi_y + C\xi_y^2 = (A\lambda_1^2 + B\lambda_1 + C)\xi_y^2 = 0,$$
$$\bar{C} = A\eta_x^2 + B\eta_x\eta_y + C\eta_y^2 = (A\lambda_2^2 + B\lambda_2 + C)\eta_y^2 = 0.$$

Thus, (5) reduces to

$$\bar{B}^2 = (B^2 - AC)(\xi_x\eta_y - \xi_y\eta_x)^2 > 0$$

as $B^2 - 4AC > 0$. Note that (6) is a first-order linear PDE in ξ and η whose characteristics curves are satisfy the first-order ODEs

$$\frac{dy}{dx} + \lambda_i(x, y) = 0, \quad i = 1, 2. \tag{7}$$

Let the family of curves determined by the solution of (7) for $i = 1$ and $i = 2$ be

$$f_1(x, y) = c_1 \quad \text{and} \quad f_2(x, y) = c_2, \tag{8}$$

respectively. These family of curves are called characteristics curves of PDE (5). With this choice, divide (4) throughout by $\bar{B}$ (as $\bar{B} > 0$) and use (7)-(8) to obtain

$$\frac{\partial^2 u}{\partial \xi \partial \eta} = \phi(\xi, \eta, u, u_\xi, u_\eta), \tag{9}$$

which is the canonical form of hyperbolic equation.

Example 1. *Reduce the equation $u_{xx} = x^2 u_{yy}$ to its canonical form.*

Solution. Comparing with (1) we find that $A = 1$, $B = 0$, $C = -x^2$.

The roots of the equations $A\alpha^2 + B\alpha + C = 0$ i.e., $\alpha^2 + x^2 = 0$ are given by $\lambda_i = \pm x$. The differential equations for the family of characteristics curves are

$$\frac{dy}{dx} \pm x = 0.$$

whose solutions are $y + \frac{1}{2}x^2 = c_1$ and $y - \frac{1}{2}x^2 = c_2$. Choose

$$\xi = y + \frac{1}{2}x^2, \quad \eta = y - \frac{1}{2}x^2.$$

An easy computation shows that

$$\begin{aligned} u_x &= u_\xi \xi_x + u_\eta \eta_x, \\ u_{xx} &= u_{\xi\xi}\xi_x^2 + 2u_{\xi\eta}\xi_x\eta_x + u_{\eta\eta}\eta_x^2 + u_\xi \xi_{xx} + u_\eta \eta_{xx} \\ &= u_{\xi\xi}x^2 - 2u_{\xi\eta}x^2 + u_{\eta\eta}x^2 + u_\xi - u_\eta, \\ u_{yy} &= u_{\xi\xi}\xi_y^2 + 2u_{\xi\eta}\xi_y\eta_y + u_{\eta\eta}\eta_y^2 + u_\xi \xi_{yy} + u_\eta \eta_{yy}, \\ &= u_{\xi\xi} + 2u_{\xi\eta} + u_{\eta\eta}. \end{aligned}$$

Substituting these expression in the equation $u_{xx} = x^2 u_{yy}$ yields

$$\begin{aligned} & 4x^2 u_{\xi\eta} = (u_\xi - u_\eta) \\ \text{or} \quad & 4(\xi - \eta)u_{\xi\eta} = \frac{1}{4(\xi - \eta)}(u_\xi - u_\eta) \\ \text{or} \quad & u_{\xi\eta} = \frac{1}{4(\xi - \eta)}(u_\xi - u_\eta) \end{aligned}$$

which is the required canonical form.

CASE II: $B^2 - 4AC = 0 \Longrightarrow$ the equation $A\alpha^2 + B\alpha + C = 0$ has two equal roots, say $\lambda_1 = \lambda_2 = \lambda$. Let $f_1(x, y) = c_1$ be the solution of $\frac{dy}{dx} + \lambda(x, y) = 0$. Take $\xi = f_1(x, y)$ and η to be the any function of x and y which is independent of ξ.

As before, $\bar{A}(\xi_x, \xi_y) = 0$ and hence from equation (5), we obtain $\bar{B} = 0$. Note that $\bar{C}(\eta_x, \eta_y) \neq 0$, otherwise η would be a function of ξ. Dividing (4) by $\bar{C}$, the canonical form of (2) is

$$u_{\eta\eta} = \phi(\xi, \eta, u, u_\xi, u_\eta). \tag{10}$$

which is the canonical form of parabolic equation.

Example 2. *Reduce the equation $u_{xx} + 2u_{xy} + u_{yy} = 0$ to canonical form.*

Solution. In this case, $A = 1$, $B = 2$, $C = 1$. The equation $\alpha^2 + 2\alpha + 1 = 0$ has equal roots $\lambda = -1$. The solution of $\frac{dy}{dx} - 1 = 0$ is $x - y = c_1$ Take $\xi = x - y$. Choose $\eta = x + y$. proceed as in Example 1 to obtain $u_{\eta\eta} = 0$ which is the canonical form of the given PDE.

CASE III: When $B^2 - 4AC < 0$, the roots of $A\alpha^2 + B\alpha + C = 0$ are complex. Following the procedure as in CASE I, we find that

$$u_{\xi\eta} = \phi_1(\xi, \eta, u, u_\xi, u_\eta). \tag{11}$$

The variables ξ, η are infact complex conjugates. To get a real canonical form use the transformation

$$\alpha = \frac{1}{2}(\xi + \eta), \quad \beta = \frac{1}{2i}(\xi - \eta)$$

to obtain

$$u_{\xi\eta} = \frac{1}{4}(u_{\alpha\alpha} + u_{\beta\beta}), \tag{12}$$

which follows from the following calculation:

$$\begin{aligned} u_\xi &= u_\alpha \alpha_\xi + u_\beta \beta_\xi = \frac{1}{2}u_\alpha + \frac{1}{2i}u_\beta \\ u_{\xi\eta} &= \frac{1}{2}(u_{\alpha\alpha}\alpha_\eta + u_{\alpha\beta}\beta_\eta) + \frac{1}{2i}(u_{\beta\alpha}\alpha_\eta + u_{\beta\beta}\beta_\eta) \\ &= \frac{1}{4}(u_{\alpha\alpha} + u_{\beta\beta}). \end{aligned}$$

The desired canonical form is

$$u_{\alpha\alpha} + u_{\beta\beta} = \psi(\alpha, \beta, u(\alpha, \beta), u_\alpha(\alpha, \beta), u_\beta(\alpha, \beta)). \tag{13}$$

Example 3. *Reduce the equation $u_{xx} + x^2 u_{yy} = 0$ to canonical form.*

Solution. In this case, $A = 1$, $B = 0$, $C = x^2$. The roots are $\lambda_1 = ix$, $\lambda_2 = -ix$. Take $\xi = iy + \frac{1}{2}x^2$, $\eta = -iy + \frac{1}{2}x^2$. Then $\alpha = \frac{1}{2}x^2$, $\beta = y$. The canonical form is

$$u_{\alpha\alpha} + u_{\beta\beta} = -\frac{1}{2\alpha}u_\alpha.$$

Practice Problems

1. Reduce the following equations to canonical/normal form:

   (A) $2u_{xx} - 4u_{xy} + 2u_{yy} + 3u = 0$.

   (B) $u_{xx} + yu_{yy} = 0$.

   (C) $u_{xy} + u_x + u_y = 2x$.

2. Show that the equation

   $$u_{xx} - 6u_{xy} + 12u_{yy} + 4u_x - u = \sin(xy)$$

   is of elliptic type and obtain its canonical form.

3. Determine the regions where Tricomi's equation $u_{xx} + xu_{yy} = 0$ is of elliptic, parabolic, and hyperbolic types. Obtain its characteristics and its canonical form in the hyperbolic region.

Unit 15

SECOND-ORDER PARTIAL DIFFERENTIAL EQUATIONS

Superposition Principle and Wellposedness

A very important fact concerning linear PDEs is the superposition principle, which is stated below.

A linear PDE can be written in the form

$$L[u] = f, \tag{1}$$

where $L[u]$ denotes a linear combination of u and some of its partial derivatives, with coefficients which are given functions of the independent variables.

Definition 1. (Superposition principle) *Let u_1 be a solution of the linear PDE*

$$L[u] = f_1$$

and let u_2 be a solution of the linear PDE

$$L[u] = f_2.$$

Then, for any any constants c_1 and c_2, $c_1u_1 + c_2u_2$ is a solution of

$$L[u] = c_1f_1 + c_2f_2.$$

That is,

$$L[c_1u_1 + c_2u_2] = c_1f_1 + c_2f_2. \tag{2}$$

In particular, when $f_1 = 0$ and $f_2 = 0$, (2) implies that if u_1 and u_2 are solutions of the homogeneous linear PDE $L[u] = 0$, then $c_1u_1 + c_2u_2$ will also be a solution of $L[u] = 0$.

Example 2. *Observe that $u_1(x, y) = x^3$ is a solution of the linear PDE $u_{xx} - u_y = 6x$, and $u_2(x, y) = y^2$ is a solution of $u_{xx} - u_y = -2y$. Then, using superposition principle, it is easy to verify that $3u_1(x, y) - 4u_2(x, y)$ will be a solution of $u_{xx} - u_y = 18x + 8y$.*

Remark 3. *Note that the principle of superposition is not valid for nonlinear partial differential equations. This failure makes it difficult to form families of new solutions from an original pair of solutions.*

Example 4. *Consider the nonlinear first order PDE $u_xu_y - u(u_x + u_y) + u^2 = 0$. Note that e^x and e^y are two solutions of this equation. However, $c_1e^x + c_2e^y$ will not be a solution, unless $c_1 = 0$ or $c_2 = 0$.*

Solution. Define $D[u] := (u_x - u)(u_y - u)$. For any $u, v \in C^1$, we have

$$\begin{aligned} D[u + v] &= (u_x + v_x - u - v)(u_y + v_y - u - v) \\ &= D[u] + D[v] + (u_y - u)(v_x - v) + (u_x - u)(v_y - v). \end{aligned}$$

The computation shows that $D[u+v] \neq D[u]+D[v]$ in general. Taking $u = c_1e^x$ and $v = c_2e^y$, an easy computation shows that

$$D[c_1e^x + c_2e^y] = D[c_1e^x] + D[c_2e^y] + (-c_1e^x)(-c_2e^y) = c_1c_2e^{x+y}.$$

Thus, $D[c_1e^x + c_2e^y] = 0$ only if $c_1 = 0$ or $c_2 = 0$.

1 Well-posed problems

A set of conditions was proposed by Hadamard (cf. [12]), who listed three requirements that must be met when formulating an initial and /or boundary value problem. A problem for which the PDE and the data lead to a solution is said to be well posed or correctly posed if the following three conditions are satisfies:

Hadamard's conditions for a well-posed problem are:

1. The solution must exist.
2. The solution should be unique.
3. The solution should depend continuously on the initial and/or boundary data.

If it fails to meet these requirements, it is incorrectly posed.

The conditions (1)-(2) require that the equation plus the data for the problem must be such that one and only one solution exists. The third condition states that a small variation of the data for the problem should cause small variation in the solution. As data are generally obtained experimentally and may be subject to numerical approximations, we require that the solution be stable under small variations in initial and/or boundary values. That is, we cannot allow large variations to occur in the solution if the data are altered slightly.

A simple example of a ill posed problem is given below.

Example 5. *Consider Cauchy's problem for Laplace's equation in $y \geq 0$:*

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \tag{3}$$

$$u(x,0) = 0, \tag{4}$$

$$u_y(x,0) = \frac{1}{n}\sin nx, \tag{5}$$

where n is a positive integer, is not well-posed.

The solution is given by $u(x,y) = \frac{1}{n^2}\sin(nx)\sinh(ny)$. Now, as $n \to \infty$, $u_y(x,0) \to 0$ so that for large n the Cauchy data $u(x,0)$ and $u_y(x,0)$ can be made arbitrarily small in magnitude. However, the solution $u(x,y)$ oscillates with an amplitude that grows exponentially like e^{ny} as $n \to \infty$. Thus, arbitrarily small data can lead to arbitrarily large variation in solutions and hence the solution is unstable. This violates the condition (3) i.e., the continuous dependence of the solution on the data.

Boundary value problems are not well posed for hyperbolic and parabolic equations. This follows because these are, in general, equations whose solutions evolve in time and their behavior at later times is predicted by their previous states.

Example 6. *Consider the hyperbolic equation*

$$u_{xy} = 0 \ \ in \ \ 0 < x < 1, \ \ 0 < y < 1$$

with the boundary conditions

$$u(x,0) = f_1(x), \ u(x,1) = f_2(x) \ \ for\, 0 \le x \le 1,$$
$$u(0,y) = g_1(y), \ u(1,y) = g_2(y) \ \ for\, 0 \le y \le 1.$$

We shall show that this problem has no solution if the data are prescribed arbitrarily. Since $u_{xy} = 0$ implies that $u_x(x,y) =$ constant, we have

$$u_x(x,0) = u_x(x,1).$$

In view of the given BC, we have

$$u_x(x,0) = f_1'(x) \quad \text{and} \quad u_x(x,1) = f_2'(x).$$

Thus, unless $f_1(x)$ and $f_2(x)$ are prescribed such that $f_1'(x) = f_2'(x)$, the BVP cannot be solved. Therefore, it is incorrectly posed.

2 Method of factorization

There is no general methods are available for obtaining the general solutions of second-order PDEs. Sometimes PDE of second-order can be factorized into two first-order equations. The equations

$$u_{\xi\eta} = 0,$$
$$yu_{xx} + (x+y)u_{xy} + xu_{yy} = 0.$$

are examples of such equation. It is often much easier to factorize an equation when in its canonical form. But, we can often factorize equations with constant coefficients directly. The method of factorization can be a useful method of solution for hyperbolic and parabolic equations.

Example 7. *The equation*

$$u_{xx} - u_{yy} + 4(u_x + u) = 0$$

can be written as

$$\left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y} + 2\right)\left(\frac{\partial}{\partial x} - \frac{\partial}{\partial y} + 2\right)u = 0.$$

It is equivalent to the pair of first order equations

$$u_x - u_y + 2u = v,$$

and

$$v_x + v_y + 2v = 0.$$

Example 8. *The hyperbolic equation*

$$acu_{xy} + au_x + cu_y + u = 0$$

can be written as

$$\left(a\frac{\partial}{\partial x} + 1\right)\left(c\frac{\partial}{\partial y} + 1\right)u = 0.$$

It is equivalent to

$$\begin{aligned} cu_y + u &= v, \\ av_x + v &= 0. \end{aligned}$$

Note: Unlike the case when the coefficients are constant, the differential operators need not commute.

Practice Problems

1. If $u_1(x, y) = x^3$ solves $u_{xx} + u_{yy} = 2$ and $u_2(x, y) = c^3 + dy^3$ solves $u_{xx} + u_{yy} = 6cx + 6dy$ for real constants c and d then find a solution of $u_{xx} + u_{yy} = ax + by + c$ for given real constants a, b and c.

2. Let $u_1(x, y)$ be the solution to the Cauchy problem

$$\begin{aligned} u_{xx} + u_{yy} &= 0, \\ u(x, 0) &= f(x), \\ u_y(x, 0) &= g(x), \end{aligned}$$

Unit 16

HEAT EQUATION

Modeling the Heat Equation

We shall derive heat equation from the principle of conservation of energy and the fact that heat flows from hot regions to cold regions.

Consider a wire or rod of length L which is made of some heat-conducting material and is insulated on the outside, except possibly over the ends at $x = 0$ and $x = L$. Let $u(x,t)$ denote the temperature at x at time t. $u(x,t)$ is assumed to be constant on each cross section at each time. By the principle of conservation of energy (heat energy), the

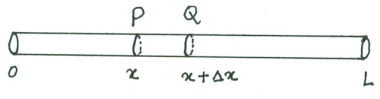


Figure 5.1: A thin rod of length L

net change of heat inside the segment PQ (between x and $x + \Delta x$) is equal to the net heat flux across the boundaries and the total heat generated inside PQ. If c is thermal capacity of the rod, ρ is the density of the rod, A is the cross-section area of the rod, k is thermal conductivity of the rod and $f(x,t)$ is the external heat source, then we calculate these terms as follows:

Total amount of heat inside the segment PQ at time $t = \int_x^{x+\Delta x} c\rho A u(\tau,t)d\tau$.

Net change of heat inside $PQ = \dfrac{d}{dt}\int_x^{x+\Delta x} c\rho A u(\tau,t)ds = c\rho A \int_x^{x+\Delta x} u_t(\tau,t)d\tau$.

Net flux of heat across the boundaries $= kA[u_x(x+\Delta x,t) - u_x(x,t)]$.

Heat generated due to external heat source inside $PQ = A\int_x^{x+\Delta x} f(\tau,t)d\tau$.

By the principle of conservation of energy, we write

$$\frac{d}{dt}\int_x^{x+\Delta x} c\rho A u(\tau,t)d\tau = c\rho A\int_x^{x+\Delta x} u_t(\tau,t)d\tau$$
$$= kA[u_x(x+\Delta x,t) - u_x(x,t)] + A\int_x^{x+\Delta x} f(\tau,t)d\tau. \qquad (1)$$

Applying Mean Value Theorem for integral[1], we obtain

$$c\rho A u_t(\xi_1, t)\Delta x = kA[u_x(x+\Delta x, t) - u_x(x,t)] + Af(\xi_2, t)\Delta x,$$

where $\xi_1, \xi_2 \in (x, x+\Delta x)$, and hence,

$$u_t(\xi_1, t) = \frac{k}{c\rho}\left[\frac{u_x(x+\Delta x, t) - u_x(x,t)}{\Delta x}\right] + \frac{1}{c\rho} f(\xi_2, t).$$

Now, letting $\Delta x \to 0$, we arrive at

$$\boxed{u_t(x,t) = \alpha^2 u_{xx}(x,t) + F(x,t),} \tag{2}$$

where $\alpha^2 = k/(c\rho)$ is called the thermal diffusivity of the rod and $F(x,t) = \frac{1}{c\rho} f(x,t)$ is called the heat source density.

Remark 1.

- When the rod is not laterally insulated and we allow the heat to flow in and out across the lateral boundary at a rate proportional to the difference between the temperature $u(x,t)$ and the surrounding medium, the conservation of heat principle yields

  $$u_t = \alpha^2 u_{xx} - \beta(u - u_0), \quad \beta > 0.$$

  The heat loss ($u > u_0$) or gain ($u < u_0$) is proportional to the difference between the temperature $u(x,t)$ of the rod and the surrounding medium u_0. Here, β is the constant of proportionality.

- If the material of the rod is uniform, then k is independent of x. For some materials, the value of k depends on the temperature u and hence the resulting heat equation

  $$u_t = \frac{1}{c\rho}\frac{\partial}{\partial x}\left\{k(u)\frac{\partial u}{\partial x}\right\}$$

  is nonlinear.

- If the material is nonhomogeneous the diffusion within the rod depends on x. For example, suppose the half of the rod is made of copper and other half is made of steel, then the PDE that describes the heat flow is given by

  $$u_t = \alpha^2(x) u_{xx}, \;\; 0 < x < L,$$

[1] If $f(x)$ is continuous on $[a,b]$, then there exists at least one number ξ in (a,b) such that

$$\int_a^b f(x)dx = f(\xi)(b-a).$$

with

$$\alpha(x) = \begin{cases} \alpha_1, & 0 < x < L/2, \\ \alpha_2, & L/2 < x < L, \end{cases}$$

where α_1 and α_2 are the thermal diffusivity coefficients of copper and steel, respectively.

Types of BCs: There are three types of boundary conditions that can occur for heat flow problems. They are

- *Dirichlet boundary conditions* (temperature is specified on the boundary):

  Consider heat flow problem in a rod ($0 \le x \le L$). The specification of the temperatures $u(0,t)$ and $u(L,t)$ at the ends are classified as Dirichlet type BC.

- *Neumann boundary conditions* (heat flow across the boundary is specified):

  The specification of the normal derivative (i.e., $\frac{\partial u}{\partial n}$, where n is the outward normal to the boundary) on the boundary is classified as Neumann type BCs. For instance, if the end points of a rod is insulated (i.e., we do not allow any flow of heat across the boundary), the BCs are

  $$u_x(0,t) = 0, \qquad u_x(L,t) = 0, \quad 0 < t < \infty.$$

- *Robin's or Mixed boundary conditions*:

  If the condition on the boundary is a mixture of both Dirichlet and Neumann types i.e.,

  $$\frac{\partial u}{\partial n} = -h(u - g(t))$$

  then it is called Robin's BCs or mixed BCs. Here, h is a constant and $g(t)$ is given function that can vary over the boundary. The mixed BCs may be interpreted as the inward flux across the boundary is proportional to the difference between the temperature u and some specified temperature g. If the temperature u on the boundary is greater than the boundary temperature, then the flow of heat is outward. If u is less than the specified boundary temperature g, then heat flows inward.

The Maximum and Minimum Principle

In this lecture, we shall prove the maximum and minimum properties of the heat equation. These properties can be used to prove uniqueness and continuous dependence on data of the solutions of these equations.

To begin with, we shall first prove the maximum principle for the inhomogeneous heat equation ($F \neq 0$).

Theorem 1. (The maximum principle) *Let $R : 0 \leq x \leq L,\ 0 \leq t \leq T$ be a closed region and let $u(x,t)$ be a solution of*

$$u_t - \alpha^2 u_{xx} = F(x,t) \quad (x,t) \in R, \tag{1}$$

which is continuous in the closed region R. If $F < 0$ in R, then $u(x,t)$ attains its maximum values on $t = 0$, $x = 0$ or $x = L$ and not in the interior of the region or at $t = T$. If $F > 0$ in R, then $u(x,t)$ attains its minimum values on $t = 0$, $x = 0$ or $x = L$ and not in the interior of the region or at $t = T$.

Proof. We shall show that if a maximum or minimum occurs at an interior point $0 < x_0 < l$ and $0 < t_0 \leq T$, then we will arrive at contradiction. Let us consider the following cases.

Case I: First, consider the case with $F < 0$. Since $u(x,t)$ is continuous in a closed and bounded region in R, $u(x,t)$ must attain its maximum in R. Let (x_0, t_0) be the interior maximum point. Then, we must have

$$u_{xx}(x_0,t_0) \leq 0, \qquad u_t(x_0,t_0) \geq 0. \tag{2}$$

Since $u_x(x_0,t_0) = 0 = u_t(x_0,t_0)$, we have

$$u_t(x_0,t_0) = 0 \quad \text{if } t_0 < T.$$

If $t_0 = T$, the point $(x_0,t_0) = (x_0,T)$ is on the boundary of R, then we claim that

$$u_t(x_0,t_0) \geq 0$$

as u may be increasing at (x_0,t_0). Substituting (2) in (1), we find that the left side of the equation (1) is non-negative while the right side is strictly negative. This leads to a contradiction and hence, the maximum must be assumed on the initial line or on the boundary.

Case II: Consider the case with $F > 0$. Let there be an interior minimum point (x_0, t_0) in R such that

$$u_{xx}(x_0, t_0) \geq 0, \qquad u_t(x_0, t_0) \leq 0. \tag{3}$$

Note that the inequalities (3) is same as (2) with the signs reversed. Again arguing as before, this leads to a contradiction, hence the minimum must be assumed on the initial line or on the boundary.

Note: When $F = 0$ i.e., for homogeneous equation, the inequalities (2) at a maximum or (3) at a minimum do not leads to a contradiction when they are inserted into (1) as u_{xx} and u_t may both vanish at (x_0, t_0).

Below, we present a proof of the maximum principle for the homogeneous heat equation.

Theorem 2. (The maximum principle) *Let $u(x, t)$ be a solution of*

$$u_t = \alpha^2 u_{xx} \quad 0 \leq x \leq L, \;\; 0 < t \leq T, \tag{4}$$

which is continuous in the closed region $R : 0 \leq x \leq L$ and $0 \leq t \leq T$. The maximum and minimum values of $u(x, t)$ are assumed on the initial line $t = 0$ or at the points on the boundary $x = 0$ or $x = L$.

Proof. Let us introduce the auxiliary function

$$v(x, t) = u(x, t) + \epsilon x^2, \tag{5}$$

where $\epsilon > 0$ is a constant and u satisfies (4). Note that $v(x, t)$ is continuous in R and hence it has a maximum at some point (x_1, t_1) in the region R.

Assume that (x_1, t_1) is an interior point with $0 < x_1 < L$ and $0 < t_1 \leq T$. Then we find that

$$v_t(x_1, t_1) \geq 0, \quad v_{xx}(x_1, t_1) \leq 0. \tag{6}$$

Since u satisfies (4), we have

$$v_t - \alpha^2 v_{xx} = u_t - \alpha^2 u_{xx} - 2\alpha^2 \epsilon = -2\alpha^2 \epsilon < 0. \tag{7}$$

Substituting (6) into (4) and using (7) now leads to

$$0 \leq v_t - \alpha^2 v_{xx} < 0,$$

which is a contradiction since the left side is non-negative and the right side is strictly negative. Therefore, $v(x, t)$ assumes its maximum on the initial line or on the boundary since v satisfies (1) with $F < 0$.

Let

$$M = \max\{u(x,t)\} \text{ on } t = 0, x = 0, \text{ and } x = L,$$

i.e., M is the maximum value of u on the initial line and boundary lines. Then

$$v(x,t) = u(x,t) + \epsilon x^2 \leq M + \epsilon L^2, \text{ for } 0 \leq x \leq L,\ 0 \leq t \leq T. \tag{8}$$

Since v has its maximum on $t = 0$, $x = 0$, or $x = L$, we obtain

$$u(x,t) = v(x,t) - \epsilon x^2 \leq v(x,t) \leq M + \epsilon L^2. \tag{9}$$

Since ϵ is arbitrary, letting $\epsilon \to 0$, we conclude that

$$u(x,t) \leq M \text{ for all } (x,t) \in R, \tag{10}$$

and this completes the proof.

Remark 3.

- The minimum principle for the heat equation can be obtained by replacing the function $u(x,t)$ by $-u(x,t)$, where $u(x,t)$ is a solution of (4). Clearly, $-u$ is also a solution of (4) and the maximum values of u correspond to the minimum values of u. Since u satisfies the maximum principle, we conclude that u assumes its min-minimum values on the initial line or on the boundary lines. In particular, this implies that if the initial and boundary data for the problem are non- negative, then the solution must be non-negative.

- In geometrical term, the maximum principle states that if a solution of the problem (4) is graphed in the xtu-space, then the surface $u = u(x,t)$ achieves its maximum height above one of the three sides $x = 0, x = L, t = 0$ of the rectangle $0 \leq x \leq L$, $0 \leq t \leq T$.

- From a physical perspective, the maximum principle states that the temperature, at any point x inside the rod at any time t $(0 \leq t \leq T)$, is less than the maximum of the initial temperature distribution or the maximum of the temperatures prescribed at the ends during the time interval $[0, T]$.

1 Uniqueness and continuous dependence

As a consequence of the maximum principle, we can show that the heat flow problem has a unique solution and depend continuously on the given initial and boundary data.

Theorem 4. (Uniqueness result) *Let $u_1(x,t)$ and $u_2(x,t)$ be solutions of the following problem*

$$\begin{aligned} PDE: \quad & u_t = \alpha^2 u_{xx}, \quad 0 < x < L, \ \ t > 0, \\ BC: \quad & u(0,t) = g(t), \quad u(L,t) = h(t), \\ IC: \quad & u(x,0) = f(x), \end{aligned} \tag{11}$$

where $f(x)$, $g(t)$ and $h(t)$ are given functions. Then $u_1(x,t) = u_2(x,t)$, for all $0 \le x \le L$ and $t \ge 0$.

Proof. Let $u_1(x,t)$ and $u_2(x,t)$ be two solutions of (11). Set $w(x,t) = u_1(x,t) - u_2(x,t)$. Then w satisfies

$$\begin{aligned} & w_t = \alpha^2 w_{xx} \quad 0 < x < L, \ \ t > 0, \\ & w(0,t) = 0, \quad w(L,t) = 0, \\ & w(x,0) = 0. \end{aligned}$$

By the maximum principle (cf. Theorem 2), we must have

$$w(x,t) \le 0 \Longrightarrow u_1(x,t) \le u_2(x,t), \ \text{ for all } 0 \le x \le L, \ \ t \ge 0.$$

A similar argument with $\bar{w} = u_2 - u_1$ yields

$$u_2(x,t) \le u_1(x,t) \ \text{ for all } 0 \le x \le L, \ \ t \ge 0.$$

Therefore, we have

$$u_1(x,t) = u_2(x,t) \ \text{ for all } 0 \le x \le L, \ \ t \ge 0,$$

and this completes the proof.

Theorem 5. (Continuous Dependence on the IC and BC) *Let $u_1(x,t)$ and $u_2(x,t)$, respectively, be solutions of the problems*

$$\begin{aligned} u_t = \alpha^2 u_{xx}; \qquad & u_t = \alpha^2 u_{xx} \\ u(0,t) = g_1(t) \ \ u(L,t) = h_1(t); \qquad & u(0,t) = g_2(t) \ \ u(L,t) = h_2(t) \\ u(x,0) = f_1(x); \qquad & u(x,0) = f_2(x), \end{aligned} \tag{12}$$

in the region $0 \le x \le L$, $t \ge 0$. If

$$|f_1(x) - f_2(x)| \le \epsilon \ \textit{ for all } x, \ 0 \le x \le L,$$

and

$$|g_1(t) - g_2(t)| \le \epsilon \ \text{ and } \ |h_1(t) - h_2(t)| \le \epsilon \ \text{ for all } t,\ 0 \le t \le T,$$

for some $\epsilon \ge 0$*, then we have*

$$|u_1(x,t) - u_2(x,t)| \le \epsilon \ \text{ for all } x \text{ and } t, \text{ where } \ 0 \le x \le L,\ 0 \le t \le T.$$

Proof. Let $v(x,t) = u_1(x,t) - u_2(x,t)$. Then $v_t = \alpha^2 v_{xx}$ and we obtain

$$\begin{aligned} |v(x,0)| &= |f_1(x) - f_2(x)| \le \epsilon, \ \ 0 \le x \le L, \\ |v(0,t)| &= |g_1(t) - g_2(t)| \le \epsilon, \ \ 0 \le t \le T, \\ |v(L,t)| &= |h_1(t) - h_2(t)| \le \epsilon, \ \ 0 \le t \le T. \end{aligned}$$

Note that the maximum of v on $t = 0$ $(0 \le x \le L)$ and $x = 0$ and $x = L$ $(0 \le t \le T)$ is not greater than ϵ. The minimum of v on these boundary lines is not less than $-\epsilon$. Hence, the maximum/minimum principle yields

$$-\epsilon \le v(x,t) \le \epsilon \implies |u_1(x,t) - u_2(x,t)| = |v(x,t)| \le \epsilon.$$

Note: (i) We observe that when $\epsilon = 0$, the problems in (12) are identical. We conclude that $|u_1(x,t) - u_2(x,t)| \le 0$ (i.e. $u_1 = u_2$). This proves the uniqueness result.

(ii) Suppose a certain initial/boundary value problem has a unique solutions. Then a small change in the initial and/or boundary conditions yields a small change in the solutions.

For the inhomogeneous equation (1), we have seen that the maximum or minimum values must be attained either on the initial line or the boundary lines and that they cannot be assumed in the interior. This result is known as a strong maximum or minimum principle.

Theorem 6. (Strong maximum principle) *Let* $u(x,t)$ *be a solution of the heat equation in the rectangle* $R \ : \ 0 \le x \le L,\ 0 \le t \le T$*. If* $u(x,t)$ *achieves its maximum at* (x^*, T)*, where* $0 < x^* < L$*, then* u *must be constant in* R*.*

Practice Problems

1. Use the maximum/minimum principle to show that the solution u of the problem

$$\begin{aligned} &u_t = u_{xx}, \ \ 0 < x < \pi, \ \ t > 0, \\ &u_x(0,t) = 0, \ u_x(\pi,t) = 0, \quad t > 0, \\ &u(x,0) = \sin(x) + \frac{1}{2}\sin(2x), \ \ 0 \le x \le \pi \end{aligned}$$

satisfies $0 \le u(x,t) \le \frac{3\sqrt{3}}{4}$, $t \ge 0$.

Unit 17

HEAT EQUATION

Method of Separation of Variables

Separation of variables is one of the oldest technique for solving initial-boundary value problems (IBVP) and applies to problems, where

- PDE is linear and homogeneous (not necessarily constant coefficients) and
- BC are linear and homogeneous.

Basic Idea: To seek a solution of the form

$$u(x,t) = X(x)T(t),$$

where $X(x)$ is some function of x and $T(t)$ in some function of t. The solutions are simple because any temperature $u(x,t)$ of this form will retain its basic "shape" for different values of time t. The separation of variables reduced the problem of solving the PDE to solving the two ODEs: One second order ODE involving the independent variable x and one first order ODE involving t. These ODEs are then solved using given initial and boundary conditions.

To illustrate this method, let us apply to a specific problem. Consider the following IBVP:

$$\text{PDE:} \qquad u_t = \alpha^2 u_{xx}, \quad 0 \le x \le L, \ 0 < t < \infty, \tag{1}$$

$$\text{BC:} \qquad u(0,t) = 0 \ \ u(L,t) = 0, \quad 0 < t < \infty, \tag{2}$$

$$\text{IC:} \qquad u(x,0) = f(x), \quad 0 \le x \le L. \tag{3}$$

Step 1:(Reducing to the ODEs) Assume that equation (1) has solutions of the form

$$\boxed{u(x,t) = X(x)T(t),}$$

where X is a function of x alone and T is a function of t alone. Note that

$$u_t = X(x)T'(t) \quad \text{and} \quad u_{xx} = X''(x)T(t).$$

Now, substituting these expression into $u_t = \alpha^2 u_{xx}$ and separating variables, we obtain

$$X(x)T'(t) = \alpha^2 X''(x)T(t)$$

$$\Rightarrow \quad \frac{T'(t)}{\alpha^2 T(t)} = \frac{X''(x)}{X(x)}.$$

Since a function of t can equal a function of x only when both functions are constant. Thus,

$$\frac{T'(t)}{\alpha^2 T(t)} = \frac{X''(x)}{X(x)} = c$$

for some constant c. This leads to the following two ODEs:

$$T'(t) - \alpha^2 cT(t) = 0, \tag{4}$$

$$X''(x) - cX(x) = 0. \tag{5}$$

Thus, the problem of solving the PDE (1) is now reduced to solving the two ODEs.

Step 2:(Applying BCs)
Since the product solutions $u(x,t) = X(x)T(t)$ are to satisfy the BC (2), we have

$$u(0,t) = X(0)T(t) = 0 \quad \text{and} \quad X(L)T(t) = 0, \quad t > 0.$$

Thus, either $T(t) = 0$ for all $t > 0$, which implies that $u(x,t) = 0$, or $X(0) = X(L) = 0$. Ignoring the trivial solution $u(x,t) = 0$, we combine the boundary conditions $X(0) = X(L) = 0$ with the differential equation for X in (5) to obtain the BVP:

$$X''(x) - cX(x) = 0, \quad X(0) = X(L) = 0. \tag{6}$$

There are three cases: $c < 0$, $c > 0$, $c = 0$ which will be discussed below. It is convenient to set $c = -\lambda^2$ when $c < 0$ and $c = \lambda^2$ when $c > 0$, for some constant $\lambda > 0$.

Case 1. ($c = \lambda^2 > 0$ for some $\lambda > 0$). In this case, a general solution to the differential equation (5) is

$$X(x) = C_1 e^{\lambda x} + C_2 e^{-\lambda x},$$

where C_1 and C_2 are arbitrary constants. To determine C_1 and C_2, we use the BC $X(0) = 0, \ X(L) = 0$ to have

$$X(0) = C_1 + C_2 = 0, \tag{7}$$

$$X(L) = C_1 e^{\lambda L} + C_2 e^{-\lambda L} = 0. \tag{8}$$

From the first equation, it follows that $C_2 = -C_1$. The second equation leads to

$$\begin{aligned} & C_1(e^{\lambda L} - e^{-\lambda L}) = 0, \\ \Rightarrow \quad & C_1(e^{2\lambda L} - 1) = 0, \\ \Rightarrow \quad & C_1 = 0. \end{aligned}$$

since $(e^{2\lambda L}-1) > 0$ as $\lambda > 0$. Therefore, we have $C_1 = 0$ and hence $C_2 = 0$. Consequently $X(x) = 0$ and this implies $u(x,t) = 0$ i.e., there is no nontrivial solution to (5) for the case $c > 0$.

Case 2. (when c=0)
The general solution solution to (5) is given by

$$X(x) = C_3 + C_4 x.$$

Applying BC yields $C_3 = C_4 = 0$ and hence $X(x) = 0$. Again, $u(x,t) = X(x)T(t) = 0$. Thus, there is no nontrivial solution to (5) for $c = 0$.

Case 3. (When $c = -\lambda^2 < 0$ for some $\lambda > 0$)

The general solution to (5) is

$$X(x) = C_5 \cos(\lambda x) + C_6 \sin(\lambda x).$$

This time the BC $X(0) = 0,\ \ X(L) = 0$ gives the system

$$\begin{aligned} C_5 &= 0, \\ C_5 \cos(\lambda L) + C_6 \sin(\lambda L) &= 0. \end{aligned}$$

As $C_5 = 0$, the system reduces to solving $C_6 \sin(\lambda L) = 0$. Hence, either $\sin(\lambda L) = 0$ or $C_6 = 0$. Now

$$\sin(\lambda L) = 0 \quad \Longrightarrow \quad \lambda L = n\pi, \quad n = 0, \pm 1, \pm 2, \ldots.$$

Therefore, (5) has a nontrivial solution ($C_6 \neq 0$) when

$$\lambda L = n\pi \quad \text{or} \quad \lambda = \frac{n\pi}{L}, \ \ n = 1, 2, 3, \ldots.$$

Here, we exclude $n = 0$, since it makes $c = 0$. Therefore, the nontrivial solutions (eigenfunctions) X_n corresponding to the eigenvalue $c = -\lambda^2$ are given by

$$X_n(x) = a_n \sin(\frac{n\pi x}{L}), \tag{9}$$

where a_n's are arbitrary constants.

Step 3:(Applying IC)

Let us consider solving equation (4). The general solution to (4) with $c = -\lambda^2 = (\frac{n\pi}{L})^2$ is

$$T_n(t) = b_n e^{-\alpha^2 (\frac{n\pi}{L})^2 t}.$$

Combing this with (9), the product solution $u(x,t) = X(x)T(t)$ becomes

$$\begin{aligned} u_n(x,t) &:= X_n(x)T_n(t) = a_n \sin(\frac{n\pi x}{L}) b_n e^{-\alpha^2(\frac{n\pi}{L})^2 t} \\ &= c_n e^{-\alpha^2(\frac{n\pi}{L})^2 t} \sin(\frac{n\pi x}{L}), \quad n = 1, 2, 3, \ldots, \end{aligned}$$

where c_n is an arbitrary constant.

Since the problem (9) is linear and homogeneous, an application of superposition principle gives

$$\boxed{u(x,t) = \sum_{n=1}^{\infty} u_n(x,t) = \sum_{n=1}^{\infty} c_n e^{-\alpha^2(\frac{n\pi}{L})^2 t} \sin(\frac{n\pi x}{L}),} \tag{10}$$

which will be a solution to (1)-(3), provided the infinite series has the proper convergence behavior.

Since the solution (10) is to satisfy IC (3), we must have

$$u(x,0) = \sum_{n=1}^{\infty} c_n \sin\left(\frac{n\pi x}{L}\right) = f(x), \quad 0 < x < L.$$

Thus, if $f(x)$ has an expansion of the form

$$f(x) = \sum_{n=1}^{\infty} c_n \sin\left(\frac{n\pi x}{L}\right), \tag{11}$$

which is called a Fourier sine series (FSS) with c_n's are given by the formula

$$c_n = \frac{2}{L} \int_0^L f(x) \sin(\frac{n\pi x}{L}) dx. \tag{12}$$

Then the infinite series (10) with the coefficients c_n given by (12) is a solution to the problem (1)-(3).

Example 1. *Find the solution to the following IBVP:*

$$u_t = 3u_{xx} \quad 0 \le x \le \pi, \ 0 < t < \infty, \tag{13}$$

$$u(0,t) = u(\pi,t) = 0, \quad 0 < t < \infty, \tag{14}$$

$$u(x,0) = 3\sin 2x - 6\sin 5x, \quad 0 \le x \le \pi. \tag{15}$$

Solution. Comparing (13) with (1), we notice that $\alpha^2 = 3$ and $L = \pi$. Using formula (10), we write a solution $u(x,t)$ as

$$u(x,t) = \sum_{n=1}^{\infty} c_n e^{-3n^2 t} \sin(nx).$$

To determine c_n's, we use IC (15) to have

$$u(x,0) = 3\sin 2x - 6\sin 5x = \sum_{n=1}^{\infty} c_n \sin(nx).$$

Comparing the coefficients of like terms, we obtain

$$c_2 = 3 \quad \text{and} \quad c_5 = -6,$$

and the remaining c_n's are zero. Hence, the solution to the problem (13)-(15) is

$$\begin{aligned} u(x,t) &= c_2 e^{-3(2)^2 t}\sin(2x) + c_5 e^{-3(5)^2 t}\sin(5x) \\ &= 3e^{-12t}\sin(2x) - 6e^{-75t}\sin(5x). \end{aligned}$$

Practice Problems

1. Solve the following IBVP:

$$\begin{aligned} &u_t = 16u_{xx}, \ \ 0 < x < 1, \ \ t > 0, \\ &u(0,t) = 0, \ u(1,t) = 0, \quad t > 0, \\ &u(x,0) = (1-x)x, \ \ 0 < x < 1. \end{aligned}$$

2. Solve the following IBVP:

$$\begin{aligned} &u_t = u_{xx}, \ \ 0 < x < \pi, \ \ t > 0, \\ &u_x(0,t) = u_x(\pi,t) = 0, \quad t > 0, \\ &u(x,0) = 1 - \sin x, \ \ 0 < x < \pi. \end{aligned}$$

Unit 18

THE WAVE EQUATION

Mathematical Formulation and Uniqueness Result

We begin by studying the one-dimensional wave equation, which describe the transverse vibrations of a string. Consider the small vibrations of a string that is fastened at each end (see, Fig. 6.1). We now make the following assumptions:

- The string is made of a homogeneous material (i.e., the mass/unit length of the string is constant).
- There is no effect of gravity and external forces.
- The vibration takes place in a plane.

The mathematical model equation under these assumptions describe small vibrations of the string. Let the forces acting on a small portion PQ of the string. Since the string

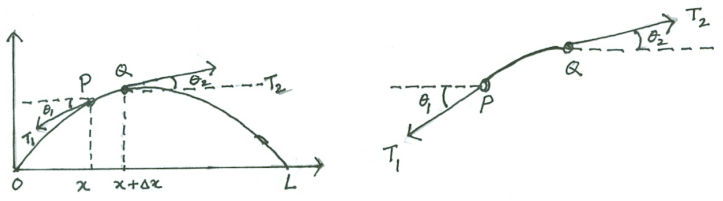


Figure 6.1: Vibrations of a string problem

does not offer resistance to bending, the tension is tangential to the curve of the string at each point. Let T_1 and T_2, respectively, be the tensions at the endpoints P and Q. Since there is no motion in horizontal direction, the horizontal components of the tension must be constant. From the Fig. 6.1, we obtain

$$T_1 \cos\theta_1 = T_2 \cos\theta_2 = T = \text{ constant.} \tag{1}$$

Let $-T_1 \sin\theta_1$ and $T_2 \sin\theta_2$ be two components of T_1 and T_2, respectively in the vertical direction. The minus sign indicates that component at P is directed downward. By Newton's second law, the resultant of these two forces is equal to the mass $\rho\Delta x$ of the portion times the acceleration u_{tt}, evaluated at some point between x and $x + \Delta x$. If ρ is the mass of the undeflected string per unit length and Δx is length of the portion of the undeflected string then we have

$$T_2 \sin\theta_2 - T_1 \sin\theta_1 = \rho\Delta x u_{tt}.$$

In view of (1), we obtain

$$\frac{T_2 \sin\theta_2}{T_2 \cos\theta_2} - \frac{T_1 \sin\theta_1}{T_1 \cos\theta_1} = \tan\theta_2 - \tan\theta_1 = \frac{\rho \Delta x}{T} u_{tt}. \tag{2}$$

Note that $\tan\theta_1$ and $\tan\theta_2$ are the slopes of the curve of the string at x and $x + \Delta x$, i,e.,

$$\tan\theta_1 = (u_x)_P, \quad \tan\theta_2 = (u_x)_Q.$$

Here, partial derivatives are used because u also depends on t. Dividing (2) by Δx, we have

$$\frac{1}{\Delta x}\left[u_x(x + \Delta x, t) - u_x(x, t)\right] = \frac{\rho}{T} u_{tt}.$$

Letting $\Delta x \to 0$, we obtain

$$\boxed{u_{tt} = c^2 u_{xx},} \tag{3}$$

where $c^2 = \frac{T}{\rho}$.

NOTE: The notation c^2 (instead of c) for the physical constant T/ρ has been chosen to indicate that this constant is positive. The constant c^2 depends on the density and tension of the string.

As the problem is linear, it is enough to prove the uniqueness of solution. The uniqueness result is proved in the following theorem.

Theorem 1. *Let $u_1(x,t)$ and $u_2(x,t)$ be two solutions of*

$$\begin{aligned} &\textit{PDE:} && u_{tt} = c^2 u_{xx}, \quad 0 \le x \le L, \ -\infty < t < \infty, \\ &\textit{BC:} && u(0,t) = a(t), \ \ u(L,t) = b(t), \\ &\textit{IC:} && u(x,0) = f(x), \ \ u_t(x,0) = g(x). \end{aligned}$$

Then $u_1(x,t) = u_2(x,t)$ for all $0 \le x \le L$, $-\infty < t < \infty$.

Proof. Let $v(x,t) = u_1(x,t) - u_2(x,t)$. Note that v satisfies

$$\begin{aligned} &v_{tt} = c^2 v_{xx}, \quad 0 \le x \le L, \ -\infty < t < \infty, \\ &v(0,t) = 0, \ \ v(L,t) = 0, \\ &v(x,0) = 0, \ \ v_t(x,0) = 0. \end{aligned}$$

with homogeneous BC and IC. Observe that $v(x,0) = 0$ and $v_t(x,0) = 0$. We need to show that $v(x,t) = 0$ for all t. We write

$$v(x,t) = v(x,t) - v(x,0) = \int_0^t v_t(x,t)dt. \tag{4}$$

We now claim that $v_t(x,t) = 0$ for all x in $[0, L]$ and for all t. Construct the function

$$H(t) = \int_0^L \{c^2 v_x^2(x,t) + v_t^2(x,t)\} dx. \tag{5}$$

Differentiating with respect to t and using $v_{tt} = c^2 v_{xx}$, we obtain

$$\begin{aligned} H'(t) &= \int_0^L \{c^2 2 v_x v_{xt} + 2 v_t v_{tt}\} dx \\ &= 2c^2 \int_0^L \{v_x v_{xt} + v_t v_{xx}\} dx \\ &= 2c^2 \int_0^L \frac{\partial}{\partial x}(v_x v_t) dx \\ &= 2c^2 \{v_x(x,t) v_t(x,t)\}\Big|_0^L \\ &= 0, \end{aligned}$$

where in the last step we have used $v_t(0,t) = \frac{d}{dt} v(0,t) = 0$ and, similarly $v_t(L,t) = 0$. Thus,

$$H'(t) = 0 \Longrightarrow H(t) = C,$$

where C is an arbitrary constant. Since $H(0) = 0$, we have $C = 0$ and, hence $H(t) = 0$. Thus, (5) becomes

$$\begin{aligned} & \int_0^L \{c^2 v_x^2(x,t) + v_t^2(x,t)\} dx = 0 \\ \Longrightarrow \quad & v_t(x,t) = 0 \quad \forall x \in [0, L], \quad \forall t \in \mathbb{R}. \end{aligned}$$

In view of (4), we obtain

$$v(x,t) = \int_0^t v_t(x,t) dt = 0 \implies u_1(x,t) = u_2(x,t).$$

This completes the proof.

The Infinite String Problem

In this lecture, we shall show that the solution of the wave equation

$$u_{tt} = c^2 u_{xx}$$

can be immediately obtained with suitable transformation of the independent variables. We shall derive D'Alembert formula for the solution of the wave equation for an infinite string ($-\infty < x < \infty$) with IC $u(x,0) = f(x)$ and $u_t(x,0) = g(x)$.

Consider the following IVP:

$$\text{PDE:} \qquad u_{tt} = c^2 u_{xx}, \quad -\infty < x < \infty, \ \ t \geq 0, \tag{1}$$

$$\text{IC:} \qquad u(x,0) = f(x) \text{ (initial displacement)}, \tag{2}$$

$$u_t(x,0) = g(x) \text{ (initial velocity)}.$$

Step 1.(Transforming to its canonical form): Introducing the transformation

$$\xi = x + ct \quad \eta = x - ct,$$

we note that

$$u_x = u_\xi \xi_x + u_\eta \eta_x = u_\xi + u_\eta.$$

$$\begin{aligned} u_{xx} &= (u_\xi + u_\eta)_x \\ &= (u_\xi + u_\eta)_\xi \xi_x + (u_\xi + u_\eta)_\eta \eta_x \\ &= u_{\xi\xi} + 2u_{\xi\eta} + u_{\eta\eta}. \end{aligned}$$

Similarly,

$$u_{tt} = c^2(u_{\xi\xi} - 2u_{\xi\eta} + u_{\eta\eta}).$$

Substituting the expression for u_{xx} and u_{tt} in $u_{tt} = c^2 u_{xx}$ yields

$$u_{\xi\eta} = 0, \tag{3}$$

which is known as canonical form of (1).

Step 2. (Solving the transformed equation (3)): Integrate (3) with respect to ξ to have

$$u_\eta(\xi, \eta) = \Phi(\eta) + \psi(\xi),$$

where $\Phi(\eta)$ is the antiderivative of $\phi(\eta)$, and $\psi(\xi)$ is any function of ξ. Thus, the general solution of $u_{\xi\eta} = 0$ is

$$u(\xi, \eta) = \phi(\eta) + \psi(\xi), \tag{4}$$

where $\phi(\eta)$, $\psi(\xi)$ are arbitrary functions of η and ξ, respectively.

Step 3. (Transforming back to the original variables x and t): Substituting $\xi = x + ct$ and $\eta = x - ct$ in (4) we get

$$u(x, t) = \phi(x - ct) + \psi(x + ct). \tag{5}$$

This is the general solution of the wave equation. We may interpret (5) as the sum of any two moving waves, each moving in opposite directions with velocity c.

Step 4. (Applying IC to the general solution): In order to solve IVP (1)-(2), the general solution $u(x, t)$ is required to satisfy the two initial conditions

$$u(x, 0) = f(x), \quad u_t(x, 0) = g(x).$$

These conditions lead to the following equations:

$$\phi(x) + \psi(x) = f(x) \tag{6}$$

$$-c\phi'(x) + c\psi'(x) = g(x). \tag{7}$$

Integrating (7) from x_0 to x, we obtain

$$-c\phi(x) + c\psi(x) = \int_{x_0}^{x} g(\tau)\, d\tau + K. \tag{8}$$

Solving for $\phi(x)$ and $\psi(x)$ from (6) and (8), we obtain

$$\phi(x) = \frac{1}{2} f(x) - \frac{1}{2c} \int_{x_0}^{x} g(\tau)\, d\tau \tag{9}$$

$$\psi(x) = \frac{1}{2} f(x) + \frac{1}{2c} \int_{x_0}^{x} g(\tau)\, d\tau \tag{10}$$

Thus, the solution to IVP (1)-(2) is given by

$$\boxed{u(x, t) = \frac{1}{2}[f(x - ct) + f(x + ct)] + \frac{1}{2c} \int_{x-ct}^{x+ct} g(\tau)\, d\tau.} \tag{11}$$

The equation (11) is known as D'Alembert solution to the IVP (1)-(2). This formula is of great interest in itself, and it avoids the problem of convergence of infinite series in the Fourier series approach.

Remark 1. *D'Alembert's formula yields a number of properties of solutions of the wave problem for the infinite string.*

- *Disturbances propagate with speed c.*

  The value $u(x_0, t_0)$ depends only on the values of g in the interval $[x_0 - ct_0, x_0 + ct_0]$ and on the values of f at the endpoints of this interval. Geometrically, this is the interval cut out by the characteristic lines that pass through the point (x_0, t_0). The interval $[x_0 - ct_0, x_0 + ct_0]$ is called the interval of dependence for the point (x_0, t_0) (since $u(x_0, t_0)$ depends only on the values $u(x, 0)$ and $u_t(x, 0)$ for x in this interval).

- *Odd initial data yields odd solution and even initial data yields even solution.*

  If $f(x)$ and $g(x)$ are odd, then $u(x, t)$ is odd in the x-variable, since

$$\begin{aligned} u(-x,t) &= \frac{1}{2}[f(-x+ct)+f(-x-ct)] + \frac{1}{2c}\int_{-x-ct}^{-x+ct} g(r)\,dr \\ &= \frac{1}{2}[-f(x-ct)-f(x+ct)] - \frac{1}{2c}\int_{x+ct}^{x-ct} g(-s)ds \\ &= -\frac{1}{2}[f(x-ct)+f(x+ct)] + \frac{1}{2c}\int_{x+ct}^{x-ct} g(s)ds \\ &= -\frac{1}{2}[f(x+ct)+f(x-ct)] - \frac{1}{2a}\int_{x-ct}^{x+ct} g(s)ds \\ &= -u(x,t). \end{aligned}$$

  Similarly, we can show that if $f(x)$ and $g(x)$ are even then $u(x, t)$ is even i.e., $u(-x, t) = u(x, t)$.

- *Periodic initial data yield periodic solutions.*

  If $f(x + 2L) = f(x)$ and $g(x + 2L) = g(x)$, then $u(x + 2L, t) = u(x, t)$. That is, if f and g are periodic of period $2L$ then $u(x, t)$ is also periodic of period $2L$ in x. This follows easily from D'Alembert's formula. This fact is useful in dealing with finite strings.

  It can be shown that if $f(x)$ and $g(x)$ are periodic of period $2L$ and

$$\int_{-L}^{L} g(x)dx = 0,$$

  then $u(x, t)$ is not only periodic in x of period $2L$, but also periodic in t of period $2L/c$.

Special cases of D'Alembert's formula:

CASE I. (*Initial velocity zero*). Suppose the string has IC

$$u(x,0) = f(x)$$
$$u_t(x,0) = 0.$$

The D'Alembert solution is

$$\boxed{u(x,t) = \frac{1}{2}[f(x-ct) + f(x+ct)].}$$

Thus, the solution u at a point (x_0, t_0) can be interpreted as the average of the initial displacement $f(x)$ at a point $(x_0 - ct_0, 0)$ and $(x_0 + ct_0, 0)$ found by backtracking the characteristic curves $x - ct = x_0 - ct_0$ and $x + ct = x_0 + ct_0$.

CASE 2. (*Initial displacement zero*) Suppose the string has the following IC:

$$u(x,0) = 0$$
$$u_t(x,0) = g(x).$$

In this case, the solution is

$$\boxed{u(x,t) = \frac{1}{2c}\int_{x-ct}^{x+ct} g(\tau)\, d\tau.}$$

The solution u at (x,t) may be interpreted as integrating the initial velocity between $x-ct$ and $x+ct$ on the initial line $t = 0$.

Let us consider the following examples.

Example 2. ***(Zero initial velocity)*** *Solve the IVP:*

$$\begin{aligned} \textit{PDE:} \quad & u_{tt} = c^2 u_{xx}, \quad -\infty < x, t < \infty, \\ \textit{IC:} \quad & u(x,0) = \sin(x), \\ & u_t(x,0) = 0. \end{aligned}$$

Solution: Applying D'Alembert's formula (11) with $f(x) = \sin(x)$ and $g(x) = 0$, we obtain

$$u(x,t) = \frac{1}{2}\left[\sin(x-ct) + \sin(x+ct)\right].$$

Example 3. ***(Zero initial displacement)*** *Consider the IVP:*

$$\begin{aligned} &\textit{PDE:} \quad u_{tt} = c^2 u_{xx}, \quad -\infty < x,\ t < \infty \\ &\textit{I.C.} \quad u(x,0) = 0, \\ &\qquad u_t(x,0) = \sin(x). \end{aligned}$$

Solution: Here the string is initially straight ($u(x,0) = 0$), but has a variable velocity at $t = 0$ ($u_t(x,0) = \sin(x)$). Thus, applying D'Alembert's formula (11) with $f(x) = 0$ and $g(x) = \sin(x)$, we obtain

$$u(x,t) = \frac{1}{2c}\int_{x-ct}^{x+ct} \sin(\tau)d\tau = -\frac{1}{2c}\left[\cos(x+ct) - \cos(x-ct)\right].$$

Practice Problems

1. Solve the following IVP:

$$\begin{aligned} &u_{tt} = 9u_{xx}, \quad -\infty < x < \infty, \quad t > 0, \\ &u(x,0) = \sin x, \quad u_t(x,0) = \cos x, \quad -\infty < x < \infty. \end{aligned}$$

2. Solve the following IVP:

$$\begin{aligned} &u_{tt} = c^2 u_{xx}, \quad -\infty < x < \infty, \quad t > 0, \\ &u(x,0) = 0, \quad u_t(x,0) = \sin^2(x), \quad -\infty < x < \infty. \end{aligned}$$

3. Let $u(x,t)$ be the solution of

$$\begin{aligned} &u_{tt} = c^2 u_{xx}, \quad 0 < x < \infty, \quad t > 0, \\ &u(x,0) = f(x), \quad u_t(x,0) = g(x), \quad -\infty < x < \infty. \end{aligned}$$

   Use D'Alembert's formula to show that u is even in x.

The Semi-Infinite String Problem

Before we introduce the semi-infinite string problem, let us look at some special cases of D'Alembert's formula derived in the previous lecture.

Example 1. *Consider the problem for the semi-infinite string ($0 \le x < \infty$) with fixed end at $x = 0$:*

$$\begin{aligned} \textit{PDE:} \quad & u_{tt} = c^2 u_{xx}, \quad 0 \le x < \infty, \ -\infty < t < \infty \\ \textit{BC:} \quad & u(0,t) = 0 \\ \textit{IC:} \quad & u(x,0) = f(x), \quad u_t(x,0) = 0. \end{aligned}$$

Solution. Note that $f(x)$ is defined for $x \ge 0$. Consider the odd extension $f_0(x)$, $-\infty < x < \infty$ as follows:

$$f_0(x) = \begin{cases} f(x) & \text{for } x \ge 0, \\ -f(-x) & \text{for } x \le 0. \end{cases}$$

The related extended problem is

$$\begin{aligned} &\text{PDE:} \quad u_{tt} = c^2 u_{xx}, \quad -\infty \le x, t < \infty \\ &\text{I.C.} \quad u(x,0) = f_0(x), \quad u_t(x,0) = 0. \end{aligned}$$

By D'Alembert's formula, the solution of this problem is

$$u(x,t) = \frac{1}{2}[f_0(x+ct) + f_0(x_0 - ct)].$$

Note that $u(x,t)$ is odd in x, since $f_0(x)$ is odd. Thus, $u(0,t) = 0$ and so $u(x,t)$ satisfies the BC.

Moreover,

$$u(x,0) = \frac{1}{2}[f_0(x + c\cdot 0) + f_0(x - c\cdot 0)] = f_0(x),$$

which is the same as $f(x)$ when $x \ge 0$.

Semi-infinite string problem: We shall find the solution of the following wave equation whose left end fixed at *zero* and has given initial conditions:

$$\begin{aligned} \text{PDE:} \quad & u_{tt} = c^2 u_{xx}, \quad 0 < x < \infty, \ 0 < t < \infty \\ \text{BC:} \quad & u(0,t) = 0, \ 0 < t < \infty, \\ \text{IC:} \quad & u(x,0) = f(x), \quad u_t(x,0) = 0, \quad 0 < x < \infty. \end{aligned}$$

Recall that the solution of the PDE (1) is given by (see (5), Lecture 2 of this module)

$$u(x,t) = \phi(x-ct) + \phi(x+ct). \tag{1}$$

Substitute the general solution into the initial conditions, we arrive at (cf. (9)-(10), Lecture 2 of this module)

$$\phi(x-ct) = \frac{1}{2}f(x-ct) - \frac{1}{2c}\int_{x_0}^{x-ct} g(\xi)\,d\xi. \tag{2}$$

$$\psi(x+ct) = \frac{1}{2}f(x+ct) + \frac{1}{2c}\int_{x_0}^{x+ct} g(\xi)\,d\xi. \tag{3}$$

Since we are looking for the solution $u(x,t)$ everywhere in the first quadrant ($x>0$, $t>0$) of the xt-plane, we must find $\phi(x-ct)\ \forall\ -\infty < x-ct < \infty$ and $\psi(x+ct)\ \forall\ 0 < x+ct < \infty$.

Using (1), (2) and (3), for $x-ct \geq 0$, it follows that

$$\begin{aligned} u(x,t) &= \phi(x-ct) + \psi(x+ct) \\ &= \frac{1}{2}[f(x-ct)+f(x+ct)] + \frac{1}{2c}\int_{x-ct}^{x+ct} g(\xi)d\xi. \end{aligned}$$

When $x < ct$, use of BC $u(0,t)=0$ leads to

$$\phi(-ct) = -\psi(ct)$$

and hence,

$$\phi(x-ct) = -\frac{1}{2}f(ct-x) - \frac{1}{2c}\int_{x_0}^{ct-x} g(\xi)\,d\xi + K.$$

Substituting this value of ϕ into the general solution

$$u(x,t) = \phi(x-ct) + \psi(x+ct).$$

yields

$$u(x,t) = \frac{1}{2}[f(x+ct)-f(ct-x)] + \frac{1}{2c}\int_{ct-x}^{x+ct} g(\xi)\,d\xi, \quad 0<x<ct.$$

Thus, for $x<ct$ and $x>ct$, we have

$$u(x,t) = \begin{cases} \frac{1}{2}[f(x-ct)+f(x+ct)] + \frac{1}{2c}\int_{x-ct}^{x+ct} g(\xi)\,d\xi & x \geq ct \\ \frac{1}{2}[f(x+ct)-f(ct-x)] + \frac{1}{2c}\int_{ct-x}^{x+ct} g(\xi)\,d\xi & x < ct. \end{cases}$$

Example 2. *Find the solution of the following IBVP:*

$$\begin{aligned} &u_{tt} = u_{xx},\ \ 0<x<\infty,\quad t>0, \\ &u(x,t) = 0,\ \ t>0, \\ &u(x,0) = |\sin x|,\ \ u_t(x,0) = 0,\ \ 0<x<\infty. \end{aligned}$$

Solution. For $x > t$,

$$\begin{aligned} u(x,t) &= \frac{1}{2}(f(x+t) + f(x-t)) \\ &= \frac{1}{2}(|\sin(x+t)| + |\sin(x-t)|). \end{aligned}$$

For $x < t$,

$$\begin{aligned} u(x,t) &= \frac{1}{2}(f(x+t) - f(t-x)) \\ &= \frac{1}{2}(|\sin(x+t)| - |\sin(t-x)|). \end{aligned}$$

Observe that $u(0,t) = 0$ is satisfied by $u(x,t)$ for $x < t$. Thus,

$$u(x,t) = \begin{cases} \frac{1}{2}(|\sin(x+t)| + |\sin(x-t)|) & x > t \\ \frac{1}{2}(|\sin(x+t)| - |\sin(t-x)|) & x < t. \end{cases}$$

Practice Problems

1. Solve the following IBVP:

$$\begin{aligned} &u_{tt} = u_{xx}, \ \ 0 < x < \infty, \quad t > 0, \\ &u_x(0,t) = 0, \ \ t \geq 0, \\ &u(x,0) = \cos x, \ \ u_t(x,0) = 0, \ \ 0 \leq x < \infty. \end{aligned}$$

2. Solve the following IBVP:

$$\begin{aligned} &u_{tt} = c^2 u_{xx}, \ \ 0 < x < \infty, \quad t > 0, \\ &u(0,t) = 0, \ \ t \geq 0, \\ &u(x,0) = x^2, \ \ u_t(x,0) = 0, \ \ 0 \leq x < \infty. \end{aligned}$$

Unit 19

THE WAVE EQUATION

The Finite Vibrating String Problem

In this lecture, we shall study the transverse vibrations of a finite string. If $u(x,t)$ represents the displacement (deflection) of the string and the ends of the string are held fixed, then the motion of the string is described by the following initial-boundary value problem (IBVP):

$$\text{PDE:} \qquad u_{tt} = c^2 u_{xx}, \quad 0 < x < L, \;\; 0 < t < \infty, \tag{1}$$

$$\text{BC:} \qquad u(0,t) = 0; \quad u(L,t) = 0, \;\; 0 < t < \infty. \tag{2}$$

$$\text{IC:} \qquad u(x,0) = f(x); \quad u_t(x,0) = g(x), \;\; 0 \le x \le L. \tag{3}$$

While studying the wave equation in a bounded region of space $0 < x < L$, it is to be noted that the waves no longer appear to be moving due to their repeated interaction with boundaries. These waves are known as standing waves (e.g., a guitar string fixed at both ends). The boundary condition in (2) reflect the fact the string is held fixed at the two end points $x = 0$ and $x = L$.

We shall apply the **method of separation of variables** to solve this problem.

Step 1. (Reducing to a system of ODEs): We seek solutions of the form

$$u(x,t) = X(x)T(t). \tag{4}$$

Substituting (4) into $u_{tt} = c^2 u_{xx}$ and separating variables, we get

$$X(x)T''(t) = c^2 X''(x)T(t).$$

or

$$\frac{T''(t)}{c^2 T(t)} = \frac{X''(x)}{X(x)} = k,$$

where the constant k can now be any number $-\infty < k < \infty$. This leads to two ODEs:

$$\boxed{\begin{array}{l} T''(t) - c^2 k T(t) = 0, \\ X''(x) - kX(x) = 0. \end{array}} \tag{5, 6}$$

The ODE $X'' - kX = 0$ is solved for $X(x)$ in a manner similar to that of heat equation (see, Lecture 3 of Module 5), but the solution of the ODE $T'' - c^2 kT = 0$ for $T(t)$ are different, because of the second-order time derivative.

Step 2. (Solving the ODEs): Investigating the solutions of these two ODEs for all different values of k lead into the following cases.

Case I: Let $k > 0$. Set $k = \lambda^2$. The soultions are given by

$$T(t) = Ae^{(c\lambda)t} + Be^{-(c\lambda)t},$$
$$X(x) = Ce^{(\lambda)x} + De^{-(\lambda)x}.$$

Application of BC yields $u \equiv 0$.

Case II: Let $k = 0$. In this case, the solutions are linear and given by

$$T(t) = At + B, \quad X(x) = Cx + D.$$

This case is of no interest because use of BC yields trivial solution $u \equiv 0$. Hence, for nontrival solution, we are left with the possibility of choosing $k < 0$.

Case III: Let $k < 0$. Set $k = -\lambda^2$ for some $\lambda \in \mathbb{R}$ and $\lambda \neq 0$.

The solutions of $T''(t) + c^2\lambda T(t) = 0$ is given by

$$T(t) = A\sin(c\lambda t) + B\cos(c\lambda t).$$

The solutions of $X''(x) + \lambda^2 X(x) = 0$ is

$$X(x) = C\sin(\lambda x) + D\cos(\lambda x),$$

where A, B, C and D are constants. Then

$$u(x,t) = [A\sin(c\lambda t) + B\cos(c\lambda t)][C\sin(\lambda x) + D\cos(\lambda x)].$$

Our goal is to find the constants A, B, C and D and the negative separation constant λ so that the expression

$$u(x,t) = [C\sin(\lambda x) + D\cos(\lambda x)][A\sin(c\lambda t) + B\cos(c\lambda t)] \tag{7}$$

satisfies the BC. As $u(x,t)$ has to satisfy the BC (2), substituting (7) into $u(0,t) = u(L,t) = 0$ gives

$$\begin{aligned} u(0,t) &= X(0)T(t) = D[A\sin(c\lambda t) + B\cos(c\lambda t)] = 0 \\ &\implies D = 0. \end{aligned}$$

$$\begin{aligned} u(L,t) &= 0 \implies X(L)T(t) = 0 \\ &= C\sin(\lambda L)[A\sin(c\lambda t) + B\cos(c\lambda t)] = 0 \\ &\implies \sin(\lambda L) = 0 \\ &\implies \lambda L = n\pi, \quad n = 0, 1, 2, \ldots \\ &\text{or } \lambda_n = \frac{n\pi}{L}, \quad n = 0, 1, 2, \ldots. \end{aligned}$$

Note that the choice of $C = 0$ in (7) would lead to $X(x)T(t) = 0$. Thus, the sequence of solutions given by

$$\begin{aligned} u_n(x,t) &= X_n(x)T_n(t) \\ &= \sin(\frac{n\pi x}{L})\left[a_n \sin(\frac{n\pi ct}{L}) + b_n \cos(\frac{n\pi ct}{L})\right], \quad n = 1, 2, 3, \cdots \end{aligned}$$

As the PDE is linear, by superposition principle we write

$$u(x,t) = \sum_{n=1}^{\infty} \sin(\frac{n\pi x}{L})\left[a_n \sin(\frac{n\pi ct}{L}) + b_n \cos(\frac{n\pi ct}{L})\right]. \tag{8}$$

These solutions are called eigenfunctions and the values $\lambda_n = \frac{n\pi}{L}$ are called the eigenvalues of the vibrating string.

Step 3. (Applying IC): Substituting (8) into IC $u(x,0) = f(x)$, $u_t(x,0) = g(x)$ yields the two equations:

$$\sum_{n=1}^{\infty} b_n \sin(\frac{n\pi x}{L}) = f(x),$$
$$\sum_{n=1}^{\infty} a_n (\frac{n\pi c}{L}) \sin(\frac{n\pi x}{L}) = g(x),$$

which represent the Fourier sine expansion of $f(x)$ and $g(x)$, respectively. The coefficients a_n and b_n are given by

$$a_n = \frac{2}{n\pi c}\int_0^L g(x)\sin(\frac{n\pi x}{L})dx, \tag{9}$$
$$b_n = \frac{2}{L}\int_0^L f(x)\sin(\frac{n\pi x}{L})dx. \tag{10}$$

Thus, the solution is

$$u(x,t) = \sum_{n=1}^{\infty} \sin(\frac{n\pi x}{L})\left[a_n \sin(\frac{n\pi ct}{L}) + b_n \cos(\frac{n\pi ct}{L})\right], \tag{11}$$

where a_n and b_n are given by (9) and (10), respectively.

Remark 1. • *The function $u(x,t)$ given by (11) with coefficients (9) and (10), is a solution of (1) that satisfies the conditions (2) and (3), provided that the series (11) converges and also that the series obtained by differentiating (11) twice (term-wise) with respect to x and t, converge and have the sums u_{xx} and u_{tt}, respectively, which are continuous.*

- *Note that each u_n in (8) represents a harmonic motion having the frequency $\lambda_n/2\pi = cn/2L$ cycles per unit time. This motion is called the nth normal mode of the string. The first normal mode is known as the fundamental mode ($n = 1$), and the others are known as overtones.*

Practice Problems

1. Solve the following IBVP:

$$\begin{aligned} &u_{tt} = u_{xx}, \;\; 0 < x < 1, \quad t > 0, \\ &u(0,t) = u(1,t) = 0, \;\; t > 0, \\ &u(x,0) = x(1-x), \;\; u_t(x,0) = 0, \;\; 0 \le x \le 1. \end{aligned}$$

2. Solve the following IBVP:

$$\begin{aligned} &u_{tt} = 4u_{xx}, \;\; 0 < x < \pi, \quad t > 0, \\ &u(0,t) = u(\pi,t) = 0, \;\; t > 0, \\ &u(x,0) = 0, \;\; u_t(x,0) = \sin x, \;\; 0 \le x \le \pi. \end{aligned}$$

The Inhomogeneous Wave Equation

Recall the Duhamel's principle for inhomogeneous heat equations that arises due to internal heat sources. We solve the inhomogeneous heat equation by solving a family of related problems in which the sources appears in the initial conditions instead of the differential equation. The same idea works for inhomogeneous wave equations. To illustrate the procedure, let us consider the following infinite string problem:

$$\text{PDE:} \qquad u_{tt} = c^2 u_{xx} + h(x,t), \quad -\infty < x,\ t < \infty, \tag{1}$$

$$\text{IC:} \qquad u(x,0) = 0, \ \ u_t(x,0) = 0. \tag{2}$$

To motivate the method of Duhamel for the string problem, let the acceleration $h(x,s)$ be applied to the string at $t = s - \Delta s$ and let the acceleration be turned off at $t = s$. The string will then acquire a velocity of $h(x,s)\Delta s$, and its position change is $h(x,s)(\Delta s)^2/2$. Assuming Δs to be small enough, the change in position can be neglected. The effect of the imposed acceleration is $v(x,t;s)\Delta s$, where $v(x,t;s)$ is the solution of

$$\text{PDE:} \qquad v_{tt} = c^2 v_{xx}, \quad -\infty < x < \infty, \ \ t \geq s, \tag{3}$$

$$\text{IC:} \qquad v(x,s;s) = 0, \ \ v_t(x,s;s) = h(x,s). \tag{4}$$

This problem has initial conditions given at the arbitrary time $t = s$, instead of $t = 0$. We can write $v(x,t;s) = \tilde{v}(x,t-s;s)$, where $\tilde{v}(x,t;s)$ solves

$$\text{PDE:} \qquad \tilde{v}_{tt} = c^2 \tilde{v}_{xx}, \quad -\infty < x < \infty, \ \ t \geq 0 \tag{5}$$

$$\text{IC:} \qquad \tilde{v}(x,s;s) = 0, \ \ \tilde{v}_t(x,s;s) = h(x,s). \tag{6}$$

By D'Alembert's formula, the solution of (5) is given by

$$\tilde{v}(x,t;s) = \frac{1}{2c}\int_{x-ct}^{x+ct} h(r,s)dr, \tag{7}$$

and hence, the solution of (3) is

$$v(x,t;s) = \tilde{v}(x,t-s;s) = \frac{1}{2c}\int_{x-c(t-s)}^{x+c(t-s)} h(r,s)dr.$$

Theorem 1. (Duhamel's principle for the wave equation[1]) *Let $h(x,t)$ be a C^1 function, $-\infty < x,\ t < \infty$. Then the unique solution of the problem (1) satisfying the conditions (2) is given by*

$$\boxed{u(x,t) = \int_0^t v(x,t;s)ds = \int_0^t \tilde{v}(x,t-s;s)ds = \frac{1}{2c}\int_0^t \int_{x-c(t-s)}^{x+c(t-s)} h(r,s)drds.} \tag{8}$$

Proof. By D'Alembert's formula, we know

$$\tilde{v}(s,t;s) = \frac{1}{2c}\int_{x-ct}^{x+ct} h(r,s)ds.$$

Note that $\tilde{v}(s,t;s)$ is in C^2 since $h(x,t)$ is assumed to be in C^1. Differentiate twice with respect to t to obtain

$$u_t(x,t) = \tilde{v}(x,0;s) + \int_0^t \tilde{v}_t(x,t-s;s)ds = \int_0^t \tilde{v}_t(x,t-s;s)ds, \tag{9}$$

and

$$\begin{aligned} u_{tt}(x,t) &= \tilde{v}_t(x,0;t) + \int_0^t \tilde{v}_{tt}(x,t-s;s)ds \\ &= h(x,t) + \int_0^t c^2\tilde{v}_{xx}(x,t-s;s)ds \\ &= h(x,t) + c^2 u_{xx}(x,t), \end{aligned}$$

where we have used (5). This shows that $u(x,t)$ is a C^2 solution of (1). By (8), we have $u(x,0) = 0$. The equation (9) yields $u_t(x,0) = 0$.

To prove the uniqueness, let u_1 and u_2 be two solutions of (1)-(2). Now, the function $v = u_1 - u_2$ satisfies $v_{tt} = c^2 v_{xx}$ with IC $v(x,0) = 0$ and $v_t(x,0) = 0$. Hence, $v \equiv 0 \implies u_1 = u_2$. This completes the proof.

Example 2. *Solve*

$$\begin{aligned} &\textit{PDE:} \quad u_{tt} - u_{xx} = x - t, \quad -\infty < x,\ t < \infty, \\ &\textit{IC:} \quad u(x,0) = x^4,\ \ u_t(x,0) = \sin(x). \end{aligned} \tag{10}$$

Solution. Splitting the problem (10) into two problems with $u_1(x,t)$ and $u_2(x,t)$ solve

$$\begin{aligned} &(u_1)_{tt} - (u_1)_{xx} = 0, \\ &u_1(x,0) = x^4, \\ &(u_1)_t(x,0) = \sin(x), \end{aligned}$$

and

$$\begin{aligned} &(u_2)_{tt} - (u_2)_{xx} = x - t, \\ &u_2(x,0) = 0, \\ &(u_2)_t(x,0) = 0. \end{aligned}$$

respectively. The solution of (8) is then $u(x,t) = u_1(x,t) + u_2(x,t)$. By D'Alembert's formula

$$u_1(x,t) = \frac{1}{2}[(x+t)^4 + (x-t)^4] - \frac{1}{2}[\cos(x+t) - \cos(x-t)].$$

Applying Theorem 1 we compute $u_2(x,t)$ as follows:

$$\begin{aligned} u_2(x,t) &= \frac{1}{2}\int_0^t \int_{x-(t-s)}^{x+(t-s)} (r-s)drds = \frac{1}{2}\int_0^t \left[\frac{r^2}{2} - sr\right]_{x-t+s}^{x+t-s} ds \\ &= \frac{1}{2}\int_0^t \left[\frac{(x+t-s)^2}{2} - \frac{(x+s-t)^2}{2} - s(x+t-s) + s(x+s-t)\right] ds \\ &= \frac{1}{2}\int_0^t \left[2s^2 - 2s(x+t) + \frac{(x+t)^2}{2} - \frac{(x-t)^2}{2}\right] ds \\ &= \frac{t^3}{3} - \frac{t^2(x+t)}{2} + t^2x = -\frac{t^3}{6} + \frac{t^2x}{2}. \end{aligned}$$

The solution $u(x,t) = u_1(x,t) + u_2(x,t)$ can easily be verified.

Remark 3. *Duhamel's principle also applies in the case of a finite string. As in Example 2, one can handle the case where both the differential equation and BC are inhomogeneous. This is done by splitting the problem into two parts and then adding the solutions of the two parts to obtain the desired solution.*

Practice Problems

1. Solve the following nonhomogeneous IBVP:

$$\begin{aligned} &u_{tt} = u_{xx} + x\sin t, \quad 0 < x < 1, \quad t > 0, \\ &u(x,0) = x(1-x), \quad u_t(x,0) = 0, \quad 0 \le x \le 1, \\ &u(0,t) = u(1,t) = 0, \quad t > 0. \end{aligned}$$

2. Solve the following nonhomogeneous IBVP:

$$\begin{aligned} &u_{tt} = u_{xx} + 2, \quad 0 < x < 1, \quad t > 0, \\ &u(x,0) = x, \quad u_t(x,0) = 0, \quad 0 \le x \le 1, \\ &u(0,t) = 0, \quad u_x(1,t) = t, \quad t \ge 0. \end{aligned}$$

Unit 20

THE LAPLACE EQUATION

Basic Concepts and The Maximum/Minimum Principle

Let Ω be an open region in $\mathbb{R}^2$. The Laplace equation in two dimension is of the form

$$\nabla^2 u(x,y) = 0, \quad (x,y) \in \Omega, \tag{1}$$

where $\nabla^2 := \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial x^2}$ is the Laplace operator or the Laplacian. The equation of the type (1) plays an important role in a variety of physical contexts such as in Gravitation theory, electrostatics, steady-state heat conduction problems and fluid flow problems.

Some examples of physical problems(cf. [10]):

Example 1. (*Gravitation theory) The force of attraction F, both inside and outside the attracting matter, can be expressed in terms of a gravitational potential u by the equation*

$$F = \nabla u.$$

In empty space u satisfies Laplace's equation

$$\nabla^2 u = 0.$$

Example 2. (*Steady-state heat flow problem) In the theory of heat conduction if the temperature u does not vary with the time, then u satisfies the equation*

$$\nabla \cdot (\kappa \nabla u) = 0,$$

where κ is the thermal conductivity. If κ is a constant throughout the medium then

$$\nabla^2 u = 0.$$

Example 3. (*Fluid flow problem) The velocity q of a perfect fluid in irrotational motion can be expressed in terms of a velocity potential u by the equation*

$$q = -\nabla u.$$

If there are no sources or sinks at all points of the fluid the function u satisfies Laplace's equation

$$\nabla^2 u = 0.$$

The inhomogeneous Laplace equation

$$\nabla^2 u(x,y) = f(x,y) \quad \text{in } \Omega,$$

where f is a given function is known as the Poisson equation.

1 Types of BVP

Because these solutions do not depend on time, initial conditions are irreverent and only boundary conditions are specified. There are three basic types of boundary conditions that are usually associated Laplace's equation. They are

- *Dirichlet BVP:* If the BC are of Dirichlet type i.e., if the solution $u(x, y)$ to Laplace equation in a domain Ω is specified on the boundary $\partial\Omega$ i.e.,

$$\boxed{u(x, y) = f(x, y) \quad \text{on } \partial\Omega,}$$

  where $f(x, y)$ is a given function. The Laplace equation together with Dirichlet BC are called the **Dirichlet problem / Dirichlet BVP**. The Dirichlet problem for Laplace equation is of the form

$$\nabla^2 u(x, y) = 0 \text{ in } \Omega; \quad u(x, y) = f(x, y) \text{ on } \partial\Omega.$$

- *Neumann BVP:* We know the BC are of Neumann type if the directional derivative $\frac{\partial u}{\partial n}$ along the outward normal to the boundary is specified on $\partial\Omega$ i.e.,

$$\boxed{\frac{\partial u}{\partial n}(x, y) = g(x, y) \quad \text{for } (x, y) \in \partial\Omega.}$$

  In physical terms, the normal component of the solution gradient is known on the boundary. In steady-state heat flow problem, Neumann BC means the rate of heat loss or gain through the boundary points is prescribed.

  The Laplace equation together with Neumann BC are called the **Neumann BVP/ Neumann problem** which is written as

$$\nabla^2 u = 0 \text{ in } \Omega; \quad \frac{\partial u}{\partial n}(x, y) = g(x, y) \quad \text{for } (x, y) \in \partial\Omega.$$

  The Neumann problem will have no solution unless we assume that the average value of the function g on $\partial\Omega$ is zero. This assumption is known as the compatibility condition

$$\int_{\partial\Omega} \frac{\partial u}{\partial n} = \int_{\partial\Omega} g = 0,$$

  which will be discussed in the next lecture.

- *Robin's BVP.* The boundary conditions are called Robin's type or mixed type if Dirichlet BC are specified on part of the boundary $\partial\Omega$ and Neumann type BC are specified on the remaining part of the boundary $\partial\Omega$. For example,

$$\boxed{\frac{\partial u}{\partial n} + c(u - g) = 0,}$$

where c is a constant and g is a given function that can vary over the boundary. The Laplace equation together with the Rabin's/Mixed BC known as **Rabin's BVP / Mixed BVP**.

2 The maximum/minimum principle

The maximum/minimum principle for Laplace's equation is stated in the following theorem.

Theorem 4. (The maximum/minimum principle for Laplace's equation) *Let $u(x,y) \in C^2(\Omega) \cap C(\bar{\Omega})$ be a solution of Laplace's equation*

$$\nabla^2 u(x,y) := u_{xx} + u_{yy} = 0 \tag{2}$$

in a bounded region Ω with boundary $\partial\Omega$. Then the maximum and minimum values of u attain on $\partial\Omega$. That is,

$$\max_{\bar{\Omega}} u(x,y) = \max_{\partial\Omega} u(x,y); \quad \textit{and} \quad \min_{\bar{\Omega}} u(x,y) = \min_{\partial\Omega} u(x,y).$$

Proof. Since u is continuous in $\bar{\Omega}$ it attains its maximum either in Ω or on $\partial\Omega$. Suppose u achieves its maximum at some point $(x_0, y_0) \in \Omega$. Let

$$u(x_0,y_0) = \max_{\Omega} u(x,y) = M_0 > M_b,$$

where $M_b = \max_{\partial\Omega} u(x,y)$. Consider the function

$$v(x,y) = u(x,y) + \epsilon[(x-x_0)^2 + (y-y_0)^2], \tag{3}$$

for some $\epsilon > 0$. Note that $v(x_0,y_0) = u(x_0,y_0) = M_0$ and

$$\max_{\partial\Omega} v(x,y) \le M_b + \epsilon d^2,$$

where d is the diameter of Ω. For such ϵ $(0 < \epsilon < (M_0 - M_b)/d^2)$, the maximum of v can not occur on $\partial\Omega$ because

$$M_0 = v(x_0,y_0) > \max_{\partial\Omega} v(x,y).$$

This implies there may be points in Ω where $v > M_0$. Let

$$v(x_1,y_1) = \max_{\Omega} v(x,y).$$

At (x_1,y_1), we must have

$$v_{xx} \le 0 \quad \text{and} \quad v_{yy} \le 0 \Longrightarrow \; v_{xx} + v_{yy} \le 0. \tag{4}$$

From (3), we observe that

$$v_{xx} + v_{yy} = u_{xx} + u_{yy} + 2\epsilon + 2\epsilon = 4\epsilon > 0,$$

where we have used the fact that $u_{xx} + u_{yy} = 0$. This led to a contradiction to (4). Thus,

$$\max_{\bar{\Omega}} v(x, y) \neq \max_{\Omega} v(x, y).$$

So, the maximum of u attains on $\partial\Omega$.

To prove that the minimum of u is also achieved on the boundary $\partial\Omega$, replace u by $-u$ in the above argument to obtain

$$\min_{\bar{\Omega}} u = \max_{\bar{\Omega}}(-u) = \max_{\partial\Omega}(-u) = \min_{\partial\Omega}(u).$$

This completes the proof.

We now discuss the maximum and minimum principle for Poisson's equation

$$\nabla^2 u(x, y) = f(x, y) \quad \text{in } \Omega. \tag{5}$$

Theorem 5. (The maximum/minimum principle for Poisson's equation) *Let Ω be a bounded domain in $\mathbb{R}^2$ with boundary $\partial\Omega$. Then the maximum values of a solution u of (5) attain on $\partial\Omega$ if $f(x, y) > 0$ in Ω and the minimum values of u occur on $\partial\Omega$ if $f(x, y) < 0$ in Ω.*

Proof. Since u is continuous in a closed and bounded domain, it must assume its maximum in Ω or in $\partial\Omega$. Suppose that the maximum is assumed at a point (x_0, y_0) in Ω, i.e.,

$$u(x_0, y_0) = \max_{\bar{\Omega}} u(x, y).$$

Suppose that $f(x, y) > 0$ in Ω. Then at $(x_0, y_0) \in \Omega$, we must have

$$u_{xx}(x_0, y_0) \leq 0, \quad u_{yy}(x_0, y_0) \leq 0.$$

As $f > 0$, it follows from (5) that

$$u_{xx} + u_{yy} > 0,$$

which is a contradiction. Hence, the maximum of $u(x, y)$ must occur on $\partial\Omega$.

To show that the minimum of $u(x, y)$ attains on $\partial\Omega$ if $f(x, y) < 0$ in Ω, replace u by $-u$ in the preceding argument. This is equivalent to replacing f by $-f$ in (4). Since $f < 0$, we obtain $-f > 0$ and conclude that $-u$ assumes its maximum on $\partial\Omega$. Therefore, u assumes its minimum on $\partial\Omega$ and this completes the proof.

The maximum/minimum principle can be used to prove uniqueness and continuous dependence of the solution for the Dirichlet's problems.

Theorem 6. *Let Ω be a bounded domain in $\mathbb{R}^2$ with boundary $\partial\Omega$. The solution of the Dirichlet's problem*

$$\nabla^2 u(x,y) = -f(x,y) \quad in\, \Omega, \quad u(x,y) = g(x,y) \quad on\, \partial\Omega \tag{6}$$

if it exists, is unique.

Proof. Let $u_1(x,y)$ and $u_2(x,y)$ be two solutions of (6). Set $v(x,y) = u_1(x,y) - u_2(x,y)$. Then v satisfies

$$\nabla^2 v = 0 \ \text{ in } \ \Omega, \quad v = 0 \ \text{ on } \ \partial\Omega.$$

The maximum/minimum principle yields (cf. Theorem 4)

$$v = 0 \ \text{ in } \ \Omega \Longrightarrow \ u_1 - u_2 = 0 \ \text{ in } \ \Omega.$$

Thus, we have

$$u_1 = u_2,$$

which proves the uniqueness.

Next, we shall prove the continuous dependence of the solution on the boundary data.

Theorem 7. *The solution of the Dirichlet problem depends continuously on the boundary data.*

Proof. Let u_i, $i = 1, 2$ be the solutions of

$$\nabla^2 u_i = F \quad \text{in } \Omega \subset \mathbb{R}^2, \quad u_i = f_i \quad \text{on } \partial\Omega.$$

Then the function $v = u_1 - u_2$ solves

$$\nabla^2 v = 0 \ \text{ in } \ \Omega \ \text{ with } \ v = f_1 - f_2 \ \text{ on } \ \partial\Omega.$$

By the maximum/minimum principle v attains its maximum/minimum on $\partial\Omega$. Thus, for all $(x,y) \in \bar{\Omega}$, we have

$$-\max_{\partial\Omega}(|f_1 - f_2|) \le \min_{\partial\Omega}(f_1 - f_2) \le v(x,y) \le \max_{\partial\Omega}(f_1 - f_2) \le \max_{\partial\Omega}(|f_1 - f_2|).$$

If $|f_1 - f_2| < \epsilon$ then

$$-\epsilon < \min_{\bar{\Omega}} v(x,y) \le v(x,y) \le \max_{\bar{\Omega}} v(x,y) < \epsilon.$$

Therefore,

$$|f_1 - f_1| < \epsilon \Longrightarrow |v(x,y)| < \epsilon$$

for all $(x,y) \in \bar{\Omega}$. This completes the proof.

Practice Problems

1. Let u satisfy the Laplace equation in a disk $\Omega = \{(x, y) \mid x^2+y^2 < 1\}$ and continuous on $\bar{\Omega}$. If $u(\cos\theta, \sin\theta) \leq \sin\theta + \cos(2\theta)$, then show that

$$u(x, y) \leq y + x^2 - y^2, \quad \forall\, (x, y) \in \bar{\Omega}.$$

2. Consider the elliptic equation

$$\nabla \cdot (\alpha \nabla u) = -F, \;\; \alpha > 0,$$

in a bounded region $\Omega \subset \mathbb{R}^2$ with the boundary $\partial\Omega$. Show that if $F < 0$ in Ω, the solution u assumes its maximum on $\partial\Omega$ and if $F > 0$ in Ω, the solution u assumes its minimum on $\partial\Omega$.

3. Let Ω be a bounded region $\mathbb{R}^2$. Use the maximum principle to prove continuous dependence on the data for the Dirichlet problem for the elliptic equation

$$\nabla \cdot (\alpha \nabla u) = -F \;\text{ in } \Omega$$

with $\alpha > 0$.

Green's Identity and Fundamental Solutions

In this lecture, we shall learn about some important identities known as Green's identities and its special forms. As a consequence of these identities we can prove the uniqueness of the solution to the Dirichlet problem and the compatibility conditions for the Neumann problems. The fundamental solutions for the Laplace equation will be discussed.

Let Ω be bounded domain in $\mathbb{R}^2$ with smooth boundary $\partial\Omega$. Recall the following Gauss divergence theorem: For $u, v \in C^1(\Omega)$

$$\int_\Omega v\frac{\partial u}{\partial x_k}dx = \int_{\partial\Omega} vu\cdot n ds - \int_\Omega u\frac{\partial v}{\partial x_k}dx, \tag{1}$$

where n is the outward unit normal the boundary $\partial\Omega$ and ds is the element of arc length.

As a consequence of Gauss divergence theorem, the following identity known as Green's identity hold true:

$$\boxed{\int_\Omega v\nabla^2 u dx = \int_{\partial\Omega} v\frac{\partial u}{\partial n}ds - \int_\Omega \nabla u\cdot\nabla v dx.} \tag{2}$$

Integrating the second term of the right hand side once more by parts we obtain

$$\boxed{\int_\Omega v\nabla^2 u dx = \int_\Omega u\nabla^2 v dx + \int_{\partial\Omega}\left(v\frac{\partial u}{\partial n} - u\frac{\partial v}{\partial n}\right)ds.} \tag{3}$$

Here, $\frac{\partial}{\partial n}$ indicates differentiation in the direction of the exterior normal to $\partial\Omega$.

From the identity (2), the special case $v = 1$ yields

$$\int_\Omega \nabla^2 u dx = \int_{\partial\Omega}\frac{\partial u}{\partial n}ds. \tag{4}$$

Another special case of interest by choosing $v = u$. In this case, the equation (2) yields the energy identity

$$\int_\Omega |\nabla u|^2 dx + \int_\Omega u\nabla^2 u dx = \int_{\partial\Omega} u\frac{\partial u}{\partial n}ds. \tag{5}$$

If $\nabla^2 u = 0$ in Ω then for $u \in C^2(\bar{\Omega})$, it follows that

$$\begin{aligned} & \int_\Omega |\nabla u|^2 dx = 0 \\ \implies \quad & \nabla u = 0 \\ \implies \quad & u = \text{constant}. \end{aligned}$$

This observation leads to uniqueness theorems for the Dirichlet problem and the Neumann problem.

Remark 1. *Using Green's identity (2), one can easily prove that:*
(i) A solution $u \in C^2(\bar{\Omega})$ of the Dirichlet problem is determined uniquely.

(ii) A solution $u \in C^2(\bar{\Omega})$ of the Neumann problem is determined uniquely within an additive constant.

Observe that the solution of the Neumann problem can only exist if the data satisfy the condition known as compatibility condition. For example, the compatibility condition for the Neumann problem:

$$\nabla^2 u = 0 \ \ in\, \Omega, \quad \frac{\partial u}{\partial n} = g \ \ on\, \partial\Omega$$

is

$$\int_{\partial\Omega} g ds = 0,$$

which immediately follows from the identity (4).

Fundamental Solutions: One of the principal features of the Laplace equation

$$\nabla^2 u = 0 \tag{6}$$

is its spherical symmetry. The Laplace equation is preserved under rotations about a point ξ. Therefore, it is reasonable to assume that there exist special solutions $v(x)$ of (6) that are invariant under rotations about ξ. Such solutions would be of the form

$$v = \psi(r), \tag{7}$$

where

$$r = |x - \xi| = \sqrt{\sum_{i=1}^{n} (x_i - \xi_i)^2}$$

represents the Euclidean distance between x and ξ. By the chain rule of differentiation we find that

$$\frac{dr}{dx_i} = \frac{1}{2}\left(\sum_{i=1}^{n}(x_i - \xi_i)^2\right)^{-1/2} \times 2x_i = \frac{x_i}{r}.$$

Further, we note that

$$v_{x_i} = \psi'(r)\frac{dr}{dx_i} = \psi'(r)(\frac{x_i}{r}), \ \ v_{x_i x_i} = \psi''(r)\frac{x_i^2}{r^2} + \left(\frac{1}{r} - \frac{x_i^2}{r^3}\right).$$

Hence,

$$\nabla^2 v = \sum_{i=1}^{n} v_{x_i x_i} = \psi''(r) + \frac{n-1}{r}\psi'(r) = 0.$$

If $\psi'(r) \neq 0$, we have

$$\frac{\psi''(r)}{\psi'(r)} = \frac{1-n}{r}.$$

On solving we arrive at $\psi'(r) = Cr^{1-n}$ and hence,

$$\psi(r) = \begin{cases} C\log r + C_1 & n = 2, \\ \frac{Cr^{2-n}}{2-n} + C_1 & n > 2, \end{cases}$$

where C and C_1 are constants.

The function $v(x) = \psi(r)$ satisfies (6) for $r > 0$, that is for $x \neq \xi$, but becomes infinite for $x = \xi$. The function v for a suitable choice of the constant C, is a fundamental solution for the operator ∇^2, satisfying the equation,

$$\boxed{\nabla^2 v = \delta(x - \xi),}$$

where δ is the Dirac delta function. The function

$$\boxed{\psi(r) = \frac{1}{2\pi}\log r, \quad r > 0}$$

is a fundamental solution to two dimensional Laplace's equation (6). For a proof, see [5].

The Poisson Integral Formula. We know the function $u \in C^2(\Omega)$ satisfying the Laplace equation $\nabla^2 u = 0$ is harmonic. The following result express the solution of the Dirichlet problem in terms of an integral known as The Poisson integral formula.

Theorem 2. (The Poisson integral formula) *Let $f(\theta)$ be a continuous function and $f(\theta + 2\pi) = f(\theta)$. Define*

$$u(r,\theta) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{(r_0^2 - r^2)f(s)}{r_0^2 - 2rr_0\cos(\theta - s) + r^2} ds, \quad r < r_0,$$
$$u(r_0,\theta) = f(\theta), \quad r = r_0.$$

Then $u(r,\theta)$ solves the following Dirichlet problem:

$$\nabla^2 u(x,y) = 0, \quad (x^2+y^2)^{1/2} < r_0,$$
$$u(r_0,\theta) = f(\theta), \quad f(\theta + 2\pi) = f(\theta),$$

where $u(r,\theta) = u(x,y) = u(r\cos\theta, r\sin\theta)$. That is, $u(r,\theta)$ is harmonic on the open disk $D = \{(x,y) \,|\, (x^2+y^2)^{1/2} < r_0\}$.

Some consequences of the Poisson integral formula are given below.

Theorem 3. *Let u be a harmonic function on some region Ω. The value of u at the center of any disk D with $D \subset \Omega$ is the average (or mean) of the values of u on the circular boundary ∂D of D.*

Note: The mean value property can be used to prove the maximum and minimum principle for solutions for Laplace's equation. It can be used to show that whenever the maximum or minimum is attained in the interior of the region, the solution u must be identically constant. This is the *strong maximum and minimum principle* for Laplace's equation.

Theorem 4. (The strong maximum/minimum principle) *Let u be a harmonic function on an open connected set Ω. Suppose that the maximum or minimum of u is attained at some point in Ω. Then u must be constant throughout Ω.*

We know by definition a harmonic function u on an open region Ω is only required to be $C^2(\Omega)$. But, u actually $C^\infty(\Omega)$ (infinitely differentiable function). Thus, we have the following result.

Theorem 5. (Regularity result) *If u is harmonic on an open region Ω, then $u \in C^\infty(\Omega)$.*

Practice Problems

1. Prove that a solution of the Neumann problem

$$\nabla^2 u = f \quad \text{in } \Omega, \quad u = g \text{ on } \partial\Omega$$

   differs from another solution by a constant.

2. Prove that $u_1(x, y) = 1 + \log(x^2 + y^2)$ and $u_2(x, y) = 1 - \log(x^2 + y^2)$ are harmonic, where defined. Note that $u_1 = u_2$ on the circle $x^2 + y^2 = 1$, but unequal inside the circle. Why does this not contradict the uniqueness theorem for the Dirichlet problem.

3. Let u be harmonic in the disk $x^2 + y^2 < r_0^2$. If u achieves its maximum at the point $(0, 0)$, then show that u must be constant throughout this disk.

The Dirichlet BVP for a Rectangle

In this lecture we shall discuss the solution of the Laplace equation with Dirichlet type BC in cartesian coordinates.

Consider the following Dirichlet problem in a rectangle:

$$\text{PDE:} \quad u_{xx} + u_{yy} = 0, \quad 0 < x < a, \;\; 0 < y < b, \tag{1}$$

$$\text{BC:} \quad u(x,0) = f_1(x), \quad u(x,b) = f_2(x), \;\; 0 \le x \le a, \tag{2}$$

$$u(0,y) = g_1(y), \quad u(a,y) = g_2(y) \;\; 0 \le y \le b.$$

We shall study how the method of separation of variables is still applicable for the BVP. Since the BC are nonhomogeneous, we are required to do some preliminary work.

By the principle of superposition, we seek the solution of the above BVP (1)-(2) as

$$\boxed{u(x,y) = u_1(x,y) + u_2(x,y) + u_3(x,y) + u_4(x,y),}$$

where each of u_1, u_2, u_3 and u_4 satisfies the PDE with one of the original nonhomogeneous BC, and the homogeneous versions of the remaining three BC. These problems are then solved by the method of separation of variables.

Let us consider solving the following example problem:

Example 1. *Solve the Dirichlet BVP:*

$$\textit{PDE:} \quad u_{xx} + u_{yy} = 0, \quad 0 < x < a, \;\; 0 < y < b, \tag{3}$$

$$\textit{BC:} \quad u(x,0) = f(x), \quad u(x,b) = 0, \quad 0 \le x \le a, \tag{4}$$

$$u(0,y) = 0, \quad u(a,y) = 0, \quad 0 \le y \le b.$$

Apply the method of separation of variables to solve this problem. The step-wise solution procedure is given below.

Step 1: (Reducing to ODEs)
Separating variables, we seek for a solution of the form

$$u(x,y) = X(x)Y(x).$$

Substituting this into (3), we obtain

$$X''Y(y) + X(x)Y''(y) = 0$$

and hence,

$$\frac{X''(x)}{X(x)} = -\frac{Y''(y)}{Y(y)} = k,$$

for some constant k, which is called the separation constant. This leads two ODEs

$$X''(x) - kX(x) = 0, \tag{5}$$

$$Y''(y) + kY(y) = 0. \tag{6}$$

Step 2: (Solving the resulting ODEs)

Case 1: When $k > 0$, set $k = \lambda^2$, where $\lambda \neq 0$. In this case, the solutions of ODEs are

$$X(x) = [Ae^{\lambda x} + Be^{-\lambda x}],$$

$$Y(y) = [C\cos(\lambda y) + D\sin(\lambda y)].$$

Therefore, the solutions of PDE $u(x, y)$ are given by

$$u(x, y) = [Ae^{\lambda x} + Be^{-\lambda x}][C\cos(\lambda y) + D\sin(\lambda y)].$$

Case 2: When $k = 0$, the solutions of ODEs are linear are given by

$$X(x) = (A + Bx), \quad Y(y) = (C + Dy).$$

Therefore,

$$u(x, y) = (A + Bx)(C + Dy).$$

Case 3: Suppose $k < 0$, set $k = -\lambda^2$, where $\lambda > 0$.
The solutions of ODEs are given by

$$X(x) = [A\cos(\lambda x) + B\sin(\lambda x)]$$

$$Y(x) = [Ce^{\lambda y} + De^{-\lambda y}].$$

Thus , the solution of PDE is

$$u(x, t) = [A\cos(\lambda x) + B\sin(\lambda x)][Ce^{\lambda y} + De^{-\lambda y}].$$

Step 3: (Applying the BC)
Using the boundary conditions $u(0, y) = 0$ and $u(a, y) = 0$ for the product solution obtained for the case $k > 0$ leads to the equations

$$A + B = 0, \quad Ae^{\lambda a} + Be^{-\lambda a} = 0,$$

which has a trivial solution $A = 0$ and $B = 0$. Thus, only the trivial solution $u(x, y) = 0$ is possible. Similarly, use of boundary conditions $u(0, y) = 0$ and $u(a, y) = 0$ also leads to a trivial solution $u(x, y) = 0$ for the case $k = 0$. Let us examine the product solution obtained in Case 3 (for $k < 0$) i.e.,

$$u(x, y) = [A\cos(\lambda x) + B\sin(\lambda x)][Ce^{\lambda y} + De^{-\lambda y}].$$

Using the boundary condition $u(0, y) = 0$ yields $A = 0$. The condition $u(a, y) = 0$ gives

$$B\sin(\lambda a)][Ce^{\lambda y} + De^{-\lambda y}] = 0.$$

For a non-trivial solution,

$$\begin{aligned} B \neq 0 \quad &\Longrightarrow \quad \sin\lambda a = 0 \\ &\Longrightarrow \quad \lambda a = n\pi \text{ or } \lambda = \frac{n\pi}{a}, \;\; n = 1, 2, 3, \ldots . \end{aligned}$$

Therefore, the sequence of non-trivial is given by

$$u_n(x, y) = \sin(\frac{n\pi x}{a})[C_n e^{\frac{n\pi y}{a}} + D_n e^{\frac{-n\pi y}{a}}]$$

Applying the BC $u(x, b) = 0$, we obtain

$$\begin{aligned} & \sin(\frac{n\pi x}{a})[C_n e^{\frac{n\pi b}{a}} + D_n e^{\frac{-n\pi b}{a}}] = 0 \\ \Longrightarrow \quad & C_n e^{\frac{n\pi b}{a}} + D_n e^{\frac{-n\pi b}{a}} = 0 \\ \Longrightarrow \quad & D_n = -C_n \frac{e^{\frac{n\pi b}{a}}}{e^{\frac{-n\pi b}{a}}}, \;\; n = 1, 2, \ldots, . \end{aligned}$$

Therefore, the solution now takes the form

$$\begin{aligned} u_n(x, y) &= \sin(\frac{n\pi x}{a}) \frac{2C_n}{e^{\frac{-n\pi b}{a}}} \left\{ e^{\frac{n\pi(y-b)}{a}} - e^{\frac{-n\pi(y-b)}{a}} \right\} /2 \\ &= \frac{2C_n}{e^{\frac{-n\pi b}{a}}} \sin(\frac{n\pi x}{a}) \sinh(\frac{n\pi(y-b)}{a}). \end{aligned}$$

Setting $c_n = \frac{2C_n}{e^{\frac{-n\pi b}{a}}}$ and using superposition principle, we obtain

$$u(x, y) = \sum_{n=1}^{\infty} c_n \sin(\frac{n\pi x}{a}) \sinh(\frac{n\pi(y-b)}{a}).$$

To satisfy the remaining nonhomogeneous BC, we must have

$$u(x, 0) = f(x) = \sum_{n=1}^{\infty} c_n \sin(\frac{n\pi x}{a}) \sinh(\frac{-n\pi b}{a}),$$

which is a half-range Fourier series. Therefore,

$$c_n \sinh(\frac{-n\pi b}{a}) = \frac{2}{a}\int_0^a f(x)\sin(\frac{n\pi x}{a})dx,$$

and this implies

$$c_n = \frac{2}{a\sinh(\frac{-n\pi b}{a})}\int_0^a f(x)\sin(\frac{n\pi x}{a})dx. \tag{7}$$

Therefore, the required solution to the problem (3)-(4) is

$$u(x,y) = \sum_{n=1}^{\infty} c_n \sin(\frac{n\pi x}{a})\sinh(\frac{n\pi(y-b)}{a})$$

with the coefficients c_n computed from (7).

As a consequence of the superposition principle we obtain the following result.

Theorem 2. *Let a_n, b_n, c_n and d_n be the Fourier coefficients of $f(x)$, $g(x)$, $h(y)$ and $k(y)$. Then solution of the Dirichlet problem*

$$\begin{aligned} \textit{PDE:} \quad & u_{xx} + u_{yy} = 0, \quad 0 < x < a, \;\; 0 < y < b, \\ \textit{BC:} \quad & u(x,0) = f(x), \quad u(x,b) = g(x) \;\; 0 \le x \le a, \\ & u(0,y) = h(y), \quad u(a,y) = k(y), \;\; 0 \le y \le b, \end{aligned}$$

is

$$\begin{aligned} u(x,y) \;=\; & \sum_{n=1}^{\infty} \Big[A_n \sin(\frac{n\pi x}{a})\sinh[\frac{n\pi(b-y)}{a}] \\ & + B_n \sin(\frac{n\pi x}{a})\sinh(\frac{n\pi y}{a}) \\ & + C_n \sin(\frac{n\pi y}{b})\sinh[\frac{n\pi(a-x)}{b}] \\ & + D_n \sin(\frac{n\pi y}{b})\sinh(\frac{n\pi x}{b})\Big], \end{aligned}$$

where

$$\begin{aligned} A_n &= a_n/\sinh(\frac{n\pi b}{a}) \quad B_n = b_n/\sinh(\frac{n\pi b}{a}) \\ C_n &= c_n/\sinh(\frac{n\pi a}{b}) \quad D_n = d_n/\sinh(\frac{n\pi a}{b}). \end{aligned}$$

Practice Problems

1. Solve the following BVP:

$$\begin{aligned} & u_{xx} + u_{yy} = 0, \quad 0 < x < 1, \;\; 0 < y < 1, \\ & u(x,0) = x(x-1), \quad u(x,1) = 0, \;\; 0 \le x \le 1, \\ & u(0,y) = 0, \quad u(1,y) = 0, \;\; 0 \le y \le 1, \end{aligned}$$

2. Solve the following BVP:

$$
\begin{aligned}
&u_{xx} + u_{yy} = 0, \quad 0 < x < \pi, \;\; 0 < y < \pi, \\
&u(x,0) = \sin x, \quad u(x,1) = \sin x, \;\; 0 \le x \le \pi, \\
&u(0,y) = \sin y, \quad u(1,y) = \sin y, \;\; 0 \le y \le \pi,
\end{aligned}
$$

The Mixed BVP for a Rectangle

we shall consider solving the mixed BVP for the Laplace equation. To begin with, let us consider the following the Neumann problem for a rectangle:

$$\begin{aligned} \text{PDE:} \quad & u_{xx} + u_{yy} = 0, \quad 0 < x < a, \ \ 0 < y < b && (1) \\ \text{BC:} \quad & u_y(x,0) = f(x), \quad u_y(x,b) = g(x), \ \ 0 \le x \le a && (2) \\ & u_x(0,y) = h(y), \quad u_x(a,y) = k(y), \ \ 0 \le y \le b. \end{aligned}$$

This problem has no solution, unless the following compatibility condition holds:

$$\int_0^a g(x)dx - \int_0^a f(x)dx + \int_0^b k(y)dy - \int_0^b h(y)dy = 0.$$

Solution. If $u(x,y)$ is a solution of (1), then

$$\begin{aligned} 0 &= \int_0^b \int_0^a (u_{xx} + u_{yy})dxdy = \int_0^b \int_0^a u_{xx}dxdy + \int_0^a \int_0^b u_{yy}dydx \\ &= \int_0^b [u_x(a,y) - u_x(0,y)]dy + \int_0^a u_y(x,b) - u_y(x,0)]dx \\ &= \int_0^b k(y)dy - \int_0^b h(y)dy + \int_0^a g(x)dx - \int_0^a f(x)dx, \end{aligned}$$

where we have used the fundamental theorem of calculus, and the Fubini's theorem.

Remark 1.

- *The compatibility condition is an immediate consequence of the following special case of Green's theorem*

$$\int_C \nabla u \cdot \mathbf{n} ds = \int_C u_x dy - u_y dx = \int\int_R (u_{xx} + u_{yy})dxdy,$$

i.e., the flux of the gradient of u through the boundary is the integral of Δu in the interior.

- *Note that we only require that u_x and u_y be continuous on the closed rectangle. Further, we do not demand that the second partial of u extend continuously to the closed rectangle.*

We now consider solving Laplace equation with mixed type of boundary conditions.

Example 2. *Solve the following BVP:*

$$\begin{aligned} \textit{PDE:} \quad & u_{xx} + u_{yy} = 0, \quad 0 < x < a, \ \ 0 < y < b, && (3) \\ \textit{BC:} \quad & u(x,0) = 0, \quad u(x,b) = 0, \ \ 0 \le x \le a, && (4) \\ & u(0,y) = g(y), \quad u_x(a,y) = h(y), \ \ 0 \le y \le b. \end{aligned}$$

Solution. The solution of this problem has a form

$$u(x,y) = u_1(x,y) + u_2(x,y),$$

where u_1 and u_2 satisfy (3) with the BC

$$(BC)_1 : \begin{cases} u_1(x,0) = u_1(x,b) = 0, & 0 \le x \le a, \\ u_1(0,y) = g(y), \;\; u_{1x}(a,y) = 0, & 0 \le y \le b, \end{cases}$$

and

$$(BC)_2 : \begin{cases} u_2(x,0) = u_2(x,b) = 0, & 0 \le x \le a, \\ u_2(0,y) = 0, \;\; u_{2x}(a,y) = h(y), & 0 \le y \le b. \end{cases}$$

We shall determine each one of u_1 and u_2 by the method of separation of variables.

Step 1.(Solving for u_1): Separating variables for $u_1(x,y) = X(x)Y(y)$ and substituting in (3) we obtain

$$\frac{X''(x)}{X(x)} + \frac{Y''(y)}{Y(y)} = 0.$$

This leads to the following ODEs:

$$X''(x) + \lambda X(x) = 0, \quad 0 < x < a, \tag{5}$$

$$Y''(y) - \lambda Y(y) = 0, \;\; 0 < y < b, \tag{6}$$

for a constant λ. Since u_1 satisfies $(BC)_1$, we must have

$$Y(0) = Y(b) = 0, \tag{7}$$

$$X'(a) = 0. \tag{8}$$

Nontrivial solutions of (6) with BC (7) are

$$Y_n(y) = \sin \frac{n\pi y}{b}$$

corresponding to

$$\lambda = \lambda_n = -\left(\frac{n\pi}{b}\right)^2, \;\; n \in \mathbb{N}.$$

The differential equation for $X(x)$

$$X''(x) - \left(\frac{n\pi}{b}\right)^2 X(x) = 0$$

has solution of the form

$$X(x) = C_1 \cosh \frac{n\pi x}{b} + C_2 \sinh \frac{n\pi x}{b}.$$

The condition (8) yields $C_2/C_1 = \tanh \frac{n\pi a}{b}$. Thus, a sequence of solutions $X(x)$ is given by

$$X_n(x) = a_n \left(\cosh \frac{n\pi x}{b} - \tanh \frac{n\pi a}{b} \sinh \frac{n\pi x}{b} \right).$$

By superposition principle the product solution u_1 is expressed by

$$u_1(x,y) = \sum_{n=1}^{\infty} a_n \left(\cosh \frac{n\pi x}{b} - \tanh \frac{n\pi a}{b} \sinh \frac{n\pi x}{b} \right) \sin \frac{n\pi y}{b}. \tag{9}$$

The boundary condition $u_1(0,y) = g(y), \;\; 0 \le y \le b$ yields

$$u_1(0,y) = \sum_{n=1}^{\infty} a_n \sin \frac{n\pi y}{b} = g(y), \;\; 0 \le y \le b,$$

with a_n's given by

$$a_n = \frac{2}{b} \int_0^b g(y) \sin \frac{n\pi y}{b} dy. \tag{10}$$

Step 2.(Solving for u_2): Suppose $u_2(x,y) = X(x)Y(y)$ satisfies (3) and $(BC)_2$. Arguing as before, we have the ODEs (5) and (6) for $X(x)$ and $Y(y)$ with the boundary conditions

$$Y(0) = Y(b) = 0; \quad X(0) = 0.$$

The non-trivial solutions corresponding to

$$\lambda = \lambda_n = -\left(\frac{n\pi}{n}\right)^2, \quad n \in \mathbb{N},$$

are

$$Y_n(y) = \sin \frac{n\pi y}{b}.$$

For $X(x)$, we have the ODE:

$$\begin{aligned} X''(x) - \left(\frac{n\pi}{b}\right)^2 X(x) &= 0, \\ X(0) &= 0. \end{aligned}$$

It has solutions of the form

$$X_n(x) = b_n \sin h \frac{n\pi x}{b}, \quad n \in \mathbf{N}.$$

Thus, $u_2(x,y)$ is given by

$$u_2(x,y) = \sum_{n=1}^{\infty} b_n \sinh \frac{n\pi x}{b} \sin \frac{n\pi y}{b} \tag{11}$$

which satisfies the boundary condition $u_{2x}(a,y) = h(y)$. This leads to

$$b_n = \frac{2}{n\pi} \frac{1}{\cosh \frac{n\pi a}{b}} \int_0^b h(y) \sin \frac{n\pi y}{b} dy. \tag{12}$$

Step 3.(Writing the solution): The solution of (3)-(4) is obtained as

$$u(x,y) = u_1(x,y) + u_2(x,y),$$

where a_n and b_n are determined by (10) and (12), respectively.

Practice Problems

1. Solve the following Neumann BVP:

$$\begin{aligned} &u_{xx} + u_{yy} = 0, \quad 0 < x < a, \;\; 0 < y < b, \\ &u_y(x,0) = 0, \quad u_y(x,b) = h(x), \;\; 0 \le x \le a, \\ &u_x(0,y) = 0, \quad u_x(a,y) = 0, \;\; 0 \le y \le b. \end{aligned}$$

   given that $g(x)$ is continuous and $\int_0^a h(x)dx = 0$. Why the assumption $\int_0^a h(x)dx = 0$ is needed?

2. Find a solution of the Neumann BVP:

$$\begin{aligned} &u_{xx} + u_{yy} = 0, \quad 0 < x < \pi, \;\; 0 < y < \pi, \\ &u_y(x,0) = \cos x, \quad u_y(x,b) = 0, \;\; 0 \le x \le \pi, \\ &u_x(0,y) = 0, \quad u_x(a,y) = 0, \;\; 0 \le y \le \pi. \end{aligned}$$

   By adding a constant, find a solution such that $u(0,0) = 0$.

3. Solve the following mixed BVP:

$$\begin{aligned} &u_{xx} + u_{yy} = 0, \quad 0 < x < a, \;\; 0 < y < b, \\ &u(x,0) = 2x, \quad u(x,b) = x^2, \;\; 0 \le x \le a, \\ &u_x(0,y) = 0, \quad u_x(a,y) = 0, \;\; 0 \le y \le b. \end{aligned}$$

The Dirichlet Problem for the Disk

The Dirichlet problem in a disk of radius r_0 and center at $(0,0)$ can be expressed as

$$\begin{aligned} \text{PDE:} \qquad & U_{rr} + \frac{U_r}{r} + \frac{U_{\theta\theta}}{r^2} = 0, \quad 0 < r < r_0, \ \ -\pi \le \theta \le \pi, \\ \text{BC:} \qquad & U(r_0, \theta) = f(\theta), \ \ -\pi \le \theta \le \pi, \end{aligned} \tag{1}$$

where $f(\theta)$ is a given periodic, continuous function of period 2π ($f(\theta + 2\pi) = f(\theta)$). To solve the above problem, we use the method of separation of variables.

Step 1.(Writing the ODEs): Seek solutions of the form

$$U(r, \theta) = R(r)T(\theta),$$

where $0 \le r \le r_0$ and $-\pi \le \theta \le \pi$. Substituting into (1) and separating variables yield

$$\begin{aligned} & R''(r)T(\theta) + r^{-1}R'(r)T(\theta) + r^{-2}R(r)T''(\theta) = 0. \\ \Longrightarrow \qquad & \frac{r^2R''(r) + rR'(r)}{R(r)} = -\frac{T''(\theta)}{T(\theta)} = k. \end{aligned}$$

Which leads to the following two ODEs:

$$T''(\theta) + kT(\theta) = 0, \tag{2}$$

$$r^2R''(r) + rR'(r) - kR(r) = 0. \tag{3}$$

Step 2.(Solving the ODEs):

Case (a): When $k < 0$, the general solution to (2) is the sum of two exponentials. Hence we have only trivial 2π-periodic solutions (see, Lecture 5).

Case (b): When $k = 0$, we find that $T(\theta) = A\theta + B$ is the solution to (2). This linear function is periodic only when $A = 0$, that is, $T_0(\theta) = B$ is the only 2π-periodic solution corresponding to $k = 0$.

Case (c): When $k > 0$, the general solution to (2) is

$$T(\theta) = A\cos(\sqrt{k}\theta) + B\sin(\sqrt{k}\theta).$$

In this case we get a nontrivial 2π-periodic solution only when $\sqrt{k} = n$, $n = 1, 2, \ldots$. Hence, we obtain the nontrivial 2π-periodic solutions

$$T_n(\theta) = A_n\cos(n\theta) + B_n\sin(n\theta) \tag{4}$$

corresponding to $\sqrt{k} = n,\ n = 1, 2, \ldots$.

Now for $k = n^2$, $n = 0, 1, 2, \ldots$, equation (3) is the Cauchy-Euler equation

$$r^2 R''(r) + rR'(r) - n^2 R(r) = 0. \tag{5}$$

When $n = 0$, the general solution is

$$R_0(r) = C + D \ln r.$$

Since $\ln r \to \infty$ as $r \to 0^+$, this solution is unbounded near $r = 0$ when $D \neq 0$. Therefore, we must choose $D = 0$ if $U(r, \theta)$ is to be continuous at $r = 0$. We now have $R_0(r) = C$ and so $U_0(r, \theta) = R_0(r)T_0(\theta) = CB$. For convenience, we write $U_0(r, \theta)$ in the form

$$U_0(r, \theta) = \frac{A_0}{2}, \tag{6}$$

where A_0 is an arbitrary constant.

When $k = n^2$, $n = 1, 2, \ldots$, the general solution of (3) is given by

$$R_n(r) = C_n r^n + D_n r^{-n}.$$

Since $r^{-n} \to \infty$ as $r \to 0^+$, we must set $D_n = 0$ in order for $u(r, \theta)$ to be bounded at $r = 0$. Thus

$$R_n(r) = C_n r^n$$

Now for each $n = 1, 2, \ldots$, we have the solutions

$$U(r, \theta) = R_n(r)T_n(\theta) = C_n r^n [A_n \cos(n\theta) + B_n \sin(n\theta)].$$

By superposition principle, we write

$$U(r, \theta) = \frac{A_0}{2} + \sum_{n=1}^{\infty} C_n r^n [A_n \cos(n\theta) + B_n \sin(n\theta)].$$

This series may be written in the equivalent form

$$U(r, \theta) = \frac{A_0}{2} + \sum_{n=1}^{\infty} \left(\frac{r}{r_0}\right)^n [A_n \cos(n\theta) + B_n \sin(n\theta)], \tag{7}$$

where the A_n's and b_n's are constants. These constants can be determined from the boundary condition. With $r = r_0$ in (7), we have

$$f(\theta) = \frac{A_0}{2} + \sum_{n=1}^{\infty} [A_n \cos(n\theta) + B_n \sin(n\theta)].$$

Since $f(\theta)$ is 2π-periodic, we recognize that A_n, B_n are Fourier coefficients. Thus

$$A_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(\theta)\cos(n\theta)d\theta, \quad n = 0, 1, \ldots, \tag{8}$$

$$B_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(\theta)\sin(n\theta)d\theta, \quad n = 1, \ldots, \tag{9}$$

We now summarize the Dirichlet problem for a disk as follows.

In the Dirichlet problem(1), if

$$f(\theta) = \frac{A_0}{2} + \sum_{n=1}^{\infty}[A_n\cos(n\theta) + B_n\sin(n\theta)],$$

then the solution is given by

$$U(r,\theta) = \frac{A_0}{2} + \sum_{n=1}^{\infty}\left(\frac{r}{r_0}\right)^n [A_n\cos(n\theta) + B_n\sin(n\theta)],$$

where A_n and B_n are given by (8) and (9), respectively.

Example 1. *Solve the following BVP*

$$\begin{aligned} \textit{PDE:} \qquad & U_{rr} + \frac{U_r}{r} + \frac{U_{\theta\theta}}{r^2} = 0, \quad 0 \le r < 1, \\ \textit{BC:} \qquad & U(1,\theta) = f(\theta), \end{aligned}$$

where $f(\theta) = 1 + r\sin\theta + \frac{r^3}{2}\sin(3\theta) + r^4\cos(4\theta)$.

Solution. Here $r_0 = 1$. Note that $f(\theta)$ is already in the form of Fourier series, with

$$A_n = \begin{cases} 2 & \text{for } n = 0 \text{ and 1 for } n = 4 \\ 0 & \text{for other } n \end{cases} \qquad B_n = \begin{cases} 1 & n = 1 \\ \frac{1}{2} & n = 3 \\ 0 & \text{for other } n \end{cases}$$

The solution of the BVP is

$$\begin{aligned} U(r,\theta) &= \frac{A_0}{2} + \sum_{n=1}^{\infty}\left(\frac{r}{r_0}\right)^n [A_n\cos(n\theta) + B_n\sin(n\theta)] \\ &= 1 + r\sin\theta + \frac{r^3}{2}\sin(3\theta) + r^4\cos(4\theta). \end{aligned}$$

Exterior Dirichlet Problem: We shall discuss the exterior Dirichlet problem i.e., the Dirichlet problem outside the circle. The exterior Dirichlet problem is given by

$$\begin{aligned} \text{PDE:} \qquad & U_{rr} + \frac{U_r}{r} + \frac{U_{\theta\theta}}{r^2} = 0, \quad 1 \le r < \infty, \\ \text{BC:} \qquad & U(1,\theta) = f(\theta), \ \ 0 \le \theta \le 2\pi. \end{aligned}$$

This problem is solved exactly in a manner similar to the interior Dirichlet problem. We assume that the solutions are bounded as $r \to \infty$. Basically, we throw out the solutions

$$r^n \cos(n\theta), \quad r^n \sin(n\theta), \quad \ln r$$

that are unbounded as $r \to \infty$.

The solution is given by

$$U(r,\theta) = \sum_{n=0}^{\infty} r^{-n}[A_n \cos(n\theta) + B_n \sin(n\theta)], \tag{10}$$

where A_n and B_n are given by

$$\begin{aligned} A_0 &= \frac{1}{2\pi}\int_0^{2\pi} f(\theta)d\theta, \\ A_n &= \frac{1}{\pi}\int_0^{2\pi} f(\theta)\cos(n\theta)d\theta, \\ B_n &= \frac{1}{\pi}\int_0^{2\pi} f(\theta)\sin(n\theta)d\theta. \end{aligned}$$

The detail procedure is thus left as an exercise.

Practice Problems

1. Solve the Dirichlet problem

$$\begin{aligned} &U_{xx} + U_{yy} = 0, \quad (x^2 + y^2 < 1), \\ &u(1,\theta) = \sin^2\theta, \quad -\pi \le \theta \le \pi, \end{aligned} \tag{11}$$

   for the disk $r \le 1$.

2. Solve the BVP

$$\begin{aligned} &U_{rr} + \frac{U_r}{r} + \frac{U_{\theta\theta}}{r^2} = 0 \quad 0 \le r < 2, \quad -\pi < \theta < \pi, \\ &U(2,\theta) = 1 + 8\sin\theta - 32\cos(4\theta) \quad -\pi < \theta < \pi. \end{aligned}$$

3. Show that the exterior Dirichlet problem

$$\begin{aligned} &U_{rr} + \frac{U_r}{r} + \frac{U_{\theta\theta}}{r^2} = 0 \quad 1 \le r < \infty, \\ &U(1,\theta) = 1 + \sin\theta + \cos(3\theta) \quad 0 < \theta < 2\pi, \end{aligned}$$

   has the solution

$$U(r,\theta) = 1 + \frac{1}{r}\sin\theta + \frac{1}{r^3}\sin(3\theta).$$

POST GRADUATE DEGREE PROGRAMME (CBCS)

M.SC. IN MATHEMATICS

SEMESTER I

SELF LEARNING MATERIAL

PAPER: COR 1.3
(Pure and Applied Streams)

Potential Theory

Abstract Algebra

Operations Research

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

Course Preparation Team

Dr. Samares Pal, Professor, Department of Mathematics, University of Kalyani

Dr. Sanjib Kumar Datta, Professor, Department of Mathematics, University of Kalyani

Dr. Saidul Islam, Assistant Professor, Department of Mathematics, University of Kalyani

Dr. Biswajit Mallick, Assistant Professor (Cont), DODL, University of Kalyani

Ms. Audrija Choudhury, Assistant Professor (Cont), DODL, University of Kalyani

Dec 2021

Directorate of Open and Distance Learning, University of Kalyani

Published by the Directorate of Open and Distance Learning

University of Kalyani, 741235, West Bengal

SYLLABUS

COR 1.3

Marks: 100; Credits: 6

Unit	Topic	Counselling Duration
Block I: Potential Theory; Marks 36 (SEE: 30; IA: 06)		
1	Concept of potential and attraction for line, surface and volume distributions of matter.	54 Mins
2	Laplace's equation, problems of attraction and potential for simple distribution of matter	54 Mins
3	Existence and continuity of first and second derivatives of potential within matter. Poisson's equation, work done by mutual attraction, problems	54 Mins
4	Integral theorem of potential theory (statement only) Green's identities, Gauss' average value theorem,	54 Mins
5	Continuity of potential and discontinuity of normal derivative of potential for a surface distribution, potential for a single and double layer, Discontinuity of potential	54 Mins
6	Boundary value problems of potential theory. Green's function, solution of Dirichlet's problem for a half-space	54 Mins
7	Solid and surface spherical harmonics	54 Mins
Block II: Abstract Algebra I; Marks 32 (SEE: 25; IA: 07)		
8	**Preliminaries**: Review of earlier related concepts-Groups and their simple properties	54 Mins
9	**Class equations on groups and related theories**: Conjugacy class equations, Cauchy's theorem,	54 Mins

10	p-Groups, Sylow theorems and their applications, simple groups	54 Mins
11	**Direct Product on groups**: Definitions, discussion on detailed theories with applications	54 Mins
12	**Solvable groups**: Related definitions and characterization theorems, examples	54 Mins
13	**Group action**: Definition and relevant theories with applications	54 Mins
Block III: Operations Research–I; Marks 32 (SEE: 25; IA: 07)		
14	**Extension of Linear Programming Methods :** Theory of Revised Simplex Method and algorithmic solution approaches to linear programs	54 Mins
15	Dual-Simplex Method, Decomposition principle and its use to linear programs for decentralized planning problems	54 Mins
16	**Integer Programming (IP) :** The concept of cutting plane for linear integer programs, Gomory's cutting plane method	54 Mins
17	Gomory's All-Integer Programming Method, Branch-and-Bound Algorithm for general integer programs	54 Mins
18	**Sequencing Models :** The mathematical aspects of Job sequencing and processing problems, Processing n jobs through Two machines, processing n jobs through m machines	54 Mins
19	**Nonlinear Programming (NLP) :** Convex analysis, Necessary and Sufficient optimality conditions, Cauchy's Steepest descent method,	54 Mins
20	Karush-Kuhn-Tucker (KKT) theory of NLP, Wolfe's and Beale's approaches to Quadratic Programs	54 Mins
Total		18 Hours

Block I

Potential Theory

Unit 1

Course structure

1. Newtonian Law of Gravitation

1 Introduction

This unit deals mainly with the Newtonian Law of Gravitation.

It resembles Coulomb's law of electrical forces, which is used to calculate the magnitude of the electrical force arising between two charged bodies. Both are inverse-square laws, where force is inversely proportional to the square of the distance between the bodies. Coulomb's law has the product of two charges in place of the product of the masses, and the electrostatic constant in place of the gravitational constant.

Newton's law has since been superseded by Albert Einstein's theory of general relativity, but it continues to be used as an excellent approximation of the effects of gravity in most applications. Relativity is required only when there is a need for extreme precision, or when dealing with very strong gravitational fields, such as those found near extremely massive and dense objects, or at very close distances (such as Mercury's orbit around the Sun). We will now state the main statement of the law in the coming section.

1.1 Newtonian Law of Gravitation

Two concentrated masses m_1, m_2 located at the Point p_1, p_2 exert on each other, a force of attraction proportional to the product of their masses and

inversely proportional to the square of the distance between them. The direction of the forces is along the line joining a masses, we can write,

$$F = G.\frac{m_1 m_2}{r^2}$$

Or,

$$F = \frac{m_1 m_2}{r^2} \quad [\text{ Considering } G = 1].$$

Field of force :- The force which act on a unit mass field at any point in space is called the value of the field of force.

$$\begin{aligned} F &= m \times 1r^2 \\ &= \frac{m}{r^2} \end{aligned}$$

Let, a mass m be Concentrated at a point $Q(\xi, \eta, \xi)$ and Consider the force which this exerts on a unit mass at $P(x, y, z)$. Let, $P\overline{Q} = r$, then the force acting on the unit mass at P in the direction P to Q.

So, the magnitude of the force,

$$F = \frac{m}{r^2}$$

Where,

$$r^2 = (\xi - x)^2 + (\eta - y)^2 + (\xi z)^2$$

The direction Cosine of the line of action PQ is

$$\frac{\xi_x}{r} = \frac{\eta - y}{r} = \frac{\zeta - z}{r}$$

Therefore, the force has the components

$$\begin{aligned} X &= \frac{m(\xi - x)}{r^3} \\ Y &= \frac{m(\eta - y)}{r^3} \\ Z &= \frac{m(\zeta - z)}{r^3} \end{aligned}$$

For several masses $m_1, m_2, \ldots, m_k$ located at the points $Q_i(\xi_i, \eta_i, \zeta_i)(I = 1, 2, \ldots, K)$.

$$F = \sum_{i=-1}^{K} \frac{m_i}{r_i^2}$$

Now, the components of forces are,

$$\begin{aligned} X &= \sum_{i=1}^{K} \frac{m_i(\xi - x)}{r_i^3} \\ Y &= \sum_{i=1}^{K} \frac{m_i(\eta - y)}{r_i^3} \\ Z &= \sum_{i=1}^{K} \frac{m_i(\zeta - z)}{r_i^3} \end{aligned}$$

We now consider a continuous distribution of mass occupied bounded region V. Let, the volume, $\delta V = \delta x.\delta y.\delta z$ at the point $Q(\xi, \eta, \zeta)$ of V containing the mass $dm = \chi dv$ [χ = density $\chi(\xi, \eta, \zeta)$].

The Components of attraction due to dm at $P(x, y, z)$ out side volume V are

$$\begin{aligned} X &= \int\int_V \int \frac{(\xi - x)}{r^3} \\ Y &= \int\int_V \int \frac{(\eta - y)}{r^3} \\ Z &= \int\int_V \int \frac{(\zeta - z)}{r^3} \end{aligned}$$

If the mass is distributed over a surface S. With density δ then,

$$\begin{aligned} x &= \int\int_S \frac{\xi - x}{r^3} \delta \, ds \\ Y &= \int\int_S \frac{\eta - x}{r^3} \delta \, ds \\ Z &= \int\int_S \frac{\zeta - x}{r^3} \delta \, ds \end{aligned}$$

If the mass is distributed over a line (curve) then,

$$\begin{aligned} X &= \int_s \frac{\xi - x}{r^3} \nu \, ds \\ Y &= \int_s \frac{\eta - y}{r^3} \nu \, ds \\ Z &= \int_s \frac{\zeta - z}{r^3} \nu \, ds \end{aligned}$$

Exercises

1. Compute the mass of the earth ,knowing the force with which it attracts a given mass on its surface, taking its radius to be 3956 miles.Hence show that the earth's mean density is about 5.5 times that of water.

2.Find the attraction of a wire of constant density having the form of an arc of a circle.

3.Find the attraction of a straight homogeneous piece of wire , at any point P of space ,not on the wire.

4.Find the attraction of a homogeneous circular wire at a point P on the axis of the wire.

Summary

In this section, we have learnt about the Newtonian Law of Gravitation and related theorems and applications.

Units 2 & 3

Course structure

- Potential of an attracted particle

Potential of an attracted particle :- Let, a mass m is situated at $Z(\xi, \eta, \zeta)$.

Let, $P(x, y, z)$ be any point on the curve C. We can calculate the work done by the attraction of the mass m when a particle of unit mass is brought from a point P_0 to a point P_1 along any regular curve C joining P_0 and P_1. If we displaced the particle from P to a neighbouring point P' on the curve C. The work done by the attraction of mass m in this displacement is,

$$X dx + Y dy + Z dz$$

$$\begin{aligned} &= \frac{m}{r^3}\left[(\xi - x)dx + (\eta - y)dy + (\zeta - z)dz\right] \\ &= -\frac{m}{r^2}dr \left[r^2 = (\xi - z)^2 + (\eta - y)^2 + (\zeta - z)^2 \text{ on differentiating } \right] \end{aligned}$$

hence, the total work done by the attraction of mass m as the particle unit mas moves from P_0 to P_1.

$$= -\int_c \frac{m}{r^2}dr = \frac{m}{r} \mid_c = \frac{m}{r_1} - \frac{m}{r_0}$$

where

$$\begin{aligned} r_1^2 &= (\xi - x_1)^2 + (\eta - y_1)^2 + (\zeta - z_1)^2 \\ r_0^2 &= (\xi - x_0)^2 + (\eta - y_0)^2 + (\zeta - z_0)^2 \end{aligned}$$

Let, the point P_0 is taken at infinity and replacing r_1 by r we get,

$$\frac{m}{r}$$

Therefore,

$$U = \frac{m}{r}$$

We can say that, the potential is the work done when a particle of unit mass is brought from infinity to a finite distance in the field of attraction of mass m situated at θ when, $r \to \infty$, $U \to 0$.

Similarly, for several masses m_i situated at $Q_i(\xi_i, \eta_i, \zeta_i)$ the potential at point $P(x, y, z)$ is

$$U = \sum_{i=1}^{n} \frac{m_i}{r_i}$$

where,

$$r_i^2 = (\xi - x)^2 + (\eta - y)^2 + (\zeta - z)^2.$$

In a similar way, we can deduce the above formula for a volume distribution.

$$U = \int\int_V \int \frac{\chi dv}{r}$$

For a surface distribution,

$$U = \int\int \frac{\sigma ds}{r}$$

For a line distribution

$$U = \int_s \frac{\gamma ds}{r}.$$

Theorem. *The potential U due a volume distribution of matter of bounded density χ contained in a region of bounded volume V has partial derivatives of all orders and the following relation hold at any point outside V.*

$$\frac{\partial U}{\partial x} = X, \ \frac{\partial u}{\partial y} = Y, \ \frac{\partial u}{\partial z} = Z,$$

$$\nabla^2 U = 0$$

Proof. We know that the potential due to a volume distribution of matter at a point $P(x, y, z)$ outside the distribution,

$$U = \int\int_{v}\int \frac{\chi dv}{r} \; r = \overline{PQ} \tag{1}$$

since, P lies outside V.

$$\begin{aligned} \frac{1}{r} \text{ and } \frac{\partial}{\partial x}\left(\frac{1}{r}\right) = -\frac{1}{r^2}.\frac{\partial r}{\partial x} & \\ &= \frac{\xi - x}{r^3} \text{ exists} \\ &= -\frac{1}{r^2} \times \left[-2(\xi - x)\right]/2r = \frac{(\xi - x}{r^3} \end{aligned}$$

and they are continuous functions in all the six variables $x, y, z, \xi, \eta, \zeta$, also χ is bounded integrable in V. Hence, differentiation under the sign of integration is permissible. Therefore, Differentiating (1) w.r.t. x we get,

$$\frac{\partial U}{\partial x} = \int\int_{V}\int \chi \frac{\partial}{\partial x}\left(\frac{1}{r}\right) dV = \int\int_{V}\int x \frac{(\xi - x)}{r^3} dV = X$$

similarly,

$$\frac{\partial U}{\partial y} = Y$$

and

$$\frac{\partial U}{\partial z} = Z$$

Again,

$$\frac{\partial}{\partial x}\left(\frac{\xi - x}{r^3}\right) = -\frac{1}{r^3} + \frac{3(\xi - x)^2}{r^5},$$

which is a continuous function of $x, y, z, \xi, \eta, \zeta$, since P lies outside V. Differentiating under the sign of integration is permissible.

Diff. (2), w.r.t. x, we get,

$$\frac{\partial^2 U}{\partial x^2} = \int\int_{v}\int \chi \left(-\frac{1}{r^3} + 3\frac{(\xi - x)^2}{r^5}\right) dV.$$

similarly,

$$\begin{aligned}\frac{\partial^2 U}{\partial y^2} &= \int\int_V \int X\left(-\frac{1}{r^3}+\frac{3(\eta-y)^2}{r^5}\right)dV \\ \frac{\partial^2 U}{\partial z^2} &= \int\int_V \int X\left(-\frac{1}{r^3}+\frac{3(\zeta-z)^2}{r^5}\right)dV.\end{aligned}$$

Adding, we get,

$$\begin{aligned}\frac{\partial^2 U}{\partial x^2}+\frac{\partial^2 U}{\partial y^2}+\frac{\partial^2 U}{\partial z^2} &= \int\int_V \int \chi\left[-\frac{3}{r^3}+\frac{3}{r^5}\times r^2\right]dv \\ &= \int\int_V \int \chi\left[-\frac{3}{r^3}+\frac{3}{r^3}\right]dv \\ &= 0\end{aligned}$$

$\Rightarrow \nabla^2 U = 0 \rightarrow$ Laplace's equation.

Theorem : *The potential U and the components of force X, Y, Z due to a volume distribution of matter of piecewise continuous density χ in the bounded volume V exists at point of*

Proof. We know that

$$\begin{aligned}U &= \int\int_V \int \frac{\chi dV}{r}, \\ Z &= \int\int_V \int \frac{\chi(\zeta-z)}{r^3}dV.\end{aligned}$$

let, V be a small region containing where $r = \overline{PQ}$ P in its interior. We shall show that

$$U' = \int\int_{V-\nu} \int \frac{\chi d\nu}{r} \text{ and } Z' = \int\int_{\nu-v} \int \frac{\chi(\zeta-z)}{r^3}dv$$

approach limits as ν shrink down on p. Now, Cauchy's test of convergence of the integral states that a necessary and sufficient condition that U' and

Z' approach limits is that corresponding to small positive number $\epsilon, \mathcal{T}$ a positive number δ such that if v and v' containing the point P and contains in a sphere $\sum$ of radius δ about P, then,

$$\left| \int\int\int_{v-\nu} \frac{\chi dv}{r} - \int\int\int_{V-\nu'} \frac{\chi dv}{r} \right| < \epsilon$$

and

$$\left| \int\int\int_{v-\nu} \frac{\chi(\zeta - z)}{r^3} dv - \int\int\int_{V-\nu'} \frac{\chi(\zeta - z)}{r^3} dv \right| < \epsilon$$

Existence of Potential :-

We have

$$\begin{aligned}
& \left| \int\int\int_{v-\nu} \frac{\chi dv}{r} - \int\int\int_{V-\nu'} \frac{\chi dv}{r} \right| \\
= & \left| \int\int\int_{v-\epsilon} \frac{\chi dv}{r} + \int\int\int_{\epsilon-v} \frac{\chi dv}{r} - \int\int\int_{v-\epsilon} \frac{\chi dv}{r} - \int\int\int_{\epsilon-\nu'} \frac{\chi dv}{r} \right| \\
= & \left| \int\int\int_{\epsilon-v} \frac{\chi dv}{r} - \int\int\int_{\epsilon-v'} \frac{\chi dv}{r^2} \right| \\
\leq & \left| \int\int\int_{\epsilon-v} \frac{\chi dv}{r} \right| + \left| \int\int\int_{\epsilon-v'} \frac{\chi dv}{r} \right| \\
< & B \int\int\int_{\epsilon-V} \frac{dv}{r} + B \int\int\int_{\epsilon-v'} \frac{dv}{r}.
\end{aligned}$$

where B is the upper bound of χ.

$$< 2B \int\int\int_{\epsilon} \frac{dv}{r}.$$

$$\left| \int\int\int_{v-\nu} \frac{\chi dv}{r} - \int\int\int_{v-\nu'} \frac{\chi dv}{r'} \right| < 2B \int\int\int_{\epsilon} \frac{dv}{r}$$

We now,

$$\left[\begin{array}{l} x = rsin\theta\cos\phi \\ y = rsin\theta\sin\phi \\ z = r\cos\theta. \end{array}\right.$$

$$\begin{aligned}
&= \alpha B\int_{r=0}^{\delta}\int_{\theta_0}^{\pi}\int_{\phi=0}^{2\pi}\frac{r^0\sin\theta dr d\theta}{r} \\
&= \pi B\delta^2 \text{ if } \delta < \frac{\sqrt{\epsilon}}{2\sqrt{\pi B}} \\
&= 2B\int_{r=0}^{\delta}\int_{\theta=0}^{\pi}\int_{\phi=0}^{2\pi} r\sin\theta dr d\theta d\phi \\
&= 2B\int_{r=0}^{\delta} rdr\int_{\theta=0}^{\pi}\sin\theta d\theta\int_{\phi=0}^{2\pi} d\phi \\
&= 2B\times\frac{\delta^2}{2}\times[-\cos\theta]_0^{\pi}\times[\theta]_0^{2\pi} \\
&= B\delta^2\times[\cos\pi+\cos 0]\times 2\pi \\
&= B\delta^2, 2\times 2\pi \\
&= 4\pi B\delta^2 < \epsilon \text{ if } \delta < \frac{\sqrt{\epsilon}}{2\sqrt{\pi}}
\end{aligned}$$

Therefore, v' approaches to v i.e., U exists at all points inside the attracting mass.

Exercises

1.Find the potential of a homogeneous straight wire.

2.Show that at a point of its axis, a homogeneous circular wire has a potential

$$U = \frac{M}{d}$$

,where d is the distance of P from a point of the wire.

3.Find the potential of a homogeneous plane rectangular lamina at a point of the normal to the lamina through one corner.

4.Calculate the potential of a circular wire of unit radius and unit mass , at a point 2 units from the center in the plane of the wire.

For the existence of force component Z at all points inside V, we find that

$$\begin{aligned}
& \left| \iiint\limits_{V-\nu} \frac{\chi(\zeta - z)}{r^3} dv - \iiint\limits_{V-v'} \frac{\chi(\zeta - z)}{r^3} dv \right| \\
= & \left| \iiint\limits_{\epsilon - v} \frac{\chi(\zeta - z)}{r^3} dv - \iiint\limits_{\epsilon - v'} \frac{\chi(\zeta - z)}{r^3} d\theta v \right| \\
\leq & \left| \iiint\limits_{\epsilon - v} \frac{\chi(\zeta - z)}{r^3} dv \right| + \left| \iiint\limits_{\epsilon - v'} \frac{\chi(\zeta - z)}{r^3} dv \right| \\
\leq & B \iiint\limits_{\epsilon - v} \frac{|\zeta - z|}{r^3} dv + B \iiint\limits_{\epsilon - v'} \frac{|\zeta - z|}{r^3} dv \\
< & 2B \iiint\limits_{\epsilon} \frac{|\zeta - z|}{r^3} dv. \left| r^2 = (\epsilon - x)^2 + (\eta - y)^2 +_(\zeta - z)^2 |\zeta - z| \leq r \right. \\
< & 2B \iiint\limits_{\epsilon} \frac{r}{r^3} dv \\
= & 2B \int_{r=0}^{\delta} \delta_{\theta=0}^{\pi} \int_{\phi=0}^{2\pi} \frac{r^2 \sin\theta dr d\theta d\phi}{r^2} \\
= & 2B \int_{r=0}^{\delta} \delta_{\theta=0}^{\pi} \int_{\phi=0}^{2\pi} \sin\theta . dr d\theta d\phi \\
= & 2B \int_{r=0}^{\delta} dr. \int_{\theta=0}^{\pi} \sin\theta d\theta. \int_{\phi=0}^{2\pi} d\phi \\
= & 2B \times \delta 2 \times 2\pi \\
= & 8B\pi\delta < \epsilon \text{ if } \delta < \frac{\epsilon}{8B\pi}.
\end{aligned}$$

Hence, z' approaches z, i.e., z exists at all point inside the attracting mass.

Similarly, the force components X and Y exists at all points with in V.

Theorem : *The potential U and the force components X, Y, Z of a volume distribution of matter of piecerise continuous density χ in the bounded volume V are continuous through out the space :*

Unit 4

Course structure

- Continuity of Potential

Continuity of Potential Let, $P(x, y, z)$ be any point in the interior of V_i and $P'(x', y', z')$ is a neighbouring point. Let, ϵ be a sphere of radius δ small enough so that ϵ lies entirely with in V. Then,

$$\begin{aligned}\left|\int\int\int_{\epsilon}\frac{\chi dv}{r}\right| &\leq \int\int\int_{\epsilon}\frac{|x|}{r}dv \\ &< B\int\int\int_{\epsilon}\frac{dv}{r} \\ &= B.\int_{r=0}^{\delta}\int_{\theta=0}^{\pi}\int_{\phi=0}^{2\pi}\frac{r^2\sin\theta dr d\theta d\phi}{r} \\ &= B\times 2\pi\delta^2 \to 0 \text{ as } \delta\to 0\end{aligned}$$

Therefore corresponding a small positive ϵ, $\mathcal{T}$ a positive number δ_1 such that

$$\left|\int\int\int_{\epsilon}\frac{\chi dv}{r}\right| < \frac{\epsilon}{3} \text{ whenever } \delta < \delta_1$$

Similarly,

$$\left|\int\int\int_{\epsilon'}\frac{\chi dv}{r}\right| \leq B\int\int\int_{\epsilon'}\frac{dv}{r'}$$

where ϵ' is the sphere of radius 2δ about p' and enclosing ϵ

$$= B\int_{\phi-0}^{2\pi}\int_{\theta=0}^{\pi}\int_{r\leq 0}^{2\delta}\frac{dv}{r'} = 8B\pi\delta^2 \to 0 \text{ as } \delta\to 0$$

Therefore, corresponding to positive number $\epsilon\mathcal{T}$ a number δ_2 such that,

$$\left|\int\int\int_{v}\frac{\chi dv}{x'}\right| < \frac{\epsilon}{3} \text{ whenever } \delta < \delta_2$$

Again, the function $\frac{1}{r}$ is continuous if P and P' lie with in ϵ and Q lies outside ϵ i.e., inside $V-\epsilon$. Therefore,

$$\left(\frac{1}{r}-\frac{1}{r'}\right)<\frac{\epsilon}{3BV}$$

whenever $\overline{PP'}<\delta_3$. Therefore,

$$\begin{aligned}
\mid U(\rho)-U(\rho')\mid &= \left|\int\int\int_V \frac{\chi dv}{r}-\int\int\int_V \frac{\chi dv}{r'}\right| \\
&= \left|\int\int\int_{V-\epsilon} \frac{\chi dv}{r}+\int\int\int_\epsilon \frac{\chi dv}{r}-\int\int\int_{v-\epsilon} \frac{\chi dv}{r'}-\int\int\int_\epsilon \frac{\chi dv}{r'}\right| \\
&\leq \left|\int\int\int_\epsilon \frac{\chi dv}{r}\right|+\left|\int\int\int_\epsilon \frac{\chi dv}{r'}\right|+\left|\int\int\int_{v-\epsilon}\chi\left(\frac{1}{r}-\frac{1}{r'}\right)dv\right| \\
&< \frac{\epsilon}{3}+\frac{\epsilon}{3}+\int\int\int_{v-\epsilon}\frac{B\epsilon}{3BV}.dv \\
&< \frac{2\epsilon}{3}+\frac{\epsilon}{3V}.V=\epsilon.
\end{aligned}$$

Therefore, $|U(\rho)-U(\rho')|<\epsilon$ whenever $\overline{\rho\rho'}<\delta$. Therefore, U is continuous at P and therefore it is continuous at any point in V. **Continuity of attraction Component :**

$$\begin{aligned}
\left|\int\int\int_\epsilon \frac{\chi(\zeta-z)}{r^3}dv\right| &= \int\int\int_\epsilon \frac{|X||\zeta-z|}{r^3}dv \\
&< B\int\int\int_\epsilon \frac{r}{r^3}dv \\
&= B\int\int\int_\epsilon \frac{dv}{r^2}. \\
&= B\int_{r=0}^{\delta}\int_{\theta=0}^{\pi}\int_{\phi=0}^{2\pi}\frac{r^2\sin\theta dr d\theta d\phi}{r^2} \\
&= B4\pi\delta\to 0 \text{ as } \delta\to 0.
\end{aligned}$$

Therefore, corresponding to arbitrary positive no ϵ there exists a $+ve$ no. δ_1 such that

$$\left|\int\int\int_\epsilon \frac{\chi(\zeta-z)}{r^3}dv\right|<\frac{\epsilon}{3} \text{ whenever } \delta<\delta_1$$

We now choose the radius δ of the sphere ϵ in such that the inequality. $\delta < \delta_1$ is satisfied. Also, we have

$$\begin{aligned}\left|\int\int\int_{\epsilon'}\frac{\chi(\zeta-z')}{r'^3}dv\right| &\leq \int\int\int_{\epsilon'}\frac{|x||\zeta-z'|}{r'^3}dv \\ &< B\int\int\int_{epsilon'}\frac{r'}{r'^3}dv \\ &= B\int\int\int_{epsilon'}\frac{dv}{r'^2} \\ &= B\int_0^{2\delta}\int_{\theta=0}^{\pi}\int_{\phi=0}^{2\pi}\frac{dvr'^2\sin\theta dr' d\theta d\phi}{r'^2} \\ &= 8\pi B\delta \to 0 \text{ as } \delta \to 0.\end{aligned}$$

Corresponding $\epsilon > 0$, there exists a positive number δ_2 such that,

$$\left|\int\int\int_{\epsilon'}\frac{\chi(\zeta-z')}{r'^3}dv\right| < \frac{\epsilon}{3}, \text{ whenever } \delta < \delta_2.$$

Again, $\frac{\zeta-z}{r^3}$ is a continuous function as long as P is inside ϵ and Q is in $V-\epsilon$. Now

$$\left|\frac{\zeta-z}{r^3}-\frac{\zeta-z'}{r'^3}\right| < \frac{\epsilon}{3Bv} \text{ whenever } PP' < \delta_3$$

where v is the volume of the region let, δ' be the smallest of $\delta_1, \delta_2, \delta_3$, then whenever $PP' < \delta'$ we have

$$\begin{aligned}&|Z(p)-Z(p')| \\ &= \left|\int\int\int_{\epsilon}\frac{\chi(\zeta-z)}{r^3}dv + \int\int\int_{v-\epsilon}\frac{\chi(\zeta-z)}{r^3}dv - \int\int\int_{\epsilon}\frac{\chi(\zeta-z')}{r'^3}dv\right. \\ &\qquad \left. - \int\int\int_{v-\epsilon}\frac{\chi(\zeta-z')}{r'^3}dv\right|\end{aligned}$$

Therefore, $|Z(\rho) - Z(\rho')| < \epsilon$ whenever$|\rho - \rho_1' < \delta$ Z is continuous similarly we can prove that the other force components X and Y are continuous at any point in V.

Holder's Condition : A function $f(Q)$ of the coordinates of Q is said to satisfy Holder's condition at P if there are three positive constants $C, A, \alpha(< 1)$ such that,

$$|f(Q) - f(P)| < Ar^{\alpha},\ r = \overline{PQ}$$

for all points Q for which $r \leq C$.

Poisson's equation : Let, U be the potential of a volume distribution of matter with piecewise continuous density χ in a regular region V, then at any interior point P of V at which χ satisfies Holder's Condition, the derivatives of second order of U exists and satisfies Poisson's equation $\nabla^2 U = -4\pi\chi(p)$.

Potential of Double Layer : Let, a mass $(-m)$ be placed at Q and a mass $(+m)$ be placed at Q' where $\overline{QQ'}$ is very small and m is very larch such that, $m\delta l$ is finite and is equal to μ. Then, such a pair forms a dipole and the direction of the vector QQ' is called the axis of the dipole. We now want to find the potential at any point P due to this dipole.

$$\begin{aligned} U &= \lim_{\delta l \to 0} \left[\frac{m}{r+\delta r} - \frac{m}{r}\right] = \lim_{\delta l \to 0} \frac{m\delta l}{\delta l}\left[\frac{1}{r+\delta r} - \frac{1}{r}\right] \\ &= \lim_{\delta l \to 0} \frac{m\delta l}{\delta l}\left[\frac{1}{r} + \left(\delta\xi.\frac{\partial}{\partial\xi} + \delta\eta.\frac{\partial}{\partial\eta} + \delta\zeta\frac{\partial}{\partial\zeta}\right)\left(\frac{1}{r} + \ldots + -\frac{1}{r}\right)\right] \\ &= \lim_{\delta l \to 0} \frac{m\delta l}{\chi}\left[\frac{\Delta\xi}{\delta l}.\frac{\partial}{\partial\xi} + \frac{\delta\eta}{\delta l}.\frac{\partial}{\partial\eta} + \frac{\delta\zeta}{\delta l}.\frac{\partial}{\partial\zeta}\right]\left(\frac{1}{r}\right) \end{aligned}$$

$U = \mu.\frac{\partial}{\partial\gamma}\left(\frac{1}{r}\right)$ where

$$\frac{\partial}{\partial\gamma} = \cos\alpha\frac{\partial}{\partial\xi} + \cos\beta\frac{\partial}{\partial\eta} + \cos\gamma\frac{\partial}{\partial}\zeta.$$

$\cos\alpha, \cos\beta, \cos\gamma$ be the direction cosines of the axes of the doublet (dipole) and μ is called moment of the dipole.

Deduce Potential of a double layer for a surface distribution :

Let us consider two surfaces S and S' at a small distance δl apart. Let, the elements be distributed there in such a way that the negative masses on S'. The axes of a element being every where normal to each other and is directed from negative masses to positive masses. We obtain the double layers a combination of two single layer with opposite density at a small distance from one another. We consider such a double layer with $\overline{QQ'}$ as a typical element when $\delta l \to 0$. Then, corresponding to the elementary area ds, element of strength $-dm$ at Q is ∂ds and $+dm$ at Q' is $+\partial ds$, where ∂ is the surface density. Therefore, the potential U of all such element distributed on the double layer at an external $P(x, y)$ is given by

$$U = \int\int_S \left(\frac{1}{r+\delta r} - \frac{1}{r}\right)\partial ds$$

or

$$\delta l \to 0 \overline{PQ} = r$$

and

$$\begin{aligned} \overline{PQ'} &= r + \delta r \\ &= \int\int_S \frac{1}{\Delta l}\left(\frac{1}{r+\Delta r} - \frac{1}{r}\right)\partial\delta l ds \text{ or } \delta l \to 0 \\ &= \int\int_S \mu.\frac{\partial}{\partial\gamma}\left(\frac{1}{r}\right) ds \text{ as } \delta l \to 0.\partial\delta l \to \mu. \end{aligned}$$

where μ is the moment of double layer and it is assumed that to be bounded and integrable and also it can be shown that the integral on the right hand side converges if P is a point on S.

Theorem : *If the moment μ of a double layer on any surface S is continuous a any point P on S, then the potential U due to the double layer at any Point P' on the normal to S at P approaches limit as $P' \to P$ from either side of surface, this limit suffers a discontinuity across 'S' and the value of discontinuity is $+4\pi\mu(P)$.*

Proof : The potential U at any point $P(x, y, z)$ is given by $U = \int\int_S \mu . \frac{\partial}{\partial \gamma}\left(\frac{1}{r}\right) dS$

$$= \int\int_S \mu \left(\cos\alpha \frac{\partial}{\partial \xi} + \cos\beta \frac{\partial}{\partial \eta} + \cos\gamma \frac{\partial}{\partial \zeta}\right)\left(\frac{1}{r}\right) ds$$

where $\cos\alpha, \cos\beta, \cos\gamma$ be the direction cosines of the normal to S at $Q(\xi, \eta, s)$

$$r = \overline{PQ} = \sqrt{(\xi - x)^2 + (\eta - y)^2 + (\zeta - z)^2}.$$

$$\begin{aligned}
\frac{\partial}{\partial x}\left(\frac{1}{r}\right) &= -\frac{1}{r^2}.\frac{\partial r}{dx} \\
&= -\frac{1}{r^2}\left(\frac{x - \xi}{r}\right) \\
&= -\left(\frac{x - \xi}{r^3}\right).
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \xi}\left(\frac{1}{r}\right) &= -\frac{1}{r^2}.\frac{\partial r}{\partial \xi} \\
&= -\frac{1}{r^2}\left(\frac{\zeta - x}{r}\right) \\
&= \frac{x - \xi}{r^3}
\end{aligned}$$

$$\frac{\partial}{\partial x}\left(\frac{1}{r}\right) = -\frac{\partial}{\partial \xi}\left(\frac{1}{r}\right)$$

similarly,

$$\frac{\partial}{\partial y}\left(\frac{1}{r}\right) = -\frac{\partial}{\partial \eta}\left(\frac{1}{r}\right).$$
$$\frac{\partial}{\partial z}\left(\frac{1}{r}\right) = -\frac{\partial}{\partial \zeta}\left(\frac{1}{r}\right).$$

Therefore,

$$\begin{aligned} U &= -\int\int_S \mu\left[\cos\alpha.\frac{\partial}{\partial x}\left(\frac{1}{r}\right)\right]ds - \int\int_S \mu\left[\cos\beta\frac{\partial}{\partial y}\left(\frac{1}{r}\right)\right]ds \\ &\qquad - \int\int_S \mu\left[\cos\gamma\frac{\partial}{\partial z}\left(\frac{1}{r}\right)\right]ds \\ &= -\frac{\partial}{\partial x}\int\int_s \frac{\mu\cos\alpha}{r}ds - \frac{\partial}{\partial y}\int\int_S \frac{\mu\cos\beta}{r}ds - \frac{\partial}{\partial z}\int\int_S \frac{\mu\cos\gamma}{r}ds \\ &= -\frac{\partial}{\partial x}U_1 - \frac{\partial}{\partial y}U_2 - \frac{\partial}{\partial z}U_3 \end{aligned}$$

$$U_1 = \int\int_S \frac{\mu\cos\alpha}{r}ds \quad U_2 = \int\int_S \frac{\mu\cos\beta}{r}ds \quad U_3 = \int\int_S \frac{\mu\cos\gamma}{r}ds.$$

let,

$$4\mu\cos\alpha = \partial\mu\cos\beta = \mu\cos\gamma$$

Then, this can be regarded as the potential of a surface distribution with surface density. Let us now take the point P as origin and the tangent plane to S at P at $\xi\eta$ plane. Therefore, $\frac{\partial U_1}{\partial x}$ and $\frac{\partial U_2}{\partial y}$, then $\frac{\partial U_1}{\partial x}$ and $\frac{\partial U_2}{\partial y}$ are the tangential derivatives of potential with surface density $\mu\cos\alpha$, $\mu\cos\beta$ respectively and $\frac{\partial U_3}{\partial z}$ is the normal derivative of potential with surface density $\mu\cos\gamma$. Since, $\cos\alpha, \cos\beta$ vanishes at P and μ is bounded, so $\mu\cos\alpha$ and $\mu\cos\beta$ satisfy Holders condition at P. Therefore, two tangential derivatives are continuous and approach to the surface S

$$\left(\frac{\partial U_1}{\partial x}\right)_+ = \left(\frac{\partial U_1}{\partial x}\right)_-$$

and

$$\left(\frac{\partial U_2}{\partial y}\right)_+ = \left(\frac{\partial U_2}{\partial y}\right)$$

where U_1 and U_2 are the potential at $P(0,0,\xi)$ due to a surface distribution of matter on S with density $\mu\cos\alpha, \mu\cos\beta$ respectively. On the other hand at this point the normal derivative $\frac{\partial U_3}{\partial z}$ approaches P' to p from either side of S and this limit suffers a discontinuity and the amount of discontinuity is

$$\left(\frac{\partial U_3}{\partial z}\right)_+ - \left(\frac{\partial U_3}{\partial z}\right)_- = -4\pi\mu(P)$$

Therefore,

$$\begin{aligned} U_+ - U_- &= \left[-\frac{\partial U_1}{\partial x} - \frac{\partial U_2}{\partial y} - \frac{\partial U_3}{\partial z}\right]_+ \\ &\quad - \left[-\frac{\partial U_1}{\partial x} - \frac{\partial U_2}{\partial y} - \frac{\partial U_3}{\partial z}\right]_- \\ &= 4\pi\mu(P) \end{aligned}$$

Exercises

1.Find the potential at interior and exterior points of a closed magnetic shell of constant moment density.

2. Find the potential of a double distribution of constant moment on an open surface.

3. Compare the potential of a homogeneous double distribution on a plane area with the component, normal to the plane, of the force due to a homogeneous plane lamina occupying the same area.

Unit 5

Course structure

- Green's Identities

Green's Identities

Green's 1st Identity : Let V be a closed regular region of space enclosed by a closed surface S. Let, ϕ and ψ be two functions of x, y, z defined in V and continuous in $V + S$ together with their 1'st order partial derivatives. Let, ψ has continuous second order derivatives in $V + S$. Putting

$$\begin{aligned} P &= \phi\frac{\partial\psi}{\partial x} \\ Q &= \phi\frac{\partial\psi}{\partial y} \\ R &= \phi\frac{\partial\psi}{\partial z} \end{aligned}$$

in Gauss divergence theorem

$$\int\int\int_V divF^{\rightarrow}dV = \int\int_S V^{\rightarrow}.nds$$

we get, where

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

$$\Rightarrow \int\int\int_V \phi\nabla^2\psi dv + \int\int\int_V (\vec{\nabla}\phi.\nabla^{\rightarrow}\psi)dv = \int\int_S \phi\frac{\partial}{\partial\nu}\psi ds \qquad (1)$$

and

$$\frac{\partial}{\partial\gamma} = l\frac{\partial}{\partial x} + m\frac{\partial}{\partial y} + \eta\frac{\partial}{\partial z}$$

where $\delta\nu$ is an element of outward directed normal to the surface ds of S. The equation (1) is known as Green's 1st identity. Again, let $\phi\&\psi$ be both continuously differentiable in V and possess continuous second order partial derivatives in $V+S$. Then interchanging ϕ and ψ in (1), we get,

$$\int\int\int_V \psi\nabla^2\phi dv + \int\int\int_V (\vec{\nabla}\phi.\vec{\nabla}\psi)dv = \int\int_S \phi\frac{\partial\phi}{\partial\gamma}ds \tag{2}$$

Subtracting 92) from (1)

$$\int\int\int_V \left(\phi\nabla^2\psi - \psi\nabla^2\phi\right)dv = \int\int_S \left(\phi\frac{\partial\psi}{\partial\gamma} - \psi\frac{\partial\phi}{\partial\gamma}\right)ds \tag{3}$$

Equation (3) is known as, **Green's 2nd identity (second) Harmonic Function :** A function $f(x,y,z)$ is said to be harmonic at a point $\rho(x,y,z)$ if its second order derivatives exists and satisfies Laplace's equation throughout some neighbourhood of the point P. A function $\rho(x,y,z)$ is said to be *regular at infinity* if $\rho f, \rho^2\frac{\partial f}{\partial x}, \rho^2\frac{\partial f}{\partial y}, \rho^2\frac{\partial f}{\partial z}$ are bounded for sufficiently large value of ρ, where $\rho^2 = (x^2+y^2+z^2)$.

Exercises

1.Given a point source of fluid in the presence of an infinite plane barrier, determine the potential of the flow, assumed to be both irrotational and solenoidal.

2.Find the distribution of the charge induced on the walls of a cuboid by a point charge in its interior.

Theorem's on Harmonic functions :

Theorem 1 : *The integral of the normal derivatives of a function* $U(x,y,z)$

vanishes when extended over the boundary S of any closed regular region V in which the function U is Harmonic and continuously differentiable. We can write $\int\int_S \frac{\partial U}{\partial \gamma} ds = 0$ **Theorem :** *Show that every regular Harmonic function can be represented as the sum of Potentials due to a surface distribution and due to a double layer on the surface.* **Proof :** Let, V be a closed region bounded by a surface S and $P(x, y, z)$ be an interior point of V. Then by Green's second identity

$$= -\int\int\int_V \left(\phi\nabla^2\psi - \psi\nabla^2\phi\right) dv = \int\int_S \left(\phi\frac{\partial\psi}{\partial\gamma} - \psi\frac{\partial\phi}{\partial}\gamma\right) ds \qquad (1)$$

let us put $\phi = U$, and $\phi = \frac{1}{r}$, then. where $r = \overline{PQ}, Q = (\xi, \eta, \zeta)$. Since P lies inside V, we surround P by a small space Σ having P has centre and radius ϵ. Let,

$$V' = V - \Sigma$$

For the region V', we have $\frac{1}{r}$ is harmonic therefore, from (1),

$$\begin{aligned} \int\int\int_V \left(U\nabla^2(\frac{1}{r}) - \frac{1}{r}\nabla^2 U\right) dV &= \int\int_S \left(U\frac{\partial}{\partial\nu}\left(\frac{1}{r}\right) - \frac{1}{r}\frac{\partial}{\partial\nu}(U)\right) ds \\ \Rightarrow -\int\int\int_V \frac{1}{r}\nabla^2 U dV &= \int\int_S \left(U\frac{\partial}{\partial\nu}(\frac{1}{r}) - \frac{1}{r}\frac{\partial}{\partial\nu}(u)\right) ds. \end{aligned}$$

$$= \int\int_S \left(U\frac{\partial}{\partial\nu}(\frac{1}{r}) - \frac{1}{r}\frac{\partial}{\partial\nu}U\right) dS + \int\int_\epsilon \left(U\frac{\partial}{\partial\nu}\left(U\frac{\partial}{\partial\nu}(\frac{1}{r}) - \frac{1}{r}\frac{\partial}{\partial\nu}\right.\right. \qquad (2)$$

Now,

$$\begin{aligned}
\int\int_{\epsilon} U\frac{\partial}{\partial\nu}\left(\frac{1}{r}\right)d\epsilon &= -\int\int_{\epsilon} U\left[l\frac{\partial}{\partial\xi}+m\frac{\partial}{\partial\eta}+\eta\frac{\partial}{\partial\zeta}\right)\left(\frac{1}{r}\right)d\epsilon. \\
&= -\int\int_{\epsilon} U\left[l.\frac{1}{r^2}\frac{\partial\gamma}{\partial\xi}+m.\frac{1}{r^2}\frac{\partial\gamma}{\partial\eta}+n.\frac{1}{r^2}\frac{\partial r}{\partial\zeta}\right]d\epsilon \\
&= \int\int_{\epsilon} U\left[\frac{\xi-x}{r}.\frac{1}{r^2}.\frac{\xi-x}{r}+\frac{\eta-y}{r}.\frac{1}{r^2}.\frac{\eta-y}{r}+\frac{\zeta-z}{r}.\frac{1}{r^2}.\frac{\zeta-z}{r}\right]d\epsilon \\
&= \int\int_{\epsilon} U.\frac{1}{r^2}d\epsilon \\
&= \int\int_{\epsilon} Ud\omega \quad \text{where } d\omega \text{ is the solid angle subtended by } d\epsilon \text{ at } P] \\
&= U(P')\int\int_{\epsilon} d\omega \\
&= 4\pi U(P') \quad \text{where } P' \text{ is a point on } \epsilon]
\end{aligned}$$

As $\epsilon \to 0, P' \to P\ U(\rho') \to U(P)$.

$$\int\int_{\epsilon} U\frac{\partial}{\partial\gamma}\left(\frac{1}{r}\right)d\epsilon \to 4\pi U(P)$$

Again

$$\begin{aligned}
\left|\int\int_{\epsilon}\frac{1}{r}\frac{\partial}{\partial\gamma}d\epsilon\right| &\leq \int\int_{\epsilon}\frac{1}{r}\left|\frac{\partial U}{\partial\nu}\right|d\epsilon \\
&\leq B\int\int_{\epsilon}\frac{1}{r}d\epsilon \\
&= B.\frac{1}{\epsilon}\int\int_{S} d\epsilon \\
&\qquad [\text{Total surface area}]
\end{aligned}$$

Therefore, from (2) we get,

$$-\int\int\int_{V}\frac{1}{r}\nabla^2 U dv-\int\int_{S}\left(U\frac{\partial}{\partial\gamma}\left(\frac{1}{r}\right)-\frac{1}{r}\left(\frac{\partial U}{\partial\nu}\right)\right)ds+4\pi U(P)$$

(Proceeding to the limit $\epsilon \to 0$).

$$U(P)=+\int\int\int_{V}\frac{-\frac{1}{4\pi}\nabla^2 U}{r}dV+\int\int_{S}-\frac{U}{4\pi}\frac{\partial}{\partial\nu}\left(\frac{1}{r}\right)ds+\int\int_{S}\frac{\frac{1}{4\pi}\frac{\partial U}{\partial\nu}}{r}ds \quad (3)$$

Equation 93) is known as Green's 3rd identity. Now, if U is harmonic in V, the $h\nabla^2 U = 0$, then

$$U(P) = \int\int_S \frac{\frac{\partial U}{\partial \nu} 4\pi}{r} ds + \int\int_S - \left(\frac{U}{4\pi}\right) \frac{\partial}{\partial \nu}\left(\frac{1}{r}\right) ds \tag{4}$$

The first term on the right hand side of (4) represents the potential of a surface distribution with density $\frac{1}{4\pi}\frac{\partial U}{\partial \nu}$. The second term represents the potential of a double layer with moment $\left(-\frac{U}{4\pi}\right)$. **Gauss integral theorem**

Theorem : *The integral of the outward drawn normal derivative of potential U across any surface S bounding region V is equal to $-4\pi M$, where M is the mass of the region V.* **Proof :** From Green's second identity,

$$\int\int\int_V (\phi\nabla^2\psi - \psi\nabla^2\phi) dv = \int\int_S \left(\phi\frac{d\psi}{d\nu} - \psi\frac{\partial\phi}{\partial\nu}\right) ds \tag{1}$$

where $\partial\nu$ is an element of the outward drawn normal to an element ds of the surface S. We put $\phi = 1$ and $\psi = U$

$$\int\int\int_V \nabla^2 U dv = \int\int_S \frac{\partial}{\partial\nu} ds \tag{2}$$

But we know that for every point inside the attracting mass $\nabla^2 U = -4\pi\chi$, χ is the density, therefore

$$\begin{aligned} \int\int_S \frac{\partial U}{\partial\nu} ds &= = -4\pi \int\int_V \chi dV \\ &= -4\pi M. \text{ Proved).} \end{aligned}$$

Gauss's Average value Theorem :

Statement : The average value over the surface of a sphere of potential U of masses lying entirely inside the sphere is independent of their distributions with in the sphere and is equal to their mass divided by the radius of the

sphere. **Proof :** Let, S be the sphere of radius 'a' and centre at $P(x, y, z)$. Then the potential U is harmonic outside S and regular at infinity. We describe a large sphere $\sum$ of radius ρ concentric with S and let V' is the region bounded by S and Σ. Let, $Q(\xi, \eta, \zeta)$ be any point and $\overline{PQ} = r$. Putting $\phi = U$ and $\psi = \frac{1}{r}$ in Green's second identity, we get,

$$\int\int\int_{V'} \left(\phi\nabla^2\psi - \psi\nabla^2\phi\right) dv = \int\int_S \left(\phi\frac{\partial\psi}{\partial\nu} - \psi\frac{\partial\phi}{\partial\nu}\right) ds + \int\int_\epsilon \left(\phi\frac{\partial\psi}{\partial\nu} - \psi\frac{\partial\phi}{\partial\nu}\right) d\epsilon \tag{1}$$

$$\Rightarrow \int\int\int_{V'} \left(U\nabla^2\left(\frac{1}{r}\right) - \frac{1}{r}\nabla^2 U\right) dv = \int\int_\epsilon U\frac{\partial}{\partial\nu}(\frac{1}{r}\frac{\partial}{\partial\nu}U)d\epsilon$$

$\frac{1}{r}, U$ are harmonic in $V', \nabla^2(\frac{1}{r} = \nabla^2 U = 0$

$$\begin{aligned}
\left|\int\int_\epsilon U\frac{\partial}{\partial\nu}(\frac{1}{r})d\epsilon\right| &= \left|\int\int_\epsilon U\frac{\partial}{\partial r}\left(\frac{1}{r}\right)d\epsilon\right| \\
&= \left|-\int\int_\epsilon - U\frac{1}{r^2}d\epsilon\right| \\
&\leq \left|-\frac{1}{\rho^2}\int\int_\epsilon U d\epsilon\right| \\
&\leq \frac{1}{\rho^2}\int\int\frac{A}{\rho}d\epsilon \\
&= \frac{A}{\rho^3}\int\int_\epsilon d\epsilon \\
&= \frac{A}{\rho^3}\times 4\pi\rho^2 \\
&= \frac{4\pi A}{\rho} \to 0 \ \rho \to \infty.
\end{aligned}$$

$$\int\int_\epsilon U\frac{\partial}{\partial\nu}\left(\frac{1}{r}\right)d\epsilon \to 0 \text{ as } \rho \to \infty.$$

Again,

$$\begin{aligned}
\left|\int\int_\epsilon \frac{1}{r}\frac{\partial U}{\partial \nu}d\epsilon\right| &= \left|\frac{1}{\rho}\int\int_\epsilon \frac{\partial U}{\partial \nu}d\epsilon\right| \\
&= \frac{1}{\rho}\left|\int\int_\epsilon \left(l\frac{\partial U}{\partial \xi}+m\frac{\partial U}{\partial \eta}+n\frac{\partial U}{\partial \zeta}\right)d\epsilon\right| \\
&\leq \frac{1}{\rho}\int\int_\epsilon \left\{|l|\left|\frac{\partial U}{\partial \xi}\right|+|m|\left|\frac{\partial U}{\partial \eta}\right||n|\left|\frac{\partial U}{\partial \zeta}\right|\right\} \\
&\leq \frac{1}{\rho^3}3B\int\int_\epsilon d\epsilon. \\
&= \frac{3B}{\rho^3}\times 4\pi\rho^2. \\
&= \frac{12\pi B}{\rho}\to 0
\end{aligned}$$

Therefore,

$$\int\int_\epsilon \frac{1}{r}\frac{\partial U}{\partial \nu}d\epsilon \to 0 \text{ as } \rho\to\infty.$$

Also,

$$\begin{aligned}
\int\int_S U\frac{\partial}{\partial \nu}(\frac{1}{r})ds &= U\frac{\partial}{\partial r}(\frac{1}{r})ds \\
&= \int\int_S \frac{U}{r^2}ds = \frac{1}{a^2}\int\int_s UdS
\end{aligned}$$

Also,

$$\int\int_S \frac{1}{r}\frac{\partial U}{\partial \nu}ds = -\int\int_S \frac{1}{r}\frac{\partial U}{\partial r}ds = -\frac{1}{a}\int\int_s \frac{\partial U}{\partial r}ds = \frac{1}{a}4\pi M \text{ (By Gauss's Integral theorem)}$$

where M is the mass of the distribution. Now, making $\rho\to\infty$, we have

$$\int\int\int_{V'}\left\{U\nabla^2(\frac{1}{r}-\frac{1}{r}\nabla^2 U)\right\}dv = \int\int_\epsilon\left\{U\frac{\partial}{\partial \nu}(\frac{1}{r})-\frac{1}{r}\frac{\partial}{\partial \nu}(U)\right\}d\epsilon \qquad (1)$$

Now, making $\rho\to\infty$ in (1).

$$\frac{1}{a^2}\int\int_S Uds = \frac{4\pi M}{a}$$

or,

$$\frac{1}{4\pi a^2}\int\int_s udS = \frac{M}{a}.$$

Units 6 & 7

Course structure

- Boundary value problem of potential theory

Boundary value problems of Potential theory :

First boundary value problem or Dirichlet problem : Let, V_i be a finite region enclosed by the surface S with continuously turning normal. Let, $f(x, y, z)$ be a function defined and continuous at all points of S. Dirichlet problem states that "to determine a solution U of Laplace's equation $\nabla^2 U = 0$ in V_i which has a continuous derivative upto second order in V_i and is continuous in $V_i + S$ and takes on prescribed values f on S." **Second boundary value problem or Neumann Problem :** The Neumann problem states that - "to determine a solution U of Laplace's equation $\nabla^2 U = 0$ in V_i bounded by a cloud surface s such that U has continuous derivatives upto 2nd order in V_i and is continuous derivatives upto 2nd order in V_i and is continuous in $v_i + s$ and its normal derivative approaches the given function f on S.

$$\left[\frac{\partial U}{\partial \nu} = f \text{ on } S\right]$$
$$[U = f \text{ on } S] \tag{1}$$

Dirichlet Principle : The 1st boundary value problem is very closely con-

nected with Dirichlet variational principle i.e., Dirichlet integral

$$\begin{aligned} D = D(w) &= \int\int\int_V \left[\left(\frac{\partial W}{\partial x}\right)^2 + \left(\frac{\partial w}{\partial y}\right)^2 + \left(\frac{\partial W}{\partial z}\right)^2\right] dv \\ &= \int\int\int_V (\nabla W)^2 dv \end{aligned} \tag{1}$$

to make a minimum. More precisely, what function among all w which has continuous derivatives in $V + s$ and takes on prescribed value ζ on S, give the smallest value for the integral D. Let, the Dirichlet variational problem has a solution U, then for all admissible function w which satisfy the above continuity and boundary condition we must have

$$D(W) > D(U) \tag{2}$$

If, F also satisfies the continuity conditions and vanishes on S, then

$$W = U + \lambda F, \text{ where } \lambda \text{ is an } D(W) = D(U + \lambda F)$$

arbitrary constant. Now,

$$\begin{aligned} &= \int\int\int_V \left(\frac{\partial U}{\partial x} + \lambda\frac{\partial F}{\partial x}\right)^2 + \left(\frac{\partial U}{\partial y} + \lambda\frac{\partial F}{\partial y}\right)^2 + \left(\frac{\partial U}{\partial z} + \lambda\frac{\partial F}{\partial z}\right)^2 \Bigg] dv \\ &= \int\int\int_V \left\{\left(\frac{\partial U}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 + \left(\frac{\partial U}{\partial z}\right)^2\right\} + \lambda^2\left\{\left(\frac{\partial F}{\partial y}\right)^2 + \left(\frac{2F}{\partial y}\right)^2 + \left(\frac{\partial F}{\partial z}\right)^4\right\} \\ &\quad +2\lambda\left\{\frac{\partial U}{\partial x}.\frac{\partial F}{\partial x} + \frac{\partial U}{\partial y}.\frac{\partial F}{\partial y} + \frac{\partial U}{\partial z}.\frac{\partial F}{\partial z}\right\}\Bigg] dv \\ &= D(U) + \lambda^2 D(F) + 2\lambda D(U, F) \end{aligned} \tag{4}$$

where

$$D(U, F) = \int\int\int_V \left(\frac{\partial U}{\partial x}.\frac{\partial F}{\partial x} + \frac{\partial U}{\partial y}.\frac{\partial F}{\partial y} + \frac{\partial U}{\partial z}.\frac{\partial F}{\partial z}\right) dv$$

Evidently,

$$D(U,U) = D(U)$$

The expression $2\lambda D(U,F)$ is known as : The first variation of $D(U)$. Now, for every function F of the form described $D(U,F) = 0$. Since, if $D(U,F) \neq 0$, we can choose the absolute value of λ in such a manner (so small) that the value of

$$2\lambda D(U,F) + \lambda^2 D(F)$$

could have the same sign as the first term and then choose the value of λ show that $\lambda D(U,F) < 0$, then we have

$$D(U, \lambda F) < D(U)$$

This is a contradiction to our hypothesis that $D(U)$ is minimum. Now, putting $\phi = F$ and $\psi = U$ in Green's first identity,

$$\int\int\int_V \phi \nabla^2 \psi dv + \int\int\int_V \vec{\nabla}\psi dv = \int\int\int_S \phi \frac{\partial \psi}{\partial \nu} ds,$$

we get,

$$\int\int\int_V F\nabla^2 U dv + \int\int\int_V \left[\frac{\partial F}{\partial x}.\frac{\partial U}{\partial x} + \frac{\partial F}{\partial y}.\frac{\partial U}{\partial y} + \frac{\partial F}{\partial z}.\frac{\partial U}{\partial z}\right] dv = \int\int_S F\frac{\partial U}{\partial \nu} ds$$

$$\int\int\int_V F\nabla^2 U dv + D(U,F) = \int\int_S F\frac{\partial U}{\gamma \nu} ds$$

$$D(U,F) = 0,\ F = 0 \text{ on } S,$$

we have,

$$\int\int\int_V F.\nabla^2 U dV = 0$$

From which we conclude, $\nabla^2 U = 0$, throughout V. **Green's function :**

According to Dirichlet problem a harmonic function can be completely determined with in any region if its values on the boundary of the region are given. We now express harmonic function at any point of the region in terms of boundary values by means of a function known as **Green's function.**
Green's function for the interior region : We know that if a potential function U is harmonic in a region bounded by a closed surface S with continuously fering normal and if the values of U and its normal derivative $\frac{\partial U}{\partial n}\frac{\partial U}{\partial \nu}$ are given on S, then the value of U at any point $P(x, y, z)$ is given by Green's third identity

$$U(P) = \frac{1}{4\pi}\int\int\left\{\frac{1}{r}\frac{\partial U}{\partial \nu} - U\frac{\partial}{\partial \nu}\left(\frac{1}{r}\right)\right\}ds \qquad 91)$$

where $r = \overline{PQ}$ and $Q(\xi, \eta, \zeta)$ is a point on S. let, H be a continuously differentiable function harmonic in V. Then putting $\phi = H$ and $\psi = U$ in Green's second identity

$$\int\int_V\int(\phi\nabla^2\psi - \psi\nabla^2\phi)dv = \int\int_S\left(\phi\frac{\partial}{\partial \nu}\psi - \psi\frac{\partial}{\partial \nu}\phi\right)ds \qquad (2)$$

$$\Rightarrow \int\int_V\int\left(H\nabla^2 U - UU\nabla^2 H\right)dv = \int\int_S\left(H\frac{\partial U}{\partial \nu} - U\frac{\partial H}{\partial \nu}\right)ds$$

Since,

$$\nabla^2 U = \nabla^2 H = 0$$

Then,

$$O = \frac{1}{4\pi}\int\int_S\left(H\frac{\partial U}{\partial \nu} - U\frac{\partial H}{\partial \nu}\right)ds \qquad (3)$$

adding (1) and (3), we get,

$$U(P) = \frac{1}{4\pi}\int\int_S \left\{\left(H + \frac{1}{r}\right)\frac{\partial U}{\partial \nu} - U\frac{\partial}{\partial \nu}\left(H + \frac{1}{r}\right)\right\} ds \tag{4}$$

Now, let, H be harmonic, such that it takes the value $-\frac{1}{r}$ at all points on S that is (i.e.), $H + \frac{1}{r} = 0$, at all points Q on S. Now putting

$$G(Q, P) = \frac{1}{r} + H(Q, p) \text{ in (4)}, \tag{5}$$

$$\begin{aligned} U(P) &= \frac{1}{4\pi}\int\int_S \left(G\frac{\partial U}{\partial \nu} - U\frac{\partial}{\partial \nu}G\right) ds \\ &= -\frac{1}{4\pi}\int\int_S U\frac{\partial}{\partial \nu}G(Q, P) ds \end{aligned} \tag{6}$$

The function $G(Q, P)$ satisfying the relation (6), is called is Green's function for the region V.

Exercises

1.Show that any function, harmonic in the region bounded by two concentric spheres is the sum of a function which is harmonic in the interior of the outer sphere, and a function which is harmonic outside the inner sphere.

2. Show that the density of the charge induced on the surface of a sphere by a point charge is inversely proportional to the cube of the distance from the point charge.

References :

1. Foundations of Potential Theory : O.D. Kellog.
2. The Theory of Potential and Spherical Harmonics : W.J. Sterrbeng and T.L. Smith.
3. An Introduction to the Theory of Newtonian Attraction - A.S. Ramsey.
4. The Theory of Poential - P.K. Ghosh

Block II

Abstract Algebra

Unit 8

Course Structure

1. Review Of earlier concepts
2. Groups and their simple Properties

Introduction to Groups

Introduction

In this unit, we will recapitulate the preliminaries of Group theory which we studied in our undergraduate courses.

Definition A group is an ordered pair (G. *) where G is a non-empty set and * is a binary operation on G such that the following properties hold:

(G1) for all a, b, c ϵ G; a * (b*c) = (a*b) *c (Associative)

(G2) there exists a ϵ G such that for any $a \epsilon G$, a*c =a=c*a (existence of an identity)

(G3) for each $a \epsilon G$, there exists $b \epsilon G$ such that a*b = c =b*a (existence of an inverse).

Thus a group is a mathematical system (G,*) satisfying axioms **G1** to **G3**.

Definition Let (G,*) be a group and H be a nonempty subset of G. Then H is called a subgroup of (G,*) if H is closed under the binary operation * and (H,*) is a group.

Note- that every group G has at least two subgroups, viz {e} and G itself. These two are called trivial sub groups. If H is a subgroup of a group G such that H≠{e} and H= G, then H is called a non-trivial subgroup of G.

Example (Q, t) is a group ƶ is a non–empty subset of Q and (ƶ, t) is a group. Therefore (ƶ, +) is a subgroup of (Q,+).

Cosets

Left cosets: Let G be a group and H be a subgroup of G Let a be an element of G. The subset {ah | hϵH} is called a left coset of H in G and is denoted by aH.

Right cosets: The subgroup {ha | h ϵH} is called a right coset of H in G and is denoted by Ha.

Let us first state Lagrange's theorem on the order of a group:

Theorem (Lagrange): Let H be a subgroup of a finite group G. Then the order of H divides the order of G. Then the order of H divides the order of G. In particular |G| = $\frac{[G:H]}{|H|}$.

Summary

In this unit we have mainly recapitulated the following:

- Definition of groups with example
- Subgroups with example
- Cosets
- Lagrange's Theorem

Units 9 & 10

Course Structure

1. Conjugacy Class Equations
2. Cauchy's Theorem
3. Sylow's theorem and their applications
4. Simple Groups

Class equation, Cauchy and Sylow's theorems

Let G be a group. On the set G, define the following relation:

$$P = \{(a, b) \epsilon GXG \mid b = xax^{-1} \; for \; some \; x \epsilon G\}.$$

It can be shown easily that P is an equivalence relation. This relation is called the conjugacy relation and the equivalence class for an element $a \epsilon G$ with respect to this relation P is called conjugacy class of a, which is denoted by cl(a). So $cl(a) = \{b \in G | bPa\} = \{b \epsilon G | b = xax^{-1} \; for \; some \; x \epsilon G\} = \{xax^{-1} | x \epsilon G\}$ Suppose, G is a finite group. Then the number of conjugacy classes are finite. If $a_1, a_2, \ldots, a_k$ are representatives from each of the distinct conjugacy classes, then

$$G = cl(a_1) \cup cl(a_2) \cup \ldots \cup cl(a_k).$$

Let a be an element of G such that $a \epsilon z$, the centre of G. Then $cl(a) = \{xax^{-1} \mid x \in G\} = \{a\}$. In this case, the conjugacy class is said to be trivial. Conversely, if $cl(a) = \{a\}, \; then \; xax^{-1} = a$ for all $x \in G$, which implies $Xa = aX$ for all $x \in G$ and so a∈z (G0. Thus we find that if $a \notin Z(G)$, then $cl(a) \cap Z(G) = \varphi$ and $a \notin Z(G)$, if and only if $cl(a) = \{a\}$. Hence we can partition G as follows $G = Z(G) \cup cl(x_1) \cup cl(x_2) \cup \ldots \cup cl(x_t)$

where $x_1, x_2, \ldots x_t$ are representatives from each of the distinct conjugacy classes $cl(x_i)$, such that $cl(x_i) \cap Z(G) = \varphi$, ie, $cl(x_i)$ are those conjugacy classes which contain more than one element. Hence,

$$|G| = |Z(G)| + \sum_{i=1}^{t} |cl(x_i)|.$$

where $x_1, x_2, \ldots x_t$, are representatives for each of the non trivial conjugacy classes. This equation is called the class equation of a finite group G.

Example Consider the group S_3. The elements of S_3 are

$$e, (1\ 2), (1\ 3), (2\ 3)\ (1\ 2\ 3)\ and\ (1\ 3\ 2).$$

Now $cl(e) = \{e\}$, and we have

$$Cl((1\ 2)) = \{x(12)x^{-1} | x \in S_3\}$$

$$= \{e(1\ 2)\ e^{-1}, (1\ 2\ (1\ 2)\ (1\ 2)^{-1}, (1\ 3)(1\ 2)(\ 1\ 3)^{-1}\ (2\ 3)(1\ 2)(2\ 3)^{-1},$$

$$(1\ 2\ 3)(1\ 2)(\ 1\ 2\ 3)^{-1}, (1\ 3\ 2)(1\ 2)(\ 1\ 3\ 2)^{-1}\}$$

$$= \{(1\ 2), (1\ 3), (2\ 3)\}$$

Similarly, $cl((1\ 2\ 3)) = \{(1\ 2\ 3), (1\ 3\ 2)\}$

Hence, $S_3 = \{e\} \cup \{(12), (13), (123)\} \cup \{(123), (132)\}$

And so,

$6 + |S_3| = |\{e\}| + |\{(12), (13), (23)\}| + |\{(123), (132)\}| = 1 + 3 + 2$

is the class equation of S_3

Definition: Let G be a group and $a \epsilon G$. Then the centralizer of a is the subset $C(a) = \{x \epsilon G | xa = ax\}$

Clearly, $e, a \in C(a)$ and it can easily be shown that $C(a)$ is a subgroup of G such that $\mathbb{Z}(G) \subseteq C(a)$

Theorem: Let G be a finite group and $a \in G$. Then $[G: C(a))] = |cl(a)|$.

Proof. Let $\mathcal{Z}$ be the set of all left cosets of $C(a)$in G, Now, $\mathcal{Z} = \{x \epsilon G\}$ and $cl(a) = \{xax^{-1} | x \epsilon G\}$, Now the function $f: \mathcal{Z} \to cl(a)$ defined by $f(xc(a) = xax^{-1}$ is a well defened bijective function. Indeed, $xC(a) = yC(a)$

$$\Leftrightarrow x^{-1}ya = ax^{-1}y$$

$$\Leftrightarrow xax^{-1} = yay^{-1}.$$

Thus it follows that $|\mathcal{Z}| = |cl(a)|$ and hence, $[G: C(a)] = |cl(a)|$.

Note that in the previous example, $\{C((1\ 2)) = \{e, (1\ 2)\}$.

Hence $[S_3 : C((1\ 2))] = 3 = |cl(1\ 2)| = |\{(1\ 2), (1\ 3), (2\ 3)\}|, again\ C((1\ 2\ 3)) = \{e, (1\ 2\ 3), (1\ 3\ 2)\}, So, [S_3: C((1\ 2\ 3\ 0)] = 2 = |cl(1\ 2\ 3)| = |\{(1\ 2\ 3), (1\ 3\ 2)\}|$.

By virtue of the above theorem, the class equation of a finite group can be written as.

$$|G| = |\mathbb{Z}(G)| + \sum_{i=1}^{t} [G: C(x_i)].$$

$$|G| = |\mathbb{Z}(G)| + \sum_{i=1}^{t} \frac{|G|}{|C(x_i)|},$$

where $x_1, x_2, \ldots x_t$ are class representatives from each of the distinct non-trivial conjugacy classes.

We point out that the class equation is an important tool to study finite groups.

Theorem If G is a group of order $P^w (n > 0)$ then Ƶ(G) $\neq$ $\{e\}$.

Proof. Consider the class equation of the group G,

$$|G| = |\mathbb{Z}(G)| + \sum_{i=1}^{t} \frac{|G|}{|c(xi)|}$$

Where $x_1, x_2, \ldots x_t$ are representatives from each conjugacy classes $cl(x)$ such that $x_i \notin \mathbb{Z}(G)$.

If $x_i \notin \mathbb{Z}(G)$, then $C(x_i) \neq G \; and \; |C(x_i)| = p^{r_i}$ whre $0 < r_i < n, (r_i \epsilon N)$. Hence, $\frac{|G|}{|c(xi)|} = p^{r_i}$, which is divisible by p divides |G| consequently, $p$divides $|\mathbb{Z}(G)|$. Since $|\mathbb{Z}(G)| \geq 1$, it follows that $|\mathbb{Z}(G)| \geq p$ and hence $\mathbb{Z}(G) \neq \{e\}$.

Lemma If G is a finite commutative group of order w such that n is divisible by a prime. p, then G contains an element of order p (whence a subgroup of order p)

Theorem (Cauchy) Let G be a finite group of order n such that n is divisible by a prime p. Then G contains an element of order p and hence a subgroup of order p.

Proof. The proof is by induction on n. If $n = 2$, then G is commutative and the result follows by the above lemma. Make the induction hypothesis that the result is true for all groups of order m such that $2 \leq m < n$. Consider the class equation.

$$|G| = |\mathbb{Z}(G)| + \sum_{a \notin Z(G)} [G : C(a)] \; for \; G$$

If $G = \mathbb{Z}(G)$, then G is cumulative and the result follows by the above lemma. If G $\neq$ Ƶ(G), Then there exists $a \epsilon$G such that $a \notin \mathbb{Z}(G)$. For such an element $a, G \neq C(a)$ and So $[G: C(a)] > 1$, where by Lagrange's theorem $|G| = [G: C(a)].|C(a)| > |C(a)|$

If p divides $|C(a)|$, then by the induction hypothesis, $C(a)$ and thus G has an element of order p. If p does not divide $|C(a)|$ for all $a \notin \mathbb{Z}(G)$ then p must divide $[G: C(a)]$ for all $a \notin \mathbb{Z}(G)$. But in the class equation, p divides $|\mathbb{Z}(G)|$. Since $\mathbb{Z}(G)$ is commutative, we have again by the above lemma that there exists $a \notin \mathbb{Z}(G)$ and hence $a \notin \mathbb{Z}(G)$ order p.

Next we apply Cauchy's theorem to prove that the converse Lagrange's theorem holds for finite commutative groups.

Theorem: Let G be a finite commutative group of order n. If m is a positive integer such that $m|n$, the G has a subgroup of order m.

(We now apply Cauchy's theorem to obtain some interesting properties of p groups)

Definition: Let p be a prime. A group G is said to be a p -group if the order of each element of G is a power of p. A subgroup H of a group G is called a p subgroup if H is a p group.

Example: The Klein 4-group is a p -group. In fact, any group of order p^n(p a þrime) is a p - group since the order of each element must divide the order of the group.

The following theorem gives a necessary and sufficient condition for a finite group to be a p —group.

Theorem H: Let G be a non trivial group. Then G is a finite p - group if and only if $|G| = p^k$ for some positive integer k.

Proof: Suppose G is a finite p-group. If q divides |G| for some prime $q \neq p$, then by Cauchy's theorem H has an element of order q, contradicting the fact that G is a p-group. Thus, p is the only prime divisor of |G|. Hence, $|G| = p^k$ for some positive integer k.

Conversely, suppose $|G| = p^k$. Then by Lagrange's theorem, the order of each element of G is a power of p.

In the next theorem, we prove that the Centre of a p -group is non-trivial.

Theorem If G is a finite p -group with $|G|$>1, then $\mathbb{Z}(G)$, the centre of G, has more than one element, i.e., if $|G| = p^k$ with $k \geq 1$ *then* $|\mathbb{Z}(G)| >1$.

Proof. Consider the class equation

$$|G| = |\mathbb{Z}(G)| + \sum_{a \notin Z(G)} [G : C(a)] \; for \; G$$

If $G = \mathbb{Z}(G)$, then the theorem is immediate. Suppose $G \supset \mathbb{Z}(G)$ and consider $a \epsilon G$ such that $a \notin \mathbb{Z}(G$. Then $C(a)$ is proper subgroup of G so that by the above theorem and by the fact that $C(a)$ is a subgroup of a p-group, $p|[G: c(a)]$ $for\ all\ a \notin \mathbb{Z}(G)$. This implies that p divides $\sum_{a \notin Z(G)}[G : C(a)]$. Since p also divides $|G|$, p divides $|\mathbb{Z}(G)|$. Hence $|\mathbb{Z}(G)| > 1$.

Corollary Let G be a group of order p^2, where p is a prime. Then G is commutative.

Proof. By the above theorem, $|\mathbb{Z}(G)| > 1$. By Lagrange's theorem, $|\mathbb{Z}(G)|$ divides p^2. Hence, $|\mathbb{Z}(G)| = p \; or \; p^2$. Suppose |Z(G)| = p. Then $\mathbb{Z}(G) \neq G$ and so there exists $a \notin G$ such that $a \notin \mathbb{Z}(G)$. Now $C(a)$ is a subgroup of G and $a \in C(a)$.

Hence, $\mathbb{Z}(G) \subset C(a)$. This implies that $|C(a)| = p^2$ and so $G = C(a)$. However, this shows that $a \notin \mathbb{Z}(G)$, a contradiction. Therefore, $|\mathbb{Z}(G)| = p^2$ and so $G = \mathbb{Z}(G)$. Thus, G is commutative.

Result. Let G be a non-commutative group of order p^3, p a prime. Prove that $|\mathbb{Z}(G)| = p$.

Solution: We write $\mathbb{Z} = \mathbb{Z}(G)$. Since $|G| = p^3$, $|\mathbb{Z}| > 1$ by the above theorem, Thus, $|\mathbb{Z}| = p, p^2 \text{or} p^3$. If $|\mathbb{Z}| = p^3$, then $G = \mathbb{Z}$ and so G is commutative, which is a contradiction. If $|\mathbb{Z}| = p^2$, then $|G|\mathbb{Z}| = p$.

Hence, $G|\mathbb{Z}$ is cyclic. But then G is commutative again a contradiction. Thus $|\mathbb{Z}| = p$.

Result. Let G be a group of order $p^2 = p$ a prime, and $w \notin \mathbb{Z}, n \geq 1$. Prove that any subgroup of G of order p^{n-1} is normal in G.

Definition Let H and K be subgroups of a group G. Let $N_k(H)$ denote the set

$$N_K(H) = \{K \epsilon K \mid KHK^{-1} = H\}$$

$N_k(H)$ is called the normalize of H in K

If follows that $N_K(H) = N_G(H) \cap K$.

Theorem (Cayley): A finite group G of order n is isomorphic to a subgroup of S_n.

Sylow's Theorems.

M. L. Sylow did work of fundamental importance in determining the structure of finite groups. We can use his results to answer the problem now posed.

If G is a finite group of order n and if H is a subgroup of G, then we Know by Lagrange's theorem that the order of H divides n. In this section, we give some answers to The question, "If m is a opositive integer, which divides n, does G contain a subgroup of order m?"

It is interesting to note that sylow's theorem was proved by sylow for permutation groups. George Frobenius established the theorem in the general setting the was influenced to do so by Cauchy's theorem.

Theorem (Sylow's First, Theorem)

Let G be a finite group of order p^rm, where p is a prime, r and m are positive integers, and p, m are relatively prime. Then G has a subgroup of order p^k for all $k, 0 \leq k \leq r$.

Corollary Let G be a finite group and p a prime. If p^w divides $|G|$, then G has a subgroup of order p^w

Definition Let G be a finite group and p a prime. A subgroup P of G is called a sylow p-subgroup if P is a p – subgroup and is not property contained in any other p –subgroup of G, i,e. P is a maximal p –subgrop of G.

Example The symmetric group S_3 has three sylow 2-subgroups, namely

$H_1 = \{(\begin{smallmatrix}123\\123\end{smallmatrix}), (\begin{smallmatrix}123\\213\end{smallmatrix})\}$,

$H_2 = \{(\begin{smallmatrix}123\\123\end{smallmatrix}), (\begin{smallmatrix}123\\321\end{smallmatrix})\}$ and

$H_3 = \{(\begin{smallmatrix}123\\123\end{smallmatrix}), (\begin{smallmatrix}123\\132\end{smallmatrix})\}$.

Thus, a Sylow p-subgroup of a given group need not be unique.

The following theorem shows the existence of Sylow p- subgroups in a finite group.

Theorem For each prime p, a finite group G has a sylow p-subgroup.

Proof If $|G| = 1$ or p does not divide $|G|$, then {e} is the required sylow p-subgroup of G. If p divides $|G|$, then by Cauchy's theorem, there is at least one subgroup H of G of order p. Since G is finite, there are a finite number of subgroups of G, which contain H. Hence, one of these subgroups is a sylow p- subgroup of G.

Theorem Let G be a finite group of order $p^r m$ where p is a prime, r and m are positive integers and p, m are relatively prime and P be a subgroup of G. Then

(i) If P is a p group, then any conjugate of P is a p-group.

(ii) If P is a Sylow p-subgroup, then any conjugate of P is a Sylow p-subgroup.

(iii) If P is the only Sylow p-subgroup of G then P is a normal subgroup of G.

The proof is omitted.

Result Let H be a normal subgroup of a group G. If H and G/H are both p-groups, then G is a p-group.

Proof Let $a \epsilon G$. Then aH ϵ G/H and so aH has order some power of p, say p^K . Thus, (aH)$p^K = H$ and so $ap^K \epsilon$ H. Now every element of H has order a power of p.

Let us say ap^k has order p^m . Thus, $(ap^K)\ p^m = e$ or $ap^{m+k} = e$. Thyis implies that $o(a)$ has order some power of p. Since a is arbitrary in G, G is a p – group. This proves the result.

Now we state sylow's second and third theorem without proof.

Sylow's second theorem:

Let G be a finite group of order $p^r m$, where p is a prime, r and m are positive integers and p, m are relatibely prime. Then any two sylow p-subgroups of G are conjugate and therefore isomorphic.

Sylow's third theorem:

Let G be a finite group of order $p^r m$ where p is a prime, r and m are positive integers and p, m are relatively prime. Then the number np of sylow p-subgroups of G is $1 + kp$ for some non negative integer k and $np | p^r m$.

Some problems:

Example: Show that every group of order 45 has a normal subgroup of order 9.

Solution: Let G be a group of order 45 =$3^2 5$ and n_3 denote the number of sylow 3-subgroups of G. Then $n_3 = 3k + 1$ for some integer $k \geq 0$ and $n_3/45$. If $k = 0$, then $n_3 = 1$, which divides 45. But for any $k \geq 1$, n_3 does not divide 45. Hence, G contains a unique Sylow 3-subgroup H of order 9. Consequently that G is simple iff G is prime order.

If this section, we apply the sylow theorems to determine some finite groups which are not simple.

Example: Let G be a group of order 10. Now 10=5.2 Let n_5 denote the number of sylow 5 subgroups of G from sylow's third theorem, $n_5 = 5k + 1$ for some integer $k \geq 0$ and n_5 divides $|G| = 10$. Thus, $n_5 = 1$ and so there exists only one sylow 5-subgroup, Say H in G. Since H is a unique sylow 5-subgroup, H is a normal subgroup of G by the following corollary, proving that G is not simple.

The following corollary is an immediate consequence of sylow's second theorem:

Corollary: Let G be a finite group and H be a sylow p-subgroup of G. Then H is a unique sylow p-subgroup of G if and only if H is a normal subgroup of G.]

Thus no group of order 10 is simple.

Example: Let G be a group of order 9. Then G is a p-group, where $p = 3$, from a previous theorm we finid that $\mathbb{Z}(G) \neq \{e\}$. If $G = \mathbb{Z}(G)$, then G is a commutative group. But commutative simple group are precisely groups of prime order. Hence, in this case G is not simple. Suppose $\mathbb{Z}(G) \neq G$. Then $\mathbb{Z}(G)$ is a now trivial normal subgroup of G. thus, we find that a group of order 9 is not a simple group.

In the above example, we showed that a group of order $9=3^2$ is not simple. In the next theorem, we prove that in general, if G is a p-group of order p^n, $n > 1$ then G is not simple.

Theorem: Let p be a prime integer and n>1 be any integer. Then no group of order p^n is simple.

Proof: Let G be a group of order p^n. Consider the centre Z(G) of G. From a previous theorem, it follows that $\mathbb{Z}(G) \neq \{e\}$. If $G = \mathbb{Z}(G)$, then G is a commutative group. If G is simple, then $|G|$ is prime, which is a contradiction. Thus, in this case G is not simple. Suppose $\mathbb{Z}(G) \neq G$. Then Z(G) is a non trivial normal subgroup of G, proving that G is not a simple group.

Theorem: Let p and q be two prime integers. Then no group of order pq is simple.

*Proof***:** Le G be a group of order pq. If $p = q$ then $|G| = p^2$ and so by a previous theorem, G is not simple. Suppose now $p \neq q$. Let $p > q$. Let n_p denote the number of sylow p-sub groups of G. Then $n_p = pk + 1$ for some interger $k \geq 0$ and n_p divides pq. Since $cd\ (1 + kp, p) = 1$, n_p does not divide p. Hence,$n_p$divides q.

Thus, $1 + kp \leq q$. But $p > q$. Therefore, $1 + kp \leq q$ holds only if $k = 0$. This implies that $n_p = 1$ and so G contains a unique sylow p-sub-group of order p, which must be normal by a previous corollary. Hence G is not simple.

We now state the following two results:

Result: Let G be a finite group and H be a proper subgroup of G of index n such that $|G|$ doest not divide n. Then G contains a non-trivial normal subgroup.

Result: Any group of order $2n$, where n is an odd integer, contains a normal subgroup of order n.

Using this result, we find that no groups of order 6 (=2.3), 18(=2.9), 50 = (2.25), 54(= 2.27) are simple.

Theorem: Let n be an integer such that. $1 \leq n < 6$ and n is not prime. The no group of order n is simple.

Let us now concentrate our discussion on $n = 60$. Sinc 60 is not prime, no commutative group of order 60 is simple. Now what is the answer if G is a non commutative group of order 60? Recall that A_5 is a simple group of order 60. Hence, we find that there exists a non commutative simple group of order 60. Next, let us ask the following question. Is A_5 the only (up to isomorphism) non commutative simple group of order 60? To answer this question, we first prove state the following result.

Result: Let G be a simple group of order 60. Then G contains a subgroup of order 12.

Result: Any simple group of order 60 is isomorphic to A_5.

From above, it follows that A_5 is the smallest non commutative simple group.

Let us now apply the sylow theorems to classify some groups of small order.

Example: Let G be a group of order 15 = 5.3. By Sylow's third theorem G has a sylow 5-subgroup A and a Sylow 3 – subgroup B. It is easy to check that A is a unique Sylow 5 – subgroup and B is a unique sylow 3 – subgroups of G. Hence, A is a normal subgroup of order 5 and B is a normal subgroup or order 3. Now A∩B = {e}. Thus, |AB|

$= \frac{|A|B|}{A \cap B|} = 15$. Hence, G = AB, $A \cap B = \{e\}$ and A, B are normal subgroups of G. Thus, $G = A \times B \simeq Z_5 \times \simeq Z_{15}$ since $gcd\ (3,5) = 1$. Hence, G is a cyclic group.

In the next theorem, we classify all groups of order pq. where p and q are distinct primes.

Theorem: Let G be a group and p, q be primes with $p > q$. If $|G| = pq$, then G is either cyclie or generated by two elements a and b satisfying the following properties: $b^p = e$, and $a^q = e$ and $a^{-1}ba = b^r$, where p does not divide $(r-1)$, but $p|(rq - 1)$. The second possibility can occur only if $q|(p-1)$.

Further we can use the sylow theorems to test the simplicity of finite groups. For example we prove the following:

Example: No group of order 175 is simple.

Proof: Let G be a group of order 175. Now $175 = 5^2.7$. Hence G has sylow 5-subgroups. Let n_5 be the number of sylow 5 – subgroups. Then by sylow's third theorem, $n_5 = 5k + 1$, $k \geq 0$ and $n_5|\ 175$. Hence $n_5 = 1$. So there exists only one sylow 5 –subgroup, say H of order 5^2 in G. Hence by Sylow's second theorem, H is a normal subgroup of order 25 in G. Thus it follows that G is not a simple group.

Example: Show that no group of order 70 is a simple group.

Solution: Let G be a group of order 70. Now $70 = 2.35$. Hence by a previous proposition (any group of order 2n where n is an odd integer, contains a normal subgroup of order n), we find that G contains a normal subgroup of order 35. Hence G is not a simple group.

Example: Let AG be a group of order 30. Since $30 = 2 \cdot 15$, G contains a normal subgroup of order 15. Hence G is not a simple group.

Test by Extended Cayley's Theorem:

We propose to give the statement of this theorem without proof. Then we shall discuss its application to test the simplicity of groups.

Theorem: Let G be a finite group and let H be a subgroup of G of index n. Then there exists a homomorphism $f: G \rightarrow S_n$ such that $Kerf \subseteq H$.

Corollary: Let G be a finite group and let H be a subgroup of a G of index $m \neq 1$. If $|G|$ does not divide m then (show that) G has a nontrivial normal subgroup.

Example: There does not exist a simple group of order 12.

Solution: Let G be a group of order 36. Now 36 = $3^2 \cdot 3^2$ Hence by sylow's first theorem, G has a subgroup H of order 3^2. Now $[G:H] = 4$.

Hence by the Extended Cayley's theorem, there exists homomorphism f: G $\rightarrow S_4$ such that $Ker\, f \subseteq H$. If $Ker\, f = \{e\}$, then f is a homomorphism and hence G is isomorphic to a subgroup K of S_4. In that case, $|K| = |G| = 36$ implies that S_4 contains a subgroup of order 36 which is impossible, since $|S_4| = 24$. Hence $1 < |Kerf| \leq |H| = 9$. Therefore G has a non trivial subgroup $Kerf$. Thus, we find that G is not a simple group.

Example: Show that no group of order 108 is a simple group.

Solution: Let G be a group of order 108. Now 108 = $3^3 \cdot 2^2$. Hence G has sylow 3 – subgroups as well as sylow 2 subgroups. Let n_3 be the number of sylow 3-subgroups. Then by sylow's third theorem n_3 =1 or 4. If $n_3 = 3k+1$, $k \geq 0$ and $n_3 | 108$. Hence n_3 = 1 or 4. If n_3=1, then G has a unique sylow 3 subgroup H of order 3^3, which must be then normal, whence G is not simple.

Suppose n_3=4. Let A and B be two distinct sylow 3-subgroups of G. Now $|A \cap B| \neq 27$. If $|A \cap B| \leq 3$, then

$|AB| = \frac{|A||B|}{|A \cap B|} \geq \frac{27.27}{3} = 243 > 108.$

Hence $|A \cap B| = 9$ and $|AB| = 81$.

Now $|A \cap B| = 3^2$ and $|A| = 3^3$. Hence $A \cap B$ is a normal subgroup in A. Similarly, $A \cap B$ is a normal subgroup in B. Hence $A \subseteq N(A \cap B)$, Where $N(A \cap B) = \{g \in G | g(A \cap B)g^{-1} \subseteq A \cap B\}$, is a subgroup of G.

Hence $AB \subseteq N(A \cap B)$. Now $81 = |AB| \leq |N(A \cap B)|$ and $|N(A \cap B)|$ divides 108. Hence $|N(A \cap B)| = 108$. This implies that $|N(A \cap B)| = G$ and so $A \cap B$ is a normal subgroup of order 9. Therefore G is not simple.

Unit 11

Course Structure

1. Direct Product Of groups: Definitions
2. Applications

Direct Product of Groups

Introduction

In this unit, we will learn about the direct product of groups. This is an important tool to construct new groups from pre-existing groups by using the operation of cross product of sets. We now formally state the definition as follows:

Definition: Let H and K be two groups. Let us consider the set G = H×K. On G we define a binary operation as follows:

(h_1, k_1) o $(h_2, k_2) = (h_1 h_2, k_1, k_2)$

It can be easily verified (verify!) that the associative law under o holds in G. Now, $(e_H, e_K) \in H \times K$ such that for any (h, k) ϵ G we have (e_H, e_K)o (h, k) $= (e_H h, e_K k) = (h, k)$ and $(h, k) \circ (e_H, e_K) = (he_H, ke_K) = (h, k)$. So (e_H, e_K) is the identity in G=H×K.

Let (h, k) ϵ G so that h ϵ H and k ϵ K. This implies $h^{-1} \in H$ and $k^{-1} \in K$. Hence $(h^{-1}, k^{-1}) \in G$ such that (h, k) o $(h^{-1}, k^{-1}) = (hh^{-1}, kk^{-1}) = (e_H, e_K)$ and (h^{-1}, k^{-1}) o (h, k) $= (h^{-1}h, k^{-1}k) = (e_H, e_K)$. Therefore, (h^{-1}, k^{-1}) is the inverse of (h, k) in G = H × K and we denote (h^{-1}, k^{-1}) by $(h, k)^{-1}$.

Thus, under the above define binary operation G = H × K is a group. This group G = H× K is called external direct product (or simply direct product) of H and K.

Definition A Group G is said to be an internal direct product of two subgroups H and K of G if (i) H and K are normal subgroups, (ii) G = HK and (iii) H∩K = {e}.

Definition A Group G is said to be an internal direct product of finite number of subgroups H_1, H_2, ..., H_n, of G if (i) H_1, H_2, ..., H_n are normal subgroups,

(ii) $G = H_1 H_2, ..., H_n$ and

(iii) $H_1 \cap (H_1 H_2, ...H_{i-1} H_{-+1}, ..., H_n) = \{e\}$ for all i = 1,2, ... ,n.

Theorem Let G be a group. H and K be two subgroups of G. Then G is an internal direct product of H and K if and only if

(i) G = HK,

(ii) ab = ba for all a ϵ H, b ϵK

(iii) $H \cap K = \{e\}$.

Proof: Let G be an internal direct product of H and K. Then H and K are normal subgroups of G, G = HK and $H \cap K = \{e\}$.

Now, let aϵH and b ϵ K. Then $aba^{-1}b^{-1} \in aKa^{-1}b^{-1} \subseteq KK \subseteq K$, as $aKa^{-1} \subseteq K \; \& b^{-1} \in K$.

Conversely, assume that the given conditions are satisfied. It remains to show that H K are normal subgroups of G.

Let h ϵ H and g ϵ G = HK. Hence $g = h_1 k_1 \epsilon K$.

$ghg^{-1} = h_1 k_1 h k_1^{-1} h_1^{-1}$

$= h_1 k_1 k_1^{-1} h h_1^{-1}$ [since $hk_1^{-1} = k_1^{-1} h$]

$= h_1 h h_1^{-1} \epsilon H$.

Therefore $gHg^{-1} \subseteq H$ for all gϵ G, $i.e.,$ H is a normal subgroup G. Consequently, G is an internal direct product of H and K.■

Theorem Let H and K be two normal subgroups of a group G. Then G is an internal direct product of H and K if and only if every element g ϵ G can be expressed uniquely as g = hk where h ϵ H &k ϵ K.

Proof: Let G be an internal direct product of H and K. So G = HK. Let g ϵ G. So g = hk for some h ϵ H, k ϵ K. If possible let g = h_1k_1 for some h_1 ϵ H, K_1 ϵ k. Therefore, h_1k_1 = hk. This implies $h^{-1}h_1 = kk_1^{-1}$ ϵ H ∩ K = {e}.

This leads to h = h_1 and k =k_1. Hence every element g ϵ G can be expressed as g = hk for unique h ϵ H & kϵ K. Therefore, G ⊆ HK. Again, H ⊆ G & K ⊆ G. So HK ⊆ G. Hence, G = HK.

Now, let x ϵ H ∩ K. Therefore, x ϵ H and x ϵ K. Now, x = ex = xe. Form unique representation of elements we have x = e. Therefore, H∩K = {e}.

Consequently, G is an internal direct product of H and K.

In general external direct product of two cyclic groups may not be cyclic. For this we consider the cyclic$\mathbb{Z}_2$ and $\mathbb{Z}_4$. But their direct product G = $\mathbb{Z}_2$ x $\mathbb{Z}_4$ is not cyclic, because G is a group or order 8 but there is no element of order 8 in G. In fact each non identity element of G is order 2 or 4. The following theorem establishes the necessary and sufficient condition such that the external direct product of two finite cyclic groups will again be a cyclic group.

Theorem Let A and B be two cyclic groups of order m & n respectively. Then A × B is cyclic if and only if gcd(m, n) =1

Proof: At first, we assume that gcd(m, n) = 1. Now, |A × B| = mn. Let (a, b) ϵ A × B. Then $(a, b)^{mn} = (a^{mn}, b^{mn}) = (a^m)^n, (b^n)^m = (e_A, e_B)$, where A = < a>& B = < b>.

Let t be any other positive integer such that $(a, b)^t = (e_A, e_B)$, i.e., $(a^t, b^t) = (e_A, e_B)$. Then $a^t = e_A$ and $b^t = e_B$, i.e., m| t and n| t. Also gcd (m, n) = 1. So mn| t, i.e., mn ≤ t. Hence o((a, b)) = mn. Consequently, A×B is cyclic.

Conversely, let A×B be a cyclic group, where A = <a> and B = <b> are two cyclic groups or order m and respectively. Let gcd(m, n) = d. Suppose d > 1.

Now, $\frac{mn}{d} = \frac{m}{d}n = \frac{n}{d}m$ *is an integer*. Then for any (x, y) ϵ A × B, we have

$(x, y)^{\frac{mn}{d}} = \left((x^m)^{\frac{n}{d}}, (y^n)^{\frac{m}{d}} = (e_A, e_B\right)$.

This shows that $o((x,y)) \leq \frac{mn}{d} < mn\ (since\ d > 1)$.Hence AxB has no element of order mn. Therefore, the group A x B is not a cyclic group → a contradiction. Hence d = 1 and consequently, gcd (m, n) = 1.

The following theorem shows the connection between internal direct product and external direct product of subgroups of a given group.

Theorem Let G be a group. Let A and B be two subgroups of G. If G is an internal direct product of A and B then G ≅ A X B.

Proof: Since, G is an internal direct product of A and B, so every element of G can be expressed uniquely as a product of elements of A and B. Let x ϵ G. Then there exist unique a ϵ A and unique b ϵ B such that x = ab.

We define *f*: G → A ×B by (x) = (a, b) where x = ab ϵ G unique a ϵ A and unique b ϵ B.

Suppose $x_1 = a_1b_1$ & $x_1 = a_2b_2$ be two elements of where $a_1, a_2 \epsilon A$ & $b_1, b_2 \epsilon B$, such that $f(x_1) = f(x_2)$, i.e., $(a_1, b_1) = (a_2, b_2)$, *i.e.*, $a_1 = a_2, b_1 = b_2$, *i.e.*, $x_1 = x_2$. Consequently, *f* is injective.

Clearly, *f* is surjective.

Let $x_1 = a_1b_1$ & $x_2 = a_2b_2$ be two elements of G. Since G is internal direct product of A and B so $b_1a_2 = a_2b_1$.

Now, $f(x_1x_2) = f(a_1b_1a_2b_2) = f(a_1a_2b_1b_2) = (a_1a_2, b_1b_2) = (a_1,b_1)(a_2,b_2) = f(x_1)\ f((x_2)$. Hence, *f* is a homomorphism.

Consequently, *f* is an isomorphism and hence G ≅ A × B.

Exercise If a group, G is the internal direct product of its subgroups H and K; then H ≅ G/K and G/H ≅ K.

Solution: Since G is an internal direct product of its two subgroups H and K so every element g ϵ G can be expressed as g = hk for unique h ϵ H and unique k ϵ K.

We now define a mapping *f* : G → K by *f*(g) = k where g = hk ϵ G for unique h ϵ H and unique k ϵ K.

Let g_1, g_2 be two elements of G. Then $g_1 = h_1k_1$ and $g_2 = h_2k_2$ for unique h_1, $h_2 \in H$ and unique k_1, $k_2 \in K$. Now,

$$f(g_1, g_2) = f\big((h_1k_1)(h_2k_2)\big)$$

$$= f(h_1(k_1h_2)k_2)$$

$$= f(h_1h_2k_1k_2) \text{ [since } k_1h_2 = h_2k_1\text{]}$$

$$= k_1k_2$$

$$= f(g_1)\, f(g_2)$$

Hence, f is homomorphism.

Clearly, f is surjective. Thus, f is an epimorphism. Hence, by First isomorphism theorem we have $G/ker\ f \cong K$.

Now

$$kerf = \{g \in G : f(g) = e\}$$

$$= \{g(= hk) \in G :\ k = e\}$$

$$= \{g \in G : g = h \in H\} = H$$

Hence $G/H \cong K$. Similarly we can show that $G/K \cong H$.

For a group G, $Aut\ G$ denotes the set of all automorphisms of G.

Exercise Find $Aut\ (\mathbb{Z}_2 \times \mathbb{Z}_2)$

Solution: Now $\mathbb{Z}_2 \times \mathbb{Z}_2 = \{([0], [0]), ([0], [1], [0]), ([1], [1])\}$. Let e =([0], [0]), a = ([0], [1], b = ([1], [0]), c = ([1], [1]). Now 2a = 2b = 2c = ([0], [0]). Also, a + b = c, b + c = a & c + a = b. Therefore, $\mathbb{Z}_2$ x $\mathbb{Z}_2 \cong K_4$.

Let $f \in Aut(\mathbb{Z}_2$ x $\mathbb{Z}_2)$. Then for any x $\in \mathbb{Z}_2$ x $\mathbb{Z}_2$, o $\big(f(x)\big) =\ o\ (x)$. Then f {a, b, c} = {a, b, c}.

Hence, f is a permutation on the set of three elements a, b, c. So we find that $|Aut(\mathbb{Z}_2$ x $\mathbb{Z}_2)| \leq 6$.

Let o be any permutation on {a, b, c}. Then we define f: $\mathbb{Z}_2$ x $\mathbb{Z}_2 \to \mathbb{Z}_2$ x $\mathbb{Z}_2$ by $f\ (e) = e, f\ (a) = a(a), f(b) = o(b), f(c) = o(c)$. Then, early f is bijective. Moreover, one can prove (prove!)

that f is a homomorphism and hence f is an automorphism. Similarly, we can show that each permutation on {a, b, c} will produce an automorphism of the group $\mathbb{Z}_2 \times \mathbb{Z}_2$. Hence, $|Aut(\mathbb{Z}_2 \times \mathbb{Z}_2)| = 6$.

Exercises

1. If $A \cong B$ and $C \cong D$ then prove that $A \times C \cong B \times D$.
2. Prove that $\mathbb{Z}_8 \ncong \mathbb{Z}_4 \times \mathbb{Z}_2$.
3. Prove that S_3 is not isomorphic to a direct product of two non-trivial groups.
4. Prove that $\mathbb{Z}_4 \ncong \mathbb{Z}_2 \times \mathbb{Z}_2$.
5. Prove that $Aut(\mathbb{Z}_2 \times \mathbb{Z}_{2)} \cong S_3$.

Let G be a group such that $G = H_1 \times H_2 \times \ldots \times H$ Where Hi is a subgroup of G for $i = 1, 2, \ldots, n$. For each $i = 1, 2, \ldots, n$ let K_i, be a normal subgroup of G such that $K_i \subseteq H_i$. Let $K = K_1 \times K_2 \times \ldots\ldots \times K_n$. Show that $G/K \cong H_1/K_1 \times H_2/K_2 \times \ldots \times H_n/K_n$.

Summary

In this unit, we have mainly dealt with Direct product of groups and relevant theorems and applications.

Unit 12

Course Structure

1. Solvable Groups: definitions and characterization theorems
2. Nilpotent groups

Solvable and Nilpotent Groups

Introduction

Ascending central series and descending central series play an important role in the literature of group theory. In this chapter we study ascending central series and descending central series in a group. Moreover we study other series of subgroups of a group. Solvability plays a vital role for solving algebraic equations over a field. We prove in this section that $n \geq 5$, S_n is not solvable. Also, in this chapter we study properties of nilpotent group.

Definition: Let G be a group and $G = H_0 \supseteq H_1 \supseteq H_2 \supseteq \ldots\ldots\ldots\ldots \supseteq H_n = \{e\}$ be a chain of subgroup of G. Then this chain is said to be a subnormal series (subnormal chain) if each H_{i+1} is normal in H_i for i = 0, 1, ……..,n-1. This chain is called a normal series if each H_i is normal in G for i = 1, 2, ……, n.

For a subnormal series $G = H_0 \supseteq H_1 \supseteq H_2 \supseteq \ldots\ldots\ldots\ldots \supseteq H_n = \{e\}$ the groups H_i / H_{i+1} (for i = 0, 1,……., n-1) are called factors of te subnormal series.

Remark: Every group G contains a subnormal series as well as a normal series, viz., $G \supseteq \{e\}$. Also, from definition it follows that every normal series is a subnormal series.

Definition: Let G be a group and $G = H_0 \supseteq H_1 \supseteq H_2 \supseteq \ldots\ldots\ldots\ldots \supseteq H_n = \{e\}$ be a subnormal series. Then this subnormal series is called a solvable series if each factor H_i / H_{i+1} is commutative. A group G is said to be a solvable group if it has a solvable series.

Example: Consider the group S_3. Now A_3 is a normal subgroup of S_3 and {e} is a normal subgroup of A_3. Hence $S_3 \supseteq A_3 \{e\}$ is a subnormal series. Also each factor S_3/A_3 and $A_3/\{e\}$ is commutative. Hence, S_3 is a solvable group.

Remark: Every commutative group is solvable. But the converse is not true in general. For example, S_3 is a solvable group which is not commutative.

Example: Consider the group S_4. Now A_4 is a normal subgroup of S_4. We consider the subgroup

K = {e,(1 2)(3 4),(1 4)(2 3), (1 3) (2 4)} of A_4. Since K is the only one subgroup of A_4 of order 4, so K is normal in A_4. Again, T = {e,(1 2)(3 4)} is a subgroup of K such that [K : T] = 2. So T is normal in K. Hence $S_4 \supseteq A_4 \supseteq K \supseteq T \supseteq \{e\}$ is a subnormal scrics. Now, $|S_4 / A_4| = 2$, $|A_4 / K| = 3$, $|K/T| = 2$ and $|T/\{e\}| = 2$. Hence each factor of the subnormal series is commutative. Consequently, S_4 is a solvable group.

Example: We consider the group S_5. Now S_5 has only two subnormal series, viz., $S_5 \supseteq \{e\}$ and $S_5 \supseteq A_5 \supseteq \{e\}$. In each case, all the corresponding factors are not commutative, since S_5 and A_5 are not commutative. Hence, S_5 is not solvable.

Note: A simple group is solvable if and only if it is commutative.

Theorem: Every subgroup of a solvable group is solvable.

Proof: Let G be a solvable group and K be a subgroup of G. Since G is solvable so it has a solvable series

$G = H_0 \supseteq H_1 \supseteq H_2 \supseteq \ldots\ldots\ldots\ldots \supseteq H_n = \{e\}$.

Set $K_i = K \cap H_i$ for i = 0,1,2,.....,n. Now to show K is solvable, it is sufficient to show that the chain

$K = K_0 \supseteq K_1 \supseteq K_2 \supseteq \ldots\ldots\ldots\ldots \supseteq K_n = \{e\}$. is a solvable series for the group K.

Now each K_i is a subgroup of K. Since each H_{i+1} is normal in H_i(i =0,1,2,.....,n-1), we find that each K_{i+1} is normal in K_i(i =0,1,2,.....,n-1).

Also, by second isomorphism theorem, we have $K_i / K_{i+1} = K_i / K_i \cap H_{i+1} \cong K_i / H_{i+1}/H_{i+1}$.

Now $K_i / H_{i+1}/H_{i+1}$ is a subgroup of H_i/H_{i+1}. Since H_i/H_{i+1} is commutative for i= 0,1, 2,.....,n-1, so its subgroup $K_i H_{i+1}/ H_{i+1}$ is commutative for i= 0,1, 2,.....,n-1. Hence K_i / K_{i+1} is commutative. Therefore,

$K = K_0 \supseteq K_1 \supseteq K_2 \supseteq \ldots\ldots\ldots \supseteq K_n = \{e\}$ is a solvable series for K. Consequently, K is solvable. Hence the theorem. ■

Theorem: Every homomorphic image of a solvable group is solvable.

Proof: Let $G = H_0 \supseteq H_1 \supseteq H_2 \supseteq \ldots\ldots\ldots\ldots \supseteq H_n = \{e_G\}$ be a solvable series of a solvable group G and let $f : G \to \bar{G}$ be an epimorphism from G onto another group $\bar{G}$. We shall now show that $\bar{G}$is also solvable. Set $\overline{H_i} = f(H_i)$ for i= 0,1, 2,.....,n. Since H_{i+1} is normal in H_i and f is an epimorphism, so $\overline{H_{i+1}} = f(H_{i+1})$ is normal in $\overline{H_i} = f(H_i)$. Also, $H_i \supseteq H_{i+1}$ for i=0,1, 2,.....,n-1implies $\overline{H_i} = f(H_i) \supseteq f(H_{i+1}) = \overline{H_{i+1}}$. Hence, $\bar{G} = \overline{H_0} \supseteq \overline{H_1} \supseteq \overline{H_2} \supseteq \cdots \supseteq \overline{H_n} = \{e_{\bar{G}}\}$ is a solvable series.

Now, we define $g : H_i \to \overline{H_1}/\overline{H_{i+1}}$ by $g(h_i) = f(h_i)\ \overline{H_{i+1}}$, for all $h_i \in H_i$. Since, f is an epimorphism so it follows that g is an epimorphism of H_i onto $\overline{H_1}/\overline{H_{i+1}}$. Let $h_{i+1} \in H_{i+1} \subseteq H_i$. Then $g(h_{i+1}) = g(h_{i+1})\overline{H_{i+1}} = \overline{H_{i+1}}$. Hence $\overline{H_{i+1}} \subseteq$ kerg. So by Fundamental theorem, we have g induces an epimorphism from H_i / H_{i+1} onto $\overline{H_1}/\overline{H_{i+1}}$. Since, H_i / H_{i+1} is commutative, it follows that $\overline{H_1}/\overline{H_{i+1}}$ is also commutative. Consequently, the subnormal series $\bar{G} = \overline{H_0} \supseteq \overline{H_1} \supseteq \overline{H_2} \supseteq \cdots \supseteq \overline{H_n} = \{e_{\bar{G}}\}$ is a solvable series. Thus $\bar{G}$ is a solvable group. Hence the theorem. ■

Corollary: Let G be a group and H be a normal subgroup of G. If G is solvable then H and G /H are both solvable.

Proof. Since every subgroup of a solvable group is solvable, it follows that H is solvable. Also there is a natural epimorphism from G onto G/H. Thus, G/H is a homomorphic image of G. Hence, G/H is solvable. ■

Theorem: Let H be a normal subgroup of a group G. If both H and G/H are solvable then G is also solvable.

Proof. Now G/H is solvable. Let $G/H = T_0 \supseteq T_1 \supseteq T_2 \supseteq \ldots\ldots\ldots\ldots \supseteq T_n = \{H\}$ be a solvable series of G/H. Since each

$T_i (i = 0,1,2,\ldots\ldots,n)$ is a subgroup of G/H, so each T_i is of the form G_i/H where each G_i is a subgroup of G such that

$H \subseteq G_i$ for i =0,1,2,.....,n. So $G/H = G_0/H \supseteq G_i/H \supseteq G_2/H \supseteq \ldots\ldots\ldots \supseteq G_n/H = \{H\}$ is a solvable series for $G/H \rightarrow (1)$.

Since, series (1) is solvable, so G_{i+1}/H is normal in G_i/H for i= 0,1, 2,.....,n-1. Hence, each G_{i+1} is normal in G_i/H for i= 0,1, 2,.....,n-1.

Again, since H is solvable so let $H = H_0 \supseteq H_1 \supseteq H_2 \supseteq \ldots\ldots\ldots\ldots \supseteq H_s = \{e\}$ be a solvable series for H. So each H_{i+1} is normal in H_i for i= 0,1, 2,.....,s-1 and each factor H_i / H_{i+1} is commutative for i= 0,1, 2,.....,s-1.

Let $g_n \in G_n$. Then $g_n H \in G_n/H = \{H\}$. So $G_n H = H$. Hence $g_n \in H$. This implies $G_n \subseteq H$ and thus $G_n = H$.

Now, we consider a subnormal series $G = G_0 \supseteq G_1 \supseteq G_2 \supseteq \ldots\ldots\ldots\ldots \supseteq G_n = H = H_0 \supseteq H_1 \supseteq H_2 \supseteq \ldots\ldots\ldots\ldots \supseteq H_s = \{e\} \rightarrow (2)$.

We show that series (2) is a solvable series for the group G. To prove this it remains so show that each factor Gi/Gi_{+1} is commutative for i= 0,1, 2,.....,n-1.

Now , by third isomorphism theorem, we have $(G_i/H) / (G_{i+1}/H) \cong G_i/ G_{i+1}$

Since series (1) is a solvable series, so $(G_i/H) / (G_{i+1}/H)$ is commutative. Thus, G_i/ G_{i+1} is commutative. Hence (2) is a solvable series and consequently, G is a solvable group. Hence the theorem. ■

Corollary: A group G is solvable if and only if G/Z(G) is solvable.

Theorem: Prove that direct product of two solvable groups is solvable.

Proof. Let H and K be two solvable groups. Let G = HxK. It is easy to verify that $G/H \cong K$. Since K is solvable so G/H is solvable. Again, H is solvable. Hence, G = H x K is solvable. ■

Exercise: Show that the group D_4 is solvable.

Solution: Now D_4 is non commutative group of order $8 - 2^3$. Therefore, $|Z(D_4)| = 2$. We now consider the group $D_4 /Z(D_4)$. The order of the group $D_4 /Z(D_4)$ is 4. So $D_4 /Z(D_4)$is commutative and hence solvable. So by Corollary 4.12 we have D_4 is solvable. ■

Exercise: Show that every group of order p^2q is solvable where p and q are primes.

Solution: Let G be a group of order p^2q . Let p>q. Now G has sylow p – subgroup. Let n_p denote the number of Sylow p – subgroup of G. Then n_p| |G| and $n_p = k_p + 1$ for some non negative integer k. From this it follows that $n_p = 1$ or q. If $n_p = q$ then kp + 1=q, i.e, p |(q-1), a contradiction since p > q. Hence $n_p = 1$ and thus the Sylow p – subgroup is normal. Let H denote the Sylow p- subgroup of G. Then H is a normal subgroup of G. We now consider the subnormal series G⊇ H ⊇ {e}. Now since |G/H| = q so G/H is cyclic and hence commutative. Therefore, the factors of the subnormal series G⊇ H ⊇ {e}are commutative.

If q>p then we consider the Sylow q – subgroup of G. Let n_q denote the number of Sylow q – subgroup of G. Then n_q| |G| and $n_q = k_{1q} + 1$ for some non negative integer k_1. From this it follows that n_q =1 or p or p^2. If n_q = p then $k_{1q} + 1$ = p, i.e., q|(p-1), A contradiction since q>p. Hence n_q =1 or p or p^2. If n_q = 1 then the Sylow q – subgroup is normal. Let K denote the Sylow q – subgroup of G. Then K is a normal subgroup of G. We now consider the subnormal series G⊇ K ⊇ {e}. Now since |G/K| = p^2 so G/K is commutative. Therefore, the factors of the subnormal series G⊇ K ⊇ {e} are commutative. So in this case G is solvable. Again if $n_q = p^2$ then $k_1q + 1 = p^2$ implies q|(p^2 – 1), i.e., either q|P -1) or q|)p + 1). Since q>p so q | (p -1). Hence q|(p = 1). This is possible only when q = 3 and p = 2. Therefore, |G| = 12 and any group of order 12 is solvable. Hence any group of order p^2q is solvable. ■

Exercise: Show that every group of order p^2q^2 is solvable where p and q are primes.

Solution. Let G be a group of order p^2p^2. Without any loss of generality we assume that p > q. now G has Sylow p – subgroup. Let n_p denote the number of Sylow p – subgroup of G. Then n_p |ıGı and n_p = kp+1 for non negative integer k. From this follows that $n_p = 1$ or q or q^2. If $n_p = q$ then kp +1 = q, i.e, p|(q-1). a contradiction since p>q. Hence $n_p = 1$ or q^2. If n_p =1 then the Sylow p-subgroup is normal. Let H denote the Sylow p – subgroup of G. Then H is a normal subgroup of G. We now consider the subnormal series G ⊇ H ⊇ $\{e\}$are commutative. So in this case G is solvable.

Let $n_p = q^2$. Then kp +1 = q^2 implies p |(q^2 -1). This is possible only when p =3 and q =2. Therefore, |G| = 36 and any group of order 36 is solvable. Hence any group of order p^2q^2 solvable. ■

Definition: Let G be a group and a, b ε G. The commutator of this two elements a and b is the element $aba^{-1}b^{-1}$.

Set A = { $aba^{-1}b^{-1}$: a, b ϵ G}. Then the subgroup of G generated by A, denoted by G is called the derived subgroup or commutator subgroup of G. One can easily verify that G is the smallest subgroup of G containing A.

Theorem: A group G is commutative if and only if G = {e}.

Proof. First suppose that G is commutative. Let a, b ε G. Then $aba^{-1}b^{-1} = e$ and A = {e}. Hence, G = {e}.

Conversely, let G' = {e}. Then for any a, b ϵ G, we have ab $a^{-1}b^{-1}$ϵ A ⊆ G = $\{e\}, i.e, aba^{-1}b^{-1} = e, i.e, ab = ba$. Hence G is commutaative. ■

Theorem: The derived subgroup G of a group G is a normal subgroup of G and G/G' is commutative.

Proof. To show G' is a normal subgroup of G, let a ϵ G' and g ϵ G. Now $gag^{-1}a^{-1} \in A \subseteq G', i.,e., gag^{-1}a^{-1} \in G'$. A;sp, a ∈ G'.

Hence $gag^{-1} = (gag^{-1}a^{-1})a \in G'$, for all a $a \in G'$ and for all $g \in G'$. Threfore, G' is a normal subgroup of G.

To show G/G' is commutative, let aG', bG' ∈ G/G'. Then a, b ∈ G. Now, $aba^{-1}b^{-1} \epsilon$ G implies $(ab)(ba)^{-1} \in G' i.e.,$ (ab)G' = (ba)G', i.e., (aG')(bG') = (bG')(aG'). Consequently, G/G' is commutative. ■

Theorem: Let G' be the derived subgroup of a group G and H be a subgroup of G. Then H ⊇ G' if and only H is a normal subgroup of G and G/H is commutative.

Proof. Let G' ⊆ H. To show H is normal in G, let g ∈G and h ∈H.

Now, $ghg^{-1}h^{-1}$ ∈G' ⊆ H. Thus, H is normal in G.

Again, to show G/H is commutative, let aH, bH ∈ G/H. Then a, b ∈ G. Now, $aba^{-1}b^{-1} \in G' \subseteq$ H implies (ab)(b $a)^{-1} \in H, il.e, (ab)H = (ba)H, i.e., (aH)(bH) = (bH)(aH)$. Consequently, G/H is commutative.

Conversely, let H be normal in G and G/H is commutative. Le a, b ∈ G. Set A = { $aba^{-1}b^{-1} \in G'$}. The G' is the smallest subgroup of G containing A. Now, let aH,bH ∈ G/H. Since G/H is commutative, so (aH)(bH) = (bH)(aH), i.e.,$(ab(ba)^{-1} \in$ H, i.e., $aba^{-1}b^{-1} \in$ H, for all a, b ∈ G. Hence, A ⊆ H. So H is a subgroup of G containing A where $G'$is the smallest subgroup of G containing A. Consequently $G' \subseteq$ H ■

Exercise: Find S_3'

Solution: Now S_3 is a non commutative group. So $S_3' \neq \{e\}$. Also, A_3 is a normal subgroup of S_3 and S_3/A_3 is commutative. Hence $S_3' \subseteq A_3$. Moreover, S_3' is a normal subgroup of S_3. Hence, $S_3' = A_3$. ■

Definition: Let G be a group and G′be the derived subgroup of G. Set $G^{(1)} = G'$and define inductively $G^{(k+1)} = G^{(k)}'$, the derived subgroup of $G^{(k)}$, k > 0. For any positive integer k, $G^{(k)}$ is called the kth derived subgroup or kth commutator subgroup of G.

Theorem: Let G be a group. Then G is solvable if and only if there exist a positive integer k such that $G^{(k)} = \{e\}$.

Proof. First suppose that there is a positive integer k such that $G^{(k)} = \{e\}$. We consider the chain

$G \supseteq G^{(1)} \supseteq G^{(2)} \supseteq \ldots\ldots\ldots \supseteq G^{(k)} = \{e\} \rightarrow (1)$ where $G^{(m+1)}$ is the derived subgroup of $G^{(m)}$ and $G^{(m)} / G^{(m+1)}$ is commutative. So the above series (1) is a solvable series and hence G is a solvable group.

Conversely, let G be a solvable group. Then there is a solvable series $G = H_0 \supseteq H_1 \supseteq H_2 \supseteq \ldots\ldots\ldots\ldots \supseteq H_s = \{e\}$ for G.

We now prove by induction on k that $G^{(k)} \subseteq H_k$ for all k = 1, 2,....., n. Now H_1 is a normal subgroup of G and G/H_1 is commutative. So $G' \subseteq H_1$, i.e., $G^{(1)} \subseteq H_1$. Hence the result is true for k = 1.

Assume that the result is true for k = p where $1 \leq p < n$. Now, H_{p+1} is a normal subgroup of H_p and H_p/H_{p+1} is commutative. Hence $H_p' \subseteq H_{p+1}$. Again, by induction hypothesis, $G^{(p)} \subseteq H_p$. This implies, $G^{(p)}{}' \subseteq H_p' \subseteq H_{p+1}$. Thus the result is true for k = p + 1. Hence by method of induction we at once have $G^{(k)} \subseteq H_k$ for all k = 1, 2, ..., n. Since $H_n = \{e\}$, so $G^{(n)} \subseteq H_n = \{e\}$, i.e., $G^{(n)} = \{e\}$. Hence, the theorem. ■

Lemma Let S_n be the symmetric group of n symbols. If $n \geq 5$, then $S_n^{(k)}$ contains all 3-cycles for any $k \in N$.

Proof. Let $(a\ b\ c) \in S_n$ be any 3 cycle. Since $n \geq 5$, so there exist d & f such that a, b, c, d and f are distinct. Now, $(a\ b\ d), (a\ c\ f) \in S_n$. So $(a\ b\ d)\ (a\ c\ f)\ (a\ b\ d)^{-1}\ (a\ c\ f)^{-1} \in S_n'$ i.e. $(a\ b\ d)\ (a\ c\ f)\ (a\ d\ b)\ (a\ f\ c) \in S_n'$ i.e. $(a\ b\ c) \in S_n'$.

Now S_n' is a normal subgroup of $\in S_n$ and $\in S_n'$ contains a 3 cycle. Again, A_n is a normal subgroup S_n and S_n/A_n is commutative. Hence $S_n' \subseteq A_n$. Therefore, S_n' is a normal subgroup of A_n such that S_n' contains a 3 cycle. Then S_n' contains all the 3 cycles and thus $S_n' = A_n$. Now A'_n is a normal subgroup of A_n which contains all the three cycles. Hence $A'_n = A_n$. Thus $S_n^{(2)} = A'_n = A_n$. Proceeding as above we can show that $S_n^{(2)} = A_n$ for any $k \in N$. Hence $S_n^{(k)}$ contains all 3 cycles for any $k \in N$. ■

Theorem The symmetric group S_n on n symbols is not solvable for $n \geq 5$.

Proof. By Lemma 4.24, we see that $S_n^{(k)} = A_n \neq \{e\}$. Consequently, $S_n (n \geq 5)$ is not solvable. ■

Exercise Let G be the group of all n x n real matrices which are invertible where $n \geq 3$. Show that G is not solvable.

Solution. Let Eij be the nxn matrix whose (i, j) th-entry is 1 and all other entries are 0. Then

$$E_{ij}E_{rs} \begin{cases} E_{is} \ if\ j = r \\ \underline{0} \ if\ j \neq r \end{cases}$$

where $\underline{0}$ denote the n x n null matrix.

Let I denote the nxn identity matrix. Now for $i \neq j$, $(I + E_{ij}) \in G$ and also $(I + E_{ij})^{-1} = (I - E_{ij})$. Let T be the subgroup of G generated by $\{I + E_{ij} : i \neq j\}$. Since $n \geq 3$, so we can find an integer k such that $1 \leq i \neq k \neq j \leq n$.

Now $(I + E_{ik})(I + E_{kj})(I + E_{ik})^{-1}(I + E_{kj})^{-1}$

$= (I + E_{ik})(I + E_{kj})(I - E_{ik})(I - E_{kj})$

$= (I + E_{ik} + E_{kj} + E_{ij})(I - E_{ik} - E_{kj} - E_{ij})$

$= I + E_{ij}$

Therefore, $I + E_{ij} \in T'$, proving that $T \in T'$. Hence $T = T'$. Thus, T is not solvable and hence G is not solvable.

Unit 13

Course Structure

1. Group Actions: Definition
2. Applications

Group Actions

Introduction

Action of a group is a formal way of interpreting the manner in which the elements of the group correspond to transformations of some space in a way that preserves the structure of that space. Common examples of spaces that groups act on are sets, vector spaces, and topological spaces.

Definition Let (G, o) be a group and S be a nonempty set. A left action of G on S is function φ : G × S → S such that

(i) $(g_1, g_2)\ \varphi\ x = g_1\ \varphi(g_2 \varphi x)$ for all $g_1,\ g_2 \in G$ and for all $x \in S$

(ii) $e\varphi x = x$ for all $x \in S$, where e is the identity element in (G, o).

If there is a left action of a group (G, o) on a nonempty set S then S is called a G-set and in this case we say that G acts on S on the left.

Exercise Let G be a group and H be a subgroup of G. Let S = {aH : a ∈ G}. We define φ : G × S → S by $g \varphi aH = (ga)H$. Then under this operation S is a G-set.

Theorem Let G be a group and S be a G-set. Then the left action of G on S induces a homomorphism from G into A(S), where A(S) is the group of all permutations on S.

Theorem Let G be a group and H be a subgroup of G. Let S = {aH : a ∈ G}. Then there exists a homomorphism θ from G into A(S) such that $ker\theta \subseteq H$.

Proof: We define $\varphi : G \times S \to S$ by $g\varphi aH = (ga)H$. Under this operation S is a G-set. Hence this left action induces a homomorphism θ from G into A(S).

Now,

$$ker\theta = \{g \in G : \theta(g) = \text{identity element of } A(S)\}$$

$$= \{g \in G : \Delta_g = \text{identity mapping from } S \to S\}$$

$$= \{g \in G : \Delta_g(aH) = aH, for\ all\ aH \in S\}$$

$$= \{g \in G : gaH = aH, \text{for all } aH \in S\}$$

$$= \{g \in G : gH = H, \text{in particular we take } eH \text{ for } aH\}$$

$$= \{g \in G : g \in H\} = H \qquad \blacksquare$$

Theorem (Cayley's Theorem) Any group G is isomorphic to a subgroup of the permutation group A(S).

Proof: Let $\Delta_g : G \to G$ by $\Delta_g(a) = ga$ for all $a \in G$, where $g \in G$.

Let $a_1, a_2 \in G$ be such that $\Delta_g(a_1) = \Delta_g(a_2)$. Then $ga_1 = ga_2, i.e., a_1 = a_2$. Thus, Δ_g is injective.

Again, let $b \in G$. Then $g^{-1}b \in G$ such that $\Delta_g(g^{-1}b) = g(g^{-1}b) = b$. Thus Δ_g is surjective. Hence Δ_g is bijective and thus $\Delta_g \in A(G)$. Also, we have $(\Delta_g)^{-1} = \Delta_{g^{-1}}$ and thus $\Delta_{g_1 g_2} = \Delta_{g_1} \circ \Delta_{g_2}$.

Now we define $\theta : G \to A(S)$ by $\theta(g) = \Delta_g$, for all $g \in G$.

Then for any $g_1, g_2 \in G$, we have $\theta(g_1 g_2) = \Delta_{g_1 g_2} = \Delta_{g_1} \circ \Delta_{g_2} = \theta(g_1) \circ \theta(g_2)$. Hence θ is a homomorphism.

Again, let $g_1, g_2 \in G$ be such that $\theta(g_1) = \theta(g_2)$. Then this leads to $\Delta_{g_1} = \Delta_{g_2}$, i.e., $\Delta_{g_1}(a) = \Delta_{g_2}(a)$, for all $a \in G$, i.e., $g_1 a = g_2 a$, for all $a \in G$, i.e., $g_1 = g_2$. Hence θ is injective. Consequently, G is isomorphic to some subgroup θ(G) of A(G). This proves the theorem. ■

Theorem Let G be a group of order *2m,* where *m* is an odd integer. Show that G has a normal subgroup of order *m*.

Proof: Since G is a group of even order so there exists an element g ϵ G such that $o\ (g) = 2$.Now by Cayley's theorem G is isomorphic to a subgroup H of of A(G) where $\theta : G \to A(G)$ is given by $\theta(x) = \Delta_x, \Delta_x(a) = xa, \forall x, a \in G$. Now $\Delta_g(ga) = g(ga) = g^2a = a$. Hence Δ_g is a product of transpositions of the form $(a\ ga)$. Since $|G| = 2m$ so the number of transpositions appearing in the factorization of $\Delta_g$is m. Thus Δ_g is an odd permutation. Hence H contains an odd permutation. We now define $f : H \to \{1, -1\}$by

$$f(o) = \begin{Bmatrix} 1 & \text{if } o \text{ is an even permutation} \\ -1 & \text{if } o \text{ is an odd permutation} \end{Bmatrix}$$

Where {1, -1} is a group under multiplication.

It is easy to verify (verify!) that $f$is an epimorphism H onto {1, -1}. Hence H/ker$f \cong$ {1, -1}.

Thus, |H/ker f |= 2, i.e., |ker f|=$\frac{|H|}{2} = \frac{2m}{2} = m$ [since $H \cong G$ so$|H| = |G| = 2m$].

So *H* has a normal subgroup, viz., $r\ f$,of order *m* and consequently, *G* has a normal subgroup of order *m*.■

Definition Let *S* be a *G* – Set, where *G* is a group. On *S* we define a relation '~' by, for all $a, b \in S$, $a \sim b$ if and only if $ga = b$ for some $g \in G$.

Since S is a G – set, so for all $a \in S, ea = a$. Hence $a \sim a$ and thus '~' is reflexive. Let $a, b \in S$ be such that $a \sim b$. Then $ga = b$ for some $g \in G$. Now, $g^{-1}b = g^{-1}(ga) = (g^{-1}g)a = ea = a$. Thus $b \sim a$ and hence '~' is symmetric.

Finally, let $a, b, c \in S$ be such that $a \sim b$ and $\sim c$. Then there exists $g_1, g_2 \in G$. such that $g_1a = b$ *and* $g_2b = c$. Hence $c = g_2b = g_2(g_1a) = (g_2g_1)a$ where $g_2g_1 \in G$ such that $g_1a = b$ and $g_2b = c$. Hence $c = g_2b = g_2(g_1a) = (g_2g_1)a$ where $g_2g_1 \in G$. Thus $a \sim c$ and hence '~' is transitive. Consequently, '~' is an equivalence relation. Then *S* can be partitioned into disjoint equivalence classes. Each class is called *orbit* and the orbit of an element $a \in S$ is denoted by $[a]$ and $[a] = \{b \in S : ga = b \text{ for some } g \in G\} = Ga$.

Definition Let G be a group and S b a G-Set. Let $a \in S$ and $g \in G$. Then a is called $fixed\ by\ a\ g$ if $ga = a$. If $ga = a\ f$or all $g \in G$ then a is called $fixed\ by\ G$.

Let S be a G – Set. We consider the subset G_a of G where $G_a = \{g \in G: ga = a\} a \in S$.

Clearly G_a is nonempty since $ca = a$ implies $c \in G_a$. Let $g_1, g_2 \in G_a$. Then $g_1 a = a \; and \; g_2 a = a$. Now, $g_2 a = a$ implis $= g_2^{-1}(g_2 a) = g_2^{-1} a$, i. e, $(g_2^{-1} g_2) a = g_2^{-1} a$, i. e., a $= g_2^{-1} a$.

Now, $(g_1 g_2^{-1}) a = g_1 (g_2^{-1} a) = g_1 a = a$. Hence $g_1 g_2^{-1} \in G_a$. Consequently, G_a is a subgroup of G. This subgroup G_a called *stabilizer* of *a* or *isotropy group of a*.

Theorem Let *G* be a group and *S* b a *G*-Set. Then for all $a \in S$, $[G : G_a] = |[a]|$.

Theorem (Burnside Theorem) Let G be a finite group and S be a finite G-set. Then the number of orbits of S is $\frac{x}{y} \sum_{g \in G} F(g)$,where $F(g)$ is the number of elements of S fixed by g.

Proof: We consider the set $T = \{(g, a) \in G \times S : ga = a\}$. Also for each $g \in G$, let $C_g = \{a \in S : ga = a\}$. Then $|C_g| = F(g)$ and hence $|T| = \sum_{g \in G} F(g) = \sum_{a \in S} |G_a|$.

Let n be the number of orbits of S. Then $S = [a_1] \cup [a_2] \cup \ldots \ldots \cup [a_n] = \bigcup_{a \in A} [a]$ where $A = \{a_1, a_2, \ldots \ldots, a_n\}$ is a subset of S containing exactly one element from each orbit.

Therefore, $\sum_{g \in G} F(g) = |T| = \sum_{a \in S} |G_a| = \sum_{a \in a_1} |G_a| + \sum_{a \in [a_2]} |G_a| + \cdots + \sum_{a \in [a_n]} |G_a|$.

Now, $[G : G_a] = |[a]| = |[a_1]| = [G : G_{a_1}]$ where $a_1 \in [a]$.

Therefore, $\frac{|G|}{|G_a|} = \frac{|G|}{|G_{a_1}|}$, i.e., $|G| = G_{a_1}$.

Hence $\sum_{a \in G} F(g) = |G_{a_1}||[a_1]| + |G_{a_2}||[a_2]| + \cdots + |G_{a_n}||[a_n]|$

$$= |G_{a_1}|[G : G_{a_1}] + |G_{a_2}|[G : G_{a_2}] + \cdots + |G_{a_n}|[G : G_{a_n}]$$

$$= |G_{a_1}| \frac{|G|}{|G_{a_1}|} + |G_{a_2}| \frac{|G|}{|G_{a_2}|} + \cdots + |G_{a_n}| \frac{|G|}{|G_{a_n}|}$$

$$= |G| + |G| + \ldots + |G| \quad (n \; times)$$

$$= n|G|$$

Therefore $n = \frac{1}{|G|} \sum_{g \in G} F(g)$. Hence, the theorem. ■

Corollary Let G be a finite group and S be a finite G-set. Then $|S| = \sum_{a \in S}[G : G_{a_1}]$, where A is a finite subset of S containing exactly one element from each orbit.

Proof: Since S is finite, so the number of orbits of S is finite.

Therefore, $S = [a_1] \cup [a_2] \cup [a_2] \ldots \ldots \cup [a_n] = \cup_{a \in A}[a]$ where $A = \{a_1, a_2, \ldots, a_n\}$ is finite subset of S containing exactly one element from each orbit.

Hence, $|S| = \sum_{a \in S}|[a]|$. But $|[a]| = [G : G_{a_1}]$. Therefore $|S| = \sum_{a \in S}[G : G_{a_1}]$.

Example Let G be a group and H be a subgroup of G of index n such that $|G|$ does not divide $n!$. Then G contains a nontrivial normal subgroup.

Solution Let $S = \{aH : a \in G\}$. Since G is finite so is S. Also $|S| = [G : H] = n$. We define $\varphi : G \times S \to S$ by $g\varphi(aH) = (ga)H$ for all $g \in G$, for all $aH \in S$. Then under this operation S is a G-set and this action induces a homomorphism $\theta : G \to A(S)$ such that $kerf\ \theta \subseteq H$. Hence $G/\ker\theta$ is isomorphic to a subgroup of A(S). So $|G/\ker\theta|$ divides $n!$. But $|G|$ does not divides $n!$. Hence $|ker\ \theta| \neq 1$, proving that $\ker\theta$ is a nontrivial subgroup of G. Hence G contains a nontrivial normal subgroup.

Exercise Let G be a finite group of order of order pn where p is prime $p > n$. If H is subgroup of G of order p, then prove that H is a normal subgroup of G.

Exercise Let G be a finite group. Let H be a subgroup of G of index p, where p is the smallest prime dividing $|G|$. Show that H is a normal subgroup of G.

Exercise Let H be a subgroup of a group G. Prove that if H has a finite index n then there is a normal subgroup K of G with $K \subseteq H$ and $[G : K] \leq n!$.

References

1. M.K. Sen, D.S. Malik, J.M. Mordeson: Fundamentals of Abstract Algebra
2. I. N. Herstein: Topics in algebra
3. David S. Dummit, Richard M. Foote: Abstract Algebra
4. Joseph A. Gallian: Contemporary Abstract Algebra

Block III

Operations Research

Units 14 & 15

Course Structure

1.0 The Revised Simplex Method

1.1.1 Introduction

1.1.2 Steps for solving Revised Simplex Method

1.1.3 Worked Examples

1.2 Dual Simplex Method

1.2.1 Introduction:

1.2.2 Computational Procedure of Dual Simplex Method

1.2.3 Worked Examples:

1.3 Dantzig–Wolfe decomposition

1.3.1 Introduction

1.3.2 Problem Reformulation:

1.3.3 The Algorithm:

1.3.4 Implementation

1.4 Summary

1.1 The Revised Simplex Method

1.1.1 Introduction

While solving linear programming problem on a digital computer by regular simplex method, it requires storing the entire simplex table in the memory of the computer table, which may not be feasible for very large problem. But it is necessary to calculate each table during the each iteration. The revised simplex method which is a modification of the original method is more economical on the computer, as it computes and stores only the relevant information needed currently for testing and / or improving the current solution. i.e., it needs only

- The net evaluation row Δ_j to determine the non-basic variable that enters the basis.
- The pivot column
- The current basis variables and their values (X_B column) to determine the minimum positive ratio and then identify the basis variable to leave the basis.

The above information is directly obtained from the original equations by making use of the inverse of the current basis matrix at any iteration.

There are two standard forms for revised simplex method

- **Standard form-I** – In this form, it is assumed that an identity matrix is obtained after introducing slack variables only.
- **Standard form-II** – If artificial variables are needed for an identity matrix, then two- phase method of ordinary simplex method is used in a slightly different way to handle artificial variables.

1.1.2 Steps for solving Revised Simplex Method

Solve by Revised simplex method

Max $Z = 2x_1 + x_2$

Subject to

$3x_1 + 4x_2 \le 6$

$6x_1 + x_2 \le 3$

and $x_1, x_2 \ge 0$

SLPP

Max $Z = 2x_1 + x_2 + 0s_1 + 0s_2$ Subject to

$3x_1 + 4x_2 + s_1 = 6$

$6x_1 + x_2 + s_2 = 3$

and $x_1, x_2, s_1, s_2 \ge 0$.

Step 1 – Express the given problem in standard form – I

- Ensure all $b_i \ge 0$
- The objective function should be of maximization
- Use of non-negative slack variables to convert inequalities to equations The objective function is also treated as first constraint equation

$Z - 2x_1 - x_2 + 0s_1 + 0s_2 = 0$

$3x_1 + 4x_2 + s_1 + 0s_2 = 6$ -- (1)

$6x_1 + x_2 + 0s_1 + s_2 = 3$

and $x_1, x_2, s_1, s_2 \ge 0$

Step 2 – Construct the starting table in the revised simplex form Express (1) in the matrix form with suitable notation

$$\begin{matrix} \beta_0^{(1)} & & & \beta_1^{(1)} & \beta_2^{(1)} \\ e_1 & a_1^{(1)} & a_2^{(1)} & a_3^{(1)} & a_4^{(1)} \end{matrix}$$

$$\begin{bmatrix} 1 & -2 & -1 & 0 & 0 \\ 0 & 3 & 4 & 1 & 0 \\ 0 & 6 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} Z \\ x_1 \\ x_2 \\ s_1 \\ s_2 \end{bmatrix} = \begin{matrix} X_B \\ \begin{bmatrix} 0 \\ 6 \\ 3 \end{bmatrix} \end{matrix}$$

Column vector corresponding to Z is usually denoted b

matrix B_1, which is usually denoted as $B_1 = [\beta_0^{(1)}, \beta_1^{(1)}, \beta_2^{(1)} \dots \beta_n]$

Hence the column $\beta_0^{(1)}, \beta_1^{(1)}, \beta_2^{(1)}$ constitutes the basis matrix B_1 (whose inverse B_1^{-1} is also B_1)

Basic variables	B_1^{-1}			X_B	X_k	X_B / X_k
	e_1 (Z)	$\beta_1^{(1)}$	$\beta_2^{(1)}$			
Z	1	0	0	0		
s_1	0	1	0	6		
s_2	0	0	1	3		

$a_1^{(1)}$	$a_2^{(1)}$
-2	-1
3	4
6	1

Step 3 – Computation of Δ_j for $a_1^{(1)}$ and $a_2^{(1)}$

Δ_1 = first row of $B_1^{-1} * a_1^{(1)} = 1 * -2 + 0 * 3 + 0 * 6 = -2$

Δ_2 = first row of $B_1^{-1} * a_2^{(1)} = 1 * -1 + 0 * 4 + 0 * 1 = -1$

Step 4 – Apply the test of optimality

Both Δ_1 and Δ_2 are negative. So find the most negative value and determine the incoming vector. Therefore most negative value is $\Delta_1 = -2$. This indicates $a_1^{(1)}(x_1)$ is incoming vector.

Step 5 – Compute the column vector X_k

$X_k = B_1^{-1} * a_1^{(1)}$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} -2 \\ 3 \\ 6 \end{bmatrix} = \begin{bmatrix} -2 \\ 3 \\ 6 \end{bmatrix}$$

Step 6 – Determine the outgoing vector. We are not supposed to calculate for Z row.

Basic variables	B_1^{-1}			X_B	X_k	X_B / X_k
	e_1 (Z)	$\beta_1^{(1)}$	$\beta_2^{(1)}$			
Z	1	0	0	0	-2	-
s_1	0	1	0	6	3	2
s_2	0	0	1	3	6 ↑ incoming	1/2→outgoing

Step 7 – Determination of improved solution

Column e_1 will never change, x_1 is incoming so place it outside the rectangular boundary

	$\beta_1^{(1)}$	$\beta_2^{(1)}$	X_B	X_1
R_1	0	0	0	-2
R_2	1	0	6	3
R_3	0	1	3	6

Make the pivot element as 1 and the respective column elements to zero.

	$\beta_1^{(1)}$	$\beta_2^{(1)}$	X_B	X_1
R_1	0	1/3	1	0
R_2	1	-1/2	9/2	0
R_3	0	1/6	1/2	1

Construct the table to start with second iteration

Basic variables	B_1^{-1}			X_B	X_k	X_B / X_k
	e_1 (Z)	$\beta_1^{(1)}$	$\beta_2^{(1)}$			
Z	1	0	1/3	1		
s_1	0	1	-1/2	9/2		
x_1	0	0	1/6	1/2		

$a_4^{(1)}$	$a_2^{(1)}$
0	-1
0	4
1	1

$\Delta_4 = 1 * 0 + 0 * 0 + 1/3 * 1 = 1/3$

$\Delta_2 = 1 * -1 + 0 * 4 + 1/3 * 1 = -2/3$

Δ_2 is most negative. Therefore $a_2^{(1)}$ is incoming vector. Compute the column vector

$$\begin{bmatrix} 1 & 0 & 1/3 \\ 0 & 1 & -1/2 \\ 0 & 0 & 1/6 \end{bmatrix} * \begin{bmatrix} -1 \\ 4 \\ 1 \end{bmatrix} = \begin{bmatrix} -2/3 \\ 7/2 \\ 1/6 \end{bmatrix}$$

Determine the outgoing vector

Basic variables	B_1^{-1}			X_B	X_k	X_B / X_k
	e_1 (Z)	$\beta_1^{(1)}$	$\beta_2^{(1)}$			
Z	1	0	1/3	1	-2/3	-
s_1	0	1	-1/2	9/2	7/2	9/7→outgoing
x_1	0	0	1/6	1/2	1/6 ↑ incoming	3

Determination of improved solution

	$\beta_1^{(1)}$	$\beta_2^{(1)}$	X_B	X_2
R_1	0	1/3	1	-2/3
R_2	1	-1/2	9/2	7/2
R_3	0	1/6	1/2	1/6

	$\beta_1^{(1)}$	$\beta_2^{(1)}$	X_B	X_2
R_1	4/21	5/21	13/7	0
R_2	2/7	-1/7	9/7	1
R_3	-1/21	8/42	2/7	0

Basic variables	B_1^{-1}			X_B	X_k	X_B / X_k
	e_1 (Z)	$\beta_1^{(1)}$	$\beta_2^{(1)}$			
Z	1	4/21	5/21	13/7		
x_2	0	2/7	-1/7	9/7		
x_1	0	-1/21	8/42	2/7		

$a_4^{(1)}$	$a_3^{(1)}$
0	0
0	1
1	0

$\Delta_4 = 1 * 0 + 4/21 * 0 + 5/21 * 1 = 5/21$

$\Delta_3 = 1 * 0 + 4/21 * 1 + 5/21 * 0 = 4/21$

Δ_4 and Δ_3 are positive. Therefore optimal solution is Max Z = 13/7, x_1= 2/7, x_2 = 9/7

1.1.3 Worked Examples:

Example 1

Max $Z = x_1 + 2x_2$

Subject to

$x_1 + x_2 \leq 3$

$x_1 + 2x_2 \leq 5$ $3x_1 +$

$x_2 \leq 6$

and $x_1, x_2 \geq 0$

Solution SLPP

Max $Z = x_1 + 2x_2 + 0s_1 + 0s_2 + 0s_3$

Subject to

$x_1 + x_2 + s_1 = 3$

$x_1 + 2x_2 + s_2 = 5$ $3x_1 + x_2$

$+ s_3 = 6$

and $x_1, x_2, s_1, s_2, s_3 \geq 0$

Standard Form-I

$Z - x_1 - 2x_2 - 0s_1 - 0s_2 - 0s_3 = 0$

$x_1 + x_2 + s_1 + 0s_2 + 0s_3 = 3$

$x_1 + 2x_2 + 0s_1 + s_2 + 0s_3 = 5$

$3x_1 + x_2 + 0s_1 + 0s_2 + s_3 = 6$

and $x_1, x_2, s_1, s_2, s_3 \geq 0$

Matrix form

$$\begin{matrix} \beta_0^{(1)} & & & \beta_1^{(1)} & \beta_2^{(1)} & \beta_3^{(1)} \\ e_1 & a_1^{(1)} & a_2^{(1)} & a_3^{(1)} & a_4^{(1)} & a_5^{(1)} \end{matrix}$$

$$\begin{bmatrix} 1 & -1 & -2 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 2 & 0 & 1 & 0 \\ 0 & 3 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} Z \\ x_1 \\ x_2 \\ s_1 \\ s_2 \\ s_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 3 \\ 5 \\ 6 \end{bmatrix}$$

Revised simplex table

Basic variables	B_1^{-1}				X_B	X_k	X_B / X_k
	e_1 (Z)	$\beta_1^{(1)}$	$\beta_2^{(1)}$	$\beta_3^{(1)}$			
Z	1	0	0	0	0		
s_1	0	1	0	0	3		
s_2	0	0	1	0	5		
s_3	0	0	0	1	6		

Additional table

$a_1^{(1)}$	$a_2^{(1)}$
-1	-2
1	1
1	2
3	1

Computation of Δ_j for $a_1^{(1)}$ and $a_2^{(1)}$

Δ_1 = first row of $B_1^{-1} * a_1^{(1)} = 1 * -1 + 0 * 1 + 0 * 1 + 0 * 3 = -1$

Δ_2 = first row of $B_1^{-1} * a_2^{(1)} = 1 * -2 + 0 * 1 + 0 * 2 + 0 * 1 = -2$

$\Delta_2 = -2$ is most negative. So $a_2^{(1)}$ (x_2) is incoming vector.

Compute the column vector X_k

$X_k = B_1^{-1} * a_1^{(1)}$

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

Basic variables	B_1^{-1}				X_B	X_k	X_B / X_k
	e_1 (Z)	$\beta_1^{(1)}$	$\beta_2^{(1)}$	$\beta_3^{(1)}$			
Z	1	0	0	0	0	-2	-
s_1	0	1	0	0	3	1	3
s_2	0	0	1	0	5	2	5/2→
s_3	0	0	0	1	6	1	6
						↑	

Improved Solution

	$\beta_1(1)$	$\beta_2(1)$		XB	X_k
R_1	0	0	0	0	-2
R_2	1	0	0	3	1
R_3	0	1	0	5	2
R_4	0	0	1	6	1

	$\beta_1(1)$	$\beta_2(1)$	$\beta_3(1)$	X_B	X_k
R_1	0	1	0	5	0
R_2	1	-1/2	0	1/2	0
R_3	0	1/2	0	5/2	1
R_4	0	-1/2	1	7/2	0

Revised simplex table for II iteration

Basic variables	B_1^{-1}				X_B	X_k	X_B / X_k
	e_1 (Z)	$\beta_1^{(1)}$	$\beta_2^{(1)}$	$\beta_3^{(1)}$			
Z	1	0	1	0	5		
s_1	0	1	-1/2	0	1/2		
x_2	0	0	1/2	0	5/2		
s_3	0	0	-1/2	1	7/2		

$a_1^{(1)}$	$a_4^{(1)}$
-1	0
1	0
1	1
3	0

$\Delta_1 = 1 * -1 + 0 * 1 + 1 * 1 + 0 * 3 = 0$

$\Delta_4 = 1 * 0 + 0 * 0 + 1 * 1 + 0 * 0 = 1$

Δ_1 and Δ_4 are positive. Therefore optimal solution is Max Z = 5, $x_1 = 0$, $x_2 = 5/2$.

1.2 Computational Procedure of Dual Simplex Method

1.2.1 Introduction:

Any LPP for which it is possible to find infeasible but better than optimal initial basic solution can be solved by using dual simplex method. Such a situation can be recognized by first expressing the constraints in '≤' form and the objective function in the maximization form. After adding slack variables, if any right hand side element is negative and the optimality condition is satisfied then the problem can be solved by dual simplex method.

Negative element on the right hand side suggests that the corresponding slack variable is negative. This means that the problem starts with optimal but infeasible basic solution and we proceed towards its feasibility.

The dual simplex method is similar to the standard simplex method except that in the latter the starting initial basic solution is feasible but not optimum while in the former it is infeasible but optimum or better than optimum. The dual simplex method works towards feasibility while simplex method works towards optimality.

1.2.2 Computational Procedure of Dual Simplex Method

The iterative procedure is as follows

Step 1 - First convert the minimization LPP into maximization form, if it is given in the minimization form.

Step 2 - Convert the '≥' type inequalities of given LPP, if any, into those of '≤' type by multiplying the corresponding constraints by -1.

Step 3 – Introduce slack variables in the constraints of the given problem and obtain an initial basic solution.

Step 4 – Test the nature of Δ_j in the starting table

- If all Δ_j and X_B are non-negative, then an optimum basic feasible solution has been attained.
- If all Δ_j are non-negative and at least one basic variable X_B is negative, then go to step 5.
- If at least Δ_j one is negative, the method is not appropriate.

Step 5 – Select the most negative X_B. The corresponding basis vector then leaves the basis set B. Let X_r be the most negative basic variable.

Step 6 – Test the nature of X_r

- If all X_r are non-negative, then there does not exist any feasible solution to the given problem.
- If at least one X_r is negative, then compute Max (Δ_j / X_r) and determine the least negative for incoming vector.

Step 7 – Test the new iterated dual simplex table for optimality.

Repeat the entire procedure until either an optimum feasible solution has been attained in a finite number of steps.

1.2.3 Worked Examples:

Example 1

Minimize $Z = 2x_1 + x_2$

Subject to

$3x_1 + x_2 \geq 3$

$4x_1 + 3x_2 \geq 6$

$x_1 + 2x_2 \geq 3$

and $x_1 \geq 0, x_2 \geq 0$

Solution

Step 1 – Rewrite the given problem in the form

Maximize $Z' = -2x_1 - x_2$

Subject to

$-3x_1 - x_2 \leq -3$

$-4x_1 - 3x_2 \leq -6$

$-x_1 - 2x_2 \leq -3$

$x_1, x_2 \geq 0$

Step 2 – Adding slack variables to each constraint

Maximize $Z' = -2x_1 - x_2$

Subject to

$-3x_1 - x_2 + s_1 = -3$

$-4x_1 - 3x_2 + s_2 = -6$

$-x_1 - 2x_2 + s_3 = -3$

$x_1, x_2, s_1, s_2, s_3 \geq 0$

Step 3 – Construct the simplex table

	$C_j \rightarrow$		-2	-1	0	0	0	
Basic variables	C_B	X_B	X_1	X_2	S_1	S_2	S_3	
s_1	0	-3	-3	-1	1	0	0	
s_2	0	-6	-4	-3	0	1	0	$\rightarrow$ outgoing
s_3	0	-3	-1	-2	0	0	1	
	$Z' = 0$		2	↑ 1	0	0	0	$\leftarrow \Delta_j$

Step 4 – To find the leaving vector

Min (-3, -6, -3) = -6. Hence s_2 is outgoing vector

Step 5 – To find the incoming vector

Max (Δ_1 / x_{21}, Δ_2 / x_{22}) = (2/-4, 1/-3) = -1/3. So x_2 is incoming vector

Step 6 –The key element is -3. Proceed to next iteration

	$C_j \rightarrow$		-2	-1	0	0	0	
Basic variables	C_B	X_B	X_1	X_2	S_1	S_2	S_3	
s_1	0	-1	-5/3	0	1	-1/3	0	→ outgoing
x_2	-1	2	4/3	1	0	-1/3	0	
s_3	0	1	5/3	0	0	-2/3	1	
	$Z' = -2$		↑ 2/3	0	0	1/3	0	$\leftarrow\Delta_j$

Step 7 – To find the leaving vector

Min (-1, 2, 1) = -1. Hence s_1 is outgoing vector

Step 8 – To find the incoming vector

Max (Δ_1 / x_{11}, Δ_4 / x_{14}) = (-2/5, -1) = -2/5. So x_1 is incoming vector

Step 9 –The key element is -5/3. Proceed to next iteration

	$C_j \rightarrow$		-2	-1	0	0	0	
Basic variables	C_B	X_B	X_1	X_2	S_1	S_2	S_3	
x_1	-2	3/5	1	0	-3/5	1/5	0	
x_2	-1	6/5	0	1	4/5	-3/5	0	
s_3	0	0	0	0	1	-1	1	
	$Z' = -12/5$		0	0	2/5	1/5	0	$\leftarrow\Delta_j$

Step 10 – $\Delta_j \geq 0$ and $X_B \geq 0$, therefore the optimal solution is Max $Z' = -12/5$, $Z = 12/5$, and $x_1=3/5$, $x_2 = 6/5$

Example 2

Minimize $Z = 3x_1 + x_2$

Subject to

$x_1 + x_2 \geq 1$

$2x_1 + 3x_2 \geq 2$

and $x_1 \geq 0$, $x_2 \geq 0$

Solution

Maximize $Z' = -3x_1 - x_2$

Subject to

$-x_1 - x_2 \leq -1$

$-2x_1 - 3x_2 \leq -2$

$x_1, x_2 \geq 0$

SLPP

Maximize $Z' = -3x_1 - x_2$

Subject to

$-x_1 - x_2 + s_1 = -1$

$-2x_1 - 3x_2 + s_2 = -2$

$x_1, x_2, s_1, s_2 \geq 0$

	$C_j \rightarrow$		-3	-1	0	0	
Basic variables	C_B	X_B	X_1	X_2	S_1	S_2	
s_1	0	-1	-1	-1	1	0	
s_2	0	-2	-2	-3	0	1	$\rightarrow$
				↑			
	$Z' = 0$		3	1	0	0	$\leftarrow\Delta_j$
s_1	0	-1/3	-1/3	0	1	-1/3	$\rightarrow$
x_2	-1	2/3	2/3	1	0	-1/3	
						↑	
	$Z' = -2/3$		7/3	0	0	1/3	$\leftarrow\Delta_j$
s_2	0	1	1	0	-3	1	
x_2	-1	1	1	1	-1	0	
	$Z' = -1$		2	0	1	0	$\leftarrow\Delta_j$

$\Delta_j \geq 0$ and $X_B \geq 0$, therefore the optimal solution is Max $Z' = -1$, $Z = 1$, and $x_1 = 0$, $x_2 = 1$.

1.3 Dantzig–Wolfe decomposition

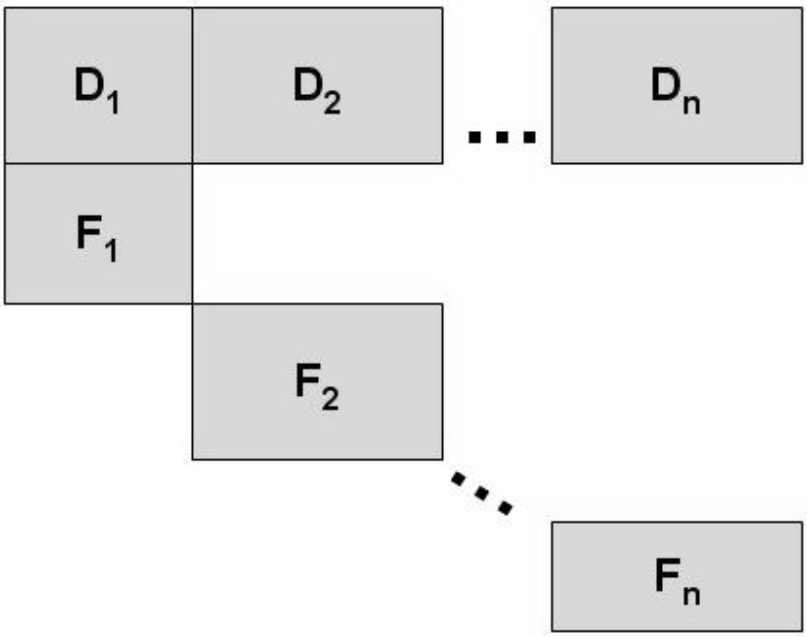
1.3.1 Introduction:

Dantzig–Wolfe decomposition is an algorithm for solving linear programming problems with special structure. It was originally developed by George Dantzig and Philip Wolfe and initially published in 1960. Many texts on linear programming have sections dedicated to discussing this decomposition algorithm.

Dantzig–Wolfe decomposition relies on delayed column generation for improving the tractability of large-scale linear programs. For most linear programs solved via the revised simplex algorithm, at each step, most columns (variables) are not in the basis. In such a scheme, a master problem containing at least the currently active columns (the basis) uses a sub-problem or sub-problems to generate columns for entry into the basis such that their inclusion improves the objective function.

1.3.1 Required Form:

In order to use Dantzig–Wolfe decomposition, the constraint matrix of the linear program must have a specific form. A set of constraints must be identified as "connecting", "coupling", or "complicating" constraints wherein many of the variables contained in the constraints have non-zero coefficients. The remaining constraints need to be grouped into independent sub-matrices such that if a variable has a non-zero coefficient within one sub-matrix, it will not have a non-zero coefficient in another sub-matrix. This description is visualized below:



The *D* matrix represents the coupling constraints and each F_i represents the independent sub-matrices. Note that it is possible to run the algorithm when there is only one *F* sub-matrix.

1.3.2 Problem Reformulation:

After identifying the required form, the original problem is reformulated into a master program and *n* subprograms. This reformulation relies on the fact that a non-empty, bounded convex polyhedron can be represented as a convex combination of its extreme points (or, in the case of an unbounded polyhedron, a convex combination of its extreme points and a weighted combination of its extreme rays).

Each column in the new master program represents a solution to one of the sub-problems. The master program enforces that the coupling constraints are satisfied given the set of sub-problem solutions that are currently available. The master program then requests additional solutions from the sub-problem such that the overall objective to the original linear program is improved.

1.3.3 The Algorithm:

While there are several variations regarding implementation, the Dantzig–Wolfe decomposition algorithm can be briefly described as follows:

1. Starting with a feasible solution to the reduced master program, formulate new objective functions for each sub-problem such that the sub-problems will offer solutions that improve the current objective of the master program.
2. Sub-problems are re-solved given their new objective functions. An optimal value for each sub-problem is offered to the master program.
3. The master program incorporates one or all of the new columns generated by the solutions to the sub-problems based on those columns' respective ability to improve the original problem's objective.

4. Master program performs x iterations of the simplex algorithm, where x is the number of columns incorporated.
5. If objective is improved, go to step 1. Else, continue.
6. The master program cannot be further improved by any new columns from the sub-problems, thus return.

1.3.4 Implementation

There are examples of the implementation of Dantzig–Wolfe decomposition available in the AMPL and GAMS mathematical modeling languages. There is a general, parallel implementation available that leverages the open source GNU Linear Programming Kit.

The algorithm can be implemented such that the sub-problems are solved in parallel, since their solutions are completely independent. When this is the case, there are options for the master program as to how the columns should be integrated into the master. The master may wait until each sub-problem has completed and then incorporate all columns that improve the objective or it may choose a smaller subset of those columns. Another option is that the master may take only the first available column and then stop and restart all of the sub-problems with new objectives based upon the incorporation of the newest column.

Another design choice for implementation involves columns that exit the basis at each iteration of the algorithm. Those columns may be retained, immediately discarded, or discarded via some policy after future iterations (for example, remove all non-basic columns every10 iterations).

A recent (2001) computational evaluation of Dantzig-Wolfe in general and Dantzig-Wolfe and parallel computation is the PhD thesis by J. R. Tebboth.

1.4 **Summary:**

The revised simplex method is another efficient method developed by G.B. Dantzig, for solving L.P.P. it is efficient in the sense that at each iteration we need not re-compute values of all the variable in the simplex table while moving from one iteration to next in such of an improved solution of an L.P.P. In the dual simplex method we always attempt to retain optimality while bringing the primal back to feasibility.

Units 16 & 17

Course Structure

2.1 Introduction
2.2. Types of Integer Programming
2.3. Integer Linear Programming
2.4 Cutting Plane Methods
2.5 Branch-and-Bound Method
2.6 Summary

2.1 Introduction

In all the previous lectures in linear programming discussed so far, the design variables considered are supposed to take any real value. However in practical problems like minimization of labor needed in a project, it makes little sense in assigning a value like 5.6 to the number of laborers. In situations like this, one natural idea for obtaining an integer solution is to ignore the integer constraints and use any of the techniques previously discussed and then round-off the solution to the nearest integer value. However, there are several fundamental problems in using this approach:

1. The rounded-off solutions may not be feasible.
2. The objective function value given by the rounded-off solutions (even if some are feasible) may not be the optimal one.
3. Even if some of the rounded-off solutions are optimal, checking all the rounded-off solutions is computationally expensive (2^n possible round-off values to be considered for an n variable problem)

2.2. Types of Integer Programming

When all the variables in an optimization problem are restricted to take only integer values, it is called an *all – integer programming problem*. When the variables are restricted to take only discrete values, the problem is called a *discrete programming problem*. When only some variable values are restricted to take integer or discrete, it is called *mixed integer or discrete programming problem*. When the variables are constrained to take values of either zero or 1, then the problem is called *zero – one programming problem*.

2.3. Integer Linear Programming

Integer Linear Programming (ILP) is an extension of linear programming, with an additional restriction that the variables should be integer valued. The standard form of an ILP is of the form,

$$\begin{aligned} \max \quad & c^T X \\ \textit{subject to} \quad & AX \le b \\ & X \ge 0 \end{aligned}$$

X must be integer valued

The associated linear program dropping the integer restrictions is called *linear relaxation LR.* Thus, LR is less constrained than ILP. If the objective function coefficients are integer, then for minimization, the optimal objective for ILP is greater than or equal to the rounded-off value of the optimal objective for LR. For maximization, the optimal objective for ILP is less than or equal to the rounded-off value of the optimal objective for LR.

For a minimization ILP, the optimal objective value for LR is less than or equal to the optimal objective for ILP and for a maximization ILP, the optimal objective value for LR is greater than or equal to that of ILP. If LR is infeasible, then ILP is also infeasible. Also, if LR is optimized by integer variables, then that solution is feasible and optimal for IP.

A most popular method used for solving all-integer and mixed-integer linear programming problems is the cutting plane method by Gomory (Gomory, 1957).

Def. (Integer Programming):A linear programming problem in which some or all of the variables in the optimal solution are restricted to assume non-negative integer values is called an Integer Programming Problem (IPP) or Integer Linear Programming.

Def. (Importance of Integer programming problem):
In LPP the values for the variables are real in the optimal solution. However in certain problems this assumption is unrealistic. For example if a problem has a solution of 81/2 cars to be produced in a manufacturing company is meaningless. These types of problems require integer values for the decision variables. Therefore IPP is necessary to round off the fractional values.

Def. (Pure IPP): In a linear programming problem, if all the variables in the optimal solution are restricted to assume non-negative integer values, then it is called the pure (all) IPP.

Def. (Mixed IPP): In a linear programming problem, if only some of the variables in the optimal solution are restricted to assume non-negative integer values, while the remaining variables are free to take any non-negative values, then it is called A Mixed IPP.

Def. (Zero-one problem): If all the variables in the optimum solution are allowed to take values either 0 or 1 as in 'do' or 'not to do' type decisions, then the problem is called Zero-one problem or standard discrete programming problem.

Methods of IPP:

a) Cutting Plane Method
b) Branch and Bound Method

2.4 Cutting Plane Methods

The first exact techniques for solving integer programming problems were *cutting plane techniques*.

General idea: Solve *linear programming relaxation*, i.e., the given problem without integrality requirements. If the optimal solution is integer, we are done. Otherwise, introduce a *cutting plane*, i.e., an additional constraint that (1) cuts off (i.e., makes infeasible) the present optimal solution, while (2) not cutting off any feasible integer point.

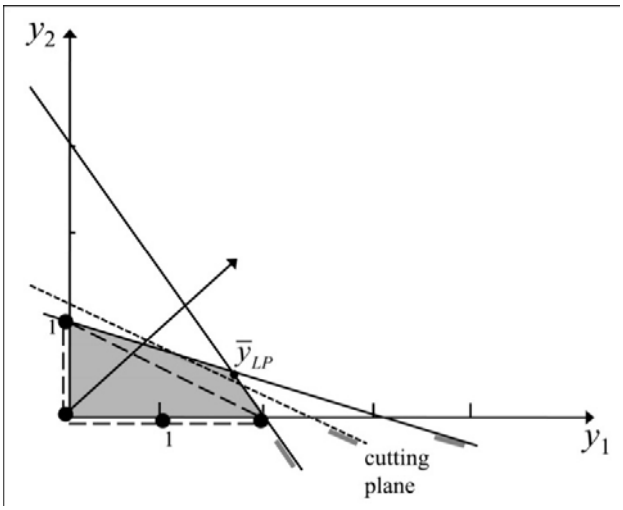
Example: Consider the all-integer programming problem:

P: Max $z = y_1 + y_2$

s.t. $3y_1 + 2y_2 \le 6$

$y_1 + 3y_2 \le 3$

$y_1, \quad y_2 \ge 0$ and integer.



The shaded area shows the feasible set of the linear programming relaxation, and $\bar{\mathbf{y}}_{LP} = (12/7, 3/7)$ is the optimal solution of the linear programming relaxation.

The triangle shown by the broken lines connecting (0, 0), (2, 0), and (0, 1) is the *convex hull* of the feasible set. The dotted line is the cutting plane $5y_1 + 10y_2 \le 12$. It is indeed a cutting plane, as the present optimal solution $\bar{\mathbf{y}}_{LP}$ is cut off as $90/7 \not\le 12$, and since all four feasible integer points satisfy the condition & are thus not cut off.

Computation performance of cutting planes has been disappointing.

Example: We use a simple *Dantzig cut*, which does not require any knowledge beyond the solution typically provided by a solver. Other, more efficient, cutting planes work on the same principle.

Given an all-integer linear programming problem. Include all slack and excess variables, so that all constraints are equations. Let there be n nonnegative variables (including the slack and excess variables) and m structural equation constraints, and assume that the present optimal solution of the linear programming relaxation has at least one non-integer component.

Separate the variables into two disjoint sets B and N, where B includes all variables that are presently positive, while N includes all variables that are presently zero. If the solution is nondegenerate, the set Bwill include exactly m variables, and the set N exactly $(n–m)$ variables. In case of primal degeneracy, the set N will include more than $(n – m)$ variables, in which case we define N as any $(n – m)$ variables presently at zero.

A *Dantzig cut* requires the sum of all variables in the set N to be at least "1." Validity: (1) Since all variables in the set N equal zero, the cutting plane invalidates the present solution. (2) Any feasible solution to the original integer problem will need to have at least one variable in N assume a positive value, which, since this is an all-integer optimization problem, must be at least one. Hence, the sum of all the variables that are presently zero, must be at least one.

Add the cut to the problem & re-solve the problem (preferably with a *warm start*). Stop, if the new solution is integer; else, repeat. The the process, the z-value cannot increase (decrease) for max (min) problems. In each step, the feasible set shrinks. Unfortunately, for Dantzig cuts, this is not necessarily finite.

Example: Consider the integer programming problem:

Max $z = 3y_1 + 2y_2$

s.t. $3y_1 + 7y_2 \leq 22$

$5y_1 + 3y_2 \leq 17$

$y_1 \geq 2$

$y_1, \quad y_2 \geq 0$ and integer.

Adding slack variables S_1 and S_2 and an excess variable E_3, we obtain the following formulation with $n = 5$ variables and $m = 3$ structural constraints:

Max $z = 3y_1 + 2y_2$

s.t. $3y_1 + 7y_2 + S_1 = 22$

$5y_1 + 3y_2 + S_2 = 17$

$y_1 - E_3 = 2$

$y_1, \quad y_2, \quad S_1, S_2, E_3 \geq 0$ and integer.

The optimal solution is $\overline{y}_1 = 2.0385$, $\overline{y}_2 = 2.2692$, $\overline{S}_1 = \overline{S}_2 = 0$, and $\overline{E}_3 = 0.0385$ with $\overline{z} = 10.65385$. Here, $N = \{S_1, S_2\}$, so that the Dantzig cut is $S_1 + S_2 \geq 1$(or $8y_1 + 10y_2 \leq 38$). Subtracting a new excess variable E_4 from the left-hand side of this cut, we obtain $S_1 + S_2 - E_4 = 1$. Adding this cut to the problem and solving it again, we obtain the new solution $\overline{y}_1 = 2.1538, \overline{y}_2 = 2.0769$, $\overline{S}_1 = 1$, $\overline{S}_2 = \overline{E}_4 = 0$ and $\overline{E}_3 = 0.1538$ with an objective value $\overline{z} = 10.61539$. Clearly, another cut is required. The sequence of cutting planes generated in the process is shown in the table below.

Optimal solution	**Cutting plane**
$\overline{y}_1 = 2.0385, \overline{y}_2 = 2.2692, \overline{S}_1 = 0, \overline{S}_2 = 0, \overline{E}_3 = 0.0385$ **with** $\overline{z} = 10.65385$ **(optimal solution of the LP relaxation).**	$S_1 + S_2 \geq 1$ **or** $S_1 + S_2 - E_4 = 1$
$\overline{y}_1 = 2.1538, \overline{y}_2 = 2.0769, \overline{S}_1 = 1, \overline{S}_2 = 0, \overline{E}_3 = 0.1538, \overline{E}_4 = 0$ **with** $\overline{z} = 10.61539.$	$S_2 + E_4 \geq 1$ **or** $S_2 + E_4 - E_5 = 1$
$\overline{y}_1 = 2.2692, \overline{y}_2 = 1.8846, \overline{S}_1 = 2, \overline{S}_2 = 0, \overline{E}_3 = 0.2692, \overline{E}_4 = 1,$ $\overline{E}_5 = 0$ **with** $\overline{z} = 10.5769.$	$S_2 + E_5 \geq 1$ **or** $S_2 + E_5 - E_6 = 1$
$\overline{y}_1 = 2.3846, \overline{y}_2 = 1.6923, \overline{S}_1 = 3, \overline{S}_2 = 0, \overline{E}_3 = 0.3846, \overline{E}_4 = 2,$ $\overline{E}_5 = 1, \overline{E}_6 = 0$ **with** $\overline{z} = 10.53846.$	$S_2 + E_6 \geq 1$ **or** $S_2 + E_6 - E_7 = 1$
$\overline{y}_1 = 2.5, \overline{y}_2 = 1.5, \overline{S}_1 = 4, \overline{S}_2 = 0, \overline{E}_3 = 0.5, \overline{E}_4 = 3, \overline{E}_5 = 2, \overline{E}_6 = 1,$ $\overline{E}_7 = 0$ **with** $\overline{z} = 10.5.$	$S_2 + E_7 \geq 1$ **or** $S_2 + E_7 - E_8 = 1$
$\overline{y}_1 = 2.6154, \overline{y}_2 = 1.3077, \overline{S}_1 = 5, \overline{S}_2 = 0, \overline{E}_3 = 0.6154, \overline{E}_4 = 4,$ $\overline{E}_5 = 3, \overline{E}_6 = 2, \overline{E}_7 = 1, \overline{E}_8 = 0$ **with** $\overline{z} = 10.46154.$	$S_2 + E_8 \geq 1$ **or** $S_2 + E_8 - E_9 = 1$
$\overline{y}_1 = 2.7308, \overline{y}_2 = 1.1154, \overline{S}_1 = 6, \overline{S}_2 = 0, \overline{E}_3 = 0.7308, \overline{E}_4 = 5,$ $\overline{E}_5 = 4, \overline{E}_6 = 3, \overline{E}_7 = 2, \overline{E}_8 = 1, \overline{E}_9 = 0$ **with** $\overline{z} = 10.42308.$	$S_2 + E_9 \geq 1$ **or** $S_2 + E_9 - E_{10} = 1$
⋮	⋮
$\overline{y}_1 = 2, \overline{y}_2 = 2, \overline{S}_1 = 2, \overline{S}_2 = 1, \overline{E}_3 = 0$ **with** $\overline{z} = 10$ **(optimal all-integer solution)**	

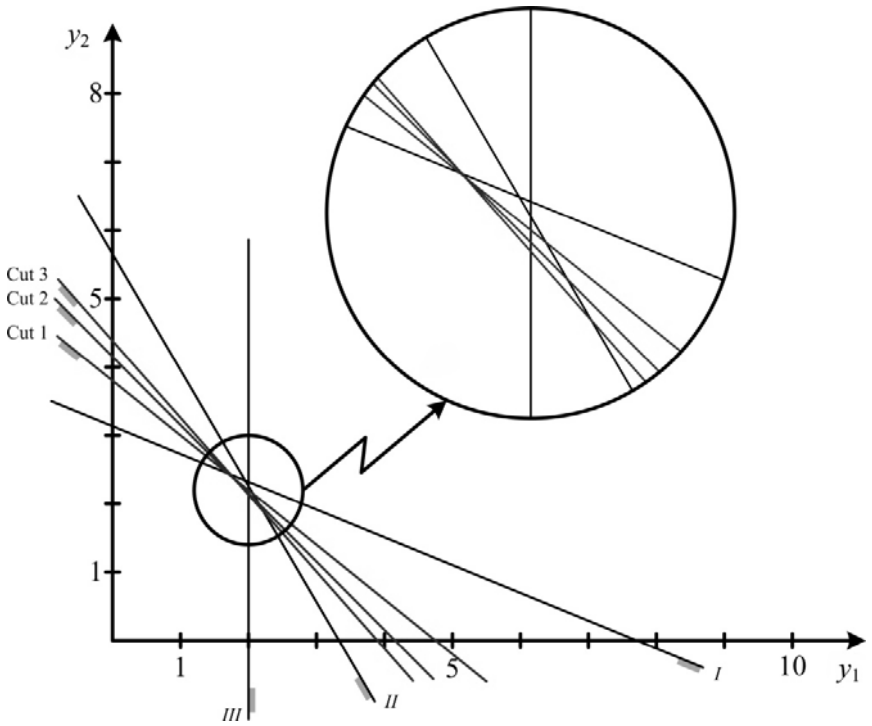
Even for this toy example, a large number of cuts are need to solve the problem. This is true in general. "Deep cuts" are much better, but cannot compete with "branch & bound methods" discussed next. One could use the objective function to derive a cut: Since all variables must be integer, the value of the objective function z =

$3y_1 + 2y_2$ must also be integer. The LP relaxation of the problem has an objective value of $\overline{z} = 10.65385$, hence $\overline{z} \leq 10$ must hold.

A cutting plane is then $3y_1 + 2y_2 + S_4 = 10$. Solving the problem with this added constraint results in the solution $\overline{y}_1 = 3.3333$, $\overline{y}_2 = 0$, $\overline{S}_1 = 12$, $\overline{S}_2 = 0.3333$, $\overline{E}_3 = 1.3333$ and $\overline{S}_4 = 0$ with $\overline{z}$ = 10. (Since the objective value has not changed, we presently encounter dual degeneracy).

The next cutting plane is then $y_2 + S_4 \geq 1$, (or, alternatively, $3y_1 + y_2 + S_5 = 9$). Adding the cut results in an optimal solution $\overline{y}_1 = 2.6667$, $\overline{y}_2 = 1$, $\overline{S}_1 = 7$, $\overline{S}_2 = 0.6667$, $\overline{E}_3 = 0.6667$, $\overline{S}_4 = \overline{S}_5 = 0$, with $\overline{z} = 10$.

The next cut is $S_4 + S_5 \geq 1$, or, rewritten in terms of the original variables and the new slack variable S_6, it is written as $6y_1 + 3y_2 + S_6 = 18$. The optimal solution is then $\overline{y}_1 = \overline{y}_2 = 2$, $\overline{S}_1 = 2$, $\overline{S}_2 = 1$, $\overline{E}_3 = \overline{S}_4 = 0$, $\overline{S}_5 = 1$, $\overline{S}_6 = 0$, with $\overline{z} = 10$. This solution is an integer optimum. The cuts are shown in the figure below.



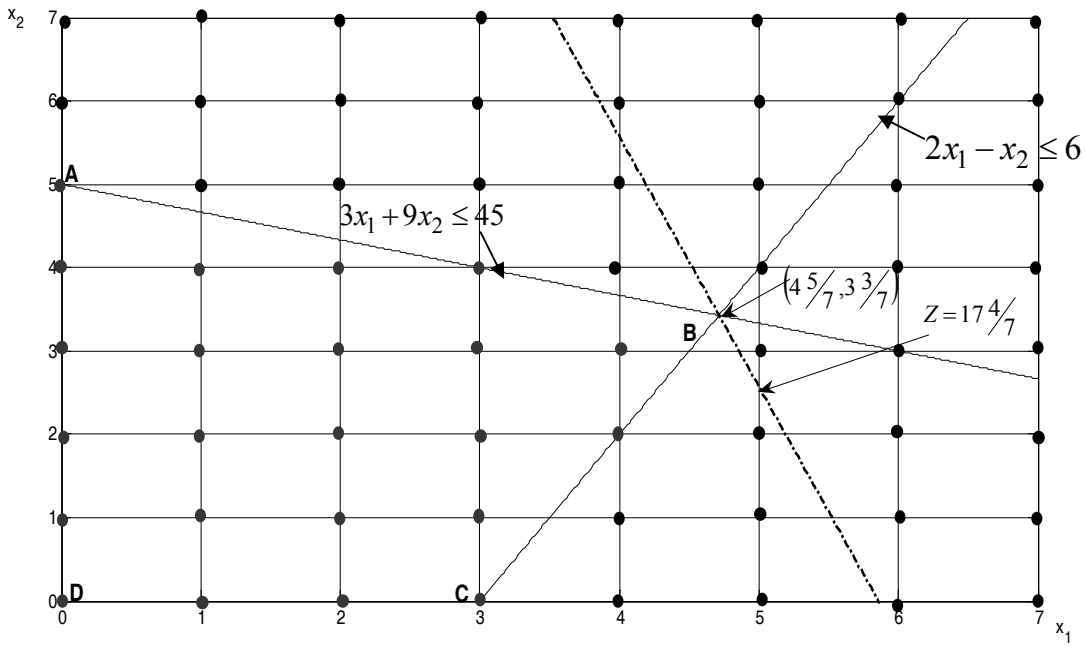
2.5 Gomory's Cutting Plane Method for All – Integer Programming

Consider the following optimization problem.

$$\begin{aligned} \text{Maximize} \quad & Z = 3x_1 + x_2 \\ \text{subject to} \quad & 2x_1 - x_2 \leq 6 \\ & 3x_1 + 9x_2 \leq 45 \\ & x_1, x_2 \geq 0 \end{aligned}$$

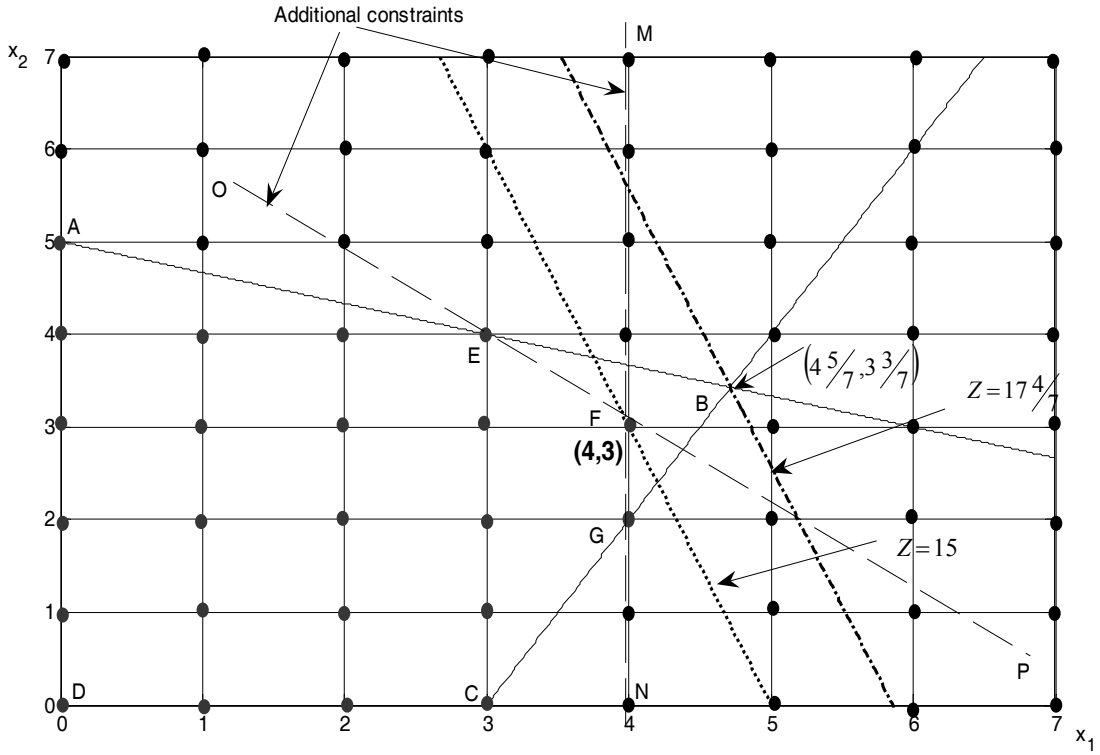
$x_1$and x_2 are integers

The graphical solution for the *linear relaxation* of this problem is shown below.



It can be seen that the solution is $x_1 = 4\frac{5}{7},\ x_2 = 3\frac{3}{7}$ and the optimal value of $Z = 17\frac{4}{7}$. The feasible solutions accounting the integer constraints are shown by red dots. These points are called integer lattice points. The original feasible region is reduced to a new feasible region by including some additional constraints such that an extreme point of the new feasible region becomes an optimal solution after accounting for the integer constraints.

The graphical solution for the example previously discussed taking x_1 and x_2 as integers are shown below. Two additional constraints (MN and OP) are included so that the original feasible region ABCD is reduced to a new feasible region AEFGCD. Thus the solution for this ILP is $x_1 = 4,\ x_2 = 3$ and the optimal value is $Z = 15$.



Gomary proposed a systematic method to develop these additional constraints known as *Gomory constraints*.

Generation of Gomory Constraints:

Let the final tableau of an LP problem consist of *n* basic variables (original variables) and *m* non basic variables (slack variables) as shown in the table below. The basic variables are represented as x_i(i=1,2,...,n) and the non-basic variables are represented as y_j (j=1,2,...,m).

Table 1

Basis	Z	Variables											b_r
		x_1	x_2	...	x_i	...	x_n	y_1	y_2	...	y_j	...	y_m	
Z	1	0	0		0		0	c_1	c_2		c_j		c_m	b
x_1	0	1	0		0		0	c_{11}	c_{12}		c_{1j}		c_{1m}	b_1
x_2	0	0	1		0		0	c_{21}	c_{22}		c_{2j}		c_{2m}	b_2
⋮														
x_i	0	0	0		1		0	c_{31}	c_{32}		c_{3j}		c_{3m}	b_i
⋮														
x_n	0	0	0		0		1	c_{41}	c_{42}		c_{4j}		c_{4m}	b_n

Choose any basic variable x_i with the highest fractional value. If there is a tie between two basic variables, arbitrarily choose any of them as x_i. Then from the i^{th} equation of table,

$$x_i = b_i - \sum_{j=1}^{m} c_{ij} y_j \qquad \text{.....(1)}$$

Express both b_i and c_{ij} as an integer value plus a fractional part.

$$b_i = \bar{b}_i + \beta_i \qquad \text{.....(2)}$$

$$c_{ij} = \bar{c}_{ij} + \alpha_{ij} \qquad \text{.....(3)}$$

Where $\bar{b}_i$, $\bar{c}_{ij}$ denote the integer part and β_i, α_{ij} denote the fractional part. β_i will be a strictly positive fraction $(0 < \beta_i < 1)$ and α_{ij} is a non-negative fraction $(0 \le \alpha_{ij} < 1)$.Substituting equations (2) and (3) in (1), equation (1) can be written as

$$\beta_i - \sum_{j=1}^{m} \alpha_{ij} y_j = x_i - \bar{b}_i - \sum_{j=1}^{m} \bar{c}_{ij} y_j \qquad \text{.....(4)}$$

For all the variables x_i and y_j to be integers, the right hand side of equation (4) should be an integer.

$$\beta_i - \sum_{j=1}^{m} \alpha_{ij} y_j = \text{integer} \qquad \text{.....(5)}$$

Since α_{ij} are non-negative integers and y_j are non-negative integers, the term $\sum_{j=1}^{m} \alpha_{ij} y_j$ will always be a non-negative number. Thus we have,

$$\left(\beta_i - \sum_{j=1}^{m} \alpha_{ij} y_j \right) \le \beta_i < 1 \qquad \text{.....(6)}$$

Hence the constraint can be expressed as

$$\beta_i - \sum_{j=1}^{m} \alpha_{ij} y_j \le 0 \qquad \text{.....(7)}$$

By introducing a slack variable s_i (which should also be an integer), the Gomory constraint can be written as

$$s_i - \sum_{j=1}^{m} \alpha_{ij} y_j = -\beta_i \qquad \text{.....(8)}$$

General procedure for solving ILP:

Solve the given problem as an ordinary LP problem neglecting the integer constraints. If the optimum values of the variables are integers itself, then there is nothing more to be done.

If any of the basic variables has fractional values, introduce the Gomory constraints as discussed in the previous section. Insert a new row with the coefficients of this constraint, to the final tableau of the ordinary LP problem (Table 1). Solve this by applying the dual simplex method. Since the value of $y_j = 0$ in Table 1, the Gomory constraint equation becomes $s_i = -\beta_i$ which is a negative value and thus infeasible. Dual simplex method is used to obtain a new optimal solution that satisfies the Gomory constraint.

Check whether the new solution is all-integer or not. If all values are not integers, then a new Gomory constraint is developed from the new simplex tableau and the dual simplex method is applied again.
This process is continued until an optimal integer solution is obtained or it shows that the problem has no feasible integer solution.

Thus, the fundamental idea behind cutting planes is to add constraints to a linear program until the optimal basic feasible solution takes on integer values. Gomory cuts have the property that they can be generated for any integer program, but has the disadvantage that the number of constraints generated can be enormous depending upon the number of variables.

2.5 Branch-and-Bound Method

The widely used search method is the Branch and Bound Technique. It starts with the continuous optimum, but systematically partitions the solution space into sub problems that eliminate parts that contain no feasible integer solution. It was originally developed by A.H.Land and A.G.Doig.

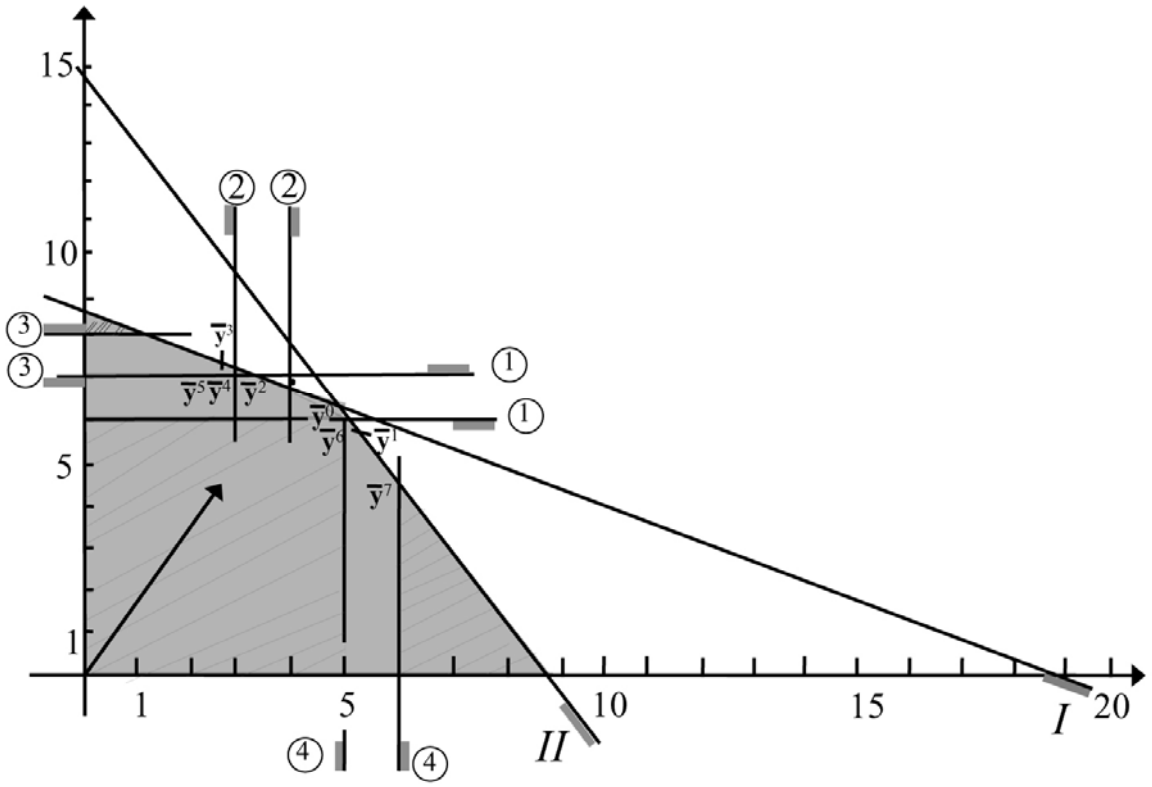
These methods are very flexible & are applicable to *AILP&MILPs*.
Idea: Starting with the LP relaxation, subdivide the problem into subproblems, whose union includes all integer solutions that are not worse than the best known integer solution.
For instance, if presently $y_3 = 5.2$, we subdivide the problem (the "parent") by adding the constraint $y_3 \leq 5$ & $y_3 \geq 6$, respectively (thus creating "children").
Example:

Max $z = 5y_1 + 9y_2$

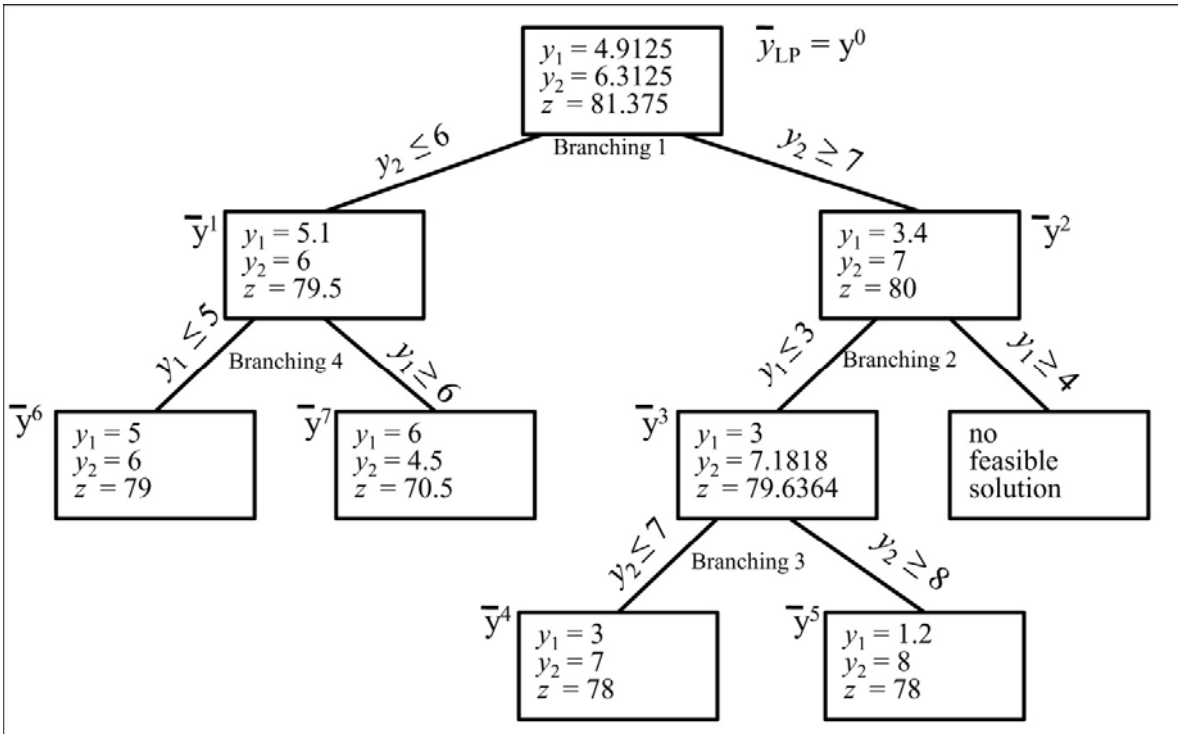
s.t. $5y_1 + 11y_2 \leq 94$ Constraint *I*

$10y_1 + 6y_2 \leq 87$ Constraint *II*

$y_1, y_2 \geq 0$ and integer.



Solution Tree



Note: Each node of the solution tree represents one linear program.

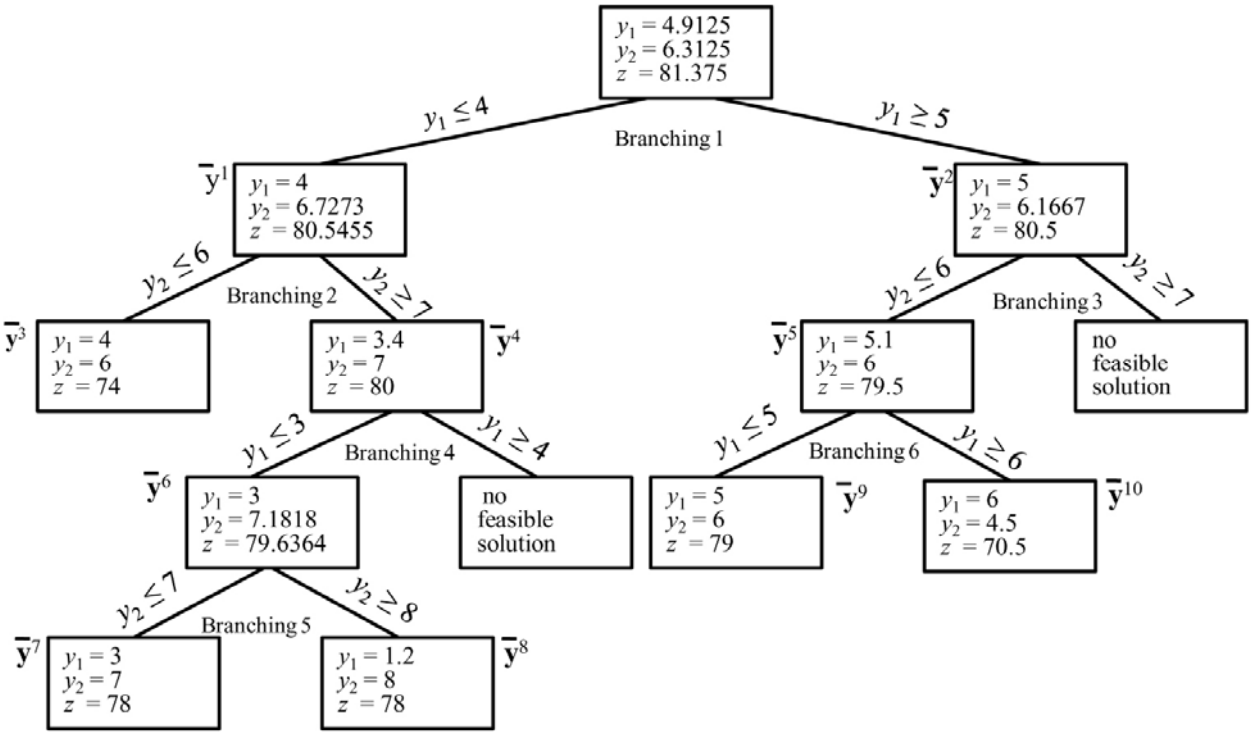
The constraints at a node are all original constraints plus all additional constraints between the root of the tree & the node in question.

As we move down the tree, the problems get to be more constrained & thus their objective values *cannot* improve.

At any stage, the problem to be worked on is the "best" active node (whose z-value is the present upper bound (for max problems, lower bound for min problems)), the best known integer solution is the present lower bound (for max problems, upper bound for min problems).

Different modes: fully automatic (specify integrality conditions & let the optimizer do its thing), fully manual (manually construct the solution tree & solve the LPs graphically), or semi-automatic (manually construct the solution tree, whose LP solutions are obtained by some LP solver).

Same example:



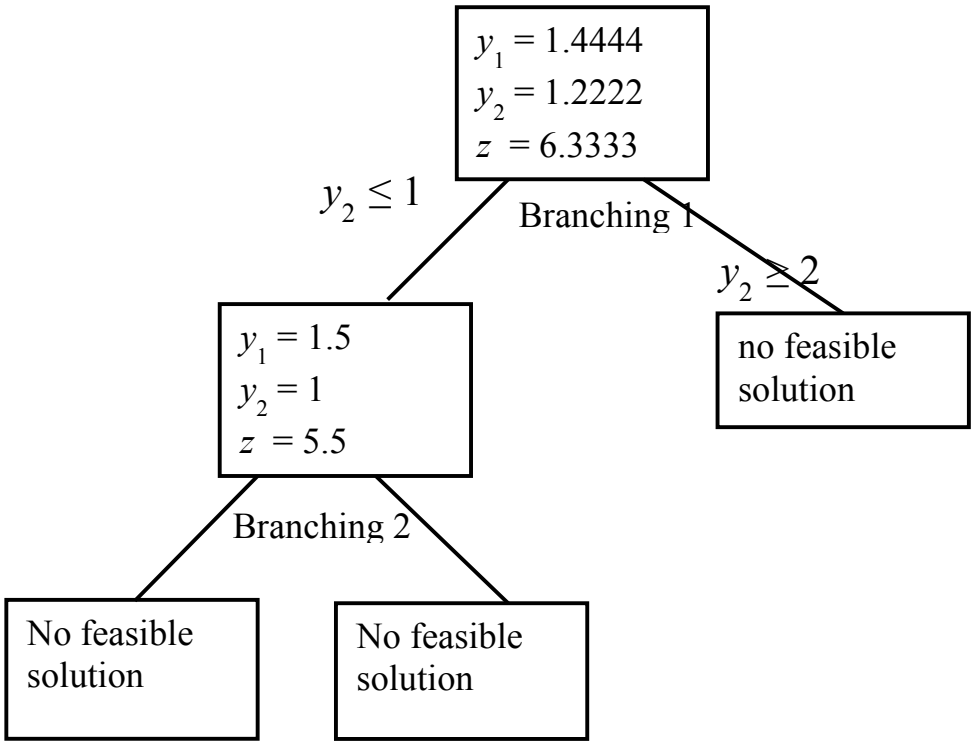
If the IP problem has no feasible solution:

P: Max $z = y_1 + 4y_2$

s.t. $28y_1 + 7y_2 \le 49$

$30y_1 - 6y_2 \ge 36$

$y_1, \quad y_2 \ge 0$ and integer.



1.6 Summary

In this chapter, an extension of linear programming, referred to as integer linear programming, was introduced where few or all variables must be an integer. If all variables of a problem are integers, then such problems are referred to as all-integer linear linear programming problems. Most integer programming applications involve 0-1 variables.

The number of applications of integer linear programming continues to grow rapidly due to the availability of integer linear programming software packages.

The study of integer linear programming is helpful when fractional values for the variables are not permitted and rounding off their values may not provide an optimal integer solution; integer LP programming facilitates developing mathematical models with variables assume either value 0 or 1. Capital budgeting , fixed cost , plant location ,etc. , are few examples where 0-1 integer programming techniques are extensively used to find an optimal solution.

Unit 18

Course Structure

3.1 INTRODUCTION

In this Chapter let us look to a problem, where we have to determine the order or sequence in which the jobs are to be processed through machines so as to minimize the total processing time. Here the total effectiveness, which may be the time or cost that is to be minimized is the function of the order of sequence. Such type of problem is known as **SEQUENCING PROBLEM.**

In case there are three or four jobs are to be processed on two machines, it may be done by trial and error method to decide the optimal sequence (*i.e.* by method of enumeration). In the method of enumeration for each sequence, we calculate the total time or cost and search for that sequence, which consumes the minimum time and select that sequence. This is possible when we have small number of jobs and machines. But if the number of jobs and machines increases, then the problem becomes complicated. It cannot be done by method of enumeration. Consider a problem, where we have '*n*' machines and '*m*' jobs then we have $(n!)^m$ theoretically possible sequences. For example, we take n = 5 and m = 5, then we have $(5!)^5$ sequences *i.e.* which works out to 25, 000,000,000 possible sequences. It is time consuming to find all the sequences and select optima among all the sequences. Hence we have to go for easier method of finding the optimal sequence. Let us discuss the method that is used to find the optimal sequence. Before we go for the method of solution, we shall define the sequencing problem and types of sequencing problem. The student has to remember that the sequencing problem is basically a **minimization problem or minimization model.**

3.2. The Mathematical Aspects of Job Sequencing and Processing Problems

A general sequencing problem may be defined as follows:

Let there be '*n*' jobs (J_1, J_2, J_3 ……J_n) which are to be processed on '*m*' machines (*A*, *B*, *C*,……), where the order of processing on machines *i.e.* for example, *ABC* means first on machine A, second on machine *B* and third on machine *C* or *CBA* means first on machine *C*, second on machine *B* and third on machine *A* etc. and the processing time of jobs on machines (actual or expected) is known to us, then our job is to find the optimal sequence of processing jobs that minimizes the total processing time or cost. Hence our job is to find that sequence out of $(n!)^m$ sequences, which minimizes the total elapsed time (*i.e.*. time taken to process all the jobs). The usual notations used in this problem are:

A_i = Time taken by i th job on machine A where i = I, 2,3…n. Similarly we can interpret for machine B and C *i.e.* B_i and C_i etc.

T = Total elapsed time which includes the idle time of machines if any and set up time and transfer time.

3.1.1. Assumptions Made in Sequencing Problems

Principal assumptions made for convenience in solving the sequencing problems are as follows:

(a) The processing times A_i and B_i etc. are exactly known to us and they are independent of order of processing the job on the machine. That is whether job is done first on the machine, last on the machine, the time taken to process the job will not vary it remains constant.

(b) The time taken by the job from one machine to other after processing on the previous machine is negligible. (Or we assume that the processing time given also includes the transfer time and setup time).

(c) Each job once started on the machine, we should not stop the processing in the middle. It is to be processed completely before loading the next job.

(d) The job starts on the machine as soon as the job and the machine both become idle (vacant).This is written as **job is next to the machine and the machine is next to the job.** (This is exactly the meaning of transfer time is negligible).

(e) No machine may process more than one job simultaneously. (This means to say that the job once started on a machine, it should be done until completion of the processing on that machine).

(f) The cost of keeping the semi-finished job in inventory when next machine on which the job is to be processed is busy is assumed to be same for all jobs or it is assumed that it is too small and is negligible. That is in process inventory cost is negligible.

(g) While processing, no job is given priority *i.e.* the order of completion of jobs has no significance. The processing times are independent of sequence of jobs.

(h) There is only one machine of each type.

3.2.2. Applicability

The sequencing problem is very much common in Job workshops and Batch production shops. There will be number of jobs which are to be processed on a series of machine in a specified order depending on the physical changes required on the job. We can find the same situation in computer center where number of problems waiting for a solution. We can also see the same situation when number of critical patients waiting for treatment in a clinic and in Xerox centers, where number of jobs is in queue, which are to be processed on the Xerox machines. Like this we may find number of situations in real world.

3.2.3. Types of Sequencing Problems :

There are various types of sequencing problems arise in real world. All sequencing problems cannot be solved. Though mathematicians and Operations Research scholars are working hard on the problem satisfactory method of solving problem is available for few cases only. The problems, which can be solved, are:

(a) ‘n’ jobs are to be processed on two machines say machine A and machine B in the order AB. This means that the job is to be processed first on machine A and then on machine B.
(b) ‘n’ jobs are to be processed on three machines A,B and C in the order ABC i.e. first on machine A, second on machine B and third on machine C.
(c) ‘n’ jobs are to be processed on ‘m’ machines in the given order
(d) Two jobs are to be processed on ‘m’ machines in the given order.

3.3 SOLUTIONS FOR SEQUENCING PROBLEMS :

Now let us take above mentioned types problems and discuss the solution methods.

3.3.1. 'N' Jobs and Two Machines

If the problem given has two machines and two or three jobs, then it can be solved by using the Gantt chart. But if the numbers of jobs are more, then this method becomes less practical. (For understanding about the Gantt chart, the students are advised to refer to a book on Production and Operations Management (chapter on Scheduling).

3.3.1.1. Analytical Method

A method has been developed by **Johnson and Bellman** for simple problems to determine a sequence of jobs, which minimizes the total elapsed time. The method:

1. '*n*' jobs are to be processed on two machines *A* and *B* in the order *AB* (*i.e.* each job is to be processed first on *A* and then on *B*) and passing is not allowed. That is whichever job is processed first on machine *A* is to be first processed on machine *B* also, whichever job is processed second on machine *A* is to be processed second on machine *B* also and so on. That means each job will first go to machine *A* get processed and then go to machine *B* and get processed. ***This rule is known as no passing rule.***

2. Johnson and Bellman method concentrates on minimizing the idle time of machines. Johnson and Bellman have proved that optimal sequence of '*n*' jobs which are to be processed on two machines *A* and *B* in the order *AB* necessarily involves the same ordering of jobs on each machine. This result also holds for three machines but does not necessarily hold for more than three machines. Thus total elapsed time is minimum when the sequence of jobs is same for both the machines.

3. Let the number of jobs be 1,2,3,……*n*

The processing time of jobs on machine *A* be $A_1, A_2, A_3 \ldots\ldots\ldots A_n$

The processing time of jobs on machine *B* be $B_1, B_2, B_3 \ldots\ldots\ldots B_n$.

Jobs	*Machining time in hours*		
	Machine A	*Machine B*	*(Order of processing is AB)*
1	A_1	B_1	
2	A_2	B_2	
3	A_3	B_3	
….. ….	…	…	
I	A_I	B_I	
…. ….	…	…	
S	A_S	B_S	
…. ….	…	…	
…. ….	…	…	
T	A_T	B_T	
…. ….	…	…	
… ….	…	…	
N	A_N	B_N	

4. Johnson and Bellman algorithm for optimal sequence states that identify the smallest element in the given matrix**.** *If the smallest element falls under column 1 i.e under machine I then do that job first*. As the job after processing on machine 1 goes to machine 2, it reduces the idle time or waiting time of machine 2. *If the smallest element falls under column 2 i.e under machine 2 then do that job last*. This reduces the idle time of machine 1. i.e. if r th job is having smallest element in first column, then do the r th job first. If s th job has the smallest element, which falls under second column, then do the s th job last. Hence the basis for Johnson and

Bellman method is to keep the idle time of machines as low as possible. Continue the above process until all the jobs are over.

5. If there are 'n' jobs, first write 'n' number of rectangles as shown. Whenever the smallest elements falls in column 1 then enter the job number in first rectangle. If it falls in second column, then write the job number in the last rectangle. Once the job number is entered, the second rectangle will become first rectangle and last but one rectangle will be the last rectangle.

6. Now calculate the total elapsed time as discussed. Write the table as shown. Let us assume that the first job starts at Zero th time. Then add the processing time of job (first in the optimal sequence) and write in out column under machine 1. This is the time when the first job in the optimal sequence leaves machine 1 and enters the machine 2. Now add processing time of job on machine 2. This is the time by which the processing of the job on two machines over. Next consider the job, which is in second place in optimal sequence. This job enters the machine 1 as soon the machine becomes vacant, i.e first job leaves to second machine. Hence enter the time in out column for first job under machine 1 as the starting time of job two on machine 1. Continue until all the jobs are over. Be careful to see that whether the machines are vacant before loading. Total elapsed time may be worked out by drawing Gantt chart for the optimal sequence.

Problem 3.1.

There are five jobs, which are to be processed on two machines *A* and *B* in the order *AB*. The processing times in hours for the jobs are given below. Find the optimal sequence and total elapsed time. *(Students has to remember in sequencing problems if optimal sequence is asked, it is the duty of* the student to find the total elapsed time also).

Jobs	1	2	3	4	5
Machine A (Time in hours)	2	6	4	8	10
Machine B (Time in hours)	3	1	5	9	7

The smallest element is 1 it falls under machine *B* hence do this job last i.e. in 5th position. Cancel job 2 from the matrix. The next smallest element is 2, it falls under machine A hence do this job first, i.e in the first position. Cancel the job two from matrix. Then the next smallest element is 3 and it falls under machine *B*. Hence do this job in fourth position. Cancel the job one from the matrix. Proceed like this until all jobs are over.

Total elapsed time :

1	3	4	5	2

Optimal Sequence	*MACHINE-A*		*MACHINE-B*		*MACHINE IDLE JOB IDLE*		*Remarks*
	IN	*OUT*	*IN*	*OUT*	*A*	*B*	
1	0	2	2	5		2	
3	2	6	6	11		1	As the Machine B Finishes Work at 5th hour will be idle for 1hour.
4	6	14	14	23		3	-do- 3 hr.
5	14	24	24	31		1	-do- 3 hr.
2	24	30	31	32	1	2	1 hr as job finished early 1 hr idle.

Total elapsed time = 32 hours. (This includes idle time of job and idle time of machines).

The procedure: Let Job 1 is loaded on machine *A* first at zero th time. It takes two hours to process on the machine. Job 1 leaves the machine *A* at two hours and enters the machine 2 at 2nd hour. Up to the time i.e first two hours, the machine *B* is idle. Then the job 1 is processed on machine *B* for 3 hours and it will be unloaded. As soon as the machine *A* becomes idle, *i.e.* at 2nd hour then next job 3 is loaded on machine *A*. It takes 4 hours and the job leaves the machine at 6th hour and enters the machine *B* and is processed for 6 hours and the job is completed by 11th hour. (Remember if the job is completed early and the Machine *B* is still busy, then the job has to wait and the time is entered in job idle column. In case the machine B completes the previous job earlier, and the machine *A* is still processing the next job, the machine has to wait for the job. This will be shown as machine idle time for machine *B*.). Job 4 enters the machine *A* at 6th hour and processed for 8 hours and leaves the machine at 14th hour. As the machine *B* has finished the job 3 by 11th hour, the machine has to wait for the next job (job 4) up to 14th hour. Hence 3 hours is the idle time for the machine *B*. In this manner we have to calculate the total elapsed time until all the jobs are over.

Problem 3.2.

There are 6 jobs to be processed on Machine *A*. The time required by each job on machine *A* is given in hours. Find the optimal sequence and the total time elapsed.

Solution

Here there is only one machine. Hence the jobs can be processed on the machine in any sequence depending on the convenience. The total time elapsed will be total of the times given in the problem. As soon as one job is over the other follows. The total time is 32 hours. The sequence may be any order.
For example: 1,2,3,4,5,6 or 6,5,4,3,2,1, or 2, 4 6 1 3 5 and so on.

Problem 3.3.

A machine operator has to perform two operations, turning and threading, on a number of different jobs. The time required to perform these operations in minutes for each job is given. Determine the order in which the jobs should be processed in order to minimize the total time required to turn out all the jobs.

Job:	1	2	3	4	5	5
Time for turning (in min.)	3	12	5	2	9	11
Time for threading (in min).	8	10	9	6	3	1

Solution

The smallest element is 1 in the given matrix and falls under second operation. Hence do the 6th job last. Next smallest element is 2 for the job 4 and falls under first operation hence do the fourth job first. Next smallest element is 3 for job 1 falls under first operation hence do the first job second. Like this go on proceed until all jobs are over. The optimal sequence is :

4	1	3	2	5	6

Optimal sequence	*Turning operation*		*Threading operation*		*Job idle*	*Machine idle.*	
	In	*Out*	*In*	*Out*		*Turning*	*Threading*
4	0	2	2	8	---	--	2
1	2	5	8	6	3		
3	5	10	16	25	6		
2	10	22	25	35	3		
5	22	31	25	38	4		
6	31	42	42	43	--	1	--
	Total elapsed time : 43 minutes						

Jobs 1 2 3 4 5 6
Time in hours (Machine A) 6 4 3 2 9 8

The Job idle time indicates that there must be enough space to store the in process inventory between two machines. This point is very important while planning the layout of machine shops.

Problem 3.4.

There are seven jobs, each of which has to be processed on machine *A* and then on Machine *B* (order of machining is *AB*). Processing time is given in hours. Find the optimal sequence in which the jobs are to be processed so as to minimize the total time elapsed.

Job:	1	2	3	4	5	6	7
	1	4	5	3	2	7	6
Machine: A (Time in hours)	3	12	15	6	10	11	9
Machine: B (Time in hours)	8	10	10	6	12	1	3

Solution

By Johnson and Bellman method the optimal sequence is:

Optimal Sequence	*Machine-A*		*Machine-B*		*Machine idle Job idle*		*Job idle time*	*Remarks*
	In	*Out*	*In*	*Out*	*A*	*B*		
1	0	3	3	11		3	--	
4	3	9	11	17			2	Job finished early
5	9	19	19	31		2		Machine A takes more time.
3	9	34	34	44		3		Machine A takes more time.
2	34	46	46	56		2		-do-
7	46	55	56	59			1	Job finished early
6	55	66	66	67	1	7		Machine A takes more time. Last is finished on machine A at 66th hour.
	Total Elapsed Time : 67 hours.							

3.3.2. SEQUENCING OF 'N' JOBS ON "M" MACHINES

A general sequencing problem of processing of '*n*' jobs through '*m*' machines $M_1, M_2, M_3, \ldots, M_n$ in the order $M_1 M_2 M_3 \ldots M_n$ can be solved by applying the following rules.

If a_{ij} where i = 1,2,3…..*n* and *j* = 1, 2, 3…….*m* is the processing time of *i th* job on *j th* machine, then find $Minimum_i\, a_{i1}$ and $Minimum_i\, a_{im}$ (*i.e.* minimum time element in the first machine and minimum time element in last Machine) and find $Maximum_i\, a_{ij}$ of intermediate machines *i.e* 2 nd machine to m-1 machine.

The problem can be solved by converting it into a two-machine problem if the following conditions are satisfied.

(a) Min $a_{i1} \geq$ Max. a_{ij} for all *j* = 1,2,3,…..*m*-1

(b) Min $a_{im} \geq$ Max a_{ij} for all *j* = 1, 2,3 …….*m-1*

At least one of the above must be satisfied. Or both may be satisfied. If satisfied, then the problem can be converted into 2- machine problem where Machine $G = a_{i1} + a_{i2} + \ldots + a_{i\,m-1}$ and Machine $G = a_{i2} + a_{i3} + \ldots + a_{i\,m}$,Where *i* = 1,2,3,….*n*.

Once the problem is *a* 2- machine problem, then by applying Johnson Bellman algorithm we can find optimal sequence and then workout total elapsed time as usual.

(Point to remember: Suppose $a_{i2} + a_{i3} + \ldots + a_{i\,m} = a$ constant number for all 'i' , we can consider two extreme machines *i.e*. machine 1 and machine -m as two machines and workout optimal sequence).

Problem 3.5.

There are 4 jobs *A*, *B*, *C* and *D*, which is to be, processed on machines M_1, M_2,M_3 and M_4 in the order $M_1M_2M_3M_4$ .The processing time in hours is given below. Find the optimal sequence.

Solution: From the data given, Min a_{i1} is 12 and Min a_{i4} is 12.

Max a_{i2} = 5 and Max a_{i3} = 10.

As Min a_{i1} is > than both Min a_{i2} and Min a_{i3}, the problem can be converted into 2 – machine problem as discussed above. Two-machine problem is:

Jobs.	*Machine (processing times in hours)*	
	G	H
A	15+5+4 = 24	5+4+14 = 23
B	12+2+10 = 24	2+10+12 = 24
C	13+3+6 = 22	3+6+15 = 24
D	16+0+3 = 19	0+3+19 = 22

Applying Johnson and Bellman rule, the optimal sequence is:

D	C	B	A

Sequence	*Machine* M_1 *Time in hours*		*Machine* M_2 *Time in hours*		*Machine* M_3 *Time in hours*		*Machine* M_4 *Time in hours*		*Job idle Time in hours*	*Machine idle Time in hours*			
	In	Out	In	Out	In	Out	In	Out		M_1	M_2	M_3	M_4
D	0	16	16	16	16	19	19	38				16	19
C	19	29	29	32	32	38	38	53			29	13	
B	29	41	41	43	43	53	53	65			9	5	
A	41	56	56	61	61	65	65	79		23	18	14	
	Total Elapsed Time : 79 hrs.												

Questions

Q.1. A bookbinder has one printing press, one binding machine and the manuscripts of a number of different books. The times required to perform printing and binding operations for ach book are known. Determine the order in which the books should be processed in order to minimize the total time required to process all the books. Find also the total time required processing all the books.

Printing time in minutes.

BOOK	A	B	C	D	E
Printing time:	40	90	80	60	50
Binding Time:	50	60	20	30	40

Suppose that an additional operation, finishing is added to the process described above and the time in minutes for finishing operation is as given below what will be the optimal sequence and the elapsed time.

Job	*Machine (processing times in hours)*			
	M_1	M_2	M_3	M_4
	a_{i1}	a_{i2}	a_{i3}	a_{i4}
A	15	5	4	14
B	12	2	10	12
C	13	3	6	15
D	16	0	3	19

BOOK	A	B	C	D	E
Finishing time (min):	80	100	60	70	110

Q.2. A ready-made garments manufacturer has to process 7 items through two stages of production, *i.e.* Cutting and Sewing. The time taken for each of these items at different stages are given in hours below, find the optimal sequence and total elapsed time.

Item:	1	2	3	4	5	6	7
Cutting time in Hrs.:	5	7	3	4	6	7	12
Sewing time in Hrs:	2	6	7	5	9	5	8

Suppose a third stage of production is added, say pressing and packing with processing time in hours as given below, find the optimal sequence and elapsed time.

Pressing time (Hrs)	10	12	11	13	12	10	11

Answers

Q.1. For two processes: sequence is : *ABEDC* and the elapsed time is 340 min.
For three processes: the optimal sequence is: *DAEBC* and the total elapsed time is 510 min.

Q.2. For two stages the sequence is : 3457261 and the time is 46 hours.
For three stages the sequence is : 1436257 and the time is 86 hours

3.4 Summary

The short-term schedules show an optimal order (sequence) and time in which jobs are processed. They also show timetables for jobs, equipment, people, materials, facilities and all other resources that are needed to support the production plan. The schedules should use the resources efficiently to give low costs and high utilizations. Other purpose of scheduling are, minimizing customers waiting time, meeting promised delivery dates, keeping stock levels low, giving preferred working pattern, minimizing waiting time of patients in a hospital for different types of tests and so on.

Units 19 & 20

Course Structure

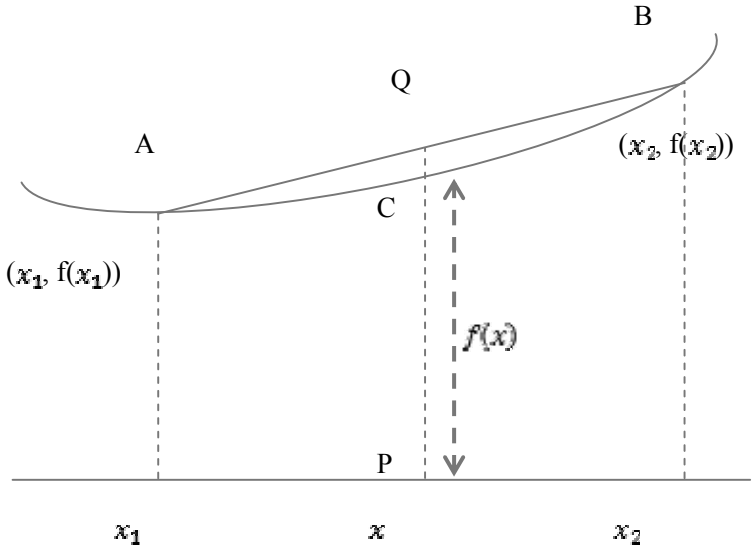
4.1 Convex Nonlinear Programming Problem

First, we define the convexity of a function which facilitates the further studies on nonlinear programming problems with equality and inequality constraints.

Definition 1. Let S be a convex set in $\mathbb{R}^n$. A function f(X) defined on S is said be convex if for any pair of points X_1, X_2 in Sand Vα: $0 \leq \alpha \leq 1$,

$$f\big((1-\alpha)X_1 + \alpha X_2\big) \leq (1-\alpha)f(X_1) + \alpha f(X_2).$$

Geometrically speaking in two dimensional plane, Definition 1 means that f(x) is convex if for any two points x_1 and x_2 in S, the chord joining the points (x_1, f(x_1)) and (x_2' f(x_2)) is above f(x), i.e., for any point $x \in [x_1, x_1]$, f(x) $\leq$ PQ, where Q is on the chord, see Fig.1.



Remarks. 1. A function f(X) is strictly convex if we have strict inequalities in Definition 1.

2. A function f(X) is concave (or strictly concave) if - f(X) is convex (or strictly convex).

3. A linear function is convex as well as concave.

Proposition 1. The sum of two convex functions is convex.

Proof. Let f_1 and f_1 be two convex functions defined on a convex set S $\subseteq$]Rn. Then, for any two points X_1 and X_2 in S, we have for all α: 0 $\leq \alpha \leq$ 1,

$$f_1\big((1-\alpha)X_1 + \alpha X_2\big) \leq (1-\alpha)f_1(X_1) + \alpha f_1(X_2).$$

$$f_2\big((1-\alpha)X_1 + \alpha X_2\big) \leq (1-\alpha)f_2(X_1) + \alpha f_2(X_2).$$

Now,

$$\begin{aligned}(f_1+f_2)\big((1-\alpha)X_1+\alpha X_2\big) &= f_1\big((1-\alpha)X_1+\alpha X_2\big) + f_2\big((1-\alpha)X_1+\alpha X_2\big)\\ &\leq (1-\alpha)f_1(X_1) + \alpha f_1(X_2) + (1-\alpha)f_2(X_1) + \alpha f_2(X_2)\\ &= (1-\alpha)(f_1+f_2)(X_1) + \alpha(f_1+f_2)(X_2).\end{aligned}$$

Proposition 2. Let f(X) = X^TAX. Then f(X) is convex in Rn if X^TAX is positive semi-definite

Corollary 1. Under the conditions of Proposition 2, if f(X) $= X^TAX$ is positive definite, then f(X) is strictly convex.

Corollary 2. Under the conditions of Proposition 2, we have

1. f(X) $\in C^2$ is convex $\Leftrightarrow$ its Hessian matrix is positive semidefinite.

2. f(X) $\in C^2$is strictly convex $\Leftrightarrow$ its Hessian matrix is positive definite.

The proof of Corollary 2 is a direct consequence of the fact that H = 2A in a quadratic form. Note that quadratic forms and quadratic functions have different meaning. Corollary 2 is very useful to decide convexity of any quadratic function.

Theorem 1. Let j(X) be a convex function defined over a convex set S in Rn. Then the local minimum is global minimum of j(X) over S.

Proof. Let X* be a point of local minimum. Hence, $f(X^*) \leq f(X^* + \delta)$, where {X : IX - X*I < δ, δ> 0} = N_δ (X*). Take any point $X_1 \in N_\delta$(X*). There exists α: 0 $\leq \alpha \leq$ 1such that X_1 = αX* + (1 - α)X for any X in S. Now, in view of convexity of f on S, we have

$$\begin{aligned}f(X_1) &= \text{J}(\alpha\text{X}^* + (1-\alpha)\text{X})\\ &\leq \alpha f(\text{X}^*) + (1-\alpha)f(\text{X})\\ &\leq \alpha f(X_1) + (1-\alpha)f(\text{X})\end{aligned}$$

This implies f(X*) $\leq$ J(X) or f(X*) $\leq$f(X_1 $\leq$ f(X). Since X is arbitrary point, and hence f(X*) $\leq$ f(X) for all X in S. So, X* is a point of the global minimum.

For developing the theory, write the problem in the format

min f(X)

s.t. g_i(X) $\leq$ 0, i = 1,2, ... , m

X $\geq$ 0,

f(X) and g_i(X) are convex functions.

The above nonlinear problem is called convex nonlinear programming problem (CNLPP). Note that f(X) and g_i(X) are convex functions over some common convex set.

Remark: Theorem 1 ensures that in a CNLPP the relative minimum or relative maximum is global minimum or global maximum.

Theorem 2. A set S = {X : g_i(X) $\leq$ 0, X $\geq$ 0} of feasible solutions of CNLPP is a convex set.

4.2 Optimal conditions

In mathematical analysis, the **maxima and minima** (the respective plurals of **maximum** and **minimum**) of a function, known collectively as **extrema** (the plural of **extremum**), are the largest and smallest value of the function, either within a given range (the **local** or **relative** extrema) or on the entire domain of a function (the **global** or **absolute** extrema).[1][2][3] Pierre de Fermat was one of the first mathematicians to propose a general technique, adequately, for finding the maxima and minima of functions.

As defined in set theory, the maximum and minimum of a set are the greatest and least elements in the set, respectively. Unbounded infinite sets, such as the set of real numbers, have no minimum or maximum.

4.2.1 Definitions

A real-valued function f defined on a domain X has a **global** (or **absolute**) **maximum point** at x^* if $f(x^*) \geq f(x)$ for all x in X. Similarly, the function has a **global** (or **absolute**) **minimum point** at x^* if $f(x^*) \leq f(x)$ for all x in X. The value of the function at a maximum point is called the **maximum value** of the function and the value of the function at a minimum point is called the **minimum value** of the function.

If the domain X is a metric space then f is said to have a **local** (or **relative**) **maximum point** at the point x^* if there exists some $\varepsilon > 0$ such that $f(x^*) \geq f(x)$ for all x in X within distance ε of x^*. Similarly, the function has a **local minimum point** at x^* if$f(x^*) \leq f(x)$ for all x in X within distance ε of x^*. A similar definition can be used when X is a topological space, since the definition just given can be rephrased in terms of neighbourhoods.

In both the global and local cases, the concept of a **strict** extremum can be defined. For example, x^* is a **strict global maximum point** if, for all x in X with $x \neq x^*$, we have $f(x^*) > f(x)$, and x^* is a **strict local maximum point** if there

exists some $\varepsilon > 0$ such that, for all x in X within distance ε of x^* with $x \neq x^*$, we have $f(x^*) > f(x)$. Note that a point is a strict global maximum point if and only if it is the unique global maximum point, and similarly for minimum points.

A continuous real-valued function with a compact domain always has a maximum point and a minimum point. An important example is a function whose domain is a closed (and bounded) interval of real numbers (see the graph above).

4.2.2 Finding maxima and Minima

Finding global maxima and minima is the goal of mathematical optimization. If a function is continuous on a closed interval, then by the extreme value theorem global maxima and minima exist. Furthermore, a global maximum (or minimum) either must be a local maximum (or minimum) in the interior of the domain, or must lie on the boundary of the domain. So a method of finding a global maximum (or minimum) is to look at all the local maxima (or minima) in the interior, and also look at the maxima (or minima) of the points on the boundary, and take the largest (or smallest) one.

Local extrema of differentiable functions can be found by Fermat's theorem, which states that they must occur at critical points. One can distinguish whether a critical point is a local maximum or local minimum by using the first derivative test, second derivative test, or higher-order derivative test, given sufficient differentiability.

For any function that is defined piecewise, one finds a maximum (or minimum) by finding the maximum (or minimum) of each piece separately, and then seeing which one is largest (or smallest).

4.3 The Method of Steepest Descent

When it is not possible to find the minimum of a function analytically, and therefore must use an iterative method for obtaining an approximate solution, Newton's Method can be an effective method, but it can also be unreliable. Therefore, we now consider another approach.

Given a function f : $\mathbb{R}^n \to \mathbb{R}$ that is differentiable at x_0, the direction of steepest descent is the vector $-\nabla f(x_0)$. To see this, consider the function

$$\varphi(t) = f(x_0 + tu),$$

where u is a unit vector; that is, $\|u\| = 1$. Then, by the Chain Rule,

$$\varphi'(t) = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \cdots\cdots + \frac{\partial f}{\partial x_n}\frac{\partial x_n}{\partial t}$$

$$= \frac{\partial f}{\partial x_1}u_1 + \cdots\cdots + \frac{\partial f}{\partial x_n}u_n$$

$$= \nabla f(x_0 + tu) \,.u.$$

and therefore

$$\varphi'(0) = \nabla f(x_0).u = \|\nabla f(x_0)\|.\cos\theta,$$

where θ is the angle between $\nabla f(x_0)$ and u. It follows that $\varphi'(0)$ is minimized when $\theta = \pi$, which yields

$$u = -\frac{\nabla f(x_0)}{\|\nabla f(x_0)\|}, \quad \varphi'(0) = -\|\nabla f(x_0)\|.$$

We can therefore reduce the problem of minimizing a function of several variables to a single-variable minimization problem, by finding the minimum of $\varphi(t)$ for this choice of u. That is, we find the value of t, for t > 0, that minimizes

$$\varphi_0(t) = f(x_0 - t\nabla f(x_0)).$$

After finding the minimizer t_0, we can set

$$x_1 = x_0 - t_0\nabla f(x_0)$$

and continue the process, by searching from x_1 in the direction of $-\nabla f(x_1)$ to obtain x_2 by minimizing $\varphi_1(t) = f(x_1 - t\nabla f(x_1))$ and so on.

This is the Method of Steepest Descent: given an initial guess x_0, the method computes a sequence of iterates $\{x_k\}$ where

$$x_{k+1} = x_k - t_k\nabla f(x_k), \quad k = 0,1,2,\ldots\ldots$$

where $t_k > 0$ minimizes the function

$$\varphi_k(t) = f(x_k - t\nabla f(x_k))$$

Example:
We apply the Method of Steepest Descent to the function f(x; y) = $4x^2$ - 4xy + $2y^2$ with initial point x_0 = (2, 3).

We first compute the steepest descent direction from

$\nabla f(x, y) = (8x - 4y, 4y - 4x)$

to obtain

$\nabla f(x_0) = \nabla f(2, 3) = (4, 4)$,

We then minimize the function

$\varphi(t) = f((2, 3) - t(4, 4)) = f(2 - 4t, 3 - 4t)$

by computing

$$\begin{aligned}\varphi'(t) &= -\nabla f(2 - 4t, 3 - 4t).(4, 4)\\ &= (8(2 - 4t) - 4(3 - 4t), 4(3 - 4t) - 4(2 - 4t)).(4, 4)\\ &= (16 - 32t - 12 + 16t, 12 - 16t - 8 + 16t).(4, 4)\\ &= -(-16t + 4, 4).(4, 4)\\ &= 64t - 32.\end{aligned}$$

This strictly convex function has a strict global minimum when $\varphi'(t) = 64t-32$, or t = 1/2, as can be seen by noting that $\varphi''(t) = 64 > 0$. We therefore set

$$x_1 = x_0 - \frac{1}{2}\nabla f(x_0) = (2,3) - \frac{1}{2}(4,4) = (0,1).$$

Continuing the process, we have

$$\nabla f(x_1) = \nabla f(0,1) = -(4,4),$$

and by defining

$\varphi(t) = f((0, 1) - t(-4; 4)) = f(4t; 1 - 4t)$,

we obtain

$$\varphi'(t) = -(8(4t) - 4(1 - 4t), 4(1 - 4t) - 4(4t))).(-4; 4) = -(48t - 4, -32t + 4).(-4; 4) = 320t - 32,$$

We have $\varphi'(t) = 0$ when t = 1/10, and because $\varphi''(t) = 320$, this critical point is a strict global minimizer. We therefore set

$$x_2 = x_1 - \frac{1}{10}\nabla f(x_1) = (0,1) - \frac{1}{10}(-4,4) = \left(\frac{2}{5}, \frac{3}{5}\right).$$

Repeating this process yields $x_3 = (0, 0.2)$. We can see that the Method of Steepest Descent produces a sequence of iterates x_k that is converging to the strict global minimizer of $f(x, y)$ at
$x^* = (0, 0)$.

The following theorems describe some important properties of the Method of Steepest Descent.

Theorem: Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable on $\mathbb{R}^n$, and let $x_0 \in D$. Let $t^* > 0$ be the minimizer of the function

$$\varphi(t) = f(x_0 - \nabla f(x_0)), t \geq 0$$

and let $x_1 = x_0 - t^* \nabla f(x_0)$.

Then $f(x_1) < f(x_0)$.

That is, the Method of Steepest Descent is guaranteed to make at least some progress toward a minimizer x^* during each iteration. This theorem can be proven by showing that $\varphi'(0) < 0$, which guarantees the existence of $\bar{t} > 0$ such that $\varphi(\bar{t}) < \varphi(0)$.

Theorem: Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable on $\mathbb{R}^n$, and let x_k and x_{k+1}, for $k \geq 0$, be two consecutive iterates produced by the Method of Steepest Descent. Then the steepest descent directions from x_k and x_{k+1} are orthogonal; that is,

$$\nabla f(x_k).\nabla f(x_{k+1}) = 0.$$

Theorem: Let $f : \mathbb{R}^n \to \mathbb{R}$ be a coercive function with continuous first partial derivatives on $\mathbb{R}^n$. Then, for any initial guess x_0, the sequence of iterates produced by the Method of Steepest Descent from x_0 contains a subsequence that converges to a critical point of f.

Theorem: Let $f : \mathbb{R}^n \to \mathbb{R}$ be a coercive, strictly convex function with continuous first partial derivatives on $\mathbb{R}^n$. Then, for any initial guess x_0, the sequence of iterates produced by the Method of Steepest Descent from x_0 converges to the unique global minimizer x^* of $f(x)$ on $\mathbb{R}^n$.

This theorem can be proved by noting that if the sequence $\{x_k\}$ of steepest descent iterates does not converge to x^*, then any subsequence that does not converge to x^* must contain a subsequence that converges to a critical point, by the previous theorem, but f(x) has only one critical point, which is x^*, which yields a contradiction.

4.4 Karush–Kuhn–Tucker (KKT) conditions

In mathematical optimization, the **Karush–Kuhn–Tucker (KKT) conditions**, also known as the **Kuhn–Tucker conditions**, are first-order necessary conditions for a solution in nonlinear programming to be optimal, provided that some regularity conditions are satisfied. Allowing inequality constraints, the KKT approach to

nonlinear programming generalizes the method of Lagrange multipliers, which allows only equality constraints. The system of equations and inequalities corresponding to the KKT conditions is usually not solved directly, except in the few special cases where a closed-formsolution can be derived analytically. In general, many optimization algorithms can be interpreted as methods for numerically solving the KKT system of equations and inequalities.

The KKT conditions were originally named after Harold W. Kuhn, and Albert W. Tucker, who first published the conditions in 1951.[2] Later scholars discovered that the necessary conditions for this problem had been stated by William Karush in his master's thesis in 1939.

Consider the following nonlinear minimization or maximization problem:

Optimize $f(x)$

subject to

$$g_i(x) \leq 0,$$
$$h_j(x) = 0,$$

where x is the optimization variable, f is the objective or utility function, $g_i \ (i = 1,2,\ldots,m)$ are the inequality constraint functions, and $h_j \ (j = 1,2,\ldots,l)$ are the equality constraint functions. The numbers of inequality and equality constraints are denoted *m* and ℓ, respectively.

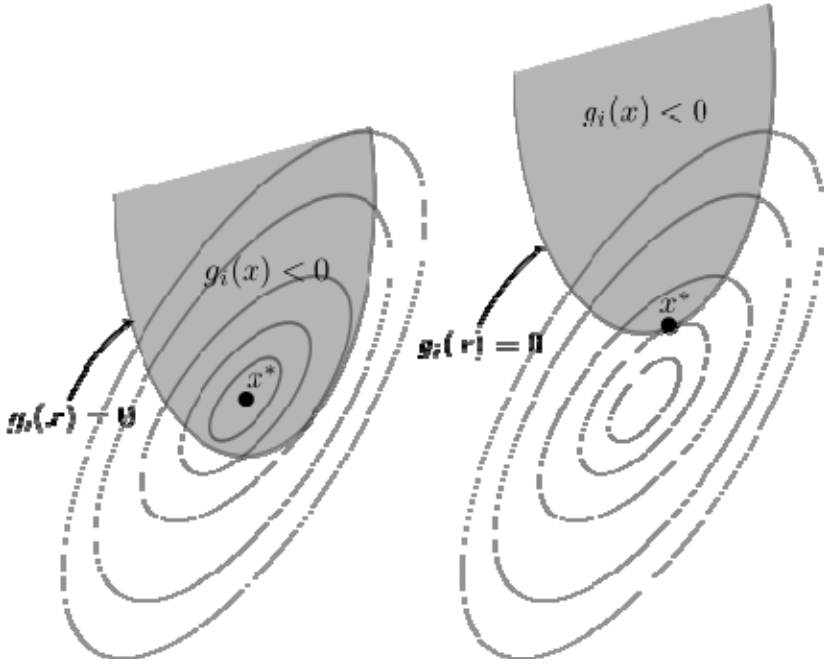
Necessary Conditions:

Suppose $f: \mathbb{R}^n \to \mathbb{R}$ that the objective function and $g_i: \mathbb{R}^n \to \mathbb{R}$ the constraint functions and $h_j: \mathbb{R}^n \to \mathbb{R}$ $h_j: \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable at a point x^*. If x^* is a local optimum and the optimization problem satisfies some regularity conditions (see below), then there exist constants $\mu_i \ (i = 1,2,\ldots,m)$ and $\lambda_j \ (j = 1,2,\ldots,l)$

$\lambda_j \ (j = 1,2,\ldots,l)$, called

KKT multipliers, such that

Stationarity



Inequality constraint diagram for optimization problems

For maximizing $f(x)$: $\nabla f(x^*) = \sum_{i=1}^{m} \mu_i \nabla g_i(x^*) + \sum_{j=1}^{l} \lambda_j \nabla h_j(x^*),$

For minimizing $f(x)$: $-\nabla f(x^*) = \sum_{i=1}^{m} \mu_i \nabla g_i(x^*) + \sum_{j=1}^{l} \lambda_j \nabla h_j(x^*),$

$$-\nabla f(x^*) = \sum_{i=1}^{m} \mu_i \nabla g_i(x^*) + \sum_{j=1}^{l} \lambda_j \nabla h_j(x^*),$$

Primal feasibility

$g_i(x^*) \leq 0, for\ i = 1,2, \ldots\ldots, m$
$h_j(x^*) \leq 0, for\ j = 1,2, \ldots\ldots, lh_j(x^*) \leq 0, for\ j = 1,2, \ldots\ldots, l$

Dual feasibility

$\mu_i \geq 0, for\ i = 1,2, \ldots\ldots, m\mu_i \geq 0, for\ i = 1,2, \ldots\ldots, m$

Complementary slackness

$\mu_i g_i(x^*) = 0, for\ i = 1,2, \ldots\ldots, m.$

In the particular case $m = 0$, i.e., when there are no inequality constraints, the KKT conditions turn into the Lagrange conditions, and the KKT multipliers are called Lagrange multipliers. If some of the functions are non-differentiable, sub-differential versions of Karush–Kuhn–Tucker (KKT) conditions are available.

In order for a minimum point x^* to satisfy the above KKT conditions, the problem should satisfy some regularity conditions; some common examples are tabulated here:

Constraint	Acronym	Statement
Linearity constraint qualification	LCQ	If g_i and h_j are affine functions, then no other condition is needed.
Linear independence constraint qualification	LICQ	The gradients of the active inequality constraints and the gradients of the equality constraints are linearly independent at x^*.

Mangasarian-Fromovitz constraint qualification	MFCQ	The gradients of the equality constraints are linearly independent at x^* and there exists a vector $d \in \mathbb{R}^n$ such that $\nabla g_i(x^*)^T d < 0$ for all active inequality constraints and $\nabla h_j(x^*)^T d = 0 \nabla h_j(x^*)^T d = 0$ for all equality constraints.[6]
Constant rank constraint qualification	CRCQ	For each subset of the gradients of the active inequality constraints and the gradients of the equality constraints the rank at a vicinity of $x^* x^*$ is constant.
Constant positive linear dependence constraint qualification	CPLD	For each subset of gradients of active inequality constraints and gradients of equality constraints, if the subset of vectors is linearly dependent at $x^* x^*$ with non-negative scalars associated with the inequality constraints, then it remain linearly dependent in a neighborhood of $x^* x^*$.

It can be shown that LICQ⇒MFCQ⇒CPLD and LICQ⇒CRCQ⇒CPLD (and the converses are not true), In practice weaker constraint qualifications are preferred since they provide stronger optimality conditions.

Kuhn-Tucker Necessary Conditions:

Maximize f(x), $X = (x_1, x_2, \ldots\ldots, x_n)$

Subject to $g_i(x) \le b_i$, i=1,2,…..,m.

Including the non-negative constraints $x \ge 0$, the necessary conditions for a local maxima at X are

i) $\frac{\partial L(\bar{x}, \bar{\lambda}, \bar{s})}{\partial x_j} = 0$, j=1,2,…..,n.

ii) $\bar{\lambda}_i[g_i(\bar{x}) - b_i] = 0$,

iii) $g_i(\bar{x}) \le b_i$,

iv) $\bar{\lambda}_i \ge 0$, i =1,2,……,m.

Where $L(\bar{x}, \bar{\lambda}, \bar{s})$ is Lagrange function L defined by:

$L(x_1, x_2, \ldots\ldots, x_n; \lambda_1, \lambda_2, \ldots\ldots, \lambda_m) = f(x_1, x_2, \ldots\ldots, x_n) + \lambda_1 g_1(x_1, x_2, \ldots\ldots, x_n) + \lambda_2 g_2(x_1, x_2, \ldots\ldots, x_n) + \ldots\ldots\ldots\ldots + \lambda_m g_m(x_1, x_2, \ldots\ldots, x_n)$.

Here $\lambda_1, \lambda_2, \ldots\ldots, \lambda_m$ are Lagrange Multipliers.

For the stationary points $\frac{\partial L}{\partial x_j} = 0$, $\frac{\partial L}{\partial \lambda_i} = 0$, for j =1,2,…..,n; i =1,2,…….,m.

Example: Consider an inventory model written as

Min $TAC(D,S,q) = \frac{SD}{q} + \frac{aHq^2}{6D} + \frac{b}{S^{\alpha} q^{\beta}}$

Subject to $w_0 q \leq W$, D, S, q > 0.

The above problem can be written as Lagrangian form

$$\mathcal{L}(S, D, q) = \frac{SD}{q} + \frac{aHq^{\beta}}{6D} + \frac{b}{S^{\gamma} q^{\gamma}} - \lambda(W - qw_0).$$

From Kuhn-Tucker necessary conditions the solution can be obtained as (doing the partially derivative w.r.t. S, D, q and λ respectively),

$$\frac{D}{q} - \frac{xb}{S^{\gamma+1} q^{\beta}} = 0,$$

$$\frac{S}{q} - \frac{aHq^{\beta}}{6D^{\beta}} = 0,$$

$$-\frac{SD}{q^{\beta}} + \frac{aHq}{3D} - \frac{\gamma b^{\gamma}}{Sq^{\gamma+1}} + \lambda w_0 = 0,$$

and $\lambda(W - w_0 q) = 0$.

Here two conditions $\lambda = 0$ or $W - w_0 q = 0$.

When $\lambda = 0$, does not satisfies all the equation of above problem, so consider $W - w_0 q = 0$.

And optimal solution is,

$$S^* = \frac{aH\,(xb)^{\frac{\beta}{2\gamma+2}}}{6\frac{W}{w_0}^{\frac{\beta}{2\gamma+2}}\left(\frac{aH}{6}\right)^{\frac{2\gamma+\beta}{2\gamma+2}}},$$

$$D^* = \frac{\left(\frac{W}{w_0}\right)^{\frac{2\gamma+4}{2\gamma+2}}\left(\frac{aH}{6}\right)^{\frac{\gamma+2}{2\gamma+2}}}{(xb)^{\frac{\beta}{2\gamma+2}}},$$

$$q^* = \frac{W}{w_0}.$$

4.5. Quadratic Programming:

Among several non-linear programming methods available for solving NLP problems, we shall discuss in this section, an NLP problem with non-linear objective function and linear constraints. Such an NLP problem is called quadratic programming problem. The general mathematical model of quadratic programming problem is as follows:

Optimize (Max or Min) $Z = \{\sum_{j=1}^{n} c_j x_j + \frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{n} x_j d_{jk} x_k\}$

subject to the constraints

$$\sum_{j=1}^{n} a_{ij} x_j \leq b_i$$

and $x_j \geq 0$ for all i and j

In matrix notations, QR problem is written as:

Optimize (Max or Min) $Z = c\,x + \frac{1}{2} x^T D\,x$

subject to the constraints

$Ax \leq b$

and$x_j \geq 0$

where $x = (x_1, x_2, \ldots, x_n)^T$; $c = (c_1, c_2, \ldots, c_n)$; $b = (b_1, b_2, \ldots, b_n)^T$

$D = [d_{jk}]$ is an n × n symmetric matrix, i.e. $d_{jk} = d_{kj}$ is an n × n matrix

If the objective function in QP problem is of minimization, then the matrix D is symmetric and positive definite (i.e. the quadratic term x^T D x in x is positive for all values of x except at x = 0) and objective function is strictly convex in x. But, if the objective function is of maximization, then matrix D is symmetric and negative-definite i.e. x^T D x < 0 for all values of x except for x = 0) and objective function is strictly concave in x. If matrix, D is null, then the QP problem reduces to the standard LP problem.

4.6. Khun-Tucker Conditions:

The necessary and sufficient Khun-Tucker conditions to get an optimal solution to the maximization QP problem subject to the linear constraints can be derived as follows:

Step 1: Introducing slack variables s_i^2 amd r_j^2 to constraints, the QP problem becomes:

Max $f(x) = \{\sum_{j=1}^{n} c_j x_j - \frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{n} x_j d_{jk} x_k\}$

subject to the constraints

$\sum_{j=1}^{n} a_{ij}x_j + s_i^2 = b_i$; i = 1, 2,…, m.

and$-x_j + r_j^2 = 0$; j = 1, 2, …, n.

Step 2: Forming the Lagrange function as follows:

$L(x, s, r, \boldsymbol{\lambda}, \mu) = f(x) - \sum_{j=1}^{n} \lambda_i(a_{ij}x_j + s_i^2 - b_i) - \sum_{j=1}^{n} \mu_j(-x_j + r_j^2)$

Step 3: Differentiate L(x, s, r, $\boldsymbol{\lambda}$, μ) partially with respect to the components of x, s, r, $\boldsymbol{\lambda}$ and μ. Then equate those derivatives with zero in order to get the required Khun-Tucker necessary conditions. That is,

(i) $c - \frac{1}{2}(2x^T D) - \boldsymbol{\lambda}A + \mu = 0$, or

$c_j - \sum_{k=1}^{n} x_k d_{jk} - \sum_{i=1}^{m} \lambda_i a_{ij} + \mu_j = 0$; j = 1, 2, …, n.

(ii) $-2\,\boldsymbol{\lambda}\, s = 0$ or $\lambda_i s_i^2 = 0$, or

$\lambda_i \{\sum_{j=1}^{n} a_{ij}x_j - b_i\} = 0$, i = 1, 2,…, m.

(iii) $-2\,\mu\, r = 0$ or $\mu_j r_j = 0$, j = 1, 2, …, n.

$\mu_j x_j = 0$, j = 1, 2, …, n.

(iv) $Ax + s_i^2 - b = 0$; i.e. $Ax \leq b$, or

$\sum_{j=1}^{n} a_{ij}x_j \leq b_i$, i = 1, 2,…, m.

(v) $-x + r^2 = 0$, i.e. $x \geq 0$, or

$x_j \geq 0$, j = 1, 2, …, n.

(vi) $\lambda_i, \mu_j, x_j, s_i, r_j \geq 0$.

These conditions, except (ii) and (iii), are linear constraints involving 2 (n+m) variables. The condition $\mu_j x_j = \lambda_i s_i = 0$ implies that both x_j and μ_j as well as s_i and λ_i cannot be basic variables at a time in a non-degenerate basic feasible solution $\mu_j x_j = 0$ and $\lambda_i s_i = 0$ are also called complementary slackness conditions.

4.7. Wolfe's Modified Simplex Method:

The Wolf's method for solving quadratic programming problem can be summarized in the following steps:

Step 1: Introducing artificial variables A_j (j = 1, 2, …, n) in the Kuhn-Tucker condition (i). Then we have

$c_j - \sum_{k=1}^{n} x_k d_{jk} - \sum_{i=1}^{m} \lambda_i a_{ij} + \mu_j + A_j = 0$

For a starting basic feasible solution we shall have $x_j = 0$, $\mu_j = 0$, $A_j = -c_j$ and $s_i^2 = b_i$. However, this solution would be desirable if and only if $A_j = 0$ for all j.

Step 2: Apply phase-I of the simplex method to check the feasibility of the constraints $Ax \leq b$. If there is no feasible solution, then terminate the solution procedure, otherwise get an initial basic feasible solution for phase-II. To obtain desired feasible solution solve the following problem:

Minimize $Z = \sum_{j=1}^{n} A_j$

subject to the constraints

$\sum_{k=1}^{n} x_k d_{jk} + \sum_{i=1}^{m} \lambda_i a_{ij} - \mu_j + A_j = c_j$, j = 1, 2, …, n.

$\sum_{j=1}^{n} a_{ij} x_j + s_i^2 = b_i$, i = 1, 2,…, m.

and $\lambda_i, \mu_j, x_j, s_i, A_j \geq 0$ for all i and j.

$\left.\begin{matrix}\lambda_i s_i = 0 \\ \mu_j x_j = 0\end{matrix}\right\}$ Complimentary slackness conditions

Thus, while deciding for a variable to enter into the basis at each iteration, the complementary slackness conditions must be satisfied.

The problem has 2(m+n) variables and (m+n) linear constraints, together with (m+n) complementary slackness conditions.

Step 3: Apply Phase-II of the simplex method to get an optimal solution to the problem given in step 2. The solution so obtained, will also an optimal solution of the quadratic programming problem.

Example: Use Wolfe's method to solve the quadratic programming problem:

Maximize $Z = 4x_1 + 6x_2 - 2x_1^2 - 2x_1x_2 - 2x_2^2$

subject to the constraint

$x_1 + 2x_2 \leq 2$ and $x_1, x_2 \geq 0$.

Solution: Consider non-negative conditions $x_1, x_2 \geq 0$ as inequality constraints. Add slack variable to all inequality constraints in order to express them as equations. The standard form of QP problem becomes:

Maximize $Z = 4x_1 + 6x_2 - 2x_1^2 - 2x_1x_2 - 2x_2^2$

subject to the constraint

(i) $x_1 + 2x_2 + s_1^2 = 2$ (ii) $-x_1 + r_1^2 = 0$ (iii) $-x_2 + r_2^2 = 0$ and $x_1, x_2, s_1, r_1, r_2 \geq 0$.

To obtain the necessary conditions, we construct the Lagrange function as follows:

$L(x_1, x_2, s_1, \boldsymbol{\lambda}_1, \mu_1, \mu_2, r_1, r_2) = (4x_1 + 6x_2 - 2x_1^2 - 2x_1x_2 - 2x_2^2) - \boldsymbol{\lambda}_1(x_1 + 2x_2 + s_1^2 - 2) - \mu_1(-x_1 + r_1^2) - \mu_2(-x_2 + r_2^2)$.

The necessary and sufficient conditions for the maximum of L and hence of Z are:

$\frac{\partial L}{\partial x_1} = 4 - 4x_1 - 2x_2 - \boldsymbol{\lambda}_1 + \mu_1 = 0.$

$\frac{\partial L}{\partial \lambda_1} = x_1 + 2x_2 + s_1^2 - 2 = 0.$

$\frac{\partial L}{\partial \mu_1} = -x_1 + r_1^2 = 0.$

$\frac{\partial L}{\partial r_1} = 2\mu_1 r_1 = 0.$

$\frac{\partial L}{\partial x_2} = 6 - 2x_1 - 4x_2 - 2\boldsymbol{\lambda}_1 + \mu_2 = 0.$

$\frac{\partial L}{\partial s_1} = 2\boldsymbol{\lambda}_1 s_1 = 0.$

$\frac{\partial L}{\partial \mu_2} = -x_2 + r_2^2 = 0.$

$\frac{\partial L}{\partial r_2} = 2\mu_2 r_2 = 0.$

After simplifying these conditions, we get:

(i) $4x_1 + 2x_2 + \boldsymbol{\lambda}_1 - \mu_1 = 4$ (ii) $2x_1 + 4x_2 + 2\boldsymbol{\lambda}_1 - \mu_2 = 6$ (iii) $x_1 + 2x_2 + s_1^2 = 2$.

$\left.\begin{matrix}\lambda_1 s_1 = 0 \\ \mu_1 x_1 = \mu_2 x_2 = 0\end{matrix}\right\}$ (Complimentary conditions)

and $x_1, x_2, \boldsymbol{\lambda}_1, \mu_1, \mu_2, s_1 \geq 0$

Introducing artificial variables A_1 and A_2 in the first two constraints respectively. Then the modified LP problem becomes:

Minimize $Z^* = A_1 + A_2$

subject to the constraints

$4x_1 + 2x_2 + \boldsymbol{\lambda}_1 - \mu_1 + A_1 = 4$

$2x_1 + 4x_2 + 2\boldsymbol{\lambda}_1 - \mu_2 + A_2 = 6$

$x_1 + 2x_2 + s_1^2 = 2$

and $x_1, x_2, \boldsymbol{\lambda}_1, \mu_1, \mu_2, A_1, A_2 \geq 0$

The initial basic feasible solution to this LP problem is shown in Table-1

Table-1

$c_j \longrightarrow$			0	0	0	0	0	0	1	1
C_B	Basic Variables B	Solution Values b(=x_B)	x_1	x_2	λ_1	μ_1	μ_2	s_1	A_1	A_2
1	A_1	4	**4**	2	1	-1	0	0	1	0
1	A_2	6	2	4	2	0	-1	0	0	1
0	s_1	2	1	2	0	0	0	1	0	0
$Z^* = 10$		$c_j - z_j$	-6	-6	-3	1	1	0	0	0

Iteration 1: in Table-1, the largest negative values among $c_j - z_j$ values is– 6 corresponding to x_1 and x_2 columns. This means of these two variables can be entered into the basis. Since $\mu_1 = 0$ (not in the basis), x_1 is considered to enter into the basis. It will replace A_1 in the basis. The new solution is shown in Table-2

Table-2

$c_j \longrightarrow$			0	0	0	0	0	0	1
C_B	Basic Variables B	Solution Values b(=x_B)	x_1	x_2	λ_1	μ_1	μ_2	s_1	A_2
0	x_1	1	1	1/2	1/4	-1/4	0	0	0
1	A_2	4	0	3	3/2	1/2	-1	0	1
0	s_1	2	0	**3/2**	-1/4	1/4	0	1	0
$Z^* = 4$		$c_j - z_j$	0	-3	-3/2	-1/2	1	0	0

Iteration 2: In Table-2, $\mu_2 = 0$ (not in the basis), therefore x_2 can be introduced into the basis to replace s_1, in the basis. The new solution is shown in Table-3

Table-3

$c_j \longrightarrow$			0	0	0	0	0	0	1
C_B	Basic Variables B	Solution Values b(=x_B)	x_1	x_2	λ_1	μ_1	μ_2	s_1	A_2
0	x_1	2/3	1	0	1/3	-1/3	0	-1/3	0
1	A_2	2	0	0	2	0	-1	-2	1
0	x_1	2/3	0	1	-1/6	1/6	0	2/3	0
$Z^* = 4$		$c_j - z_j$	0	0	-2	0	1	2	0

Iteration 2: In Table-3, $s_1 = 0$ (not in the basis), therefore λ_1 can be entered into the basis to replace A_2. The new solution is shown in Table-4

Table-4

$c_j \longrightarrow$			0	0	0	0	0	0
C_B	Basic Variables B	Solution Values b(=x_B)	x_1	x_2	λ_1	μ_1	μ_2	s_1
0	x_1	1/3	1	0	0	-1/3	1/6	0
0	λ_1	1	0	0	1	0	-1/2	-1
0	x_2	5/6	0	1	0	1/6	-1/12	1/2
$Z^* = 0$		$c_j - z_j$	0	0	0	0	0	0

In Table-4, since all $c_j - z_j = 0$, anoptimal solution for Phase-1 is reached. The optimal solution is:
$x_1 = 1/3$, $x_2 = 5/6$, $\boldsymbol{\lambda}_1 = 1$,$\boldsymbol{\lambda}_2 = 0$, $\mu_1 = \mu_2 = 0$, $s_1 = 0$
This solution also satisfied the complimentary conditions: $\lambda_1 s_1 = 0$; $\mu_1 x_1 = \mu_2 x_2 = 0$ and the restriction on the signs of Lagrange multipliers, λ_1,μ_1, μ_2.

Further, as $Z^* = 0$, this implies that the current solution is also feasible. Thus, the maximum value of the given quadratic programming problem is:

Max $Z = 4 x_1 + 6 x_2 - 2 x_1^2 - 2 x_1 x_2 - 2 x_2^2$
$= 4(1/3) + 6(5/6) - 2(1/3)^2 - 2(1/3)(5/6) - 2(5/6)^2 = 25/6$

4.8. Beals's Method:

In this method, instead of Kuhn-Tucker conditions, results based on calculus are used for solving a given quadratic programming problem.

Minimize $Z = c x + \frac{1}{2} x^T D x$ (1)

subject to the constraints

$A x = b$ (2)

and $x \geq 0$ (3)

where $x \in E^n$, $b \in E^m$, $c \in E^n$, D is symmetric n×n matrix and A is m×n matrix.

Beale's method starts with the partitioning of n variables in QP problem into the basic and non-basic variables at each iteration of the solution process, and expressing the basic variables as well as objective function in terms of non-basic variables. Let B be any m×n non-singular matrix that contains columns of A corresponding to the basic variables, $x_B \in$ m× (n-m) matrix that contains columns of A corresponding to basic variables, $x_N \in E^{n-m}$. Eqn. (2) can then be written as:

$[B, N] \begin{bmatrix} x_B \\ x_N \end{bmatrix} = b$ or $Bx_B + Nx_N = b$ or $x_B = B^{-1}b - B^{-1} N x_N$

Or $x_{B_i} = y_{i0} - \sum_{j=1}^{n-m} y_{ij} x_{N_j}$; i= 1, 2,……., m (4)

where $y_{i0} = (y_{10}, y_{20}, \ldots\ldots\ldots, y_{m0})^T = B^{-1}b$ and $y_{ij} = B^{-1} N$

For the current basic feasible solution $x_{N_j} = 0$ (j=1,2,……..,n - m), we have $x_{B_j} = y_{i0}$, (i=1,2,…….,m).

Assuming that $y_{i0} \geq 0$.

The objective function (1) in terms of x_B and x_N can be written as:

$$Z = [\, c_B \,, c_N] \begin{bmatrix} x_B \\ x_N \end{bmatrix} + \frac{1}{2} [x_B^T, x_N^T] \begin{bmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{bmatrix} \begin{bmatrix} x_B \\ x_N \end{bmatrix}$$

Expressing Z in terms of the remaining (n – m) non-basic variables x_N only, and simplifying, we get:

$Z = Z_0 + \alpha x_N + x_N^T G x_N$ (5)

where Z_0 = value of objective function Z when $x_N = 0$ and $x_{B_i} = y_{i0}$

G = symmetric matrix of order (n – m) x (n – m)

$\alpha = \alpha_1, \alpha_2, \ldots\ldots\ldots, \alpha_{n-m}$ (constant)

The Procedure

Step 1: Evaluate the partial derivatives of Z with respect to non-basic variables, $x_N = 0$. Thus, from Eqn.(5)
We get:

$\frac{\partial Z}{\partial x_{N_j}} = \alpha_j + 2 \sum_{k=1}^{n-m} g_{jk} x_{Nk}$; j = 1,2,……..,n – m (6)

Step 2: See the nature of $\left| \frac{\partial Z}{\partial x_{N_j}} \right|_{x_N} = \alpha_j$; k = 1,2,……..,n – m

(a) If $\alpha_j < 0$, for all j, then the current solution is also an optimal solution

(b) But if at least one $\alpha_j > 0$, one of the non-basic variables, which is currently at zero level, corresponding to the largest positive value of α_j, will be selected to enter the basis.

Step 3: $\left.\frac{\partial Z}{\partial x_N}\right|_{x_N=0}$ = α_j (largest), then choose non-basic variables x_r, for entering the basis. For this it will be profitable to go on increasing its value from zero till a point where either :

(a) any one of the present basic variables becomes negative, or

(b) $\partial Z/\partial x_{N_j}$ reduces to zero and is about to become negative.

Step 4: For maintaining the feasibility of the solution we must consider only that value of non-basic variables x_r, say β_1 , which has only a positive coefficient. In this case, the first basic variable selected to leave the basis should satisfy the usual minimum ratio rule of the simplex method and will be given by:

$$\beta_1 = \begin{cases} Min\left\{\frac{y_{i0}}{y_{ij}}\,;y_{ij}>0\right\} \\ \infty \;\;;\; y_{ij} \le 0,\; j=1,2,\dots\dots,n-m \end{cases} \qquad (7)$$

where $y_{i0} = x_{B_l}$

Since it is not desirable to increase the value of the non-basic variable x_r beyond the point where $\partial Z/\partial x_{N_j}$ becomes zero, the critical value of x_r say β_2, at which $\partial Z/\partial x_{N_j}$ becomes zero is given by :

$$\beta_2 = \begin{cases} \frac{|\alpha_j|}{2\,g_{ij}}\;;\; g_{ij}>0 \\ \infty\;;\; g_{ij} \le 0 \end{cases}$$

where g_{ij} is the element of matrix **G .**

Hence the value of non-basic variable x_r must be determined by taking the minimum between β_1 and β_2, that is, x_r = Min { β_1, β_2}. However if $\beta_1 = \beta_2 = \infty$, the value of x_r can be increased indefinitely without violating either the condition (a) or (b) of Step-3 and the condition that QP problem must have an unbounded solution . Moreover,

(i) If the entering variable x_r is increased up to only β_1 and at least one basic variable is reduced to Zero, then a new basic feasible solution can be obtained by the usual simplex method. But if by entering x_r into the basis two or more basic variables are reduced to zero, then the new solution, so obtained, will be degenerate and thus cycling can occur.

(ii) If the entering variable is increased up $\beta_2 (< \infty)$, then we may have more than m variables at positive level at any iteration. This stage comes when the new (non-basic) feasible solution occurs where $\partial Z/\partial x_{N_j} = 0$. At this stage we define a new variable (unrestricted) u_j as :

$$u_j = \frac{\partial Z}{\partial x_r} = \alpha_j + 2\sum_{k=1}^{n-m} g_{jk} x_{Nk}$$

The variable u_j is also called free variable. Clearly, we now have m+1 non-zero variables and m + 1 constraints. These variables forms a basic feasible solution to the new set of constraints:

Ax = b

$$u_j - 2\sum_{k=1}^{n-m} g_{jk} x_{Nk} = \alpha_j$$

The variable u_j is introduced in the set of constraints only for computational purposes and its value is zero at the next basic feasible solution. Now, the variables x_B and u_j are treated as basic variables. The new set of constraints is again expressed in terms of non-basic variables for obtaining the new basic feasible solution.

Step 5: Go to step 1 and repeat the entire procedure of getting a new basic feasible solution until no further improvement in the objective function can be obtained by making any permitted changes in one of the non-basic

variables. The permitted changes here include increase in all variables and decrease in free variables. In other words, the procedure terminate when:

$$\frac{\partial Z}{\partial x_{N_j}}\begin{cases}\leq 0, & \text{if } x_{N_j} \text{ is a restricted (non } - \text{ negative) variable}\\ = 0, & \text{if } x_{N_j} \text{ is a free variable.}\end{cases} \quad (8)$$

The necessary conditions (8) for terminating the produce are also sufficient for a global minimum if *D* is positive semi-definite or positive definite.

Remarks: 1. While evaluating $\partial Z/\partial u_j$, both increase and decrease must be checked, as u_j is unrestricted sign

2. If at any iteration a free variable becomes a basic variable is non-zero, then drop the new constraint containing it. This should be done because it is a free variable, and therefore, will neither be chosen to leave the basis nor will appear in the selection of leaving variable.

Example: Use Beale's method to solve quadratic programming problem:

Maximize $Z = 2\,x_1 + 3x_2 - 2\,x_2^2$

subject to the constraints

(i) $x_1 + 4x_2 \leq 4$ (ii) $x_1 + x_2 \leq 2$

and $x_1, x_2 \geq 0$

Solution: After introducing slack variables s_1 and s_2, the given constraints can be written as:

(i) $x_1 + 4x_2 + s_1 = 4$ (ii) $x_1 + x_2 + s_2 = 2$

and $x_1, x_2, s_1, s_2 \geq 0$.

Consider s_1 and $s_2$are basic variables in initial solution and express these in terms of non-basic variables s_1 and $s_2$as follows:

$s_1 = 4 + 1\,(-x_1) + 4\,(-x_2)$and $s_2 = 2 + 1\,(-x_1) + 2\,(-x_2)$

the initial basic feasible solution: $x_1 = x_2 = 0$; $s_1 = 4$ and $s_2 = 2$, is shown in Table-1

Table-1

Basic Variables B	Solution Values b(=x_B)	$\mathbf{x_1}$	$\mathbf{x_2}$	$\mathbf{s_1}$	$\mathbf{s_2}$
s_1	1/3	1	4	1	0
$\mathbf{s_2}$	1	1	1	0	1

The value of the objective function at this solution is Z = 0. Also $x_B = (s_1, s_2) = (4, 2)$ and $x_N = (x_1, x_2) = (0, 0)$. Expressing Z in terms of non-basic variables x_1 and x_2 we get:

$Z = 2\,x_1 + 3x_2 - 2\,x_2^2$ and $\frac{\partial Z}{\partial x_1} = 2$, $\frac{\partial Z}{\partial x_2} = 3 - 4\,x_2$

At the current basic feasible solution evaluate those partial derivatives of Z with respect to $\mathbf{x_N} = 0$, i.e. $x_1 = x_2 = 0$.

$\left.\frac{\partial Z}{\partial x_1}\right|_{\substack{x_1=0\\x_2=0}} = 2$ and $\left.\frac{\partial Z}{\partial x_2}\right|_{\substack{x_1=0\\x_2=0}} = 3$

Here $\alpha_1 = 2$ and $\alpha_2 = 3$. Since both of these are positive, therefore chose x_2 (due to most positive value α_2) to enter into the basis in order to improve the value of the objective function. Using Table-1, the critical value $\beta_2$of x_2 is given by:

(i) Largest value of x_2 without deriving any basic variable s_1 and s_2 to zero. Since

(a)$s_1 = 4 - x_1 - 4\,x_2$ (b)$s_2 = 2 - x_1 - x_2$

Therefore β_1 min {4/4, 2/1} = 1, (corresponding to y_{22})

(ii) The partial derivatives $\frac{\partial Z}{\partial x_2}$ becomes zero at $x_2 = ¾$ ($x_1 = 0$). Therefore

$\beta_2 = \frac{|\alpha_2|}{2\, g_{22}} = \frac{|3|}{2\,(2)} = \frac{3}{4}$

The new value of the entering variable x_2 is given by:

$X_2 = \min\{\beta_1, \beta_2\} = \{1, 3/4\} = ¾$

This value of x_2 corresponding to β_2; therefore case (ii) applies and neither of the current basic variables become zero. Consequently we introduce a free variables u_1 and the new constraint:

$u_1 = \frac{\partial Z}{\partial x_2} = 3 - 4x_2$ or $4x_2 + u_1 = 3$

as shown in table-2

It may be noted from tabe-2 that now $x_B = (s_1, s_2, u_1)$ and $x_N = (x_1, x_2)$.

Table-2

Basic Variables B	Solution Values b(=x_B)	**x_1**	**x_2**	**s_1**	**s_2**	**u_1**
s_1	4	1	4	1	0	0
s_2	2	1	1	0	1	0
u_1	**3**	0	4	0	0	1

Introducing x_2 into the basis and remove u_1 from the basis in Table-2. The new solution is shown in Table-3

Table-3

Basic Variables B	Solution Values b(=x_B)	**x_1**	**x_2**	**s_1**	**s_2**	**u_1**
s_1	1	1	0	1	0	0
s_2	5/4	1	0	0	1	1/4
x_2	3/4	0	1	0	0	-1/4

The new set of basic and non-basic variables is:

$x_B = (s_1, s_2, x_2) = (1, 5/4, 3/4)$; $x_N = (x_1, u_1) = (0, 0)$

Expressing basic variables x_2, s_1 and s_2 in terms of non-basic variables x_1 and u_1 as follows:

(i) $x_2 = \frac{3}{4} - \frac{1}{4}u_1$; (ii) $s_1 = 1 - x_1 - u_1$ (iii) $s_2 = \frac{5}{4} - x_1 - \frac{1}{4}u_1$

Also by eliminating the basic variable x_2 from the objective function and expressing it in terms of non-basic variables x_1 and u_1 we get:

$Z = 2x_1 + 3\left(\frac{3}{4} - \frac{u_1}{4}\right) - 2\left(\frac{3}{4} - \frac{u_1}{4}\right)^2 = \frac{9}{8} + 2x_1 - \frac{u_1^2}{8}$

Computing the partial derivatives of Z with respect to x_1 and u_1 we have

$\frac{\partial Z}{\partial x_1} = 2$; $\frac{\partial Z}{\partial u_1} = -\frac{u_1}{4}$

At the current solution we get:

$\left.\frac{\partial Z}{\partial x_1}\right|_{\substack{u_1=0\\x_1=0}} = 2$ and $\left.\frac{\partial Z}{\partial u_1}\right|_{\substack{u_1=0\\x_1=0}} = 0$

Since $\alpha_1 = 2$ and $\alpha_2 = 0$, choose x_1 to enter the basis. Using Table-3, the critical value β_1 of x_1 is given by

(i) Largest value of x_1 without deriving any basic variable s_1, s_2 and x_2 to zero. Since

(i) $x_2 = \frac{3}{4} - \frac{1}{4}u_1$; (ii) $s_1 = 1 - x_1 - u_1$ (iii) $s_2 = \frac{5}{4} - x_1 - \frac{1}{4}u_1$

therefore $\beta_1 = \min\{\frac{1}{1}, \frac{(5/4)}{1}\} = 1$.

(ii) Since partial derivative $\frac{\partial Z}{\partial x_1}$ is non-zero, therefore $\beta_2 = 0$.

Thus the new value of entering variable x_1 is: $x_1 = \min\{\beta_1, \beta_2\} = 1$. This value of x_1 corresponds to β_1, therefore case (i) applies and the new optimal solution is shown in Table-4

Table-4

Basic Variables B	Solution Values b(=x_B)	x_1	x_2	s_1	s_2	u_1
x_1	1	1	0	1	0	0
s_2	1/4	1	0	-1	1	-3/4
x_2	3/4	0	1	0	0	-1/4

Now we have$x_B = (x_1, s_2, x_2) = (1, 1/4, 3/4)$; $x_N = (s_1, u_1) = (0, 0)$

Expressing basic variables x_1, $x_2$and$s_2$in terms of non-basic variables s_1 and u_1 as follows:

(i) $x_1 = 1 - s_1 - u_1$; (ii) $s_2 = \frac{1}{4} + s_1 + \frac{3}{4}u_1$ (iii) $x_2 = \frac{3}{4} + \frac{1}{4}u_1$

Also expressing objective function Z in terms of non-basic variables $s_1$and u_1, we get:

$Z = \frac{9}{8} + 2(1 - s_1 - u_1) - \frac{1}{8}u_1^2 = \frac{25}{8} - 2s_1 - 2u_1 - \frac{1}{8}u_1^2$

Computing partial derivative of Z with respect to $s_1$and u_1, we have:

$\frac{\partial Z}{\partial s_1} = -2$; $\frac{\partial Z}{\partial u_1} = -2 - \frac{u_1}{4}$

But the current solution, we have:

$\left.\frac{\partial Z}{\partial s_1}\right|_{\substack{s_1=0\\u_1=0}} = -2$ and $\left.\frac{\partial Z}{\partial u_1}\right|_{\substack{u_1=0\\x_1=0}} = -2$

Since both $\alpha_j <$ (j = 1, 2), the optimal solution is: $x_1 = 1$, $x_2 = \frac{3}{4}$ and Max Z = 25/8.

Exercise

1. Use Wolfe's method to solve the quadratic programming problem:

   Maximize $Z = 2x_1 + x_2 - x_1^2$

   subject to the constraint

   (i)$2x_1 + 3x_2 \leq 6$(ii)$2x_1 + x_2 \leq 4$ and $x_1, x_2 \geq 0$.

2. Use Beale's method to solve following quadratic programming problem:

Maximize $Z = -4x_1 + x_1^2 - 2x_1x_2 + 2x_2^2$

subject to the constraints

(i) $2x_1 + x_2 \geq 6$ (ii) $x_1 - 4x_2 \geq 0$ and $x_1, x_2 \geq 0$

4.9 Summary

Linear programming required the objective function and constrains to be linear. However, if either of these are not linear, then non-linear programming methods are used to find optimal value of the objective function with or without constraints. In the more general procedure, conditions necessary for an optimum value of a function subject to inequality constraints, are known as Kuhn-Tucker conditions. Beale's and Wolf's methods have also been demonstatrated to solve quadratic programming problems.

In case the objective function and constrains are separable, the separable programming technique is used for solving a NL programming problem. Sometimes, functions that are not separable by using the approximation methods.

Geometric programming is used to solve NL programming problems that involve special type of functions called polynomials. The GP approach first finds the optimal value of the objective function by solving its dual problem and then determines the solution to the given NLP problem from the optimal solution of the dual.

If true values of the LP model parameters are not known, then in such a case stochastic programming approach is used to solve the LP model by making a few decisions by selecting model parameters at different points in time. This is done to consider random effects on the parameters explicitly in the solution of the model.

References

1. Linear Programming – G. Hadley.
2. Mathematical Programming Techniques – N. S. Kambo.
3. Nonlinear and Dynamic Programming – G. Hadley.
4. Operations Research – K. Swarup, P. K. Gupta and Man Mohan.
5. Operations Research – H. A. Taha.
6. Operations Research – S. D. Sharma.

POST GRADUATE DEGREE PROGRAMME (CBCS)

M.SC. IN MATHEMATICS

SEMESTER I

SELF LEARNING MATERIAL

PAPER: DSE 1.4
(Pure Stream)

Differential Geometry I

Topology I

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

Course Preparation Team	
Dr. Avijit Sarkar, Associate Professor, Department of Mathematics, University of Kalyani	Dr. Dibyendu De, Professor, Department of Mathematics, University of Kalyani
Dr. Biswajit Mallick, Assistant Professor (Cont), DODL, University of Kalyani	Ms. Audrija Choudhury, Assistant Professor (Cont), DODL, University of Kalyani

Dec 2021

Directorate of Open and Distance Learning, University of Kalyani

Published by the Directorate of Open and Distance Learning

University of Kalyani, 741235, West Bengal

SYLLABUS

DSE 1.4 (Pure Stream)

Marks: 100; Credits: 6

Unit	Topic	Counselling Duration
Block I: Differential Geometry I; Marks 50 (SEE: 40; IA: 10)		
1	Vector valued functions, Directional Derivatives, Total derivatives,	54 Mins
2	Statement of Inverse and Implicit Function Theorems, Curvilinear coordinate system in E3.	54 Mins
3	Reciprocal base system. Riemannian space. Reciprocal metric tensor, Christoffel symbols, Covariant differentiation of vectors and tensors of rank 1 and 2.	54 Mins
4	Riemannian curvature tensor, Rieci tensor and scalar curvature. Space of constant curvature, Einstein space	54 Mins
5	On the meaning of covariant derivative. Intrinsic differentiation. Parallel vector field.	54 Mins
6	Tensor Algebra on finite dimensional vector spaces, Inner product spaces, matrix representation of an inner product ,	54 Mins
7	Linear functional, r-forms, Exterior product, Exterior derivative	54 Mins
8	Regular curves, curvature, torsion, curves in plane, signed curvature, curves in spaces,	54 Mins
9	Serret Frenet formulae, Isoperimetric inequality, four vertex theorem	54 Mins
10	Introduction to surface, Definition example, first fundamental form of surfaces	54 Mins
Block II: Topology I; Marks 50 (SEE: 40; IA: 10)		
11	Definition and examples of topological spaces.	54 Mins
12	Basis for a given topology, necessary and sufficient condition for two bases to be equivalent,	54 Mins
13	Sub-base, topologizing of two sets from a sub base	54 Mins
14	Closed sets, closure and interior, their basic properties and their relations	54 Mins
15	Neighbourhoods, exterior and boundary, dense sets. Accumulation points and derived sets. Subspace topology	54 Mins
16	Continuous, open, closed mappings, examples and counter examples,	54 Mins
17	Their different characterizations and basic properties	54 Mins
18	Pasting lemma, homeomorphism, topological properties.	54 Mins
19	The countability axioms, Separation axioms	54 Mins

20	Urysohn's lemma and Tietze extension theorem (Statements only) and some of their applications.	54 Mins
Total		18 Hours

Block I
Differential Geometry I

Units 1 & 2

Course structure

- Differentiability of a map from R^2 to R^3
- Inverse Function Theorem

Objective

The object of this unit is to give the idea that what differential geometry is and what are needed to study differential geometry.

1 Introduction

Differential geometry is a subject, where we study geometric properties with the help of calculus. To have the idea of this subject, we need some basic concept of calculus of several variables. We also need some knowledge of basic linear algebra.

1.1 Definition

Let f be a map from an open set of R to R^n given by

$$f(t) = (f_1(t), f_2(t), \cdots, f_n(t))$$

The map will be called continuous at a point if the each component $f(t) = (f_1(t), f_2(t), \cdots, f_n(t)$ is continuous at that point. Similarly differentiability of f is determined by the differentiability of the components.

1.2 Differentiability of a map from R^2 to R^3

There is a precise definition of differentiability of valued maps. For details, readers may consult some standard books of calculus of several variables. In the following, we give an working idea of differentiability of maps from $R^2 \to R^3$.

Let $f : U \to R^3$, U is open in R^2 be given by

$$f(x, y) = (f_1(x, y), f_2(x, y), f_2(x, y))$$

Now

$$J(f) = \begin{bmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} \\ \frac{\partial f_3}{\partial x} & \frac{\partial f_3}{\partial y} \end{bmatrix}$$

The map f will be called differentiable at a point of U, if al that point $J(f)$ is maximal rank.

1.3 Inverse function theorem

Let $f : U \to R^n$ be a smooth map defined on an open subset U of R^n $(n \geq 1)$. Assume that at some point $x_0 \in U$, the Jacobian matrix $J(f)$ is invertible. Then there is an open subset V of R^n and a smooth map $g : V \to R^n$ such that

(i) $y_0 = f(x_0) \in V$

(ii) $g(y_0) = x_0$

(iii) $g(V) \subseteq U$

(iv) $g(V)$ is an open subset of R^n

(v) $f(g(y)) = y$ for all $y \in V$.

In particular, $g : V \to g(V)$ and $f : g(V) \to V$ are inverse bijection.

Thus the inverse function theorem says that, if $J(f)$ is invertible at some point, then f is bijective near that point and its inverse map is smooth.

1.4 Summary

In this unit, we have learnt about the continuity of vector-valued functions and their differentiability dtermined by the conitnuity and differentiability of the component functions. Also, we learnt the Inverse function theorem.

Units 3 & 4

Course structure

• Parametric curves

Objective

The object of this unit is to study parametric curves and their reparametrization.

2 Introduction

In our undergraduate classes, we read about the curvature of curves in two dimensional plane R^2. Here we shall study curvature of curves in R^2 and curvature and torsion of curves in R^3. In this chapter, we shall use parametric equation of curves. We know that parametric equation of the circle $x^2 + y^2 = 1$ is $x = cos\ t,\ y = sin\ t$. In analytical notation, we represent a curve in plane by a mapping from an open interval of R to R^2. For instance, we express a circle by a map $\gamma : [0, 2\pi] \to R^2$ by

$$\gamma(t) = (cos\ t, sin\ t).$$

Now if we take $(0, 2\pi)$ instead of $[0, 2\pi]$, we get the circle excluding the point $(0, 1)$.

Since, we shall use analysis to study curves, we have to use concept of differentiability. So we prefer $(0, 2\pi)$ instead of $[0, 2\pi]$. In the following we give formal definition of parametric curves.

2.1 Parametrized curve

A parametrized curve in R^n is a map $\gamma : (\alpha, \beta) \to R^n$, for some $\alpha,\ \beta$ with $-\infty \leq \alpha < \beta \leq \infty$.

The symbol (α, β) denotes the open interval

$$(\alpha, \beta) = \{t \in R : \alpha < t < \beta$$

A parametrized curve whose image is contained in a level curve C is called a parametrization of part of C.

2.2 Smooth curve

Let $\gamma : (\alpha, \beta) \to R^n$ be given by

$$\gamma(t) = \Big(\gamma_1(t), \gamma_2(t), \cdots, \gamma_n(t)\Big)$$

The curve γ will be called smooth if each of the components $\gamma_1, \gamma_2, \cdots, \gamma_n$ of γ is smooth.

2.3 Tangent to a curve

If $\gamma(t)$ is a parametrized curve, its derivative $\frac{d\gamma}{dt}$ is called the tangent vector of γ at the point $\gamma(t)$.

2.4 Arc length of a parametrized curve

The arc length of a curve γ starting at the point $\gamma(t_0)$ is the function $s(t)$ given by

$$s(t) = \int_0^t ||\dot{\gamma}(n)||dn$$

For logarithmic spiral

$$\gamma(t) = (e^t cos\ t,\ e^t sint)$$

We have

$$\begin{aligned} \dot{\gamma} &= \Big(e^t(cos\ t - sin\ t),\ e^t(sin\ t + cos\ t)\Big) \\ ||\dot{\gamma}|| &= e^{2t}(cos\ t - sin\ t)^2 + e^{2t}(sin\ t + cos\ t)^2 \\ &= 2e^{2t} \end{aligned} \tag{2.1}$$

Hence the arc length of γ starting at $\gamma(0) = (1, 0)$ is

$$\begin{aligned} s &= \int_0^t \sqrt{2e^{2n}dn} \\ &= \sqrt{2}(e^t - 1) \end{aligned} \tag{2.2}$$

Note: If s is the arc length of a curve γ starting at $\gamma(t_0)$, we have

$$\begin{aligned} \frac{ds}{dt} &= \frac{d}{dt}\int_{t_0}^t ||\gamma(u)||du \\ &= ||\dot{\gamma}(t)|| \end{aligned}$$

Thinking of $\gamma(t)$ as the position of a moving point at time t, $\frac{ds}{dt}$ is the speed of the point (rate of change of distance along the curve). For this reason, we make the following definition.

2.5 Speed of curve

If $\gamma : (\alpha, \beta) \to R^n$ is parametrized curve, its speed at that point $\gamma(t)$ is $||\dot{\gamma(t)}||$, and γ is said to be a unit speed curve if $||\dot{\gamma}(t)||$ is a unit vector for all $t \in (\alpha, \beta)$.

2.6 Proposition

Let $n(t)$ be a unit vector that is a smooth function of a parameter t. Then the dot product $\dot{n}(t) \cdot \dot{n}(t) = 0$ for all t, i.e., $\dot{n}(t)$ is zero or perpendicular to $n(t)$ for all t.

In particular, if γ is a unit speed curve, then $\ddot{\gamma}$ is zero or perpendicular to $\dot{\gamma}$.

Proof: We use the product formula for differentiating dot products of vector valued functions $a(t)$ and $b(t)$

$$\frac{d}{dt}(a \cdot b) = \frac{da}{dt} \cdot b + a \cdot \frac{db}{dt}$$

Using this to differentiate both sides of the equation $n \cdot n = 1$, with respect to t gives

$$\dot{n} \cdot n + n \cdot \dot{n} = 0$$

So $2\dot{n}n = 0$. The last part follows by taking $n = \dot{\gamma}$. We know that parametric equation of a curve is not unique. So we can parametrize a curve in several ways but the curve is same. In the following, we give the definition of reparametrization.

2.7 Reparametrization

A parametrized curve $\widetilde{\gamma} : (\widetilde{\alpha}, \widetilde{\beta}) \to R^n$ is a parametrization of a parametrized curve $\gamma : (\alpha, \beta) \to R^n$ if there is a smooth bijective map $\phi : (\widetilde{\alpha}, \widetilde{\beta}) \to (\alpha, \beta)$ such that the inverse map

$$\phi^{-1} : (\alpha, \beta) \to (\widetilde{\alpha}, \widetilde{\beta})$$

is also smooth and

$$\widetilde{\gamma}(\widetilde{t}) = \gamma\Big(\phi(\widetilde{t})\Big) \quad \text{for all} \quad \widetilde{t} \in (\widetilde{\alpha}, \widetilde{\beta})$$

Two curves that are reparametrizations of each other have the same image, so they should have the same geometric properties.

2.8 Proposition

Any reparametrization of a regular curve is regular.

Proof: Suppose that γ and $\widetilde{\gamma}$ are related as in definition 2.7. Let $t = \phi(\widetilde{t})$ and let $\psi = \phi^{-1}$ so that $\widetilde{t} = \psi(t)$. Differentiating both sides of the equation $\phi\Big(\psi(t)\Big) = t$ with respect to t and using the chain rule gives

$$\frac{d\phi}{d\widetilde{t}} \cdot \frac{d\psi}{dt} = 1$$

This shows that $\frac{d\phi}{d\widetilde{t}}$ is never zero. Since $\widetilde{\gamma}(\widetilde{t}) = \gamma\Big(\phi(\widetilde{t})\Big)$, another application of chain rule gives

$$\frac{d\widetilde{\gamma}}{d\widetilde{t}} = \frac{d\gamma}{dt} \cdot \frac{d\phi}{d\widetilde{t}} \tag{2.3}$$

which shows that $\frac{d\widetilde{\gamma}}{d\widetilde{t}}$ is never zero if $\frac{d\gamma}{dt}$ is never zero.

2.9 Observation

See that $f : R^2 - \{(0,0)\} \to R$ defined by $f(x,y) = \sqrt{x^2 + y^2}$ is smooth in $R^2 - (0,0)$.

2.10 Proposition

If $\gamma(t)$ is regular curve, its arc length s starting at any point γ is a smooth function of t.

Proof: We have already seen that s is a differentiable function of t and $\frac{ds}{dt} = ||\dot{\gamma}(t)||$. To simplify the notation, assume from now on that γ is plane curve, say

$$\gamma(t) = \Big(u(t), v(t)\Big), \tag{2.4}$$

where u and v are smooth functions of t. Define $f : R^2 \to R$ by

$$f(u, v) = \sqrt{u^2 + v^2}$$

so that

$$\frac{ds}{dt} = f(\dot{u}, \dot{v}) \tag{2.5}$$

The crucial point is that f is smooth on $R^2 - \{(0, 0)\}$, which means that all the partial derivatives of f of all orders exist and are continuous functions except at the origin $(0, 0)$. For example

$$\frac{\partial f}{\partial u} = \frac{u}{\sqrt{u^2 + v^2}}, \quad \frac{\partial f}{\partial v} = \frac{v}{\sqrt{u^2 + v^2}}$$

are well defined and continuous expect where $u = v = 0$ and similarly for higher derivatives. Since γ is regular $\dot{u}$ and $\dot{v}$ are never both zero. So the chain rule and equation (2.5) shows that $\frac{ds}{dt}$ is smooth.

For example,

$$\frac{d^2 s}{dt^2} = \frac{\partial f}{\partial u}\ddot{u} + \frac{\partial f}{\partial v}\ddot{v},$$

and similarly for the higher order derivatives of s.

2.11 Proposition

A parametrized curve has unit speed reparametrization if and only if it is regular.

Proof: Suppose first that a parametrized curve $\gamma : (\alpha, \beta) \to R^n$ has a unit speed reparametrization $\widetilde{\gamma}$, with reparametrization map ϕ. Letting $t = \phi(\widetilde{t})$, we have

$$\begin{aligned} \widetilde{\gamma}(\widetilde{t}) &= \gamma(t) \\ \Rightarrow \frac{d\widetilde{\gamma}}{d\widetilde{t}} &= \frac{d\gamma}{dt} \cdot \frac{dt}{d\widetilde{t}} \\ \therefore \left|\left|\frac{d\widetilde{\gamma}}{d\widetilde{t}}\right|\right| &= \left|\left|\frac{d\gamma}{dt}\right|\right| \cdot \left|\left|\frac{dt}{d\widetilde{t}}\right|\right| \end{aligned}$$

Since $\widetilde{\gamma}$ is unit speed, $\left|\left|\frac{d\widetilde{\gamma}}{d\widetilde{t}}\right|\right| = 1$, so clearly $\frac{dt}{d\widetilde{t}}$ is not zero.

Conversely, suppose that the tangent vector $\frac{d\gamma}{dt}$ is never zero. By the note 2.4 $\frac{ds}{dt} > 0$ for all t, where s is the arc length of γ starting at any point of the curve and by proposition 2.10 s is smooth function of t. It follows from inverse function theorem of multivariable calculus that $s : (\alpha, \beta) \to R$ is injective, that its image is an open interval $(\widetilde{\alpha}, \widetilde{\beta})$ and that the inverse map $s^{-1} : (\widetilde{\alpha}, \widetilde{\beta}) \to (\alpha, \beta)$ is smooth. We take $\phi = s^{-1}$ and let $\widetilde{\gamma}$ be the corresponding reparametrization of γ, so that

$$\begin{aligned}
&\widetilde{\gamma}(s) = \gamma(t) \\
&\Rightarrow \frac{d\widetilde{\gamma}}{ds} \cdot \frac{ds}{dt} = \frac{d\gamma}{dt} \\
&\Rightarrow \left\| \frac{d\widetilde{\gamma}}{ds} \right\| \cdot \frac{ds}{dt} = \left\| \frac{d\gamma}{dt} \right\| = \frac{ds}{dt} \\
&\therefore \left\| \frac{d\widetilde{\gamma}}{ds} \right\| = 1
\end{aligned}$$

The above proof shows that the arc length is essentially the only unit speed parameter on a regular curve.

2.12 Exercise

(i) Show that $\gamma(t) = \left(cos^2 t - \frac{1}{2}, sin\, t\, cos\, t,\ sin\, t \right)$ is a parametrization of the curve of intersection of the circular cylinder of radius $\frac{1}{2}$ and axis the z-axis with the sphere of radius 1 and centre $(\frac{1}{2}, 0, 0)$.

(ii) Show that the curve γ given by

$$\gamma(t) = \left(\frac{1}{3}(1+t)^{3/2}, \frac{1}{3}(1-t)^{3/2}, \frac{t}{\sqrt{2}} \right)$$

is unit speed.

(iii) Test whether the following curve regular

$$\gamma(t) = \left(cos^2 t, sin^2 t \right) \text{ for } -\infty < t < \infty$$

2.13 Summary

In this unit, we have known about parametrization of a curve in R^n. Reparametrization is explained. Necessary and sufficient condition for reparametrization is proved.

Suggested Readings: Elementary differential geometry by Andrew Pressley.

Units 5 & 6

Course structure

• Curvature of curves

Objective

The object of this present unit is to give the idea of curvature of plane curves, signed curvature of plane curves and curvature and torsion of space curves.

3 Introduction

In our undergraduate classes, we have studied about curvature of plane curves. We know that the curvature of a circle in plane is constant.

In this unit we shall study how to find curvature of plane and space curves.

Recall that a curve γ parametrized by t is called a unit speed curve if $||\dot{\gamma}(t)|| = 1$.

We also have seen in the previous unit that if we reparametrize a curve by arc length, the curve becomes unit speed. In the following we shall give two definitions of curvature of curves. One for unit speed curve parametrized by arc length and another for any regular curve.

3.1 Curvature for unit speed curve parametrized by arc length

If a curve γ is unit speed and is parametrized by arc length s, then its curvature $k(s)$ is defined as

$$k(s) = \left|\left|\frac{d^2\gamma(s)}{ds^2}\right|\right|.$$

Observe that $\frac{d\gamma(s)}{ds}$ denotes the tangent to the curve γ at $\gamma(s)$. Hence second derivative of γ denotes the rate of change of tangent. Thus the definition conforms with our usual sense of curvature as a rate of change of tangents. Again there are some other reasoning to define curvature like the above definition. For details see any standard book of Differential Geometry of curves and surfaces.

However, to give unit speed reparametrization to a curve using arc length is theoretically true but practically some times it become a hard work. So there is another definition of curvature for regular curves.

3.2 Curvature of regular curve

Let $\gamma(t)$ be a regular curve in R^3, then its curvature $k(t)$ is defined as

$$k(t) = \frac{||\ddot{\gamma} \times \dot{\gamma}||}{||\dot{\gamma}||^3}$$

where the $\times$ denotes the vector product and $\cdot$ denote $\frac{d}{dt}$.

3.3 The definition 3.2 can be deduced from 3.1 as following:

Let $\widetilde{\gamma}$ (with parameter s) be a unit speed reparametrization of γ, and let us denote $\frac{d}{ds}$ by a dash ($'$). Then by chain rule

$$\widetilde{\gamma}' \frac{ds}{dt} = \dot{\gamma}$$

so

$$\begin{aligned} k = ||\widetilde{\gamma}''|| &= \left|\left| \frac{d}{ds}\left(\frac{\dot{\gamma}}{\frac{ds}{dt}}\right) \right|\right| \\ &= \left|\left| \frac{\frac{d}{dt}\left(\frac{\dot{\gamma}}{\frac{ds}{dt}}\right)}{\frac{ds}{dt}} \right|\right| \\ &= \left|\left| \frac{\ddot{\gamma}\frac{ds}{dt} - \dot{\gamma}\frac{d^2s}{dt^2}}{\frac{ds}{dt}} \right|\right| \end{aligned} \tag{3.1}$$

Now

$$\left(\frac{ds}{dt}\right)^2 = ||\dot{\gamma}||^2 = \dot{\gamma} \cdot \dot{\gamma}$$

and differentiating with respect to t gives

$$\frac{ds}{dt} \cdot \frac{d^2s}{dt^2} = \dot{\gamma} \cdot \ddot{\gamma}$$

Using this and above equation (3.1), we get

$$\begin{aligned} k &= \left|\left| \frac{\ddot{\gamma}\left(\frac{ds}{dt}\right)^2 - \dot{\gamma}\frac{d^2s}{dt^2}\frac{ds}{dt}}{\left(\frac{ds}{dt}\right)^4} \right|\right| \\ &= \frac{||\ddot{\gamma}(\dot{\gamma} \cdot \dot{\gamma}) - \dot{\gamma}(\dot{\gamma} \cdot \ddot{\gamma})||}{||\dot{\gamma}||^4} \end{aligned}$$

Using the vector triple product identity

$$a \times (b \times c) = (a \cdot c)b - (a \cdot b)c, \quad (a, b, c \in R^3)$$

we get

$$\dot{\gamma} \times (\ddot{\gamma} \times \dot{\gamma}) = \dddot{\gamma}(\dot{\gamma} \cdot \dot{\gamma}) - \dot{\gamma}(\dot{\gamma} \cdot \ddot{\gamma})$$

Again $\dot{\gamma}$ and $\ddot{\gamma} \times \dot{\gamma}$ are perpendicular vectors so

$$||\dot{\gamma} \times (\ddot{\gamma} \times \dot{\gamma})|| = ||\dot{\gamma}||\, ||\ddot{\gamma} \times \dot{\gamma}||,$$

Hence

$$\frac{||\dddot{\gamma}(\dot{\gamma} \cdot \dot{\gamma}) - \dot{\gamma}(\dot{\gamma} \cdot \ddot{\gamma})||}{||\dot{\gamma}||^4} = \frac{||\dot{\gamma} \times (\ddot{\gamma} \times \dot{\gamma})||}{||\dot{\gamma}||^4} = \frac{||\ddot{\gamma} \times \dot{\gamma}||}{||\dot{\gamma}||^3}$$

If γ is non-regular curve its curvature is not defined, in general.

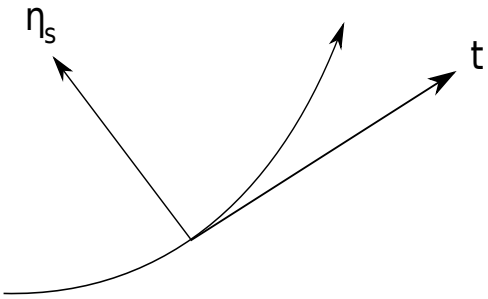
3.4 Concept of signed curvature

For plane curves, it is possible to refine the definition of curvature slightly and give it an appealing geometric interpretation.

Suppose that $\gamma(s)$ is a unit speed curve in R^2. Denoting $\frac{d}{ds}$ by a dot, let

$$t = \dot{\gamma}$$

be the tangent vector of γ; note that t is a unit vector. There are two unit vectors perpendicular to t; we make a choice by defining η_s, the signed unit normal of γ, to be the unit vector obtained by rotating t anticlockwise by $\frac{\pi}{2}$.



We know that $\dot{t} = \ddot{\gamma}$ is perpendicular to t, and hence parallel to η_s. Thus, there is a number k_s such tat

$$\ddot{\gamma} = k_s \eta_s.$$

The scalar k_s is called the signed curvature of γ (it can be positive, negative or zero).

Note that $||\eta_s|| = 1$, we have

$$k = ||\ddot{\gamma}|| = ||k_s \eta_s|| = |k_s|,$$

so the curvature of γ is the absolute value of its signed curvature.

3.5 Proposition

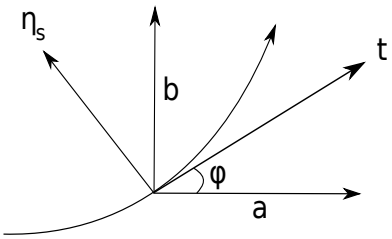
Let $\gamma(s)$ be a unit speed plane curve, and let $\phi(s)$ be the angle through which a fixed unit vector must be rotated anticlockwise to bring it into coincidence with the unit tangent vector t of γ. Then

$$k_s = \frac{d\phi}{ds}.$$

Proof: Let a be the fixed unit vector and let b be the unit vector obtained by rotating a anticlockwise by $\frac{\pi}{2}$. Then,

$$\begin{aligned} t &= a\,cos\,\phi + b\,sin\,\phi, \\ \Rightarrow \dot{t} &= (-a\,sin\,\phi + b\,cos\,\phi)\frac{d\phi}{ds} \\ \therefore \dot{t}.a &= -a\,sin\,\phi\frac{d\phi}{ds} \\ \therefore k_s(\eta_s.a) &= -sin\,\phi\frac{d\phi}{ds} \end{aligned}$$

But the length between η_s and a is $\phi + \frac{\pi}{2}$, since t must be rotated anticlockwise by $\frac{\pi}{2}$ to bring it into coincidence with η_s.



Hence

$$\eta_s.a = cos\,(\phi + \frac{\pi}{2}) = -sin\,\phi.$$

Hence

$$k_s = \frac{d\phi}{ds}.$$

3.6 Theorem

Let $\kappa : (\alpha, \beta) \to R$ be any smooth function. Then, there is a unit speed curve $\gamma : (\alpha, \beta) \to R^2$ whose signed curvature is κ.

Further, if $\widetilde{\gamma} : (\alpha, \beta) \to R^2$ is any other unit speed curve whose signed curvature is κ, there is a rigid motion M of R^2 such that

$$\widetilde{\gamma}(s) = M(\gamma(s)) \quad \text{for all } s \in (\alpha, \beta)$$

Proof: For the first part, fix $s_0 \in (\alpha, \beta)$ and define, for any $s \in (\alpha, \beta)$,

$$\gamma(s) = \left(\int_{s_0}^{s} cos\,\phi(v)dv, \int_{s_0}^{s} sin\,\phi(v)dv \right),$$

where $\phi(s) = \int_{s_0}^{s} k(u)du$.

Then, the tangent vector of γ is

$$\dot{\gamma}(s) = (cos\,\phi(s), sin\,\phi(s)),$$

which is a unit vector making an angle $\phi(s)$ with the x-axis. Then γ is a unit speed and , by proposition 3.5 its signed curvature is

$$\frac{d\phi}{ds} = \frac{d}{ds}\int_{s_0}^{s} k(u)du = k(s)$$

For second part, let $\widetilde{\phi}(s)$ be the angle between the x-axis and the unit tangent $\dot{\widetilde{\gamma}}$ of $\widetilde{\gamma}$. Thus

$$\begin{aligned}\dot{\widetilde{\gamma}}(s) &= (cos\ \widetilde{\phi}(s), sin\ \widetilde{\phi}(s)),\\ \therefore \widetilde{\gamma}(s) &= \left(\int_{s_0}^{s} cos\ \widetilde{\phi}(v)dv, \int_{s_0}^{s} sin\ \widetilde{\phi}(v)dv\right) + \widetilde{\gamma}(s_0)\end{aligned}$$

By previous result, we have

$$\begin{aligned}\frac{d\widetilde{\phi}}{ds} &= k(s)\\ \therefore \widetilde{\phi}(s) &= \int_{s_0}^{s} k(u)du + \widetilde{\phi}(s_0) = \phi(s) + \widetilde{\phi}(s_0)\end{aligned}$$

Form the above equations and writing a for the constant vector $\widetilde{\gamma}(s_0)$ and θ for the constant scalar $\widetilde{\phi}(s_0)$, we get

$$\begin{aligned}\widetilde{\gamma}(s) &= T_a\left(\int_{s_0}^{s} cos\ (\phi(v)+\theta)dv, \int_{s_0}^{s} sin\ (\phi(v)+\theta)dv\right)\\ &= T_a\left(cos\ \theta\int_{s_0}^{s} cos\ \phi(v)dv - sin\ \theta\int_{s_0}^{s} sin\ \phi(v)dv,\ sin\ \theta\int_{s_0}^{s} cos\ \phi(v)dv + cos\ \theta\int_{s_0}^{s} sin\ \phi(v)dv\right)\\ &= T_aR_\theta\left(\int_{s_0}^{s} cos\ \phi(v)dv, \int_{s_0}^{s} sin\ \phi(v)dv\right)\\ &= M(\gamma(s))\end{aligned}$$

3.7 Space curves

In the previous section, we have seen that a plane curve is completely determined by signed curvature. But this is not true for space curves. To describe space curves, we need a space curve we need two curvatures. One is called simply curvature and the second one is called torsion.

We shall prove that curvature and torsion completely determine the curve in space.

Let $\gamma(s)$ be a unit speed curve in R^3, and let $t = \dot{\gamma}$ be its unit tangent vector. If the curvature $\kappa(s)$ is non-zero, we define the principal normal of γ at the point $\gamma(s)$ to be the vector

$$\eta(s) = \frac{1}{\kappa(s)}\dot{t}(s)$$

Since $||\dot{t}|| = \kappa$, η is unit vector. Further we know $\dot{t}\cdot\dot{t} = 0$. So t is perpendicular to η. It follows that

$$t \times \eta = b \quad \text{(say)}$$

is a unit vector perpendicular to both t and η. The vector $b(s)$ is called binormal vector of γ at the point $\gamma(s)$. Thus $\{t, \eta, b\}$ is an orthonormal basis of R^3 and is right handed, i.e.,

$$b = t \times \eta, \quad \eta = b \times t, \quad t = \eta \times b$$

Since $b(s)$ is unit vector for all s, b is perpendicular to b. Now we use the product rule for differentiating the vector product of vector valued functions u and v of a parameter s

$$\frac{d}{ds}(u \times v) = \frac{du}{ds} \times v + u \times \frac{dv}{ds}$$

Applying this to $b = t \times \eta$ gives

$$\begin{aligned} \dot{b} &= \dot{t} \times \eta + t \times \dot{\eta} \\ &= t \times \dot{\eta} \end{aligned}$$

Since by definition of η,

$$\dot{t} \times \eta = \kappa(\eta \times \eta) = 0.$$

Hence $\dot{b}$ is perpendicular to t. Being perpendicular to both t and b, $\dot{b}$ must be parallel to η. So

$$\dot{b} = -\tau\eta,$$

for some scalar τ, which is called the torsion of γ. Note that the torsion is only defined if the curvature is non-zero.

Of course, we define the torsion of an arbitrary regular curve γ to be the torsion of a unit speed reprametrization of γ of the form

$$u = \pm s + c$$

where c is a constant, then τ is unchanged. But this change of parameter has the following effect on the vectors introduced above: $t \to \pm t$, $\dot{t} \to \dot{t}$, $\eta \to \eta$, $b \to \pm b$, $\dot{b} \to b$.

Hence $\tau \to \tau$.

3.8 Proposition

Let $\gamma(t)$ be a regular curve in R^3 with nowhere vanishing curvature. Then denoting $\frac{d}{dt}$ by a dot, its torsion is given by

$$\tau = \frac{(\dot{\gamma} \times \ddot{\gamma}) \cdot \dddot{\gamma}}{||\dot{\gamma} \times \ddot{\gamma}||^2}$$

3.9 Example

Compute the torsion of the circular helix

$$\gamma(\theta) = (a\ cos\ \theta, a\ sin\ \theta, b\theta)$$

Solution:

$$\begin{aligned} \dot{\gamma}(\theta) &= (-a\ sin\ \theta, a\ cos\ \theta, b), \\ \ddot{\gamma}(\theta) &= (-a\ cos\ \theta, -a\ sin\theta, 0), \\ \dddot{\gamma}(\theta) &= (-a\ sin\ \theta, -a\ cos\ \theta, 0) \end{aligned}$$

Hence

$$\dot{\gamma} \times \ddot{\gamma} = (ab\ sin\ \theta, -ab\ cos\ \theta, a^2)$$
$$\therefore ||\dot{\gamma} \times \ddot{\gamma}||^2 = a^2(a^2 + b^2)$$
$$\therefore (\dot{\gamma} \times \ddot{\gamma})\dddot{\gamma} = a^2 b$$

and so the torsion

$$\begin{aligned} \tau &= \frac{(\dot{\gamma} \times \ddot{\gamma})\dddot{\gamma}}{||\dot{\gamma} \times \ddot{\gamma}||^2} \\ &= \frac{a^2 b}{a^2(a^2 + b^2)} \\ &= \frac{b}{a^2 + b^2} \end{aligned}$$

3.10 Theorem

Let γ be a unit speed curve in R^3 with nowhere vanishing curvature. Then

$$\begin{aligned} \dot{t} &= \kappa\eta \\ \dot{\eta} = -\kappa t + \tau b & \\ \dot{b} = -\tau\eta & \end{aligned}$$

3.11 Theorem

Let $\gamma(s)$ and $\widetilde{\gamma}(s)$ be two unit speed curves in R^3 with the same curvature $\kappa(s) > 0$ and the same torsion $\tau(s)$ for all s. Then, there is a rigid motion M of R^3 such that

$$\widetilde{\gamma}(s) = M\Big(\gamma(s)\Big) \quad \text{for all } s$$

Further, if κ and τ are smooth functions with $\kappa > 0$ everywhere, there is a unit speed curve in R^3 whose curvature is κ and whose torsion is τ.

3.12 Exercise

(i) Show that a circle with centre (x_0, y_0) and radius R has curvature $\frac{1}{R}$.

(ii) Compute the curvature of the curve

$$\gamma(t) = \left(\frac{1}{3}(1+t)^{3/2}, \frac{1}{3}(1-t)^{3/2}, \frac{t}{\sqrt{2}}\right)$$

(iii) Show that, if the curvature $\kappa(t)$ of a regular curve $\gamma(t) > 0$ everywhere, then $\kappa(t)$ is a smooth function of t. Give an example to show that this may not be the case without the assumption that $\kappa > 0$.

(iv) Prove that signed curvature is a smooth function.

(v) Prove that any regular plane curve whose curvature is a positive constant is part of a circle.

3.13 Summery

In this unit we have studied about curvature and torsion of curves in R^3. We have studied about signed curvature of plane curves. We observed that for a plane curves, signed curvature dtermine the curve completely and a space curve is determined by curvature and torsion.

3.14 Suggested Reading

Elementary Differential Geometry by Andrew Pressley.

Units 7 & 8

Course structure

• Global properties of plane curves

Objective

The object of this present unit is to study some global properties of plane curves. We mainly study isoperimetric inequality and four vertex theorem.

4 Introduction

In the previous chapters, we have studied some properties of curves which depend on a point of the curve. These properties are called local properties. There are some properties which depend on the total shape of the curve. These properties are known as global properties.

4.1 Simple closed curve

Let $a \in R$ be a positive constant. A simple closed curve in R^2 with period a is a (regular) curve $\gamma : R \to R^2$ such that $\gamma(t) = \gamma(t')$ if and only if $(t' - t) = \kappa\, a$ for some integer κ.

4.2 Length of simple closed curve

The length of a curve γ is defined as

$$l(\gamma) = \int_0^a ||\dot{\gamma}(t)||dt,$$

where a is period of the curve.

4.3 Orientation

If the signed normal of a curve points into the interior of the curve γ, then the curve γ is called positively oriented.

4.4 Note

In the next section, we shall be interested in the area contained by a single closed curve γ, i.e.,

$$\mathcal{A}\Big(int(\gamma)\Big) = \int\int_{int(\gamma)} dx\, dy$$

This can be computed by using Green's Theorem, which says that, for all smooth functions $f(x, y)$ and $g(x, y)$

$$\int\int_{int(\gamma)} \left(\frac{\partial g}{\partial x} - \frac{\partial f}{\partial y}\right) = \int_\gamma \Big(f(x, y)dx + g(x, y)dy\Big)$$

if γ is positively oriented simple closed curve.

4.5 Proposition

If $\gamma(t) = \Big(x(t), y(t)\Big)$ is a positively oriented simple closed curve in R^2 with period a, then

$$\mathcal{A}\Big(Int(\gamma)\Big) = \frac{1}{2}\int_0^a (x\dot{y} - y\dot{x})dt$$

Proof: Taking $f = -\frac{1}{2}y$, $g = \frac{1}{2}x$ in Green's Theorem, we get

$$\mathcal{A}\Big(Int(\gamma)\Big) = \frac{1}{2}\int_\gamma (x\ dy - y\ dx)$$

which proves the proposition.

4.6 Isoperimetric inequality

Let γ be a simple closed curve, let $l(\gamma)$ be its length and let $\mathcal{A}\Big(int(\gamma)\Big)$ be the area of its interior. Then

$$\mathcal{A}\Big(int(\gamma)\Big) \leq \frac{1}{4\pi}l(\gamma)^2,$$

with equality holding if and only if γ is a circle.

To prove the theorem, we need the following inequality:

4.7 Wirtinger Inequality

Let $F : [0, \pi] \mapsto R$ be a smooth function such that $F(0) = F(\pi) = 0$. Then

$$\int_0^\pi \left(\frac{dF}{dt}\right)^2 dt \geq \int_0^\pi F(t)^2 dt,$$

with equality holding if and only if $F(t) = A\sin t$ for all $t \in [0, \pi]$, where A is a constant.

Proof: Let

$$G(t) = \frac{F(t)}{\sin t}$$

Then denoting $\frac{d}{dt}$ by a dot as usual,

$$\begin{aligned}\int_0^\pi \dot{F}^2 dt &= \int_0^\pi (\dot{G}\sin t + G\cos t)^2 dt \\ &= \int_0^\pi \dot{G}^2 \sin^2 t dt + 2\int_0^\pi G\dot{G}\sin t\cos t dt + \int_0^\pi G^2\cos^2 t dt\end{aligned}$$

Integrating by parts,

$$\begin{aligned} 2\int_0^\pi G\dot{G}\sin t\cos t dt &= G^2\sin t\cos t\big|_0^\pi - \int_0^\pi G^2(\cos^2 t - \sin^2 t)dt \\ &= \int_0^\pi G^2(\sin^2 t - \cos^2 t)dt \end{aligned}$$

So,

$$\begin{aligned} \int_0^\pi \dot{F}^2 dt &= \int_0^\pi \dot{G}^2\sin^2 t dt + \int_0^\pi G^2(\sin^2 t - \cos^2 t)dt + \int_0^\pi G^2\cos^2 t dt \\ &= \int_0^\pi (G^2 + \dot{G}^2)\sin^2 t dt \\ &= \int_0^\pi F^2 dt + \int_0^\pi \dot{G}^2\sin^2 t dt \end{aligned}$$

and so,

$$\int_0^\pi \dot{F}^2 dt - \int_0^\pi F^2 dt = \int_0^\pi \dot{G}^2\sin^2 t dt$$

The integral on the right hand side is obviously greater than 0, and it is zero if and only if $\dot{G} = 0$ for all t, that is, if and only if $G(t)$ is equal to a constant, say A, for all t. Then $F(t) = A\sin t$, as required.

We now prove the isoperimetric inequality.

Proof: We start by making some assumptions about γ that will simplify the proof.

First, we can, if we wish, assume that γ is parametrized by arc length s. However, because of the π that appears in the statement, it turns out to be more convenient to assume that the period of γ is π. If we change the parameter of γ from s to

$$t = \frac{\pi s}{l(\gamma)} \tag{4.1}$$

the resulting curve is still simple closed, and has period π because when s increases by $l(\gamma)$, t increases by π. We shall therefore assume that γ is parametrized using the parameter t in (4.1) from now on.

For the second simplification, we note that both $l(\gamma)$ and $\mathcal{A}(\gamma)$ are unchanged if γ is subjected to a translation

$$\gamma(t) \to \gamma(t) + b$$

where b is any constant vector. Taking $b = -\gamma(0)$, we might as well assume that $\gamma(0) = 0$ to begin with, that is, we assume that γ begins and ends at the origin.

To prove the theorem, we shall calculate $l(\gamma)$ and $\mathcal{A}(int(\gamma))$ by using polar coordinates

$$x = r\cos\theta,\ y = r\sin\theta$$

Using the chain rule, it is easy to show that

$$\begin{aligned} \dot{x}^2 + \dot{y}^2 &= \dot{r}^2 + r^2\dot{\theta}^2 \\ x\dot{y} - y\dot{x} &= r^2\dot{\theta} \end{aligned}$$

where $\frac{d}{dt}$ is denoted by a dot. Then using equation (4.1),

$$\begin{aligned} \dot{r}^2 + r^2\dot{\theta}^2 &= \left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 \\ &= \left[\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2\right]\left(\frac{ds}{dt}\right)^2 \\ &= \frac{l(\gamma)^2}{\pi^2} \end{aligned}$$

Since

$$\left(\frac{dx}{ds}\right)^2 + \left(\frac{dy}{ds}\right)^2 = 1$$

further by $\mathcal{A}(int(\gamma)) = \frac{1}{2}\int_0^a (x\dot{y} - y\dot{x})dt$, we have,

$$\begin{aligned} \mathcal{A}(int(\gamma)) &= \frac{1}{2}\int_0^\pi (x\dot{y} - y\dot{x})dt \\ &= \frac{1}{2}\int_0^\pi r^2\dot{\theta}dt \end{aligned} \tag{4.2}$$

Now, to prove the theorem, we have to show that

$$\frac{l(\gamma)^2}{4\pi} - \mathcal{A}(int(\gamma)) \geq 0$$

with equality holding if and only if γ is a circle. By equation (4.2),

$$\int_0^\pi (\dot{r}^2 + r^2\dot{\theta}^2)dt = \frac{l(\gamma)^2}{\pi}$$

Hence using equation (4.2),

$$\begin{aligned} \frac{l(\gamma)^2}{4\pi} - \mathcal{A}(int(\gamma)) &= \frac{1}{4}\int_0^\pi (\dot{r}^2 + r^2\dot{\theta}^2)dt - \frac{1}{2}\int_0^\pi r^2\dot{\theta}^2 dt \\ &= \frac{1}{4}I \end{aligned}$$

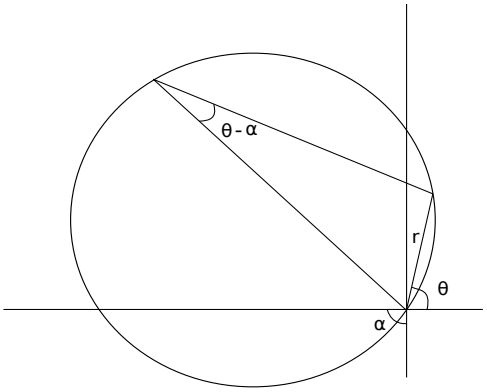
where,

$$I = \int_0^\pi (\dot{r}^2 + r^2\dot{\theta}^2 - 2r^2\dot{\theta})dt \tag{4.3}$$

Thus to prove the theorem, we have to show that, $I \geq 0$ and that $I = 0$ if and only if γ is a circle.

By simple algebra,

$$I = \int_0^\pi r^2(\dot{\theta} - 1)^2 dt + \int_0^\pi (\dot{r}^2 - r^2) dt \qquad (4.4)$$



The first integral on the right hand side of (4.4) is obviously greater than or equal to zero and hence the second integral is greater than or equal to zero by Wirtinger's inequality(we are taking $F = r$: note that $r(0) = r(\pi) = 0$, since $\gamma(0) = \gamma(\pi) = 0$). Hence, $I \geq 0$. Further, since both the integrals on the right hannd side of equation (4.4) are ≥ 0, their sum I is zero if and only if both of these integrals are zero. But the first integral is zero only if $\dot{\theta} = 1$ for all t, and the second is zero only if $r = A \sin t$ for some constant A(by Wirtinger inequality). So, $\theta = t + \alpha$, where α is a constant, and hence $r = A \sin(\theta - \alpha)$. It is easy to see that this is the polar equation of a circle of diameter A. Hence the proof is complete.

4.8 Convex Curve

A simple closed curve γ is called a convex curve if its interior $int(\gamma)$ is convex, in the usual sense that the straight line segment joining any two points of $int(\gamma)$ is contained entirely on $int(\gamma)$.

4.9 Vertex

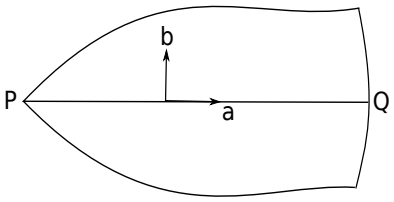
A vertex of a curve $\gamma(t)$ in R^2 is a point where its signed curvature κ_s has a stationary point, that is, where $\frac{d\kappa_s}{ds} = 0$.

4.10 Four Vertex Theorem

Every convex simple closed curve in R^2 has at least four vertices.

Proof:

We might as well assume that the curve $\gamma(t)$ is unit speed so that its period is the length l of γ. We consider the integral



$$\int_0^l \dot{\kappa}_s(t)\gamma(t)dt$$

where a dot denotes the derivative.

Integrating by parts and using the equation

$$\dot{\eta}_s = -\kappa_s t$$

we get

$$\begin{aligned}\int_0^l \dot{\kappa}_s \gamma dt &= -\int_0^l \kappa_s \dot{\gamma} dt \\ &= \int_0^l \kappa_s t dt \\ &= \int_0^l \dot{\eta}_s dt \\ &= \eta_s(l) - \eta_s(0) \\ &= 0 \end{aligned} \tag{4.5}$$

Now, κ_s attains all its values on the closed interval $[0, l]$, so κ_s must attain its maximum and minimum values at some points P and Q of γ, say. We can assume that $P \neq Q$. Since otherwise κ_s would be constant, γ would be a circle and every point of γ would be a vertex. Let a be a unit vector perallel to the vector PQ, and let b be the vector obtained by rotating a anticlockwise by $\pi/2$. Taking the dot product of the integral in equation (4.5) with constant vector b gives

$$\int_0^l \dot{\kappa}_s (\gamma.b) dt = 0 \tag{4.6}$$

Suppose that P and Q are the only vertices of γ. Since γ is convex, the straight line joining P and Q divides γ into two segments, and since there are no other vertices, we must have $\dot{\kappa}_s > 0$ on one segment and $\dot{\kappa}_s < 0$ on the other. But then the integrand on the left hand side of (4.6) is either always > 0 or always < 0 (except at P and Q where it vanishes), so the integral is definitely > 0 or < 0, a contradiction.

Hence there must be at least one more vertex, say R. If there are no other vertices, the points P , Q and R divide γ into three segments, on each of which, $\dot{\kappa}_s$ is either always > 0 or always < 0. But then $\dot{\kappa}_s$ must have the same sign on two adjacent segments. Hence there is a straight line that divides γ into two segments, on one of which $\dot{\kappa}_s$ is always positive, and on the other, $\dot{\kappa}_s < 0$. The argument in the preceeding paragraph shows that this is not possible. So there must be a fourth vertex.

4.11 Exercises

1. Show that the definition of a vertex of a plane curve is independent of its parametrization.

2. Find the signed curvature $\kappa(s)$ of the curve

$$\gamma(t) = (a\cos t, b\sin t)$$

Find at how many points $\frac{d\kappa(s)}{ds}$ vanishes. Hence verify four vertex theorem for this problem.

4.12 Summary

In this unit, we have learnt about the simple closed curves and their orientations leading up to isoperimetric inequality and Wirtinger Inequality and Four Vertex Theorem.

Units 9 & 10

Course structure

- Surface: Definition
- Transition map, Normal to surface, Orientable surface

Objective

The objective of the present unit is to study surfaces in R^3 in parametric form. The notion of tangent space and orientability have also been discussed with examples.

5 Introduction

In this chapter, we introduce several ways to formulate mathematically the notion of a surface.

5.1 What is a Surface?

A surface is a subset of R^3 that looks like a piece of R^2 in the vicinity of any given point, just as the surface of the earth, although actually nearly spherical, appears to be a flat plane to an observer on the surface who sees only to the horizon. To make the phrases 'looks like' and in the vicinity precise, we must first introduce some preliminary material. We describe this for R^n for any $n \geq 1$, although we shall need it only for $n = 1, 2, \ or \ 3$.

First, a subset U of R^n is called open if whenever a is a point in U, there is a positive number ϵ such that every point $u \in R^n$ within a distance of ϵ of a is also in U:

$$a \in U \ and \ ||u - a|| < \epsilon \implies u \in U$$

For example, the whole of R^n is an open set, as in

$$D_r(a) = \{u \in R^n; ||u - a|| < r\}$$

the open ball with centre a and radius $r > 0$. (If $n = 1$, an open ball is called an open interval; if $n = 2$, it is called an open disc.) However,

$$\bar{D}_r(a) = \{u \in R^n; ||u - a|| \leq r\}$$

is not open, because however small the positive number ϵ is, there is a point within a distance ϵ of the point $(a_1 + r, a_2, \ldots, a_n) \in \bar{D}_n$(say) that is not in $\bar{D}_r(a)$ (for example the point $(a_1 + r + \epsilon/2, a_2, \ldots, a_n)$)

Next if X and Y are subsets of R^m and R^n, respectively a map $f : X \mapsto Y$ is said to be continuous at a point $a \in X$ if points in X near a are mapped by f onto points in Y near $f(a)$. More precisely, f is continuous at a if, given any number $\epsilon > 0$, there is a number $\delta > 0$ such that

$$u \in X \; and \; ||u - a|| < \delta$$
$$\implies \; ||f(u) - f(a)|| < \epsilon$$

Then f is said to be continuous if it is continuous at every point of X. Composites of continuous maps are also continuous.

In view of the definition of open sets, this is equivalent to the following:

f is continuous if and only if, for any open set V of R^n, there is an open set U of R^m such that f maps $U \cap X$ into $V \cap Y$.

If $f : X \mapsto Y$ is continuous and bijective and if its inverse function $f^{-1} : Y \mapsto X$ is also continuous, then f is called a homeomorphism and X and Y are said to be homeomorphic.

In the following, we give the definition of surface in R^3.

5.2 Definition

A subset S of R^3 is a surface, if for every point $p \in S$, there is an open set U in R^2 and an open set W in R^3 containing p such that $S \cap W$ is homeomorphic to U.

Thus, a surface comes equipped with a collection of homeomorphisms $\sigma : U \mapsto S \cap W$, which we call surface patches or parametrizations. The collection of all these surface patches is called atlas of S.

Every point of S lies in the image of at least one surface patch in the atlas of S. The reason for this terminology will become clear from the following example.

5.3 Example

Prove that the unit sphere

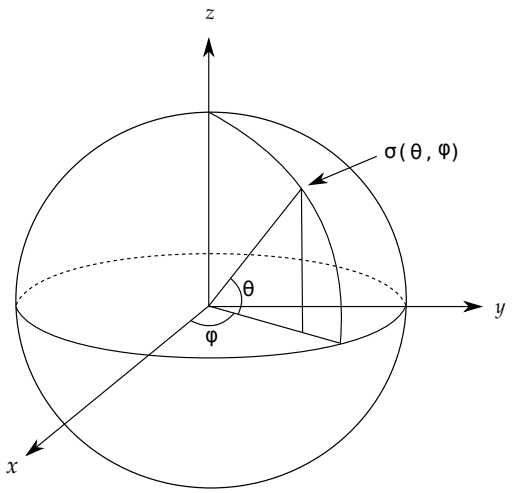
$$S^2 = \{(x, y, z) \in R^3; x^2 + y^2 + z^2 = 1\}$$

is a surface.

The most obvious parametrization is probably that given by latitude θ and longitude ϕ.

$$\sigma(\theta, \phi) = (\cos\theta \cos\phi, \cos\theta \sin\phi, \sin\theta)$$

Without some restriction on (θ, ϕ), σ is not injective and so it is not homeomorphism. To cover the whole sphere, it is clearly sufficient to take $-\pi/2 \leq \theta \leq \pi/2, \; 0 \leq \phi \leq 2\pi$.



However, the set of points (θ, ϕ) satisfying these inequalities is not an open subset of R^2 and so cannot be used as the domain of a surface. The largest open set consistent with the above inequalities is

$$U = \{(\theta, \phi); -\pi/2 < \theta < \pi/2,\ 0 < \phi < 2\pi\}$$

but now the image of $\sigma : U \mapsto R^3$ is not the whole of the sphere, but rather the complement of the great semicircle C consisting of the points of the sphere of the form $(x, 0, z)$ with $x \geq 0$. Hence, $U \to R^3$ covers only a patch of the sphere. Again we shall not verify in detail that σ is a homeomorphism from U to the intersection of the sphere with the open set

$$W = \{(x, y, z) \in R^3; x < 0 \text{ or } y \neq 0\}$$

To show that the sphere is a surface, we must therefore produce at least one more surface patch covering the part of the sphere omitted by σ. For example, let $\tilde{\sigma}$ be the patch obtained by rotating σ by π about the z-axis and then by $\pi/2$ about the x-axis. Explicitly, $\tilde{\sigma} : U \mapsto R^3$ is given by

$$\tilde{\sigma}(\theta, \phi) = (-\cos\theta\cos\phi, -\sin\theta, -\cos\theta\sin\phi)$$

The image $\tilde{\sigma}$ is the component of the great semi-circle $\tilde{C}$ consisting of the points of the sphere of the form $(x, y, 0)$ with $x \leq 0$.

It is clear that C and $\tilde{C}$ do not intersect, so the union of the images of σ and $\tilde{\sigma}$ is the whole sphere. Note that most points of the sphere are in the images of both surface patches.

5.4 Transition Map

As the example of the sphere shows, a point a of a surface S will generally lie in the image of more than one surface patch. Suppose then that $\sigma : U \mapsto S \cap W$ and $\tilde{\sigma} : \tilde{U} \mapsto S \cap \tilde{W}$ are two patches such that $a \in S \cap W \cap \tilde{W}$. Since σ and $\tilde{\sigma}$ are homeomorphisms, $\sigma^{-1}(S \cap W \cap \tilde{W})$ and $\tilde{\sigma}^{-1}(S \cap W \cap \tilde{W})$ are open sets $V \subseteq U$ and $\tilde{V} \subseteq \tilde{U}$, respectively. The composite homeomorphism $\sigma^{-1} \circ \tilde{\sigma}^{-1} : \tilde{V} \mapsto V$ is called the transition map from σ to $\tilde{\sigma}$. If we denote this map by Φ, we have

$$\tilde{\sigma}(\tilde{u}, \tilde{v}) = \sigma(\Phi(\tilde{u}, \tilde{v}))$$

for all $(\tilde{u}, \tilde{v}) \in \tilde{V}$.

5.5 Smooth Surface

In differential geometry we use calculus to analyse surface (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 R^m, we say that a map $f : U \mapsto R^n$ is smooth if each of the n-components of f, which are functions $U \to R$, have continuous partial derivatives of all orders. The partial derivatives of f are then computed component-wise. 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

$$\begin{aligned} \frac{\partial f}{\partial u} &= \left(\frac{\partial f_1}{\partial u}, \frac{\partial f_2}{\partial u}, \frac{\partial f_3}{\partial u}\right) \\ \frac{\partial f}{\partial v} &= \left(\frac{\partial f_1}{\partial v}, \frac{\partial f_2}{\partial v}, \frac{\partial f_3}{\partial v}\right) \end{aligned}$$

and similarly for higher derivatives.

5.6 Definition

A surface patch $\sigma : U \mapsto R^3$ 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.

5.7 Definition

A smooth surface is a surface σ whose atlas consists of regular surface patches.

For the unit sphere S^2, it is again clear that σ and $\tilde{\sigma}$ are smooth. As for regularity, we compute

$$\begin{aligned} \sigma_\theta &= (-\sin\theta\cos\phi, -\sin\theta\sin\phi, \cos\theta) \\ \sigma_\phi &= (-\cos\theta\sin\phi, \cos\theta\cos\phi, 0) \end{aligned}$$

which gives

$$\sigma_\theta \times \sigma_\phi = (-\cos^2\theta\cos\phi, -\cos^2\theta\sin\phi, -\sin\theta\cos\phi)$$

and hence,

$$||\sigma_\theta \times \sigma_\phi|| = |\cos\theta|.$$

But if, $(\theta, \phi) \in U$, then $-\pi/2 < \theta < \pi/2$, so $\cos\theta \neq 0$. Similarly, one checks that $\tilde{\sigma}$ is regular.

5.8 Proposition

Let U and $\tilde{U}$ be open subsets of R^2 and let $\sigma : U \mapsto R^3$ be a regular surface patch. Let $\phi : \tilde{U} \mapsto U$ be a bijective smooth map with smooth inverse map $\phi^{-1} : U \mapsto \tilde{U}$. Then

$$\tilde{\sigma} \equiv \sigma \circ \phi : \tilde{U} \mapsto R^3$$

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,

$$\begin{aligned} \tilde{\sigma}_{\tilde{u}} &= \frac{\partial u}{\partial \tilde{u}}\sigma_u + \frac{\partial v}{\partial \tilde{u}}\sigma_v \\ \tilde{\sigma}_{\tilde{v}} &= \frac{\partial u}{\partial \tilde{v}}\sigma_u + \frac{\partial v}{\partial \tilde{v}}\sigma_v \\ \tilde{\sigma}_{\tilde{u}} \times \tilde{\sigma}_{\tilde{v}} &= \left(\frac{\partial u}{\partial \tilde{u}}\frac{\partial v}{\partial \tilde{v}} - \frac{\partial u}{\partial \tilde{v}}\frac{\partial v}{\partial \tilde{u}}\right)\sigma_u \times \sigma_v \end{aligned} \tag{5.1}$$

The scalar on the right hand side of this equation is the determinant of the Jacobian matrix

$$J(\phi) = \begin{bmatrix} \frac{\partial u}{\partial \tilde{u}} & \frac{\partial u}{\partial \tilde{v}} \\ \frac{\partial v}{\partial \tilde{u}} & \frac{\partial v}{\partial \tilde{v}} \end{bmatrix}$$

of ϕ. We recall from calculus that, if ψ and $\tilde{\psi}$ are two maps between open sets in R^2,

$$J(\tilde{\psi}.\psi) = J(\tilde{\psi})J(\psi).$$

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 (5.1) shows that $\tilde{\sigma}$ is regular.

5.9 Theorem

Transition maps of smooth maps are smooth.

Proof: We shall use inverse function theorem to prove the theorem. We want to show that if $\sigma : U \mapsto R^3$ and $\tilde{\sigma} : \tilde{U} \mapsto R^3$ are two regular patches in the atlas of a surface S, the transition map from σ to $\tilde{\sigma}$ is smooth where it is defined.

Suppose that a point P lies in both patches, say

$$\sigma(u_0, v_0) = \tilde{\sigma}(\tilde{u}_0, \tilde{v}_0) = P$$

write

$$\sigma(u,v) = (f(u,v), g(u,v), h(u,v))$$

Since σ_u and σ_v are linearly independent, the Jacobian matrix

$$\begin{bmatrix} f_u & f_v \\ g_u & g_v \\ h_u & h_v \end{bmatrix}$$

of σ has rank 2 everywhere. Hence, at least one of its three 2×2 submatrices is invertible at each point. Suppose that the submatrix

$$\begin{bmatrix} f_u & f_v \\ g_u & g_v \end{bmatrix}$$

is invertible at P. (The proof is similar in other two cases) By the inverse function theorem applied to the map $F : U \mapsto R^2$ given by

$$F(u,v) = (f(u,v), g(u,v))$$

there is an open subset V of R^2 containing $F(u_0, v_0)$ and an open subset W of U containing (u_0, v_0) such that $F : W \mapsto V$ is bijective with a smooth inverse function given by $\pi(x,y,z) = (x,y)$ is also bijective, since $\pi = F \circ \sigma^{-1}$ on $\sigma(W)$. It follows that $W = \tilde{\sigma}^{-1}(\sigma(W))$ is an open subset of $\tilde{U}$ and that $\sigma^{-1} \circ \tilde{\sigma} \equiv F^{-1} \circ \tilde{F}$ on $\tilde{W}$, where $\tilde{F} = \pi \circ \tilde{\sigma}$. Since F^{-1} and $\tilde{F}$ are smooth on $\tilde{W}$, so is the transition map $\sigma^{-1} \circ \tilde{\sigma}$. Since $\sigma^{-1} \circ \tilde{\sigma}$ is smooth on an open set containing any point (u_0, v_0) where it is defined,it is smooth.

5.10 Definition

If $\gamma : (\alpha, \beta) \mapsto R^3$ is contained in the image of a surface patch $\sigma : U \mapsto R^3$ in the atlas of S, there is a map $(\alpha, \beta) \to U$, say, $t \to (u(t), v(t))$ such that

$$\gamma(t) = \sigma(u(t), v(t))$$

The tangent space at a point P of a surface S is the set of tangent vectors at P of all curves in S passing through P.

5.11 Proposition

Let $\sigma : U \mapsto R^3$ be a patch of a surface S containing a point P of S, and let (u,v) be coordinates in U. The tangent space to S at P is the vector subspace of R^3 spanned by the vectors σ_u and σ_v.
Proof: Let γ be a smooth curve in S, say

$$\gamma(t) = \sigma(u(t), v(t))$$

Denoting $\frac{d}{dt}$ by 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 subspace of R^2 spanned by σ_u and σ_v is of the form $\psi\sigma_u + \eta\sigma_v$ for some scalars ψ and η.
Define

$$\gamma(t) = \sigma(u_0 + \psi t, v_0 + \eta t)$$

Then γ is smooth curve in S and at $t = 0$, that is, at the point P in S, we have

$$\dot{\gamma} = \psi\sigma_u + \eta\sigma_v$$

This shows that every vector in the span of σ_u and σ_v is the tangent vector at P of some curves in S.

5.12 Normal to the surface

Let σ_u and σ_v be two linearly independent tangents at a point P on a surface. Then normal to the surface at P is defined as

$$N(\sigma) = \frac{\sigma_u \times \sigma_v}{||\sigma_u \times \sigma_v||}$$

5.13 Orientable Surface

An orientable surface is a surface with an atlas having the property that, if ϕ is the transition map between any two charts in the atlas, then $\det\{J(\phi)\} > 0$, where ϕ is defined.

A Möbius band is not orientable.

5.14 Exercises

1. Show that an open disc in the xy-plane is a surface.
2. Show that the circular cylinder

   $$S = \{(x, y, z) \in R^3 : x^2 + y^2 = 1\}$$

   can be covered by a single surface patch and so is to 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. Show that a Möbius band is not orientable.

5.15 Summary

In this unit, we have defined smooth surface, transition map, tangent and normal to a surface. We have also defined orientability.

5.16 Suggested Reading

(i) Elementary differential geometry- Andrew Pressley.

References:

1. Elementary differential geometry, Andrew Pressley.
2. Differential geometry of curves and surfaces, M. P do Carmo.
3. Multivariate Calculus and geometry, S. Dineen.
4. Differential Geometry of curves and surfaces,(Tensor Approach), U. C. De.
5. Differential Geometry, C. B$\ddot{a}$r.

Block II
Topology I

Unit 11

Course Structure

1. Topological spaces: definition and examples

1 Introduction

One way to describe the subject of Topology is to say that it is qualitative geometry. The idea is that if one geometric object can be continuously transformed into another, then the two objects are to be viewed as being topologically the same. For example, a circle and a square are topologically equivalent, a sphere and a hollow box are equivalent . Physically, a rubber band can be stretched into the form of either a circle or a square, as well as many other shapes which are also viewed as being topologically equivalent. On the other hand, a figure eight curve formed by two circles touching at a point is to be regarded as topologically distinct from a circle or square. A qualitative property that distinguishes the circle from the figure eight is the number of connected pieces that remain when a single point is removed: When a point is removed from a circle what remains is still connected, a single arc, whereas for a figure eight if one removes the point of contact of its two circles, what remains is two separate arcs, two separate pieces. The term used to describe two geometric objects that are topologically equivalent is homeomorphic. Thus a circle and a square are homeomorphic. Concretely, if we place a circle C inside a square S with the same center point, then projecting the circle radially outward to the square defines a function $f : C \to S$, and this function is continuous: small changes in x produce small changes in $f(x)$. The function f has an inverse $f^{-1} : C \to S$ obtained by projecting the square radially inward to the circle, and this is continuous as well. One says that f is a homeomorphism between C and S. One of the basic problems of Topology is to determine when two given geometric objects are non homeomorphic. This can be quite difficult in general. Our first goal will be to define exactly what the 'geometric objects' are that one studies in Topology. These are called topological spaces. The definition turns out to be extremely general, so that many objects that are topological spaces are not very geometric at all, in fact.

1.1 Topological Spaces

Rather than jump directly into the definition of a topological space we will first spend a little time motivating the definition by discussing the notion of continuity of a function. One could say that topological spaces are the objects for which continuous functions can be defined.

For the sake of simplicity and concreteness let us talk about functions $f : \mathbb{R} \to \mathbb{R}$. There are two definitions of continuity for such a function that the you may already be familiar with, the $\epsilon - \delta$ definition and the definition in terms of limits. But it is a third definition, equivalent to these two, that is the one we want here. This definition is expressed in terms of the notion of an open set in $\mathbb{R}$, generalizing the familiar idea of an open interval (a, b).

Definition 1. A subset O of $\mathbb{R}$ is open if for each point $x \in O$ there exists an interval (a, b) that contains x and is contained in O.

With this definition an open interval certainly qualifies as an open set. Other examples are:

- $\mathbb{R}$ itself is an open set, as are semi-infinite intervals (a, ∞) and $(-\infty, a)$.
- The complement of a finite set in $\mathbb{R}$ is open.
- If $A = \{1/n : n = 1, 2, \ldots\} \cup \{0\}$ $\mathbb{R} \setminus A$ is open.
- Any union of open intervals is an open set. The preceding examples are special cases of this. The converse statement is also true: every open set O is a union of open intervals since for each $x \in O$ there is an open interval (a_x, b_x) with $x \in (a_x, b_x) \subset O$, and O is the union of all these intervals (a_x, b_x) .
- The empty set $\emptyset$ is open, since the condition for openness is satisfied vacuously as there are no points x where the condition could fail to hold.

Now for the nice definition of a continuous function in terms of open sets:

Definition 2. A function $f : \mathbb{R} \to \mathbb{R}$ is continuous if for each open set O in $\mathbb{R}$ the inverse image$f^{-1}(O) = \{x \in \mathbb{R} : f(x) \in O\}$ is also an open set.

To see that this corresponds to the intuitive notion of continuity, consider what would happen if this condition failed to hold for a function f. There would then be an open set O for which $f^{-1}(O)$ was not open. This means there would be a point $x_0 \in f^{-1}(O)$ for which there was no interval (a, b) containing x_0 and contained in $f^{-1}(O)$. This is equivalent to saying there would be points x arbitrarily close to x_0 that are in the complement of $f^{-1}(O)$. For x to be in the complement of $f^{-1}(O)$ means that $f(x)$ is not in O. On the other hand, x0 was in $f^{-1}(O)$ so $f(x_0)$ is in O. Since O was assumed to be open, there is an interval (c, d) about $f(x_0)$ that is contained in O. The points $f(x)$ that are not in O are therefore not in (c, d) so they remain at least a fixed positive distance from $f(x_0)$. To summarize: there are points x arbitrarily close to x_0 for which $f(x)$ remains at least a fixed positive distance away from $f(x_0)$. This certainly says that f is discontinuous at x_0. This reasoning can be reversed. A reasonable interpretation of discontinuity of f at x_0 would be that there are points x arbitrarily close to x_0 for which $f(x)$ stays at least a fixed positive distance away from $f(x_0)$. Call this fixed positive distance ϵ. Let O be the open set $(f(x_0)-\epsilon,\ f(x_0)+\epsilon)$. Then $f^{-1}(O)$ contains x_0 but it does not contain any points x for which $f(x)$ is not in O , and we are assuming there are such points x arbitrarily close to x_0, so $f^{-1}(O)$ is not open since it does not contain all points in some interval (a, b) about x_0.

In trying to find a satisfactory definition of a topological space we shall have two aims in mind. The definition should be general enough to allow a wide range of different structures as spaces. We would like to consider a finite, discrete set of points as a space, or equally a whole uncountable continuum of points such as the real line; one of our nice geometrical surfaces should qualify under the definition, and so too should a function space such as the set of continuous complex-valued functions defined on the unit circle in the complex plane. We would like to be able to perform simple constructions with our spaces, such as taking the cartesian product of two spaces, or identifying some of the points of a space in order to form a new one (think of the construction of the Mobius strip outlined earlier). On the other hand, the definition of a space should contain enough information so that we can define the notion of continuity for functions between spaces. It is really this second consideration which leads to the abstract definition given below.

Definition 3. A topological space is a set X together with a collection $\mathcal{O}$ of subsets of X, called open sets, such that:

(i) Both $\emptyset$ and X are in $\mathcal{O}$.
(ii) The union of any collection of sets in O is in $\mathcal{O}$.
(iii) The intersection of any finite collection of sets in O is in $\mathcal{O}$.
The collection $\mathcal{O}$ of open sets is called a topology on X.
Complements of open sets are called closed sets.

Exercise 4. Prove that so called open sets in $\mathbb{R}$ produces a topology $\mathbb{R}$.

Exercise 5. Give an example to show that intersection of infinitely many open sets may not be open.

It is always possible to construct at least two topologies on every set X by choosing the collection $\mathcal{O}$ of open sets to be as large as possible or as small as possible: The collection $\mathcal{O}$ of all subsets of X defines a topology on X called the discrete topology. If we let $\mathcal{O}$ consist of just X itself and $\emptyset$, this defines a topology, the trivial topology.

Thus we have three different topologies on $\mathbb{R}$, the usual topology, the discrete topology, and the trivial topology. The following one contains, fewer open sets than the usual topology on $\mathbb{R}$.

Exercise 6. Let $\mathcal{O} = \{A \subset \mathbb{R} : \mathbb{R} \setminus A \text{ is finite}\}$. Prove that $\mathcal{O}$ is a topology on $\mathbb{R}$ containing fewer open sets than the usual topology on $\mathbb{R}$.

The following one contains, more open sets than the usual topology on $\mathbb{R}$.

Exercise 7. Let $\mathcal{O}$ consists of the sets $A \subset \mathbb{R}$ such that for each $x \in A$ there exists $a, b \in \mathbb{R}$ so that $x \in [a, b) \subset A$. Prove that $\mathcal{O}$ is a topology on $\mathbb{R}$ containing more open sets than the usual topology on $\mathbb{R}$.

SELF ASSESSMENT

1. Prove that the collection of all open sets in R^2 obtained by Euclidean metric generates a topology.

2. What will by the interior of the set $\{\frac{1}{n} : n \in \mathbb{N}\}$ in $\mathbb{R}$?

3. What will by the closure of the set $\{\frac{1}{n} : n \in \mathbb{N}\}$ in $\mathbb{R}$?

4. Write the interior and closure of the following set $\{\frac{1}{n} + \frac{1}{m} : n, m \in \mathbb{N}\}$.

5. Show that if U is open in X and A is closed in X, then $U \setminus A$ is open in X, and $A \setminus U$ is closed in X.

6. If A is a dense subset of a space X, and if O is open in X, show that $O \subset \overline{A \cap O}$.

7. Prove that τ_{cc} is in fact a topology on any uncountable space.

8. Prove that τ_{cf} is becomes discrete topology if the space is countable.

9. Verify each of the following for arbitrary subsets A, B of a topological space X:

(a) $\overline{A \cup B} = \overline{A} \cup \overline{B}$, (b) $\overline{A \cap B} \subset \overline{A} \cap \overline{B}$.

(c)$\text{int}(A \cap B) = \text{int}(A) \cap \text{int}(B)$, (d) $\text{int}(A \cap B) \subset \text{int}(A) \cap \text{int}(B)$.

Give examples where equality fails to hold in (b) and (d).

10. 2. Specify the interior and closure of the of the following subsets :

(a) $\{(x, y) : 1 < x^2 + y^2 \leq 2\}$,

(b) $\mathbb{R}^2$ with both axes removed,

(c) $\mathbb{R}^2 - \{(x, \sin 1/x) : x > 0\}$.

1.2 Summary

This section has made us acquainted with open sets which are the main foundations of Topological Spaces; seen some examples; learnt about the continiuty of real functions in terms of open sets in the real line.

Unit 12

Course Structure

1. Basis for a topology

2 Introduction

Many arguments with open sets in $\mathbb{R}$ reduce to looking at what happens with open intervals since open sets are defined in terms of open intervals. A similar statement holds for $\mathbb{R}^2$ and $\mathbb{R}^n$ with open disks and balls in place of open intervals. In each case arbitrary open sets are unions of the special open sets given by open intervals, disks, or balls. This idea is expressed by the following terminology:

Definition 8. A collection $\mathcal{B}$ of open sets in a topological space X is called a basis for the topology if every open set in X is a union of sets in $\mathcal{B}$.

A topological space can have many different bases. Most common example is that $\mathbb{R}$ with usual topology has the following two bases :

1. $\{(a, b) : a, b \in \mathbb{R}\}$,
2. $\{(a, b) : a, b \in \mathbb{Q}\}$

In $\mathbb{R}^2$ another basis besides the basis of open disks is the basis of open squares with edges parallel to the coordinate axes. Or we could take open squares with edges at 45 degree angles to the coordinate axes, or all open squares without restriction. Many other shapes besides squares could also be used.

The following Lemma provides some neighbourhood like properties of base.

Lemma 9. *If $\mathcal{B}$ is a basis for a topology on X , then $\mathcal{B}$ satisfies the following two properties:*

(1) Every point $x \in X$ lies in some set $B \in \mathcal{B}$.

(2) For each pair of sets B_1, B_2 in $\mathcal{B}$ and each point $x \in B_1 \cap B_2$ there exists a set B_3 in $\mathcal{B}$ with $x \in B_3 \subset B_1 \cap B_2$.

Proof. The first statement holds since X is open and is therefore a union of sets in $\mathcal{B}$. The second statement holds since $B_1 \cap B_2$ is open and hence is a union of sets in $\mathcal{B}$. □

The following properties shows that if a a class of subsets of a set X has the above mentioned two properties then the class has the power to generate a topology.

Proposition 10. *If $\mathcal{B}$ is a collection of subsets of a set X satisfying (1) and (2) then $\mathcal{B}$ is a basis for a topology on X.*

Proof. Do yourself. □

Exercise 11. (a) Prove that the collection $\{[a, b) : a, b \in \mathbb{R}\}$ is a basis for some topology on $\mathbb{R}$. This topology on $\mathbb{R}$ called the lower limit topology and denoted by $\mathbb{R}_l$. This is also called Sorgenfrey line.

(b) Prove that $\mathbb{R}_l$ has a basis each of whose member is closed as well as open.

(c) Whether $\{[a, b) : a, b \in \mathbb{Q}\}$ generates the lower limit topology on $\mathbb{R}$.

(d) Prove that every metric space is a topological spaces with the set $\{B(x, r) : x \in \mathbb{R}, r > 0\}$ as a basis.

Definition 12. A neighborhood of a point x in a topological space X is any set $A \subset X$ that contains an open set O containing x. Dually x is said to be an interior point of A.

The more restricted kind of neighborhood can then be described as an open neighborhood, that is neighborhoods which are open.

Definition 13. A topological space X is said to be **Hausdorff** if for any two distinct points x, y there exists disjoint neighborhoods of x and y.

Exercise 14. (a) Prove that the space defined in Exercise 6 is not Hausdorff.

(b) Prove that every finite Hausdorff space is discrete.

(c) Prove that $\mathbb{R}_u$, $\mathbb{R}_l$, any metric spaces are Hausdorff.

(d) Proved that every finite set in a Hausdorff space is closed. Does there exist any non Hausdorff space in whic every finite set is closed. (In fact such sets are called T_1-spaces).

From now on unless otherwise stated any space will be considered as Hausdorff. We already defined that complements of open sets are called closed sets. In the following discussions we introduce the notion of limit points, closure.

Definition 15. Let A be a subset of a topological space X and $x \in X$. Then x is said to be a limit point of A if any neoghborhood of x meets A at a point other that x. If x is not a limit point of A then it is an isolated point of A. Hence isolated point of X is just an open set.

The set of all limit points of a set A is called derived set of A and denoted by A'. By the closure of A we mean the set A along with its limit points, that is $A \cup A'$ and denoted by $\overline{A}$ or $c\ell_X A$. The set of all interior points of A called interior of A and denoted by A°.

Proposition 16. *Let X be a topological and $A \subset X$. Then the followings hold:*

(1) A is open if and only if $A = A^\circ$.

(2) A° is the largest open set contained in A.

(3) A is closed if and only if $A = \overline{A}$.

(4) $\overline{A}$ is the smallest closed set containing A.

Proof. Do yourself. □

We can define convergency of sequence in topological spaces analogus to metric space.

Definition 17. Let X be a topological space and $x \in X$. A sequence $(x_n)_{n=1}^{\infty}$ is said to converge at x if for any neighborhood N_x of x there exists some $n_0 \in \mathbb{N}$ such that $(\forall n \geq n_0)(x_n \in N_x)$.

It is obvious that any sequence in any Hausdorff space may converge to at most one point.

Example 18. (a) Give an example of non Hausdorff space where a sequence may converge to more that one point.

(b) Construct an example of a topological space where convergency of a sequence may not explain limit points.

SELF ASSESMENT

1. Give examples of two dierent bases for the usual topology of $\mathbb{R}$.

2. Show that the collection of all open rectangles form a base for $\mathbb{R}^2$.

3. Show that if $\mathcal{B}$ is a basis for a topology on X, then the topology generated by $\mathcal{B}$ equals the intersection of all topologies on X that contain $\mathcal{B}$.

4. Does topologies of $\mathbb{R}$ and $\mathbb{R}_l$ are comparable?

5. Prove that basic open sets in Sorjenfrey Line are also closed.

6. Let Y be a subspace of X. Given A $\subset Y$. Prove that $\text{int}_X A \subset \text{int}_Y(A)$ and give an example to show the two may not be equal.

7. Prove that any metric gives a base for some topology on a set.

8. Does the intersection of lower and upper limit topologies equal with usual topology on $\mathbb{R}$?

Summary

In this section, we have learnt about the Basis of a topological space which make up the open sets in them. For example, in the real line, the open sets are countable union of disjoint open intervals, which in turn, serve as the basis of the usual topology.We have also learnt about the Hausdorff property of topological spaces.

Unit 13

Course Structure

1. Subspace: Definition and example
2. Continuity and Homeomorphism

3 Introduction

We turn now to a topic which seems simple enough at first glance, but turns out to be a source of many headaches until one finally becomes comfortable with it. But once reader will be comfortable with it, can constract many examples of topological spaces.

Given a topology $\mathcal{O}$ on a space X and a subset $A \subset X$, we would like to use the topology on X to define a topology $\mathcal{O}_A$ on A. There is an easy way to do this: Just define a set $O \subset X$ to be in $\mathcal{O}_A$ if there exists an open set O' in $\mathcal{O}$ such that $O = A \cap O'$.

Exercise 19. Prove that $\mathcal{O}_A$ is in fact a topology on A.

The topology $\mathcal{O}_A$ on A is called the subspace topology, and A with this topology is called a subspace of X. For example, if we take X to be $\mathbb{R}^2$ with its usual topology, then every subset of $\mathbb{R}^2$ becomes a topological space. In particular, geometric figures such as circles and polygons can now be viewed as topological spaces. Likewise, geometric figures in $\mathbb{R}^3$ such as spheres and polyhedra become topological spaces, with the subspace topology from the usual topology on $\mathbb{R}^3$.

In case the space X is a metric space, any subset $A \subset X$ becomes a metric space by restricting the metric $X \times X \longrightarrow \mathbb{R}$ to $A \times A$, since the three defining properties of ametric obviously still hold for the restricted distance function. The following Proposition gives some strong evidence that the subspace topology is a natural topology to use on subsets.

Proposition 20. *The metric topology on a subset A of a metric space X is the same as the subspace topology.*

Proof. Do yourself. □

For a subspace $A \subset X$, a subset of A which is open or closed in A need not be open or closed in X. However, we have the following fact.

Lemma 21. *For a subspace $Y \subset X$ which is open (resp. closed) in X, a subset $A \subset Y$ is open in Y (resp. closed) if and only if it is open (resp. closed) in X.*

Proof. Do youeself. □

Exercise 22. Give an example of a topological space to establish that open or closed for the subspace Y in the above example can not be removed.

Proof. Closures behave nicely with respect to subspaces: □

Theorem 23. *Given a space X, a subspace Y, and a subset $A \subset Y$, then the closure of A in the space Y is the intersection of the closure of A in X with Y . That is a point $y \in Y$ is a limit point of A in Y (i.e. using the subspace topology on Y) if and only if y is a limit point of A in X.*

Proof. For a point $y \in Y$ to be a limit point of A in X means that every open set O in X that contains y meets A. Since $A \subset Y$, this is equivalent to $O \cap Y$ meeting A, or in other words, that every open set in Y containing y meets A. □

Exercise 24. Show by an example that analogus statement of the above theorem is not true for interiors.

4 Continuity and Homeomorphisms

Recall the definition: a function $f : X \to Y$ between topological spaces is continuous if $f^{-1}(O)$ is open in X for each open set O in Y. For brevity, continuous functions are sometimes called maps.

Proposition 25. *A function $f : X \to Y$ is continuous if and only if $f^{-1}(C)$ is closed in X for each closed set C in Y.*

Proof. Do yourself. □

The following proposition shows that continuous maps behave well under formation of closure.

Proposition 26. *A function $f : X \to Y$ is continuous if and only if $f(c\ell_X A) \subset c\ell_Y f(A)$ for any $A \subset X$.*

Proof. If x is a limit point of A then $f(x)$ is a limit point of $f(A)$ (verify). Hence $f(c\ell_X A) \subset c\ell_Y f(A)$.

Conversely let K be a closed set in Y and $A = f^{-1}(K)$. Then$f(c\ell_X A) \subset c\ell_Y f(A) = c\ell_Y f((f^{-1}(K))$. This implies that $f(c\ell_X A) \subset c\ell_Y K = K$, so that $A \subset c\ell_X A \subset f^{-1}(K) = A$. Hence $c\ell_X A = f^{-1}(K)$, that is $f^{-1}(K)$ is closed and therefore f is continuous. □

The following fact again shows that information about base is sufficiant for any topological spaces.

Lemma 27. *Given a function $f : X \to Y$ and a basis $\mathcal{B}$ for Y, then f is continuous if and only if $f^{-1}(B)$ is open in X for each $B \in \mathcal{B}$.*

Proof. Left as exercise.

The following lemma shows that continuous maps behave well under composition. □

Lemma 28. *If $f : X \to Y$ and $g : Y \to Z$ are continuous, then their composition $g \circ f : X \to Zis$ also continuous.*

Proof. Obvious. □

Lemma 29. *If $f : X \to Y$ is continuous and A is a subspace of X, then the restriction $f_{/A}$ of f to A is continuous as a function $A \to Y$. Further the inclusion map $i_A : A \hookrightarrow X$ is also continuous.*

Proof. Obvious.

We introduce the following definition in comtrust to continuous mapping. □

Definition 30. A function $f : X \to Y$ between topological spaces is said to be open if for any open (resp. closed) set A, in X, $f(A)$ is open (resp. closed) in Y.

Example 31. (a) Give an example of a continuous function which is neither open nor closewd.

(b) Give an example of a open map which is not continuous.

(c) Give an example of a closed map which is not continuous.

(d) Give an example of a open map which is not closed.

(b) Give an example of a closed map which is not open.

Now we are in a situation to define structure preserving mapping for topological spaces.

Definition 32. A continuous map $f : X \to Y$ between topological spaces is said to be homeomorphism if there exists a continuous map $g : Y \to X$ such that $g \circ f = 1_X$ and $f \circ g = 1_Y$.

Theorem 33. *A mapping $f : X \to Y$ between topological spaces is a homeomorphism if and only if it is continuous and open or closed.*

Exercise 34. (a) Prove that the circrle $\{(x, y) : x^2 + y^2 = 1\}$ and $\{(x, y) : x^2 + y^2 = 4\}$ are homeomorphic.

(b) Prove that the unit disk $\{(x, y) : x^2 + y^2 \leq 1\}$ and the unit square $\{(x, y) : 0 \leq x, y \leq 1\}$ are homeomorphic.

SELF ASSESMENT

1. Wheather $[0, \frac{1}{2})$ is an open set in the subspace topology on $[0, 1]$ induced from the usual topology from $\mathbb{R}$.

2. Describe the basic open set on the unit circle in the subspace topology induced from $\mathbb{R}^2$.

3. Describe the basic open set on the unit sphere in the subspace topology induced from $\mathbb{R}^3$.

4. Show that if Y is a subspace of X, and A is a subset of Y, then the topology A inherits as a subspace of Y is the same as the topology it inherits as a subspace of X.

5. Prove that the set of rational numbers as a subspace of $\mathbb{R}$ is not discrete but $\mathbb{Z}$ as a subspace of $\mathbb{R}$ is discrete.

6. Whether $\{0\} \cup \{\frac{1}{n} : n \in \mathbb{N}\}$ is a discrete subspace of $\mathbb{R}$?

7. Let Y be a subspace of X. If A is open (closed) in Y, and if Y is open (closed) in X, show that A is open (closed) in X.

8. Give an example of two closed sets A and B in $\mathbb{R}$ such that $A + B$ is not closed in $\mathbb{R}$.

9. Prove that all bounded closed intervals are homeomorphic.

10. Prove that open interval $(0; 1)$ is homeomorphic to $\mathbb{S}^1 \setminus \{i\}$.

11. Let X denote the set of all real numbers with the finite-complement topology, and dene $f : \mathbb{R} \to X$. Show that f is continuous, but is not a homeomorphism.

12. Suppose $f : X \to Y$ is a homeomorphism and $U \subset X$ is an open subset. Show that $f(U)$ is open in Y and the restriction $f|_U$ is a homeomorphism from U to $f(U)$.

13. There is a generalization of homeomorphisms that is often useful. We say that a map $f : X \to Y$ between topological spaces is a local homeomorphism if every point $x \in X$ has a neighborhood $U \subset X$ such that $f(U)$ is an open subset of Y and $f|_U : U \to f(U)$ is a homeomorphism. prove that

(a) Every homeomorphism is a local homeomorphism.

(b) Every local homeomorphism is continuous and open.

(c) Every bijective local homeomorphism is a homeomorphism.

14. A map $f : X \to Y$ is said to be open if $f(O)$ is open in Y whenever O is open in X. Similarly, $f : X \to Y$ is said to be closed if $f(C)$ is closed in Y whenever C is closed in X.

(a) Give an example of a map that is open but not closed, and an example of a map that is closed but not open.

(b) Determine whether the projection map $\mathbb{R}^2 \to \mathbb{R}$ sending (x, y) to x is open or closed.

(c) The exponential map $x \to e^{ix}$ from the real line to the circle.

(d) The folding map $f : \mathbb{R}^2 \to \mathbb{R}^2$ given by $f(x, y) = (x, |\mathrm{y}|)$.

(e) The map which winds the plane three times on itself given, in terms of complex numbers, by $z \to z^3$.

15. Show the twomaps $\mathbb{R}^2 \to \mathbb{R}$ sending (x,y) to $x + y$ and $x \cdot y$ are continuous, using only definitions and results from this class, not results from calculus for example.

16. Prove that $\mathbb{S}^n \setminus \{N\}$ is homeomorphic to $\mathbb{R}^n$.

17. If $f : \mathbb{R} \to \mathbb{R}$ is a map (i.e., a continuous function), show that the set of points which are left fixed by f is a closed subset of $\mathbb{R}$. If g is a continuous real-valued function on X show that the set $x : g(x) = 0$ is closed.

18. Prove that the function $h(x) = \frac{e^x}{1+e^x}$ is a homeomorphism from the real line to the open interval $(0, 1)$.

Summary

In this unit, we have learnt that when can we call a subset of a topological space, a subspace of it. This gives us an important tool to create noew topological spaces from the existing ones with the help of the existing open sets. ALso we have become acquainted with continuity and homeomorphism.

Unit 14

Course Structure

1. Product Spaces: Definition and examples

5 Introduction

In this unit we will learn about an important tool to create new topological spaces from existing ones by the simple operation of cartesian product of two sets X and Y. The open sets in the resulting set will be defined likewise. We give the formal definition of the product topology as follows:

5.1 Product Spaces

If X and Y are topological spaces, we can define a topology on $X \times Y$ by saying that a basis consists of the subsets $U \times V$ as U ranges over open sets in X and V ranges over open sets in Y . Verify that the collection is infact a basis. The topology generated by this base is called the product topology on $X \times Y$.

More generally one can define the product$X = X_1 \times \ldots \times X_n$ to consist of all ordered n-tuples $(x_1, \ldots, x_n)$ with $x_i \in X_i$ for each i. A basis for the product topology on $X = X_1 \times \ldots \times X_n$ consists of all products $U_1 \times \ldots \times U_n$ as each U_i ranges over open sets in X_i , or just over a basis for the topology on X_i . Thus $\mathbb{R}^n$ with its usual topology is also describable as the product of n copies of $\mathbb{R}$, with basis the open 'boxes' $(a_1, b_1) \times \ldots \times (a_n, b_n)$. The map $\pi_i : X \to X_i$ defined by $\pi_i(x_1, \ldots, x_n) = x_i$ is called i-th projection map. It is easy to verify that projection maps are continuous.

Example 35. (a). If we view points in the unit circle $\mathbb{S}^1$ in $\mathbb{R}^2$ as angles θ, then polar coordinates give a homeomorphism $f : \mathbb{S}^1 \times (0, \infty) \to \mathbb{R}^2 \setminus \{0\}$ defined by $f(\theta, r) = (r \cos \theta, r \sin \theta)$. This is one-to-one and onto since each point in $\mathbb{R}^2$, other than the origin has unique polar coordinates (θ, r). To see that f is a homeomorphism, just observe that it takes a basis set $U \times V$, where U is an open interval (θ_0, θ_1) of values and V is an open interval (r_0, r_1) of r values, to an open polar rectangle and such rectangles form a basis for the topology on $\mathbb{R}^2 \setminus \{0\}$. as a subspace of $\mathbb{R}^2$. By restricting f to a product $\mathbb{S}^1 \times [a, b]$ for $0 < a < b$ we obtain a homeomorphism from this product to a closed annulus in $\mathbb{R}^2$, the region between two concentric circles.

More generally, $\mathbb{R}^n \setminus \{0\}$ is homeomorphic to $\mathbb{S}^{n-1} \times (0, \infty)$ where $\mathbb{S}^{n-1}$ is the unit sphere in $\mathbb{R}^n$. Using vector notation, a homeomorphism $f : \mathbb{S}^{n-1} \times (0, \infty) \to \mathbb{R}^n \setminus \{0\}$ is given by$f(v, r) = rv$, with inverse $f^{-1}(v) = (\frac{v}{|v|}, |v|)$. The continuity of f and f^{-1} are clear.

Example 36. A product $\mathbb{S}^1 \times [1, 2]$ is homeomorphic to a cylinder as well as to an annulus. If we use cylindrical coordinates (r, θ, z) in $\mathbb{R}^3$ then a cylinder is specified by taking r to be a constant 1, letting range over the circle $\mathbb{S}^1$, and restricting z to an interval $[1, 2]$.

Example 37. The product $\mathbb{S}^1 \times \mathbb{S}^1$ is homeomorphic to a torus, say the torus $\mathbb{T}$ in $\mathbb{R}^3$ obtained by taking a circle $\mathcal{C}$ in the yz-plane disjoint from the z-axis and rotating this circle about the z-axis. We can parametrize points on $\mathbb{T}$ by a pair of angles (θ_1, θ_2) where θ_1 is the angle between the horizontal radial vector of $\mathcal{C}$ pointing away from the z-axis and the radial vector to a given point of $\mathcal{C}$ and θ_2 is the angle through which the yz-plane has been rotated around z-axis (One can think of θ_1 and θ_2 as longitude and latitude on $\mathbb{T}$). A basic open set $U \times V$ in $\mathbb{S}^1 \times \mathbb{S}^1$ is a product of two open arcs, and this corresponds to an open curvilinear rectangle on $\mathbb{T}$. Such rectangles form a basis for the topology on $\mathbb{T}$ as a subspace of $\mathbb{R}^3$, so it follows that $\mathbb{T}$ is homeomorphic to $\mathbb{S}^1 \times \mathbb{S}^1$.

Definition 38. A product space $X \times Y$ has two projection maps $p_1 : X \times Y \to X$ and $p_1 : X \times Y \to Y$ defined by $p_1(x, y) = x$ and $p_2(x, y) = y$. These maps are continuous since if $U \subset X$ is open then so is $p_1^{-1}(U) = U \times Y$, and if $V \subset Y$ is open then so is $p_2^{-1}(V) = X \times V$.

Theorem 39. *A function $f : Z \to X \times Y$ defined by $f(z) = (f_1(z), f_2(z))$ is continuous if and only if its component functions $f_1 : Z \to X$ and $f_2 : Z \to Y$ are both continuous.*

Proof. Will be done in class. □

Observe that the method we have used to construct for product topology with two topological spaces can be easily generalised to any number of finite topological spaces. One can observe that if we take countably many spaces $(X_n)_{n=1}^{\infty}$ the above method can still be generalised.

Definition 40. If $(X_n)_{n=1}^{\infty}$ be a sequence of topological spaces then we define $X = \prod_{n=1}^{\infty} X_n = \{f : \mathbb{N} \to \bigcup_{n=1}^{\infty} X_n : f(i) \in X_i\}$.

Exercise 41. Let $(X_n)_{n=1}^{\infty}$ be a sequence of topological spaces then $\mathcal{B} = \{\prod_{n=1}^{\infty} B_n : B_n \text{ is a basic open set in } X_n\}$ is a base for some topology on X.

In the following example we show that the topology created in Exercise 41 is not a successful generalization for infinite product.

Example 42. Let $\mathbb{R}^{\mathbb{N}}$ be endowded with the topology defined as in Exercise 41. And let $f : \mathbb{R} \to \mathbb{R}^{\mathbb{N}}$ defined by

$$f(x) = (x, x, \ldots, x) \text{ for all } x \in \mathbb{R}.$$

Now $f(0) = (0, 0, \ldots, 0)$ and $U = \prod_{n=1}^{\infty}(-\frac{1}{n}, \frac{1}{n})$ is an open neighbourhood of $\bar{0} = (0, 0, \ldots, 0)$ by the topology defined as in Exercise 41. But $f^{-1}(U) = \{0\}$(verify) is not open in $\mathbb{R}$. Hence $f : \mathbb{R} \to \mathbb{R}^{\mathbb{N}}$ is not continuous. (You have to write in details the example).

The above example hints the necessity of a new definition for infinite product so that Theorem 39 remains valid.

Definition 43. Let $(X_\alpha)_\alpha$ be a collection of topological spaces. We define product of the collection $(X_\alpha)_{\alpha\in\Lambda}$ ... modulo Zorn's Lemma, which the reader is kindly encouraged to accept as follows

$$X = \prod_{\alpha\in\Lambda} X_\alpha = \{f : \Lambda \to \bigcup_{\alpha\in\Lambda} X_\alpha : f(\alpha) \in X_\alpha\}.$$

For each $\alpha \in \Lambda$, let $\mathcal{S}_\alpha = \{\pi_\alpha^{-1}(U) : U \text{ is open in } X_\alpha\}$ and set $\mathcal{S} = \bigcup_{\alpha\in\Lambda} \mathcal{S}_\alpha$. Observe that for each $x \in X$ there exists some $S \in \mathcal{S}$ containing x. Let $\mathcal{B}$ be the set of all possible finite intersections of members of $\mathcal{S}$'s, that is

$$\mathcal{B} = \left\{ \bigcap_{\alpha\in F} U_\alpha : \text{ where } U_\alpha \text{ is open in } X_\alpha \text{ and } F \text{ is a finite subset of } \Lambda \right\}.$$

Remark 44. Observe that if $U \in \mathcal{S}$ then there exist some open set say U_α in X_α such that $U = \pi_\alpha^{-1}(U_\alpha)$ and if $B \in \mathcal{B}$ then there exist $\alpha_1, \alpha_2 \ldots, \alpha_k \in \Lambda$ such that

$$B = \pi_{\alpha_1}^{-1}(U_{\alpha_1}) \cap \pi_{\alpha_2}^{-1}(U_{\alpha_2}) \cap \ldots \cap \pi_{\alpha_k}^{-1}(U_{\alpha_k}).$$

For any $\alpha \in \Lambda$ and open set U in X_α, $\pi_\alpha^{-1}(U) = \prod_{\beta\in\Lambda} U_\beta$, where $U_\beta = U$ when $\beta = \alpha$, and $U_\beta = X_\beta$ for $\beta \neq \alpha$. This easily shows that

$\pi_{\alpha_1}^{-1}(U_{\alpha_1}) \cap \pi_{\alpha_2}^{-1}(U_{\alpha_2}) \cap \ldots \cap \pi_{\alpha_k}^{-1}(U_{\alpha_k}) = \prod_{\beta\in\Lambda} U_\beta$ where $U_\beta = U_{\alpha_i}$ for $\beta \in \{\alpha_1, \alpha_2 \ldots, \alpha_k\}$ and $U_\beta = X_\beta$ for $\beta \in \Lambda \setminus \{\alpha_1, \alpha_2 \ldots, \alpha_k\}$.

Exercise 45. Prove that $\mathcal{B}$ is a basis for some topology on $\prod_{\alpha\in\Lambda} X_\alpha$. The topology described as above called the product topology.

Exercise 46. Prove that the topology described in Exercise 45 is the smallest topology on $\prod_{\alpha\in\Lambda} X_\alpha$ which makes each projection map π_α continuous.

Exercise 47. Let A be a topological space and $(X_\alpha)_\alpha$ be a collection of topological spaces. Further let $f_\alpha : A \to X_\alpha$ be continuous maps for each $\alpha \in \Lambda$. Define $f : A \to \prod_{\alpha\in\Lambda} X_\alpha$ defined by $f(a) = (f_\alpha(a))_{\alpha\in\Lambda}$. Prove that f is continuous if and only if each f_α is continuous.

Proof. First part follows from the equality $f_\alpha = \pi_\alpha \circ f$.

To prove the converse part it is sufficient to work with any open set of the form $\pi_\alpha^{-1}(U)$ (verify), where U is an open set in X_α. Now $f^{-1}(\pi_\alpha^{-1}(U))$. But $f^{-1}(\pi_\alpha^{-1}(U)) = (\pi_\alpha \circ f)^{-1}(U)$. But since $f_\alpha = \pi_\alpha \circ f$ we have $f^{-1}(\pi_\alpha^{-1}(U)) = f_\alpha^{-1}(U)$ which is an open set in A since each f_α is continuous. This completes the proof. □

Exercise 48. Let $(X_\alpha)_\alpha$ be a collection of Hausdorff topological spaces. Prove that then $\prod_{\alpha\in\Lambda} X_\alpha$ is also Hausdorff.

Exercise 49. Let $A_\alpha \subset X_\alpha$ where for each $\alpha \in \Lambda$, $X_\alpha$is a topological space. Then prove that

$$c\ell(\prod_{\alpha\in\Lambda} A_\alpha) = \prod_{\alpha\in\Lambda} c\ell A_\alpha.$$

SELF ASSESSMENT

1. Let $f : \mathbb{R} \to \mathbb{R}$ be a map and define its graph $G_f : \mathbb{R} \to \mathbb{R}^2$ by $G_f(x) = (x, f(x))$. Show that G_f is continuous and that its image (taken with the topology induced from $\mathbb{R}^2$) is homeomorphic to $\mathbb{R}^1$.

1. Prove that $\mathbb{R}_l \times \mathbb{R}$ are $\mathbb{R} \times \mathbb{R}_l$are homeomorphic.

2. Prove that $(X_1 \times X_2) \times X_3$ are homeomorphic to $X_1 \times (X_2 \times X_3)$.

3. Let A be an nn orthogonal matrix. Prove that $A : \mathbb{R}^n \to \mathbb{R}^n$ is a homeomorphism.

4. Given sequence (a_n) and (b_n) be two sequences of real numbers with $a_i > 0$ for all i. Dene $h : \mathbb{R}^\omega \to \mathbb{R}^\omega$ by $h((a_n)) = ((a_n x_n + b_n))$. If $\mathbb{R}^\omega$ is given product topology prove that h is a homeomorphism.

5. What happens if we consider the box topology in the above example?

6. Whether every convergent sequence in $\mathbb{R}^\omega$ with product topology will converge in box topology?

7. Whether every convergent sequence in $\mathbb{R}^\omega$ with box topology will converge in product topology?

8. Show that the countable collection

$$\{(a, b) \times (c, d) : a < b \text{ and } c < d \text{ and } a, b, c, d \text{ are rationals}\}$$

is a basis for $\mathbb{R}^2$

Summary

In this unit we have learnt about product spaces from existing spaces and related important properties and examples.

Unit 15

Course Structure

1. Metrizable Spaces: Definition and properties

6 Introduction

The concepts we are going to introduce now, arise from a deeper study of topology itself. Such problems as imbedding a given space in a metric space basically problems of topology rather than analysis. In the study of metric space we have observed that a point is a limit point of a set if and only if the the set in question contains a sequence which converges to that point, called sequence lemma. Let us start with the following example which shows that this fact may not hold for arbitrary topological space.

Example 50. An uncountable product of $\mathbb{R}$ with itself is does not posses sequence lemma.

Proof. Let Λ be an uncountable index set; we show that $\mathbb{R}^\Lambda$ does not satisfy the sequence lemma (in the product topology). Let A be the subset of $\mathbb{R}^\Lambda$ consisting of all points $(x_\alpha)_{\alpha\in\Lambda}$ such that $x_\alpha = 1$ for all but finitely many values of α. Let $\overline{0}$ be the "origin" in $\mathbb{R}^\Lambda$, the point each of whose coordinates is 0. We assert that $\overline{0}$ belongs to the closure of $c\ell A$. Let $\prod U_\alpha$ be a basis element containing $\overline{0}$. Then $U_\alpha = \mathbb{R}$, for only finitely many values of α say for $\alpha = \alpha_1, \alpha_2 \ldots, \alpha_k$. Let $(x_\alpha)_{\alpha\in\Lambda}$ be the point of A defined by letting $x_\alpha = 0$ for $\alpha = \alpha_1, \alpha_2 \ldots, \alpha_k$ and $x_\alpha = 1$ for all other values of α; then $(x_\alpha) \in A \cap \prod U_\alpha$ and therefore $\overline{0}$ is a limit point of A.

But we claim that there is no sequence of points of A converging to $\overline{0}$. For let $(a_n)_{n=1}^\infty$ be a sequence of points of A. Given n, let $\Lambda_n = \{\alpha \in \Lambda : \pi_\alpha(a_n)) \neq 1$. The union of all the sets Λ_n is a countable union of finite sets and therefore countable. Because Λ itself is uncountable, there is an index, say β in Λ, such that $\beta \notin \cup_{n\in\mathbb{N}}\Lambda_n$. This means that $\pi_\beta(a_n) = 1$ for all $n \in \mathbb{N}$.

Now let U_β be the open interval $(-1/2, 1/2)$ in $\mathbb{R}$, and consider the open set $\pi_\beta^{-1}(U_\beta)$ in $\mathbb{R}^\Lambda$. The set $\pi_\beta^{-1}(U_\beta)$ is a neighbourhood of $\overline{0}$ that contains none of the points a_n; therefore, the sequence $(a_n)_{n=1}^\infty$ cannot converge to $\overline{0}$. □

6.1 Metrizable Space

One of the important class of topological spaces are those which are generated by some metric. In this section we shall investigate some properties of such topological spaces. We assume that readers already have been already introduced with the definition of metric space (X, d), open ball $B(x, r) = \{y \in X : d(x, y) < r\}$, closed ball $\overline{B_d(x, r)} = \{y \in X : d(x, y) < r\}$. So let us start with the following proposition.

Proposition 51. *Let (X, d) be a metric space. Then the set $\{B(x, r) : r > 0\}$ forms a base for some topology on X.*

Proof. Do yourself. □

Definition 52. A topological space X is said to be metrizable if there exist a metric d on X such that the basis $\{B(x, r) : x \in X, r > 0\}$ generates the topology of the given topological space X.

Let us first observe that metrizability is a topological invariant.

Theorem 53. *Metrizability is a topological invariant.*

Proof. Let $f : X \to Y$ be a homeomorphism and d be a metric on X which generates the topology of X. For any two points u, v in Y we define a metric ρ on Y such that $\rho(u, v) = d(x, y)$ where $u = f(x)$ and $v = f(y)$. It is easy to verify that ρ is a metric on Y. Hence it remains to show that ρ generates the topology of Y. For this let V be an open set in Y and $v \in V$. Then $f^{-1}(v) \in f^{-1}(V)$. Since d generates the topology of X there exist $r > 0$ such that $f^{-1}(v) \in B_d(f^{-1}(v), r) \subseteq f^{-1}(V)$. This implies that $v \in B_\rho(v, r) \subseteq V$. This proves that the metric ρ generates the topology of Y. □

Remark 54. Recall that an isometric or rigid motion between two metric spaces (X, d) and (Y, ρ) is a bijection $f : X \to Y$ which preserves distance between them that is for any $x, y \in X$ we have $\rho(f(x)\, f(y)) = d(x, y)$. In the construction of the above proof the given homeomorphism becomes an isometry. But it is not true in general that homemorphism between two metrizable spaces will always be an isometry, which will be clear from the following discussions.

Definition 55. A subset A of a metric space (X, d) is said to be bounded if there exists some $r > 0$ such that $d(x, y) \leq r$ for all $x, y \in A$. If A is a bounded and non-empty subset of X then diameter of A is defined to be the number

$$\text{diam}A = \sup\{d(x, y) : x, y \in A\}$$

Theorem 56. *Given a metric space (X, d), the $\bar{d} : X \times X \rightarrowtail X$ defined by $\bar{d}(x, y) = min\{d(x, y), 1\}$ is a metric on X which generates the same topology on X as d.*

Proof. Do yourself. □

Remark 57. Justify yourself by the fact that Theorem **??** proves that metrizable spaces may be homeomorphic but not isomorphic.

Example 58. Produce a concrete example of two homeomorphic spaces which are not isometric.

We know that well known euclidean metric on $\mathbb{R}^n$ is defined by the formula $d(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^2\right)^{\frac{1}{2}}$. If we replace 2 by any $p \geq 1$ we again get a metric $\mathbb{R}^n$ that is $d_p(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}$ defines a metric on $\mathbb{R}^n$(verify). $\mathbb{R}^n$ with this metric is generally denoted by l_p^n.

Example 59. Prove that the Euclidean metric d and the metric d_p defined as above generates same topology $\mathbb{R}^n$.

Definition 60. Let us define another metric ρ, $\mathbb{R}^n$ by the formula $\rho(x, y)$=max$\{|x_1 - y_{1|}, |x_2 - y_2|, \ldots, |x_n - y_{n|}|\}$. This is called square metric on $\mathbb{R}^n$.

Example 61. Verify that ρ is infact a metric on $\mathbb{R}^n$.

Observe that on $\mathbb{R}$ both the above metric generate the usual topology of $\mathbb{R}$. The following Theorem shows that this fact can be generalized for any $n \in \mathbb{N}$.

Theorem 62. *Both the metrics d and ρ generates the product topology on $\mathbb{R}^n$.*

Proof. See Munkresh for the complete proof. □

SELF ASSESSMENT

1. Let $1 \leq p < \infty$. Consider the set of all sequences $(x_n)_n$ in $\mathbb{R}$ such that

$$\sum_{n=1}^{\infty} |x_n|^p < \infty$$

Denote this set by l^p. For $x = (x_n)_n$ and $y = (y_n)_n$ in l^p, define

$$d_p(x, y) = \left\{ \sum_{n=1}^{\infty} |x_n - y_n|^p \right\}^{\frac{1}{p}}.$$

Prove that $(X,\ d_p)$ is a metric space.

2. Let l^∞ be the set of all bounded sequences in $\mathbb{R}$; that is,

$$l^\infty = \{(x_{n)_n} \subset \mathbb{R} : \sup_{1 \leq n \leq \infty} |x_n| < \infty\}.$$

For $x\,,\ y \in l^\infty$, define $d_\infty(x, y) = \sup_{1 \leq n \leq \infty} |x_n - y_n|$. Then Prove that d_∞ is a metric on l^∞.

3. Let $C[a, b]$ be the set of all real valued continuous functions defined on $[a, b]$. For $x\,, y\ \in C[a, b]$, define

$$d_\infty(x, y)\ =\ \sup_{t \in [a,b]} |x(t) - y(t)|.$$

Then prove that $(C[a, b]\,,\ d_\infty)$ is a metric space.

4. For $x\,, y\ \in C[a, b]$, define

$$d_1(x, y)\ =\ \int_a^b |x(t) - y(t)| dt.$$

Then prove that $(C[a, b]\,,\ d_1)$ is a metric space.

5. For $x\,, y\ \in C[a, b]$, define

$$d_p(x, y)\ = \left\{ \int_a^b |x(t) - y(t)|^p dt \right\}^{\frac{1}{p}}.$$

Then prove that $(C[a, b]\,,\ d_p)$ is a metric space.

6. Let (X, d) be a metric space. For $x\,, y\, \in X$, define

$$d^*(x, y)\, = \frac{d(x, y)}{1 + d(x, y)}.$$

   Proved that d^* is a metric on X.

7. Let (X, d) be a metric space. Prove that for any $A \subseteq X$ and $x \in X$ the real valued function defined by $f(x) = d(x, A)$, where $d(x, A) = \inf\{d(x, A) : x \in A\}$ is a continuous function.

Summary

In this unit, we have got to know about the metrizability of topological spaces and related properties.

Unit 16

Course Structure

1. Countability Axioms.

7 Introduction

Countability axioms is the common name used to refer to a set of properties of a topological space which have to do with the existence of countable sets, or countable families of open sets, satisfying certain conditions.

They are not axioms in the strict sense of the word, but they are usually named as such because one may think of them as additional basic properties that one can ask from a topological space.

7.1 The Countability Axioms

Definition 63. A space X is said to have a countable basis at x if there is a countable collection $\mathcal{B}$ of neighbourhoods of x such that each neighbourhood of x contains at least one of the elements of $\mathcal{B}$. A space that has a countable basis at each of its points is said to satisfy the first countability axiom, or to be first-countable.

Exercise 64. Give an example of a first countable space which is not second countable.

It is clear that every metrizable space satisfies this axiom for example $\{B(x, 1/n); n \in \mathbb{N}\}$ is a countable basis at x. The most useful fact concerning spaces that satisfy this axiom is the fact that in such a space, convergent sequences are adequate to detect limit points of sets and to check continuity of functions.

Theorem 65. *Let X be a topological space.*

(a) Let A be a subset of X. If there is a sequence of points of A converging to x, then x is a limit point of A; the converse holds if X is first-countable.

(b) Let $f : X \to Y$. If f is continuous, then for every convergent sequence $(x_n)_n$ converging to x in X, the sequence $f(x_n)$ converges to $f(x)$. The converse holds if X is first- countable.

Proof. The main idea of the proof of this theorem is to construct a decreasing sequence of basis at each point x. This follows from the first-countability axioms. In fact let $\mathcal{B}_x = \{U_1, U_2, \ldots, \}$ be a countable base at x. Let $V_1 = U_1$ and let $V_1, V_2, \ldots, V_n$ be constructed such that $V_1 \supseteq V_2 \supseteq \ldots, \supseteq V_n$. Now take $V_{n+1} = V_1 \cap V_2 \cap \ldots, \cap V_n$. Then by induction hypothesis we get a decreasing sequence of neighbourhoods $(V_n)_n$. □

Theorem 66. *A subspace of a first-countable space is first countable, and a countable product of first countable spaces is first-countable. A subspace of a second- countable space is second countable, and a countable product of second-countable spaces is second-countable.*

Proof. Do yourself. □

We can observe that, $\mathbb{R}^n$ has a countable dense subset for any $n \in \omega$. This leads to us the following definition.

Definition 67. A topological space is called separable if it has a countable dense subset.

Example 68. Give an example of a non separable metric space a separable space which is non metrizable.

Proposition 69. *Prove that any separable metric space is second countable.*

Proof. Let (X, d) be a separable metric space and let us take $A = \{a_1, a_2, \ldots\}$ as a countable dense subset of X. Let $\mathcal{B} = \{B(a_i, r) : i \in \mathbb{N} \text{ and } r \in\in \mathbb{Q}^+\}$. Then $\mathcal{B}$ is countable and will be the required countable basis of X (Complete the proof). □

Exercise 70. Whether the above proposition holds for all separable first countable space?

Does it happen for all separable first countable space also?

It is known that $\mathbb{R}^n$ is Lindeloff, that is every open cover of $\mathbb{R}^n$ has a countable subcover. Observe that this property does not hold for every topological space, even not for all metric space. For example uncountable discrete metric space is not Lindeloff.

Definition 71. A topological space X is said to be a Lindeloff space if every open cover of it self has a countable sub-cover.

For example each $\mathbb{R}^n$ is Lindeloff space but uncountable discrete metric space is not Lindeloff. The following theorem shows that every secod countable space satisfies all the countability axioms.

Theorem 72. *If X is a secountable space the the following holds:*
(a) X is first countable;
(b) X is separable;
(c) X is Lindeloff.

Proof. (a) It is obvious.

(b) Take a countable base $\{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_n, \ldots\}$ and construct a set say A choosing a point from each $\mathcal{B}_i$.

(c) Let $\mathcal{A}$ be an open covering of X and

$$J = \{n \in \mathbb{N} : \text{there exists } A \in \mathcal{A}, \text{ with } A \supseteq B_n\}.$$

If $n \in J$, let us denote the corresponding element of $\mathcal{A}$ by A_n and let $\mathcal{A}'$ be the collection of this form. Since J is countable $\mathcal{A}'$ is countable.. Furthermore, it covers X: given a point $x \in X$, we can choose an element A of $\mathcal{A}$, containing x. Since A is open, there is a basis element $\mathcal{B}_n$ such that $x \in B_n \subset A$. Because B_n lies in an element of $\mathcal{A}$, the index n belongs to the set J, so A_n is defined; since A_n contains B_n, it contains x. Thus $\mathcal{A}'$ is a countable subcollection of $\mathcal{A}$, that covers X. □

The following proposition shows that Lindeloffness is closed hereditary property.

Proposition 73. *Closed subspace of Lindeloff space is Lindeloff.*

Proof. Done in the class. □

In the following discussion we shall show that $\mathbb{R}_l$ satisfies all the countability axioms except the second.

First Countability : $[x, x + 1/n) : n \in \mathbb{N}\}$ is a countable local base at x.

Second countability : Let $\mathcal{B}$ be a basis of $\mathbb{R}_l$ and $x \neq y$ in $\mathbb{R}$. Consider basic open sets $[x, x + 1)$ and $[y, y + 1)$. Then there exist B_x and B_y such that $x \in B_x \subset [x, x + 1)$ and $y \in B_y \subset [y, y + 1)$. Obviously $B_x \neq B_y$ and $x = \inf B_x$ and $y = \inf B_y$.

$\mathbb{R}_l$ is Lindeloff : It will suffice to show that every open covering of $\mathbb{R}_l$ by basic open sets contains a countable subcollection covering $\mathbb{R}_l$. Let

$\{[a_\alpha, b_\alpha) : \alpha \in J\}$ be a basic open cover of $\mathbb{R}_l$. Let $C = \cup_{\alpha \in J}(a_\alpha, b_\alpha)$. Then $\{(a_\alpha, b_\alpha) : \alpha \in J\}$ becomes an open cover of C in the usual topology of $\mathbb{R}$. Since any subspace of $\mathbb{R}$ is Lindeloff we have C is Lindeloff. Hence there exists a countable subcover say $\{(a_n, b_n) : n \in \mathbb{N}\}$ of $\{(a_\alpha, b_\alpha) : \alpha \in J\}$. It remains to show that $\mathbb{R} \setminus C$ is countable.

Let x be a point of $\mathbb{R} \setminus C$. We know that x does not belong to any open interval (a_α, b_α); therefore $x = a_\alpha$ for some index β. Choose such a β and then choose q_x be a rational number belonging to the interval (a_α, b_α). Because (a_α, b_α) is contained in C, so is the interval $(a_\alpha, q_x) = (x, q_x)$. It follows that if x and y are two points of $\mathbb{R} \setminus C$ with $x < y$, then $q_x < q_y$. (For otherwise, we would have $x < y < q_y = q_x$, so that y would lie in lhe interval (x, q_x) and hence in C.) Therefore the map $x \to q_x$ of $\mathbb{R} \setminus C$ into $\mathbb{Q}$ is injective, so that $\mathbb{R} \setminus C$ is countable.

Example 74. Using Proposition 73 prove that Linedeloffness is even finitely productive.

SELF ASSESSMENT:

1. A G_δ set in a space X is a set A that equals a countable intersection of open sets of X. Show that in a first-countable T_1 space, everyone-point set is a G_δ set. There is a familiar space in which everyone-point set is a G_δ set, which nevertheless does not satisfy the first countability axiom. What is it ?

2. Show that if X has a countable basis $\{B_n\}$, then every basis $\mathcal{C}$ for X contains a countable basis for X. (Hint: For every pair of indices n, m for which it is possible, choose $C_{n,m} \in \mathcal{C}$ such that $B_n \subset C_{n,m} \subset B_m$).

3. Let X have a countable basis; let A be an uncountable subset of X. Show that uncountably many points of A are limit points of A.

4. Show that every compact metrizable space X has a countable basis. (Hint: Let $\mathcal{A}_n$ be a finite covering of X by $1/n$-balls.).

5. (a) Show that every metrizable space with a countable dense subset has a countable basis.
   (b) Show that every metrizable Lindelof space has a countable basis.

6. Let A be a closed subspace of X. Show that if X is Lindelof, then A is Lindelof. Show by example that if X has a countable dense subset, A need not have a countable dense subset.

7. Show that if X is a countable product of spaces having countable dense subsets, then X has a countable dense subset.

8. Let $f; X \to Y$ be continuous. Show that if X is Lindelof, or if X has a countable dense subset. then $f(X)$ satisfies the same condition.

9. Let $f; X \to Y$ be a continuous open map. Show that if X satisfies the first or the second countability axiom, then $f(X)$ satisfies the same axiom.

10. Show that if X has a countable dense subset, every collection of disjoint open sets in X is countable.

Summary

In this unit, we have mainly learnt about countability axioms and related properties. We have also seen various applications.

Unit 17

Course Structure

1. Regularity of metrizable spaces.
2. Normal Spaces.

8 Introduction

We have already introduced the notion of Hausdorff property. T_1 property is also a separation property in topological spaces. In fact it can be easily observed that a space X is T_1 if and only if for any two distinct points there exist open sets, such that each contains one but not the other. In this section we will introduce some other stronger stronger separation axioms.

Definition 75. A T_1 topological space X is said to be regular if for any point x in X and a closed set K not containing x there exist disjoint open sets U and V such that $x \in U$ and $K \subset V$. Regular space is also called T_3.

So any regular space has this separation property and as well as it is T_1. Since in a T_1 space every singletoned set is closed every regular space is Hausdorff.

The following gives another formulation of regularity in terms of closed neighbourhood.

Proposition 76. *Let X be a T_1 topological space. Then X is regular if and only if given a point x of X and a neighbourhood U of x, there is a neighbourhood V of x such that $x \in V \subset \overline{V} \subset U$.*

Proof. Let X be a T_1 space, which is regular and consider an open neighbourhood U of x. Then $K = X \setminus U$ is a closed set not containing x. Now we have to apply the definition of regularity.
For the converse let the given hypothesis be true and K be a closed set not containing a point x. Then $x \in X \setminus K$, and $X \setminus K$ is open. Then choose a neighbourhood of x contained in $X \setminus K$ with it's closure. □

Theorem 77. *Subspace of regular space is regular.*

Proof. Let X be a regular space and Y be a subspace of it. Y being a subspace of T_1 space X is T_1. Now let $y \in U$ where U is an open set in Y. Then there exists an open set O in X such that $U = O \cap Y$. Therefore by regularity property there exists an open set V in X such that $x \in V \subset \overline{V} \subset O$. Then $V \cap Y$ is open in Y. Also $\mathrm{cl}_Y(V \cap Y) = \mathrm{cl}_X V \cap Y \subset O \cap Y = U$. This proves the claim. □

Theorem 78. *Product of regular spaces is regular.*

Proof. Let $\{X_\alpha : \alpha \in \Gamma\}$ be a collection of regular spaces and $X = \prod_{\alpha\in\Gamma} X_\alpha$. Since product of T_1 spaces is T_1, we have that X is T_1. We use the regularity criteria to prove this result. To prove that X is regular let $B = \prod_\alpha U_\alpha$ be a basic open set containing a point $\mathbf{x} = (x_\alpha)_\alpha$. Then there exists $\alpha_1, \alpha_2, \ldots, \alpha_n$ such that $U_\alpha = X_\alpha$ for all $\alpha \neq \alpha_1, \alpha_2, \ldots, \alpha_n$. Therefore for each α_i there exist there exists open sets V_{α_i} in X_{α_i} such that $x_{\alpha_i} \in V_{\alpha_i} \subset \overline{V_{\alpha_i}} \subset U_{\alpha_i}$. Let us put $V = \prod_\alpha V_\alpha$. Then we know that $\prod_\alpha \overline{V_\alpha} = \overline{\prod_\alpha V_\alpha}$. This proves that $\overline{x} \in V \subset \overline{V} \subset U$. □

It is easy to observe that every regular space is Hausdorff. Recall the K-topology on $\mathbb{R}$, where $K = \{\frac{1}{n} : n \in \mathbb{N}\}$ and the subset

$$(a, b) \setminus K = \{x \in (a, b) : x \neq \frac{1}{n} \text{ for any integer } n \in \mathbb{N}\}$$

of the open interval (a, b): The collection

$$B_1 = \{(a, b) \subset \mathbb{R} : a, b \in \mathbb{R}\} \cup \{(a, b) \setminus K \subset \mathbb{R} : a, b \in \mathbb{R}\}$$

is a basis for a topology on $\mathbb{R}$.
We shall prove that 0 and the closed set K can not be strongly separated in this topology. If so then there exist disjoint open sets U and V containing 0 and V respectively. Choose a basis element containing 0 and lying in U. It must of the form $(a, b) - K$, since each basis element of the form (a, b) containing 0 intersects K, choose n large enough that $\frac{1}{n} \in (a, b)$. Then choose a basis element about $\frac{1}{n}$ contained in V, it must be a basis element of the form (c, d). Finally, choose z so that $z < \frac{1}{n}$ and $z > \max\{c, \frac{1}{n+1}\}$. Then z belongs to both U and V, so they are not disjoint.

Example 79. If X is the real line with the topology generated by the subbase consisting of all the open intervals and the set $\mathbb{Q}$, show that X is Hausdorff but not regular.

9 Regularity of metrizable spaces

To prove that every metrizable space is regular we need the following metric space version of Urysohn lemma.

Lemma 80. *If F is a closed subset of X and G is an open set containing F, then there is a continuous function $f : X \to \mathbb{R}$ such that $0 \leq f(x) \leq 1$ for all x in X, $f(x) = 1$ when $x \in F$, and $f(x) = 0$ when $x \in X \setminus G$.*

Take $f(x) = \frac{d(x,X\setminus G)}{d(x,F)+d(x,X\setminus G)}$.

Now assume we have a metric space (X, d). If F is closed and $x \notin F$, then F and $\{x\}$ are disjoint closed sets. By above Lemma there is a continuous function $f : X \to \mathbb{R}$ with $f(x) = 1$ and $f(y) = 0$ for every y in F. Hence $U = \{y : f(y) > 1\}$ and $V = \{y : f(y) < 1\}$ are disjoint open sets that separate x from F.

10 Normal space

Definition 81. A T_1 topological space X is said to be normal if for any two disjoint closed sets H and K there exist disjoint open sets U and V such that $H \subset U$ and $K \subset V$. Normal spaces are also called T_4.

So any regular space has separation property for disjoint closed sets and as well as as it is T_1. Since in a T_1 space every singletoned set is closed every normal space is regular.
In a short while we shall produce examples of regular spaces which are not normal. We shall also prove by examples that unlike other separation properties normality is not a productive property. The following gives another handy formulation of Normality.

Lemma 82. *Let X be a T_1 topological space. Then X is normal if and only if given a closed set K of X and an open set U containing K, there is an open set V such that $K \subset V \subset \overline{V} \subset U$.*

Proof. Let X be a T_1 space, which is normal and consider an open set U containing K. Then $H = X \setminus U$ is a closed set disjoint from K. Now we have to use the definition.
For the converse let the given hypothesis be true and K and H be disjoint closed sets. Then $H \subset X \setminus K$, and $X \setminus K$ is open. Then choose open set containing H contained in $X \setminus K$ with it's closure. □

Example 83. The space $\mathbb{R}_l$ is an useful example of normal space which is not metrizable. Since the topology of $\mathbb{R}_l$ is finer than that of $\mathbb{R}_u$, it is immediate that singleton sets are closed in $\mathbb{R}_l$. To check normality, suppose that A and B are disjoint closed sets in $\mathbb{R}_l$. For each point a of A choose a basis element $[a, x_a)$ not intersecting B, and for each point b of B choose a basis element $[b, x_b)$ not intersecting A. The open sets

$$U = \bigcup_{a \in A} [a, x_a) \text{ and } V = \bigcup_{b \in B} [b, x_b).$$

are disjoint open sets about A and B, respectively.

We now show that every Metrizable space is normal.

Let X be a metrizable and consider two disjoint closed sets A and B. For each $a \in A$, choose r_a so that the ball $B(a, r_a)$ does not meet B. Similarly, for each b in B, choose r_b so that the ball $B(b, r_b)$ does not intersect A. Define

$$U = \bigcup_{a \in A} B(a, r_a) \quad \text{and} \quad V = \bigcup_{b \in B} B(b, r_b).$$

Then U and V are open sets containing A and B, respectively. For if $z \in U \cap V$ then there exist $a \in A$ and $b \in B$ such that $z \in B(a, \frac{r_a}{2}) \cap B(b, \frac{r_b}{2})$. By triangle inequality we have that $d(a, b) < \frac{r_a + r_b}{2}$. If $r_b \leq r_a$, then $d(a, b) < r_a$, so that the open ball $B(a, r_a)$ contains the point b, a contradiction. Hence U and V are disjoint.

SELF ASSESSMENT:

1. Prove that metric space is regular.

2. Let X be a Hausdorff space and each $x \in X$ has a neighbourhood U such that $\overline{U}$ is regular. Then prove that X is regular.

3. Give an example of non metrizable regular space.

4. If X is the real line with the topology generated by the subbase consisting of all the open intervals and the set $\mathbb{Q}$, show that X is Hausdorff but not regular.

5. Show that if X is regular, every pair of points of X have neighbourhoods whose closures are disjoint.

6. Show that if X is normal, every pair of disjoint closed sets have neighbourhoods whose closures are disjoint.

7. Show that every order topology is regular.

8. Let $f, g : X \mapsto Y$ be continuous; assume that Y is Hausdorff. Show that $\{x : f(x) = g(x)\}$ is closed in X.

Summary

This section has made us acquainted with a special property of the topological spaces, namely the countability axioms which provide us with knowledge of countable sets, countable open sets, etc.

Unit 18

Course Structure

1. Heredity of Normality
2. Linearly Ordered topological space

11 Introduction

We start this section by showing that normality is closed heredity property.As we go further, we will learn about linearly ordered topological spaces.

11.1 Heredity of Normality

Theorem 84. *Closed subspace of a normal spaces is normal.*

Proof. Let X be a normal space and Y be a closed subspace of X. Then any two disjoint closed sets in Y is also closed in X. Then normality of X can be used to prove the normality of Y. □

In the following example we show that normality is not even finitely productive property.

Example 85. In a previous example we have observed that $\mathbb{R}_l$ is normal. Now we shall show that ${\mathbb{R}_l}^2$ is not normal. Which will further produce an example of a regular space which is not normal, as product of regular spaces is regular and $\mathbb{R}_l$ being normal is regular.

If possible let ${\mathbb{R}_l}^2$ be normal. Let L be the subspace of ${\mathbb{R}_l}^2$, consisting of all points of the form $(x, -x)$. Then L is closed in ${\mathbb{R}_l}^2$ and L has the discrete topology.

Hence every subset A of L, being closed in L, is closed in ${\mathbb{R}_l}^2$. Because $L \setminus A$ is also closed in ${\mathbb{R}_l}^2$, this means that for every non-empty proper subset A of L, one can find disjoint open sets U_A and V_A containing A and $L - A$, respectively. Let $D = \mathbb{Q} \times \mathbb{Q}$. Then D is dense in ${\mathbb{R}_l}^2$. We define a map f that assigns, to each subset of the line L, a subset of the set D, as follows :

$$\begin{array}{rcll} f(A) & = & D \cap U_A & \emptyset \subsetneq A \subsetneq L \\ f(\emptyset) & = & \emptyset & \\ f(L) & = & D & \end{array} .$$

We claim that $f : \mathcal{P}(L) \to \mathcal{P}(D)$ is injective.
Let A be a proper non-empty subset of L. Then $f(A) = D \cap U_A$ is neither empty as U_A is open and D is dense in ${\mathbb{R}_l}^2$, nor all of D since $D \cap V_A$ is non-empty. It remains to show that if B is another proper non-empty subset of L, then $f(A) \neq f(B)$. One of the sets A, B contains a point not in the other; suppose that $x \in A$ and $x \notin B$. Then $x \in L - B$, so that $x \in U_A \cap V_B$; since the latter set is open and non-empty, it must contain points of D. These points belong to U_A and not to U_B; therefore, $D \cap U_A \neq D \cap U_B$, as desired. Thus f is injective.

Next as D is countable and L has cardinality of the continuum, there exist a bijection $\varphi : \mathcal{P}(D) \to L$. Then $\varphi \circ f : \mathcal{P}(L) \to L$ is an injective mapping, which is a contradiction.

12 Linearly ordered topological space

Theorem 86. *Every order space is normal.*

Proof. We assert that every interval of the form $(x, y]$ is open in a well-ordered set X. In fact if X has a largest element and y is that element, $(x, y]$ is just a basis element about y. If y is not the largest element of X, then $(x, y]$ equals the open set $(x, \overline{y})$, where $\overline{y}$ is the immediate successor of y. Now let A and B be disjoint closed sets in X, assume for the moment that neither A nor B contains the smallest element a_0 of X. For each $a \in A$, there exists a basis element about a disjoint from B it contains some interval of the form $(x, a]$, as a is not the smallest element of X. Choose, for each $a \in A$, such an interval $(x_a, a]$ disjoint from B. Similarly, for each $b \in B$, choose an interval $(y_b, b]$ disjoint from A. The sets

$$U = \bigcup_{a \in A} (x_a, a] \quad \text{and} \quad V = \bigcup_{b \in B} (y_b, b].$$

are open sets containing A and B, respectively; we claim that they are disjoint. In fact if $z \in U \cap V$. Then $z \in (x_a, a] \cap (y_b, b]$ for some $a \in A$ and some $b \in B$. Assume that $a < b$. Then if $a \leq y_b$, the two intervals are disjoint, while if $a > y_b$, we have $a \in (y_b, b]$, contrary to the fact that $(y_b, b]$ is disjoint from A. A similar contradiction occurs if $b < a$.
Finally, assume that A and B are disjoint closed sets in X and A contains

the smallest element a_0 of X. The set $\{a_0\}$ is both open and closed in X. Then by above there exist disjoint open sets U and V containing the closed sets $A - \{a_0\}$ and B, respectively. Then $U \cup \{a_0\}$ and V are disjoint open sets containing A and B, respectively. □

In the coming section we shall see that Linearly ordered topological spaces satisfies more stronger separation property called completely normal. One of the useful properties of the set $\mathbb{N}$ of positive integers is the fact that each of its nonempty subsets has a smallest element. This property leads to the following definition.

Definition 87. A set X with an order relation $\leq$ is said to be well-ordered if every nonempty subset of X has a smallest element.

Theorem 88. *Every nonempty finite ordered set has the order type of a section $\{1, 2, \ldots, n\}$ of $\mathbb{N}$, so it is well-ordered.*

Theorem 89. *(Well-ordering theorem). If A is a set, there exists an order relation on A that is a well-ordered.*

Corollary 90. *There exists an uncountable well-ordered set.*

For any well ordered set X and given $\alpha \in X$, let S_α let us denote the set $S_\alpha = \{\beta \in X : \beta < \alpha\}$. It is called the section of X by α.

Theorem 91. *There exists a well-ordered set A having a largest element Ω, such that the section S_Ω of A by Ω is uncountable but every other section of A is countable.*

In fact if B be an uncountable well-ordered and S be the well-ordered set $\{1, 2\} \times B$ in the dictionary order; then some section of C is uncountable. Let Ω be the smallest element of C for which the section of C by Ω is uncountable. Then let A consist of this section along with the element Ω.

Note that S_Ω is an uncountable well-ordered set every section of which is countable. Its order type is in fact uniquely determined by this condition. We shall denote the well-ordered set $A = S_\Omega \cup \{\Omega\}$ by the symbol $S_\Omega + 1$.

The most useful property of the set S_Ω for our purposes is expressed in the following theorem:

Theorem 92. *If A is a countable subset of S_Ω, then A has an upper bound in S_Ω.*

With these few properties of ordinal numbers, which will be useful for us for further discussions, we end this section.

SELF ASSESSMENT:

1. Prove that every finite Hausdorff space is normal.

2. Prove that a metric space is normal space.

3. Give an example of non metrizable normal space.

4. Is normality preserved in a finer topological space?

5. Let $p : X \mapsto Y$ be a closed continuous surjective map. Show that if X is normal, then so is Y.

6. Let Y be normal and $F_1, F_2, \ldots, F_n$ be closed subsets such that $\cap_{i=1}^{n} F_i = \emptyset$. Prove that there exist open sets $V_i \supset F_i$ such that $\cap_{i=1}^{n} \overline{V_i} = \emptyset$.

7. Let X be normal, $A \subset X$ be closed, and U an open set containing A. Prove that there exists an open F_σ-set V such that $A \subset V \subset X$.

8. Let X be the upper half of $\mathbb{R}^2$ including x-axis. Give the portion $\{(x, y) : y > 0\}$ the subspace topology induced from Euclidean topology. Let us define the neighbourhood points of $(x, 0)$ to be $\{(x, 0)\} \cup \{(x, y) : y > 0\}$ tangent to the x-axis at $(x, 0)$. Prove that this space is normal.

Summary

In this unit, we have mainly seen properties of normality; linearly ordered topological spaces and their properties.

Unit 19

Course Structure

1. Product of Normal Spaces

13 Introduction

We have already proved that $\mathbb{R}_l \times \mathbb{R}_l$ is not normal. In the following we shall provide another example to show product of normal spaces may not be normal. This example is taken from J. R. Munkres.

Example 93. Consider the well-ordered set $S_\Omega + 1$, in the order topology, and consider the subset S_Ω, in the subspace topology (which is the same as the order topology). Both spaces are normal, as linearly ordered topological spaces are normal. We wish to prove that $(S_\Omega + 1) \times S_\Omega$ is not normal.

First, we consider the space $(S_\Omega + 1) \times (S_\Omega + 1)$, and its "diagonal" $\triangle = \{(x, x) : x \in S_\Omega + 1\}$. Because $S_\Omega + 1$ is Hausdorff, $\triangle$ is closed in $(S_\Omega + 1) \times (S_\Omega + 1)$. If U and V are disjoint neighbourhoods of x and y, respectively, then $U \times V$ is a neighbourhood of $x \times y$ that does not intersect $\triangle$. Therefore, in the subspace $S_\Omega \times (S_\Omega + 1)$, the set

$$A = \triangle \cap (S_\Omega \times (S_\Omega + 1)) = \triangle - \{\Omega \times \Omega\}$$

is closed. Similarly the set $B = S_\Omega \times \{\Omega\}$ is closed in $S_\Omega \times (S_\Omega + 1)$. The sets A and B are disjoint.

Assuming there exist disjoint open sets U and V of $S_\Omega \times (S_\Omega + 1)$ containing A and B, respectively, we shall derive a contradiction.
Given $x \in S_\Omega$, consider the vertical slice $x \times (S_\Omega + 1)$. We assert that there is some point β with $x < \beta < \Omega$ such that (x, β) lies outside U. In fact if U is contained all points (x, β) for $x < \beta < \Omega$ then the top point (x, Ω) of the slice would be a limit point of U, which it is not because V is an open set disjoint from U containing this top point
Choose $\beta(x)$ to be such a point; just to be definite, let $\beta(x)$ be the smallest element of $S_(\Omega)$ such that $x < \beta(x) < \Omega$ and $(x, \beta(x))$ lies outside U.

Let us define a sequence of points of $S_(\Omega)$ as follows: Let x_1 be any point of $S_(\Omega)$. Let $x_2 = \beta(x_1)$, and in general,$x_{n+1} = \beta(x_n)$. We have

$$x_1 < x_2 < \ldots,$$

because $\beta(x) > x$ for all x.

The set $\{x_n\}$ is countable and therefore has an upper bound in $S_(\Omega)$; let $b \in S_\Omega$ be its least upper bound. Because the sequence is increasing, it must converge to its least upper bound; thus $x_n \to b$. But $\beta(x_n) = x_n + 1$ so that $\beta(x_n) \longrightarrow b$ also. Then

$$(x_n, \beta(x_n) \longrightarrow (b, b)$$

in the product space. Now we have a contradiction, for the point (b, b) lies in the set A, which is contained in the open set U; and U contains none of the points $(x_n, \beta(x_n)$.

Now we shall enter into some deeper study of normal spaces. Two important Theorems to note here are Urysohn Lemma ant Tietz extension lemma. The first one gives information about rich source of continuous functions and the second one tells about the extension of continuous functions. Let us start with the following theorem

Theorem 94. *Every regular space with a countable basis is normal.*

Proof. Let X be a regular space with a countable basis say $\mathcal{B} = \{B_n; n \in \mathbb{N}\}$. Let A and B be disjoint closed subsets of X. The idea of the proof is to cover A with some basic open sets and to cover B with some basic open sets and then to to eliminate the overlapped parts. Each point x of A has a neighbourhood U not intersecting B. Using regularity, choose a neighbourhood V of x whose closure lies in U; we can do these operations taking elements from $\mathcal{B}$. By choosing such a basis element for each x in A, we construct a countable covering of A by open sets whose closures do not intersect B. Since this covering of A is countable, we can index it with the positive integers; let us denote it by $\{U_n\}$. Similarly, choose a countable collection $\{V_n\}$ of open sets covering B, such that each set V_n is disjoint from A. The sets $U = \bigcup_n U_n$ and $V = \bigcup_n V_n$ are open sets containing A and B. Given $n \in \mathbb{N}$, define

$$U'_n = U_n \setminus \left(\bigcup_n \overline{V}_n \right) \text{ and } V'_n = V_n \setminus \left(\bigcup_n \overline{U}_n \right).$$

Then each U'_n and V'_n are open sets. It is easy to observe that $\{U'_n\}$ is a cover of A and $\{V'_n\}$ is a cover of B. The collection $\{U'_n\}$ is a cover of A because each x in A belongs to U_n for some n, and x belongs to none of the sets $\overline{V}_i$. Similarly, the collection $\{V'_n\}$ covers B. Now let us set

$$U' = \bigcup_n U'_n \text{ and } V' = \bigcup_n V'_n.$$

It remains to show that U' and V' are disjoint. For if $x \in U' \cap V'$ then $x \in U'_j \cap V'_k$ for some j and k. Suppose that $j < k$. It follows from the definition of U' that $x \in U_j$, and since $j \leq k$ it follows from the definition of V'_k that $x \notin \overline{U}_j$. A similar contradiction arises if $j \geq k$. □

As an immediate application of the above Theorem we can say that real line with usual topology is normal.

Quite same construction like the above Theorem proves the following Theorem.

Theorem 95. *Every regular Lindelöf space is normal.*

Next we shall introduce another notion of separability. We know that normality is not heredity property. But there are normal spaces all whose subspaces are normal. This motivates the following definition.

Definition 96. A space X is said to be completely normal if every subspace of it is normal.

Definition 97. In a space X two sets A and B will be said separable if $\overline{A} \cap B = \emptyset$ and $A \cap \overline{B} = \emptyset$.

Theorem 98. *A space is completely normal iff every pair of separated subsets can be separated by neighbourhoods.*

Proof. Suppose A and B is a pair of separated subsets of X. Then $Y = X - (\overline{A} \cap \overline{B})$ is an open subset of X that contains both A and B. $\overline{A}_Y \cap \overline{B}_Y = Y \cap \overline{A} \cap \overline{B} = \emptyset$. Thus, $\overline{A}_Y$ and $\overline{B}_Y$ can be separated by open neighbourhoods in Y. Since Y is open, these neighbourhoods are also open in X .
Conversely take a set $Y \subset X$ and two disjoint subsets $A, B \subset Y$ closed in Y. $\overline{A}_X \cap B = \overline{A}_X \cap Y \cap B = \overline{A}_Y \cap B = \emptyset$. Similarly, $\overline{B}_X \cap A = \emptyset$. Therefore, A and B can be separated by neighbourhoods in X and their intersections with Y separate A and B in Y. □

In the following we shall examine the complete normality of some spaces.

Exercise 99. Does every subspace of a completely normal space is completely normal?

Let Y be a subspace of a completely normal space X. Then any subspace of Y is a subspace of the completely normal space X therefore, is normal. This means that Y is completely normal.

Exercise 100. Is every well-ordered set is Completely normal in the order topology.

If a set is well-ordered then every element has a successor and $\{(a, x]\}$ is a basis at x where $a < x$ or $a = -\infty$. Therefore, for a pair of separated sets we can cover each set with such neighbourhoods that do not intersect the other set. Moreover, the neighbourhoods belonging to one set do not intersect the neighbourhoods belonging to the other set.

Exercise 101. Is $\mathbb{R}_l$ Completely normal?

Indeed, the proof of the fact that $\mathbb{R}_l$ is Completely normal, is extremely similar to the previous example: both use the fact that there is a basis at x with sets of the form $[x, a)$. This shows that every point in one closed set has such a neighbourhood that does not intersect the other set. Then coverings by such basis sets are disjoint automatically. So, we obtain the disjoint open neighbourhoods for any pair of sets such that neither one contains limit points of the other one.

Exercise 102. Does every metric space Completely normal?

Every subspace is metrizable as well, therefore, normal

Exercise 103. Does every regular space with a countable basis is completely normal?

Every regular second-countable space is normal. Also every its subspace is also regular and second-countable. Therefore, every regular second-countable space is completely normal.

Exercise 104. The product of two completely normal spaces.

$\mathbb{R}_l$ is completely normal, but $(\mathbb{R}_l)^2$ is not even normal.

SELF ASSESSMENT:

1. Prove that closed subspace of normal space is normal.

2. Prove if product is normal space then each factor space is normal.

3. Give an example of non metrizable normal space.

4. Prove that every regular Lindelöf space is normal.

5. Is $\mathbb{R}^\omega$ normal in product topology?

6. Is $\mathbb{R}^\omega$ normal in uniform topology?

7. Prove that unit interval $[0, 1]$ is completely normal.

8. If J is an uncountable set prove that $\mathbb{R}^J$ is not normal. (This is an extremely difficult question. See J. R. Munkres for proof)

Summary

In this unit, we have become acquainted with product of normal spaces and various relevant properties.

Unit 20

Course Structure

1. Urysohn's Lemma
2. Completely Regular Space

14 Introduction

Now we shall gradually move to Urysohn Metrization Theorem. In this course we shall require the famous Urysohn's Lemma. We have already given a version of Urysohn's lemma for metric space. But that depends completely on the metric. In this module we shall present a general version of Urysohn's lemma.

14.1 Urysohn's Lemma

Theorem 105 (Urysohn's Lemma)**.** *A space X is normal if and only if for any two disjoint closed sets E and F there exists a continuous functions $f : X :\to \mathbb{R}$ such that*

$$f(E) = 0 \text{ and } f(F) = 0.$$

Proof. Set $V = X \setminus F$, an open set containing E. Then by normality criteria there exists an open set $U_{\frac{1}{2}}$ such that

$$E \subset U_{\frac{1}{2}} \subset \overline{U_{\frac{1}{2}}} \subset V.$$

Successive application of normality open sets $U_{\frac{1}{4}}$ and $U_{\frac{3}{4}}$ such that

$$E \subset U_{\frac{1}{4}} \subset \overline{U}_{\frac{1}{4}} \subset U_{\frac{1}{2}} \subset \overline{U_{\frac{1}{2}}} \subset U_{\frac{3}{4}} \subset \overline{U}_{\frac{3}{4}} \subset V.$$

Continuing in this manner, we have for each dyadic rational number $r \in (0, 1)$, an open set U_r such that

$$U_r \subset U_s,\ 0 < r < s < 1,$$

$$E \subset U_r,\ 0 < r < 1,$$

$$U_r \subset V,\ 0 < r < 1.$$

With this information we shall now go to define the function f.

$$f(x) = \begin{cases} 0 & \text{if } x \in U_r \text{ for all } r > 0 \\ \sup\{r : x \notin U_r\} & \end{cases}$$

Evidently $0 \leq f \leq 1$, $f = 0$ on E and $f = 1$ on F. It suffices to show that f is continuous.

Let $x \in X$. For convenience, we assume that $0 < f(x) < 1$, the case $f(x) = 0$ and $f(x) = 1$ are not difficult. Let $\epsilon > 0$. Choose dyadic rational number $0 < r < s < 1$ and

$$f(x) - \epsilon < r < f(x) < f(x) + \epsilon.$$

Then $x \notin U_t$ for dyadic rational numbers between r and $f(x)$, so that $x \notin \overline{U}_r$. On the other hand $x \in U_s$. Hence $W = U_s \setminus \overline{U}_r$ is an open neighbourhood of x. If $y \in W$, then from the definition of f we see that $r \leq f(x) \leq s$. In particular, $|f(x) - f(y)| < \epsilon$ for $y \in W$, so that f is continuous at x. □

This is a deep Theorem, both from the point of view of its proof, which involves really original idea, and also from the point of view of its application.

Recall that A is a G_δ set in a space X if A is the intersection of a countable collection of open sets of X. In metric space all closed sets are G_δ sets. But this is not true in general. In normal space we have the following Theorem.

Theorem 106. *Let X be normal space. There exists a continuous function $f : X \to [0, 1]$ such that $f(x) = 0$ for $x \in A$, and $f(x) > 0$ for $x \notin A$, if and only if A is a closed G_δ set in X.*

Proof. Suppose there exists a continuous function $f : X \to [0, 1]$ such that $f(x) = 0$ for $x \in A$, and $f(x) > 0$ for $x \notin A$. Then $A = f^{-1}(0)$ must be closed. Now $A = \bigcap_n f^{-1}(-\frac{1}{n}, \frac{1}{n})$ and each $f^{-1}(-\frac{1}{n}, \frac{1}{n})$ is an open set. Hence A is a G_δ set also.

Conversely let A be a closed G_δ set. Then there exists a sequence (U_n) of open sets such that $A = \bigcap_n U_n$. Then for each n there exists a continuous function f_n which vanishes on A and equal to 1 on $X - U_n$. Now take

$$f = \sum_n \frac{|f_n|}{2^n}.$$

Then clearly f is continuous and serves our purpose. □

Now we are in a position to prove the coveted Urysohn Metrization Theorem.

Theorem 107. *Every regular second countable space is metrizable.*

Proof. We know that any regular second countable space X is normal so that we can invoke Urysoh's lemma. Let $\mathcal{B} = \{B_n : n \in \mathbb{N}\}$ be a countable base. Then for any $m, n \in \mathbb{N}$ if $\overline{B_m} \subset B_n$ there exists some $f \in C^*(X)$ such that $f(\overline{B_m}) = 0$ and $f(B_n) = 1$. In this way we get a countable collention of functions say $\{f_n : n \in \mathbb{N}\}$ which does the above job. We define $e : X \to \mathbb{R}^{\mathbb{N}}$ to be $\pi_n(e(x)) = f_n(x)$. We shall prove that e is an embedding. Continuity follows from the fact that each factor map f_n is continuous. If $x \neq y$ then there exists $B_m, B_n \in \mathcal{B}$ such that $x \in B_m$, $y \notin B_n$ and $\overline{B_m} \subset B_n$. So there exists some f_k such that $f_k(x) = 0$ and $f_k(y) = 1$. It's remain to show that $e : X \to e(X)$ is open. Let U be open in X, then we need to show that set $e(U)$ is open in $e(X)$. Let $v \in e(U)$. We have to find an open set W of $e(X)$ such that $v \in W \subset e(U)$. Let u be the point of U such that $e(u) = v$. Choose an index N for which $f_N(u) > 0$ and $f_N(X - U) = \{0\}$. Take the open ray $(0, +\infty)$ in $\mathbb{R}$, and let V be the open set $\pi_N^{-1}((0, +\infty))$ of $\mathbb{R}^{\mathbb{N}}$. We claim that $v \in W \subset e(U)$ First, $v \in W$ because $\pi_N(u) = \pi_N(e(u)) = f_N(u) > 0$. Second, $W \subset e(U)$. For if $z \in W$, then $z = e(x)$ for some $x \in X$, and $\pi_N(z) \in (0, +\infty)$. Since $\pi_N(z) = \pi_N(e(x)) = f_N(x)$, and f_N vanishes outside U the point x must be in U. Then $z = e(x)$ is in $e(U)$ as desired.

Thus e is an imbedding of X in $\mathbb{R}^{\mathbb{N}}$. □

14.2 Completely Regular Space

Definition 108. If E and F be two disjoint closed sets in a space X and there exists a continuous functions $f : X :\to \mathbb{R}$ such that $f(E) = 0$ and $f(F) = 0$. We say that E and F can be separated by a continuous function.

The Urysohn lemma says that if every pair of disjoint closed sets in X can be separated by disjoint open sets, then each such pair can be separated by a continuous function. The converse is trivial, for if

$$f(E) = 0 \text{ and } f(F) = 0.$$

is the function, then $f^{-1}[0, \frac{1}{3})$ and and $f^{-1}(\frac{2}{3}, 0]$ are disjoint open sets containing A and B, respectively.

Theorem 109. *A subspace of a completely regular space is completely regular.*

Proof. Let X be a completely regular space and let Y be a subspace of X. Let x be a point of Y, and let K be a closed set of Y disjoint from x. We choose a closed set in X such that $K = H \cap Y$. Therefore, $x \notin H$. Since X is completely regular, we can choose a continuous function $f : X \to \mathbb{R}$ such that $f(x) = 1$ and $f(H) = 0$. The restriction of f to Y is the desired continuous function on Y. □

Theorem 110. *Product of completely regular spaces is completely regular.*

Proof. Let $X = \prod X_\alpha$ be a product of completely regular spaces. Let $x = (x_\alpha)$ be a point of X and let U be an open set of X containing x. Then $U = \prod U_\alpha$, where each U_α is open in X_α and $U_\alpha = X_\alpha$ except for finitely many α's, say $\alpha_1, \alpha_2, \ldots, \alpha_k$. Then for each α_i, we can choose a continuous function $f_i : X \to \mathbb{R}$ such that $f_i(x_{\alpha_i}) = 1$ and $f_i(X_{\alpha_i} \setminus U_{\alpha_i}) = 0$. Now, let us set $\phi_i = f_i \circ \pi_{\alpha_i}$. Then each ϕ_i is continuous from X to $\mathbb{R}$. If we set $\phi = \phi_i.\phi_2 \ldots \phi_k$, then it is easy to observe that $\phi(x) = 1$ and $\phi(X \backslash U) = 0$. □

SELF ASSESSMENT

1. Give a direct proof of Urysohn's Lemma for metric spaces.

2. Give an example showing that a second countable Hausdorff space need not satisfy Urysohn's Lemma.

3. Give an example showing that a second countable space need not be metrizable.

4. Is every regular Lindelöf space metrizable?

5. Give an example to show that a completely normal, first countable, Lindelöf separable space need not be metrizable.

6. Prove that every F_σ subset of a normal space is normal.

7. Show that a regular Lindelöf space is metrizable if it is locally metrizable, where a space X is locally metrizable if each point x of X has a neighbourhood that is metrizable in the subspace topology.

8. Show that a space X is completely regular if and only if it is homeomorphic to a subspace of $[0,1]^J$ for some J.

Summary

In this unit, we have learnt about Urysohn's Lemma and Completely regular spaces along with their properties.

References

[1] M A Armstrong. Basic Topology. Springer, 1983.

[2] Dugundji. Topology. Allyn and Bacon, Boston, 1966.

[3] TW Gamelin and RE Greene. Introduction to Topology. 2nd ed. Dover Publications, 1999.

[4] P. R. Halmos. Naive Set Theory. Van Nostrand Reinhold Co., New York, 1960.

[5] K J¨anich. Topology. Springer, 1984.

[6] J. L. Kelley. General Topology. Springer-Verlag, New York, 1991.

[7] J. R. Munkres. Elements of Algebraic Topology. Perseus Books, Reading, Mass., 1993.

[8] George E. Simmons. Introduction to topology and modern analysis. McGraw Hill, New York, 1963.

[9] S. Willard. General Topology. Addison-Wesley Publishing Company, Inc., Reading, Mass., 1970.

Students␣are␣also␣suggested␣to␣visit␣the␣followings␣web␣links
1.␣http://www.topologywithouttears.net/topbook.pdf
2.␣https://nilesjohnson.net/point-set-topology-topics.html
3.␣http://www-bcf.usc.edu/~sheelgan/math440/
4.␣http://mathworld.wolfram.com/Point-SetTopology.html

POST GRADUATE DEGREE PROGRAMME (CBCS)

M.SC. IN MATHEMATICS

SEMESTER I

SELF LEARNING MATERIAL

PAPER : DSE 1.4
(Applied Stream)

Mechanics of Solids
Non-linear Dynamics

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

Course Preparation Team

1. Dr. Samares Pal Professor Department of Mathematics University of Kalyani	2. Dr. Amiya Das Assistant Professor Department of Mathematics University of Kalyani
3. Dr. Biswajit Mallick Assistant Professor (Cont.) DODL, University of Kalyani	4. Ms. Audrija Choudhury Assistant Professor (Cont.) DODL, University of Kalyani

December, 2021

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 weel as each and every member of PG-BOS (DODL), University of Kalyani, Heartfelt thanks is also due to the Course Writers-faculty members at the DODL, subject-experts serving at University Post Graduate departments and also to the authors and academicians whose academic contributions have enriched the SLMs. We humbly acknowledge their valuable academic contributions. I would especially like to convey gratitude to all other University dignitaries and personnel involved either at the conceptual or operational level of the DODL of University of Kalyani. Their persistent and co-ordinated efforts have resulted in the compilation of comprehensive, learner-friendly, flexible texts that meet the curriculum requirements of the Post Graduate Programme through Distance Mode. Self Learning Materials (SLMs) have been published by the Directorate of Open and Distance Learning, University of Kalyani, Kalyani-741235, West Bengal and all the copyright reserved for University of Kalyani. No part of this work should be reproduced in any from without permission in writing from the appropriate authority of the University of Kalyani.

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

Director

Directorate of Open and Distance Learning

University of Kalyani

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

APPLIED STREAM

DSE 1.4
Marks : 100 (SEE : 80; IA : 20); Credit : 6

Mechanics of Solids (Marks : 50 (SEE: 40; IA: 10))
Non-linear Dynamics (Marks : 50 (SEE: 40; IA: 10))

Syllabus

Block I

- **Unit 1:** Brief discussion of tensor transformation, symmetric tensor, alternating tensor. Analysis of strain, Normal strain, shearing strain and their geometrical interpretations
- **Unit 2:** Strain quadratic of Cauchy, Principal strains, Invariants, Saint-Venant's equations of compatibility, equivalence of Eulerian and Lagrangian components of strain in infinitesimal deformation
- **Unit 3:** Analysis of stress, stress tensor, Equations of equilibrium and motion. Stress quadric of Cauchy. Principal stress and invariants, strain energy function
- **Unit 4:** Graphical representation of elastic deformation. Equations of elasticity. Generalized Hooke's law. Homogeneous isotropic media. Elastic modulii for isotropic media.
- **Unit 5:** Equilibrium and dynamical equations for an isotropic elastic solid.Connections of the strain energy function with Hooke's Law, uniqueness of solutions. Clapeyron's Theorem, Beltrami-Michell compatibility equations, Saint-Venant's principle.
- **Unit 6:** Equilibrium of isotropic elastic solid: Deformations under uniform pressure. Deformations of prismatical bar stretched by its own weight and a cylinder immersed in a fluid, twisting of circular bar by couples at the ends
- **Unit 7:** Torsion : Torsion of cylindrical bars, Torsional rigidity, Torsion function, Lines of shearing stress, simple problems related to circle, ellipse and equilateral triangle
- **Unit 8:** Two-dimensional problems: Plane strain, Plane stress, Generalised plane stress, Airy's stress function, General solution of biharmonic equation.

- **Unit 9:** Stresses and displacements in terms of complex potentials. Simple problems, stress function appropriate to problems of plane stress
- **Unit 10:** Waves: Propagation of waves in an isotropic elastic medium, waves of dilatation and distortion. Plane waves

Block II

- **Unit 11:** Linear autonomous systems: Linear autonomous systems, existence, uniqueness and continuity of solutions, diagonalization of linear systems,
- **Unit 12:** Fundamental theorem of linear systems, the phase paths of linear autonomous plane systems
- **Unit 13:** Complex eigen values, multiple eigen values, similarity of matrices and Jordon canonical form, stability theorem
- **Unit 14:** Reduction of higher order ODE systems to first order ODE systems, linear systems with periodic coefficients
- **Unit 15:** Linearization of dynamical systems: Two, three and higher dimension.
- **Unit 16:** Population growth. Lotka-Volterra system
- **Unit 17:** Stability: Asymptotic stability (Hartman's theorem), Global stability (Liapunov's second method)
- **Unit 18:** Limit set, attractors, periodic orbits, limit cycles
- **Unit 19:** Bendixon criterion, Dulac criterion, Poincare-Bendixon Theorem.
- **Unit 20:** Stability and bifurcation: Saddle-Node, transcritical and pitchfork bifurcations. Hopf- bifurcation

Contents

Unit 1

Course Structure

- Preliminaries to continuum mechanics
- Body and surface force
- Analysis of stress, stress Vector and stress tensor
- Cauchy's fundamental theorem for stress
- Equation of equilibrium

1.1 Notion of a continuum

Materials such as solids, liquids and gases are composed of molecules separated by "empty" space. On a microscopic scale, materials have cracks and discontinuities. However, certain physical phenomena can be modeled assuming that the materials exist as a continuum. Continuum means the matter in the body is continuously distributed and fills the entire region of space it occupies with no empty space. This ensures that it possesses unique physical properties such as unique density, unique displacement, unique velocity at every point of space which can be expressed as continuous functions of position and time. A continuum is a body that can be continually subdivided into infinitesimal elements with properties being those of the bulk materials.

Thus matter is idealized as a continuum, which has two properties:

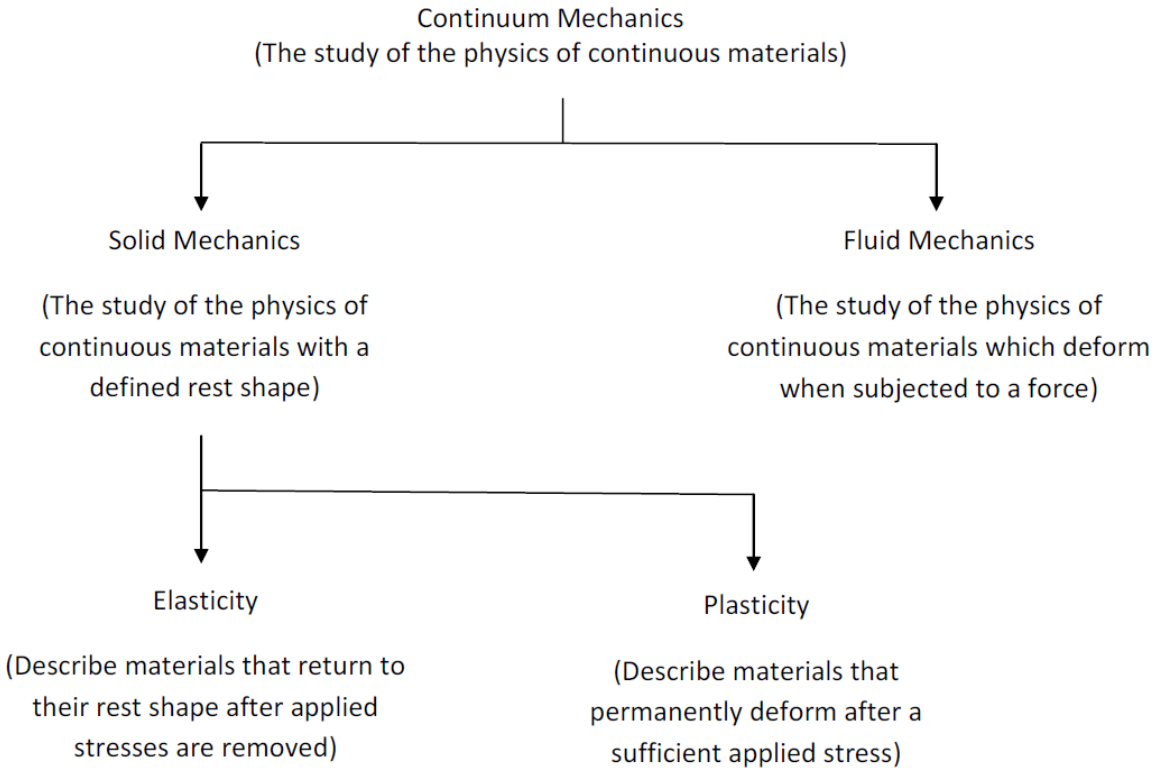
- it is infinitely divisible and
- it is locally homogeneous, in other words if we subdivide it sufficiently many times, all subdivisions have identical properties (e.g. mass, density etc).

A continuum can be thought as an infinite set of vanishingly small particles connected together.

1.2 Continuum Mechanics

Continuum mechanics, a scientific discipline, is a branch of mechanics that deals with the analysis of the kinematics and the mechanical behaviour of substances under the influence of external agents that produce changes in the state of medium. These changes may appear in the form of contact forces such as chemical, electrical, mechanical or any other type of disturbances.

The major areas of continuum mechanics are:



1.3 Configuration

The configuration of a solid is a region of space occupied by the solid. When we describe motion, we namely choose some convenient configuration of the solid to use as reference. This is often the initial, undeformed solid but it can be any convenient region that could be occupied by the solid. The material changes its shape under the action of external loads and at some time t occupies a new region which is called the deformed or current configuration of the solid.

1.4 Deformation

Let a continuum body occupies a certain region of space B_0 at time t_0. When external forces are applied to the body, the material points of B_0 move so that they occupy some other region of space

B after time t. Consequently there are changes in the positions of all the material points of the body. The body is then said to be deformed and the transformation of the body from its initial configuration to subsequent configuration is called deformation. There are two types of deformations:

- rigid body deformation
- strain deformation

1.4.1 Rigid body deformation

When the deformation is such that there are no changes in the relative positions of constituent material points of the continuum firmly bound together so that the length of any line joining any two material points does not change, then the deformation is a combination of translation and rotation about an axis at any point causing a change in configuration and orientation of the body only and is called rigid body deformation and the body is called rigid body.

1.4.2 Strain deformation

When the deformation is such that there are changes in the relative positions of constituent material points of the continuum body so that the length and orientation of any line joining any two points changes, then the deformation causes a change in the shape of the body only and is called strain deformation and the body is called deformed body. The existence of strain deformation depends on the occurrence of relative displacement of points in the medium with respect to each other.

In a continuum body, a deformation field results from a stress field induced by applied forces or is due to changes in temperature field inside the body. The relation between stresses and induced strains is expressed by constitutive equations, e.g., Hooke's law for linear elastic materials. Deformations which are recovered after the stress field has been removed are called elastic deformation. In this case, the continuum completely recovers its original configuration. On the other hand, irreversible deformations remain even after stresses have been removed. One type of irreversible deformation is plastic deformation, which occurs in material bodies after stresses have attained a certain threshold value known as the elastic limit.

1.5 Linear elastic solid or Hookean solid

By linear elastic solid, we mean continuous materials which undergoes very small change of shape when subjected to forces of reasonable magnitude. It has the property that the body recovers its original shape upon the removal of forces causing deformation provided the forces are not too large. It is restricted to the case in which the deformation and gradients are small. Linear elastic solid shows the linear relations between the stress components and strain components.

1.6 Introduction to Stress

Stress is a measure of force intensity, either within or on the boundary surface of a body subjected to loads. It should be noted that in continuum mechanics a body is considered to be stress free if

the only forces present are those interatomic forces required to hold the body together. Therefore it follows that the stresses that concern us here are those which result from the application of forces by an external agent.

1.7 Body and Surfaces forces

The motion of a material body is produced by the action of externally applied forces which are assumed to be of two kinds:

- body forces (F_B),
- surface forces (F_C).

1.7.1 Body force

Body forces are forces originating from sources outside of the body that act on all volume elements (or mass) of the body and distributed throughout the body. These forces arise from the presence of the body in force fields, e.g., gravitational field or electromagnetic field, from inertial forces when bodies are in motion. Body forces are specified by vector fields which are assumed to be continuous over the entire volume of the body i.e. acting on every point in it. The total body force applied to a continuous body is expressed as

$$F_B = \iiint_V \rho b \, \mathrm{d}\, V \,, \tag{1.7.1}$$

where ρ denotes force per unit volume and b is force per unit of mass.

1.7.2 Surface force

The forces which act upon and are distributed in some fashion over a surface element of the body regardless of whether that element is part of the bounding surface or an arbitrary element of surface within the body are called surface forces or contact forces. Surface forces are expressed as force per unit area.

The total contact force on the particular internal surface S is then expressed as the sum of the contact forces on all differential surfaces $\mathrm{d}\, S$ as

$$F_C = \iint_S T^{(n)} \, \mathrm{d}\, S \,. \tag{1.7.2}$$

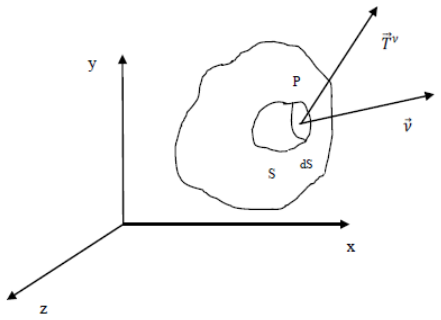
Hence the total force F applied to a body on a portion of the body can be expressed as

$$F = F_B + F_C = \iiint_V \rho b \, \mathrm{d}\, V + \iint_S T^{(n)} \, \mathrm{d}\, S \,. \tag{1.7.3}$$

Examples of body forces are gravitational and magnetic forces, while that of surface forces are hydrostatic pressure of liquid or pressures of one solid body on another due to contact.

1.8 Stress vector and stress tensor

The concept of stress arises from the consideration of the internal forces which the particles of one part of the deformed body exerts on the particles of the adjacent part through the separating boundary surface in the form of restoring forces.



Consider a continuous medium B occupy a volume V at some time t. Imagine a closed surface S within V. Let $\mathrm{d}\,S$ be a small elementary area surrounding a point P of the surface. Let us draw a normal $\vec{\nu}$ at P to the surface element $\mathrm{d}\,S$ in a specified sense. The components of surface force across $\mathrm{d}\,S$ which the material on the side of $\mathrm{d}\,S$ towards which normal $\vec{\nu}$ is drawn exerts on the material on the other side are expressed by $\tau_{\nu x}\,\mathrm{d}\,S$, $\tau_{\nu y}\,\mathrm{d}\,S$, $\tau_{\nu z}\,\mathrm{d}\,S$. These are the components of forces along x, y, z axes respectively. If the direction of $\vec{\nu}$ is in the direction of x axis, the components of surface force are $\tau_{xx}\,\mathrm{d}\,S$, $\tau_{xy}\,\mathrm{d}\,S$, $\tau_{xz}\,\mathrm{d}\,S$. If $\vec{\nu}$ is drawn in the direction of y axis, the components of surface force are $\tau_{yx}\,\mathrm{d}\,S$, $\tau_{yy}\,\mathrm{d}\,S$, $\tau_{yz}\,\mathrm{d}\,S$. Similarly for z axis, the components are $\tau_{zx}\,\mathrm{d}\,S$, $\tau_{zy}\,\mathrm{d}\,S$, $\tau_{zz}\,\mathrm{d}\,S$. If $\mathrm{d}\,S$ is unity, the components of surface force are $\tau_{\nu x}$, $\tau_{\nu y}$, $\tau_{\nu z}$. These are called the component of stress at a point P and the vector $\vec{T}^{(\nu)}$ of which these are the components is called the stress vector corresponding to the unit area P is in normal in the direction of $\vec{\nu}$. Therefore

$$\vec{T}^{(\nu)} = \tau_{\nu x}\hat{i} + \tau_{\nu y}\hat{j} + \tau_{\nu z}\hat{k}\,. \tag{1.8.1}$$

Taking $\vec{\nu}$ in the direction of x, y, z axes respectively, we have

$$\begin{aligned}
\vec{T}^{x} &= \tau_{xx}\hat{i} + \tau_{xy}\hat{j} + \tau_{xz}\hat{k} \\
\vec{T}^{y} &= \tau_{yx}\hat{i} + \tau_{yy}\hat{j} + \tau_{yz}\hat{k} \\
\vec{T}^{z} &= \tau_{zx}\hat{i} + \tau_{zy}\hat{j} + \tau_{zz}\hat{k}\,.
\end{aligned} \tag{1.8.2}$$

There are nine stress vector components on the right hand side of (1.8.2) which are the components of a second order Cartesian tensor known as the stress tensor.

The matrix representation of stress tensor is

$$\begin{bmatrix} \tau_{xx} & \tau_{xy} & \tau_{xz} \\ \tau_{yx} & \tau_{yy} & \tau_{yz} \\ \tau_{zx} & \tau_{zy} & \tau_{zz} \end{bmatrix}\,. \tag{1.8.3}$$

The components perpendicular to the coordinate planes i.e., τ_{xx}, τ_{yy}, τ_{zz} are called normal stress and those tangential to the plane i.e., non diagonal elements are called shear stress.

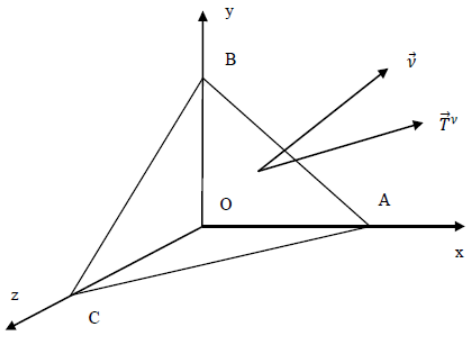
1.9 Cauchy's fundamental theorem for stress

The stress vector at a point on any arbitrary plane surface is a linear function of three stress vectors acting on any three mutually perpendicular planes through that point.

Proof: Let us consider a tetrahedron of the continuum with one corner at the point O and edges OA, OB, OC parallel to the coordinate axes and of infinitely small length and face ABC perpendicular to the direction $\vec{\nu}$ whose direction cosines are l, m, n and these direction are drawn outwards.

Let $\mathrm{d}\,S_1$, $\mathrm{d}\,S_2$, $\mathrm{d}\,S_3$ and $\mathrm{d}\,S$ be the areas of the faces OBC, OCA, OAB and ABC respectively. Therefore

$$\mathrm{d}\,S_1 = l\,\mathrm{d}\,S,\ \mathrm{d}\,S_2 = m\,\mathrm{d}\,S,\ \mathrm{d}\,S_3 = n\,\mathrm{d}\,S,\ . \qquad (1.9.1)$$



Let $\mathrm{d}\,V$ be the volume of the material in the tetrahedron. Also let f_x, f_y and f_z be the component of acceleration of the continuum in the direction of x, y and z axes respectively and ρX, ρY, ρZ are the components of body forces per unit volume of the material.

Hence the x component of equation of motion of the material in the tetrahedron is

$$\begin{aligned}
&\rho\,\mathrm{d}\,V f_x = \rho X\,\mathrm{d}\,V + \tau_{\nu x}\,\mathrm{d}\,S - \tau_{xx}\,\mathrm{d}\,S_1 - \tau_{yx}\,\mathrm{d}\,S_2 - \tau_{zx}\,\mathrm{d}\,S_3 \\
\Rightarrow\ &\rho\,\mathrm{d}\,V f_x = \rho X\,\mathrm{d}\,V + (\tau_{\nu x} - l\tau_{xx} - m\tau_{yx} - n\tau_{zx})\,\mathrm{d}\,S \\
\Rightarrow\ &\rho f_x \frac{\mathrm{d}\,V}{\mathrm{d}\,S} = \rho X \frac{\mathrm{d}\,V}{\mathrm{d}\,S} + (\tau_{\nu x} - l\tau_{xx} - m\tau_{yx} - n\tau_{zx})\,.
\end{aligned}$$

We now make the dimension of the tetrahedron tends to zero in such a manner that face ABC remains parallel to itself i.e., the direction of normal remains itself and in the limit face ABC tends to pass through O. Therefore $\tau_{\nu x}$, $\tau_{\nu y}$, $\tau_{\nu z}$ can be taken to be the values of the corresponding quantities at the point O. Since the linear dimension of the tetrahedron $\frac{\mathrm{d}\,V}{\mathrm{d}\,S} \to 0$, we have

$$\tau_{\nu x} = l\tau_{xx} + m\tau_{yx} + n\tau_{zx}. \qquad (1.9.2)$$

In a similar manner we can obtain

$$\begin{aligned}
\tau_{\nu y} &= l\tau_{xy} + m\tau_{yy} + n\tau_{zy}\,, \\
\tau_{\nu z} &= l\tau_{xz} + m\tau_{yz} + n\tau_{zz}\,. \qquad (1.9.3)
\end{aligned}$$

This is the relation between the component of the stress vector $\vec{T}^{\nu}$ with the components of stress tensor at any point. Thus we conclude that the stress at any point O is completely determined by the nine components of stress tensor at that point.

Example: *Let the components of the stress tensor at P be given in matrix form by*

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

in units of mega-pascals, where b is a constant. Determine b so that stress vector on some plane at the point will be zero. Also determine the direction cosines of the normal to the plane.

Solution: Let $l,\ m,\ n$ be the direction cosines of the normal to the plane. Then due to Cauchy's stress formula

$$\begin{bmatrix} \tau_{\nu x} \\ \tau_{\nu y} \\ \tau_{\nu z} \end{bmatrix} = \begin{bmatrix} \tau_{xx} & \tau_{yx} & \tau_{zx} \\ \tau_{xy} & \tau_{yy} & \tau_{zy} \\ \tau_{xz} & \tau_{yz} & \tau_{zz} \end{bmatrix} \begin{bmatrix} l \\ m \\ n \end{bmatrix}$$

By the given condition

$$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 2 \\ 1 & b & 1 \\ 2 & 1 & 0 \end{bmatrix} \begin{bmatrix} l \\ m \\ n \end{bmatrix}$$

For non trivial solution of the above system of linear equation, we have

$$\begin{vmatrix} 0 & 1 & 2 \\ 1 & b & 1 \\ 2 & 1 & 0 \end{vmatrix} = 0 \quad \Rightarrow b = 1\,.$$

For $b = 1$, the system of linear equations reduces to

$$\frac{l}{1} = \frac{m}{-2} = \frac{n}{1}\,.$$

Hence the direction cosines of the normal to the plane is given by $\vec{\nu} = (l,\ m,\ n) = \frac{1}{\sqrt{6}}(1, -2, 1)$.

Example: *The state of stress throughout a continuum is given with respect to cartesian axes $OX_1,\ OX_2,\ OX_3$ by*

$$(T_{ij}) = \begin{bmatrix} 3x_1x_2 & 5x_2^2 & 0 \\ 5x_2^2 & 0 & 2x_3 \\ 0 & 2x_3 & 0 \end{bmatrix}\,.$$

Determine the stress vector acting at a point $P(2, 1, \sqrt{3})$ on the plane tangent to the cylindrical surface $x_2^2 + x_3^2 = 4$ at the point $P(2, 1, \sqrt{3})$.

Solution: The cylindrical surface can be expressed as

$$\phi(x_1, x_2, x_3) \equiv x_2^2 + x_3^2 - 4 = 0\,.$$

Then

$$\nabla\phi(x_1, x_2, x_3) = (0, 2x_2, 2x_3) = (0, 2, 2\sqrt{3}) \quad \text{at } P(2, 1, \sqrt{3})\,.$$

The unit normal to the surface at $P(2, 1, \sqrt{3})$ is given by

$$\vec{\nu} = (l, m, n) = \frac{\nabla\phi}{|\nabla\phi|} = \frac{1}{2}(0, 1, \sqrt{3})\,.$$

At the point P, the state of stress throughout a continuum is given by

$$(T_{ij}) = \begin{bmatrix} 3x_1x_2 & 5x_2^2 & 0 \\ 5x_2^2 & 0 & 2x_3 \\ 0 & 2x_3 & 0 \end{bmatrix} = \begin{bmatrix} 6 & 5 & 0 \\ 5 & 0 & 2\sqrt{3} \\ 0 & 2\sqrt{3} & 0 \end{bmatrix} \quad \text{at } P(2, 1, \sqrt{3})\,.$$

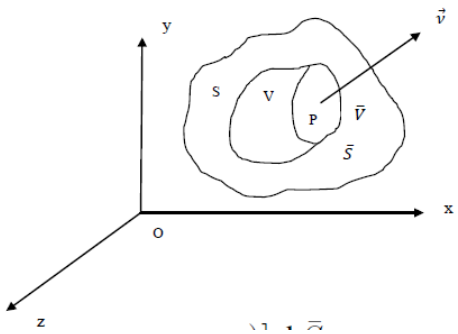
Due to Cauchy's stress formula, the stress vector at P on the plane perpendicular to $\vec{\nu} = (l, m, n) = \frac{1}{2}(0, 1, \sqrt{3})$ is given by

$$\begin{bmatrix} \tau_{\nu x} \\ \tau_{\nu y} \\ \tau_{\nu z} \end{bmatrix} = \begin{bmatrix} \tau_{xx} & \tau_{yx} & \tau_{zx} \\ \tau_{xy} & \tau_{yy} & \tau_{zy} \\ \tau_{xz} & \tau_{yz} & \tau_{zz} \end{bmatrix} \begin{bmatrix} l \\ m \\ n \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 6 & 5 & 0 \\ 5 & 0 & 2\sqrt{3} \\ 0 & 2\sqrt{3} & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ \sqrt{3} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 5 \\ 6 \\ 2\sqrt{3} \end{bmatrix} .$$

1.10 Equation of equilibrium of a continuum

Consider a continuous body every portion of which is contained in the volume V and bounded by the closed surface S is in equilibrium. Let $P(x, y, z)$ be any point in the volume V. We consider a small closed surface $\bar{S}$ enclosing the point P and lie entirely within V. Let $\bar{V}$ be the volume enclosed by the surface $\bar{S}$ and $\vec{\nu}$ is drawn outward and normal to the surface whose direction cosines are l, m, n.

Let X, Y, Z be the components of body force per unit mass. Hence $\rho X, \rho Y, \rho Z$ are the components of body force per unit volume where ρ denotes the density of the solid. For equilibrium of the matter within volume $\bar{V}$, the resultant of the body forces within $\bar{V}$ and surface forces along $\bar{S}$ must vanish together. We consider the components of body and surface forces along x direction and have



$$\iiint_{\bar{V}} \rho X \, \mathrm{d}\bar{V} + \iint_{\bar{S}} \tau_{\nu x} \, \mathrm{d}\bar{S} = 0 \, . \tag{1.10.1}$$

Now

$$\tau_{\nu x} = l\tau_{xx} + m\tau_{yx} + n\tau_{zx} \, .$$

Hence

$$\begin{aligned} \iint_{\bar{S}} \tau_{\nu x} \, \mathrm{d}\bar{S} &= \iint_{\bar{S}} (l\tau_{xx} + m\tau_{yx} + n\tau_{zx}) \, \mathrm{d}\bar{S} \, , \text{using divergence theorem, we have} \\ &= \iiint_{\bar{V}} \left[\frac{\partial}{\partial x}(\tau_{xx}) + \frac{\partial}{\partial y}(\tau_{yx}) + \frac{\partial}{\partial z}(\tau_{zx}) \right] \mathrm{d}\bar{V} \, . \end{aligned} \tag{1.10.2}$$

Using this in (1.10.1), we have

$$\iiint_{\bar{V}} \left[\rho X + \frac{\partial}{\partial x}(\tau_{xx}) + \frac{\partial}{\partial y}(\tau_{yx}) + \frac{\partial}{\partial z}(\tau_{zx}) \right] \mathrm{d}\bar{V} = 0 \, .$$

We consider the dimension of $\bar{S}$ tends to zero in such a manner that always encloses the point P and we have

$$\rho X + \frac{\partial}{\partial x}(\tau_{xx}) + \frac{\partial}{\partial y}(\tau_{yx}) + \frac{\partial}{\partial z}(\tau_{zx}) = 0 \, . \tag{1.10.3}$$

In a similar manner, considering the components of forces in y and z directions respectively we can obtain

$$\rho Y + \frac{\partial}{\partial x}(\tau_{xy}) + \frac{\partial}{\partial y}(\tau_{yy}) + \frac{\partial}{\partial z}(\tau_{zy}) = 0 \, , \tag{1.10.4}$$

$$\rho Z + \frac{\partial}{\partial x}(\tau_{xz}) + \frac{\partial}{\partial y}(\tau_{yz}) + \frac{\partial}{\partial z}(\tau_{zz}) = 0\,. \tag{1.10.5}$$

The above three equations are called *equation of equilibrium of a continuum*.

If the material within the volume $\bar{V}$ enclosed by the surface $\bar{S}$ be in equilibrium, then the moments of the body and surface forces about x, y and z axes vanishes separately. We consider the moment about x axis and have

$$\iint_{\bar{S}} (y\tau_{\nu z} - z\tau_{\nu y})\,\mathrm{d}\,\bar{S} + \iiint_{\bar{V}} (y\rho Z - z\rho Y)\,\mathrm{d}\,\bar{V} = 0\,. \tag{1.10.6}$$

Now

$$\begin{aligned}
&\iint_{\bar{S}} (y\tau_{\nu z} - z\tau_{\nu y})\,\mathrm{d}\,\bar{S} = \iint_{\bar{S}} [y(l\tau_{xz} + m\tau_{yz} + n\tau_{zz}) - z(l\tau_{xy} + m\tau_{yy} + n\tau_{zy})]\,\mathrm{d}\,\bar{S} \\
&= \iiint_{\bar{V}} \left[\frac{\partial}{\partial x}(y\tau_{xz} - z\tau_{xy}) + \frac{\partial}{\partial y}(y\tau_{yz} - z\tau_{yy}) + \frac{\partial}{\partial z}(y\tau_{zz} - z\tau_{zy})\right]\mathrm{d}\,\bar{V} \\
&= \iiint_{\bar{V}} \left[y\left(\frac{\partial}{\partial x}\tau_{xz} + \frac{\partial}{\partial y}\tau_{yz} + \frac{\partial}{\partial z}\tau_{zz}\right) - z\left(\frac{\partial}{\partial x}\tau_{xy} + \frac{\partial}{\partial y}\tau_{yy} + \frac{\partial}{\partial z}\tau_{zy}\right) + (\tau_{yz} - \tau_{zy})\right]\mathrm{d}\,\bar{V} \\
&= \iiint_{\bar{V}} [y\{-\rho Z - z(-\rho Y)\} + (\tau_{yz} - \tau_{zy})]\,\mathrm{d}\,\bar{V}\,, \quad \text{using (1.10.4), (1.10.5)} \\
&= \iiint_{\bar{V}} [\rho(zY - yZ) + (\tau_{yz} - \tau_{zy})]\,\mathrm{d}\,\bar{V}\,.
\end{aligned} \tag{1.10.7}$$

Substituting this result in (1.10.6), we obtain

$$\begin{aligned}
&\iiint_{\bar{V}} [\rho(zY - yZ) + (\tau_{yz} - \tau_{zy})]\,\mathrm{d}\,\bar{V} + \iiint_{\bar{V}} (y\rho Z - z\rho Y)\,\mathrm{d}\,\bar{V} = 0 \\
\Rightarrow &\iiint_{\bar{V}} (\tau_{yz} - \tau_{zy})\,\mathrm{d}\,\bar{V} = 0\,.
\end{aligned} \tag{1.10.8}$$

We consider the dimension of $\bar{V}$ tends to zero in such a manner so that it always represents the point P. Hence we have

$$\tau_{yz} = \tau_{zy}\,.$$

In a similar manner considering the moments of the forces about y and z axes respectively, we obtain

$$\tau_{zx} = \tau xz\,, \quad \tau_{xy} = \tau_{yx}\,.$$

Therefore, the matrix representation of the stress tensor $\begin{bmatrix} \tau_{xx} & \tau_{yx} & \tau_{zx} \\ \tau_{xy} & \tau_{yy} & \tau_{zy} \\ \tau_{xz} & \tau_{yz} & \tau_{zz} \end{bmatrix}$ is symmetric.

Unit 2

Course Structure

- Transformation rule of stress components
- Principal stress and principal axes of stresses
- Stress invariants, Stress quadric of Cauchy

2.1 Rule of transformation of stress components

Let Ox, Oy, Oz be a set of rectangular axes and Ox', Oy', Oz' are another set of rectangular axes through O such that the direction cosines of these axes are (l_1, m_1, n_1), (l_2, m_2, n_2) and $l_3, m_3, n_3)$ respectively with respect to Ox, Oy, Oz.

Let P be any point whose coordinate referred to Ox, Oy, Oz as axes are (x, y, z) and if (x', y', z') be the coordinates of the same point referred to Ox', Oy', Oz' as axes, then the scheme of transformation from one set of coordinates to another is given by

	x	y	z
x'	l_1	m_1	n_1
y'	l_2	m_2	n_2
z'	l_3	m_3	n_3

We know that if we draw a unit area at O and draw the normal $\vec{\nu}$ to the surface, then the force exerted by the material on the side of the surface towards which normal $\vec{\nu}$ is drawn to the material on the other side of the surface across the unit area has its components $\tau_{\nu x}, \tau_{\nu y}, \tau_{\nu z}$, where

$$\begin{aligned} \tau_{\nu x} &= l\tau_{xx} + m\tau_{yx} + n\tau_{zx}\,, \\ \tau_{\nu y} &= l\tau_{xy} + m\tau_{yy} + n\tau_{zy}\,, \\ \tau_{\nu z} &= l\tau_{xz} + m\tau_{yz} + n\tau_{zz}\,, \end{aligned} \tag{2.1.1}$$

where (l, m, n) are the direction cosines of the normal $\vec{\nu}$. If we choose $\vec{\nu}$ in the direction of Ox' axis then

$$\begin{aligned}\tau_{x'x} &= l_1\tau_{xx} + m_1\tau_{yx} + n_1\tau_{zx}\\ \tau_{x'y} &= l_1\tau_{xy} + m_1\tau_{yy} + n_1\tau_{zy}\\ \tau_{x'z} &= l_1\tau_{xz} + m_1\tau_{yz} + n_1\tau_{zz}\,.\end{aligned} \tag{2.1.2}$$

In matrix form it can be expressed as

$$\begin{bmatrix}\tau_{x'x}\\ \tau_{x'y}\\ \tau_{x'x}\end{bmatrix} = \begin{bmatrix}\tau_{xx} & \tau_{yx} & \tau_{zx}\\ \tau_{xy} & \tau_{yy} & \tau_{zy}\\ \tau_{xz} & \tau_{yz} & \tau_{zz}\end{bmatrix}\begin{bmatrix}l_1\\ m_1\\ n_1\end{bmatrix}\,. \tag{2.1.3}$$

If we choose $\vec{\nu}$ in the direction of Oy' axis then we have

$$\begin{bmatrix}\tau_{y'x}\\ \tau_{y'y}\\ \tau_{y'x}\end{bmatrix} = \begin{bmatrix}\tau_{xx} & \tau_{yx} & \tau_{zx}\\ \tau_{xy} & \tau_{yy} & \tau_{zy}\\ \tau_{xz} & \tau_{yz} & \tau_{zz}\end{bmatrix}\begin{bmatrix}l_2\\ m_2\\ n_2\end{bmatrix}\,. \tag{2.1.4}$$

If we choose $\vec{\nu}$ in the direction of Oz' axis then we have

$$\begin{bmatrix}\tau_{z'x}\\ \tau_{z'y}\\ \tau_{z'x}\end{bmatrix} = \begin{bmatrix}\tau_{xx} & \tau_{yx} & \tau_{zx}\\ \tau_{xy} & \tau_{yy} & \tau_{zy}\\ \tau_{xz} & \tau_{yz} & \tau_{zz}\end{bmatrix}\begin{bmatrix}l_3\\ m_3\\ n_3\end{bmatrix}\,. \tag{2.1.5}$$

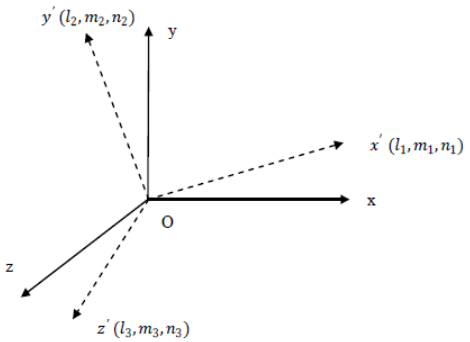
Combining (2.1.3), (2.1.4) and (2.1.5), we have

$$\begin{bmatrix}\tau_{x'x} & \tau_{y'x} & \tau_{z'x}\\ \tau_{x'y} & \tau_{y'y} & \tau_{z'y}\\ \tau_{x'z} & \tau_{y'z} & \tau_{z'z}\end{bmatrix} = \begin{bmatrix}\tau_{xx} & \tau_{yx} & \tau_{zx}\\ \tau_{xy} & \tau_{yy} & \tau_{zy}\\ \tau_{xz} & \tau_{yz} & \tau_{zz}\end{bmatrix}\begin{bmatrix}l_1 & l_2 & l_3\\ m_1 & m_2 & m_3\\ n_1 & n_2 & n_3\end{bmatrix}\,. \tag{2.1.6}$$

Here $\tau_{x'x}, \tau_{x'y}, \tau_{x'z}$ are the components of force per unit area which the material on the positive side of x' axis exerts on the material on the negative side across the plane $x' =$ constant.

Now if $\tau_{x'x'}$ be the component of the force per unit area in the direction of x' axis which the material on the positive side of x' axis exerts on the material on the negative side across the plane $x' =$ constant. Then

$$\begin{aligned}\tau_{x'x'} &= l_1\tau_{xx'} + m_1\tau_{yx'} + n_1\tau_{zx'}\\ \tau_{x'y'} &= l_1\tau_{xy'} + m_1\tau_{yy'} + n_1\tau_{zy'}\\ \tau_{x'z'} &= l_1\tau_{xz'} + m_1\tau_{yz'} + n_1\tau_{zz'}\,.\end{aligned} \tag{2.1.7}$$



It can be expressed in the matrix form

$$\begin{bmatrix}\tau_{x'x'}\\ \tau_{x'y'}\\ \tau_{x'z'}\end{bmatrix} = \begin{bmatrix}l_1 & m_1 & n_1\end{bmatrix}\begin{bmatrix}\tau_{xx'} & \tau_{xy'} & \tau_{xz'}\\ \tau_{yx'} & \tau_{yy'} & \tau_{yz'}\\ \tau_{zx'} & \tau_{zy'} & \tau_{zz'}\end{bmatrix}\,. \tag{2.1.8}$$

In a similar manner we obtain

$$\begin{bmatrix} \tau_{y'x'} \\ \tau_{y'y'} \\ \tau_{y'z'} \end{bmatrix} = \begin{bmatrix} l_2 & m_2 & n_2 \end{bmatrix} \begin{bmatrix} \tau_{xx'} & \tau_{xy'} & \tau_{xz'} \\ \tau_{yx'} & \tau_{yy'} & \tau_{yz'} \\ \tau_{zx'} & \tau_{zy'} & \tau_{zz'} \end{bmatrix} . \tag{2.1.9}$$

$$\begin{bmatrix} \tau_{z'x'} \\ \tau_{z'y'} \\ \tau_{z'z'} \end{bmatrix} = \begin{bmatrix} l_3 & m_3 & n_3 \end{bmatrix} \begin{bmatrix} \tau_{xx'} & \tau_{xy'} & \tau_{xz'} \\ \tau_{yx'} & \tau_{yy'} & \tau_{yz'} \\ \tau_{zx'} & \tau_{zy'} & \tau_{zz'} \end{bmatrix} . \tag{2.1.10}$$

Combining (2.1.8), (2.1.9) and (2.1.10), we get

$$\begin{bmatrix} \tau_{x'x'} & \tau_{y'x'} & \tau_{z'x'} \\ \tau_{x'y'} & \tau_{y'y'} & \tau_{z'y'} \\ \tau_{x'z'} & \tau_{y'z'} & \tau_{z'z'} \end{bmatrix} = \begin{bmatrix} l_1 & m_1 & n_1 \\ l_2 & m_2 & n_2 \\ l_3 & m_3 & n_3 \end{bmatrix} \begin{bmatrix} \tau_{xx'} & \tau_{xy'} & \tau_{xz'} \\ \tau_{yx'} & \tau_{yy'} & \tau_{yz'} \\ \tau_{zx'} & \tau_{zy'} & \tau_{zz'} \end{bmatrix} . \tag{2.1.11}$$

Using (2.1.6) we have

$$\begin{bmatrix} \tau_{x'x'} & \tau_{y'x'} & \tau_{z'x'} \\ \tau_{x'y'} & \tau_{y'y'} & \tau_{z'y'} \\ \tau_{x'z'} & \tau_{y'z'} & \tau_{z'z'} \end{bmatrix} = \begin{bmatrix} l_1 & m_1 & n_1 \\ l_2 & m_2 & n_2 \\ l_3 & m_3 & n_3 \end{bmatrix} \begin{bmatrix} \tau_{xx} & \tau_{yx} & \tau_{zx} \\ \tau_{xy} & \tau_{yy} & \tau_{zy} \\ \tau_{xz} & \tau_{yz} & \tau_{zz} \end{bmatrix} \begin{bmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \\ n_1 & n_2 & n_3 \end{bmatrix} . \tag{2.1.12}$$

These are the laws of transformation of stress tensor.

Example: *The state of stress at a point with respect to Cartesian axes OX_1, OX_2, OX_3 is given by*

$$(T_{ij}) = \begin{bmatrix} 15 & -10 & 0 \\ -10 & 5 & 0 \\ 0 & 0 & 20 \end{bmatrix} .$$

Determine the stress tensor T'_{ij} for related axes OX'_1, OX'_2, OX'_3 for which transformation matrix is

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

Solution: By using the stress transformation laws we get

$$\begin{aligned} \begin{bmatrix} T'_{11} & T'_{21} & T'_{31} \\ T'_{12} & T'_{22} & T'_{32} \\ T'_{13} & T'_{23} & T'_{33} \end{bmatrix} &= \begin{bmatrix} l_1 & m_1 & n_1 \\ l_2 & m_2 & n_2 \\ l_3 & m_3 & n_3 \end{bmatrix} \begin{bmatrix} T_{11} & T_{21} & T_{31} \\ T_{12} & T_{22} & T_{32} \\ T_{13} & T_{23} & T_{33} \end{bmatrix} \begin{bmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \\ n_1 & n_2 & n_3 \end{bmatrix} \\ &= \begin{bmatrix} 3/5 & 0 & -4/5 \\ 0 & 1 & 0 \\ 4/5 & 0 & 3/5 \end{bmatrix} \begin{bmatrix} 15 & -10 & 0 \\ -10 & 5 & 0 \\ 0 & 0 & 20 \end{bmatrix} \begin{bmatrix} 3/5 & 0 & 4/5 \\ 0 & 1 & 0 \\ -4/5 & 0 & 3/5 \end{bmatrix} \\ &= \begin{bmatrix} 91/5 & -6 & -1/5 \\ -6 & 5 & -8 \\ -12/5 & -8 & 84/5 \end{bmatrix} . \end{aligned}$$

2.2 Principal stresses and principal axes of stresses

Generally, stress vector do not act in a direction perpendicular to the plane element on which it is acting. Particularly when stress vector acts entirely in a direction perpendicular to the element of plane on which it acts is called *principal stress*. The element of plane area on which principal stress is acting is called principal plane and the direction of principal stress is called *principal direction of stress* or *principal axis of stress*.

2.2.1 Determination of principal stress and principal direction of stress

Consider a unit area in a continuum and let $\vec{\nu}$ be the unit vector in the direction of the normal to it in a specified sense. Let the components of the stress vector $\vec{T}^{\nu}$ at a point are $\tau_{\nu x}, \tau_{\nu y}, \tau_{\nu z}$. In general the direction of $\vec{T}^{\nu}$ is different from the direction of $\vec{\nu}$, but if the orientation of the unit area be such that $\vec{T}^{\nu}$ is in the direction of $\vec{\nu}$ then these directions are called principal direction of stress.

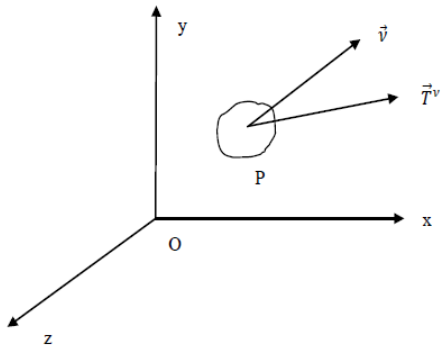
Thus for a principal stress direction,

$$\vec{T}^{\nu} = \sigma\vec{\nu}\,,$$

where σ is the magnitude of the stress vector and is called *principal stress value*. Explicitly

$$\begin{aligned} &\vec{T}^{\nu} = \sigma\vec{\nu} \\ \Rightarrow\ &\tau_{\nu x} = \sigma l, \tau_{\nu y} = \sigma m, \tau_{\nu z} = \sigma n \end{aligned}$$

where l, m, n are the direction cosines of $\vec{\nu}$. Therefore, we have



$$\begin{aligned} l\tau_{xx} + m\tau_{yx} + n\tau_{zx} &= \sigma l \\ l\tau_{xy} + m\tau_{yy} + n\tau_{zy} &= \sigma m \\ l\tau_{xz} + m\tau_{yz} + n\tau_{zz} &= \sigma n\,. \end{aligned} \tag{2.2.1}$$

It can be written as

$$\begin{aligned} l(\tau_{xx} - \sigma) + m\tau_{yx} + n\tau_{zx} &= 0 \\ l\tau_{xy} + m(\tau_{yy} - \sigma) + n\tau_{zy} &= 0 \\ l\tau_{xz} + m\tau_{yz} + n(\tau_{zz} - \sigma) &= 0\,. \end{aligned} \tag{2.2.2}$$

The system of equation (2.2.2) has a non-vanishing solution l, m, n if and only if the determinant of the coefficient vanishes i.e.,

$$\begin{vmatrix} \tau_{xx} - \sigma & \tau_{yx} & \tau_{zx} \\ \tau_{xy} & \tau_{yy} - \sigma & \tau_{zy} \\ \tau_{xz} & \tau_{yz} & \tau_{zz} - \sigma \end{vmatrix} = 0\,. \tag{2.2.3}$$

This is a cubic equation in σ and is called the characteristic equation. It has three roots $\sigma_1, \sigma_2, \sigma_3$ which are called *principal stress values*.

Corresponding to each of these principal stresses $\sigma_1, \sigma_2, \sigma_3$ there are three principal stress directions $l_1, m_1, n_1; l_2, m_2, n_2; l_3, m_3, n_3$ respectively such that

$$\begin{aligned} l_1\tau_{xx} + m_1\tau_{yx} + n_1\tau_{zx} &= \sigma_1 l_1 \\ l_1\tau_{xy} + m_1\tau_{yy} + n_1\tau_{zy} &= \sigma_1 m_1 \\ l_1\tau_{xz} + m_1\tau_{yz} + n_1\tau_{zz} &= \sigma_1 n_1 \end{aligned} \tag{2.2.4}$$

$$\begin{aligned} l_2\tau_{xx} + m_2\tau_{yx} + n_2\tau_{zx} &= \sigma_2 l_2 \\ l_2\tau_{xy} + m_2\tau_{yy} + n_2\tau_{zy} &= \sigma_2 m_2 \\ l_2\tau_{xz} + m_2\tau_{yz} + n_2\tau_{zz} &= \sigma_2 n_2 \end{aligned} \tag{2.2.5}$$

and another set of equations can be obtained by replacing l_1, m_1, n_1 by l_3, m_3, n_3 and σ_1 by σ_3 in (2.2.4).

We now prove that the three principal stress values $\sigma_1, \sigma_2, \sigma_3$ are real and that the corresponding principal stress directions are mutually orthogonal.

Multiplying three equations of (2.2.4) by l_2, m_2, n_2 respectively and adding we get

$$\begin{aligned} l_1l_2\tau_{xx} + m_1m_2\tau_{yy} + n_1n_2\tau_{zz} + (m_1n_2 + m_2n_1)\tau_{yz} + (n_1l_2 + n_2l_1)\tau_{zx} + (l_1m_2 + l_2m_1)\tau_{xy} \\ = \sigma_1(l_1l_2 + m_1m_2 + n_1n_2)\,. \end{aligned} \tag{2.2.6}$$

Similarly, multiplying three equations of (2.2.5) by l_1, m_1, n_1 respectively and adding we get

$$\begin{aligned} l_1l_2\tau_{xx} + m_1m_2\tau_{yy} + n_1n_2\tau_{zz} + (m_1n_2 + m_2n_1)\tau_{yz} + (n_1l_2 + n_2l_1)\tau_{zx} + (l_1m_2 + l_2m_1)\tau_{xy} \\ = \sigma_2(l_1l_2 + m_1m_2 + n_1n_2)\,. \end{aligned} \tag{2.2.7}$$

Subtracting these two equations, we obtain

$$(\sigma_1 - \sigma_2)(l_1l_2 + m_1m_2 + n_1n_2) = 0\,. \tag{2.2.8}$$

In a similar manner we obtain other two equations

$$(\sigma_2 - \sigma_3)(l_1l_2 + m_1m_2 + n_1n_2) = 0\,. \tag{2.2.9}$$

$$(\sigma_3 - \sigma_1)(l_1l_2 + m_1m_2 + n_1n_2) = 0\,. \tag{2.2.10}$$

Let the roots of (2.2.3) are $\sigma_1, \sigma_2, \sigma_3$ and assume that the equation has complex root. Since it is a cubic equation with real coefficient, then another root must also be a complex which is the complex conjugate of the former. Hence the set of roots can be written as

$$\sigma_1 = \alpha + \mathrm{i}\,\beta\,, \quad \sigma_2 = \alpha - \mathrm{i}\,\beta\,, \quad \text{and} \;\; \sigma_3\,. \tag{2.2.11}$$

We consider the complex conjugate of (2.2.5) and obtain

$$\begin{aligned} l_2\bar{\tau}_{xx} + m_2\bar{\tau}_{yx} + n_2\bar{\tau}_{zx} &= \bar{\sigma}_2 l_2 \\ l_2\bar{\tau}_{xy} + m_2\bar{\tau}_{yy} + n_2\bar{\tau}_{zy} &= \bar{\sigma}_2 m_2 \\ l_2\bar{\tau}_{xz} + m_2\bar{\tau}_{yz} + n_2\bar{\tau}_{zz} &= \bar{\sigma}_2 n_2 \end{aligned} \tag{2.2.12}$$

Since $\bar{\sigma}_2 = \sigma_1$ and τ_{xx}, τ_{xy} etc. are real, then the above equation becomes

$$\begin{aligned} l_2\tau_{xx} + m_2\tau_{yx} + n_2\tau_{zx} &= \sigma_1 l_2 \\ l_2\tau_{xy} + m_2\tau_{yy} + n_2\tau_{zy} &= \sigma_1 m_2 \\ l_2\tau_{xz} + m_2\tau_{yz} + n_2\tau_{zz} &= \sigma_1 n_2 \end{aligned} \tag{2.2.13}$$

The coefficient of l_2, m_2, n_2 in (2.2.13) are the complex conjugate of the coefficient of l_1, m_1, n_1 in (2.2.4). Therefore, the values of l_2, m_2, n_2 determined from (2.2.13) are the complex conjugate of the values of l_1, m_1, n_1 determined from (2.2.4). Hence if

$$l_1 = a_1 + \mathrm{i}\, b_1\,, m_1 = a_2 + \mathrm{i}\, b_2\,, n_1 = a_3 + \mathrm{i}\, b_3$$

then

$$l_2 = a_1 - \mathrm{i}\, b_1\,, m_2 = a_2 - \mathrm{i}\, b_2\,, n_2 = a_3 - \mathrm{i}\, b_3\,.$$

Therefore

$$l_1 l_2 + m_1 m_2 + n_1 n_2 = a_1^2 + a_2^2 + a_3^2 + b_1^2 + b_2^2 + b_3^2 \neq 0\,.$$

Then it follows from (2.2.8) that

$$\begin{aligned} & \sigma_1 = \sigma_2 \\ \Rightarrow\ & \alpha + \mathrm{i}\,\beta = \alpha - \mathrm{i}\,\beta \\ \Rightarrow\ & \beta = 0\,, \end{aligned} \tag{2.2.14}$$

which contradicts the assumption that the roots are complex. Hence the roots of (2.2.3) are not complex and $\sigma_1, \sigma_2, \sigma_3$ are real. Now, if $\sigma_1 \neq \sigma_2 \neq \sigma_3$ then from (2.2.8),(2.2.9),(2.2.10) we find that the principal stress directions are mutually orthogonal. Again if $\sigma_1 = \sigma_2 \neq \sigma_3$ then l_3, m_3, n_3 are fixed and we can determine an infinite number of values of all direction cosines l_1, m_1, n_1 and l_2, m_2, n_2 orthogonal to l_3, m_3, n_3. If $\sigma_1 = \sigma_2 = \sigma_3$ then any set of orthogonal system may be taken as the principal stress direction.

Example: *For the state of stress*

$$(T_{ij}) = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}\,,$$

determine the principal stress and their directions.

Solution: Let σ be the principal stress. Then it satisfy the characteristic equation

$$\begin{vmatrix} 0-\sigma & 1 & 1 \\ 1 & 0-\sigma & 1 \\ 1 & 1 & 0-\sigma \end{vmatrix} = 0\,.$$

$$\Rightarrow \sigma = -1, -1, 2\,.$$

Since two principal stress values are equal i.e., $\sigma_1 = \sigma_2 = -1$, then the principal stress directions corresponding to $\sigma_3 = 2$ is unique and any two directions perpendicular to this direction are principal stress directions associated with σ_1 and σ_2.

If l, m, n be the direction of principal stress corresponding to $\sigma_3 = 2$, then

$$\begin{aligned} l(0-\sigma_3) + m\cdot 1 + n\cdot 1 &= 0 \\ l\cdot 1 + m(0-\sigma_3) + n\cdot 1 &= 0 \\ l\cdot 1 + m\cdot 1 + n(0-\sigma_3) &= 0 \end{aligned}$$

$$\begin{aligned} \Rightarrow -2l + m + n &= 0 \\ l - 2m + n &= 0 \\ l + m - 2n &= 0 \end{aligned}$$

$$\Rightarrow l = m = n = \frac{1}{\sqrt{3}}\,.$$

Hence one principal direction is $\left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right)$. Any pair of axes perpendicular to each other and perpendicular to $\left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right)$ may serve as principal axes.

2.3 Stress invariants

Let Ox, Oy, Oz be a system of orthogonal axes with respect to which stress tensor is

$$\begin{bmatrix} \tau_{xx} & \tau_{yx} & \tau_{zx} \\ \tau_{xy} & \tau_{yy} & \tau_{zy} \\ \tau_{xz} & \tau_{yz} & \tau_{zz} \end{bmatrix}\,.$$

Let $\vec{\nu}$ be the direction of principal stress at O, then $\vec{T}^{\nu}$ is in the direction of $\vec{\nu}$. If l, m, n be the direction cosines of this line and σ be the magnitude of principal stress then

$$\tau_{\nu x} = \sigma l, \tau_{\nu y} = \sigma m, \tau_{\nu z} = \sigma n\,,$$

where $\tau_{\nu x}, \tau_{\nu y}, \tau_{\nu z}$ are the components of $\vec{T}^{\nu}$ which are given by

$$\begin{aligned} \tau_{\nu x} &= l\tau_{xx} + m\tau_{yx} + n\tau_{zx} \\ \tau_{\nu y} &= l\tau_{xy} + m\tau_{yy} + n\tau_{zy} \\ \tau_{\nu z} &= l\tau_{xz} + m\tau_{yz} + n\tau_{zz}\,. \end{aligned}$$

Hence we have following three linear equations in l, m, n

$$\begin{aligned} l(\tau_{xx}-\sigma) + m\tau_{yx} + n\tau_{zx} &= 0 \\ l\tau_{xy} + m(\tau_{yy}-\sigma) + n\tau_{zy} &= 0 \\ l\tau_{xz} + m\tau_{yz} + n(\tau_{zz}-\sigma) &= 0\,. \end{aligned}$$

Eliminating l, m, n we have

$$\begin{vmatrix} \tau_{xx}-\sigma & \tau_{yx} & \tau_{zx} \\ \tau_{xy} & \tau_{yy}-\sigma & \tau_{zy} \\ \tau_{xz} & \tau_{yz} & \tau_{zz}-\sigma \end{vmatrix} = 0\,.$$

This is a cubic equation in σ and can be expressed as

$$\sigma^3 - J_1\sigma^2 + J_2\sigma - J_3 = 0\,, \tag{2.3.1}$$

in which the coefficients have the following values

$$J_1 = \tau_{xx} + \tau_{yy} + \tau_{zz}\,, \tag{2.3.2}$$

$$J_2 = \begin{vmatrix} \tau_{xx} & \tau_{xy} \\ \tau_{xy} & \tau_{yy} \end{vmatrix} + \begin{vmatrix} \tau_{yy} & \tau_{yz} \\ \tau_{yz} & \tau_{zz} \end{vmatrix} + \begin{vmatrix} \tau_{zz} & \tau_{zx} \\ \tau_{zx} & \tau_{xx} \end{vmatrix}$$
$$= \tau_{xx}\tau_{yy} + \tau_{yy}\tau_{zz} + \tau_{zz}\tau_{xx} - \tau_{xy}^2 - \tau_{yz}^2 - \tau_{zx}^2\,, \tag{2.3.3}$$

$$J_3 = \begin{vmatrix} \tau_{xx} & \tau_{yx} & \tau_{zx} \\ \tau_{xy} & \tau_{yy} & \tau_{zy} \\ \tau_{xz} & \tau_{yz} & \tau_{zz} \end{vmatrix}\,. \tag{2.3.4}$$

All the three roots of (2.3.1) are real and they have the values of principal stresses $\sigma_1, \sigma_2, \sigma_3$.

Since the principal stresses characterize the physical state of stress of a point, they are independent of any coordinate of reference.

Therefore

$$J_1 = \tau_{xx} + \tau_{yy} + \tau_{zz} = \sigma_1 + \sigma_2 + \sigma_3$$

is invariant under any coordinate transformation and it is called *first invariant of stress*.

Again

$$J_2 = \tau_{xx}\tau_{yy} + \tau_{yy}\tau_{zz} + \tau_{zz}\tau_{xx} - \tau_{xy}^2 - \tau_{yz}^2 - \tau_{zx}^2$$
$$= \sigma_1\sigma_2 + \sigma_2\sigma_3 + \sigma_3\sigma_1$$

is also invariant under any coordinate transformation and it is called *second invariant of stress*.

Finally

$$J_3 = \begin{vmatrix} \tau_{xx} & \tau_{yx} & \tau_{zx} \\ \tau_{xy} & \tau_{yy} & \tau_{zy} \\ \tau_{xz} & \tau_{yz} & \tau_{zz} \end{vmatrix} = \sigma_1\sigma_2\sigma_3$$

is also invariant under any coordinate transformation and it is called *third invariant of stress*.

Example: *Evaluate directly stress invariants from stress tensor*

$$(T_{ij}) = \begin{bmatrix} 6 & -3 & 0 \\ -3 & 6 & 0 \\ 0 & 0 & 8 \end{bmatrix}\,.$$

Also determine the principal stresses for this state of stress. Verify that the stress invariants calculated from the principal stresses are same.

Solution: The stress invariants are given as follows.

The first stress invariant is

$$J_1 = \tau_{xx} + \tau_{yy} + \tau_{zz} = 6 + 6 + 8 = 20\,.$$

The second stress invariant is

$$\begin{aligned} J_2 &= \begin{vmatrix} \tau_{xx} & \tau_{xy} \\ \tau_{xy} & \tau_{yy} \end{vmatrix} + \begin{vmatrix} \tau_{yy} & \tau_{yz} \\ \tau_{yz} & \tau_{zz} \end{vmatrix} + \begin{vmatrix} \tau_{zz} & \tau_{zx} \\ \tau_{zx} & \tau_{xx} \end{vmatrix} \\ &= \begin{vmatrix} 6 & -3 \\ -3 & 6 \end{vmatrix} + \begin{vmatrix} 6 & 0 \\ 0 & 8 \end{vmatrix} + \begin{vmatrix} 8 & 0 \\ 0 & 6 \end{vmatrix} \\ &= 123\,. \end{aligned}$$

The third stress invariant is

$$J_3 = \begin{vmatrix} \tau_{xx} & \tau_{yx} & \tau_{zx} \\ \tau_{xy} & \tau_{yy} & \tau_{zy} \\ \tau_{xz} & \tau_{yz} & \tau_{zz} \end{vmatrix} = \begin{vmatrix} 6 & -3 & 0 \\ -3 & 6 & 0 \\ 0 & 0 & 8 \end{vmatrix} = 216\,.$$

The characteristic equation is given by

$$\begin{aligned} &\begin{vmatrix} 6-\sigma & -3 & 0 \\ -3 & 6-\sigma & 0 \\ 0 & 0 & 8-\sigma \end{vmatrix} = 0 \\ &\Rightarrow \sigma^3 - 20\sigma^2 + 123\sigma - 216 = 0 \\ &\Rightarrow \sigma = 3, 8, 9\,. \end{aligned}$$

Therefore the values of principal stresses are $\sigma_1 = 3, \sigma_2 = 8, \sigma_3 = 9$. Then the stress invariants calculated from principal stresses are given by

$$\begin{aligned} J_1 &= \sigma_1 + \sigma_2 + \sigma_3 = 3 + 8 + 9 = 20\,. \\ J_2 &= \sigma_1\sigma_2 + \sigma_2\sigma_3 + \sigma_3\sigma_1 = 3 \cdot 8 + 8 \cdot 9 + 3 \cdot 9 = 123 \\ J_3 &= \sigma_1\sigma_2\sigma_3 = 3 \cdot 8 \cdot 9 = 216 \end{aligned}$$

Hence the results are verified.

2.4 Few Probable Questions

1. Stress tensor at P are given in appropriate units.

$$(T_{ij}) = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 0 \\ 2 & 0 & -2 \end{bmatrix},$$

   Find the principal stresses and show that the principal directions which correspond to largest and smallest principal stresses are both perpendicular to y-axis. [Ans: 2, 1, -3; 2, 0,1; 0,1,0; 1,0,-2]

2. Stress tensor at a point is given by

$$(T_{ij}) = \begin{bmatrix} 5 & 0 & 0 \\ 0 & -6 & -12 \\ 0 & -12 & 1 \end{bmatrix},$$

Determine the maximum shear stress. [Ans: 12.5]

3. Stress tensor at a point P is given by

$$(T_{ij}) = \begin{bmatrix} 7 & 0 & -2 \\ 0 & 5 & 0 \\ -2 & 0 & 4 \end{bmatrix},$$

Determine stress vector on the plane at P whose unit normal is $\frac{2}{3}$, $-\frac{2}{3}$, $\frac{1}{3}$. [Ans: 4, $-\frac{10}{3}$, 0]

4. Stress tensor at a point P is given by

$$(T_{ij}) = \begin{bmatrix} -10 & 9 & 5 \\ 9 & 0 & 0 \\ 5 & 0 & 8 \end{bmatrix},$$

Find principal stresses and their directions.

$$\left[\text{Ans:} 4, 10.08, -16.08;\ n_i^{(1)} = \left(\tfrac{4}{\sqrt{122}}, \tfrac{9}{\sqrt{122}}, -\tfrac{5}{\sqrt{122}}\right);\ n_i^{(2)} = \tfrac{9}{4} n_i^{(1)};\ n_i^{(3)} = \tfrac{-5}{4} n_i^{(1)}\right]$$

5. Determine the principal stresses for

$$a)\quad (T_{ij}) = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \qquad b)\quad (T_{ij}) = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$

and show that both have the same principal directions but do not have same principal stresses. [Ans: (a) 2,-1,-1; (b) 4,1,1]

Unit 3

Course Structure

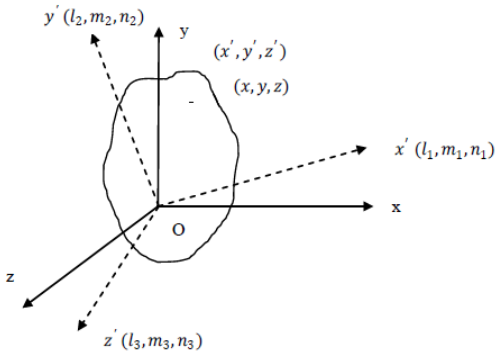
- Stress quadric of Cauchy
- Normal and shearing stress
- Mohr's circle for stress

3.1 Stress quadric of Cauchy

Let Ox, Oy, Oz be a set of rectangular axes through O. We suppose a quadric surface

$$\tau_{xx}x^2 + \tau_{yy}y^2 + \tau_{zz}z^2 + 2\tau_{yz}yz + 2\tau_{zx}zx + 2\tau_{xy}xy = \pm k^2\,, \tag{3.1.1}$$

where k is a constant and τ_{xx}, τ_{xy} etc. are the components of stress tensor at O referred to Ox, Oy and Oz axes. The quadric surface is called *stress quadric of Cauchy*.



Let us make a transformation of axis according to the following scheme.

	x	y	z
x'	l_1	m_1	n_1
y'	l_2	m_2	n_2
z'	l_3	m_3	n_3

Hence the equation of quadric referred to these new set of axes Ox', Oy' and Oz' becomes

$$\tau_{xx}(l_1x' + l_2y' + l_3z')^2 + \tau_{yy}(m_1x' + m_2y' + m_3z')^2 + \tau_{zz}(n_1x' + n_2y' + n_3z)^2$$
$$+2\tau_{yz}(m_1x' + m_2y' + m_3z')(n_1x' + n_2y' + n_3z') + 2\tau_{zx}(n_1x' + n_2y' + n_3z')(l_1x' + l_2y' + l_3z')$$
$$+2\tau_{xy}(l_1x' + l_2y' + l_3z')(m_1x' + m_2y' + m_3z') = \pm k^2 \,. \tag{3.1.2}$$

We denote the coefficient of x'^2 on the left hand side of (3.1.2) as $\tau_{x'x'}$ and obtain

$$\tau_{x'x'} \equiv \tau_{xx}l_1^2 + \tau_{yy}m_1^2 + \tau_{zz}n_1^2 + 2\tau_{yz}m_1n_1 + 2\tau_{zx}n_1l_1 + 2\tau_{xy}l_1m_1 \,.$$

In a similar manner we can obtain the coefficients of y'^2 and z'^2 on the left hand side of (3.1.2) which are denoted by $\tau_{y'y'}$ and $\tau_{z'z'}$ respectively. Again we denote the coefficient of $2x'y'$ on the left hand side of (3.1.2) as $\tau_{x'y'}$ and obtain

$$\tau_{x'y'} = l_1l_2\tau_{xx} + m_1m_2\tau_{yy} + n_1n_2\tau_{zz} + \tau_{yz}(m_1n_2 + m_2n_1) + \tau_{zx}(n_1l_2 + n_2l_1) + \tau_{xy}(l_1m_2 + l_2m_1) \,.$$

In a similar manner we can obtain the coefficients of $2y'z'$ and $2z'x'$ on the left hand side of (3.1.2) which are denoted by $\tau_{y'z'}$ and $\tau_{z'x'}$ respectively. Hence the equation of quadric surface given by (3.1.1) when referred to $Ox', Oy;, Oz'$ axes takes the form

$$\tau_{x'x'}x'^2 + \tau_{y'y'}y'^2 + \tau_{z'z'}z'^2 + 2\tau_{y'z'}y'z' + 2\tau_{z'x'}z'x' + 2\tau_{x'y'}x'y' = \pm k^2 \,. \tag{3.1.3}$$

It is clearly seen from (3.1.3) that the coefficients give the components of stress with respect to the Ox', Oy', Oz' axes. If the quadric were referred to its principal axes, the tangential stresses across the coordinate planes would vanish. Hence the equation of the stress quadric referred to its principal axes as the coordinate axes OX, OY and OZ takes the form

$$\tau_{xx}X^2 + \tau_{yy}Y^2 + \tau_{zz}Z^2 = \pm k^2 \,,$$

which shows that the principal axes of the stress quadric will coincide with the principal stress direction at O.

3.1.1 Properties of stress quadric of Cauchy

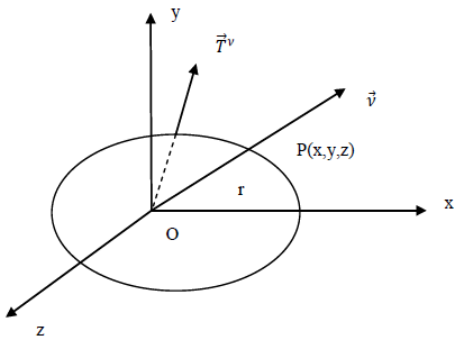
Property 1: *The normal stress across any plane through its centre is inversely proportional to the square of that radius vector of the quadric which is normal to the plane.*

Consider any unit area through a point O and let $\vec{\nu}$ be the unit normal to this elementary area whose direction cosines are l, m, n.

Let us draw the radius vector $\vec{OP}$ to the quadric in this direction such that $P(x, y, z)$ is a point on the quadric and $OP = r$. Therefore,

$$\frac{x}{r} = l, \quad \frac{y}{r} = m \quad \text{and} \quad \frac{z}{r} = n\,.$$

The stress exerted by the material on the side toward which normal $\vec{\nu}$ is drawn on the material on the opposite side across the unit area has its components $\tau_{\nu x}, \tau_{\nu y}$ and $\tau_{\nu z}$.



The component of the stress in the direction of normal $\vec{\nu}$, i.e., the normal component of stress is given by

$$\begin{aligned}
& l\tau_{\nu x} + m\tau_{\nu y} + n\tau_{\nu z} \\
&= l(l\tau_{xx} + m\tau_{yx} + n\tau_{zx}) + m(l\tau_{xy} + m\tau_{yy} + n\tau_{zy}) + n(l\tau_{xz} + m\tau_{yz} + n\tau_{zz}) \\
&= l^2\tau_{xx} + m^2\tau_{yy} + n^2\tau_{zz} + 2mn\tau_{yz} + 2nl\tau_{zx} + 2lm\tau_{xy} \\
&= \frac{1}{r^2}\left[\tau_{xx}x^2 + \tau_{yy}y^2 + \tau_{zz}z^2 + 2\tau_{yz}yz + 2\tau_{zx}zx + 2\tau_{xy}xy\right] \\
&= \pm\frac{k^2}{r^2}\,, \text{using (3.1.1)}\,,
\end{aligned}$$

as (x, y, z) is a point on the quadric.

Hence, if the quadric surface with centre at O given by (3.1.1) can be drawn then the normal stress on any unit area can easily be derived drawing the normal to the unit area. If the normal cuts the quadric surface at a distance r from O then $+\frac{k^2}{r^2}$ is the normal stress.

Property 2: *The normal to the quadric surface* (3.1.1) *at the end of the radius vector* $\vec{OP}$ *is parallel to the stress vector* $\vec{T}^{\nu}$ *at* O *where* $\vec{OP}$ *is in the direction of* $\vec{\nu}$ *and* $P(x, y, z)$ *is a point on the quadric surface.*

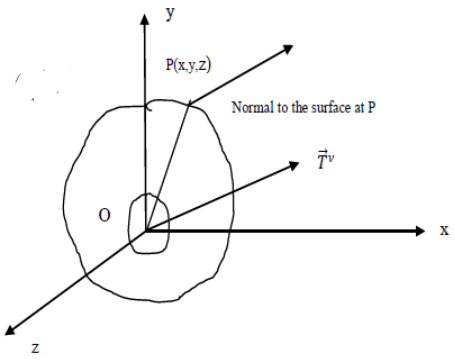
First note that the stress exerted by the material on the side of $\vec{\nu}$ across unit area at O of which the normal $\vec{\nu}$ to the material on the other side is $\vec{T}^{\nu}$. Its components are $\tau_{\nu x}, \tau_{\nu y}$ and $\tau_{\nu z}$. Hence the direction ratios of $\vec{T}^{\nu}$ are $\tau_{\nu x}, \tau_{\nu y}$ and $\tau_{\nu z}$.

Now the equation of the quadric surface can be expressed as

$$f(x, y, z) \equiv \tau_{xx}x^2 + \tau_{yy}y^2 + \tau_{zz}z^2 + 2\tau_{yz}yz + 2\tau_{zx}zx + 2\tau_{xy}xy \mp k^2 = 0\,.$$

Then the direction ratios of the normal to the surface at $P(x, y, z)$ is given by

$$\begin{aligned}
\frac{\partial f}{\partial x} &= 2(x\tau_{xx} + y\tau yx + z\tau_{zx}) \\
&= 2r(l\tau_{xx} + m\tau yx + n\tau_{zx}) \\
&= 2r\tau_{\nu x}
\end{aligned}$$



In a similar manner,

$$\frac{\partial f}{\partial y} = 2r\tau_{\nu y}\,, \quad \frac{\partial f}{\partial y} = 2r\tau_{\nu z}\,.$$

This shows that the normal to the quadric surface at P is parallel to the stress vector $\vec{T}^{\nu}$.

Example: *Determine the Cauchy's stress quadric at a point P for a state of stress*

$$(T_{ij}) = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix}\,,$$

where a, b, c are all of same sign.

Solution: Consider a point $P(x, y, z)$ in the deformed state of a continuum body. The stress tensor T_{ij} at P with respect to a system of axes Ox, Oy, Oz fixed in space are given by

$$(T_{ij}) = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix}\,.$$

Since all the shearing stresses $T_{ij}\,(i \neq j)$ vanishes, the coordinate axes are the principal axes of stresses.

The equation of the stress quadric with respect to the principal axes is given by

$$\begin{aligned} & T_{11}x^2 + T_{22}y^2 + T_{33}z^2 = \text{constant} \\ \Rightarrow \quad & ax^2 + by^2 + cz^2 = k\,(\text{say}) \\ \Rightarrow \quad & \frac{x^2}{k/a} + \frac{y^2}{k/b} + \frac{z^2}{k/c} = 1\,, \end{aligned}$$

which is the required Cauchy's stress quadric at P. It represents an ellipsoid.

3.2 Extreme normal and shearing stresses

To determine the extreme normal and shearing stresses at any point O of a continuum, we consider the coordinate axes at O as Ox, Oy and Oz which are in the direction of principal stresses. Let

$\sigma_1, \sigma_2, \sigma_3$ be the corresponding stresses such that

$$\tau_{xx} = \sigma_1, \tau_{yy} = \sigma_2, \tau_{zz} = \sigma_3, \tau_{xy} = \tau_{yz} = \tau_{zx} = 0\,.$$

We consider any arbitrary unit area through O with normal $\vec{\nu}$ determined by the direction cosines l, m, n. The stress vector $\vec{T}^{\nu}$ at O has components $\tau_{\nu x}, \tau_{\nu y}, \tau_{\nu z}$ where

$$\begin{aligned}
\tau_{\nu x} &= l\tau_{xx} + m\tau_{yx} + n\tau_{zx} = \sigma_1 l \\
\tau_{\nu y} &= l\tau_{xy} + m\tau_{yy} + n\tau_{zy} = \sigma_2 m \\
\tau_{\nu z} &= l\tau_{xz} + m\tau_{yz} + n\tau_{zz} = \sigma_3 n
\end{aligned} \tag{3.2.1}$$

The resultant stress is given by

$$(\vec{T}^{\nu})^2 = l^2\sigma_1^2 + m^2\sigma_2^2 + n^2\sigma_3^2\,. \tag{3.2.2}$$

The normal stress N can be expressed as

$$\begin{aligned}
N &= l\tau_{\nu x} + m\tau_{\nu y} + n\tau_{\nu z} \\
&= l^2\sigma_1 + m^2\sigma_2 + n^2\sigma_3\,.
\end{aligned} \tag{3.2.3}$$

(i) **(Extreme normal stress)** *The extremum values of normal stress at a point of a continuum are principal stresses.*

We choose principal stresses $\sigma_1, \sigma_2, \sigma_3$ in such a manner that $\sigma_1 > \sigma_2 > \sigma_3$. We are to extremize the value of N subject to the constraint

$$l^2 + m^2 + n^2 = 1\,. \tag{3.2.4}$$

Construct a Lagrangian function

$$\begin{aligned}
F(l, m, n) &= N - \lambda(l^2 + m^2 + n^2 - 1) \\
&= l^2\sigma_1 + m^2\sigma_2 + n^2\sigma_3 - \lambda(l^2 + m^2 + n^2 - 1)\,,
\end{aligned} \tag{3.2.5}$$

where λ is a parameter. For extremum value of F, we have

$$\begin{aligned}
&\frac{\partial F}{\partial l} = 0\,, \qquad \frac{\partial F}{\partial m} = 0\,, \qquad \frac{\partial F}{\partial n} = 0\,. \\
\Rightarrow\ & l(\sigma_1 - \lambda) = 0\,, \quad m(\sigma_2 - \lambda) = 0\,, \quad n(\sigma_3 - \lambda) = 0\,.
\end{aligned} \tag{3.2.6}$$

Therefore

$$\begin{aligned}
& l^2(\sigma_1 - \lambda) + m^2(\sigma_2 - \lambda) + n^2(\sigma_3 - \lambda) = 0 \\
\Rightarrow\ & \lambda(l^2 + m^2 + n^2) = l^2\sigma_1 + m^2\sigma_2 + n^2\sigma_3 \\
\Rightarrow\ & \lambda = N\,.
\end{aligned} \tag{3.2.7}$$

Hence

$$l(\sigma_1 - N) = 0\,, \quad m(\sigma_2 - N) = 0\,, \quad n(\sigma_3 - N) = 0\,. \tag{3.2.8}$$

The above equations determines three unknowns l, m, n for which N is extremum. The trivial zero solution $l = m = n = 0$ of (3.2.8) is not compatible with the constraint (3.2.4). One type of nontrivial solution can be considered as

$$l = m = 0, n \neq 0\,.$$

Then (3.2.4) gives

$$n^2 = 1 \quad \Rightarrow n = \pm 1\,.$$

The first two equations of (3.2.8) are identically satisfied and the third equation gives $N = \sigma_3$. In a similar manner, for the solution $l = 0, m \neq 0, n = 0$, (3.2.8) gives $N = \sigma_2$ and for the solution $l \neq 0, m = n = 0$, (3.2.8) gives $N = \sigma_1$. Since $\sigma_1 > \sigma_2 > \sigma_3$, maximum value of the normal stress $N = \sigma_1$ and minimum value of $N = \sigma_3$.

Therefore, extremum values of the normal stress at a point are always principal stresses acting across planes for which shearing stress components vanishes identically.

(ii) **(Extreme shearing stress)** *The maximum shearing stress acts on the plane that bisects the angle between the greatest and the smallest principal stress planes. Its value is one half the difference between the greatest and the smallest principal stresses.*

Let S be the magnitude of the shearing stress. Then

$$\begin{aligned} & S^2 + N^2 = (\vec{T}^{\nu})^2 \\ \Rightarrow\ & S^2 = (\vec{T}^{\nu})^2 - N^2 \\ \Rightarrow\ & S^2 = l^2\sigma_1^2 + m^2\sigma_2^2 + n^2\sigma_3^2 - (l^2\sigma_1 + m^2\sigma_2 + n^2\sigma_3)^2\,. \end{aligned} \tag{3.2.9}$$

Since $l^2 + m^2 + n^2 = 1$, the expression for shearing stress can be written in terms of two variables l and m only. We use

$$n^2 = 1 - l^2 - m^2$$

in relation (3.2.9) and obtain

$$S^2 = (\sigma_1^2 - \sigma_3^2)l^2 + (\sigma_2^2 - \sigma_3^2)m^2 + \sigma_3^2 - [(\sigma_1 - \sigma_3)l^2 + (\sigma_2 - \sigma_3)m^2 + \sigma_3]^2\,. \tag{3.2.10}$$

The maximum value of shearing stress can be obtained by equating the partial derivatives of S with respect to l and m to zero. At the values of l and m, for which S is maximum, S^2 will also be maximum. Hence

$$\frac{\partial S^2}{\partial l} = 0 \quad \text{and} \quad \frac{\partial S^2}{\partial l} = 0\,,$$

which implies

$$(\sigma_1^2 - \sigma_3^2)l - 2[(\sigma_1 - \sigma_3)l^2 + (\sigma_2 - \sigma_3)m^2 + \sigma_3](\sigma_1 - \sigma_3)l = 0 \tag{3.2.11}$$

$$(\sigma_2^2 - \sigma_3^2)m - 2[(\sigma_1 - \sigma_3)l^2 + (\sigma_2 - \sigma_3)m^2 + \sigma_3](\sigma_2 - \sigma_3)m = 0 \tag{3.2.12}$$

First, we consider the most general case $\sigma_1 \neq \sigma_2 \neq \sigma_3$.

Dividing (3.2.11) by $(\sigma_1 - \sigma_3)$ and (3.2.12) by $(\sigma_2 - \sigma_3)$, we get

$$\{(\sigma_1 - \sigma_3) - 2[(\sigma_1 - \sigma_3)l^2 + (\sigma_2 - \sigma_3)m^2]\}l = 0 \tag{3.2.13}$$

$$\{(\sigma_2 - \sigma_3) - 2[(\sigma_1 - \sigma_3)l^2 + (\sigma_2 - \sigma_3)m^2]\}m = 0\,. \tag{3.2.14}$$

We have two equations of degree three in l and m and accordingly we shall obtain three solutions. The simplest one is $l = 0, m = 0, n = 1$. Corresponding to these set of values, we find from (3.2.10)

$$S = 0\,.$$

This merely verified the known fact that plane element normal to the principal direction is shear stress free. Therefore the minimum value of $|S|$ is associated with the principal stress direction. As we are seeking for maximum shearing stress, so we discard these values of l, m, n and have other three possibilities

$$(i)\, l \neq 0, m = 0,\ (ii)\, l = 0, m \neq 0,\ (iii)\, l \neq 0, m \neq 0\,.$$

The last case is impossible as then cancelling l and m from (3.2.14) and (3.2.14) respectively and subtracting the resulting equations, we obtain $\sigma_1 = \sigma_2$, which contradicts our assumption $\sigma_1 \neq \sigma_2 \neq \sigma_3$.

Now consider the first case i.e., $l \neq 0, m = 0$. Then from (3.2.14), we get

$$(\sigma_1 - \sigma_3)(1 - 2l^2) = 0 \quad \Rightarrow l = \pm\frac{1}{\sqrt{2}}\,.$$

Hence

$$l = \pm\frac{1}{\sqrt{2}},\ m = 0 \text{ and } n = \pm\frac{1}{\sqrt{2}}\,,\ \text{as } l^2 + m^2 + n^2 = 1\,. \tag{3.2.15}$$

Considering the second case i.e., $l = 0, m \neq 0$, we get from (3.2.14)

$$(\sigma_2 - \sigma_3)(1 - 2m^2) = 0 \quad \Rightarrow m = \pm\frac{1}{\sqrt{2}}\,.$$

Hence

$$l = 0,\ m = \pm\frac{1}{\sqrt{2}} \text{ and } n = \pm\frac{1}{\sqrt{2}}\,. \tag{3.2.16}$$

If at the outset we eliminate m instead of n from (3.2.9) and repeat the same analysis, we can obtain one more solution

$$l = \pm\frac{1}{\sqrt{2}},\ m = \pm\frac{1}{\sqrt{2}},\ n = 0\,. \tag{3.2.17}$$

Hence the extremum values of shearing stresses can be obtained by substituting these values in (3.2.10). Then from (3.2.10) and (3.2.15), we have

$$\begin{aligned}(S_1^2)_{\text{extremum}} &= \frac{\sigma_1^2 - \sigma_3^2}{2} + \sigma_3^2 - \left[\frac{\sigma_1 - \sigma_3}{2} + \sigma_3\right]^2 \\ &= \frac{\sigma_1^2 + \sigma_3^2}{2} - \left[\frac{\sigma_1 + \sigma_3}{2}\right]^2 \\ &= \left[\frac{\sigma_1 - \sigma_3}{2}\right]^2 \\ (S_1)_{\text{extremum}} &= \pm\frac{1}{2}(\sigma_1 - \sigma_3)\,.\end{aligned}$$

In a similar manner, from (3.2.16), (3.2.10) and (3.2.17), (3.2.10) we have

$$(S_2)_{\text{extremum}} = \pm\frac{1}{2}(\sigma_2 - \sigma_3) \text{ and}$$
$$(S_3)_{\text{extremum}} = \pm\frac{1}{2}(\sigma_1 - \sigma_2)\,.$$

Now if $\sigma_1 > \sigma_2 > \sigma_3$, the maximum value of $|S|$ is

$$|S|_{\text{max}} = \frac{1}{2}(\sigma_1 - \sigma_3)\,.$$

In this case we find from (3.2.15) that the maximum shearing stress acts on the plane containing the y axis and bisecting the angle between x and z axis.

Therefore, we conclude that the maximum shearing stress is equal to the half of the difference between the greatest and least of $\sigma_1, \sigma_2, \sigma_3$ and acts on the plane which bisects the angle between the direction of largest and smallest.

Example: *Let the components of the stress tensor at P be given in the matrix form*

$$(T_{ij}) = \begin{bmatrix} 3 & 1 & 1 \\ 1 & 0 & 2 \\ 1 & 2 & 0 \end{bmatrix}\,.$$

Determine principal stresses and principal directions. Also find the magnitude of the maximum shearing stress.

Solution: The principal stresses $\sigma_1, \sigma_2, \sigma_3$ at the point P are the roots of the characteristic equation

$$\begin{vmatrix} 3-\sigma & 1 & 1 \\ 1 & 0-\sigma & 2 \\ 1 & 2 & 0-\sigma \end{vmatrix} = 0$$
$$\Rightarrow \sigma^3 - 3\sigma^2 - 6\sigma + 8 = 0$$
$$\Rightarrow \sigma = -2, 1, 4\,.$$

The principal direction of stress at P are given by the set of equations

$$(3-\sigma)l + m + 2n = 0$$
$$l - \sigma n + 2n = 0$$
$$l + 2m - \sigma n = 0\,.$$

For $\sigma = -2$, the above system of equations becomes

$$5l + m + 2n = 0$$
$$l + 2m + 2n = 0$$
$$l + 2m + 2n = 0\,.$$

The solution is

$$(l, m, n) = (0, 1, -1)\,.$$

For $\sigma = 1$, the above system of equations becomes

$$\begin{aligned} 2l + m + 2n &= 0 \\ l - m + 2n &= 0 \\ l + 2m - n &= 0\,. \end{aligned}$$

The solution is

$$(l, m, n) = (1, -1, -1)\,.$$

For $\sigma = 4$, the above system of equations becomes

$$\begin{aligned} -l + m + 2n &= 0 \\ l - 4m + 2n &= 0 \\ l + 2m - 4n &= 0\,. \end{aligned}$$

The solution is

$$(l, m, n) = (2, 1, 1)\,.$$

Thus the principal directions are given by

$$\frac{1}{\sqrt{2}}(0, 1, -1),\ \frac{1}{\sqrt{3}}(1, -1, 1),\ \frac{1}{\sqrt{6}}(2, 1, 1)\,.$$

As $4 > 1 > -2$, $\sigma_1 = 4$ is the largest and $\sigma_3 = -2$ is the smallest principal stress value.

Thus the maximum value of the shearing stress is given by

$$|S| = \frac{1}{2}(\sigma_1 - \sigma_3) = \frac{1}{2}\{4 - (-2)\} = 3\,.$$

Finally, the maximum value of the normal stress is the maximum principal stress value which is 4.

Example: *At point P, there are three principal stresses $\sigma_1, \sigma_2, \sigma_3$ such that $2\sigma_2 = \sigma_1 + \sigma_3$. Determine the unit normal for the plane on which normal stress is σ_2 and shearing stress is $\frac{1}{4}(\sigma_1 - \sigma_3)$.*

Solution: Let l, m, n be the direction cosines of the unit normal for the plane on which normal and shearing stresses are given. Since $\sigma_1, \sigma_2, \sigma_3$ are the principal stresses, then normal stress N is given by

$$\begin{aligned} N &= l^2\sigma_1 + m^2\sigma_2 + n^2\sigma_3 \\ \Rightarrow \sigma_2 &= l^2\sigma_1 + m^2\sigma_2 + n^2\sigma_3 \\ \Rightarrow (1 - m^2)\sigma_2 &= l^2\sigma_1 + n^2\sigma_3\,. \end{aligned} \tag{3.2.18}$$

Let S be the magnitude of shearing stress, then

$$\begin{aligned} S^2 &= l^2\sigma_1^2 + m^2\sigma_2^2 + n^2\sigma_3^2 - N^2 \\ \Rightarrow \left(\frac{\sigma_1 - \sigma_3}{4}\right)^2 &= l^2\sigma_1^2 + (m^2 - 1)\sigma_2^2 + n^2\sigma_3^2\,. \end{aligned} \tag{3.2.19}$$

Also

$$l^2 + m^2 + n^2 = 1\,, \tag{3.2.20}$$

and

$$\sigma_2 = \frac{\sigma_1 + \sigma_3}{2}\,. \tag{3.2.21}$$

Eliminating σ_2 from (3.2.18) and (3.2.21), we get

$$\begin{aligned}
&(1-m^2)\frac{\sigma_1+\sigma_3}{2} = l^2\sigma_1 + n^2\sigma_3 \\
\Rightarrow\ &(l^2+n^2)\frac{\sigma_1+\sigma_3}{2} = l^2\sigma_1 + n^2\sigma_3 \\
\Rightarrow\ &\frac{n^2-l^2}{2}\sigma_1 + \frac{l^2-n^2}{2}\sigma_3 = 0 \\
\Rightarrow\ &(l^2-n^2)(\sigma_3-\sigma_1) = 0 \\
\Rightarrow\ &l = n\,.
\end{aligned}$$

Then

$$m^2 = 1 - l^2 - n^2 = 1 - 2l^2 = 1 - 2n^2\,.$$

Again eliminating σ_2 from (3.2.19) and (3.2.21), we get

$$\begin{aligned}
&\left(\frac{\sigma_1-\sigma_3}{4}\right)^2 = l^2\sigma_1^2 + (m^2-1)\left(\frac{\sigma_1+\sigma_3}{2}\right)^2 + n^2\sigma_3^2 \\
\Rightarrow\ &\left(\frac{\sigma_1-\sigma_3}{4}\right)^2 = l^2\sigma_1^2 + l^2\sigma_3^2 + (1-2l^2-1)\left(\frac{\sigma_1+\sigma_3}{2}\right)^2, \quad \text{as } l=n,\ m^2 = 1-2l^2 \\
\Rightarrow\ &\left(\frac{\sigma_1-\sigma_3}{4}\right)^2 = \frac{l^2}{2}(\sigma_1-\sigma_3)^2 \\
\Rightarrow\ &\left(\frac{1}{16} - \frac{l^2}{2}\right)(\sigma_1-\sigma_3)^2 = 0 \\
\Rightarrow\ &l^2 = \frac{1}{8} \\
\Rightarrow\ &l = \frac{1}{2\sqrt{2}}\,.
\end{aligned}$$

Hence

$$n = \frac{1}{2\sqrt{2}}\,,$$

and

$$m^2 = 1 - 2\cdot\frac{1}{8} = \frac{3}{4} \quad \Rightarrow m = \frac{\sqrt{3}}{2}\,.$$

Hence the required unit normal is $\left(\frac{1}{2\sqrt{2}}, \frac{\sqrt{3}}{2}, \frac{1}{2\sqrt{2}}\right)$.

3.3 Mohr's circles for stress

Mohr's circle is a two dimensional graphical representation of the transformation law for the Cauchy stress tensor and it is a useful technique for finding principal stresses and strains in materials.

3.3.1 Mohr's circle for three dimensional state of stress

To construct the Mohr's circle for a general three dimensional case of stresses at a point, the values of the principal stresses $\sigma_1, \sigma_2, \sigma_3$ must be evaluated.

Let the principal stresses be ordered according to $\sigma_1 > \sigma_2 > \sigma_3$. If N and S denote the normal and shear stress at P, then

$$N = l^2\sigma_1 + m^2\sigma_2 + n^2\sigma_3\,, \tag{3.3.1}$$

and

$$S^2 = l^2\sigma_1^2 + m^2\sigma_2^2 + n^2\sigma_3^2 - N^2\,. \tag{3.3.2}$$

Also

$$l^2 + m^2 + n^2 = 1\,. \tag{3.3.3}$$

Using Gauss elimination method we obtain

$$l^2 = \frac{(N-\sigma_2)(N-\sigma_3) + S^2}{(\sigma_1-\sigma_2)(\sigma_1-\sigma_3)}\,, \tag{3.3.4}$$

$$m^2 = \frac{(N-\sigma_3)(N-\sigma_1) + S^2}{(\sigma_2-\sigma_3)(\sigma_2-\sigma_1)}\,, \tag{3.3.5}$$

$$n^2 = \frac{(N-\sigma_1)(N-\sigma_2) + S^2}{(\sigma_3-\sigma_1)(\sigma_3-\sigma_2)}\,. \tag{3.3.6}$$

In the above three equations $\sigma_1, \sigma_2, \sigma_3$ are known and N, S are functions of l, m, n.

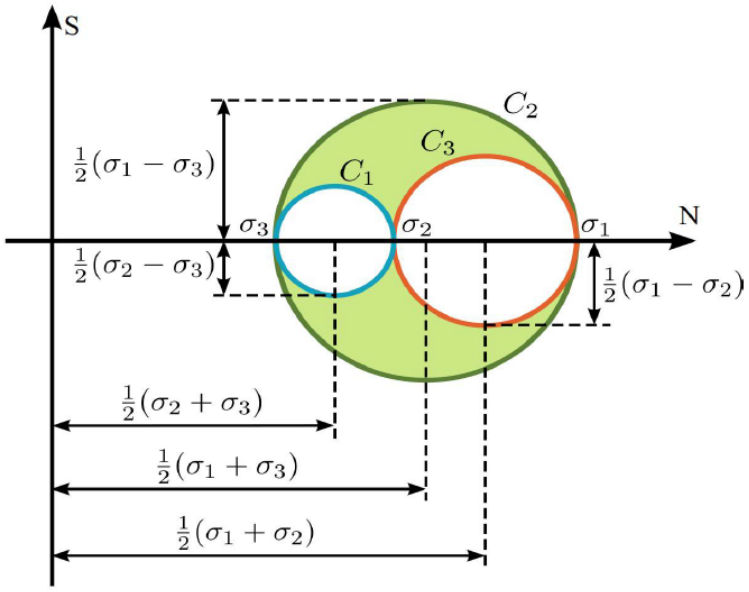
We would like to interpret these equations graphically by representing conjugate pairs of N, S values, which satisfy (3.3.4), (3.3.5) and (3.3.6) as a point on the stress plane having N as abscissa and S as ordinate.

To develop the graphical interpretation of the three dimensional stress state in terms of N and S, note that the denominator for the expression of l^2 in (3.3.4) is positive since both $\sigma_1 - \sigma_2 > 0$ and $\sigma_1 - \sigma_3 > 0$ and $l^2 \geq 0$. Therefore, we have

$$(N-\sigma_2)(N-\sigma_3) + S^2 \geq 0\,. \tag{3.3.7}$$

In case of equality, (3.3.7) can be written as

$$\begin{aligned} &(N-\sigma_2)(N-\sigma_3) + S^2 = 0 \\ \Rightarrow\ &N^2 - (\sigma_2+\sigma_3)N + S^2 + \sigma_2\sigma_3 = 0 \\ \Rightarrow\ &\left(N - \frac{\sigma_2+\sigma_3}{2}\right)^2 + S^2 = \left(\frac{\sigma_2-\sigma_3}{2}\right)^2\,. \end{aligned}$$



It represents a circle in the $N,\ S$ plane with its centre at $\dfrac{\sigma_2+\sigma_3}{2}$ on N axis having radius $\dfrac{\sigma_2-\sigma_3}{2}$. We label this circle by C_1 as shown in figure. For the case in which the inequality sign holds for (3.3.7), we observe that conjugate pairs of values of N and S which satisfy this relationship result in stress points having coordinates exterior to circle C_1. Thus combinations of N and S which satisfy (3.3.4) lie on or exterior to circle C_1.

Again from (3.3.5), note that the denominator is negative since $\sigma_2 - \sigma_3 < 0$ and $\sigma_2 - \sigma_1 > 0$. Also $m^2 \geq 0$. Then we have

$$(N-\sigma_3)(N-\sigma_1) + S^2 \leq 0\,. \tag{3.3.8}$$

In case of equality, (3.3.8) can be written as

$$\begin{aligned} &(N-\sigma_3)(N-\sigma_1) + S^2 = 0 \\ \Rightarrow\ &N^2 - (\sigma_1+\sigma_3)N + S^2 + \sigma_1\sigma_3 = 0 \\ \Rightarrow\ &\left(N - \frac{\sigma_1+\sigma_3}{2}\right)^2 + S^2 = \left(\frac{\sigma_1-\sigma_3}{2}\right)^2 . \end{aligned}$$

It represents a circle in the $N,\ S$ plane with its centre at $\dfrac{\sigma_1+\sigma_3}{2}$ on N axis having radius $\dfrac{\sigma_1-\sigma_3}{2}$. We label this circle by C_2 as shown in figure. The stress points which satisfy the inequality of (3.3.8) lie interior to circle C_2.

Following the same procedure, we obtain the third circle C_3 extracting from (3.3.6) which is given by

$$\left(N - \frac{\sigma_1+\sigma_2}{2}\right)^2 + S^2 = \left(\frac{\sigma_1-\sigma_2}{2}\right)^2 .$$

The admissible stress points in the $N,\ S$ plane lie on or exterior to this circle. The three circles defined above and shown in figure are called *Mohr's circles for stress*. All possible pairs of values

of N and S at P which satisfy (3.3.4), (3.3.5), (3.3.6) lie on these circles or within the shaded area enclosed by them. In addition it is clear from Mohr's circles diagram that the maximum shear stress value at P is the radius of the circle C_2 which confirms the result $\dfrac{\sigma_1 - \sigma_3}{2}$.

3.4 Few Probable Questions

1. The state of stress of a point is given by

$$(T_{ij}) = \begin{bmatrix} a & 2 & 1 \\ 2 & 0 & 2 \\ 1 & 2 & 0 \end{bmatrix},$$

where a is a constant. Determine a such that there is atleast one plane through the point in such a why that resultant stress on that plane is zero. Determine the direction cosines of the normal to the plane. $\left[\text{Ans}: \ 2; \ \pm\frac{2}{3}, \ \mp\frac{1}{3}, \ \mp\frac{2}{3}\right]$.

2. The stress tensor at a point is

$$(T_{ij}) = \begin{bmatrix} -a & 0 & d \\ 0 & b & c \\ d & e & e \end{bmatrix},$$

Determine the unit normal of a plane parallel to z-axis on which the resultant stress vector is tangential to the plane. $\left[\text{Ans}: \ \sqrt{\frac{b}{a+b}}, \sqrt{\frac{b}{a+b}}, 0\right]$.

3. The stress tensor at a point is given by

$$(T_{ij}) = \begin{bmatrix} 0 & 1 & 2 \\ 0 & b & 1 \\ 2 & 1 & 0 \end{bmatrix},$$

where b is a constant. Find b so that stress vector on some place at the point will be zero. Determine the direction cosines of the normal to the plane. $\left[\text{Ans}: \ 1; \ \left(\frac{1}{\sqrt{6}}, -\frac{2}{\sqrt{6}}, \frac{1}{\sqrt{6}}\right)\right]$.

4. The state of stress at a point is given by

$$(T_{ij}) = \begin{bmatrix} 4 & 0 & 2 \\ 0 & 8 & 0 \\ 2 & 0 & -12 \end{bmatrix},$$

Compute the magnitude of the maximum shearing stress.

5. At the point P, principal stresses are $4, \ 5, \ 6$. Determine the unit normal for the plane upon which normal stress is 5 and shearing stress is $\dfrac{1}{2}$.

Unit 4

Course Structure

- Deformation, Method of description: Lagrangian & Eulerian
- Deformation gradient and finite strain tensor

4.1 Deformation

In a rigid body for all possible motion of it, the distances between any pair of particles of it remain constant for all times. A deformable body is that in which the distance between any two particles of it under the influence of external effect changes. Thus when the relative position of points in a continuous body is altered, we say that the body is strained. The change in the relative position of points is called deformation.

The deformations which are recovered after the stress field has been removed are called elastic deformation. In this case, the continuum completely recovers its original configuration. Here we will describe all motions and deformations by expressing positions of points in both undeformed and deformed solids as components in a cartesian reference frame (which is also taken to be an inertial frame). Thus x_i will denote components of the position vector of a material particle before deformation and $X_i(x_k)$ will be components of its position vector after deformation.

Mathematically, we describe a deformation as a 1:1 mapping which transforms points from the reference configuration of a solid to the deformed configuration. More specifically, let ϵ_i be three numbers specifying the position of some points in the undeformed solid (these could be the three components of position vector in a Cartesian coordinate system). As the solid deforms, each the value of the coordinates change to different numbers. We can write this in general form as

$$\eta_i = f_i(\epsilon_i, t). \tag{4.1.1}$$

This is called deformation mapping.

4.2 Method of Description

When analyzing the deformation or motion of solids, or the flow of fluids, it is necessary to describe the sequence or evolution of configurations throughout time. There are two methods of analyzing the properties of the deformation, viz.

- Lagrangian or material method,
- Eulerian or spatial method.

4.2.1 Lagrangian description or material method

In the material or Lagrangian method, our object of study is material points of a continuum body. In this method, we identity individual material points and describe the motion of each individual material point of fixed identity for all time by following its motin throughout its course. In this approach, individual material points are possessed with physical properties, which may be changed in two ways.

- They change as we pass form one material point to another and
- they change as time changes for a fixed material point.

In other words, these properties are considered as functions of time and data which identify the material points. These are normally denoted by uppercase variables X, Y and Z and are used to label material particles. For such data we usually take the rectangular cartesian coordinates X_1, X_2, X_3 of the position of a material point of the continuum body in its initial undeformed state. We identify the given material point by (X_1, X_2, X_3). It is given a fixed identity by specifying its initial position. All physical properties associated with this material point will then the functions of X_1, X_2, X_3 and time t. If (x_1, x_2, x_3) be the rectangular cartesian coordinates of this position, then

$$x_i = x_i(X_1, X_2, X_3, t),\ i = 1, 2, 3\,. \tag{4.2.1}$$

This equation describes motion of the material point completely in material method giving the subsequent position at time t. The coordinates X_1, X_2, X_3 are independent coordinates and are called material coordinates or Lagrangian coordinates, whereas x_1, x_2, x_3 are dependent coordinates and are called spatial coordinates.

4.2.2 Eulerian description or spatial method

In the spatial or Eulerian method, our object of study strictly speaking, is not moving material points but fixed spatial point. We identify the spatial points and describe the motion of the medium at each spatial point at different times without considering the whereabouts of individual material points. We focus our attention on a fixed spatial point in space occupied by different material points at different times and observe that changes of various properties are taking place at the spatial point. In this approach, the physical properties change in two ways.

- When we pass from one spatial point to another point and

- with time at a fixed spatial point.

If a material point which was at a position (X_1, X_2, X_3) in the undeformed state at $t = 0$ happens to occupy the position (x_1, x_2, x_3) at subsequent time t, then coordinates x_1, x_2, x_3 identify the spatial point in the deformed state. The physical properties will be functions of (x_1, x_2, x_3) and time t. In particular,

$$X_i = X_i(x_1, x_2, x_3, t),\ i = 1, 2, 3\,, \tag{4.2.2}$$

which traces the material point occupying spatial position (x_1, x_2, x_3).

The distinction between material and spatial description is basic, in the former X_1, X_2, X_3, t are independent variables whereas in the latter x_1, x_2, x_3, t are independent variables. An elastic solid has undeformed state to which it always return whenever loads are removed. To describe this property of elasticity, the undeformed state must be marked by identifying material points. Thus material description is natural for elastic body. On the other hand, the fluid has no past undeformed configuration. The response of the fluid is determined solely by instantaneous values of time rates of deformation. For this reason, it is natural to use spatial description for fluid.

4.3 Displacement

A change in the configuration of a continumm body results in a displacement. The displacement of a body has two components, a rigid body displacement and a deformation . A rigid body displacement consists of a simultaneous translation and rotation of the body about an axis without changing its shape or size. Deformation implies the change of its configuration and orientation as well as in shape.

The displacement of a typical material point from its position (X_1, X_2, X_3) in the undeformed state at $t = 0$ to its position (x_1, x_2, x_3) in the deformed state at time t is defined by

$$u_i = x_i - X_i,\ i = 1, 2, 3\,. \tag{4.3.1}$$

In the *material description*, u_i and x_i are regarded as functions of X_1, X_2, X_3 and t so that displacement

$$u_i(X_1, X_2, X_3, t) = x_i(X_1, X_2, X_3, t) - X_i\,. \tag{4.3.2}$$

In the *spatial description*, u_i and X_i are regarded as functions of x_1, x_2, x_3 and t so that displacement

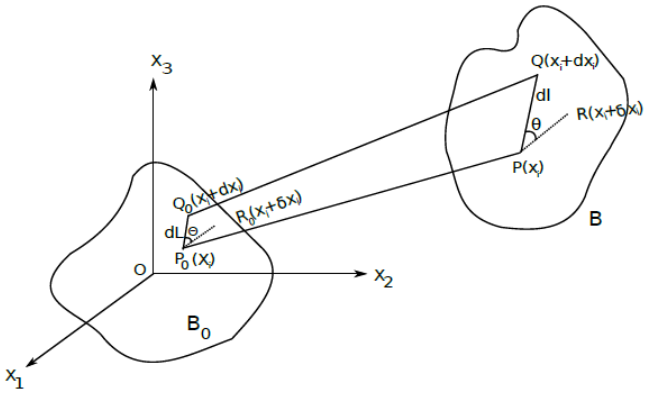
$$u_i(x_1, x_2, x_3) = x_i - X_i(x_1, x_2, x_3, t)\,. \tag{4.3.3}$$

4.4 Deformation gradients, Finite strain tensor

The deformation gradient is the fundamental measure of deformation. It is a second order tensor which maps line elements in the reference configuration into line elements in the current configuration.

4.4.1 Lagrangian finite strain tensor (Change in the length of a line element in material method)

Consider a material line element P_0Q_0, joining a pair of neighbouring points $P_0,\ Q_0$ of length dL oriented in the direction $N = (N_1, N_2, N_3)$ in the initial underformed region B_0 at time $t = 0$. If P_0



has coordinates (X_1, X_2, X_3) and Q_0 has coordinates $(X_1 + dX_1, X_2 + dX_2, X_3 + dX_3)$ with respect to an orthogonal set of coordinate axes fixed in space, then

$$\begin{aligned} dL^2 &= dX_1^2 + dX_2^2 + dX_3^2 \\ &= dX_i \cdot dX_i \\ &= \delta_{ij} \cdot dX_i \cdot dX_j \,, \end{aligned} \tag{4.4.1}$$

and

$$N_i = \frac{dX_i}{dL} \,, \tag{4.4.2}$$

where δ_{ij} is a Kronecker delta defined by

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

When the body undergoes deformation, the same material points which lie on P_0Q_0 at $t = 0$ will lie on a new line element PQ of length dl oriented in the direction (n_1, n_2, n_3) in current region B at time t.

If P has coordinates $x = (x_1, x_2, x_3)$ and Q has coordinates $(x_1 + dx_1, x_2 + dx_2, x_3 + dx_3)$ with respect to the same fixed set of coordinate axes, then

$$\begin{aligned} dl^2 &= dx_1^2 + dx_2^2 + dx_3^2 \\ &= dx_k \cdot dx_k \\ &= \delta_{kl}\, dx_k\, dx_l \end{aligned} \tag{4.4.3}$$

and

$$n_i = \frac{dx_i}{dl} \,. \tag{4.4.4}$$

In material method, the deformation is characterized by the equation

$$x_k = x_k(X_1, X_2, X_3, t) \tag{4.4.5}$$

Since $x_k + dx_k$ are coordinates of Q at the same time t, thus

$$\begin{aligned} dx_k &= \frac{\partial x_k}{\partial X_1} dX_1 + \frac{\partial x_k}{\partial X_2} dX_2 + \frac{\partial x_k}{\partial X_3} dX_3 \\ &= \frac{\partial x_k}{\partial X_j} dX_j \\ &= \frac{\partial x_k}{\partial X_i} dX_i \\ &= x_{k,i}\, dX_i\,. \end{aligned} \tag{4.4.6}$$

The quantity $x_{k,i} = \frac{\partial x_k}{\partial X_i}$ is called *deformation gradient tensor* or simply the *deformation gradient* and is denoted by F_{ki}. Sometimes the notation $\mathbf{F}$ is used to express the deformation gradient as

$$\mathbf{F} = \frac{\partial x}{\partial X} = \text{Grad}\; x\,. \tag{4.4.7}$$

The capital 'G' is used on"Grad" to emphasize that this is a gradient with respect to the material coordinates, the material gradient. Now using Eq.(4.4.6) in Eq.(4.4.3) we have

$$\begin{aligned} dl^2 &= \delta_{kl} \frac{\partial x_k}{\partial X_i} dX_i \frac{\partial x_l}{\partial X_j} dX_j \\ &= \frac{\partial x_k}{\partial X_i} \frac{\partial x_k}{\partial X_j} dX_i\, dX_j \;(\text{as } \delta_{kl} = 1\,,\; if\; k = l)\,. \end{aligned} \tag{4.4.8}$$

The difference $dl^2 - dL^2$ is a measure of change of length of line element. Therefore,

$$\begin{aligned} dl^2 - dL^2 &= \frac{\partial x_k}{\partial X_i} \frac{\partial x_k}{\partial X_j} dX_i\, dX_j - \delta_{ij} dX_i\, dX_j \\ &= \left[\frac{\partial x_k}{\partial X_i} \frac{\partial x_k}{\partial X_j} - \delta_{ij} \right] dX_i\, dX_j \\ &= 2 r_{ij} dX_i\, dX_j\,, \end{aligned} \tag{4.4.9}$$

where

$$\begin{aligned} r_{ij} &= \frac{1}{2} \left[\frac{\partial x_k}{\partial X_i} \frac{\partial x_k}{\partial X_j} - \delta_{ij} \right] \\ &= \frac{1}{2} [c_{ij} - \delta_{ij}]\,, \end{aligned} \tag{4.4.10}$$

in which we have a symmetric tensor

$$\begin{aligned} c_{ij} &= \frac{\partial x_k}{\partial X_i} \frac{\partial x_k}{\partial X_j} \\ \text{i.e.}\,,\; C &= \mathbf{F}^T \cdot \mathbf{F}\,, \end{aligned} \tag{4.4.11}$$

known as the *Green's deformation tensor*. From this, we immediately define the Lagrangian finite strain tensor r_{ij} as

$$\begin{aligned} 2r_{ij} &= c_{ij} - \delta_{ij} \\ \Rightarrow 2R &= C - I \,. \end{aligned}$$

Therefore we can write,

$$\frac{dl^2 - dL^2}{dL^2} = 2r_{ij}\frac{dX_i}{dL}\frac{dX_j}{dL} = 2r_{ij}N_iN_j \,. \tag{4.4.12}$$

The deformation of a body is completely described by the displacement vector. It is possible to express Lagrangian finite strain tensor r_{ij} in terms of the displacement. If u_i be the displacement of a material point from its position P_0 to P, then

$$u_i = x_i - X_i \,. \tag{4.4.13}$$

If $u_i + du_i$ be the displacement of the material point from its position Q_0 to Q, then

$$\begin{aligned} u_i + du_i &= (x_i + dx_i) - (X_i + dX_i) \\ \Rightarrow (x_i - X_i) + du_i &= (x_i - X_i) + (dx_i - dX_i) \\ \Rightarrow du_i &= dx_i - dX_i \\ \Rightarrow dx_i &= du_i + dX_i \\ \Rightarrow dx_k &= du_k + dX_k \end{aligned} \tag{4.4.14}$$

Therefore,

$$\begin{aligned} \frac{\partial x_k}{\partial X_i} &= \frac{\partial u_k}{\partial X_i} + \frac{\partial X_k}{\partial X_i} \\ \Rightarrow x_{k,i} &= u_{k,i} + \delta_{ki} \end{aligned} \tag{4.4.15}$$

and

$$\begin{aligned} \frac{\partial x_k}{\partial X_j} &= \frac{\partial u_k}{\partial X_j} + \frac{\partial X_k}{\partial X_j} \\ \Rightarrow x_{k,j} &= u_{k,j} + \delta_{kj} \,. \end{aligned} \tag{4.4.16}$$

Therefore form Eq.(4.4.10), we have

$$\begin{aligned} r_{ij} &= \frac{1}{2}\left[\frac{\partial x_k}{\partial X_i}\frac{\partial x_k}{\partial X_j} - \delta_{ij}\right] \\ &= \frac{1}{2}\left[\left(\frac{\partial u_k}{\partial X_i} + \delta_{ik}\right)\cdot\left(\frac{\partial u_k}{\partial X_j} + \delta_{jk}\right) - \delta_{ij}\right] \\ &= \frac{1}{2}\left[\frac{\partial u_k}{\partial X_i}\frac{\partial u_k}{\partial X_j} + \frac{\partial u_k}{\partial X_j}\delta_{ik} + \frac{\partial u_k}{\partial X_i}\delta_{jk} + \delta_{ik}\delta_{jk} - \delta_{ij}\right] \\ &= \frac{1}{2}\left[\frac{\partial u_i}{\partial X_j} + \frac{\partial u_j}{\partial X_i} + \frac{\partial u_k}{\partial X_i}\frac{\partial u_k}{\partial X_j}\right], \quad (\text{as } \delta_{ik}\delta_{jk} = \delta_{ij}) \\ &= \frac{1}{2}[u_{i,j} + u_{j,i} + u_{k,i}\cdot u_{k,j}] \end{aligned} \tag{4.4.17}$$

The quantities r_{ij}'s are called *Lagrangian strains* expressed in terms of displacement components.

For infinitesimal deformation, u_i's are small. Therefore, the second order terms can be neglected giving

$$u_{k,i} \cdot u_{k,j} \approx 0 \,.$$

Therefore,

$$r_{ij} \approx \frac{1}{2}[u_{i,j} + u_{j,i}] = e_{ij} \text{ (say)} \tag{4.4.18}$$

4.5 Change in the angle between two line elements in material method

Here we consider change in angle between two material line elements P_0Q_0 and P_0R_0 at P_0 inclined at an angle Θ, where P_0Q_0 is of length dL oriented in the direction (N_1, N_2, N_3) and P_0R_0 is of length δL oriented in the direction (M_1, M_2, M_3) in the region B. If Q_0 has coordinates $(X_i + dX_i)$ and R_0 has coordinates $X_i + \delta X_i$, then

$$M_i = \frac{\delta X_i}{\delta L}, \quad N_i = \frac{dX_i}{dL} \tag{4.5.1}$$

and

$$\cos\Theta = \frac{dX_i}{dL}\frac{\delta X_i}{\delta L} = N_i M_i \,. \tag{4.5.2}$$

when the body undergoes deformation, the two line elements P_0Q_0 and P_0R_0 at P_0 will defrom into two other line elements PQ and PR at P of length dl and δl, oriented in the direction (n_1, n_2, n_3) and (m_1, m_2, m_3) and inclined at an angle θ is the region B_0. If Q has coordinates $(x_i + dx_i)$ and R has coordinates $(x_i + \delta x_i)$, then

$$m_i = \frac{\delta x_i}{\delta l}, \; n_i = \frac{dx_i}{dl} \text{ and } cos\theta = \frac{dx_i}{dl}\frac{\delta x_i}{\delta l} = n_i m_i \,. \tag{4.5.3}$$

Also, we have

$$dx_k = \frac{\partial x_k}{\partial X_j} dX_j, \quad \delta x_k = \frac{\partial x_k}{\partial X_j}\delta X_j \,. \tag{4.5.4}$$

Therefore, we can write

$$\frac{\delta l^2 - \delta L^2}{\delta L^2} = 2r_{ij}\frac{\delta X_i}{\delta L}\frac{\delta X_j}{\delta L} = 2r_{ij}M_iM_j \,, \tag{4.5.5}$$

and

$$\frac{dl^2 - dL^2}{dL^2} = 2r_{ij}\frac{dX_i}{dL}\frac{dX_j}{dL} = 2r_{ij}N_iN_j \,. \tag{4.5.6}$$

Again

$$\begin{aligned} dx_i\delta x_i - dX_i\delta X_i &= dx_k\delta x_k - dX_i\delta X_i \\ &= \frac{\partial x_k}{\partial X_i}dX_i\frac{\partial x_k}{\partial X_j}\delta X_j - \delta_{ij}dX_i\delta X_j \\ &= 2r_{ij}dX_i\delta X_j \,, \end{aligned} \tag{4.5.7}$$

where r_{ij} is given by (4.4.17).

Hence

$$\begin{aligned}
&\frac{dx_i}{dL}\frac{\delta x_i}{\delta L} - \frac{dX_i}{dL}\frac{\delta X_i}{\delta L} = 2r_{ij}\frac{dX_i}{dL}\frac{\delta X_j}{\delta L} \\
\Rightarrow\ &\frac{dx_i}{dl}\frac{\delta x_i}{\delta l}\frac{dl}{dL}\frac{\delta l}{\delta L} - \frac{dX_i}{dL}\frac{\delta X_i}{\delta L} = 2r_{ij}\frac{dX_i}{dL}\frac{\delta X_j}{\delta L} \\
\Rightarrow\ &\frac{dl}{dL}\frac{\delta l}{\delta L}\cos\theta - \cos\Theta = 2r_{ij}N_iM_j\,. \qquad (4.5.8)
\end{aligned}$$

Now equation (4.4.12), (4.5.5) and (4.5.6) show that if $r_{ij} = 0$, then $dl = dL,\ \delta l = \delta L,\ \theta = \Theta$. Thus when $r_{ij} = 0$, length of a line element and angle between two line elements remain unchanged during deformation and the body has undergone only rigid body deformation. Therefore, the necessary and sufficient condition for rigid body deformation at each point is $r_{ij} = 0$.

Note: From Eq.(4.4.12) we observe that $2r_{ij}N_iN_j$ is a scalar. But the product N_iN_j of two vector components is known to be a tensor of order two. Therefore by quotient law of tensor r_{ij} is a second order tensor known as Lagrangian finite strain tensor. Further,

$$\begin{aligned}
r_{ij} &= \frac{1}{2}\left[\frac{\partial u_j}{\partial X_i} + \frac{\partial u_i}{\partial X_j} + \frac{\partial u_k}{\partial X_j}\frac{\partial u_k}{\partial X_i}\right] \\
&= \frac{1}{2}\left[\frac{\partial u_i}{\partial X_j} + \frac{\partial u_j}{\partial X_i} + \frac{\partial u_k}{\partial X_i}\frac{\partial u_k}{\partial X_j}\right] \\
&= r_{ji} \qquad (4.5.9)
\end{aligned}$$

so that r_{ij} is symmetric.

Example: Given that displacement field $x_1 = X_1 + 2X_3,\ x_2 = X_2 - 2X_3,\ x_3 = X_3 - 2X_1 + 2X_2$. Determine the deformation gradient. Green's deformation tensor and Lagrangian finite strain tensor.

Solution: The deformation gradient F has the matrix form

$$(F_{ki}) = \begin{bmatrix} \frac{\partial x_1}{\partial X_1} & \frac{\partial x_1}{\partial X_2} & \frac{\partial x_1}{\partial X_3} \\ \frac{\partial x_2}{\partial X_1} & \frac{\partial x_2}{\partial X_2} & \frac{\partial x_2}{\partial X_3} \\ \frac{\partial x_3}{\partial X_1} & \frac{\partial x_3}{\partial X_2} & \frac{\partial x_3}{\partial X_3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & -2 \\ -2 & 2 & 1 \end{bmatrix} \qquad (4.5.10)$$

The Green's deformation Tensor $C = F^T \cdot F$ has the matrix

$$(C_{ij}) = \begin{bmatrix} 1 & 0 & -2 \\ 0 & 1 & 2 \\ 2 & -2 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & -2 \\ -2 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 5 & -4 & 0 \\ -4 & 5 & 0 \\ 0 & 0 & 9 \end{bmatrix}$$

The displacement components $u_i = x_i - X_i$ of a material point are given by

$$\begin{aligned}
u_1 &= x_1 - X_1 = X_1 + 2X_3 - X_1 = 2X_3 \\
u_2 &= x_2 - X_2 = X_2 - 2X_3 - X_2 = -2X_3 \\
u_3 &= x_3 - X_3 = X_3 - 2X_1 + 2X_2 - X_3 = -2X_1 + 2X_2
\end{aligned}$$

Now the second order Lagrangian finite strain tensor r_{ij} are given by

$$\begin{aligned} r_{11} &= \frac{1}{2}\left[\frac{\partial u_1}{\partial X_1}+\frac{\partial u_1}{\partial X_1}+\frac{\partial u_k}{\partial X_1}\frac{\partial u_k}{\partial X_1}\right] \\ &= \frac{\partial u_1}{\partial X_1}+\frac{1}{2}\left[\left(\frac{\partial u_1}{\partial X_1}\right)^2+\left(\frac{\partial u_2}{\partial X_1}\right)^2+\left(\frac{\partial u_3}{\partial X_1}\right)^2\right] \\ &= 0+\frac{1}{2}\left[0^2+0^2+(-2)^2\right]=2 \end{aligned}$$

$$\begin{aligned} r_{22} &= \frac{1}{2}\left[\frac{\partial u_2}{\partial X_2}+\frac{\partial u_2}{\partial X_2}+\frac{\partial u_k}{\partial X_2}\frac{\partial u_k}{\partial X_2}\right] \\ &= \frac{\partial u_2}{\partial X_2}+\frac{1}{2}\left[\left(\frac{\partial u_1}{\partial X_2}\right)^2+\left(\frac{\partial u_2}{\partial X_2}\right)^2+\left(\frac{\partial u_3}{\partial X_2}\right)^2\right] \\ &= 0+\frac{1}{2}\left[0^2+0^2+2^2\right]=2 \end{aligned}$$

$$\begin{aligned} r_{33} &= \frac{1}{2}\left[\frac{\partial u_3}{\partial X_3}+\frac{\partial u_3}{\partial X_3}+\frac{\partial u_k}{\partial X_3}\frac{\partial u_k}{\partial X_3}\right] \\ &= \frac{\partial u_1}{\partial X_1}+\frac{1}{2}\left[\left(\frac{\partial u_1}{\partial X_1}\right)^2+\left(\frac{\partial u_2}{\partial X_1}\right)^2+\left(\frac{\partial u_3}{\partial X_1}\right)^2\right] \\ &= 0+\frac{1}{2}\left[0^2+0^2+(-2)^2\right]=2 \end{aligned}$$

$$\begin{aligned} r_{12} &= \frac{1}{2}\left[\frac{\partial u_1}{\partial X_2}+\frac{\partial u_3}{\partial X_3}+\frac{\partial u_k}{\partial X_3}\frac{\partial u_k}{\partial X_3}\right] \\ &= \frac{1}{2}\left[\frac{\partial u_1}{\partial X_2}+\frac{\partial u_2}{\partial X_1}+\frac{\partial u_1}{\partial X_1}\frac{\partial u_1}{\partial X_2}+\frac{\partial u_2}{\partial X_1}\frac{\partial u_2}{\partial X_2}+\frac{\partial u_3}{\partial X_1}\frac{\partial u_3}{\partial X_2}\right] \\ &= \frac{1}{2}\left[0+0+0+0+(-2)2\right]=-4=r_{21} \end{aligned}$$

$$\begin{aligned} r_{13} &= \frac{1}{2}\left[\frac{\partial u_1}{\partial X_3}+\frac{\partial u_3}{\partial X_1}+\frac{\partial u_k}{\partial X_1}\frac{\partial u_k}{\partial X_3}\right] \\ &= \frac{1}{2}\left[\frac{\partial u_1}{\partial X_3}+\frac{\partial u_3}{\partial X_1}+\frac{\partial u_1}{\partial X_1}\frac{\partial u_1}{\partial X_3}+\frac{\partial u_2}{\partial X_1}\frac{\partial u_2}{\partial X_3}+\frac{\partial u_3}{\partial X_1}\frac{\partial u_3}{\partial X_3}\right] \\ &= \frac{1}{2}\left[2+(-2)+0\cdot 2+0\cdot(-2)+(-2)\cdot 0\right]=0=r_{31} \end{aligned}$$

$$\begin{aligned} r_{23} &= \frac{1}{2}\left[\frac{\partial u_2}{\partial X_3}+\frac{\partial u_3}{\partial X_2}+\frac{\partial u_k}{\partial X_2}\frac{\partial u_k}{\partial X_3}\right] \\ &= \frac{1}{2}\left[\frac{\partial u_2}{\partial X_3}+\frac{\partial u_3}{\partial X_2}+\frac{\partial u_1}{\partial X_2}\frac{\partial u_1}{\partial X_3}+\frac{\partial u_2}{\partial X_2}\frac{\partial u_2}{\partial X_3}+\frac{\partial u_3}{\partial X_2}\frac{\partial u_3}{\partial X_3}\right] \\ &= \frac{1}{2}\Big[(-2)+2+0\cdot 2+0\cdot(-2)+2\cdot 0\Big] = 0 = r_{32} \end{aligned}$$

In matrix notation, the second order Lagrangian finite strain tensor r_{ij} are given by

$$(r_{ij}) = \begin{bmatrix} 2 & -2 & 0 \\ -2 & 2 & 0 \\ 0 & 0 & 4 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 5 & -4 & 0 \\ -4 & 5 & 0 \\ 0 & 0 & 9 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \frac{1}{2}C - \frac{1}{2}I$$

Unit 5

Course Structure

- Eulerian finite strain tensor
- Infinitesimal strain component
- Infinitesimal rotation tensor

5.1 Eulerian finite strain tensor

In the spatial method of description of deformation (x_1, x_2, x_3) are regard as independent variables and the equation characterizing the defromation can be written as

$$X_k = X_k(x_1, x_2, x_3, t);\ k = 1, 2, 3 \tag{5.1.1}$$

where X_1, X_2, X_3 represents the material coordinates of a material particle. SInce $X_k + dX_k$ are co-ordinate of Q_0 at the same time t, Therefore

$$dX_k = \frac{\partial x_k}{\partial x_i} dx_i = \frac{\partial x_k}{\partial x_j} dx_j = X_{k,j} dx_j \tag{5.1.2}$$

The quantity $x_{k,,j} = \frac{\partial X_k}{\partial x_j}$ is called deformation gradient tensor or simply the deformation gradient and is dented by F^{-1}_{kj}. Then

$$dX = \vec{F}\, dx \tag{5.1.3}$$

Also we have

$$dL^2 = dx_k dx_k = \frac{\partial X_k}{\partial x_i} \frac{\partial X_k}{\partial x_j} dx_i dx_j \tag{5.1.4}$$

and

$$dl^2 = dx \cdot dx = dx_i \cdot dx_i = \delta_{ij} dx_i dx_j \tag{5.1.5}$$

Therefore, a measure of change of length of a line element

$$\begin{aligned} &= dl^2 - dL^2 \\ &= \delta_{ij} dx_i dx_j - \frac{\partial x_k}{\partial x_i}\frac{\partial x_k}{\partial x_j} dx_i dx_j \\ &= \left[\delta_{ij} - \frac{\partial x_k}{\partial x_i}\frac{\partial x_k}{\partial x_j}\right] dx_i dx_j \\ &= 2\eta_{ij} dx_i dx_j, \end{aligned} \tag{5.1.6}$$

where

$$\begin{aligned} \eta_{ij} &= \frac{1}{2}\left[\delta_{ij} - \frac{\partial x_k}{\partial x_i}\frac{\partial x_k}{\partial x_j}\right] \\ &= \frac{1}{2}\left[\delta_{ij} - c_{ij}\right] \end{aligned} \tag{5.1.7}$$

in which we have a symmetric tensor

$$c_{ij} = \frac{\partial x_k}{\partial x_i}\frac{\partial x_k}{\partial x_j}$$

that is,

$$c = (\vec{F})^T F^{-1}$$

which is known as the Cauchy's deformation tensor. From this, we immediately define the Eulerian finite strain tensor η_{ij} as

$$2\eta_{ij} = \delta_{ij} - c_{ij}$$

or,

$$2n = I - C$$

Now, we can write

$$\frac{dl^2 - dL^2}{dl^2} = 2n_{ij}\frac{dx_i}{dl}\frac{dx_j}{dl} = 2\eta_{ij} n_i n_j \tag{5.1.8}$$

The Eulerian finite strain tensor expressed by Eq. (5.1.7) is given in terms of the appropriate deformation gradients. These same tensors may also be developed in terms of displacement gradients.

In Component notation, the material description is

$$u_i = x_i - X_i$$

The deformation of a body is completely described by the displacement vector. It is possible to express η_{ij} in terms of the displacement u_i of a spatial point from its position from P_0 to P, then

$$u_i = x_i - X_i, \quad i.e., \quad X_k = x_k - u_k$$

If $u_i + du_i$ be the displacement of the spatial point from its position Q_0 to Q, then

$$\begin{aligned} u_i + du_i &= (x_i + dx_i) - (X_i + dX_i) \\ \text{or, } (x_i - X_i) + du_i &= (x_i - X_i) + (dx_i - dX_i) \\ \text{or, } du_i &= dx_i - dX_i \\ \text{i.e, } dX_k &= dx_k - du_k \end{aligned}$$

Differentiating with respect to x_i, we get

$$\begin{aligned} X_{k,i} &= \frac{\partial X_k}{\partial x_i} \\ &= \frac{\partial x_k}{\partial x_i} - \frac{\partial u_k}{\partial x_i} \\ &= \delta_{ki} - \frac{\partial u_k}{\partial x_i} \\ &= \delta_{ki} - u_{k,i} \end{aligned}$$

Similarly, differentiating with respect to x_j, we get

$$\begin{aligned} X_{k,j} &= \frac{\partial x_k}{\partial x_j} - \frac{\partial u_k}{\partial x_j} \\ &= \delta_{kj} - u_{k,j} \end{aligned}$$

Then from (5.1.7), the expression for η_{ij} in terms of the displacement u_i of a material point from its position P_0 to P is given by

$$\begin{aligned} n_{ij} &= \frac{1}{2}\left[\delta_{ij} - \frac{\partial X_k}{\partial x_i}\frac{\partial X_k}{\partial x_j}\right] \\ &= \frac{1}{2}[\delta_{ij} - X_{k,i}X_{k,j}] \\ &= \frac{1}{2}[\delta_{ij} - (\delta ki - u_{k,i})(\delta_{kj} - u_{k,j})] \\ &= \frac{1}{2}[\delta_{ij} - (\delta ki - \frac{\partial u_k}{\partial x_i})(\delta_{kj} - \frac{\partial u_k}{\partial x_j})] \\ &= \frac{1}{2}\left[\delta_{ij} - \delta_{ki}\delta_{kj} + \frac{\partial u_k}{\partial x_i}\delta_{kj} + \frac{\partial u_k}{\partial x_j}\delta_{ki} - \frac{\partial u_k}{\partial x_i}\frac{\partial u_k}{\partial x_j}\right] \\ &= \frac{1}{2}\left[\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} - \frac{\partial u_k}{\partial x_i}\frac{\partial u_k}{\partial x_j}\right] \; as\; \delta_{ki}\delta_{kj} = \delta ij \\ &= \frac{1}{2}[u_{i,j} + u_{j,i} - u_{k,i}u_{k,j}] \end{aligned} \tag{5.1.9}$$

5.2 Change in the angle between two line elements in spatial method

Here we consider change in angle between two material line elements P_0Q_0 and P_0R_0 at P_0 inclined at an angle Θ, where P_0Q_0 and is of length dL oriented in the direction (N_1, N_2, N_3) and P_0R_0 is of

length δL oriented in the direction (M_1, M_2, M_3) in the region B_0. If Q_0 has coordinates $(X_i + dX_i)$ and R_0 has coordinates $(X_i + \delta X_i)$, then

$$M_i = \frac{\delta X_i}{\delta L},\ N_i = \frac{dX_i}{dL}\ and\ \cos\Theta = \frac{dx_i}{dL}\frac{\delta X_i}{\delta L} = N_i M_i \tag{5.2.1}$$

when the body undergoes deformation the two line elements P_0Q_0 and P_0R_0 at P_0 will deform into two other line elements PQ and PR at P of length dl and δl, oriented in the direction (n_1, n_2, n_3) and (m_1, m_2, m_3) and inclined at an angle θ in the region B_0.

If Q has co-ordinantes $(x_i + dx_i)$ and R has coordinates $(x_i + \delta x_i)$, then

$$m_i = \frac{\delta X_i}{\delta L},\ n_i = \frac{dX_i}{dL},\ \cos\theta = \frac{dx_i}{dl}\frac{\delta x_i}{\delta l} = n_i m_i \tag{5.2.2}$$

Also we have

$$dX_k = \frac{\partial X_k}{\partial X_i}dx_i, \quad \text{i.e.,} \quad \delta_{X_k} = \frac{\partial X_k}{\partial x_i}\delta x_i.$$

Therefore,

$$\frac{\delta l^2 - \delta L^2}{\delta L^2} = 2\eta_{ij}\frac{\delta x_i}{\delta L}\frac{\delta X_j}{\delta L} = 2\eta_{ij}M_iM_j$$
$$\frac{dl^2 - dL^2}{dL^2} = 2\eta_{ij}\frac{dx_i}{dL}\frac{dX_j}{dL} = 2\eta_{ij}N_iN_j$$

Again

$$\begin{aligned} dx_i\delta x_i - dX_i\delta X_i &= \delta_{ij}dx_i\delta x_j - dx_k\delta x_k \\ &= \delta_{ij}dx_i\delta x_j - \frac{\delta X_k}{\delta X_i}\delta x_i\frac{\delta X_k}{\delta X_j}\delta X_j \\ &= 2\eta_{ij}dx_i\delta_j \quad \text{where} \quad \eta_{ij} = \frac{1}{2}\left[\delta_{ij} - \frac{\partial X_k}{\partial X_i}\frac{\partial X_k}{\partial X_j}\right] \end{aligned}$$

Therefore,

$$\begin{aligned} &\frac{dx_i}{dl}\frac{\delta x_i}{\delta l} - \frac{dX_i}{dl}\frac{\delta X_i}{\delta l} = 2\eta_{ij}\frac{dx_i}{dl}\frac{\delta x_i}{\delta l} \\ \Rightarrow\ &\frac{dx_i}{dl}\frac{\delta x_i}{\delta l} - \frac{dX_i}{dL}\frac{dL}{dl}\frac{\delta X_i}{\delta L}\frac{\delta L}{\delta l} = 2\eta_{ij}n_im_j \\ \Rightarrow\ &\cos\theta - \cos\Theta\frac{dL}{dl}\frac{\delta L}{\delta l} = 2\eta_{ij}n_im_j \end{aligned} \tag{5.2.3}$$

We observed that if $\eta_{ij} = 0$, then $dL = dl$, $\delta L = \delta l$, $\theta = \Theta$. Thus, when $\eta_{ij} = 0$, length of a line element and angle between two line elements remain unchanged during deformation and the body undergone only rigid body deformation.

- Thus the necessary and sufficient condition for rigid body deformation at each point is $\eta_{ij} = 0$.

- The knowledge of η_{ij} at a point enables us to determine the change in length of a line element and change in angle between two line elements. Therefore η_{ij} can be taken as the measure of strain deformation.
- It can be easily shown that η_{ij} is a symmetric tensor of order two. It is known as Eulerian finite strain tensor.

Example: Given the displacement field $x_1 = X_1 + 2X_3$, $x_2 = X_2 - 2X_3$, $x_3 = X\_2X_1 + 2X_2$. Determine the deformation gradient, Cauchy's deformation tensors and Eulerian finite strain tensor.

Solution: We have

$$\begin{aligned} x_1 &= X_1 + 2X_3, && (5.2.4)\\ x_2 &= X_2 - 2X_3, && (5.2.5)\\ x_3 &= X_2X_1 + 2X_2 && (5.2.6) \end{aligned}$$

$$\text{From (5.2.4) and (5.2.5)}, \; x_1 + x_2 = X_1 + X_2, \qquad (5.2.7)$$

$$\text{From (5.2.5) and (5.2.6)}, \; x_2 + 2x_3 = -4X_1 + 5X_2, \qquad (5.2.8)$$

From (5.2.7) and (5.2.8)

$$4(x_1 + x_2) + (x_2 + 2x_3) = 4X_2 + 5X_2 \quad \Rightarrow \quad X_2 = \frac{1}{9}[4x_1 + 5x_2 + 2x_3] \qquad (5.2.9)$$

Therefore,

$$\begin{aligned} X_1 &= x_1 + x_2 - X_2 \quad \text{From (5.2.7)}\\ \Rightarrow X_1 &= x_1 + x_2 - \frac{1}{9}[4x_1 + 5x_2 + 2x_3]\\ \Rightarrow X_1 &= \frac{1}{9}[5x_1 + 4x_2 - 2x_3] \qquad (5.2.10) \end{aligned}$$

and

$$\begin{aligned} X_3 &= x_3 + 2x_1 - 2x_2\\ \Rightarrow X_3 &= x_3 + \frac{2}{9}[5x_1 + 4x_2 - 2x_3] - \frac{2}{9}[4x_1 + 5x_2 + 2x_3]\\ \Rightarrow X_3 &= \frac{1}{9}[2x_1 - 2x_2 + x_3] \qquad (5.2.11) \end{aligned}$$

The deformation gradient $\vec{F}$ has the matrix form

$$(\vec{F_{ki}}) = \begin{bmatrix} \frac{\partial x_1}{\partial X_1} & \frac{\partial x_1}{\partial X_2} & \frac{\partial x_1}{\partial X_3} \\ \frac{\partial x_2}{\partial X_1} & \frac{\partial x_2}{\partial X_2} & \frac{\partial x_2}{\partial X_3} \\ \frac{\partial x_3}{\partial X_1} & \frac{\partial x_3}{\partial X_2} & \frac{\partial x_3}{\partial X_3} \end{bmatrix} = \frac{1}{9}\begin{bmatrix} 5 & 4 & -2 \\ 4 & 5 & 2 \\ 2 & -2 & 1 \end{bmatrix}$$

The Cauchy's deformation tensor, $C = (\vec{F}^T \cdot \vec{F})$ has the matrix form

$$(\vec{C_{ij}}) = \frac{1}{9}\begin{bmatrix}5 & 4 & -2\\ 4 & 5 & 2\\ 2 & -2 & 1\end{bmatrix}^T \cdot \frac{1}{9}\begin{bmatrix}5 & 4 & -2\\ 4 & 5 & 2\\ 2 & -2 & 1\end{bmatrix} = \frac{1}{81}\begin{bmatrix}45 & 36 & 0\\ 36 & 45 & 0\\ 0 & 0 & 9\end{bmatrix}$$

The displacement components $u_i = x_i - X_i$ of a material point are given by

$$\begin{aligned}
u_1 &= x_1 - X_1 = x_1 - \frac{1}{9}(5x_1 + 4x_2 - 2x_3) = \frac{1}{9}(4x_1 - 4x_2 + 2x_3),\\
u_2 &= x_2 - X_2 = x_2 - \frac{1}{9}(4x_1 + 5x_2 + 2x_3) = \frac{1}{9}(-4x_+4x_2 - 2x_3),\\
u_3 &= x_3 - X_3 = x_3 - \frac{1}{9}(2x_1 - 2x_2 + x_3) = \frac{1}{9}(-2x_1 + 2x_2 + 8x_3).
\end{aligned}$$

Now, the Eulerian finite strain tensor are given by

$$\begin{aligned}
\eta_{11} &= \frac{1}{2}\left[\frac{\partial u_1}{\partial x_1} + \frac{\partial u_1}{\partial x_1} - \frac{\partial u_k}{\partial x_1}\frac{\partial u_k}{\partial x_1}\right] = \frac{\partial u_1}{\partial x_1} - \frac{1}{2}\left[\left(\frac{\partial u_1}{\partial x_1}\right)^2 + \left(\frac{\partial u_2}{\partial x_1}\right)^2 + \left(\frac{\partial u_3}{\partial x_1}\right)^2\right]\\
&= \frac{4}{9} - \frac{1}{2}\left[\left(\frac{4}{9}\right)^2 + \left(-\frac{4}{9}\right)^2 + \left(-\frac{2}{9}\right)^2\right] = \frac{2}{9}
\end{aligned}$$

$$\begin{aligned}
\eta_{22} &= \frac{1}{2}\left[\frac{\partial u_2}{\partial x_2} + \frac{\partial u_2}{\partial x_2} - \frac{\partial u_k}{\partial x_2}\frac{\partial u_k}{\partial x_2}\right] = \frac{\partial u_2}{\partial x_2} - \frac{1}{2}\left[\left(\frac{\partial u_1}{\partial x_2}\right)^2 + \left(\frac{\partial u_2}{\partial x_2}\right)^2 + \left(\frac{\partial u_3}{\partial x_2}\right)^2\right]\\
&= \frac{4}{9} - \frac{1}{2}\left[\left(-\frac{4}{9}\right)^2 + \left(\frac{4}{9}\right)^2 + \left(\frac{2}{9}\right)^2\right] = \frac{2}{9}
\end{aligned}$$

$$\begin{aligned}
\eta_{33} &= \frac{1}{2}\left[\frac{\partial u_3}{\partial x_3} + \frac{\partial u_3}{\partial x_3} - \frac{\partial u_k}{\partial x_3}\frac{\partial u_k}{\partial x_3}\right] = \frac{\partial u_3}{\partial x_3} - \frac{1}{2}\left[\left(\frac{\partial u_1}{\partial x_3}\right)^2 + \left(\frac{\partial u_2}{\partial x_3}\right)^2 + \left(\frac{\partial u_3}{\partial x_3}\right)^2\right]\\
&= \frac{8}{9} - \frac{1}{2}\left[\left(\frac{2}{9}\right)^2 + \left(-\frac{2}{9}\right)^2 + \left(\frac{8}{9}\right)^2\right] = \frac{4}{9}
\end{aligned}$$

$$\begin{aligned}
\eta_{12} &= \frac{1}{2}\left[\frac{\partial u_1}{\partial x_2} + \frac{\partial u_2}{\partial x_1} - \frac{\partial u_k}{\partial x_1}\frac{\partial u_k}{\partial x_2}\right] = \frac{1}{2}\left[\frac{\partial u_1}{\partial x_2} + \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_1}\frac{\partial u_1}{\partial x_2} - \frac{\partial u_2}{\partial x_1}\frac{\partial u_2}{\partial x_2} - \frac{\partial u_3}{\partial x_1}\frac{\partial u_3}{\partial x_2}\right]\\
&= \frac{1}{2}\left[\left(-\frac{4}{9}\right) + \left(-\frac{4}{9}\right) - \left(\frac{4}{9}\right)\left(-\frac{4}{9}\right) - \left(\frac{4}{9}\right)\left(-\frac{4}{9}\right) - \left(\frac{2}{9}\right)\left(-\frac{2}{9}\right)\right] = -\frac{2}{9} = \eta_{21}
\end{aligned}$$

$$\begin{aligned}
\eta_{13} &= \frac{1}{2}\left[\frac{\partial u_1}{\partial x_3} + \frac{\partial u_3}{\partial x_1} - \frac{\partial u_k}{\partial x_1}\frac{\partial u_k}{\partial x_3}\right] = \frac{1}{2}\left[\frac{\partial u_1}{\partial x_3} + \frac{\partial u_3}{\partial x_1} - \frac{\partial u_1}{\partial x_1}\frac{\partial u_1}{\partial x_3} - \frac{\partial u_2}{\partial x_1}\frac{\partial u_2}{\partial x_3} - \frac{\partial u_3}{\partial x_1}\frac{\partial u_3}{\partial x_3}\right]\\
&= \frac{1}{2}\left[\left(\frac{2}{9}\right) + \left(-\frac{2}{9}\right) - \left(\frac{4}{9}\right)\left(\frac{2}{9}\right) - \left(-\frac{4}{9}\right)\left(-\frac{2}{9}\right) - \left(-\frac{2}{9}\right)\left(\frac{8}{9}\right)\right] = 0 = \eta_{31}
\end{aligned}$$

$$\eta_{23} = \frac{1}{2}\left[\frac{\partial u_2}{\partial x_3}+\frac{\partial u_3}{\partial x_2}-\frac{\partial u_k}{\partial x_1}\frac{\partial u_k}{\partial x_3}\right] = \frac{1}{2}\left[\frac{\partial u_2}{\partial x_3}+\frac{\partial u_3}{\partial x_2}-\frac{\partial u_1}{\partial x_2}\frac{\partial u_1}{\partial x_3}-\frac{\partial u_2}{\partial x_2}\frac{\partial u_2}{\partial x_3}-\frac{\partial u_3}{\partial x_2}\frac{\partial u_3}{\partial x_3}\right]$$
$$= \frac{1}{2}\left[\left(-\frac{2}{9}\right)+\left(\frac{2}{9}\right)-\left(-\frac{4}{9}\right)\left(\frac{2}{9}\right)-\left(\frac{4}{9}\right)\left(-\frac{2}{9}\right)-\left(\frac{2}{9}\right)\left(\frac{8}{9}\right)\right] = 0 = \eta_{32}$$

In matrix notation, the second order Eulerian finite strain tensors η_{ij} are given by

$$(\eta_{ij}) = \frac{1}{9}\begin{bmatrix}2 & -2 & 0\\ -2 & 2 & 0\\ 0 & 0 & 4\end{bmatrix} = \frac{1}{2}\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix} - \frac{1}{2}\frac{1}{81}\begin{bmatrix}45 & 36 & 0\\ 36 & 45 & 0\\ 0 & 0 & 1\end{bmatrix}$$

Hence, $\eta = \frac{1}{2}I - \frac{1}{2}C$.

5.3 Infinitesimal strain component

There are many important engineering problems that involve structural members or machine parts for which the deformation is very small (mathematically treated as infinitesimal). In some common materials, like metals, concrete, wood etc. undergo small changes of shape when forces of reasonable magnitude are applied to them.

If the displacement gradients are small and the squares and products of the partial derivatives of $u_i's$ are negligible then the Lagrangian finite strain tensor reduced to infinitesimal Lagrangian strain tensor denoted by $E_{ij},\ i,j = 1,2,3$.

In this case we have,

$$F_{ij}(X_1,X_2,X_3) \approx \frac{1}{2}[u_{i,j}+u_{j,i}] = \frac{1}{2}\left[\frac{\partial u_i}{\partial X_j}+\frac{\partial u_j}{\partial X_i}\right]$$

Therefore, Normal strains are given by

$$F_{11} = u_{1,1} = \frac{\partial u_1}{\partial X_1},\ \ F_{22} = u_{2,2} = \frac{\partial u_2}{\partial X_2},\ \ F_{33} = u_{3,3} = \frac{\partial u_3}{\partial X_3}$$

and shearing strains are given by

$$F_{23} = \frac{1}{2}[u_{2,3}+u_{3,2}] = \frac{1}{2}\left[\frac{\partial u_2}{\partial X_3}+\frac{\partial u_3}{\partial X_2}\right],$$
$$F_{31} = \frac{1}{2}[u_{3,1}+u_{1,3}] = \frac{1}{2}\left[\frac{\partial u_3}{\partial X_1}+\frac{\partial u_3}{\partial X_3}\right],$$
$$F_{12} = \frac{1}{2}[u_{1,2}+u_{2,1}] = \frac{1}{2}\left[\frac{\partial u_1}{\partial X_2}+\frac{\partial u_2}{\partial X_1}\right],$$

Similarly the Eulerian finite strain tensors reduced to infinitesimal Eulerian strain tensor denoted by $e_{ij}(i,j=1,2,3)$.

Therefore,

$$e_{ij}(x_1,x_2,x_3) \approx \frac{1}{2}\left[\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}\right]$$

Hence, the normal stains are given by

$$e_{11}=\frac{\partial u_1}{\partial x_1},\ e_{22}=\frac{\partial u_2}{\partial x_2},\ e_{33}=\frac{\partial u_3}{\partial x_3},$$

and the shearing strains are given by

$$e_{23}=\frac{1}{2}\left[\frac{\partial u_2}{\partial x_3}+\frac{\partial u_3}{\partial x_2}\right],\ e_{31}=\frac{1}{2}\left[\frac{\partial u_3}{\partial x_1}+\frac{\partial u_1}{\partial x_3}\right],\ e_{12}=\frac{1}{2}\left[\frac{\partial u_1}{\partial x_2}+\frac{\partial u_2}{\partial x_1}\right]$$

We now show that both the displacement components are their gradient are small then $E_{ij}(X_1,X_2,X_3)$ and $e_{ij}(x_1,x_2,x_3)$ are identical.

We have $x_i=X_i+u_i$. By Taylor's series,

$$\begin{aligned} u_i(x_1,x_2,x_3) &= u_i(X_1+u_1,X_2+u_2,X_3+u_3) \\ &= u_i(X_1,X_2,X_3)+u_i\frac{\partial u_i}{\partial X_j}+\cdots \\ &\approx u_i(X_1,X_2,X_3) \end{aligned}$$

(Neglecting the product terms $u_i\frac{\partial u_i}{\partial X_j}$ and small quantities of higher order)

Therefore,

$$\begin{aligned} \frac{\partial u_i}{\partial X_j}(X_1,X_2,X_3) &\approx \frac{\partial u_i}{\partial X_j}(x_1,x_2,x_3) \\ &= \frac{\partial u_i}{\partial x_k}(x_1,x_2,x_3)\frac{\partial x_k}{\partial X_j} \\ &= \frac{\partial u_i}{\partial x_k}(x_1,x_2,x_3)\left[\frac{\partial u_k}{\partial X_j}+\delta_{kj}\right]\ (since\ x_i=X_i+u_i) \\ &\approx \frac{\partial u_i}{\partial x_k}(x_1,x_2,x_3)\delta_{kj}\ \text{(Neglecting the product term)} \\ &= \frac{\partial u_i}{\partial x_j}(x_1,x_2,x_3) \end{aligned}$$

Therefore in Cartesian co-ordinate

$$\begin{aligned} E_{ij} &= \frac{1}{2}\left[\frac{\partial u_i}{\partial X_j}+\frac{\partial u_j}{\partial X_i}\right] \\ &= \frac{1}{2}\left[\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}\right] \\ &= e_{ij} \end{aligned}$$

We observe that in infinitesimal deformation case the distinction between Lagrangian and Eulerian strain components disappears. This is because of the fact that it is quite immaterial whether the derivatives are to be evaluated at the position of a point before or after deformation.

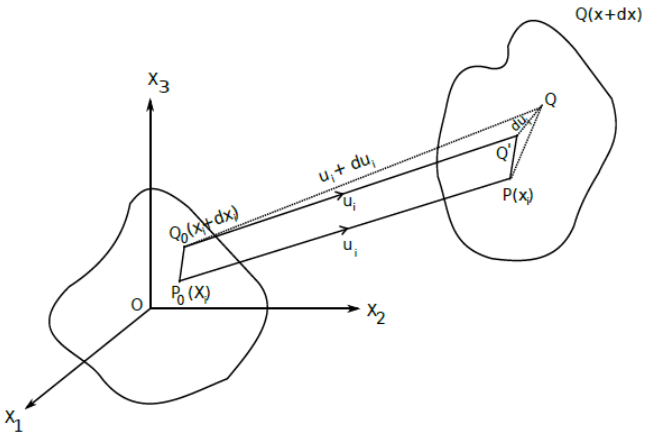
Note: In matrix notation the infinitesimal Lagrangian Strain tensor E in terms of the components of the displace gradients in rectangular cartesian coordinates is given by

$$[E] = \begin{bmatrix} \frac{\partial u_1}{\partial X_1} & \frac{1}{2}\left[\frac{\partial u_1}{\partial X_2} + \frac{\partial u_2}{\partial X_1}\right] & \frac{1}{2}\left[\frac{\partial u_1}{\partial X_3} + \frac{\partial u_3}{\partial X_1}\right] \\ \frac{1}{2}\left[\frac{\partial u_1}{\partial X_2} + \frac{\partial u_2}{\partial X_1}\right] & \frac{\partial u_2}{\partial X_2} & \frac{1}{2}\left[\frac{\partial u_2}{\partial X_3} + \frac{\partial u_3}{\partial X_2}\right] \\ \frac{1}{2}\left[\frac{\partial u_1}{\partial X_3} + \frac{\partial u_3}{\partial X_1}\right] & \frac{1}{2}\left[\frac{\partial u_2}{\partial X_3} + \frac{\partial u_3}{\partial X_3}\right] & \frac{\partial u_3}{\partial X_3} \end{bmatrix}$$

Exercise: Show that the expressions in Lagrangian and Eulerian description of deformation of a continuous medium are identical in infinitesimal theory.

5.4 Infinitesimal Rotation tensor

Consider two neighbouring material points at the positions P_0 and Q_0 of the continuum in the undeformed state with cooordinates X_i and $X_i + dX_i$ respectively. As a result of deformation, the



material point at P_0 undergoes a displacement u_i and moves to the position P and let the material point at position Q_0 experiences a displacement $u_i + du_i$ and moves to the position Q in the deformed state. If we draw Q_0Q' equal and parallel to P_0P then the relative displacement of material point originally at Q_0 with respect to the material point originally at P_0 will be represented by $\vec{Q'Q}$. Now, $\vec{P_0P} = \vec{Q_0Q} = \vec{u}$, $\vec{u}$ and $\vec{u} + d\vec{u}$ being displacement of material points at P_0 and Q_0 respectively, and $\vec{Q_0Q} = \vec{u} + d\vec{u}$. Hence,

$$\begin{aligned} &\vec{Q_0Q'} + \vec{Q'Q} = \vec{u} + d\vec{u} \\ \Rightarrow\ &\vec{u} + \vec{Q'Q} = \vec{u} + d\vec{u} \\ \Rightarrow\ &\vec{Q'Q} = d\vec{u} \end{aligned}$$

For material method of description

$$u_i = F_i(X_1, X_2, X_3) \tag{5.4.1}$$

Therefore we can write

$$u_i + du_i = F_i(X_1 + dX_1, X_2 + dX_2, X_3 + dX_3)$$

Since P_0 and Q_0 are very closed together, dx_i are small. Using Taylor's series and neglecting higher powers of dx_i, we have

$$\begin{aligned} u_i + du_i &= F_i(X_1, X_2, X_3) + \frac{\partial F_i}{\partial X_1}dX_1 + \frac{\partial F_i}{\partial X_2}dX_2 + \frac{\partial F_i}{\partial X_3}dX_3 \\ \Rightarrow u_i + du_i &= F_i(X_1, X_2, X_3) + \frac{\partial F_i}{\partial X_j}dX_j \\ \Rightarrow u_i + du_i &= u_i + \frac{\partial F_i}{\partial X_j}dX_j \\ \Rightarrow du_i &= \frac{\partial F_i}{\partial X_j}dX_j \end{aligned} \tag{5.4.2}$$

which can be expressed in the form

$$\begin{aligned} du_i &= \frac{\partial u_i}{\partial X_j}dX_j \\ &= \left[\frac{1}{2}\left(\frac{\partial u_i}{\partial X_j} + \frac{\partial u_j}{\partial X_i}\right) + \frac{1}{2}\left(\frac{\partial u_i}{\partial X_j} - \frac{\partial u_j}{\partial X_i}\right)\right]dX_j \\ &= (R_{ij} + E_{ij})dX_j \\ &= R_{ij}dX_j + E_{ij}dX_j \\ &= du_i^{(1)} + du_i^{(2)} \end{aligned} \tag{5.4.3}$$

where

$$\begin{aligned} E_{ij} &= \frac{1}{2}\left[\frac{\partial u_i}{\partial X_j} + \frac{\partial u_j}{\partial X_i}\right] \\ &= \text{Symmetric small strain tensor of order 2} \\ &= E_{ji} \end{aligned}$$

and

$$\begin{aligned} R_{ij} &= \frac{1}{2}\left[\frac{\partial u_i}{\partial X_j} - \frac{\partial u_j}{\partial X_i}\right] \\ &= \text{Skew-symmetric tensor of order 2} \\ &= -R_{ji} \end{aligned}$$

Thus, the relative displacement du_i consists of two parts

$$du_i^{(1)} = R_{ij}dX_j \text{ and } du_i^{(2)} = E_{ij}dX_j$$

In order to study $du_i^{(1)}$ we form a vector R_i by setting $R_i = e_{ijk}R_{kj}$. Therefore

$$\begin{aligned} e_{ijk}R_i &= e_{ijk}e_{ipq}R_{qp} \\ &= (\delta_{jp}\delta_{kq} - \delta_{jq}\delta_{kp})R_{qp} \\ &= R_{kj} - R_{jk} = 2R_{kj} \text{ (since, } R_{jk} = -R_{kj}) \end{aligned}$$

where e_{ijk} is alternating symbol (or Levi-Civita symbol) defined by

$$\begin{aligned} e_{ijk} &= 0, \quad \text{if any two of } i,\ j,\ k \text{ are equal} \\ &= 1, \quad \text{if } i,\ j,\ k \text{ are even permutation of } 1,\ 2,\ 3 \\ &= -1, \quad \text{if } i,\ j,\ k \text{ if } i,\ j,\ k \text{ are odd permutation of } 1,\ 2,\ 3 \end{aligned}$$

Therefore, $R_{kj} = \frac{1}{2}e_{ijk}R_i$ and $du_k^{(1)}$ becomes

$$du_k^{(1)} = R_{kj}dX_j = \frac{1}{2}e_{ijk}R_i dX_j$$

In general, $du_k^{(1)} = \frac{1}{2}\vec{R} \times d\vec{X}$, where $d\vec{X}$ is the vector connecting the position P_0 and Q_0 of the continuum.

Now,

$$\begin{aligned} R_i &= e_{ijk}R_{kj} \\ &= \frac{1}{2}e_{ijk}\left[\frac{\partial u_k}{\partial X_j} - \frac{\partial u_j}{\partial X_k}\right] \\ &= \frac{1}{2}(e_{ijk}u_{k,j} - e_{ijk}u_{j,k}) \\ &= \frac{1}{2}(e_{ijk}u_{k,j} - e_{ikj}u_{k,j}) \text{ (In second term interchanging the dummy indices } j \text{ and } k) \\ &= \frac{1}{2}(e_{ijk}u_{k,j} + e_{ijk}u_{k,j}) \end{aligned}$$

Therefore, $R_i = e_{ijk}u_{k,j} = \text{rot}(\vec{u})_i$. Hence $\vec{R} = \text{rot}\vec{u}$. Therefore, the part $du_i^{(1)} = R_{ij}dX_j$ represents a relative displacement involving small rigid body rotation of the neighbouring element of P_0 through and angle $\frac{1}{2}\vec{R} = \frac{1}{2}(rot\vec{u})$. The $\vec{R} = \text{rot}\vec{u}$ is called small rotation vector and R_{ij} are called small rotation vector and R_{ij} are called small rotation tensor. Now pure rotation does not bring about any strain deformation in the body. The part $du_i^{(2)} = E_{ij}dX_j$ represents a relative displacement involving a strain deformation causing a change in shape in constrain to rigid body deformation. Therefore the absolute displacement of a material point at $Q_0(X_i + dX_i)$ in the nbd of $P_0(X_i)$ is given by

$$\begin{aligned} u_i + du_i &= u_i + du_i^{(1)} + du_i^{(2)} \\ &= u_i + R_{ij}dX_j + E_{ij}dX_j \end{aligned}$$

which is decomposed into three parts, viz.

- The displacement due to rigid body translation which carries the element as a whole with the displacement u_i

- The displacement due to rigid body rotation determined by R_{ij} which rotates the element as a whole through and angle $\frac{1}{2}rot\vec{u}$.
- The displacement due to straining determined by E_{ij} which causes change in the length and circulation of every line element causing a change in shape.

In particular when the displacement component $du_i^{(1)}$ due to rotation vanishes, i.e., rot$\vec{u} = \vec{0}$, displacement is called irrotational. In this case, there exists a scaler potential function ϕ, called displacement potential, such that

$$\vec{u} = -\vec{\nabla}\phi$$

Example 5.4.1. The displacement field for small deformation theory is given by $u_1 = (X_1 - X_2)^2$, $u_2 = (X_2 + X_3)^2$, $u_3 = -X_1X_2$. Determine infinitesimal strain tensor, rotation tensor at the point $(0, 2, -1)$.

Solution: The infinitesimal strain tensor are given by

$$E_{ij} = \frac{1}{2}\left[\frac{\partial u_i}{\partial X_j} + \frac{\partial u_j}{\partial X_i}\right]$$

Therefore, $E_{11} = \dfrac{\partial u_1}{\partial X_1} = 2(X_1 - X_3)$, $E_{22} = \dfrac{\partial u_2}{\partial X_2} = 2(X_2 + X_3)$ and $E_{33} = \dfrac{\partial u_3}{\partial X_3} = 0$.

$$E_{13} = \frac{1}{2}\left[\frac{\partial u_1}{\partial X_3} + \frac{\partial u_3}{\partial X_1}\right] = -X_1 - \frac{X_2}{2} + X_3 = E_{31}$$

$$E_{23} = \frac{1}{2}\left[\frac{\partial u_2}{\partial X_3} + \frac{\partial u_3}{\partial X_2}\right] = X_2 - \frac{X_1}{2} + X_3 = E_{32}$$

$$E_{12} = \frac{1}{2}\left[\frac{\partial u_1}{\partial X_2} + \frac{\partial u_2}{\partial X_1}\right] = 0 = E_{21}$$

Thus at the point $(0, 2, -1)$ the infinitesimal strain tensors are given by

$$(E_{ij}) = \begin{bmatrix} 2 & 0 & -2 \\ 0 & 2 & 1 \\ -2 & 1 & 0 \end{bmatrix}$$

The infinitesimal rotation tensors are given by

$$R_{ij} = \frac{1}{2}\left[\frac{\partial u_i}{\partial X_j} - \frac{\partial u_j}{\partial X_i}\right]$$

Therefore,

$$R_{11} = \tfrac{1}{2}\left[\tfrac{\partial u_1}{\partial X_1} - \tfrac{\partial u_1}{\partial X_1}\right] = 0,$$

$$R_{22} = \tfrac{1}{2}\left[\tfrac{\partial u_2}{\partial X_2} - \tfrac{\partial u_2}{\partial X_2}\right] = 0,$$

$$R_{33} = \frac{1}{2}\left[\frac{\partial u_3}{\partial X_3} - \frac{\partial u_3}{\partial X_3}\right] = 0,$$

$$R_{12} = \frac{1}{2}\left[\frac{\partial u_1}{\partial X_2} - \frac{\partial u_2}{\partial X_1}\right] = 0 = R_{21}$$

$$R_{13} = \frac{1}{2}\left[\frac{\partial u_1}{\partial X_3} - \frac{\partial u_3}{\partial X_1}\right] = -X_1 + \frac{X_2}{2} + X_3 = -R_{31}$$

$$R_{23} = \frac{1}{2}\left[\frac{\partial u_2}{\partial X_3} - \frac{\partial u_3}{\partial X_2}\right] = \frac{X_1}{2} + X_2 + X_3 = -R_{32}$$

Thus at the points $(0, 2, 1)$ the infinitesimal rotation tensor are given by

$$(R_{ij}) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix}$$

5.5 Geometrical Interpretation of infinitesimal strain components

A geometrical meaning for the strains is provided by considering the length and angle changes as a result of the deformation. In analysing the state of strain in undeformed body, it is natural to use the coordinates of the initial state as independent variables and follow the material description of deformation throughout. To give a geometrical interpretation of strains E_{11}, E_{22}, E_{33}. We first consider the change in length of a material line element.

5.5.1 Diagonal element of (E_{ij})

Consider a material line element P_0Q_0 of length dL at $P_0(X_1, X_2, X_3)$ oriented in the direction of (N_1, N_2, N_3) in the undeformed body. After deformation line element P_0Q_0 denotes into a line element PQ of length dl at $P(x_1, x_2, x_3)$ in the deformed body. We know that

$$\frac{dl^2 - dL^2}{dL^2} = 2E_{ij}N_iN_j \tag{5.5.1}$$

where $E_{ij} = \frac{1}{2}\left[\frac{\partial u_i}{\partial X_j} + \frac{\partial u_j}{\partial X_i}\right]$=Infinitesimal strain tensor at $P_0(X_1, X_2, X_3)$. From Eq.(5.5.1) we obtain

$$\begin{aligned} \frac{dl^2}{dL^2} - 1 &= 2E_{ij}N_iN_j \\ \Rightarrow \frac{dl}{dL} &= (1 + 2E_{ij}N_iN_j)^{1/2} \\ \Rightarrow \frac{dl}{dL} &= 1 + E_{ij}N_iN_j + \cdots \end{aligned}$$

When strain components are so small that we can neglect squares and products of E_{ij}. Therefore,

$$\begin{aligned} \frac{dl}{dL} &= 1 + E_{ij}N_iN_j \\ \Rightarrow \frac{dl}{dL} - 1 &= E_{ij}N_iN_j \\ \Rightarrow \frac{dl - dL}{dL} &= E_{ij}N_iN_j \end{aligned} \tag{5.5.2}$$

Now left hand side of Eq.(5.5.2) is the extension per unit original length of a line element oriented in the direction N_1, N_2, N_3 and is called small extensional strain denoted by $E_{(}N)$.

The small extension strain $E_{(}N) = E_{ij}N_iN_j$.

5.5.2 Geometrical interpretation of E_{11}, E_{22}, E_{33}

Consider a line element initially parallel to X_1-axis. Then we have $N_1 = 1, N_2 = 0, N_3 = 0$. Therefore, $E_{(1)} = E_{11}$.

Thus E_{11} is the extension per unit original length of a line element which is initially parallel to X_1-axis. Similarly, E_{22}, E_{33} represent the extension of a line element per unit original length which are initially parallel to X_2- and X_3- axes, respectively. These components E_{11}, E_{22}, E_{33} are called extensional strain or normal strain.

To give the geometrical interpretation of strain E_{23}, E_{31}, E_{12} we consider the change in angle between orthogonal line elements.

5.5.3 The off diagonal elements of (E_{ij})

Consider two orthogonal material line elements P_0Q_0 and P_0R_0 of length dL and δL at $P_0(X_1, X_2, X_3)$ in the undeformed state of the body oriented in the direction (N_1, N_2, N_3) and (M_1, M_2, M_3) respectively. After deformatin two line elements P_0Q_0 and P_0R_0 deform into another two line elements PQ and PR at P of length dl and δl respectively inclined at an angle θ in the deformed state of the body. We know that

$$\begin{aligned} \frac{dl}{dL}\frac{\delta l}{\delta L}\cos\theta - \cos\left(\frac{\pi}{2}\right) &= 2E_{ij}N_iM_j \\ \Rightarrow \frac{dl}{dL}\frac{\delta l}{\delta L}\sin\left(\frac{\pi}{2} - \theta\right) &= 2E_{ij}N_iM_j \\ \Rightarrow \sin\left(\frac{\pi}{2} - \theta\right) &= \frac{2E_{ij}N_iM_j}{\frac{dl}{dL}\frac{\delta l}{\delta L}}, \end{aligned}$$

where $E_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial X_j} + \frac{\partial u_j}{\partial X_i}\right)$. Now $\left(\frac{\pi}{2} - \theta\right)$ is the decrease in right angle between two orthogonal lines P_0Q_0 and P_0R_0 in the undeformed state and is called *shear* along two lines. If ν_{NM} denote

the shear along two orthogonal line elements initially oriented in the direction (N_1, N_2, N_3) and (M_1, M_2, M_3), then

$$\begin{aligned} \nu_{(NM)} &= \frac{\pi}{2} - \theta \\ \text{and } \sin\ \nu_{(NM)} &= \frac{2E_{ij}N_iM_j}{\frac{dl}{dL}\frac{\delta l}{\delta L}} \end{aligned} \tag{5.5.3}$$

If E_1 denotes the extension of P_0Q_0 and E_2 that of P_0R_0, then $E_1 = \dfrac{dl - dL}{dL} \Rightarrow dl = (1 + E_1)dL$ and $E_2 = \frac{\delta l - \delta L}{\delta L} \Rightarrow \delta l = (1 + E_2)\delta L$. Substituting these values in Eq.(5.5.3), we get

$$sin\ \nu_{(NM)} = \frac{2E_{ij}N_iM_j}{(1 + E_1)(1 + E_2)} \tag{5.5.4}$$

$$\begin{aligned} \Rightarrow sin\ \nu_{(NM)} &= 2E_{ij}N_iM_j(1 + E_1)^{-1}(1 + E_2)^{-1} \\ \Rightarrow sin\ \nu_{(NM)} &= 2E_{ij}N_iM_j(1 + E_1 + E_2 + E_1E_2)^{-1} \end{aligned}$$

$$\Rightarrow \nu_{(NM)} = 2E_{ij}N_iM_j \tag{5.5.5}$$

(since for small deformation, $\sin \nu_{(NM)} \approx \nu_{(NM)}$ and neglecting squares and products of small quantities.)

5.5.4 Geometrical Interpretation Of E_{23}, E_{31}, E_{12}

If we consider a part of orthogonal line elements initially parallel to X_2, X_3 axes respectively, then we have $N_1 = 0$, $N_2 = 1$, $N_3 = 0$ and $M_1 = 0$, $M_2 = 0$, $M_3 = 1$. Therefore,

$$\begin{aligned} \nu_{(23)} &= 2E_{23} \\ \text{or, } E_{23} &= \frac{1}{2}\nu_{(23)} \end{aligned}$$

Thus, E_{23} represents one half of the shear between two linear elements which are initially parallel to X_2 and X_3 axes. Similar interpretations can be made in regard to E_{31} and E_{12}. Also, E_{23}, E_{31}, E_{12} are called shearing strains. Thus, E_{ij} denotes increase in length of a line element per unit original length or decrease in right angle between two line lements.

Example 5.5.1. For the displacement field $u_1 = (X_1 - X_2)^2$, $u_2 = (X_2 + X_3)^2$, $u_3 = -X_1X_2$, determine the extension of a line lement in the direction of $\left(\frac{8}{9}, -\frac{1}{9}, \frac{4}{9}\right)$ and compute the change in right angle between $\vec{N} = \frac{1}{9}\left(8\hat{e_1} - \hat{e_2} + 4\hat{e_3}\right)$ and $\vec{M} = \frac{1}{9}\left(4\hat{e_1} - 4\hat{e_2} + 7\hat{e_3}\right)$ at the point $(0, 2, -1)$.

Solution: The infinitesimal strain tensor are given by

$$E_{ij} = \frac{1}{2}\left[\frac{\partial u_i}{\partial X_j} + \frac{\partial u_j}{\partial X_i}\right]$$

Therefore, $E_{11} = \frac{\partial u_1}{\partial X_1} = 2(X_1 - X_2) = -4$

$E_{22} = \frac{\partial u_2}{\partial X_2} = 2(X_2 + X_3) = 2$

$E_{33} = \frac{\partial u_3}{\partial X_3} = 0$

$E_{12} = \frac{1}{2}\left[\frac{\partial u_1}{\partial X_2} + \frac{\partial u_2}{\partial X_1}\right] = 2 = E_{21}$

$E_{13} = \frac{1}{2}\left[\frac{\partial u_1}{\partial X_3} + \frac{\partial u_3}{\partial X_1}\right] = -1 = E_{31}$

$E_{23} = \frac{1}{2}\left[\frac{\partial u_2}{\partial X_3} + \frac{\partial u_3}{\partial X_2}\right] = 1 = E_{32}$

Thus at the point $(0, 2, -1)$ the infinitesimal strain tensor are given by

$$(E_{ij}) = \begin{bmatrix} -4 & 2 & -1 \\ 2 & 2 & 1 \\ -1 & 1 & 0 \end{bmatrix}$$

The material line element at $P_0(0, 2, -1)$ is origented in the direction of $\vec{N} = (N_1, N_2, N_3) = (\frac{8}{9}, -\frac{1}{9}, \frac{4}{9})$. Then small extensional strain

$$\begin{aligned} E_{(N)} &= E_{ij}N_iN_j \\ &= E_{11}N_1^2 + E_{22}N_2^2 + E_{33}N_3^2 + 2(E_{12}N_1N_2 + E_{13}N_1N_3 + E_{23}N_3N_3) \\ &= -\frac{358}{87} \end{aligned}$$

The change of right angle between $\vec{N}$ and $\vec{M}$ at the point $(0, 2, -1)$ is given by

$$\begin{aligned} \nu_{MN} &= 2E_{ij}N_iM_j \\ &= 2(E_{11}N_1M_1 + E_{22}N_2M_2 + E_{33}N_3M_3 + 2E_{12}(N_1M_2 + N_2M_1) \\ &\quad +2E_{13}(N_1M_3 + N_3M_1) + 2E_{23}(N_1M_3 + N_3M_2)) \\ &= -\frac{574}{81} \end{aligned}$$

5.6 Few Probable Questions

1. Deduce the expression for Lagrangian strain components in the form $2r_{ij} = u_{i,j} + u_{j,i} + u_{k,i}u_{k,j}$, $(i, j, k = 1, 2, 3)$, where u_i's are the components of displacement vector at a point of the medium.

2. Deduce the expression for Eulerian strain components in a continuum medium in the form $2\eta_{ij} = u_{i,j} + u_{j,i} + u_{k,i}u_{k,j}$, $(i, j, k = 1, 2, 3)$, where u_i's are the components of displacement vector.

3. The strain tensor at a point is given by

$$(E_{ij}) = \begin{bmatrix} 5 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

Determine the extension of a line element in the direction of $\frac{1}{3}(2, 2, 1)$. What is the change of angle between two perpendicular line elements in the directions of $\frac{1}{3}(2, 2, 1)$ and $\frac{1}{\sqrt{5}}(1, 0, -2)$. [Ans: $\frac{58}{9}$, $\frac{32}{3\sqrt{5}}$.]

4. The strain tensor at a point is

$$(E_{ij}) = \begin{bmatrix} 1 & -3 & \sqrt{2} \\ -3 & 1 & -\sqrt{2} \\ \sqrt{2} & -\sqrt{2} & 4 \end{bmatrix}$$

Determine
(a) extension of a line element in the direction $\left(\frac{1}{2}, -\frac{1}{2}, \frac{1}{\sqrt{2}}\right)$
(b) shear between the directions $\left(\frac{1}{2}, -\frac{1}{2}, \frac{1}{\sqrt{2}}\right)$ and $\left(-\frac{1}{2}, -\frac{1}{2}, \frac{1}{\sqrt{2}}\right)$ and
(c) principal strains, maximum normal strain, maximum shearing strain and strain invariants.

[Ans: $E_{(N)} = 6$, $\nu_{(MN)} = 0$, $E_1 = 6,\ E_2 = 2,\ E_3 = -2$. Maximum normal strain=6, Maximum shearing strain=4, $\theta_1 = 6$, $\theta_2 = -4$, $\theta_3 = -24$]

5. For a given strain field

$$(E_{ij}) = \begin{bmatrix} K_1X_2 & 0 & 0 \\ 0 & -K_2X_2 & 0 \\ 0 & 0 & -K_2X_2 \end{bmatrix}$$

find the relation between K_1 and K_2 such that there will be no volume change. [Ans:$K_1 = 2K_2$].

6. The deformation of a body is defined by displacement components

$$u_1 = k(3x_1^2 + x_2),\ u_2 = k(2x_2^2 + x_3),\ u_3 = k(4x_3 + x_1)$$

where k is a positive constant. Find the extension of a line element that passes through the point $(1, 1, 1)$ in the direction $\frac{1}{\sqrt{3}}(1, 1, 1)$. [Ans: $\frac{17}{3}k$.]

Unit 6

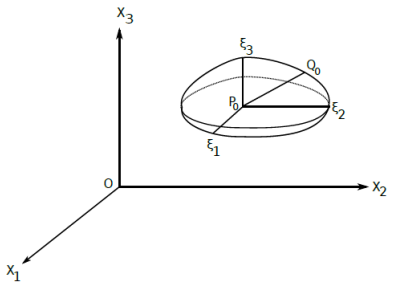
Course Structure

- Strain quadric, Principal strains
- Strain Invariants, Geometrical Interpretatoin
- Compatibility equations

6.1 The Strain Quadric

The state of local deformations in the neighbourhood of a point P_0 in undeformed state of a continuum body can be understood more clearly by a geometrical treatment.

Consider a point $P_0(X_i)$ in the undeformed state of a continuum body. Let E_{ij} be the small strain tensor at P_0 with respect to a system of axes OX_1, OX_2, OX_3 fixed in space. We introduce a local system of axes $P_0\xi_1, P_0\xi_2, P_0\xi_3$ with the origin at P_0 and axes parallel to OX_1, OX_2, OX_3 respectively. For a given set of strain tensor E_{ij} we can construct a quadric surface with its centre at P_0 given by $E_{ij}\xi_i\xi_j = 1$. This quadric surface is known as strain quadric and every straight line meets the quadratic surface in two points.



6.1.1 Properties of strain quadric

1. *The extensional strain $(E_{(N)})$ of a line element through the centre of a strain quadric in the direction of any central radius vector is equal to the inverse of the square of the radius vector.*

Consider a point $P_0(X_i)$ in the undeformed state. Let E_{ij} be the small tensor at P_0 referred to a fixed system of axes X_i. We introduce a local system of axes $P_0\xi_i$ parallel to OX_i system with origin at P. Let the equation of a strain quadric with its centre at P_0 be

$$E_{ij}\xi_i\xi_j = 1. \tag{6.1.1}$$

Draw any line P_0Q_0 through the centre P_0 to intersect the quadric surface at Q_0. Let L denotes the length P_0Q_0, (N_1, N_2, N_3) be direction cosines of P_0Q_0. Let (ξ_1, ξ_2, ξ_3) be the coordinates of Q_0 and $E_{(N)}$ be the extension of line element P_0Q_0 in the direction of P_0Q_0. Therefore $E_{(N)} = E_{ij}N_iN_j$. Also for the point Q_0, $\xi_i = LN_i \Rightarrow N_i = \dfrac{\xi_i}{L}$. Therefore

$$E_{(N)} = E_{ij}\frac{\xi_i}{L}\frac{\xi_j}{L} \tag{6.1.2}$$

Again since Q lies on strain quadric, its coordinates $\xi_1, x\xi_2, \xi_3$ satisfy Eq.(6.1.1), i.e., $E_{ij}\xi_i\xi_j = 1$. Therefore we have, $E_N = \frac{1}{L^2}$. Thus the result follows.

2. *The displacement of a material point at any point on the strain quadric relative to that at the centre is directed along the normal to the surface of the quadric at that point.*

Consider a point $P_0(X_i)$ in the undeformed state of a continuum body. Let E_{ij} be the small strain tensor at $P_0(X_i)$ with respect to a fixed system of axes OX_1, OX_2, OX_3 fixed in space. We introduce a local system of axes $P_0\xi_1, P_0\xi_2, P_0\xi_3$ with origin at $P_0(X_i)$ and parallel to the axes OX_1, OX_2, OX_3 respectively.

Let the equation of the strain quadric with its centre at $P_0(X_i)$ be $E_{ij}\xi_i\xi_j = 1$. Draw any line P_0Q_0 through the centre P_0 to intersect the quadric surface at $Q_0(\xi_1, \xi_2, \xi_3)$. Let $\overline{u_i}$ be the displacement of the material point at $Q_0(\xi_1, \xi_2, \xi_3)$ relative to that at $P_0(X_1, X_2, X_3)$ due to strain deformation only. Now since (ξ_1, ξ_2, ξ_3) are the relative coordinates of Q_0 relative to that at $P_0(X_1, X_2, X_3)$, we have $\overline{u_i} = E_{ij}\xi_j$.

Let us consider the quadratic function

$$2G(\xi_1, \xi_2, \xi_3) = E_{ij}\xi_i\xi_j$$

So the strain quadratic reduces to

$$2G(\xi_1, \xi_2, \xi_3) = 1$$

It follows from the above result that

$$\begin{aligned}
\frac{\partial}{\partial \xi_i}[2G(\xi_1,\xi_2,\xi_3)] &= \frac{\partial}{\partial \xi_i}[E_{kl\xi_k\xi_l}] \\
&= E_{kl}\left[\frac{\partial \xi_k}{\partial \xi_i}\xi_l + \xi_k\frac{\partial \xi_l}{\partial \xi_i}\right] \\
&= E_{kl}[\delta_{ki}\xi_l + \xi_k\delta_{li}] \\
&= E_{kl}\delta_{ki}\xi_l + E_{kl}\xi_k\delta_{li} \\
&= E_{il}\xi_l + E_{ki}\xi_k \\
&= E_{ij}\xi_j + E_{ji}\xi_j \\
&= 2E_{ij}\xi_j
\end{aligned}$$

Therefore,

$$\frac{\partial}{\partial \xi_i}[G(\xi_1,\xi_2,\xi_3)] = E_{ij}\xi_j = \overline{u_i}.$$

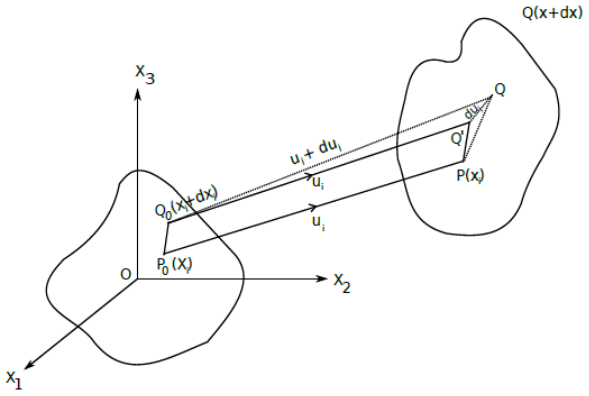
But $\dfrac{\partial G}{\partial \xi_i}$ are direction ratios of the normal to the quadric surface $2G(\xi_1,\xi_2,\xi_3) = 1$ at the point $Q_0(\xi_1,\xi_2,\xi_3)$. It follows that relative displacement is directed along the normal to the quadric surface at $Q_0(\xi_1,\xi_2,\xi_3)$.

6.2 Principal strains and principal axis of strains

In general, a line element changes its direction due to strain deformation. In particular when the direction of a line element at a given point of a continuum remains unchanged by strain deformation then that direction is called *principal direction of strain* or *principal axis of strain* and the extension that occurs along the direction is called *principal strain*.

Consider two neighbouring material points at the positions $P_0(X_1,X_2,X_3)$ and $Q_0(X_1+dX_1, X_2+dX_2, X_3+dX_3)$ of the continuum in the undeformed state, with respect to an orthogonal set of coordinate axes fixed in space. The material line element P_0Q_0 has the length dL oriented in the direction (N_1,N_2,N_3) in the initial undeformed state of a continuum body, then

$$N_i = \frac{dX_i}{dL},\ i = 1,2,3 \text{ and } N_iN_j = 1$$



Let $E_{ij} = \frac{1}{2}\left[\frac{\partial u_i}{\partial X_j} + \frac{\partial u_j}{\partial X_i}\right]$=Infinitesimal strain tensor at $P_0(X_1,X_2,X_3)$.

When the body undergoes strain deformation, the material point at $P_0(X_1, X_2, X_3)$ undergoes a displacement u_i and move to the position $P(x_1, x_2, x_3)$ and the material point $Q_0(X_i + dX_i)$ undergoes a displacement $u_i + du_i$ and moves to the position $Q_(X_1 + dX_1, X_2 + dX_2, X_3 + dX_3)$. Also the material points which form the line element P_0Q_0 in the initial state will form a new line element PQ in the deformed state.

Let E_{ij} be the small strain tensor at $P_0(X_1, X_2, X_3)$. If the line element P_0Q_0 is to be the principal direction of strain at P_0, then $\overline{P_0Q_0}$ must be parallel to $\overline{PQ}$. If u_i be the displacement of P_0 and $u_i + du_i$ be the displacement of Q_0 then du_i lie along PQ, then du_i will be proportional to dX_i, i.e.,

$$du_i = \lambda dX_i \tag{6.2.1}$$

where λ is the constant of proportionality. Thus

$$\lambda = \frac{du_i}{dX_i} = \frac{dx_i - dX_i}{dX_i}$$

Here λ is the extension of the line element $\overline{P_0Q_0}$ in the direction of $\overline{P_0Q_0}$ and λ is called the *principal strain*. We know that the strain vector is given by

$$E_i^{(N)} = \frac{du_i}{dL} = \lambda \frac{dX_i}{dL} = \lambda N_i \tag{6.2.2}$$

Also the strain vector is related to the strain tensor by the equation

$$E_i^{(N)} = E_{ij} N_j \tag{6.2.3}$$

Thus from Eq.(6.2.2) and (6.2.3) we get

$$\begin{aligned} E_{ij}N_j &= \lambda N_i = \lambda \delta_{ij} N_j \\ \Rightarrow (E_{ij} - \lambda \delta_{ij}) N_i &= 0, \; j = 1, 2, 3 \end{aligned} \tag{6.2.4}$$

By expanding we get

$$\begin{aligned} (E_{11} - \lambda)N_1 + E_{12}N_2 + E_{13}N_3 &= 0 \\ E_{21}N_1 + (E_{22} - \lambda)N_2 + E_{23}N_3 &= 0 \\ E_{31}N_1 + E_{32}N_3 + (E_{33} - \lambda)N_3 &= 0 \end{aligned} \tag{6.2.5}$$

This is a set of three homogeneous linear equation for N_1, N_2, N_3 which has to satisfy the condition

$$N_1^2 + N_2^2 + N_3^2 = 1 \tag{6.2.6}$$

The condition for the existence for a non-trivial solution of the Eq.(6.2.2) is

$$\begin{vmatrix} E_{11} - \lambda & E_{12} & E_{13} \\ E_{21} & E_{22} - \lambda & E_{23} \\ E_{31} & E_{32} & E_{33} - \lambda \end{vmatrix} = 0 \tag{6.2.7}$$

The above cubic equation is called the characteristic equation which has three roots $\lambda_1, \lambda_2, \lambda_3$ are called the *principal strain*. Corresponding to each λ_i we can solve the system of Eq.(6.2.5) subject to (6.2.6) and find (N_1, N_2, N_3) which gives the corresponding *principal direction*.

Note 6.2.1. 1. Since E_{ij} is symmetric the roots of the characteristic equation are real and hence all principal strain and direction are real.

2. Also since E_{ij} is symmetric there exist at least three mutually perpendicular directions N_1, N_2, N_3 with respect to which the matrix E_{ij} is diagonal. Geometrically, this means that infinitesimal line elements in the principal directions remains mutually perpendicular after deformation. These directions are known as principal directions.

3. The principal directions of strain corresponding to distinct principal strain are orthogonal to each other.

4. When two roots are equal we calculate the principal axes corresponding to the third and any two mutually perpendicular line which are perpendicular the third axis may be taken as principal axes.

5. When all roots are equal any three mutually perpendicular lines through the point P_0 may be taken as principal axis.

Theorem 6.2.2. All principal strains are real.

Proof. Here we are to show that the three roots E_1, E_2, E_3 of the Eq.(6.2.7) and corresponding N_i vectors are all real.

Let one of the roots of Eq.(6.2.7), say E_1 be complex. Since the coefficients of the Eq.(6.2.7) are all real so the complex conjugate E_1^* of E_1 is also a root of Eq.(6.2.7). Corresponding to these roots we obtain complex direction $N_i^{(1)}$ and its complex conjugate $N_i^{*(1)}$ satisfying the system of linear Eq.(6.2.5)

$$E_{ij}N_j^{(1)} = E_1 N_i^{(1)} \tag{6.2.8}$$

and

$$E_{ij}N_j^{*(1)} = E_1^* N_i^{*(1)}, \quad i = 1, 2, 3 \tag{6.2.9}$$

Multiplying both sides of the Eq.(6.2.8) by $N_i^{*(1)}$ and Eq.(6.2.9) by $N_i^{(1)}$, we get

$$E_{ij}N_j^{(1)}N_i^{*(1)} = E_1 N_i^{(1)}N_i^{*(1)} \tag{6.2.10}$$

$$\text{and} \quad E_{ij}N_j^{*(1)}N_i^{(1)} = E_1^* N_i^{*(1)}N_i^{(1)} \tag{6.2.11}$$

Now

$$\begin{aligned} E_{ij}N_j^{*(1)}N_i^{(1)} &= E_{ji}N^{*(1)}N_j^{(1)}, \text{ interchanging dummy indices} \\ &= E_{ij}N_i^{*(1)}N_j^{(1)}, \text{ since } E_{ij} \text{ is symmetric} \end{aligned}$$

Thus it follows from (6.2.10) and (6.2.11) that

$$\begin{aligned} & E_1 N_i^{(1)}N_i^{*(1)} = E_1^* N_i^{*(1)}N_i^{(1)} \\ \Rightarrow\ & (E_1 - E_1^*)N_i^{(1)}N_i^{*(1)} = 0 \end{aligned} \tag{6.2.12}$$

Since $N_i^{*(1)} N_i^{(1)}$ is a sum of squares of real number, it cannot be zero. Hence $E_1 = E_1^*$. Therefore, E_1 is real. Therefore, the roots E_1, E_2, E_3 of the Eq.(6.2.7) are all real and the corresponding values of $N_i^{(1)}$ of Eq.(6.2.5) are all real. Thus we have shown that at any point $P_0(X_1, X_2, X_3)$ in the undeformed body there exist three real directions of strain whose orientation is left unchanged by strain deformation. □

Theorem 6.2.3. Principal directions of strain corresponding to distinct principal strains are orthogonal to each other.

Proof. We consider the following three cases,

Case I: Three principal strains E_1, E_2, E_3 are distinct.

First let E_1 and E_2 be any two distinct real roots of the characteristic Eq.(6.2.7); $N_i^{(1)}$ and $N_i^{(2)}$ be corresponding direction cosines obtained from Eq.(6.2.5). Thus we have

$$E_{ij} N_j^{(1)} = E_1 N_i^{(1)} \; and \; E_{ij} N_j^{(2)} = E_2 N_i^{(2)}, \; i = 1, 2, 3 \tag{6.2.13}$$

Multiplying both sides of the first equation of (6.2.13) by $N_i^{(2)}$ and second by $N_i^{(1)}$, we get

$$\begin{aligned} E_{ij} N_j^{(1)} N_i^{(2)} &= E_1 N_i^{(1)} N_i^{(2)}, \; i = 1, 2, 3 \\ \text{and } E_{ij} N_j^{(2)} N_i^{(1)} &= E_2 N_i^{(2)} N_i^{(1)}, \; i = 1, 2, 3 \end{aligned} \tag{6.2.14}$$

But

$$\begin{aligned} E_{ij} N_j^{(2)} N_i^{(1)} &= E_{ji} N_i^{(2)} N_j^{(1)} \text{ (interchanging dummy indices)} \\ &= E_{ij} N_i^{(2)} N_j^{(1)} \text{ (since } E_{ij} \text{ is symmetric)} \end{aligned} \tag{6.2.15}$$

Thus from Eq.(6.2.14) it follows that

$$\begin{aligned} & E_1 N_i^{(1)} N_i^{(2)} = E_2 N_i^{(2)} N_i^{(1)} \\ \Rightarrow \; & (E_1 - E_2) N_i^{(1)} N_i^{(2)} = 0 \\ \Rightarrow \; & N_i^{(1)} N_i^{(2)} = 0 \; (\because \; E_1 \neq E_2) \end{aligned}$$

This shows that $N_i^{(1)}$ and $N_i^{(2)}$ are orthogonal. Thus two principal directions of strain corresponding to two distinct principal strains are orthogonal.

Similar results are obtained for other set of pair of roots consequently three principal directions of strain are mutually perpendicular provided three principal strains are distinct.

Case II: Two principal strains are equal.

Let the roots E_1 and E_2 of the characteristic equation (6.2.7) be equal so that we can write $E_1 = E_2$. We know that Eq.(6.2.7) has at least one real root, say E_3. For $E = E_3$ the Eq.(6.2.5) has one real solution $N_i^{(3)}$. This solution defines the third principal direction of strain.

Choose the associated direction as X_3 axis and any pair of mutually perpendicular lines each of which is perpendicular to X_3 axis are taken as X_1 and X_2 axes.

In the new system of coordinates

$$N_1^{(3)} = 0,\ N_2^{(3)} = 0,\ N_3^{(3)} = 0.$$

It follows from equation $E_{ij}N_j = EN_i$ that

$$E_{ij}N_j^{(3)} = E_3N_i^{(3)}.$$

Therefore,

$$E_{1j}N_j^{(3)} = E_3N_1^{(3)} = 0,\ i = 1, \Rightarrow E_{13} = 0.$$

Similarly $E_{2j}N_j^{(3)} = E_3N_2^{(3)} = 0,\ i = 2$, i.e., $E_{23} = 0$ and $E_{3j}N_j^{(3)} = E_3N_3^{(3)} = E_3,\ i = 3$, i.e., $E_{33} = E_3$.

Therefore the Eq.(6.2.7) reduces to

$$\begin{vmatrix} E_{11} - E & E_{12} & 0 \\ E_{21} & E_{22} - E & 0 \\ 0 & 0 & E_3 - E \end{vmatrix} = 0$$

$$\Rightarrow (E_3 - E)\{(E_{11} - E)(E_{22} - E) - E_{12}E_{21}\} = 0$$

Thus $E = E_3$ is one of the root of the equation. The other two roots of this cubic equation are given by

$$\begin{aligned} &(E_{11} - E)(E_{22} - E) - E_{12}E_{21} = 0 \\ &\Rightarrow E^2 - (E_{11} + E_{22})E + (E_{11}E_{22} - E_{12}^2) = 0\ [\because\ E_{12} = E_{21}] \end{aligned}$$

Since $E_1 = E_3$ the above equation would have equal roots if

$$\begin{aligned} &(E_{11} + E_{22})^2 - 4(E_{11}E_{22} - E_{12}^2) = 0 \\ &\Rightarrow (E_{11} - E_{22})^2 + 4E_{12}^2 = 0 \\ &\Rightarrow E_{11} - E_{22} = 0 \ \text{ and } \ E_{12} = 0 \end{aligned} \tag{6.2.16}$$

Hence we have $E_{13} = E_{23} = E_{12} = 0$.

It follows that for $E_1 = E_2$ coordinate system X_i and in consequence any coordinate system containing the third principal axis $N_i^{(3)}$ corresponding to E_3 as X_3 defines one parameter family of systems of principal directions of strain.

Thus when two roots are equal, any two mutually orthogonal directions lying in a plane perpendicular to the principal direction corresponding to simple roots may be taken as corresponding principal directions of strain.

Case III: All the principal strains are equal.

Consider the case when all the three principal strains E_1, E_2, E_3 are equal. When $E_1 = E_2$, any system of coordinate axis with third principal axis $N_i^{(3)}$ corresponding to E_3 as X_3 defines a system of principal directions of strain. When $E_2 = E_3$, any system of coordinate axis with third principal axis $N_i^{(1)}$ corresponding to E_1 as X_1 defines a system of principal direction of strain. When $E_3 = E_1$, any system of coordinate axis with third principal axis $N_i^{(2)}$ corresponding to E_2 as X_2 defines a system of principal directions of strain.

Therefore, for $E_1 = E_2 = E_3$, every system of space is a principal direction of strain. □

6.3 Strain Invariants

There are a number of constraints of strain tensors E_{ij} which remains unaltered by the rotation of the coordinate system. They are called *strain invariants*.

We know that the three principal strains E_1, E_2, E_3 are roots of the characteristic equation

$$\begin{vmatrix} E_{11} - E & E_{12} & 0 \\ E_{21} & E_{22} - E & 0 \\ 0 & 0 & E_{33} - E \end{vmatrix} = 0 \tag{6.3.1}$$

Expanding, we get

$$E^3 - \theta_1 E^2 + \theta_2 E - \theta_3 = 0 \tag{6.3.2}$$

where

$$\theta_1 = E_{11} + E_{22} + E_{33} \tag{6.3.3}$$

$$\theta_2 = \begin{vmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{vmatrix} + \begin{vmatrix} E_{22} & E_{23} \\ E_{32} & E_{33} \end{vmatrix} + \begin{vmatrix} E_{33} & E_{31} \\ E_{13} & E_{11} \end{vmatrix} \tag{6.3.4}$$

$$= E_{11}E_{22} + E_{22}E_{33}E_{33}E_{11} - E_{12}^2 - E_{23}^2 - E_{31}^2 \tag{6.3.5}$$

and

$$\theta_3 = \begin{vmatrix} E_{11} & E_{12} & E_{13} \\ E_{21} & E_{22} & E_{23} \\ E_{31} & E_{32} & E_{33} \end{vmatrix} \tag{6.3.6}$$

The coefficients θ_1, θ_2 and θ_3 are called the principal scalar invariants of the strain tensor.

Since E_1, E_2, E_3 are the roots of Eq.(6.3.1), by the relation between the roots and coefficients of the equation, we have

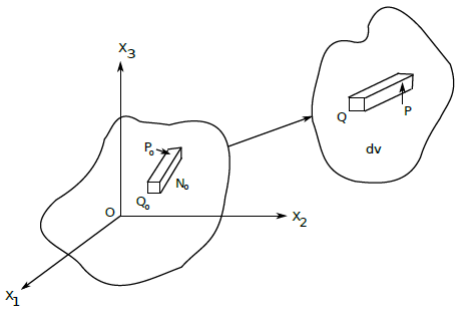
$$\begin{aligned} \theta_1 &= E_1 + E_2 + E_3 \\ \theta_2 &= E_1E_2 + E_2E_3 + E_3E_1 \\ \theta_3 &= E_1E_2E_3 \end{aligned}$$

Since the principal strains E_1, E_2, E_3 at a point have a geometrical meaning independent of the choice of coordinate system, it is clear from Eq.(6.3.1) that θ_1, θ_2 θ_3 given by Eq.(6.3.3), (6.3.5) and (6.3.6) are invariant with respect to orthogonal transformations of coordinates and are respectively called first, second and third strain invariants as they have the same values in all coordinate system.

6.4 Geometrical Interpretation of the First Strain Invariants

Let us consider the change in volume element. Let P_0 be a point in the initial state. Let E_1, E_2, E_3 be three principal strains at P_0. Consider a volume element of continuum occupying a rectangular parallelopiped of volume dV_0, with one of its vertices at P_0, with edges parallel to the principal direction of strains at P_0 and of lengths of L_1, L_2, L_3 so that $dV_0 = L_1L_2L_3$.

After deformation this element becomes again a rectangular parallelopiped of volume dV because orienttation of principal directions of strain remain unchanged. If l_1, l_2, l_3 be the lengths of its edges then



$$dV = l_1l_2l_3$$

Now,

$$\begin{aligned} l_1 &= (1+E_1)L_1 \\ l_2 &= (1+E_2)L_2 \\ l_3 &= (1+E_3)L_3 \end{aligned}$$

Hence,

$$\begin{aligned} \frac{dV - dV_0}{dV_0} &= \frac{l_1l_2l_3 - L_1L_2L_3}{L_1L_2L_3} \\ &= \frac{(1+E_1)(1+E_2)(1+E_3)L_1L_2L_3 - L_1L_2L_3}{L_1L_2L_3} \\ &= (1+E_1)(1+E_2)(1+E_3) - 1 \\ &= E_1 + E_2 + E_3 \\ &= E_{11} + E_{22} + E_{33} \\ &= \theta_1 \end{aligned}$$

Then the first strain invariant $\theta_1 = E_{11} + E_{22} + E_{33}$ represents the change in volume per unit original volume.

Note:

1. For small deformation $\theta_1 = E_1 + E_2 + E_3$ =first principal scalar invariant
   In general,

$$\begin{aligned}\theta_1 &= E_1 + E_2 + E_3 \\ &= E_{11} + E_{22} + E_{33} \\ &= \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} + \frac{\partial u_3}{\partial x_3} \\ &= \text{div } u\end{aligned}$$

This unit volume change is known as *dilations*.

2. If the change in volume element remains unaltered then the deformation is called *Isochoric deformation*. Thus, for isochoric deformation,

$$E_{11} + E_{22} + E_{33} = 0$$

Example 6.4.1. Determine the principal direction and principal strains for

$$(E_{ij}) = \begin{bmatrix} e & e & e \\ e & e & e \\ e & e & e \end{bmatrix}$$

Solution: The principal strains E_1, E_2, E_3 at the point P are the roots of the characteristic equation

$$\begin{vmatrix} e-E & e & e \\ e & e-E & e \\ e & e & e-E \end{vmatrix} = 0 \tag{6.4.1}$$

$$\begin{aligned}&\Rightarrow E^2(3e - E) = 0 \\ &\Rightarrow E = 0,\ 0,\ 3e\end{aligned}$$

The principal directions of stain at P are given by

$$\begin{bmatrix} e-E & e & e \\ e & e-E & e \\ e & e & e-E \end{bmatrix} \begin{bmatrix} N_1^{(1)} \\ N_2^{(1)} \\ N_3^{(1)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{6.4.2}$$

For $E = 3e$, the above system of Eq.(6.4.2) becomes

$$\begin{bmatrix} e-3e & e & e \\ e & e-3e & e \\ e & e & e-3e \end{bmatrix} \begin{bmatrix} N_1^{(1)} \\ N_2^{(1)} \\ N_3^{(1)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Hence

$$\begin{aligned}-2eN_1^{(1)} + eN_2^{(1)} + eN_3^{(1)} &= 0 \\ eN_1^{(1)} - 2eN_2^{(1)} + eN_3^{(1)} &= 0 \\ eN_1^{(1)} + eN_2^{(1)} - 2eN_3^{(1)} &= 0\end{aligned}$$

Therefore,

$$\begin{bmatrix} N_1^{(1)} \\ N_2^{(1)} \\ N_3^{(1)} \end{bmatrix} = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

For $E = 0$, the above system of Eq.(6.4.2) becomes

$$\begin{aligned} eN_1^{(2)} + eN_2^{(2)} + eN_3^{(2)} &= 0 \\ eN_1^{(2)} - 2eN_2^{(2)} + eN_3^{(2)} &= 0 \\ eN_1^{(2)} + eN_2^{(2)} + eN_3^{(1)} &= 0 \end{aligned}$$

These equations together with $(N_1^{(2)})^2 + (N_2^{(2)})^2 + (N_3^{(2)})^2 = 1$ are insufficient to determine the principal direction corresponding to $E = 0$. Since two principal strains are equal, any pair of lines perpendicular to each other and each perpendicular to $\left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right)$ may be taken as principal directions of strain.

Example 6.4.2. Given $E_{11} = k(X_1^2 - X_2^2),\ E_{12} = -kX_1X_2,\ E_{22} = kX_1X_2,\ E_{13} = E_{33} = E_{23} = 0$, a possible state of strain. Find the displacement components.

Solution: From the relation $E_{11} = k(X_1^2 - X_2^2)$ we get

$$\begin{aligned} E_{11} &= \frac{\partial u_1}{\partial X_1} = k(X_1^2 - X_2^2) \\ \Rightarrow u_1 &= k(\frac{1}{3}X_1^3 - X_1X_2^2) + f(X_2), \text{ where } f \text{ is arbitrary} \end{aligned}$$

Similarly from $E_{22} = kX_1X_2$, we get

$$\begin{aligned} E_{22} &= \frac{\partial u_2}{\partial X_2} = kX_1X_2, \\ \Rightarrow u_2 &= \frac{k}{2}X_1X_2^2 + g(X_1), \text{ where } g \text{ is arbitrary} \end{aligned}$$

The functions f and g are to be determined. Now using the formula

$$\begin{aligned} 2E_{12} &= \frac{\partial u_1}{\partial X_2} + \frac{\partial u_2}{\partial X_1} \\ \Rightarrow -2kX_1X_2 &= -2kX_1X_2 + f'(X_2) + \frac{k}{2}X_2^2 + g'(X_1) \\ \Rightarrow g'(X_1) + f'(X_2) &= -\frac{k}{2}X_2^2 \\ \Rightarrow g'(X_1) &= -f'(X_2) - \frac{k}{2}X_2^2. \end{aligned}$$

Since the left hand side is a function of X_1 only and the right hand side is a function of X_2 alone, each side mus be constant equal to c, say. Thus

$$\begin{aligned} g'(X_1) = c \text{ and } &-f'(X_2) - \frac{k}{2}X_2^2 = c, \\ \Rightarrow g(X_1) = cX_1 + c_1 \text{ and } &f(X_2) = -\frac{k}{6}X_2^3 - cX_2 + c_2 \end{aligned}$$

where c_1 and c_2 are arbitrary constant. Finally the displacement components are given by

$$\begin{aligned} u_1 &= k(\frac{1}{3}X_1^3 - X_1X_2^2) - \frac{k}{6}X_2^3 - cX_2 + c_2, \\ u_2 &= \frac{k}{2}X_1X_2^2 + cX_1 + c_1, \\ u_3 &= 0 \end{aligned}$$

Omitting the linear parts of the displacements (which corresponds to rigid motion), we have for pure deformation

$$\begin{aligned} u_1 &= k(\frac{1}{3}X_1^3 - X_1X_2^2) - \frac{k}{6}X_2^3, \\ u_2 &= \frac{k}{2}X_1X_2^2, \\ u_3 &= 0. \end{aligned}$$

6.5 Compatibility equations for strain components

If the strain components E_{ij} are given explicitly as functions of coordinates, the six independent equations

$$E_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial X_j} + \frac{\partial u_j}{\partial X_i}\right), \quad i = 1, 2, 3$$

may be viewed as a system of six linear partial differential equations for determining three unknown displacement components u_i. The system of equations is an over determined system as number of equations (six) is more than number of unknowns (three). To ensure the existence of single valued displacement solutions, strain components cannot arbitrarily prescribed as function of coordinates but must necessarily be subjected to additional restrictions or conditions.

The necessary and sufficient condition for the existence of single valued displacement is that strain components must satisfy the compatibility equation

$$E_{ij,kl} + E_{kl,ij} - E_{ik,jl} - E_{jl,ik} = 0.$$

The above system has $3^4 = 81$ equations but only six are algebraically independent. These six equations are

$$\begin{aligned} E_{22,33} + E_{33,22} &= 2E_{23,23} \\ E_{33,11} + E_{11,33} &= 2E_{31,31} \\ E_{11,22} + E_{22,11} &= 2E_{12,12} \\ E_{11,23} &= (-E_{23,1} + E_{31,2} + E_{12,3})_{,1} \\ E_{22,31} &= (-E_{23,1} - E_{31,2} + E_{12,3})_{,2} \\ E_{33,12} &= (E_{23,1} + E_{31,2} - E_{12,3})_{,3} \end{aligned}$$

6.6 Few Probable Questions

1. Calculate the strain invariants from strain tensor

$$(E_{ij}) = \begin{bmatrix} 5 & -1 & -1 \\ -1 & 4 & 0 \\ -1 & 0 & 4 \end{bmatrix}$$

Determine principal strains. Find strain invariants from them. Show the equivalence of strain invariants. [Ans: $\theta_1 = 13,\ \theta_2 = 54,\ \theta_3 = 72,\ E_1 = 6,\ E_2 = 4,\ E_3 = 3$].

2. Determine principal strains and corresponding direction from strain tensors

$$(a)(E_{ij}) = \begin{bmatrix} 1 & 1 & 3 \\ 1 & 5 & 1 \\ 3 & 1 & 1 \end{bmatrix}$$

[Ans: 6, 3, 2; $\frac{1}{\sqrt{6}}(1,2,1),\ \frac{1}{\sqrt{3}}(1,1,1), \frac{1}{\sqrt{2}}(1,0,-1)$]

$$(b)(E_{ij}) = \begin{bmatrix} 5 & 2 & 2 \\ 2 & 2 & -4 \\ 2 & -4 & 2 \end{bmatrix}$$

[Ans: 6, 6, -3; $\frac{1}{\sqrt{3}}(2,2,-1),\ \frac{1}{3}(2,-1,2), \frac{1}{3}(-1,2,2)$]

3. The strain tensor at a point is given by

$$(E_{ij}) = \begin{bmatrix} a & b & 0 \\ b & -a & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Find principal axes of strain and corresponding direction ratios of principal strains.
[Ans:$E_1, E_2 = \sqrt{a^2+b^2}, E_3 = -\sqrt{a^2+b^2},\ (0,0,1),\ \left(\frac{a+\sqrt{a^2+b^2}}{b}, 1, 0\right),\ \left(\frac{a-\sqrt{a^2+b^2}}{b}, 1, 0\right)$.

4. A body undergoes deformation

$$\begin{aligned} x_1 &= \sqrt{2}X_1 + \frac{3}{4}\sqrt{2}X_2 \\ x_2 &= -X_1 + \frac{3}{4}X_2 + \frac{\sqrt{2}}{4}X_3 \\ x_3 &= X_1 - \frac{3}{4}X_2 + \frac{1}{4}\sqrt{2}X_3 \end{aligned}$$

Find the direction after deformation of a line element with initial direction ratios $1 : 1 : 1$. [Ans: $7\sqrt{2} : \sqrt{2} - 1 : \sqrt{2} + 1$].

5. The displacement in an elastic solid is given by

$$\begin{aligned} u_1 &= k(X_1 + 2X_2 + 3X_3) \\ u_2 &= k(-2X_1 + X_2) \\ u_3 &= k(X_1 + 4X_2 + 2X_3) \end{aligned}$$

where k is a small quantity. Find dilation, rotation, shear, principal strain and corresponding principal axes.

6. Find r_{ij} and η_{ij} for deformation

$$\begin{aligned} x_1 &= a_1(X_1 + \alpha X_2) \\ x_2 &= a_2 X_2 \\ x_3 &= a_3 X_3 \end{aligned}$$

where a_1, a_2, a_3 are constants.

Ans.

$$(r_{ij}) = \frac{1}{2}\begin{bmatrix} a_1^2 & \alpha a_1^2 & 0 \\ \alpha a_1^2 & \alpha a_1^2 + a_2^2 & 0 \\ 0 & 0 & a_3^2 \end{bmatrix} - \frac{1}{2}\delta_{ij}$$

$$(\eta_{ij}) = \frac{1}{2}\delta_{ij} - \frac{1}{2}\begin{bmatrix} a_1^{-2} & -\alpha a_1^{-2} a_2^{-2} & 0 \\ -\alpha a_1^{-2} a_2^{-2} & (1+\alpha^2) a_2^{-2} & 0 \\ 0 & 0 & a_3^{-2} \end{bmatrix}$$

7. The deformation of a body is given by

$$\begin{aligned} u_1 &= AX_1 + BX_1(X_1^2 + X_2^2)^{-1} \\ u_2 &= AX_2 + BX_2(X_1^2 + X_2^2)^{-1} \\ u_3 &= CX_3 \end{aligned}$$

where A, B, C are constants. Find E_{ij} and R_{ij}. Find principal values and principal axes of E_{ij}.
Ans:

$$(E_{ij}) = \begin{bmatrix} A - B\frac{X_1^2 - X_2^2}{(X_1^2+X_2^2)^2} & -2B\frac{X_1 X_2}{(X_1^2+X_2^2)^2} & 0 \\ -2B\frac{X_1 X_2}{(X_1^2+X_2^2)^2} & A + B\frac{X_1^2 - X_2^2}{(X_1^2+X_2^2)^2} & 0 \\ 0 & 0 & C \end{bmatrix}$$

$\mathrm{R}_{ij} = 0$, principal strains: $A \pm B(X_1^2 + X^2)^{-1}$; C; direction ratios: $(X_2, -X_1, 0)$, $(X_1, X_2, 0)$, $(0, 0, 1)$.

8. Show that $E_{11} = k(X_1^2 + X_2^2)$, $E_{22} = kX_2^2$, $E_{12} = kX_1X_2$, $E_{33} = E_{23} = E_{31} = 0$ is a possible state of strain rate while $E_{11} = kX_3(X_1^2 + X_2^2)$, $E_{22} = kX_2^2X_3$, $E_{12} = kX_1X_2$, $E_{33} = E_{23} = E_{31} = 0$ is not a possible one.

Unit 7

Course Structure

- Motion of deformable bodies
- Lagrangian and Eulerian description
- Flow: Path line and stream line
- Boundary surface
- Conservation of mass: Equation of continuity

7.1 Introduction

In order to study the motion of the continuum are generally follow the one of the following two methods.

(a) Lagrangian description of motion or material description of motion of a continuum.

(b) Eulerian description of motion or particle description of motion of a continuum.

7.2 Lagrangian description of motion of a continuum

We consider a fixed frame of reference OX_1, OX_2, OX_3. Let a material particle which is initially at $P_0(X_1, X_2, X_3)$ move to another point $P(x_1, x_2, x_3)$ after time t. The coordinate (x_1, x_2, x_3) will be functions of their initial values X_1, X_2, X_3 and t. Thus $x_i = f_i(X_1, X_2, X_3, t)$. The components of velocity of particle at time t whose initial coordinates are (X_1, X_2, X_3) are

$\frac{\partial x_1}{\partial t}, \frac{\partial x_2}{\partial t}, \frac{\partial x_3}{\partial t}$ and acceleration components $\frac{\partial^2 x_1}{\partial t^2}, \frac{\partial^2 x_2}{\partial t^2}, \frac{\partial^2 x_3}{\partial t^2}$.

It is to be understood that in the differentiation with respect to $'t'$ the initial coordinated (X_1, X_2, X_3) of the particle are kept unaltered. Such a differentiation is frequently referred to as particle differentiation or differentiation following the particle. Lagrangian description or material description of motion is usually used in elastic solids. IF u_i be the displacement of the particle at time t then its coordinates after time t is $x_i = X_i + u_i$. The displacement components are obviously functions of X_1, X_2, X_3 and t so its velocity components are $\frac{\partial x_i}{\partial t}$ which are equivalent to $\frac{\partial u_i}{\partial t}$ and the acceleration components are $\frac{\partial^2 u_i}{\partial t^2}$.

7.2.1 Eulerian description of motion of a continuum

In the material description or Lagrangian description every particle is identified by its initial coordinates at $t = 0$ and time t. This is not always convenient, when it describes the flow of watch in a river we do not desire to identify the location from where every particle of water comes. In this case we are generally interested in the instantaneous velocity field and its change with time. In the Eulerian method a particle point in the space occupied by the fluid is selected. We denote this point by its coordinates (x_1, x_2, x_3). In this case x_1, x_2, x_3 and t are independent. So expressions like $\dot{x}, \ddot{x}, \cdots$ etc do not occur.

In Eulerian method the velocity of fluid at a point is measured as follows:

Consider a fixed point $P(x_1, x_2, x_3)$ in space, at this point suppose an apparatus is placed to record the velocity for different time t. This measurement yields velocity at P as a function of t. However by locating the recording apparatus at all points of the medium and assembling the resulting data we obtain the velocity yield as a function of t and x_1, x_2, x_3. Therefore velocity components v_1, v_2, v_3 as a function of x_1, x_2, x_3 and t are known. In order to obtain the expression for acceleration in Eulerian method we assume that $v_1 = F(x_1, x_2, x_3, t)$.

Let after an interval of infinitesimal time δt, the material point move on to a neighbouring position $(x_1 + \delta x_1, x_2 + \delta x_2, x_3 + \delta x_3)$. Thus

$$\begin{aligned} \delta x_1 &= v_1\, \delta t \quad \text{in} \quad x_1 \quad \text{direction} \\ \delta x_2 &= v_2\, \delta t \quad \text{in} \quad x_2 \quad \text{direction} \\ \delta x_3 &= v_3\, \delta t \quad \text{in} \quad x_3 \quad \text{direction}\,. \end{aligned} \tag{7.2.1}$$

If δv_1 be the change in particle the x_1 - component of velocity by this time then

$$\begin{aligned} v_1 + \delta v_1 &= F(x_1 + v_1\delta t, x_2 + v_2\delta t, x_3 + v_3\delta t, t + \delta t) \\ &= F(x_1, x_2, x_3) + \delta t\left[v_1\frac{\partial F}{\partial x_1} + v_2\frac{\partial F}{\partial x_2} + v_3\frac{\partial F}{\partial x_3} + \frac{\partial F}{\partial t}\right] \\ &\quad + \cdots \text{ the terms containing high power of } \delta t \\ &= v_1 + \delta t\left[v_1\frac{\partial F}{\partial x_1} + v_2\frac{\partial F}{\partial x_2} + v_3\frac{\partial F}{\partial x_3} + \frac{\partial F}{\partial t}\right] \end{aligned} \tag{7.2.2}$$

Hence the x_1-component of acceleration

$$\begin{aligned} \frac{dv_1}{dt} &= \lim_{\delta t\to 0}\frac{\delta v_1}{\delta t} = v_1\frac{\partial v_1}{\partial x_1} + v_2\frac{\partial v_1}{\partial x_2} + v_3\frac{\partial v_1}{\partial x_3} + \frac{\partial v_1}{\partial t} \\ &= \left[v_1\frac{\partial}{\partial x_1} + v_2\frac{\partial F}{\partial x_2} + v_3\frac{\partial}{\partial x_3} + \frac{\partial}{\partial t}\right]v_1 \\ &= \frac{Dv_1}{Dt} \text{ where } \frac{D}{Dt} \equiv v_1\frac{\partial}{\partial x_1} + v_2\frac{\partial F}{\partial x_2} + v_3\frac{\partial}{\partial x_3} + \frac{\partial}{\partial t} \\ &= \frac{\partial}{\partial t} + v_k\frac{\partial}{\partial x_k} = \frac{\partial}{\partial t} + (\vec{v}\cdot\vec{\nabla}) \end{aligned} \tag{7.2.3}$$

Similarly components of acceleration in x_2 and x_3 directions are respectively $\frac{Dv_2}{Dt}$ and $\frac{Dv_3}{Dt}$.

Example: *Motion of a particle is given by* $x_1 = X_1 + X_2t + X_3t^2, x_2 = X_2 + X_3t + X_1t^2, x_3 = X_3 + X_1t + X_2t^2$.
(i) Find at time t, the velocity and acceleration of a particle which was at $(1, 1, 1)$ *at* $t = 0$.
(ii) Find at time t, the velocity and acceleration of a particle which is at $(1, 1, 1)$ *at time t.*

Solution: Here $x_i = x_i(X_1, X_2, X_3, t)$. Therefore the velocity components of the particle which was as $(1, 1, 1)$ at $t = 0$ are given by

$$\begin{aligned} v_1 &= \frac{\partial x_1}{\partial t} = X_2 + 2X_3t = 1 + 2t \quad \text{at} \quad (1, 1, 1) \\ v_2 &= \frac{\partial x_2}{\partial t} = X_3 + 2X_1t = 1 + 2t \quad \text{at} \quad (1, 1, 1) \\ v_3 &= \frac{\partial x_3}{\partial t} = X_1 + 2X_2t = 1 + 2t \quad \text{at} \quad (1, 1, 1) \end{aligned}$$

The acceleration of a particle which was at $(1, 1, 1)$ at $t = 0$ are given by

$$\begin{aligned} f_1 &= \frac{\partial^2 x_1}{\partial t^2} = 2X_3 = 2 \quad \text{at} \quad (1, 1, 1) \\ f_2 &= \frac{\partial^2 x_2}{\partial t^2} = 2X_1 = 2 \quad \text{at} \quad (1, 1, 1) \\ f_3 &= \frac{\partial^2 x_3}{\partial t^2} = 2X_2 = 2 \quad \text{at} \quad (1, 1, 1) \end{aligned}$$

Now using the given relations we have

$$x_1 - x_2 t = X_1(1 - t^3)$$
$$\Rightarrow X_1 = \frac{x_1 - x_2 t}{1 - t^3}$$

Similarly, $X_2 = \dfrac{x_2 - x_3 t}{1 - t^3}$, $X_3 = \dfrac{x_3 - x_1 t}{1 - t^3}$.
Therefore,

$$v_1 = X_2 + 2X_3 t = \frac{x_2 + x_3 t - 2x_1 t2}{1 - t^3} = \frac{1 + t - 2t^2}{1 - t^3} \text{ at } x_i = 1$$
$$v_2 = X_3 + 2X_1 t = \frac{x_3 + x_1 t - 2x_2 t2}{1 - t^3} = \frac{1 + t - 2t^2}{1 - t^3} \text{ at } x_i = 1$$
$$v_3 = X_1 + 2X_2 t = \frac{x_1 + x_2 t - 2x_3 t2}{1 - t^3} = \frac{1 + t - 2t^2}{1 - t^3} \text{ at } x_i = 1$$

Now the acceleration components

$$\begin{aligned}
f_1 &= \frac{\partial v_1}{\partial t} + v_1\frac{\partial v_1}{\partial x_1} + v_2\frac{\partial v_1}{\partial x_2} + v_3\frac{\partial v_1}{\partial x_3} \\
&= \frac{\partial}{\partial t}\left[\frac{x_2 + x_3 t - 2x_1 t^2}{1 - t^3}\right] + v_1\left[-\frac{2t^2}{1 - t^3}\right] + v_2\left[\frac{1}{1 - t^3}\right] + v_3\left[\frac{t}{1 - t^3}\right] \\
&= \frac{-4x_1 t + x_3}{1 - t^3} + (x_2 + x_3 t - 2x_1 t^2)\frac{3t^2}{(1 - t^3)^2} + \frac{1}{1 - t^3}[v_1 \cdot (-2t^2) + v_2 + v_3 t] \\
&= \frac{-4t + 1}{1 - t^3} + \frac{(1 + t - 2t^2)3t^2}{(1 - t^3)^2} + \frac{1 + t - 2t^2}{(1 - t^3)^2}(1 + t - 2t^2) \text{ (as } x_1 = x_2 = x_3 = 1) \\
&= \frac{1 - 4t}{1 - t^3} + \frac{1 + t - 2t^2}{(1 - t^3)^2}[1 + t - 2t^2 + 3t^2] \\
&= \frac{(1 - 4t)(1 - t^3) + (1 + t - 2t^2)(1 + t + t^2)}{(1 - t^3)^2} \\
&= \frac{2t^4 - 2t^3 - 2t + 2}{(1 - t^3)^2} = \frac{(1 - t^3)(2 - 2t)}{(1 - t^3)^2} = \frac{2 - 2t}{1 - t^3}.
\end{aligned}$$

Similarly

$$f_2 = f_3 = \frac{2 - 2t}{1 - t^3}$$

7.3 Flow

A deformable body is called a fluid if the deformation increases indefinitely with the continued application of any force, how small it may be, on the body and can not return to its original configuration when the force is withdrawn. This continuous shear deformation is called flow.

Geometrically, in the continuum physical properties including the velocity field is dependent on time. The motion is then call unsteady. When the physical properties and the velocity field does not change in time, the motion is said to be steady.

7.4 Path line and Stream line

A path line is a curve in the continuum followed by a given particle during tis motion.

Let $\vec{v}$ be the velocity of a given particle at any point $P(x_1, x_2, x_3)$ of the path line at time t and $\vec{r}$ is the position vector of P. If v_1, v_2, v_3 be components of $\vec{v}$ and dx_1, dx_2, dx_3 be components of $d\vec{r}$, then the differential equation of the path line is

$$\vec{v} = \frac{d\vec{r}}{dt}$$

$$i.e., \quad \frac{dx_1}{v_1(x_1, x_2, x_3, t)} = \frac{dx_2}{v_2(x_1, x_2, x_3, t)} = \frac{dx_3}{v_3(x_1, x_2, x_3, t)} = dt$$

A stream line is a curve drawn in the continuum at any given instant of time such that tangent at every point of it is in the instantaneous direction of the velocity of the particle at that point. Therefore the differential equation of the stream line at any given instant t is

$$\frac{dx_1}{v_1(x_1, x_2, x_3, t)} = \frac{dx_2}{v_2(x_1, x_2, x_3, t)} = \frac{dx_3}{v_3(x_1, x_2, x_3, t)}$$

The stream line shows how each particle is moving at a given instant of time while the path line shows how a given particle is moving at each instant.

Experimentally, by dropping a small visible floating particle and taking a long time exposer, we see the trace of the particle revealing its path line. A short time exposure of a fluid onto which many floating particles are dropped will show the instantaneous direction of the velocity field, this revealing the stream lines.

Example: *Find the stream line and path line of a continuum particle for the velocity field* $v_1 = \frac{x_1}{1+t}, v_2 = x_2, v_3 = 0$.

Solution: The differential equation of the stream line at a given instant t is given by

$$\begin{aligned} & \frac{dx_1}{v_1} = \frac{dx_2}{v_2} = \frac{dx_3}{v_3} \\ \Rightarrow \quad & \frac{dx_1}{x_1/(1+t)} = \frac{dx_2}{x_2} = \frac{dx_3}{0} \quad \text{where } t \text{ is a constant.} \end{aligned} \tag{7.4.1}$$

Therefore

$$\begin{aligned} & (1+t)\frac{dx_1}{x_1} = \frac{dx_2}{x_2} \\ \Rightarrow \quad & (1+t)\log\ x_1 = \log\ x_2 + \log\ c_1 \\ \Rightarrow \quad & x_1^{(1+t)} = c_1 x_2 \end{aligned}$$

Also $dx_3 = 0$ gives $x_3 =$ constant $= c_2$ (say). Therefore the required equation of the streamline at a given instant t is

$$x_2 = \frac{x_1^{(1+t)}}{c_1} \text{ and } x_3 = c_2. \tag{7.4.2}$$

The differential equation of the path line is given by

$$\frac{dx_1}{dt} = v_1 = \frac{x_1}{1+t}, \quad \frac{dx_2}{dt} = v_2 = x_2 \ \ and \ \ \frac{dx_3}{dt} = v_3 = 0, \ \ where\ t\ is\ a\ variable. \tag{7.4.3}$$

Now

$$\begin{aligned} & \frac{dx_1}{dt} = \frac{x_1}{1+t} \\ \Rightarrow \quad & \log\left(\frac{x_1}{c_3}\right) = \log(1+t) \\ \Rightarrow \quad & x_1 = c_2(1+t) \end{aligned}$$

Now

$$\frac{dx_2}{dt} = x_2 \Rightarrow x_2 = c_4 e^t \ \text{ and } \ \frac{dx_3}{dt} = 0 \Rightarrow x_3 = \text{constant} = c_5$$

Similarly, the equation of path is $x_2 = c_4 e^{(\frac{x_1}{c_3} - 1)}$ and $x_3 = c_5$.

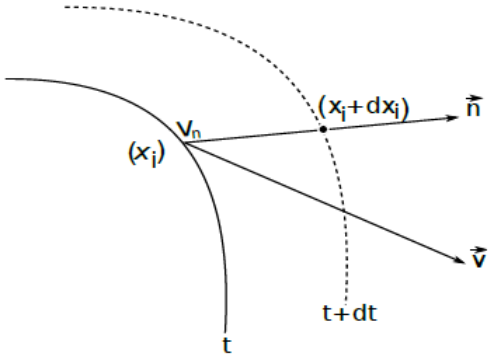
7.5 Boundary Surface

A surface is a boundary surface when it always consists of the same set of particles.

Theorem: *A necessary and sufficient condition for a given surface $F(x_1, x_2, x_3, t) = 0$ to be a boundary surface is that*

$$\dot{F} = \frac{DF}{Dt} = \frac{\partial F}{\partial t} + v_1\frac{\partial F}{\partial x_1} + v_2\frac{\partial F}{\partial x_2} + v_3\frac{\partial F}{\partial x_3} = 0, \ \ i.e., \ \ \frac{\partial F}{\partial t} + v_k F_{,k} = 0$$

Proof: Necessary Part:



Let $F(x_1, x_2, x_3, t) = 0$ be a boundary surface. The necessary condition implies that the component of the velocity of a particle on the surface along the normal to the boundary surface must be equal to the normal velocity of the surface itself.

If v_n be the velocity of the surface normal to itself at any point (x_1, x_2, x_3) , n_1, n_2, n_3 the direction cosines of the normal to the surface and (v_1, v_2, v_3) be the velocity components of the particle intravenously occupying the point (x_1, x_2, x_3), then mathematical formulation of the condition is

$$v_n = n_i v_i$$

Since v_n is the normal component of velocity, the displacement components (dx_1, dx_2, dx_3) of the point $P(x_1, x_2, x_3)$ of the surface are given by

$$dx_i = n_i v_i dt$$

At time t, let $P(x_1, x_2, x_3)$ be the position of the particle of the surface $F(x_1, x_2, x_3, t) = 0$. At time $t + dt$, the point moves to the position $Q(x_1 + dx_1, x_2 + dx_2, x_3 + dx_3)$. Since the particle continues to lie on the surface $F(x_1, x_2, x_3, t) = 0$, we must have

$$\begin{aligned}
& F(x_1 + dx_1, x_2 + dx_2, x_3 + dx_3, t + dt) = 0 \\
\Rightarrow \quad & F(x_1, x_2, x_3, t) + \frac{\partial F}{\partial x_1}dx_1 + \frac{\partial F}{\partial x_2}dx_2 + \frac{\partial F}{\partial x_3}dx_3 + \frac{\partial F}{\partial t}dt = 0 \\
& \qquad\qquad [By\ \ Taylor\ \ series\ \ expansion] \\
\Rightarrow \quad & \frac{\partial F}{\partial x_1}dx_1 + \frac{\partial F}{\partial x_2}dx_2 + \frac{\partial F}{\partial x_3}dx_3 + \frac{\partial F}{\partial t}dt = 0 \quad (as\ \ F(x_1, x_2, x_3, t) = 0) \\
\Rightarrow \quad & \frac{\partial F}{\partial x_i}dx_i + \frac{\partial F}{\partial t}dt = 0 \\
\Rightarrow \quad & \left(\frac{\partial F}{\partial x_i}n_i\right)v_n dt + \frac{\partial F}{\partial t}dt = 0 \\
\Rightarrow \quad & v_n = \frac{-\frac{\partial F}{\partial t}}{n_i\frac{\partial F}{\partial x_i}}
\end{aligned}$$

But the direction cosines n_i of the normal to the surface $F(x_1, x_2, x_3, t) = 0$ are given by

$$n_i = \frac{1}{R}\frac{\partial F}{\partial x_i} \text{ where } R = \sqrt{\left(\frac{\partial F}{\partial x_1}\right)^2 + \left(\frac{\partial F}{\partial x_2}\right)^2 + \left(\frac{\partial F}{\partial x_3}\right)^2}$$

Therefore

$$v_n = \frac{-\frac{\partial F}{\partial t}}{\frac{1}{R}\frac{\partial F}{\partial x_i}\frac{\partial F}{\partial x_i}} = -\frac{\frac{\partial F}{\partial t}}{\frac{R^2}{R}} = -\frac{1}{R}\frac{\partial F}{\partial t}$$

Hence

$$\begin{aligned} & -\frac{1}{R}\frac{\partial F}{\partial t} = n_i v_i = \frac{1}{R}\frac{\partial F}{\partial x_i}v_i \\ \Rightarrow \quad & \frac{\partial F}{\partial t} + v_i\frac{\partial F}{\partial x_i} = 0 \\ \Rightarrow \quad & \frac{DF}{Dt} = 0 \end{aligned}$$

Hence the condition is necessary.

Sufficient Part:

Let $F(x_1, x_2, x_3, t) = 0$ satisfies the condition

$$\begin{aligned} & \frac{DF}{Dt} = 0 \\ i.e, \quad & \frac{\partial F}{\partial t} + v_1\frac{\partial F}{\partial x_1} + v_2\frac{\partial F}{\partial x_2} + v_3\frac{\partial F}{\partial x_3} = 0 \end{aligned} \tag{7.5.1}$$

which is the first order partial differential equation. The differential equation of the paths of the particle is given by

$$\frac{dx_1}{v_1} = \frac{dx_2}{v_2} = \frac{dx_3}{v_3} = dt$$

The integrals of these equation have the form $x_i = x_i(X_1, X_2, X_3, t)$ where X_i are three arbitrary constants which identify the particle.

Therefore the general solution of the Eq.(7.5.1) is given by $F = \Phi(X_1, X_2, X_3)$ where Φ is arbitrary function. This shows that when $F = 0$, a particle once on the surface remains on the surface throughout the motion.

Example: *Find the condition that* $\frac{x^2}{a^2}f_1(t) + \frac{y^2}{b^2}f_2(t) + \frac{z^2}{c^2}f_3(t) = 1$ *is a possible form of a boundary surface for an incompressible flow.*

Solution: The given surface can be written as

$$F(x,y,z,t) = \frac{x^2}{a^2}f_1(t) + \frac{y^2}{b^2}f_2(t) + \frac{z^2}{c^2}f_3(t) - 1 = 0$$

Now the surface $F(x,y,z,t) = 0$ can be a possible boundary surface, if it satisfies the boundary condition

$$\begin{aligned}
&\frac{\partial F}{\partial t} + u\frac{\partial F}{\partial x} + v\frac{\partial F}{\partial y} + w\frac{\partial F}{\partial z} = 0 \\
\Rightarrow\ &\frac{x^2}{a^2}f_1'(t) + \frac{y^2}{b^2}f_2'(t) + \frac{z^2}{c^2}f_3'(t) + u\frac{2x}{a^2}f_1 + v\frac{2y}{b^2}f_2 + w\frac{2z}{c^2}f_3 = 0 \\
\Rightarrow\ &\frac{2x}{a^2}f_1\left(u + \frac{x}{2}\frac{f_1'}{f_1}\right) + \frac{2y}{a^2}f_2\left(v + \frac{y}{2}\frac{f_2'}{f_2}\right) + \frac{2z}{c^2}f_3\left(w + \frac{z}{2}\frac{f_3'}{f_3}\right) = 0
\end{aligned}$$

This equation is identically satisfied if we take

$$\begin{aligned}
&u + \frac{x}{2}\frac{f_1'}{f_1} = 0, \quad v + \frac{y}{2}\frac{f_2'}{f_2} = 0, \quad w + \frac{z}{2}\frac{f_3'}{f_3} = 0 \\
\Rightarrow\ &u = -\frac{x}{2}\frac{f_1'}{f_1}, \quad v = -\frac{y}{2}\frac{f_2'}{f_2}, \quad w = -\frac{z}{2}\frac{f_3'}{f_3}
\end{aligned}$$

which are the expressions for velocity components. Since the flow is incompressible, so u, v, w satisfy the equation of continuity

$$\begin{aligned}
&\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0 \\
\Rightarrow\ &-\frac{1}{2}\left[\frac{f_1'}{f_1} + \frac{f_2'}{f_2} + \frac{f_3'}{f_3}\right] = 0 \\
\Rightarrow\ &\log(f_1 f_2 f_3) = \text{constant} = k \text{ (say)}.
\end{aligned}$$

7.6 Material derivative of volume integral

Let $I(t)$ be the volume integral of a continuously derivable function $A(x,y,z,t)$ which may be density, pressure, components of velocity for any physical quantity defined over volume V occupied by a given set of particles at time t. Therefore

$$I(t) = \iiint_V A(x,y,z,t)\, dx\, dy\, dz$$

The rate of change of $I(t)$ w.r.t time denoted by $\dfrac{dI}{dt}$ or $\dfrac{DI}{Dt}$ is called the material derivative of I and is defined for a given set of particles.

The expression of $\dfrac{DI}{Dt}$ can be deduced as

$$I(t) = \iiint_V \left[\frac{DA}{Dt} + A\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}\right)\right] dx\, dy\, dz$$

7.7 Conservation of mass

Every material body, as well as every portion of such a body is possessed with a non-negative scalar measure, called the mass of the body or of the portion under consideration. Physically, the mass is associated with the inertial property of the body, i.e., its tendency to resist a change in motion. The measure of mass may be a function of the space variables and time.

If Δm is the mass of a small volume ΔV inn the current configuration and if we assume that Δm is absolutely continuous, the limit

$$I(t) = \lim_{\Delta V \to 0} \frac{\Delta m}{\Delta V}$$

define the scalar field $\rho = \rho(x, t)$ called the mass density of the body for that configuration at time t. Therefore the mass m of the entire body is given by

$$m = \iiint_V \rho(x, t) dV$$

The law of conservation of mass asserts that the total mass of any portion of a continuum medium remains unchanged during the motion, i.e., remains constant in every configuration.

7.8 Equation of continuity in Lagrangian Method

In this method the principle of conservation of mass in the form that the mass of a specific portion of the moving continuum enclosed in a volume does not change as it move.

Consider a specific portion of the continuum occupying at the initial instant $t = 0$ an arbitrary volume V_0 in the undeformed state. Let $P_0(X_1, X_2, X_3)$ be any point in it and $\rho_0 = \rho(X_1, X_2, X_3)$ be the density at P_0. Let dV_0 be the element of the volume at P_0 and mass of this element is $\rho_0 dV_0$. The total mass of the continuum which fills the volume V_0 at $t = 0$ is $\iiint_{V_0} \rho_0 \, dV_0$. At subsequent time $t > 0$, different particles of the continuum forming the volume V_0 move in such a manner that they form some other volume V in the deformed state. Let the particle at the initial position P_0 occupy the subsequent position $P(x_1, x_2, x_3)$ in V and let ρ be the density of the medium at P.

In Lagrangian description we have

$$x_i = x_i(X_1, X_2, X_3, t) \quad \text{and} \quad \rho = \rho(X_1, X_2, X_3, t)$$

The total mass of the continuum which fills the volume V at time t is $\iiint_V \rho \, dV$. Also we know that

$$dV = J \, dV_0 = \frac{\partial(x_1, x_2, x_3)}{\partial(X_1, X_2, X_3)} dV_0$$

is the Jacobian. By principle of conservation of mass, the masses of the material within these two volume must be equal. Therefore

$$\begin{aligned}
&\iiint_{V_0} \rho_0 dV_0 = \iiint_V \rho dV \\
&\Rightarrow \iiint_{V_0} \rho_0 dV_0 = \iiint_{V_0} \rho \frac{\partial(x_1, x_2, x_3)}{\partial(X_1, X_2, X_3)} dV_0 \\
&\Rightarrow \iiint_{V_0} \left[\rho_0 - \rho \frac{\partial(x_1, x_2, x_3)}{\partial(X_1, X_2, X_3)} \right] dV_0 = 0
\end{aligned}$$

Since the volume V_0 is arbitrary, it follows that

$$\begin{aligned}
&\rho_0 - \rho \frac{\partial(x_1, x_2, x_3)}{\partial(X_1, X_2, X_3)} = 0 \\
&\Rightarrow \rho_0 = \rho J
\end{aligned}$$

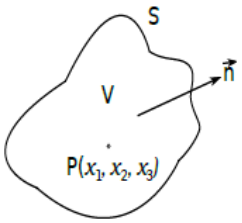
which implies that ρJ is independent of time t. Therefore

$$\frac{d}{dt}(\rho J) = 0$$

which is the equation of continuity in Lagrangian form.

7.9 Equation of continuity in Eulerian Method

In this method the principal of conservation of mass is expressed in the form that the rate at which the mass of the continuum within any fixed closed surface increases is equal to the rate at which the net mass of the continuum flows in across the boundary surface.



Let $P(x_1, x_2, x_2)$ be any point in the continuum and ρ be the density of the material at time t, so that $\rho = \rho(x_1, x_2, x_3, t)$. Also let $\vec{v}$ be the velocity at this point with components v_1, v_2, v_3.

Let us describe a closed surface $'s'$ including the point P and let V be the volume within S. Let $\vec{n}$ be the outward drawn normal to the surface S at any point on it. Then the normal

component of velocity $\vec{v}$ along the direction of $\vec{n}$ is $\vec{n} \cdot \vec{v}$. Therefore, the rate at which the material is entering within the volume bounded by S across the boundary surface is equal to

$$-\iint_S \rho\vec{n} \cdot \vec{v}\, dS$$

Also the rate at which the material is accumulating within the volume is $\iiint_V \frac{\partial \rho}{\partial t} dV$ where dV is the elementary volume. Now from the principle of conservation of mass these two rates are equal. Therefore,

$$\begin{aligned}
&\iiint_V \frac{\partial \rho}{\partial t} dV = -\iint_S \rho\vec{n} \cdot \vec{v} dS \\
&\Rightarrow \iiint_V \frac{\partial \rho}{\partial t} dV = -\iint_S \vec{n} \cdot (\rho\vec{v}) dS \\
&\Rightarrow \iiint_V \frac{\partial \rho}{\partial t} dV = -\iiint_V \vec{\nabla} \cdot (\rho\vec{v}) dV \text{ (Applying Gauss divergence theorem)} \\
&\Rightarrow \iiint_V \left[\frac{\partial \rho}{\partial t} + \vec{\nabla} \cdot (\rho\vec{v})\right] dV = 0
\end{aligned}$$

This is true for any volume V which contains the point P in its interior. Making the dimension of the volume tends to zero in a manner so as to enclose the point P always. Therefore the integrated must vanish at P. Hence

$$\begin{aligned}
&\frac{\partial \rho}{\partial t} + \vec{\nabla} \cdot (\rho\vec{v}) = 0 \\
&\Rightarrow \frac{\partial \rho}{\partial t} + \vec{v} \cdot (\vec{\nabla}\rho) + \rho\vec{\nabla} \cdot \vec{v} = 0 \\
&\Rightarrow \left[\frac{\partial \rho}{\partial t} + v_1\frac{\partial \rho}{\partial x_1} + v_2\frac{\partial \rho}{\partial x_2} + v_3\frac{\partial \rho}{\partial x_3}\right] + \rho\left[\frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \frac{\partial v_3}{\partial x_3}\right] = 0 \\
&\Rightarrow \frac{D\rho}{Dt} + \rho\left[\frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \frac{\partial v_3}{\partial x_3}\right] = 0 \qquad (7.9.1)
\end{aligned}$$

where $\frac{D}{Dt} \equiv$ differentiation following the motion of the continuum.

This Eq.(7.9.1) is the Euler's equation of motion, known as equation of continuity sometimes called equation of conservation of mass.

For incompressible material, the material derivaive of the density is zero i.e., $\frac{D\rho}{Dt} = 0$. In this case the equation of continuity becomes

$$\begin{aligned}
&\frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \frac{\partial v_3}{\partial x_3} = 0 \\
&\Rightarrow \vec{\nabla} \cdot \vec{v} = 0
\end{aligned}$$

7.10 Equivalence of equation of continuity in Lagrangian and Eulerian form

In Lagrangian form the equation of continuity is

$$\frac{d}{dt}(\rho J) = 0 \quad where \quad J = \frac{\partial(x_1, x_2, x_3}{\partial(X_1, X_2, X_3)}$$
$$\Rightarrow J\frac{d\rho}{dt} + \rho\frac{dJ}{dt} = 0 \tag{7.10.1}$$

If we want to pass from Lagrangian to Eulerian form we replace $\frac{d}{dt}$ by $\frac{D}{Dt}$ and $\dot{x_1}, \dot{x_2}, \dot{x_3}$ by v_1, v_2, v_3 respectively. Now

$$\begin{aligned}\frac{dJ}{dt} &= \frac{d}{dt}\begin{vmatrix}\frac{\partial x_1}{\partial X_1} & \frac{\partial x_2}{\partial X_1} & \frac{\partial x_3}{\partial X_1}\\ \frac{\partial x_1}{\partial X_2} & \frac{\partial x_2}{\partial X_2} & \frac{\partial x_3}{\partial X_2}\\ \frac{\partial x_1}{\partial X_3} & \frac{\partial x_2}{\partial X_3} & \frac{\partial x_3}{\partial X_3}\end{vmatrix}\\ &= \begin{vmatrix}\frac{\partial}{\partial X_1}\left(\frac{dx_1}{dt}\right) & \frac{\partial x_2}{\partial X_1} & \frac{\partial x_3}{\partial X_1}\\ \frac{\partial}{\partial X_2}\left(\frac{dx_1}{dt}\right) & \frac{\partial x_2}{\partial X_2} & \frac{\partial x_3}{\partial X_2}\\ \frac{\partial}{\partial X_3}\left(\frac{dx_1}{dt}\right) & \frac{\partial x_2}{\partial X_3} & \frac{\partial x_3}{\partial X_3}\end{vmatrix} + \begin{vmatrix}\frac{\partial x_1}{\partial X_1} & \frac{\partial}{\partial X_1}\left(\frac{dx_2}{dt}\right) & \frac{\partial x_3}{\partial X_1}\\ \frac{\partial x_1}{\partial X_2} & \frac{\partial}{\partial X_2}\left(\frac{dx_2}{dt}\right) & \frac{\partial x_3}{\partial X_2}\\ \frac{\partial x_1}{\partial X_3} & \frac{\partial}{\partial X_3}\left(\frac{dx_2}{dt}\right) & \frac{\partial x_3}{\partial X_3}\end{vmatrix} + \begin{vmatrix}\frac{\partial x_1}{\partial X_1} & \frac{\partial x_2}{\partial X_1} & \frac{\partial}{\partial X_1}\left(\frac{dx_3}{dt}\right)\\ \frac{\partial x_1}{\partial X_2} & \frac{\partial x_2}{\partial X_2} & \frac{\partial}{\partial X_2}\left(\frac{dx_3}{dt}\right)\\ \frac{\partial x_1}{\partial X_3} & \frac{\partial x_2}{\partial X_3} & \frac{\partial}{\partial X_3}\left(\frac{dx_3}{dt}\right)\end{vmatrix}\end{aligned} \tag{7.10.2}$$

Putting $\frac{dx_1}{dt} = v_1$, first determinant of (7.10.2) becomes

$$\begin{aligned}\frac{\partial(v_1, x_2, x_3)}{\partial(X_1, X_2, X_3)} &= \frac{\partial(v_1, x_2, x_3)}{\partial(x_1, x_2, x_3)} \cdot \frac{\partial(x_1, x_2, x_3)}{\partial(X_1, X_2, X_3)}\\ &= J \cdot \begin{vmatrix}\frac{\partial v_1}{\partial X_1} & 0 & 0\\ \frac{\partial v_1}{\partial X_2} & 1 & 0\\ \frac{\partial v_1}{\partial X_3} & 0 & 1\end{vmatrix}\\ &= J \cdot \frac{\partial v_1}{\partial x_1}\end{aligned}$$

Similarly 2^{nd} and 3^{rd} determinant of (7.10.2) are $J\frac{\partial v_2}{\partial x_2}$, $J\frac{\partial v_3}{\partial x_3}$ respectively. So equation (7.5.1) becomes

$$J\frac{D\rho}{Dt} + \rho J\left(\frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \frac{\partial v_3}{\partial x_3}\right) = 0$$
$$\Rightarrow \frac{D\rho}{Dt} + \rho\left(\frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \frac{\partial v_3}{\partial x_3}\right) = 0 \tag{7.10.3}$$

which is the equation of continuity in Eulerian form.

7.11 Few Probable Questions

(a) Given Eulerian velocity distribution at any time t in a fluid is

$$\mathbf{q} = (-z + \cos at)\mathbf{j} + (y + \sin at)\mathbf{k},$$

a is a constant ($\neq \pm 1$). Find the streamlines and pathlines. Discuss the

$$(E_{ij}) = \begin{bmatrix} 5 & -1 & -1 \\ -1 & 4 & 0 \\ -1 & 0 & 4 \end{bmatrix}$$

Determine principal strains. Find strain invariants from them. Show the equivalence of strain invariants. [Ans: $\theta_1 = 13,\ \theta_2 = 54,\ \theta_3 = 72,\ E_1 = 6,\ E_2 = 4,\ E_3 = 3$].

(b) Show that

$$\frac{x^2}{a^2} f(t) + \frac{y^2}{b^2}\phi(t) + \frac{z^2}{c^2}\psi(t) = 1$$

is a possible form of the boundary surface if $f(t)\phi(t)\psi(t) = 1$.

(c) Show that

$$\frac{x^2}{a^2}\tan^2 t + \frac{y^2}{b^2}\cot^2 t = 1$$

is a possible form for the boundary surface of a liquid. Find an expression for the normal velocity.

(d) Show that the variable ellipsoid

$$\frac{x^2}{a^2 e^{-t}\cos(t + \pi/4)} + \frac{y^2}{b^2 e^{t}\sin(t + \pi/4)} + \frac{z^2}{c^2 \sec 2t} = 1$$

is a possible form of boundary surface of a liquid at any time t. Determine the velocity $\mathbf{q}$ of any particle on this velocity. Show that the equation of continuity is satisfied.

(e) Show that the ellipsoid

$$\frac{x^2}{a^2 k^2 t^{2n}} + kt^n \left\{ \left(\frac{y^2}{b^2} + \frac{z^2}{c^2} \right) \right\} = 1$$

is a possible form of the boundary surface of a liquid.

Unit 8

Course Structure

- Momentum principles and equation of motion
- Energy balance, laws of thermodynamics
- Constitutive equation: Generalised Hooke's law
- Isotropy and elastic moduli, Stress-Strain relation

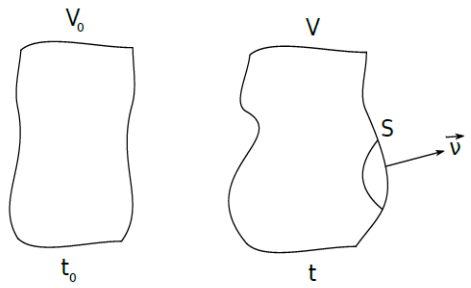
8.1 Momentum principles and equation of motion

The principle of balance of linear momentum states that the time rate of change of linear momentum of any portion of a continuum in motion is equal to the total applied force acting on that portion.

8.1.1 Equation of motion of a continuum applying the principle of linear momentum:

Let a given set of material particle of a continuum occupy the volume V_0 at time t_0 which now occupy the volume V at time t. Let S be the boundary surface of V. Let (v_1, v_2, v_3) be the components of the velocity and ρ be the density at any point (x_1, x_2, x_3) within V at time t. Let $\rho X_1, \rho X_2, \rho X_3$ be the components of the body force per unit volume of the material within V and $\tau_{\nu_{x_1}}, \tau_{\nu_{x_2}}, \tau_{\nu_{x_3}}$ are the components of surface force at any point on S where $\vec{\nu}$ is the outward drawn normal to the surface whore direction cosines are l, m, n.

By Newton's 2^{nd} law the rate of change of components of the linear momentum of material



within V in any direction must be equal to the resultant of the forces on the material in V in that direction. By considering components in the x-direction we have

$$\underbrace{\frac{D}{Dt}\iiint_V \rho v_1 dV}_{Change\ in\ momentum\ in\ x_1-direction} = \underbrace{\iiint_V \rho X_1 dV}_{Body\ force} + \underbrace{\iint_S \tau_{\nu_{x_1}} dS}_{Surface\ force} \tag{8.1.1}$$

$$\Rightarrow \iiint_V \left[\frac{D}{Dt}(\rho v_1) + \rho v_1\left(\frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \frac{\partial v_3}{\partial x_3}\right)\right] dV = \iiint_V \rho X_1 dV + \iint_S (l\tau_{x_1x_1} + m\tau_{x_2x_1} + n\tau_{x_3x_1}) dS \tag{8.1.2}$$

Now by Gauss Theorem

$$\iint_S (l\tau_{x_1x_1} + m\tau_{x_2x_1} + n\tau_{x_3x_1}) dS = \iiint_V \left[\frac{\partial}{\partial x_1}(\tau_{x_1x_1}) + \frac{\partial}{\partial x_2}(\tau_{x_2x_1}) + \frac{\partial}{\partial x_3}(\tau_{x_3x_1})\right] dV$$

Also

$$\begin{aligned}
&\frac{D}{Dt}(\rho v_1) + \rho v_1\left[\frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \frac{\partial v_3}{\partial x_3}\right] \\
&= \left[\frac{\partial}{\partial t}(\rho v_1) + v_1\frac{\partial}{\partial x_1}(\rho v_1) + v_2\frac{\partial}{\partial x_2}(\rho v_1) + v_3\frac{\partial}{\partial x_3}(\rho v_1)\right] + \rho v_1\left[\frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \frac{\partial v_3}{\partial x_3}\right] \\
&= \rho\left[\frac{\partial v_1}{\partial t} + v_1\frac{\partial v_1}{\partial x_1} + v_2\frac{\partial v_1}{\partial x_2} + v_3\frac{\partial v_1}{\partial x_3}\right] + v_1\left[\frac{\partial \rho}{\partial t} + v_1\frac{\partial \rho}{\partial x_1} + v_2\frac{\partial \rho}{\partial x_2} + v_3\frac{\partial \rho}{\partial x_3}\right] \\
&\qquad + \rho v_1\left[\frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \frac{\partial v_3}{\partial x_3}\right] \\
&= \rho\frac{Dv_1}{Dt} + v_1\left[\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x_1}(\rho v_1) + \frac{\partial}{\partial x_2}(\rho v_2) + \frac{\partial}{\partial x_3}(\rho v_3)\right] \\
&= \rho\frac{Dv_1}{Dt} \text{ (As the quantity in the third bracket is zero by the equation of continuity)}
\end{aligned}$$

The equation (8.1.2) becomes

$$\iiint_V \left[\rho\frac{Dv_1}{Dt} - \rho X_1 - \left(\frac{\partial \tau_{x_1x_1}}{\partial x_1} + \frac{\partial \tau_{x_2x_1}}{\partial x_2} + \frac{\partial \tau_{x_3x_1}}{\partial x_3}\right)\right] dV = 0$$

Since the equation holds for any arbitrary volume V, so the integral must vanish. Therefore

$$\rho\frac{Dv_1}{Dt} = \frac{\partial \tau_{x_1x_1}}{\partial x_1} + \frac{\partial \tau_{x_2x_1}}{\partial x_2} + \frac{\partial \tau_{x_3x_1}}{\partial x_3} + \rho X_1 \tag{8.1.3}$$

Similarly considering momentum and forces in x_2 and x_3 directions respectively, we obtain other two relations as below

$$\rho\frac{Dv_2}{Dt} = \frac{\partial \tau_{x_1x_2}}{\partial x_1} + \frac{\partial \tau_{x_2x_2}}{\partial x_2} + \frac{\partial \tau_{x_3x_2}}{\partial x_3} + \rho X_2 \tag{8.1.4}$$

$$\rho\frac{Dv_3}{Dt} = \frac{\partial \tau_{x_1x_3}}{\partial x_1} + \frac{\partial \tau_{x_2x_3}}{\partial x_2} + \frac{\partial \tau_{x_3x_3}}{\partial x_3} + \rho X_3 \tag{8.1.5}$$

Equations (8.1.3), (8.1.4), (8.1.5) are called the Eulerian equation of motion of continuum or the stress equation of motion of continuum. Writtin $\tau_{11}, \tau_{22}, \tau_{12}$ etc. for $\tau_{x_1x_1}, \tau_{x_2x_2}, \tau_{x_1x_2}$ etc. equations (8.1.3), (8.1.4), (8.1.5) together can be written as

$$\rho\frac{Dv_i}{Dt} = \frac{\partial \tau_{ji}}{\partial x_j} + \rho X_i \;\; (i, j = 1, 2, 3) \tag{8.1.6}$$

This equation is known as the Cauchy equation of motion for a continuum.

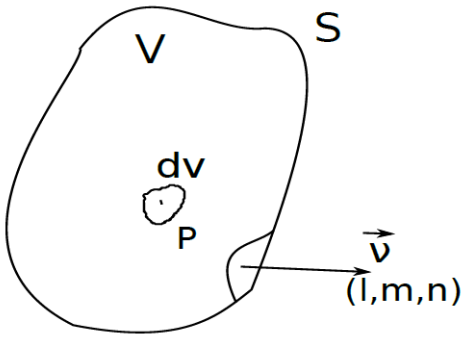
8.2 Energy balance, Laws of Thermodynamics

8.2.1 Principle of conservation of Energy

Let a given set of material particles of the continuum which occupied the volume V_0 at time $t = 0$, occupy the volume V at time t. Let S be the boundary surface of V.

Let v_1, v_2, v_3 be the components of velocity and ρ be the density at any point (x_1, x_2, x_3) with in V at time t. $\rho X_1, \rho X_2, \rho X_3$ are the components of body force. $(\rho\vec{F})$ per unit volume of the material within V and $\tau_{\nu_1}, \tau_{\nu_1}, \tau_{\nu_1}$ are the components of surface stress vector at any point on S where $\vec{\nu}$ is the outward drawn normal to the surface S with direction cosines l, m, n.

The principle of conservation of energy, also known as *first law of thermodynamics*, states that time rate of change of kinetic and internal energy (total energy) of the material within V must be equal to the rate of work done by the body and surface forces plus any non-mechanical energy supplied to the material within V per unit time. (Non-mechanical energy may include thermal, chemical or eletromagnetical energy, but we shall only consider the thermal energy



change.) So if K and E be the K.E and internal energy respectively of the material within V then energy principle gives

$$\frac{DK}{Dt} + \frac{DE}{dt} = \text{rate of work done by the body and surface forces} + \text{rate of increase of total heat within the material in V.} \tag{8.2.1}$$

Now

$$\begin{aligned}
\frac{DK}{Dt} &= \frac{D}{Dt}\iiint_V \frac{1}{2}\rho q^2 dV \quad where \quad q^2 = v_1^2 + v_2^2 + v_3^2 \\
&= \iiint_V \left[\frac{D}{Dt}\left(\frac{1}{2}\rho q^2\right) + \frac{1}{2}\rho q^2\left(\frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \frac{\partial v_3}{\partial x_3}\right)\right] dV \\
&= \iiint_V \left[\frac{1}{2}q^2\frac{D\rho}{Dt} + \frac{\rho}{2}\frac{D}{Dt}(q^2) + \frac{1}{2}q^2\rho\frac{\partial v_i}{\partial x_i}\right] dV \\
&= \iiint_V \left[\frac{1}{2}q^2\left\{\frac{D\rho}{Dt} + \rho\frac{\partial v_i}{\partial x_i}\right\} + \frac{\rho}{2}\frac{D}{Dt}(v_1^2 + v_2^2 + V_3^2)\right] dV \\
&= \iiint_V \rho\left[v_1\frac{Dv_1}{Dt} + v_2\frac{Dv_2}{Dt} + v_3\frac{Dv_3}{Dt}\right] dV, \quad \text{since by equation of continuity}
\end{aligned}$$

Therefore,

$$\frac{DK}{Dt} = \iiint_V \rho\left[v_j\frac{Dv_j}{Dt}\right] dV \tag{8.2.2}$$

Let e be the internal energy per unit mass. Therefore,

$$\begin{aligned}
\frac{DE}{Dt} &= \frac{D}{Dt}\iiint_V \rho e dV \\
&= \iiint_V \left[\frac{D(\rho e)}{Dt} + \rho e\left\{\frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \frac{\partial v_3}{\partial x_3}\right\}\right] dV \\
&= \iiint_V \left[e\left\{\frac{D\rho}{Dt} + \rho\frac{\partial v_i}{\partial x_i}\right\} + \rho\frac{De}{Dt}\right] dV \\
&= \iiint_V \rho\frac{De}{Dt} dV \quad \text{since by equation of continuity} \quad \frac{D\rho}{Dt} + \rho\frac{\partial v_i}{\partial x_i} = 0 \quad (8.2.3)
\end{aligned}$$

Now the rate of work done by body and surface forces on the material within V

$$\begin{aligned}
&= \iiint_V [(\rho X_1)v_1 + (\rho X_2)v_2 + (\rho X_3)v_3]dV + \iint_S [\tau_{\nu_1}v_1 + \tau_{\nu_2}v_2 + \tau_{\nu_3}v_3]dS \\
&= \iiint_V X_j v_j \rho dV + \iint_S [(l\tau_{11} + m\tau_{21} + n\tau_{31})v_1 + (l\tau_{12} + m\tau_{22} + n\tau_{32})v_2 \\
&\qquad + (l\tau_{13} + m\tau_{23} + n\tau_{33})v_3]dS \\
&= \iiint_V X_j V_j \rho dV + \iint_S [l(\tau_{11}v_1 + \tau_{12}v_2 + \tau_{13}v_3) + m(\tau_{21}v_1 + \tau_{22}v_2 + \tau_{23}v_3) \\
&\qquad + n(\tau_{31}v_1 + \tau_{23}v_2 + \tau_{33}v_3)]dS \\
&= \iiint_V X_j v_j \rho dV + \iint_S [l\tau_{1j}v_j + m\tau_{2j}v_j + n\tau_{3j}v_j]dS \\
&= \iiint_V X_j v_j \rho dV + \iiint_V \frac{\partial}{\partial x_i}(\tau_{ij})dV \quad \text{(By Gauss theorem)} \\
&= \iiint_V X_i v_j \rho dV + \iiint_V \left[v_j\frac{\partial \tau_{ij}}{\partial x_i} + \tau_{ij}\frac{\partial v_j}{\partial x_i}\right] dV \qquad (8.2.4)
\end{aligned}$$

If h be the body heat energy or radiant heat energy generated per unit mass per unit time and $\vec{c} = c_1\hat{i} + c_2\hat{j} + c_3\hat{k}$ represents the flow of heat per unit area across a surface per unit time then the rate of increase of total heat energy in to the continuum enclosed in V is equal to

$$\begin{aligned}
&\iiint_V \rho h dV - \iint_S (lc_1 + mc_2 + nc_3)dS \\
&= \iiint_V \rho h dV - \iiint_V \frac{\partial c_i}{\partial x_i} dV \quad \text{(By Gauss Theorem)} \qquad (8.2.5)
\end{aligned}$$

Substituting the results of (8.2.2),(8.2.3),(8.2.4),(8.2.5) in equation (8.2.1) we get,

$$\begin{aligned}
\iiint_V v_j\frac{Dv_j}{Dt}dV + \iiint_V \rho\frac{De}{Dt}dV = \iiint_V \rho v_j X_j dV + \iiint_V \left[v_j\frac{\partial \tau_{ij}}{\partial x_i} + \tau_{ij}\frac{\partial v_j}{\partial x_i}\right] dV \\
+ \iiint_V \rho h dV - \iiint_V \frac{\partial c_i}{\partial x_i} dV \qquad (8.2.6)
\end{aligned}$$

$$\Rightarrow \iiint_V v_j \left[\rho \frac{Dv_j}{Dt} - \rho X_j - \frac{\partial \tau_{ij}}{\partial x_i}\right] dV + \iiint_V \left[\rho \frac{De}{Dt} - \rho h + \frac{\partial c_i}{\partial x_i} - \tau_{ij} \frac{\partial v_j}{\partial x_i}\right] dV = 0 \quad (8.2.7)$$

$$\Rightarrow \iiint_V \left[\rho \frac{De}{Dt} - \rho h + \frac{\partial c_i}{\partial x_i} - \tau_{ij} \frac{\partial v_j}{\partial x_i}\right] dV = 0 \quad (8.2.8)$$

[Since $\rho \frac{Dv_j}{Dt} = \rho X_j + \frac{\partial \tau_{ij}}{\partial x_i}$ by Cauchy equation of motion for a continuum]

Since integral is zero, for an arbitrary volume V so we must have

$$\rho \frac{De}{Dt} = \rho h - \frac{\partial c_i}{\partial x_i} + \tau_{ij} \frac{\partial v_j}{\partial x_i} \quad (8.2.9)$$

$$\text{Now } \tau_{ij} \frac{\partial v_j}{\partial x_i} = \frac{1}{2} \tau_{ij} \left(\frac{\partial v_j}{\partial x_i} + \frac{\partial v_i}{\partial x_j}\right) + \frac{1}{2} \tau_{ij} \left(\frac{\partial v_j}{\partial x_i} - \frac{\partial v_i}{\partial x_j}\right) \quad (8.2.10)$$

$$= \tau_{ij} d_{ij} + 0 \quad (8.2.11)$$

where $d_{ij} = \frac{1}{2}\left(\frac{\partial v_j}{\partial x_i} + \frac{\partial v_i}{\partial x_j}\right)$ is the strain rate tenor. The last term on the R.H.S of (8.2.10) vanishes because it is the product of a symmetric tensor and with an antisymmetric tensor. Using the result (8.2.11) in equation (8.2.9) we obtain the final form of the energy equation

$$\rho \frac{De}{Dt} = \rho h - \frac{\partial c_i}{\partial x_i} + \tau_{ij} d_{ij} \quad (8.2.12)$$

The scalar quantity $\tau_{ij} d_{ij}$ is called stress power.

8.3 Constitutive Equations

As equation which describes a property of a material is called a constitutive equation of that material. A stress strain relation describes a mechanical property of the material and therefore this relation is constitutive equation.

8.3.1 Generalized Hooke's Law

In a continuous medium, the state of stress is completely determined by the stress tensor τ_{ij} and the state of deformation by the strain tensor e_{ij}.

In general it may be expressed as

$$\begin{aligned} &T = f(e) \\ &i.e, \quad \tau_{ij} = f_{ij}(e_{kl}), \quad i,j,k,l = 1,2,3 \end{aligned}$$

where, f is a symmetric tensor valued function of various strain tensor e.

This is the generalized Hooke's law, which stated that at each point of a continuous medium at a fixed temperature each of the six components is a function of the six strain components.

Expanding the equation by Taylors series theorem, we obtain

$$\begin{aligned} \tau_{ij} &= f_{ij}(0,0,\cdots,0) + \left(\frac{\partial f_{ij}}{\partial e_{kl}}\right) e_{kl} + \cdots \\ &= b_{ij} + a_{ijkl}e_{kl} + \cdots \end{aligned}$$

For linear elastic body, the stresses are liner function of infinitesimal strains, the constitutive equation can be written as

$$\tau_{ij} = b_{ij} + a_{ijkl}e_{kl} \tag{8.3.1}$$

Since in the initial unstrained state, the body will be unstressed, $\tau_{ij} = 0$ when all $e_{ij} = 0$. Therefore we must have $f_{ij}(0,0,\cdots,0) = b_{ij} = 0$ for $i,j = 1,2,3$. Thus the constitutive equations for a linear elastic solid takes the form

$$\tau_{ij} = a_{ijkl}e_{kl}, \quad i,j,k,l = 1,2,3 \tag{8.3.2}$$

i.e,

$$\begin{aligned} \tau_{11} &= a_{1111}e_{11} + a_{1112}e_{12} + \cdots + a_{1133}e_{33} \\ \tau_{12} &= a_{1211}e_{11} + a_{1212}e_{12} + \cdots + a_{1233}e_{33} \\ &\cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \\ &\cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \\ \tau_{33} &= a_{3311}e_{11} + a_{3312}e_{12} + \cdots + a_{3333}e_{33} \end{aligned} \tag{8.3.3}$$

where the tensor of elastic coefficients a_{ijkl} has $3^4 = 81$ components.

However, due to the symmetry of both the stress and strain tensors, it is clear that

$$a_{ijkl} = a_{jikl} = a_{ijlk}$$

which reduces the 81 possibilities to 36 distinct coefficients at most.

The constitutive linear law for the relation (8.3.2) between stress and strain is known as the *generalized Hooke's law for linear elastic solid*. The coefficients a_{ijkl} are called the *elastic*

constant or *elastic moduli* and they are describing the elastic properties of the body.

Note: A particular case of relation (8.3.2) is

$$\tau_{11} = Ee_{11}$$

If the coefficient E is a constant then the above relation state that for a linear elastic solid the normal stress in x_1 direction is directly proportional to the normal strain in the same direction (within the strain limit). This relation is known as *Hooke's law.*

8.4 Isotropy and Elastic moduli

8.4.1 Constitutive equation of linearly elastic isotropic solid

A linearly elastic solid is said to be *isotropic* if it exhibits same elastic symmetry in all directions. For such materials, the constitutive equation has only two elastic constants.

Now for a linear elastic solid we have from generalized Hooke's law

$$\tau_{ij} = a_{ijkl}e_{kl}; \quad i,j,k,l = 1,2,3 \tag{8.4.1}$$

where a_{ijkl} are elastic constants. Now a_{ijkl} can be written as

$$\begin{aligned} a_{ijkl} &= \frac{1}{2}(a_{ijkl} - a_{ijlk}) + \frac{1}{2}(a_{ijkl} + a_{ijlk}) \\ &= b_{ijkl} + c_{ijkl} \end{aligned}$$

where $b_{ijkl} = \dfrac{1}{2}(a_{ijkl} - a_{ijlk})$ and $c_{ijkl} = \frac{1}{2}(a_{ijkl} + a_{ijlk})$.
Now

$$\begin{aligned} b_{ijkl} &= \frac{1}{2}(a_{ijkl} - a_{ijlk}) \\ &= -\frac{1}{2}(a_{ijkl} - a_{ijlk}) \\ &= b_{ijlk}, \end{aligned}$$

and $c_{ijkl} = \dfrac{1}{2}\dfrac{1}{2}(a_{ijkl} + a_{ijlk}) = c_{ijlk}$. Hence b_{ijkl} is skew-symmetric and c_{ijkl} is symmetric. Therefore we have

$$\begin{aligned} \tau_{ij} &= a_{ijkl}e_{kl} \\ &= b_{ijkl}e_{kl} + c_{ijkl}e_{kl} \\ &= 0 + c_{ijkl}e_{kl} \end{aligned}$$

Therefore

$$\tau_{ij} = c_{ijkl}e_{kl} \quad i,j,k,l = 1,2,3 \tag{8.4.2}$$

Since τ_{ij} and e_{ij} are second order tensors, therefore c_{ijkl} must be a tensor of order 4. Now for isotropic elastic medium the elastic constant c_{ijkl} remains the same under all orthogonal transformation of the coordinate axes. Now any fourth order tensor can be represented in the form

$$c_{ijkl} = \alpha\delta_{ij}\delta_{kl} + \beta\delta_{ik}\delta_{jl} + \gamma\delta_{il}\delta jk \quad \text{where } \alpha, \beta, \gamma \text{ are scalars.} \tag{8.4.3}$$

Also

$$c_{ijlk} = \alpha\delta_{ij}\delta_{lk} + \beta\delta_{il}\delta_{jk} + \gamma\delta_{ik}\delta jl \tag{8.4.4}$$

Since c_{ijkl} is symmetric therefore using the relation $c_{ijkl} = c_{ijlk}$, we have

$$\begin{aligned} &\alpha\delta_{ij}\delta_{kl} + \beta\delta_{ik}\delta_{jl} + \gamma\delta_{il}\delta jk = \alpha\delta_{ij}\delta_{lk} + \beta\delta_{il}\delta_{jk} + \gamma\delta_{ik}\delta jl \\ &\Rightarrow (\beta - \gamma)(\delta_{ik}\delta_{jl} - \delta_{il}\delta_{jk}) = 0 \end{aligned} \tag{8.4.5}$$

The relation (8.4.5) is true for all values of i, j, k, l. If we take $i = k = 1$ and $j = l = 2$, the equation (8.4.5) becomes

$$\begin{aligned} &(\beta - \gamma)(\delta_{11}\delta_{22} - \delta_{12}\delta_{21}) = 0 \\ &\Rightarrow (\beta - \gamma)(1 - 0) = 0 \\ &\Rightarrow \beta = \gamma \end{aligned}$$

Thus equation (8.4.3) can be written as

$$c_{ijkl} = \alpha\delta_{ij}\delta_{kl} + \beta(\delta_{il}\delta_{jk} + \delta_{ik}\delta jl) \tag{8.4.6}$$

Therefore from equation (8.4.2) we can write

$$\begin{aligned} \tau_{ij} &= \alpha\delta_{ij}\delta_{kl}e_{kl} + \beta(\delta_{il}\delta_{jk} + \delta_{ik}\delta_{jl})e_{kl} \\ &= \alpha\delta_{ij}e_{kk} + \beta(\delta_{il}e_{jl} + \delta_{ik}e_{kj}) \\ &= \alpha\delta_{ij}e_{kk} + \beta(e_{ji} + e_{ij}) \\ &= \alpha\delta_{ij}e_{kk} + 2\beta e_{ij} \quad [\text{as } e_{ij} = e_{ji}] \end{aligned}$$

This can be written as

$$\tau_{ij} = \lambda\theta\delta_{ij} + 2\mu e_{ij} \tag{8.4.7}$$

where $\theta = e_{kk} = e_{11} + e_{22} + e_{33}$, $\lambda = \alpha$ and $\mu = \beta$. These relation represents the generalized Hooke's law for a linear isotropic elastic solid. The constants λ and μ are called *Lame constant* or *Lame moduli*. The relation (8.4.7) can also be rewritten in the form of a matrix equation as follows

$$\begin{bmatrix} \tau_{11} \\ \tau_{22} \\ \tau_{33} \\ \tau_{12} \\ \tau_{23} \\ \tau_{31} \end{bmatrix} = \begin{bmatrix} \lambda + 2\mu & \lambda & \lambda & 0 & 0 & 0 \\ \lambda & \lambda + 2\mu & \lambda & 0 & 0 & 0 \\ \lambda & \lambda & \lambda + 2\mu & 0 & 0 & 0 \\ 0 & 0 & 0 & 2\mu & 0 & 0 \\ 0 & 0 & 0 & 0 & 2\mu & 0 \\ 0 & 0 & 0 & 0 & 0 & 2\mu \end{bmatrix} \begin{bmatrix} e_{11} \\ e_{22} \\ e_{33} \\ e_{12} \\ e_{23} \\ e_{31} \end{bmatrix}$$

Also we can write

$$\begin{aligned}
\tau_{11} &= \lambda\theta + 2\mu\frac{\partial u_1}{\partial x_1} \\
\tau_{22} &= \lambda\theta + 2\mu\frac{\partial u_2}{\partial x_2} \\
\tau_{33} &= \lambda\theta + 2\mu\frac{\partial u_3}{\partial x_3} \\
\tau_{12} &= \mu\left(\frac{\partial u_1}{\partial x_2} + \frac{\partial u_2}{\partial x_1}\right) \\
\tau_{23} &= \mu\left(\frac{\partial u_1}{\partial x_3} + \frac{\partial u_1}{\partial x_3}\right) \\
\tau_{31} &= \mu\left(\frac{\partial u_1}{\partial x_3} + \frac{\partial u_1}{\partial x_3}\right)
\end{aligned} \tag{8.4.8}$$

where $\theta = \dfrac{\partial u_1}{\partial x_1} + \dfrac{\partial u_2}{\partial x_2} + \dfrac{\partial u_3}{\partial x_3}$.

Example 8.4.1. Find the strain and stress components for the displacement field given by $u_1 = 3x_1x_2^2$, $u_2 = 2x_1x_3$, $u_3 = x_3^2 - x_1x_2$.

Solution: Let us take the principle direction of strain at some point of the body as coordinate axes. Let e_{ij} be strain tensor and τ_{ij} be stress tensor at that point. Then

$$\begin{aligned}
e_{11} &= \frac{\partial u_1}{\partial x_1} = 3x_1^2, \quad e_{22} = \frac{\partial u_2}{\partial x_2} = 0, \quad e_{33} = \frac{\partial u_3}{\partial x_3} = 2x_3, \\
e_{23} &= \frac{1}{2}\left(\frac{\partial u_2}{\partial x_3} + \frac{\partial u_3}{\partial x_2}\right) = x_1, \\
e_{31} &= \frac{1}{2}\left(\frac{\partial u_1}{\partial x_3} + \frac{\partial u_3}{\partial x_1}\right) = -x_2, \\
e_{12} &= \frac{1}{2}\left(\frac{\partial u_1}{\partial x_2} + \frac{\partial u_2}{\partial x_1}\right) = 6x_1x_2 + 2x_3
\end{aligned} \tag{8.4.9}$$

Therefore $\theta = e_{11} + e_{22} + e_{33} = 3x_2^2 + 2x_3$.

The constitutive equation for isotropic elastic body given by the following equation

$$\tau_{ij} = \lambda\theta\delta_{ij} + 2\mu e_{ij}$$

where λ and μ are Lame's constant.

Therefore, the stress components are

$$\begin{aligned}
\tau_{11} &= \lambda\theta\delta_{11} + 2\mu e_{11} = (3\lambda + 6\mu)x_2^2 + 2\lambda x_3 \\
\tau_{22} &= \lambda\theta\delta_{22} + 2\mu e_{22} = \lambda(3x_2^2 + 2x_3) \\
\tau_{33} &= \lambda\theta\delta_{33} + 2\mu e_{33} = 3\lambda x_2^2 + (2\lambda + 4\mu)x_3 \\
\tau_{12} &= \lambda\theta\delta_{12} + 2\mu e_{12} = 2\mu e_{12} = 2\mu(6x_1x_2 + 2x_3 \\
\tau_{23} &= \lambda\theta\delta_{23} + 2\mu e_{23} = 2\mu e_{23} = 2\mu x_1 \\
\tau_{31} &= \lambda\theta\delta_{31} + 2\mu e_{31} = 2\mu e_{31} = -2\mu x_2
\end{aligned}$$

Note 8.4.2. The principle directions of strain at each point of a linearly elastic isotropic body are coincident with the principle directions of stress.

Proof: Let us take the principle direction of strain at some point of the body as coordinate axes. Let e_{ij} be strain tensor and τ_{ij} be stress tensor at that point. Then $e_{31} = 0,\ e_{12} = 0,\ e_{23} = 0.$

Now from constitutive equation for isotropic linearly elastic solid body is

$$\tau_{ij} = \lambda\theta\delta_{ij} + 2\mu e_{ij}$$

Thus

$$\begin{aligned}
\tau_{12} &= 2\mu e_{12} = 0 \\
\tau_{23} &= 2\mu e_{23} = 0 \\
\tau_{31} &= 2\mu e_{31} = 0
\end{aligned}$$

Hence coordinate axes must be along the principle directions of stress.

8.5 Strains in terms of Stresses

From stress-strain relations for isotropic linear elastic solid, we have

$$\tau_{ij} = \lambda\theta\delta_{ij} + 2\mu e_{ij} \tag{8.5.1}$$

where λ and μ are Lame's constant.

Putting $j = i$ and summing for $i = 1, 2, 3$ we get

$$\begin{aligned}
\tau_{ij} &= 3\lambda\theta + 2\mu e_{ii} \\
&= 3\lambda\theta + 2\mu\theta \\
\Theta &= (3\lambda + 2\mu)\theta
\end{aligned}$$

Hence

$$\theta = \frac{\Theta}{(3\lambda + 2\mu)} \tag{8.5.2}$$

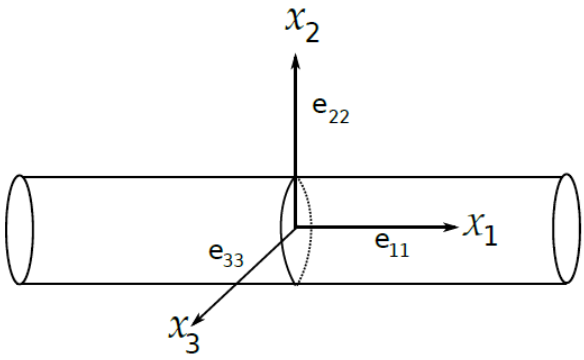
where $\Theta = \tau_{kk}$ =sum of normal stresses=$\tau_{11} + \tau_{22} + \tau_{33}$. Also we have

$$\begin{aligned} e_{ij} &= \frac{\tau_{ij}}{2\mu} - \frac{\lambda\theta\delta_{ij}}{2\mu} \\ \Rightarrow e_{ij} &= \frac{\tau_{ij}}{2\mu} - \frac{\lambda}{2\mu}\left(\frac{\Theta}{3\lambda+2\mu}\right)\delta_{ij} \\ & \quad \text{for } \mu \neq 0 \text{ and } 3\lambda + 2\mu \neq 0;\ \Theta = \tau_{kk} \end{aligned} \tag{8.5.3}$$

Equation (8.5.3) is the *inversion of Hooke's law* and give us the strain stress relation.

8.6 Elastic Moduli

In order to find the physical meaning of the elastic constants/moduli appearing in the Hooke's law, we consider the following particular cases.



Suppose that the stress tensor has only one non zero component τ_{11}. Such a stress system occurs in a beam placed along the x_1 axis and subjected to a longitudinal stress. Then from stress strain relations,

$$e_{ij} = \frac{1}{2\mu}\left[\tau_{ij} - \frac{\lambda}{3\lambda+2\mu}\delta_{ij}\tau_{kk}\right]$$

we have

$$\begin{aligned} e_{11} &= \frac{1}{2\mu}\left[\tau_{11} - \frac{\lambda}{3\lambda+2\mu}\delta_{11}\tau_{11}\right] = \frac{\lambda+\mu}{\mu(3\lambda+2\mu)}\tau_{11} \\ e_{22} &= \frac{1}{2\mu}\left[\tau_{22} - \frac{\lambda}{3\lambda+2\mu}\delta_{22}(\tau_{11}+\tau_{22}+\tau_{33})\right] = -\frac{\lambda}{2\mu(3\lambda+2\mu)}\tau_{11} \\ e_{33} &= \frac{1}{2\mu}\left[\tau_{33} - \frac{\lambda}{3\lambda+2\mu}\delta_{33}(\tau_{11}+\tau_{22}+\tau_{33})\right] = -\frac{\lambda}{2\mu(3\lambda+2\mu)}\tau_{11} \\ e_{12} &= e_{23} = e_{31} = 0 \end{aligned}$$

If we set

$$E = \frac{\mu(3\lambda + 2\mu)}{\lambda + \mu} \tag{8.6.1}$$

$$\text{and } \sigma = \frac{\lambda}{2(\lambda + \mu)} \tag{8.6.2}$$

Then we have

$$\frac{\tau_{11}}{e_{11}} = E \quad \text{and} \quad \frac{e_{22}}{e_{11}} = \frac{e_{33}}{e_{11}} = -\sigma$$

Since τ_{11} represents tension, $\tau_{11} > 0$ also tensile stress will produce an extension in the direction of the axis of cylinder and a contraction in the cross-section. Thus $\tau_{11} > 0$ implies $e_{11} > 0$ and $e_{22},\ e_{33} < 0$. It follows that $E > 0$ and $\sigma > 0$.

$$\text{Therefore, } E = \frac{\tau_{11}}{e_{11}} = \frac{\text{longitudinal stress}}{\text{longitudinal strain}}$$

This constant E is called *Young's Modulus*.

Also

$$\sigma = \left|\frac{e_{22}}{e_{11}}\right| = \left|\frac{e_{33}}{e_{11}}\right| = \text{ratio of the contraction of the linear element}$$

is a transverse direction to the corresponding extension in the longitudinal direction. This ratio is known as *Poissons ratio.*

Note: Solving the relation (8.6.1) and (8.6.2) for λ and μ we get

$$1 + \sigma = 1 + \frac{\lambda}{2(\lambda + \mu)} = \frac{3\lambda + 2\mu}{2(\lambda + \mu)} = \frac{E}{2\mu}$$
$$\Rightarrow \mu = \frac{E}{2(1 + \sigma)} > 0, \quad \text{since } E,\ r > 0.$$

Now

$$1 - 2\sigma = 1 - \frac{2\lambda}{2(\lambda + \mu)} = \frac{\mu}{\lambda + \mu}, \quad 1 + \sigma = \frac{3\lambda + 2\mu}{2(\lambda + \mu)}$$
$$(1 - 2\sigma)(1 + \sigma) = \frac{\mu}{\lambda + \mu}\frac{3\lambda + 2\mu}{2(\lambda + \mu)} = \frac{E}{2\mu}\frac{\mu}{\lambda + \mu} = \frac{E}{\lambda}\sigma \Rightarrow \lambda = \frac{E\sigma}{(1 + \sigma)(1 - 2\sigma)}$$

8.7 Stress-Strain relation in terms of E and σ

We know that

$$\tau_{ij} = \lambda\theta\delta_{ij} + 2\mu e_{ij} \tag{8.7.1}$$

$$\text{But} \quad \lambda = \frac{E\sigma}{(1+\sigma)(1-2\sigma)} \text{ and } \mu = \frac{E}{2(1+\sigma)}$$

$$\text{Therefore } \tau_{ij} = \frac{E\sigma}{(1+\sigma)(1-2\sigma)}\theta\delta_{ij} + \frac{E}{(1+\sigma)}e_{ij}$$

$$\Rightarrow \tau_{ij} = \frac{E}{(1+\sigma)}\left[e_{ij} + \frac{\sigma}{1-2\sigma}\delta_{ij}e_{kk}\right] \text{ (as } \theta = e_{kk}) \tag{8.7.2}$$

Also we have

$$e_{ij} = \frac{1}{2\mu}\left[\tau_{ij} - \frac{\lambda\Theta}{3\lambda+2\mu}\delta_{ij}\right] \tag{8.7.3}$$

$$\text{Also } E = \frac{(3\lambda+2\mu)\mu}{\lambda+\mu} \text{ and } \sigma = \frac{\lambda}{2(\lambda+\mu)}$$
$$\therefore \frac{\sigma}{E} = \frac{\lambda}{2(\lambda+\mu)}\frac{(\lambda+\mu)}{\mu(3\lambda+2\mu)} = \frac{\lambda}{2\mu(3\lambda+2\mu)}$$
$$\therefore \frac{\lambda}{3\lambda+2\mu} = \frac{\sigma}{E}\cdot 2\mu = \frac{\sigma}{E}\cdot\frac{E}{1+\sigma} = \frac{\sigma}{1+\sigma} \quad \left[\text{since } \mu = \frac{E}{2(1+\sigma)}\right] \tag{8.7.4}$$

From (8.7.3) we have

$$e_{ij} = \frac{1+\sigma}{E}\left[\tau_{ij} - \frac{\sigma}{1+\sigma}\Theta\delta_{ij}\right]$$
$$\Rightarrow e_{ij} = \frac{1}{E}\left[(1+\sigma)\tau_{ij} - \sigma\Theta\delta_{ij}\right]$$

Thus

$$\begin{aligned} e_{11} &= \frac{1}{E}\left[(1+\sigma)\tau_{11} - \sigma(\tau_{11}+\tau_{22}+\tau_{33})\right] \quad (\text{since } \Theta = \tau_{11}+\tau_{22}+\tau_{33}) \\ &= \frac{1}{E}\left[\tau_{11} - \sigma(\tau_{22}+\tau_{33})\right] \end{aligned} \tag{8.7.5}$$

Similarly,

$$
\begin{aligned}
e_{22} &= \frac{1}{E}\Big[\tau_{22} - \sigma(\tau_{33} + \tau_{11})\Big] \\
e_{33} &= \frac{1}{E}\Big[\tau_{33} - \sigma(\tau_{11} + \tau_{22})\Big] \\
e_{23} &= \frac{1}{E}\Big[(1+\sigma)\tau_{23} - 0\Big] = \frac{1+\sigma}{E}\tau_{23} \\
e_{31} &= \frac{1+\sigma}{E}\tau_{31} \quad \text{and} \quad e_{12} = \frac{1+\sigma}{E}\tau_{12}
\end{aligned}
$$

Unit 9

Course Structure

- Linearised elasticity, Equation of motion and equilibrium
- Compatibility of strain components, Beltrami-Michell compatibility equation
- Strain energy density function, Saint Venant's principle
- Boundary value problems of elasticity, Clapeyron's theorem

9.1 Equation of Motion and Equilibrium in terms of Displacement

Using the principle of balance of linear momentum, the stress equation of equilibrium of a continuum under external body forces per unit volume is given by

$$\tau_{ij,j} + F_i = 0,\ i,j = 1,2,3 \tag{9.1.1}$$

where τ_{ij} are the stress components. From the stress-strain relation for a linear isotropic solid is given by

$$\tau_{ij} = \lambda\theta\delta_{ij} + 2\mu e_{ij},\ \theta = e_{kk},\ i,j = 1,2,3 \tag{9.1.2}$$

where,

$$2e_{ij} = \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) = u_{i,j} + u_{j,i} \tag{9.1.3}$$

u_i being the displacement component.
Then, from (9.1.2) and (9.1.3), we get

$$\tau_{ij} = \lambda\theta\delta_{ij} + \mu[u_{i,jj} + (u_{j,i})_j] \tag{9.1.4}$$

Differentiating (9.1.4) partially with respect to x_j and summing over j we get

$$\begin{aligned}\tau_{ij,j} &= \lambda\delta_{ij}\theta_{,j} + \mu[u_{i,jj} + (u_{j,i})_j] \\ &= \lambda\theta_{,j} + \mu[\nabla^2 u_i + (u_{j,j})_i] \\ &= \lambda\theta_{,j} + \mu\nabla^2 u_i + \mu\theta_{,1i} \end{aligned} \tag{9.1.5}$$

since

$$\begin{aligned}\theta &= e_{kk} \\ &= \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} + \frac{\partial u_3}{\partial x_3} \\ &= u_{j,j}\end{aligned}$$

Hence, from (9.1.1) and (9.1.5), we get,

$$\mu\nabla^2 u_i + (\lambda+\mu)\theta_{,i} + F_i = 0,\ i = 1,2,3 \tag{9.1.6}$$

These are called the displacement equation of equilibrium, also called *Navier's equations*.
The equation of motion in terms of displacements can similarly be obtained from the stress equations of motion namely

$$\tau_{ij,j} + F_i = \rho\ddot{u}_i,$$

$$\mu\nabla^2 u_i + (\lambda+\mu)\theta_{,i} + F_i = \rho\frac{\partial^2 u}{\partial t^2}$$

In vector form,

$$\mu\nabla^2\vec{u} + (\lambda+\mu)\text{grad}(\text{div }\vec{u}) + \vec{F} = \rho\ddot{\vec{u}}$$

where, $\theta = \text{div }\vec{u}$.
Note: For one-dimensional deformation, in absence of body forces, we have (Resolving in the x-direction)

$$\begin{aligned}&\mu\nabla^2 u(x,t) + (\lambda+\mu)\frac{\partial\theta}{\partial x} = \rho\frac{\partial^2 u}{\partial t^2}(x,t) \\ &\Rightarrow\ \mu\frac{\partial^2 u}{\partial x^2} + (\lambda+\mu)\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\right) = \rho\frac{\partial^2 u}{\partial t^2} \\ &\Rightarrow\ (\lambda+2\mu)\frac{\partial^2 u}{\partial x^2} = \rho\frac{\partial^2 u}{\partial t^2} \\ &\Rightarrow \frac{\partial^2 u}{\partial x^2} = \frac{\rho}{\lambda+2\mu}\frac{\partial^2 u}{\partial t^2} \\ &\Rightarrow \frac{\partial^2 u}{\partial x^2} = \frac{1}{\frac{\lambda+2\mu}{\rho}}\frac{\partial^2 u}{\partial t^2} \\ &\Rightarrow \frac{\partial^2 u}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} \quad \text{where} \quad c = \sqrt{\frac{\lambda+\mu}{\rho}}\end{aligned}$$

This equation is called *one-dimensional wave equation*. c is the speed of the dilational wave or compressional wave.

9.2 Compatibility of Strain Components

In three-dimension, the strain-tensor has six components and the displacement vector has three components. If the displacement components are given, then from the strain displacement relations,

$$e_{ij} = \frac{1}{2}(u_{i,j} + u_{j,i}),\ i,j = 1,2,3 \tag{9.2.1}$$

The strain components e_{ij} can easily be determined. On the other hand, if six components e_{ij} are given, the displacement field requires the solution of a system of six partial differential equation (9.2.1).

For the determinaton of three unknowns $u_i (i = 1,2,3)$, such a system is overdetermined and for the existence of a single valued continuous displacement field, certain restrictions must be imposed. These conditions are known as *compatibility equations for strains*.

It is known that the six compatibility equations are

$$e_{ij,kl} + e_{kl,ij} - e_{ik,jl} - e_{jl,ik} = 0,\ (i,j,k = 1,2,3)$$

This system of equations consist of $3^4 = 81$ equations, but some of these are identically satisfied and some others are repeatitions becauseof the symmetry in the indices i,j and k,l. then only six of these 81 equations remain and they are

$$\begin{aligned} e_{11,22} + e_{22,11} &= 2e_{12,12} \\ e_{22,33} + e_{33,22} &= 2e_{23,23} \\ e_{33,11} + e_{33,22} &= 2e_{31,31} \\ (-e_{23,1} + e_{31,2} + e_{12,3})_{,1} &= e_{11,23} \\ (e_{23,1} - e_{31,2} + e_{12,3})_{,1} &= e_{22,31} \\ (e_{23,1} + e_{31,2} - e_{12,3})_{,1} &= e_{33,12} \end{aligned}$$

Example 9.2.1. Test whether the following system of strain components is possible in an elastic body:

$$e_{xx} = k(x^2 + y^2),\ \ e_{yy} = k(z^2 + y^2),\ \ e_{xy} = kxyz,$$
$$e_{yz} = e_{zx} = e_{xy} = e_{zz} = 0,\ k \neq 0$$

Solution: Here,

$$\frac{\partial e_{xx}}{\partial y} = 2ky,\ \frac{\partial^2 e_{xx}}{\partial y^2} = 2k,$$
$$\frac{\partial e_{yy}}{\partial x} = 0,\ \frac{\partial^2 e_{yy}}{\partial x^2} = 0,$$
$$\frac{\partial^2 e_{xy}}{\partial x \partial y} = \frac{\partial}{\partial x}\left(\frac{\partial e_{xy}}{\partial y}\right) = \frac{\partial}{\partial x}(kxz) = kz$$

Hence,

$$\frac{\partial^2 e_{xx}}{\partial y^2} + \frac{\partial^2 e_{yy}}{\partial x^2} \neq 2\frac{\partial^2 e_{xy}}{\partial x \partial y}$$

Thus, the compatibility equations are not all satisfied and hence, the system of strain components is not possible in an elastic body.

9.3 Beltrami-Michell Compatibility Equations

Let us consider the problems in which surface (stress) are prescribed everywhere on the boundary.
The equations of compatibility of stress are

$$e_{ij,kl} + e_{kl,ij} - e_{ik,jl} - e_{jl,ik} = 0. \tag{9.3.1}$$

Putting $l = k$ and summing over k we get

$$e_{ij,kk} + e_{kk,ij} - e_{ik,jk} - e_{jk,ik} = 0. \tag{9.3.2}$$

From stress-strain relations for isotropic elastic body

$$e_{ij} = \frac{1+\sigma}{E}\left[\tau_{ij} - \frac{\sigma}{1+\sigma}\Theta\delta_{ij}\right] \tag{9.3.3}$$

where,

$$\sigma = \frac{\lambda}{2(\lambda+\mu)},\ E = \frac{\mu(3\lambda+2\mu)}{\lambda+\mu}.$$

Using the results

$$\tau_{ij,kk} = \frac{\partial^2 \tau_{ij}}{\partial x_k^2} = \nabla^2 \tau_{ij}$$

$$\tau_{kk,ij} = \frac{\partial^2 \tau_{kk}}{\partial x_i \partial x_j} = \Theta_{,ij}$$

From equation (9.3.2), we get,

$$\begin{aligned}
e_{ij,kk} &= \frac{1+\sigma}{E}\left[\tau_{ij,kk} - \frac{\sigma}{1+\sigma}\Theta_{,kk}\delta_{ij}\right] = \frac{1+\sigma}{E}\left[\nabla^2\tau_{ij} - \frac{\sigma}{1+\sigma}\Theta_{,kk}\delta_{ij}\right] \\
e_{kk,ij} &= \frac{1+\sigma}{E}\left[\tau_{kk,ij} - \frac{\sigma}{1+\sigma}\Theta_{,ij}\delta_{kk}\right] = \frac{1+\sigma}{E}\left[\Theta_{,ij} - \frac{3\sigma}{1+\sigma}\Theta_{,ij}\right] \\
e_{ik,jk} &= \frac{1+\sigma}{E}\left[\tau_{ik,jk} - \frac{\sigma}{1+\sigma}\Theta_{,jk}\delta_{ik}\right] \\
e_{jk,ik} &= \frac{1+\sigma}{E}\left[\tau_{jk,ik} - \frac{\sigma}{1+\sigma}\Theta_{,ik}\delta_{jk}\right]
\end{aligned}$$

Using these relations the equation (9.3.2) reduces to

$$\frac{1+\sigma}{E}\left[\nabla^2\tau_{ij} - \frac{\sigma}{1+\sigma}\Theta_{,kk}\delta_{ij} + \Theta_{,ij} - \frac{3\sigma}{1+\sigma}\Theta_{,ij} - \tau_{ik,jk} + \frac{\sigma}{1+\sigma}\Theta_{,jk}\delta_{ik} - \tau_{jk,ik} + \frac{\sigma}{1+\sigma}\Theta_{,ik}\delta_{jk}\right] = 0$$

Hence,

$$\begin{aligned}\nabla^2\tau_{ij}+\Theta_{,ij}-\tau_{ik,jk}-\tau_{jk,ik} &= \frac{\sigma}{1+\sigma}[\Theta_{,kk}\delta_{ij}+3\Theta_{,ij}-\Theta_{,jk}\delta_{ik}-\Theta_{,ik}\delta_{jk}]\\ &= \frac{\sigma}{1+\sigma}[\Theta_{,kk}\delta_{ij}+3\Theta_{,ij}-\Theta_{,ji}-\Theta_{,ij}]\end{aligned}$$

$$\Rightarrow,\ \nabla^2\tau_{ij}+\left[1-\frac{3\sigma}{1+\sigma}+\frac{2\sigma}{1+\sigma}\right]\Theta_{,ij}-\frac{\sigma}{1+\sigma}\nabla^2\Theta\delta_{ij}=\tau_{ik,jk}+\tau_{jk,ik}$$

$$\Rightarrow\ \nabla^2\tau_{ij}+\frac{1}{1+\sigma}\Theta_{,ij}-\frac{\sigma}{1+\sigma}\delta_{ij}\nabla^2\Theta=\tau_{ik,jk}+\tau_{jk,ik} \tag{9.3.4}$$

Now, from the stress equation of equilibrium, we have,

$$\tau_{ik,k}+\rho F_i=0,$$

where F_i's are the body force componenets per unit mass.
Differentiating partially with respect to x_j we get,

$$\begin{aligned}&\tau_{ik,jk}+\rho F_{i,j}=0\\ &\Rightarrow\ \tau_{ik,jk}=-\rho F_{i,j}\end{aligned}$$

Similarly differentiating with respect to x_i, we get

$$\tau_{jk,ik} = -\rho F_{j,i}$$

Substituting these values in equation 9.3.4, we get,

$$\nabla^2\tau_{ij}+\frac{1}{1+\sigma}\Theta_{,ij}-\frac{\sigma}{1+\sigma}\delta_{ij}\nabla^2\Theta=-\rho[F_{i,j}+F_{j,i}] \tag{9.3.5}$$

Putting $j=i$ and summing up with respect to i, we get

$$\begin{aligned}\nabla^2\tau_{ii}+\frac{1}{1+\sigma}\Theta_{,ii}-\frac{\sigma}{1+\sigma}\delta_{ii}\nabla^2\Theta &= -\rho[F_{i,i}+F_{i,i}]\\ \nabla^2\Theta+\frac{1}{1+\sigma}\nabla^2\Theta-\frac{3\sigma}{1+\sigma}\delta_{ii}\nabla^2\Theta &= -2\rho F_{i,i}[\Theta=\tau_{kk}]\end{aligned}$$

Hence,

$$\nabla^2\Theta=-\frac{1+\sigma}{1-\sigma}div\vec{F} \tag{9.3.6}$$

Substituting equation 9.3.6 in equation 9.3.5 we get

$$\nabla^2\tau_{ij}+\frac{1}{1+\sigma}\Theta_{,ij}=-\frac{\sigma}{1-\sigma}\rho\delta_{ij}\text{div}\vec{F}-\rho[F_{i,j}+F_{j,i}],\ (i,j=1,2,3) \tag{9.3.7}$$

The above equation contains six independent equations. They are known as the *Beltrami-Michell compatibility equations for stresses*.

More generally, Beltrami-Michell equations become,

$$\begin{aligned}
\nabla^2\tau_{xx} + \frac{1}{1+\sigma}\frac{\partial^2\Theta}{\partial x^2} &= -\frac{\rho\sigma}{1-\sigma}.div\vec{F} - 2\rho\frac{\partial F_x}{\partial x} \\
\nabla^2\tau_{yy} + \frac{1}{1+\sigma}\frac{\partial^2\Theta}{\partial y^2} &= -\frac{\rho\sigma}{1-\sigma}.div\vec{F} - 2\rho\frac{\partial F_y}{\partial y} \\
\nabla^2\tau_{zz} + \frac{1}{1+\sigma}\frac{\partial^2\Theta}{\partial z^2} &= -\frac{\rho\sigma}{1-\sigma}.div\vec{F} - 2\rho\frac{\partial F_z}{\partial z} \\
\nabla^2\tau_{yz} + \frac{1}{1+\sigma}\frac{\partial^2\Theta}{\partial y\partial z} &= -\rho(F_{y,z} + F_{z,y}) \\
\nabla^2\tau_{zx} + \frac{1}{1+\sigma}\frac{\partial^2\Theta}{\partial z\partial x} &= -\rho(F_{z,x} + F_{x,z}) \\
\nabla^2\tau_{xy} + \frac{1}{1+\sigma}\frac{\partial^2\Theta}{\partial x\partial y} &= -\rho(F_{x,y} + F_{y,x})
\end{aligned}$$

Result 9.3.1. If the body forces are constants, the invariants Θ and θ are harmonic functions and the stress components τ_{ij} and strain components e_{ij} are bi-harmonic functions.

Proof. The Beltrami-Michell compatibility equation for stresses are given by

$$\nabla^2\tau_{ij} + \frac{1}{1+\sigma}\Theta_{,ij} = -\frac{\sigma}{1-\sigma}\rho\delta_{ij}\text{div}\vec{F} - \rho[F_{i,j} + F_{j,i}]$$

If $\vec{F}$ be a constant vector, then $\text{div}\vec{F} = \vec{0}$, $F_{i,j} = 0 = F_{j,i}$. Therefore, the above equations become

$$\nabla^2\tau_{ij} + \frac{1}{1+\sigma}\Theta_{,ij} = 0$$

Putting $j = i$ and summing over i, we get

$$\begin{aligned}
& \nabla^2\tau_{ii} + \frac{1}{1+\sigma}\Theta_{,ii} = 0 \\
\implies \quad & \nabla^2\Theta + \frac{1}{1+\sigma}\nabla^2\Theta = 0 \\
\implies \quad & \frac{2+\sigma}{1+\sigma}\nabla^2\Theta = 0 \\
\implies \quad & \nabla^2\Theta = 0
\end{aligned}$$

which shows that Θ is harmonic. Again, since $\Theta = (3\lambda + 2\mu)\theta$, we have

$$\nabla^2\Theta = \nabla^2(3\lambda + 2\mu)\theta = 0$$

$$\implies \nabla^2\theta = 0$$

which shows that θ is harmonic also. Now, from Beltrami-Michell equation,

$$\nabla^2\tau_{ij} + \frac{1}{1+\sigma}\Theta_{,ij} = 0$$

Operating on both sides by ∇^2, we get,

$$\nabla^4\tau_{ij} + \frac{1}{1+\sigma}(\nabla^2\Theta)_{,ij} = 0$$
$$\implies \nabla^4\tau_{ij} = 0 \ \text{[since } \nabla^2\Theta = 0\text{]}$$

which shows that the stress components τ_i are bi-harmonic.
Further, we have

$$\tau_{ij} = \lambda\theta\delta_{ij} + 2\mu e_{ij} \ \text{[for isotropic elastic body]}$$

Operating on both sides by ∇^2, we get

$$\nabla^2\tau_{ij} = \lambda\nabla^2\theta\delta_{ij} + 2\mu\nabla^2 e_{ij} = 2\mu\nabla^2 e_{ij}, \quad \text{since}\nabla^2\theta = 0.$$

Again, operating ∇^2 on both sides,

$$\nabla^4\tau_{ij} = 2\mu\nabla^4 e_{ij} = 0 \ \text{[as } \nabla^4\tau_{ij} = 0\text{]}$$

Hence,

$$\nabla^4 e_{ij} = 0$$

which shows that e_{ij} are bi-harmonic. □

Example 9.3.2. Show that the following stress components are not the solutions of the problem in elasticity, even though they satisfy the equations of equilibrium with zero body forces:

$$\tau_{11} = \alpha[x_2^2 + \sigma(x_1^2 - x_2^2)], \ \ \tau_{12} = -2\alpha\sigma x_1 x_2, \ \ \tau_{13} = 0,$$
$$\tau_{22} = \alpha[x_1^2 + \sigma(x_2^2 - x_1^2)], \ \ \tau_{23} = 0, \ \ \tau_{33} = \alpha\sigma(x_1^2 + x_2^2)$$

Solution: The quantity

$$\Theta = \tau_{kk} = \tau_{11} + \tau_{22} + \tau_{33}$$

is given by,

$$\begin{aligned}\Theta = \tau_{kk} &= \alpha[x_2^2 + \sigma(x_1^2 - x_2^2)] + \alpha[x_1^2 + \sigma(x_2^2 - x_1^2)] + \alpha\sigma(x_1^2 + x_2^2) \\ &= \alpha[(x_1^2 + x_2^2) + \sigma(x_1^2 + x_2^2)] \\ &= \alpha(1+\sigma)(x_1^2 + x_2^2)\end{aligned}$$

In absence of both forces, the Beltrami-Michell compatibility equations are

$$\nabla^2\tau_{11} + \frac{1}{1+\sigma}\Theta_{,ij} = 0$$

For $i = 1, 2, 3$, $j = 1, 2, 3$ this equation becomes,

$$\begin{aligned}\nabla^2\tau_{ij} + \frac{1}{1+\sigma}\Theta_{,11} &= \frac{\partial^2\tau_{11}}{\partial x_1^2} + \frac{\partial^2\tau_{11}}{\partial x_2^2} + \frac{\partial^2\tau_{11}}{\partial x_3^2} + \frac{1}{1+\sigma}\frac{\partial^2\Theta}{\partial x_1^2} \\ &= 2\alpha\sigma + (2\alpha - 2\alpha\sigma) + 0 + \frac{1}{1+\sigma}.2\alpha(1+\sigma) \\ &= 4\alpha \\ &\neq 0\end{aligned}$$

$$\begin{aligned}\nabla^2\tau_{22} + \frac{1}{1+\sigma}\Theta_{,22} &= \frac{\partial^2\tau_{22}}{\partial x_1^2} + \frac{\partial^2\tau_{22}}{\partial x_2^2} + \frac{\partial^2\tau_{22}}{\partial x_3^2} + \frac{1}{1+\sigma}\frac{\partial^2\Theta}{\partial x_2^2} \\ &= (2\alpha - 2\alpha\sigma) + 2\alpha\sigma + 0 + \frac{1}{1+\sigma}.2\alpha(1+\sigma) \\ &= 4\alpha \\ &\neq 0\end{aligned}$$

$$\begin{aligned}\nabla^2\tau_{33} + \frac{1}{1+\sigma}\Theta_{,33} &= \frac{\partial^2\tau_{33}}{\partial x_1^2} + \frac{\partial^2\tau_{33}}{\partial x_2^2} + \frac{\partial^2\tau_{33}}{\partial x_3^2} + \frac{1}{1+\sigma}\frac{\partial^2\Theta}{\partial x_3^2} \\ &= 2\alpha\sigma + 2\alpha\sigma + 0 + \frac{1}{1+\sigma}.0 \\ &= 4\alpha\sigma \\ &\neq 0\end{aligned}$$

$$\begin{aligned}\nabla^2\tau_{12} + \frac{1}{1+\sigma}\Theta_{,12} &= \frac{\partial^2\tau_{12}}{\partial x_1^2} + \frac{\partial^2\tau_{12}}{\partial x_2^2} + \frac{\partial^2\tau_{12}}{\partial x_3^2} + \frac{1}{1+\sigma}\frac{\partial^2\Theta}{\partial x_1\partial x_2} \\ &= 0\end{aligned}$$

Similarly,

$$\begin{aligned}\nabla^2\tau_{13} + \frac{1}{1+\sigma}\Theta_{,13} &= 0 \\ \nabla^2\tau_{23} + \frac{1}{1+\sigma}\Theta_{,23} &= 0\end{aligned}$$

Thus, Beltrami-Michell compatibility equations are not all satisfied. Hence the given stress components are not solutions of the problem in elasticity.
The stress equations of equilibrium for a continuous medium is given by

$$\tau_{ij,j} + \rho F_i = 0, \;\; i, j = 1, 2, 3.$$

In absence of body forces, the stress equations of equilibrium becomes $\tau_{ij,j} = 0, \;\; i, j = 1, 2, 3.$ Therefore,

$$\tau_{1j,j} = \frac{\partial\tau_{11}}{\partial x_1} + \frac{\partial\tau_{12}}{\partial x_2} + \frac{\partial\tau_{13}}{\partial x_3} = 2\alpha\sigma x_1 - 2\alpha\sigma x_1 + 0 = 0$$

$$\tau_{2j,j} = \frac{\partial \tau_{21}}{\partial x_1} + \frac{\partial \tau_{22}}{\partial x_2} + \frac{\partial \tau_{23}}{\partial x_3} = -2\alpha\sigma x_2 + 2\alpha\sigma x_2 + 0 = 0$$

$$\tau_{3j,j} = \frac{\partial \tau_{31}}{\partial x_1} + \frac{\partial \tau_{32}}{\partial x_2} + \frac{\partial \tau_{33}}{\partial x_3} = 0 + 0 + 0 = 0$$

Hence, in absence of body forces, the given stress components satisfy the equations of equilibrium.

9.4 Strain Energy Density Function

The work done in deforming a body by the surface forces, that is the stress, is transformed completely into potential energy which is stored in that body. This potential energy due to deformation or strain is called the *strain energy or the stress potential of the elastic body*.

Let W represents potential energy per unit volume stored up in the body by strain deformation alone, then W must be a function of components of strain so that we can write

$$W = W(e_1, e_2, \ldots, e_6) \text{ such that } \tau_i = \frac{\partial W}{\partial e_i},\ i = 1, 2, \ldots, 6$$

where,

$$\tau_1 = \tau_{11},\ \tau_2 = \tau_{22},\ \tau_3 = \tau_{33},$$
$$\tau_4 = \tau_{23} = \tau_{32},\ \tau_5 = \tau_{31} = \tau_{13},\ \tau_6 = \tau_{12} = \tau_{21}$$

Similarly,

$$e_1 = e_{11},\ e_2 = e_{22},\ e_3 = e_{33},$$
$$e_4 = e_{23} = e_{32},\ e_5 = e_{31} = e_{13},\ e_6 = e_{12} = e_{21}$$

Now, expanding W in a power series about the origin, we have,

$$\begin{aligned} W &= W(0, 0, \ldots, 0) + \left(\frac{\partial W}{\partial e_i}\right)_0 e_i + \frac{1}{2}\left(\frac{\partial^2 W}{\partial e_i \partial e_j}\right)_0 e_i e_j + \cdots \\ &= c_0 + c_i e_i + \frac{1}{2} c_{ij} e_i e_j \text{[Neglecting the third and other higher order terms]} \end{aligned}$$

where,

$$c_0 = W(0, 0, \ldots, 0) = \text{a constant.}$$

$$c_i = \left(\frac{\partial W}{\partial e_i}\right)_0,\ c_{ij} = \left(\frac{\partial^2 W}{\partial e_i \partial e_j}\right)_0$$

Since the magnitude of the derivatives of W does not depend on the order of differentiation,

$$\begin{aligned} \frac{\partial^2 W}{\partial e_i \partial e_j} &= \frac{\partial^2 W}{\partial e_j \partial e_i} \\ \implies c_{ij} &= c_{ji} \end{aligned}$$

We disregard the constant c_0, since we are interested in the derivative of W. We can set $c_0 = 0$. Then,

$$W = c_i e_i + \frac{1}{2} c_{ij} e_i e_j.$$

We have,

$$\begin{aligned} \frac{\partial}{\partial e_i}(c_{pq} e_p e_q) &= c_{pq}\left(e_p \frac{\partial e_q}{\partial e_i} + \frac{\partial e_p}{\partial e_i} e_q\right) \\ &= c_{pq}(e_p \delta_{qi} + \delta_{pi} e_q) \\ &= c_{pi} e_p + c_{iq} e_q \\ &= (c_{ji} + c_{ij}) e_j. \end{aligned}$$

So, differentiating W with respect to e_i, we get,

$$\begin{aligned} \frac{\partial W}{\partial e_i} &= c_i + \frac{1}{2}(c_{ji} + c_{ij}) e_j \\ \implies \tau_i &= c_i + \frac{1}{2}(c_{ji} + c_{ij}) e_j \end{aligned}$$

Now, τ_i vanishes when the strains are zero. Therefore, we must have $c_i = 0$. Thus,

$$\begin{aligned} W &= \frac{1}{2} c_{ij} e_i e_j \;\; i, j = 1, 2, \ldots, 6 \\ 2W &= (c_{ij} e_j) e_i \\ &= \tau_i e_i \;\; i = 1, 2, \ldots, 6 \end{aligned}$$

since $\tau_i = c_{ij} e_j$ which is the generalized Hooke's law for linear elastic solid in which stresses are linear functions of strain.

Returning to double suffix symbols

$$2W = \tau_{ij} e_{ij} \;\; i, j = 1, 2, 3$$

For isotropic solids, Hooke's law gives

$$\tau_{ij} = \lambda \theta \delta_{ij} + 2\mu e_{ij}$$

Therefore,

$$\begin{aligned} 2W &= (\lambda \theta \delta_{ij} + 2\mu e_{ij}) e_{ij} \\ &= \lambda \theta \delta_{ij} e_{ij} + 2\mu e_{ij} e_{ij} \;\; i, j = 1, 2, 3 \\ &= \lambda \theta e_{ii} + 2\mu (e_{11}^2 + e_{22}^2 + e_{33}^2 + 2e_{23}^2 + 2e_{31}^2 + 2e_{12}^2) \\ &= \lambda \theta^2 + 2\mu (e_{11}^2 + e_{22}^2 + e_{33}^2 + 2e_{23}^2 + 2e_{31}^2 + 2e_{12}^2) \end{aligned}$$

Since λ and μ are positive, W is a positive definite quadratic form in the strains.

Different forms:

(a) *W in terms of invariants of strain tensor*.

The following expressions are the two invariants among three of the strain tensor:

$$I_1 = e_{ii} = e_{11} + e_{22} + e_{33} = \theta$$

$$\begin{aligned} I_2 &= \begin{vmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \end{vmatrix} + \begin{vmatrix} e_{22} & e_{23} \\ e_{32} & e_{33} \end{vmatrix} + \begin{vmatrix} e_{33} & e_{31} \\ e_{13} & e_{11} \end{vmatrix} \\ &= (e_{11}e_{22} - e_{12}^2) + (e_{22}e_{33} - e_{23}^2) + (e_{33}e_{11} - e_{13}^2) \end{aligned}$$

Now,

$$\begin{aligned} 2W &= \lambda\theta^2 + 2\mu(e_{11}^2 + e_{22}^2 + e_{33}^2 + 2e_{23}^2 + 2e_{31}^2 + 2e_{12}^2) \\ &= (\lambda + 2\mu)\theta^2 + 2\mu[e_{11}^2 + e_{22}^2 + e_{33}^2 + 2e_{23}^2 + 2e_{31}^2 + 2e_{12}^2 - (e_{11} + e_{22} + e_{33})^2] \\ &= (\lambda + 2\mu)\theta^2 - 2\mu.2[e_{11}e_{22} + e_{22}e_{33} + e_{33}e_{11} - e_{23}^2 - e_{31}^2 - e_{12}^2] \\ &= (\lambda + 2\mu)\theta^2 - 4\mu[(e_{11}e_{22} - 2e_{12}^2) + (e_{22}e_{33} - e_{23}^2) + (e_{33}e_{11} - e_{31}^2) \\ &= (\lambda + 2\mu)\theta^2 - 4\mu I_2 \\ &= (\lambda + 2\mu)I^2 - 4\mu I_2 \end{aligned}$$

(b) *W in terms of stresses*:

We have the expression of strains in terms of stresses as

$$e_{ij} = \frac{1}{E}[(1+\sigma)\tau_{ij} - \sigma\Theta\delta_{ij}] = \frac{1+\sigma}{E}\tau_{ij} - \frac{\sigma}{E}\Theta\delta_{ij}$$

Thus,

$$2W = \tau_{ij}e_{ij}$$

gives

$$\begin{aligned} 2W &= \tau_{ij}\left[\frac{1+\sigma}{E}\tau_{ij} - \frac{\sigma}{E}\Theta\delta_{ij}\right] \\ &= \frac{\sigma}{E}\Theta\tau_{ij}\delta_{ij} + \frac{1+\sigma}{E}\tau_{ij}\tau_{ij} \\ &= \frac{\sigma}{E}\Theta^2 + \frac{1+\sigma}{E}[\tau_{11}^2 + \tau_{22}^2 + \tau_{33}^2 + 2\tau_{23}^2 + 2\tau_{31}^2 + 2\tau_{12}^2] \end{aligned}$$

where, E is the *Young's modulus* and σ is *Poission's ratio*.

9.5 Saint Venant's Principle

In the application of theory of elasticity, we shall often refer to a principle into Saint-Venant's, the essence of which can be stated as follows:

If a system of forces acting on a small portion of the surface of an elastic body is repeated by another "statically equivalent" system of forces acting on the same portion of the surface, this redistribution of loading produces substantial changes in the stresses only in the immediate neighbourhood of hte loading and the stresses are essentially same in the parts of ht ebody which are at large distances in comparison with the linear dimension of the surface on which the forces are changed.

The phrase "statically equivalent" means that the two distributions of forces have the same resultant force and the same resultant moment.

As and illustration of the principle we can consider a long beam, one end of which is fixed in a rigid wall while the other is acted by a distribution of forces that gives rise to a resultant force F and a couple of moment M. Now there are infinitely many distributions of forces that may act on the end of the beam and that will have the same resultant moment M.

The principle of Saint-Venant asserts that while the distributions of stresses near the region of application may differ greatly, the eccentricities of the load distribution will have no appreciable effect on the state of stress far enough from the points of application, so that the system of applied forces are statically equivalent.

9.6 Boundary Value Problems of Static and Dynamic Elasticity

In elastostatics or elasto-dynamics where a linear elastic solid is in equilibrium or in motion, the shape of the bodyand the distribution of the external body forces throughout the body are known. The problem is: to find the 15 unknowns:6 stresses τ_{ij}, 6 strains e_{ij} and 3 displacement functions u_i which satisfy the basic 15 appropriate field equation.

(a) Three equation sof equilibrium:

$$\tau_{ij,j} + \rho F_i = 0, \;\; i = 1, 2, 3.$$

(b) Six stress-strain constitutive relations

$$\tau_{ij} = \lambda\theta\delta_{ij} + 2\mu e_{ij}, \;\; \theta = e_{kk}, \;\; i, j = 1, 2, 3.$$

or six equivalent strain stress constitutive relations

$$e_{ij} = \frac{1+\sigma}{E}\left[\tau_{ij} - \frac{\sigma}{1+\sigma}\Theta\delta_{ij}\right], \;\; \theta = \tau_{kk}, \;\; i, j = 1, 2, 3.$$

(c) and six displacement kinematic relations

$$e_{ij} = \frac{1}{2}[u_{i,j} + u_{j,i}], \;\; i, j = 1, 2, 3.$$

at all interior points of linearly elastic body.

The solution of these partial differential equations involve arbitrary functions by using a set of boundary or initial conditions(as the case may be). Accordingly, it is necessary to formulate proper boundary-value problems with appropriate boundary conditions for both elastostatical and elastodynamical problems.

9.6.1 Fundamental Boundary value problems in elastostatics

(a) *First fundamental boundary value problem:(stresses are prescribed)*

To determine the distribution of displacements and stresses in the interior of an elastic solid in equilibrium under prescribed body forces, when the distribution of the forces acting on the surface of the body is known.

(b) *Second Fundamental boundary value problem:(displacement are prescribed)*

To determine the distribution of displacements and stresses in the interior of an elastic solid in equilibrium under prescribed body forces, when the displacements on the surface of the body are prescribed functions.

(c) *Mixed boundary value problem:(stresses are prescribed over a portion of the boundary and displacements are prescribed over the remaining part)*

To determine the distributions of the displacements and stresses in the interior of an elastic solid in equilibrium under prescribed body forces when the distribution of forces on the parts $\sum_\tau$ of the surface $\sum$ and the displacement of points on the remaining part $\sum_u$ of $\sum$ are also prescribed functions.

Problem: State the fundamental boundary value problems of elasto-statics.

9.6.2 Uniqueness of solutions of fundamental boundary value problems in elastostatic cases

Before we proceed to prove the uniqueness of solutions of Fundamental boundary value problems in elastostatics, we establish an important theorem concerning the strain energy function.

Theorem 9.6.1. (Clapeyron's Theorem) If a body is in equilibrium under a given system of external body forces and surface forces, then the work done by the external forces of the equilibrium state in deforming the body from unstressed state to the state of equilibrium is equal to twice the strain energy of deformation.

Proof. Consider a linearly elastic body in a deformed state of rest under the action of body force F_i per unit mass and the surface force $\vec{T}^{(\nu)}$ per unit area. The work done by the above forces during the displacement u_i is

$$W = \iiint_V \rho F_i u_i dV + \iint_S T_i^{(\nu)} u_i ds \tag{9.6.1}$$

Now

$$\begin{aligned}
\iint_S T_i^{(\nu)} u_i ds &= \iint_S \tau_{ij} n_j u_i ds \\
&= \iiint_V (\tau_{ij} u_i)_{,j} dV \text{ [By Gauss-divergence theorem]} \\
&= \iiint_V (\tau_{ij,j} u_i dV + \tau_{ij} u_{i,j}) dV \\
&= \iiint_V \tau_{ij,j} u_i dV + \iiint_V \tau_{ij} e_{ij} dV + \iiint_V \tau_{ij} r_{ij} dV
\end{aligned} \tag{9.6.2}$$

where,

$$e_{ij} = \frac{1}{2}[u_{i,j} + u_{j,i}] = e_{ji}$$

$$r_{ij} = \frac{1}{2}[u_{i,j} - u_{j,i}] = -r_{ji}$$

Also,

$$\tau_{ij} = \tau_{ji}$$

Therefore,

$$\begin{aligned}
&\tau_{ij} r_{ij} = \tau_{ji} r_{ji} = -\tau_{ji} r_{ij} = -\tau_{ij} r_{ij} \\
&\Rightarrow \quad 2\tau_{ij} r_{ij} = 0
\end{aligned} \tag{9.6.3}$$

Hence, equation (9.6.2) becomes

$$\iint_S T_i^{(\nu)} u_i ds = \iiint_V (\tau_{ij,j} u_i dV + \iiint_V \tau_{ij} e_{ij} dV \tag{9.6.4}$$

Substituting (9.6.4) in (9.6.1), we get

$$\begin{aligned}
W &= \iiint_V (\rho F_i + \tau_{ij,j}) u_i dV + \iiint_V \tau_{ij} e_{ij} dV \\
&= \iiint_V \tau_{ij} e_{ij} dV \text{ [From the equation of equilibrium, } \rho F_i + \tau_{ij,j} = 0]
\end{aligned} \tag{9.6.5}$$

Now, the strain energy per unit volume

$$\mathscr{W} = \frac{1}{2}\tau_{ij} e_{ij} \text{ [If the deformation takes place isothermally or adiabetically by Clapeyron's formula]} \tag{9.6.6}$$

From (9.6.4) and (9.6.5)

$$W = \iiint_V 2\mathscr{W} dV$$

This completes the proof. □

Uniqueness: To prove the uniqueness of solutions, consider an elastic body in a state of rest subjected to a specific given body force F_i. In addition to body forces either surface forces $T_i^{(n)}$ or surface displacement u_i are prescribed on the boundary.

Let us assume that it is possible to obtain two stes of solutions $u_i^{'}$, $T_i^{(n)'}$ and $u_i^{"}$, $T_i^{(n)"}$ which satisfy 15 basic equations of elasticity and boundary conditions.

Let us define

$$u_i = u_i^{'} - u_i^{"}, \quad \tau_{ij} = \tau_{ij}^{'} - \tau_{ij}^{"}$$
$$e_{ij} = e_{ij}^{'} - e_{ij}^{"}, \quad T_i^{(n)} = T_i^{(n)'} - T_i^{(n)"}$$

For the first state of stress, we have

$$\tau_{ij,j}^{'} + \rho F_i = 0$$

as well as the following boundary condition

$$\begin{aligned} \tau_{ij}^{'} n_j &= f_i(x_1, x_2, x_3) \text{ [If surface forces are prescribed on the boundary]} \\ \Rightarrow \quad u_i^{'} &= g_i(x_1, x_2, x_3) \text{ [If displacement are prescribed on the boundary]} \end{aligned}$$

Similarly, for second state of stress

$$\tau_{ij,j}^{''} + \rho F_i = 0$$

with boundary condition

$$\begin{aligned} \tau_{ij,j}^{''} n_j &= f_i(x_1, x_2, x_3) \\ or, \quad u_i^{''} &= g_i(x_1, x_2, x_3) \end{aligned}$$

Subtracting, we get

$$\tau_{ij,j}^{'} - \tau_{ij,j}^{''} = 0$$

and either

$$\tau_{ij,j}^{'} n_j - \tau_{ij,j}^{''} n_j = 0$$

or,

$$u_i^{'} - u_i^{''} = 0$$

on the boundary.
In other words,

$$\tau_{ij,j} = 0$$

at every interior point and either

$$T_i^{(n)} = \tau ijn_j = \tau_{ij,j}^{'} n_j - \tau_{ij,j}^{''} n_j = 0$$

or,

$$u_i = u_i^{'} - u_i^{''} = 0$$

on the boundary.

Thus, we have a new state of stress in which body forces are absent and either surface forces or surface displacements vanish.

On the surface of the body, boundary conditions are either $T_i^{(n)} = 0$ or, $u_i = 0$. In either case, $T_i^{(n)} u_i = 0$ at every point on the boundary.

Also by Clapeyron's theorem,

$$\iiint_V 2\mathscr{W} dV = \iiint_V \rho F_i u_i dV + \iint_S T_i^{(n)} u_i ds$$

For new state of stress, $F_i = 0$ in V and $T_i^{(n)} u_i = 0$ on S. Hence we have,

$$\iiint_V \mathscr{W} dV = 0$$

But

$$\mathscr{W} = \frac{1}{2}\lambda\theta^2 + \mu(e_{11}^2 + e_{22}^2 + e_{33}^2 + 2e_{12}^2 + 2e_{23}^2 + 2e_{13}^2)$$

is a positive definite quadratic form in components of strain. Hence the integral can vanish only when $\mathscr{W} = 0$, that is, when $e_{ij} = 0$. Also, from

$$\tau_{ij} = \lambda\theta\delta_{ij} + 2\mu e_{ij}$$

it follows that

$$\tau_{ij} = 0$$

Therefore,

$$e_{ij}^{'} = e_{ij}^{''} \ \ and \ \ \tau_{ij}^{'} = \tau_{ij}^{''}$$

Consequently, components of strain tensor and components of stress tensor are identical. As regards the uniqueness of displacements, we know that the displacements are solutions of the equation

$$u_{i,j} + u_{j,i} = 2e_{ij} = 0$$

and are determined within the quantities representing the rigid body motion. Hence the problem follows.

9.7 Fundamental boundary value problems in Elastodynamics

In elastodynamics, the equilibrium equations must be replaced by the equations of motion in the system of basic field equations. Therefore, all field quantities are now considered as functions of time as well as of the coordinates. In elastodynamics, the problem is to find 15 unknowns: 6 stresses τ_{ij}, 6 strains e_{ij} and 3 displacement functions $u_i = u_i(x,t)$ which staisfy the basic 15 equations

$$\begin{aligned} \tau_{ij,j} + \rho F_i &= \rho \ddot{u}_i,\ i = 1,2,3 \\ \tau_{ij} = \lambda\theta\delta_{ij} + 2\mu e_{ij},\ \theta &= e_{kk},\ i,j = 1,2,3 \\ e_{ij} &= \frac{1}{2}(u_{i,j} + u_{j,i}) \end{aligned} \tag{9.7.1}$$

at all interior points of linearly elastic body.

(a) **First fundamental boundary value problem in elastodynamics**(stress vector is given at each point on the boundary)

When the surface function $f_i(x_1, x_2, x_3, t)$ are prescribed on the boundary surface of the body at the time t, representing the stress vector acting on surface element with normal n_i the stresses τ_{ij}, in addition must satisfy 3 boundary conditions

$$\tau_{ij} n_j = f_i(x_1, x_2, x_3, t),\ i = 1,2,3 \tag{9.7.2}$$

To these conditions, it is necessary to adjoin the initial conditions specifying displacement and velocity of a point of the body at initial time $t = 0$, that is,

$$\begin{aligned} u_i(x_1, x_2, x_3, 0) &= F_i(x_1, x_2, x_3) \\ \frac{\partial u_i}{\partial t}(x_1, x_2, x_3, 0) &= G_i(x_1, x_2, x_3) \end{aligned} \tag{9.7.3}$$

throughout the volume. Thus, the problem of obtaining the displacements, strains and stresses in linearly elastic isotropic solid body in equilibrium which satisfy 15 basic equations (9.7.1) in addition to the boundary condition (9.7.2) together with the initial conditions (9.7.3) is known as the first fundamental boundary value problem in elastodynamics.

(b) **Second fundamental boundary value problem in elastodynamics** (The displacement is prescribed at each point on the boundary)

When the displacement functions $g_i(x_1, x_2, x_3, t)$ are prescribed on the boundary at time t then the displacement u_i, in addition, must satisfy 3 boundary conditions

$$u_i = g_i(x_1, x_2, x_3, t),\ \ i = 1,2,3 \tag{9.7.4}$$

The problem of obtaining the displacements, strains and stresses which satisfy the 15 basic equations (9.7.1) in addition with the boundary conditions (9.7.4) together with the initial conditions (9.7.3) is known as the second fundamental boundary value problem in elastodynamics.

9.8 Uniqueness of solutions of fundamental BVP in elastodynamics

Before we proceed to prove, we establish an important result.

Theorem 9.8.1. The time rate of change of work done by external forces in altering the configuration of the natural state of an elastic body to the current state is equal to the sum of time rate of change of kinetic energy and the time rate of change of strain energy.

Proof. Suppose a body is acted on by a surface force $T_i^{(n)}$ per unit area and a body force F_i per unit mass. Let u_i be the displacement of the point at time t. The displacement of the point during the time interval dt is $\dfrac{\partial u_i}{\partial t}dt$.
If dW is the work done by external forces during time dt, then

$$\begin{aligned} dW &= \iiint_V \rho F_i \dot{u}_i dt dV + \iint_S T_i^{(n)} \dot{u}_i dt ds \\ \text{or, } \frac{dW}{dt} &= \iiint_V \rho F_i \dot{u}_i dV + \iint_S T_i^{(n)} \dot{u}_i ds \end{aligned} \tag{9.8.1}$$

Now,

$$\begin{aligned} \iint_S T_i^{(n)} \dot{u}_i ds &= \iint_S \tau_{ij} n_j \dot{u}_i ds \\ &= \iiint_V (\tau_{ij}\dot{u}_j)_{,j} dV \\ &= \iiint_V (\tau_{ij,j}\dot{u}_i + \tau_{ij}\dot{u}_{i,j}) dV \\ &= \iiint_V \tau_{ij,j}\dot{u}_i dV + \iiint_V \tau_{ij} d_{ij} dV + \iiint_V \tau_{ij} w_{ij} dV \end{aligned} \tag{9.8.2}$$

where

$$d_{ij} = \frac{1}{2}(\dot{u}_{i,j} + \dot{u}_{j,i}) = e_{ij}$$

and

$$w_{ij} = \frac{1}{2}(\dot{u}_{i,j} - \dot{u}_{j,i})$$

Also,

$$\tau_{ij} w_{ij} = 0$$

Hence, equation (9.8.1) becomes,

$$\begin{aligned}
\frac{d\mathscr{W}}{dt} &= \iiint_V \rho F_i \dot{u}_i dV + \iiint_V \tau_{ij,j}\dot{u}_i dV + \iiint_V \tau_{ij} d_{ij} dV \\
&= \iiint_V (\rho F_i + \tau_{ij,j})\dot{u}_i dV + \iiint_V \tau_{ij} d_{ij} dV \\
&= \iiint_V \rho \ddot{u}_i \dot{u}_i dV + \iiint_V \tau_{ij} d_{ij} dV \;\; [as,\; \rho F_i + \tau_{ij,j} = \rho \ddot{u}_i] \\
&= \frac{dK}{dt} + \iiint_V \tau_{ij} d_{ij} dV
\end{aligned}$$

where,

$$\begin{aligned}
K &= \text{Kinetic energy of the body} \\
&= \frac{1}{2}\iiint \rho \dot{u}_i . \dot{u}_i . dV.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{d\mathscr{W}}{dt} &= \frac{dK}{dt} + \iiint_V \tau_{ij}\frac{\partial e_{ij}}{\partial t} dV \; as \; d_{ij} = \dot{e}_{ij} = \frac{\partial e_{ij}}{\partial t} \\
&= \frac{dK}{dt} + \iiint_V \frac{\partial W}{\partial e_{ij}}\frac{\partial e_{ij}}{\partial t} dV \; as \; \tau_{ij} = \frac{\partial W}{\partial e_{ij}} \\
&= \frac{dK}{dt} + \frac{d}{dt}\iiint_V W dV
\end{aligned}$$

where, $\iiint_V W dV$ is strain energy and W is starin energy per unit volume. This completes the proof. □

Uniqueness To prove the uniqueness of solutions,consider an elastic body in motion sunjected to a specific given body force F_i. In addition to body force, either surface force $T_i^{(n)}$ or surface displacements u_i are prescribed on the boundary.

Let us assume that it is possible to obtain two sets of solutions $u_i^{'}$, $e_{ij}^{'}$, $\tau_{ij}^{'}$ and $u_i^{''}$, $e_{ij}^{''}$, $\tau_{ij}^{''}$, which satisfy 15 basic equations of elasticity and boundary conditions. Let us define

$$u_i = u_i^{'} - u_i^{''},\; \tau_{ij} = \tau_{ij}^{'} - \tau_{ij}^{''},\; e_{ij} = e_{ij}^{'} - e_{ij}^{''}$$

For the first state of stress, we have

$$\rho F_i + \tau_{ij,j}^{'} = \rho \ddot{u}_i^{'}$$

and the boundary condition

$$\begin{aligned}
\tau_{ij,j}^{'} n_j &= f_i(x_1, x_2, x_3, t) \;\; \text{[if surface forces are prescribed]} \\
\text{or, } u_i^{'} &= g_i(x_1, x_2, x_3, t) \;\; \text{[if boundary displacement are prescribed]}
\end{aligned}$$

Similarly, for the second state of stress,

$$\rho F_i + \tau''_{ij,j} = \rho \ddot{u}''_i$$

and

$$\begin{aligned} \tau''_{ij} n_j &= f_i(x_1, x_2, x_3, t) \\ u''_i &= g_i(x_1, x_2, x_3, t) \end{aligned}$$

Subtracting, we get,

$$\tau'_{ij,j} - \tau''_{ij,j} = \rho(\ddot{u}'_i - \ddot{u}''_i)$$

and either

$$\tau'_{ij} n_j - \tau'''_{ij} n_j = 0$$

or,

$$\ddot{u}'_i - \ddot{u}''_i = 0$$

on the boundary. In other words,

$$\tau_{ij,j} = \rho \ddot{u}_i$$

at every interior point and either

$$\begin{aligned} T_i^{(n)} = \tau_{ij} n_j = 0 \\ or,\ u_i = 0 \ \ for\ t \geq 0 \end{aligned}$$

on the boundary.

Thus, we have a new state in which body forces are absent and surface forces or surface displacements vanish. When on the surface of the body,

$$u_i = 0 \ \ for\ t \geq 0$$

We must have $\dfrac{\partial u_i}{\partial t} = 0$ on the surface for $t \geq 0$. On the surface of the body, the boundary conditions are either $T_i^{(n)} = 0$ for $t \geq 0$, or, $\dfrac{\partial u_i}{\partial t} = 0$ for $t \geq 0$. In either case, $T_i^{(n)} \dfrac{\partial u_i}{\partial t} = 0$ on the surface for $t \geq 0$. Since both the solutions of the problem must satisfy the same initial condition we have

$$\begin{aligned} u_i &= 0 \ \ for\ t = 0 \\ \frac{\partial u_i}{\partial t} &= 0 \ \ for\ t = 0 \end{aligned}$$

Now, we know that

$$\begin{aligned}\frac{d\mathscr{W}}{dt} &= \frac{dK}{dt} + \frac{d}{dt}\iiint_V W dV \\ &= \iiint_V \rho F_i \dot{u}_i dV + \iint_S T_i^{(n)} \dot{u}_i ds\end{aligned}$$

Since for the new state body forces are absent, that is, $F_i = 0$ and $T_i^{(n)} \dfrac{\partial u_i}{\partial t} = 0$ on the surface for $t \geq 0$. So, we have

$$\begin{aligned}&\frac{d\mathscr{W}}{dt} = 0 \\ &\implies \frac{dK}{dt} + \frac{d}{dt}\iiint_V W dV = 0 \\ &\implies K + \iiint_V W dV = \text{constant}\end{aligned}$$

Since $u_i = 0, \dot{u}_i = 0$ for $t = 0$, constant of integration must be zero. Hence,

$$K + \iiint_V W dV = 0$$

Since both kinetic energy K and W are essentially positive definite, we obtain,

$$K = 0 \ \ and \ \ W = 0 \ \ \forall t \geq 0$$

It follows that,

$$\begin{aligned}&\frac{\partial u_i}{dt} = \dot{u}_i = 0 \ \ and \ e_{ij} = 0 \ \forall t \geq 0 \\ \implies\ & u_i = \text{independent of time, and } \frac{1}{2}[u_{i,j} + u_{j,i}] = 0\end{aligned}$$

That is, the solution can represent only rigid body displacement of the body. But the displacement $u_i = 0$ at $t = 0$. Hence, the rigid body displacement must be zero at all parts of the body and at all time. Hence, two solutions are completely identical.

9.9 Few Probable Questions

(a) Express the strain energy density function W in the form

$$W = -\frac{\sigma}{2E}\Theta^2 + \frac{1+\sigma}{2E}\tau_{ij}\tau_{ij}, \quad i, j = 1, 2, 3$$

where $\Theta = \tau_{ii}$= sum of the normal stresses and τ_{ij} = stress tensor.

(b) State the fundamental boundary value problems of elastostatics.

Unit 10

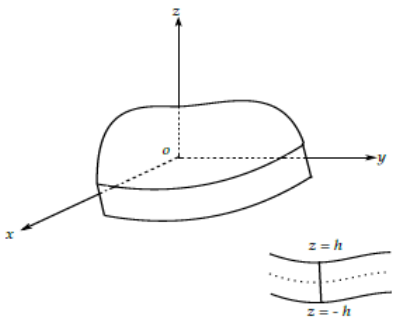
Course Structure

- Two dimensional problems
- Plane stress, Airy's stress function
- .Elastic waves, Waves of dilation distortion

10.1 Plane Stress

A state of stress in a plate is said to be a plane stress if the stress vector on planes parallel to the base is zero throughout its volume. A body is in the plane stress parallel to the xy-plane where these stress components τ_{xz}, τ_{yz}, τ_{zz} vanish.

Let the middle plane of the plate of thickness $2h$ be taken as coordinate plane xy. By



definition, $\tau_{xz} = \tau_{yz} = \tau_{zz} = 0$. Hence, the system of equilibrium equations become

$$\begin{aligned}
\frac{\partial \tau_{xx}}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} + \rho F_x &= 0 \ [\text{since } \tau_{ij,j} + \rho F_i = 0] \\
\frac{\partial \tau_{yx}}{\partial x} + \frac{\partial \tau_{yy}}{\partial y} + \rho F_y &= 0 \\
F_z &= 0
\end{aligned}$$

From constitutive equation we have,

$$\begin{aligned}
\tau_{zz} &= \lambda\theta + 2\mu e_{zz} \\
\implies 0 &= \lambda\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}\right) + 2\mu\frac{\partial w}{\partial z}, \ \text{ since } \tau_{zz} = 0 \text{ and } u, v, w \text{ displacements} \\
\implies \frac{\partial w}{\partial z} &= -\frac{\lambda}{2\mu + \lambda}\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right)
\end{aligned}$$

Substituting these values of $\dfrac{\partial w}{\partial z}$ in the components of stress and strain, we get,

$$\begin{aligned}
\tau_{xx} &= \lambda\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}\right) + 2\mu\frac{\partial u}{\partial x} \\
&= (\lambda + 2\mu)\frac{\partial u}{\partial x} + \lambda\left(\frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}\right) \\
&= (\lambda + 2\mu)\frac{\partial u}{\partial x} + \lambda\left[\frac{\partial v}{\partial y} - \frac{\lambda}{2\mu + \lambda}\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right)\right] \\
&= \left(\lambda - \frac{\lambda^2}{2\mu + \lambda}\right)\frac{\partial u}{\partial x} + \left(\lambda - \frac{\lambda^2}{2\mu + \lambda}\right)\frac{\partial v}{\partial y} + 2\mu\frac{\partial u}{\partial x} \\
&= \frac{2\lambda\mu}{2\mu + \lambda}\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right) + 2\mu\frac{\partial u}{\partial x} \\
&= \lambda^*\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right) + 2\mu\frac{\partial u}{\partial x} \quad \text{where } \lambda^* = \frac{2\lambda\mu}{2\mu + \lambda}
\end{aligned}$$

Similarly,

$$\begin{aligned}
\tau_{yy} &= \lambda\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}\right) + 2\mu\frac{\partial v}{\partial y} \\
&= (\lambda + 2\mu)\frac{\partial v}{\partial y} + \lambda\left(\frac{\partial u}{\partial x} + \frac{\partial w}{\partial z}\right) \\
&= (\lambda + 2\mu)\frac{\partial v}{\partial y} + \lambda\left[\frac{\partial u}{\partial x} - \frac{\lambda}{2\mu + \lambda}\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right)\right] \\
&= \left(\lambda - \frac{\lambda^2}{2\mu + \lambda}\right)\frac{\partial u}{\partial x} + \left(\lambda - \frac{\lambda^2}{2\mu + \lambda}\right)\frac{\partial v}{\partial y} + 2\mu\frac{\partial v}{\partial y} \\
&= \frac{2\lambda\mu}{2\mu + \lambda}\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right) + 2\mu\frac{\partial v}{\partial y} \\
&= \lambda^*\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right) + 2\mu\frac{\partial v}{\partial y}
\end{aligned}$$

Also, we can find τ_{xy}.

10.2 Airy's Stress Function

In the solution of a plane problem, the stress components τ_{xx}, τ_{yy}, τ_{xy} must satisfy the equation of equilibrium

$$\begin{aligned} \frac{\partial \tau_{xx}}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} + \rho F_x &= 0 \\ \frac{\partial \tau_{yx}}{\partial x} + \frac{\partial \tau_{yy}}{\partial y} + \rho F_y &= 0 \\ F_z &= 0 \end{aligned} \tag{10.2.1}$$

In the absence of body force, the above equation of equilibrium for plane problem reduces to

$$\begin{aligned} \frac{\partial \tau_{xx}}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} &= 0 \\ \frac{\partial \tau_{yx}}{\partial x} + \frac{\partial \tau_{yy}}{\partial y} &= 0 \end{aligned} \tag{10.2.2}$$

It is observed that these equations will be identically satisfied by choosing a representation

$$\tau_{xx} = \frac{\partial^2 \phi}{\partial y^2}, \ \ \tau_{yy} = \frac{\partial^2 \phi}{\partial x^2}, \ and \ \tau_{xy} = -\frac{\partial^2 \phi}{\partial x \partial y} \tag{10.2.3}$$

where, $\phi = \phi(x, y)$ is called *Airy's Stress function*. This stress function may be Algebraic function, polynomial, trigonometric function, complex function, etc.

The compatibility equation for strain is

$$\frac{\partial^2 e_{xx}}{\partial y^2} + \frac{\partial^2 e_{yy}}{\partial x^2} = 2\frac{\partial^2 e_{xy}}{\partial x \partial y} \tag{10.2.4}$$

Now, from strain-stress relation,

$$\begin{aligned} e_{xx} &= \frac{1}{2\mu}\left[\tau_{xx} - \frac{\lambda}{2(\lambda+\mu)}(\tau_{xx} + \tau_{yy})\right] \\ e_{yy} &= \frac{1}{2\mu}\left[\tau_{yy} - \frac{\lambda}{2(\lambda+\mu)}(\tau_{xx} + \tau_{yy})\right] \\ e_{xy} &= \frac{1}{2\mu}\tau_{xy} \end{aligned} \tag{10.2.5}$$

Therefore,

$$\begin{aligned} \frac{\partial^2 e_{xx}}{\partial y^2} &= \frac{1}{2\mu}\left[\left(1 - \frac{\lambda}{2(\lambda+\mu)}\right)\frac{\partial^2 \tau_{xx}}{\partial y^2} - \frac{\lambda}{2(\lambda+\mu)}\frac{\partial^2 \tau_{yy}}{\partial y^2}\right] \\ \frac{\partial^2 e_{yy}}{\partial x^2} &= \frac{1}{2\mu}\left[\left(1 - \frac{\lambda}{2(\lambda+\mu)}\right)\frac{\partial^2 \tau_{yy}}{\partial x^2} - \frac{\lambda}{2(\lambda+\mu)}\frac{\partial^2 \tau_{xx}}{\partial x^2}\right] \\ \frac{\partial^2 e_{xy}}{\partial x \partial y} &= \frac{1}{2\mu}\frac{\partial^2 \tau_{xy}}{\partial x \partial y} \end{aligned} \tag{10.2.6}$$

Now, using (10.2.6), from (10.2.4), we get

$$\left(1-\frac{\lambda}{2(\lambda+\mu)}\right)\left[\frac{\partial^2\tau_{xx}}{\partial y^2}+\frac{\partial^2\tau_{yy}}{\partial x^2}\right]-\frac{\lambda}{2(\lambda+\mu)}\left[\frac{\partial^2\tau_{yy}}{\partial y^2}+\frac{\partial^2\tau_{xx}}{\partial x^2}\right]$$
$$=2\frac{\partial^2\tau_{xy}}{\partial x\partial y}=\frac{\partial^2\tau_{xy}}{\partial x\partial y}+\frac{\partial^2\tau_{xy}}{\partial x\partial y}=-\frac{\partial^2\tau_{xx}}{\partial x^2}-\frac{\partial^2\tau_{yy}}{\partial y^2}\text{ using (10.2.2)}$$
$$\Rightarrow\ \left(1-\frac{\lambda}{2(\lambda+\mu)}\right)\left[\frac{\partial^2\tau_{xx}}{\partial x^2}+\frac{\partial^2\tau_{xx}}{\partial y^2}+\frac{\partial^2\tau_{yy}}{\partial x^2}+\frac{\partial^2\tau_{yy}}{\partial y^2}\right]=0$$
$$\Rightarrow\ (1-\sigma)\nabla^2(\tau_{xx}+\tau_{yy})=0,\ \text{where},\ \nabla^2\equiv\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}$$
$$\Rightarrow\ \nabla^2(\tau_{xx}+\tau_{yy})=0 \tag{10.2.7}$$

This is the compatibility equation in terms of stresses for the plane stress problem in the absence of body forces. Substituting

$$\tau_{xx}=\frac{\partial^2\phi}{\partial y^2},\ \tau_{yy}=\frac{\partial^2\phi}{\partial x^2}$$

we get,

$$\frac{\partial^4\phi}{\partial x^2\partial y^2}+\frac{\partial^4\phi}{\partial x^4}+\frac{\partial^4\phi}{\partial y^4}+\frac{\partial^4\phi}{\partial x^2\partial y^2}=0$$
$$\Rightarrow\ \frac{\partial^4\phi}{\partial x^4}+2\frac{\partial^4\phi}{\partial x^2\partial y^2}+\frac{\partial^4\phi}{\partial y^4}=0$$
$$\Rightarrow\ \nabla^4\phi=0. \tag{10.2.8}$$

The above compatibility equation is in terms of stress function in the absence of body forces for plane problem. This equation is also known as fourth degree Biharmonic equation.

Thus, the problem of determination of stress distribution in an elastic body in the case of plane stress in the absence of body forces is reduced to that of finding a stress function which satisfies the biharmonic equation.

10.3 Solution by polynomials

In this section, we shall use the inverse method to obtain the solution of some simple plane problem by stress functions in the form of polynomials, assuming that there are no body forces.

Let us assume a polynomial of second degree as

$$\phi(x,y)=\frac{a^2}{2}x^2+b_2xy+\frac{c_2}{2}y^2 \tag{10.3.1}$$

Then from the following equations in the absence of body forces,

$$\tau_{xx}=\frac{\partial^2\phi}{\partial y^2},\ \tau_{yy}=\frac{\partial^2\phi}{\partial x^2},\ \tau_{xy}=-\frac{\partial^2\phi}{\partial x\partial y}$$

we get

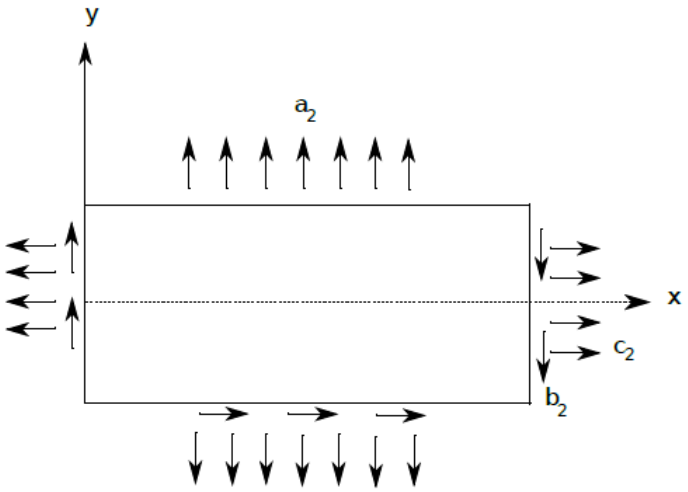
$$\tau_{xx} = c_2,\ \tau_{yy} = a_2,\ \tau_{xy} = -b_2$$

Clearly,

$$\nabla^4 \phi = 0$$

Thus, the stresses are constant throughout the body, that is, the stress function ϕ represents a combination of uniform normal stresses(uniform tension or compression, according as a_2, b_2, c_2 be positive or negative, in two perpendicular directions and uniform shear stresses with no body force.

For a rectangular plate with sides parallel to the coordinate axes, the forces are shown in the figure.



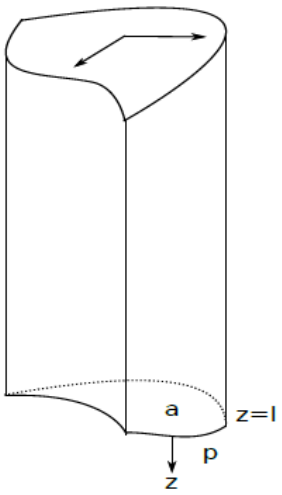
Example: Extension of a beam by longitudinal forces.

Let a force T, along the z axis, be applied at the centre of gravity of the area a of the cross-sectional base $z = l$ of a cylinder. If the stresses giving rise to the force are assumed to be uniformly distributed, then

$$\begin{aligned} \tau_{zz} &= \frac{T}{a} = p, \text{ constant} \\ \tau_{zx} &= \tau_{zy} = 0 \quad \text{with} \quad z = l \end{aligned} \tag{10.3.2}$$

If we assume in absence of body forces

$$\tau_{zz} = p,\ \tau_{xx} = \tau_{yy} = \tau_{xy} = \tau_{yz} = \tau_{zx} = 0$$



throughout this cylinder, then the equations of equilibrium and boundary conditions are obviously satisfied. The Beltrami-Michell compatibility equations are also satisfied.

Now, the displacement components u, v, w can be calculated as

$$\begin{aligned}
\frac{\partial u}{\partial x} &= \frac{1+\sigma}{E}\tau_{xx} - \frac{\sigma}{E}\tau_{zz} \\
&= 0 - \frac{\sigma}{E}\tau_{kk} \\
&= -\frac{\sigma}{E}p \qquad (10.3.3) \\
\frac{\partial v}{\partial y} &= \frac{\sigma}{E}p \qquad (10.3.4) \\
\frac{\partial w}{\partial z} &= \frac{1+\sigma}{E}p - \frac{\sigma}{E}p \\
&= \frac{p}{E} \qquad (10.3.5)
\end{aligned}$$

Now,

$$\begin{aligned}
e_{xy} = 0 &\implies \tfrac{\partial v}{\partial x} + \tfrac{\partial u}{\partial y} = 0 \qquad (10.3.6) \\
e_{yz} = 0 &\implies \tfrac{\partial w}{\partial y} + \tfrac{\partial v}{\partial z} = 0 \qquad (10.3.7) \\
e_{zx} = 0 &\implies \tfrac{\partial u}{\partial z} + \tfrac{\partial w}{\partial x} = 0 \qquad (10.3.8)
\end{aligned}$$

Integrating (10.3.5), we get

$$w = \frac{p}{E}z + w_0(x,y), \qquad (10.3.9)$$

where w_0 is an arbitrary function of x and y.

From (10.3.8), we have,

$$\begin{aligned} \frac{\partial u}{\partial z} &= = -\frac{\partial w}{\partial x} \\ &= -\frac{\partial w_0}{\partial x} \\ \implies u &= -z\frac{\partial w_0}{\partial x} + u_0(x,y) \end{aligned} \tag{10.3.10}$$

where, u_0 is an arbitrary function.
Similarly, from (10.3.7), we get,

$$\begin{aligned} \frac{\partial v}{\partial z} &= = -\frac{\partial w}{\partial y} \\ &= -\frac{\partial w_0}{\partial y} \\ \implies v &= -z\frac{\partial w_0}{\partial y} + v_0(x,y) \end{aligned} \tag{10.3.11}$$

Substituting (10.3.10) in (10.3.3) we get

$$-z\frac{\partial^2 w_0}{\partial x^2} + \frac{\partial u_0}{\partial x} = -\frac{\sigma p}{E}$$

This implies that,

$$\begin{aligned} \frac{\partial^2 w_0}{\partial x^2} = 0, \text{ and } \frac{\partial u_0}{\partial x} &= -\frac{\sigma p}{E} \\ \implies u_0 &= -\frac{\sigma p}{E}x + f_1(y) \end{aligned} \tag{10.3.12}$$

Substituting (10.3.11) in (10.3.4), we get

$$-z\frac{\partial^2 w_0}{\partial y^2} + \frac{\partial v_0}{\partial y} = -\frac{\sigma p}{E}$$

This implies that,

$$\begin{aligned} \frac{\partial^2 w_0}{\partial y^2} = 0, \text{ and } \frac{\partial v_0}{\partial y} &= -\frac{\sigma p}{E} \\ \implies v_0 &= -\frac{\sigma p}{E}y + f_2(x) \end{aligned} \tag{10.3.13}$$

Now, equations (10.3.10), (10.3.11), (10.3.12), (10.3.13) give,

$$\begin{aligned} u &= -z\frac{\partial w_0}{\partial x} + u_0(x,y) \\ &= -z\frac{\partial w_0}{\partial x} - \frac{\sigma p}{E}x + f_1(y) \end{aligned} \tag{10.3.14}$$

$$v = -z\frac{\partial w_0}{\partial y} - \frac{\sigma p}{E}y + f_2(x) \tag{10.3.15}$$

Substituting these values in (10.3.6), we get

$$\frac{\partial v}{\partial x}+\frac{\partial u}{\partial y}=0$$
$$\Longrightarrow\ -z\frac{\partial^2 w_0}{\partial x\partial y}-0+f_2^{'}(x)-z\frac{\partial^2 w_0}{\partial y\partial x}-0+f_1^{'}(y)=0$$
$$\Longrightarrow\ -2z\frac{\partial^2 w_0}{\partial x\partial y}+f_2^{'}(x)+f_1^{'}(y)=0$$

This holds if

$$\frac{\partial^2 w_0}{\partial x\partial y}=0,\ f_2^{'}(x)+f_1^{'}(y)=0$$
$$\Longrightarrow\ \frac{df_2}{dx}=-\frac{df_1}{dy}=\nu\ \text{(say)}$$
$$\Longrightarrow\ f_2=\nu x+b \text{ and } f_1=-\nu y+c$$

Again since

$$\frac{\partial^2 w_0}{\partial x^2}=0,\ \frac{\partial^2 w_0}{\partial y^2}=0,\ \frac{\partial^2 w_0}{\partial x\partial y}=0$$

therefore, w_0 is linear in x and y.

Let $w_0=\beta x+\gamma y+d$. From (10.3.14) and (10.3.15), we have

$$\begin{aligned} u &= -z(-\beta)-\frac{\sigma p}{E}x+(-\gamma y+c)\\ &= -\frac{\sigma p}{E}x-\gamma y+\beta z+c\\ v &= -\frac{\sigma p}{E}y-\gamma z+\nu x+b\\ w &= \frac{p}{E}z-\beta x+\gamma y+d \end{aligned}$$

The parts of u, v, w containing b, c, d and β, γ, ν correspond to rigid motion. Therefore, for pure deformation (in which rigid body motion is absent) we must have

$$u=-\frac{\sigma p x}{E},\ v=-\frac{\sigma p y}{E}\quad\text{and}\quad w=\frac{px}{E}.$$

10.4 Elastic waves

If a disturbance is produced at any point of an alstic medium, waves radiate from that point in all directions. Material particles of the medium undergo small displacements due to this disturbance. If the particle motions occur parallel to the direction of wave propagation, the wave is termed as longitudinal wave while if the particle motions take place perpendicular to the direction of wave propagation, it is called shear wave. The speed of propagation of the two types of plane waves depend on elastic properties of the medium. The waves undergo no dispersion in isotropic homogeneous elastic media.

10.5 Propagation of wave in isotropic elastic media

Waves of dilation and distortion In the absence of body force the equation of motion in isotropic elastic medium are

$$\rho\frac{\partial^2 u}{\partial t^2} = (\lambda+\mu)\vec{\nabla}(\vec{\nabla}.\vec{u}) + \mu\vec{\nabla}^2\vec{u} \tag{10.5.1}$$

Taking vector operation of divergence on both sides of (10.5.1), we get

$$\rho\frac{\partial^2}{\partial t^2}(\vec{\nabla}.\vec{u}) = (\lambda+\mu)\vec{\nabla}.\vec{\nabla}(\vec{\nabla}.\vec{u}) + \mu\vec{\nabla}.(\vec{\nabla}^2\vec{u}) \tag{10.5.2}$$

Now, div grad $\equiv \nabla^2$ and

$$\begin{aligned}
\vec{\nabla}.(\vec{\nabla}^2\vec{u}) &= \frac{\partial}{\partial x}(\vec{\nabla}^2\vec{u}) + \frac{\partial}{\partial y}(\vec{\nabla}^2\vec{u}) + \frac{\partial}{\partial z}(\vec{\nabla}^2\vec{w}) \\
&= \vec{\nabla}^2\frac{\partial u}{\partial x} + \vec{\nabla}^2\frac{\partial u}{\partial y} + \vec{\nabla}^2\frac{\partial w}{\partial z} \\
&= \vec{\nabla}^2\left(\frac{\partial u}{\partial x} + \frac{\partial u}{\partial y} + \frac{\partial w}{\partial z}\right) \\
&= \vec{\nabla}^2\theta \\
&= \vec{\nabla}^2(\vec{\nabla}.\vec{u})
\end{aligned}$$

So, the equation (10.5.2) becomes

$$\begin{aligned}
\rho\frac{\partial^2}{\partial t^2}(\vec{\nabla}.\vec{u}) &= (\lambda+\mu)\vec{\nabla}^2.(\vec{\nabla}.\vec{u}) + \mu\vec{\nabla}^2(\vec{\nabla}.\vec{u}) \\
&= (\lambda+2\mu)\vec{\nabla}^2.(\vec{\nabla}.\vec{u}) \\
\implies \frac{\partial^2\theta}{\partial t^2} &= \frac{\lambda+2\mu}{\rho}(\vec{\nabla}^2\theta) \quad [\text{where } \vec{\nabla}.\vec{u} = \theta] \\
\implies \vec{\nabla}^2\theta &= \frac{1}{c_1^2}\frac{\partial^2\theta}{\partial t^2} \quad \text{where } c_1^2 = \frac{\lambda+2\mu}{\rho}
\end{aligned} \tag{10.5.3}$$

We thus conclude that a change in volume or dilatational disturbance will propagate at the velocity

$$c_1 = \sqrt{\frac{\lambda+2\mu}{\rho}}$$

and such a wave is known as *dilational wave* or *primary wave(p-wave)*.

Next we perform on vector operation curl $\vec{u}$ on both sides of (10.5.1), we obtain

$$\rho\frac{\partial^2}{\partial t^2}(\text{curl}\vec{u}) = (\lambda+\mu)\text{curl grad}(\vec{\nabla}.\vec{u}) + \mu\ \text{curl}(\vec{\nabla}^2\vec{u}) \tag{10.5.4}$$

We know that curl grad$\phi = 0$ and

$$\text{curl}(\vec{\nabla}^2\vec{u}) = \vec{\nabla}^2(\text{curl } \vec{u})$$

Therefore, the above equation becomes,

$$\begin{aligned} \rho\frac{\partial^2}{\partial t^2}(\text{curl } \vec{u}) &= \mu\vec{\nabla}^2(\text{curl } \vec{u}) \\ \implies \vec{\nabla}^2(\text{curl } \vec{u}) &= \frac{1}{c_2^2}\frac{\partial^2}{\partial t^2}(\text{curl } \vec{u}) \text{ where } c_2^2 = \frac{\mu}{\rho} \end{aligned} \tag{10.5.5}$$

Let curl $\vec{u} = 2\vec{w}$, then equation (10.5.5) becomes

$$\vec{\nabla}^2\vec{w} = \frac{1}{c_2^2}\frac{\partial^2\vec{w}}{\partial t^2} \tag{10.5.6}$$

showing that the rotational wave propagate at the velocity

$$c_2 = \sqrt{\frac{\mu}{\rho}}$$

Now suppose that tha displacement vectors are given in such a way that $\theta = 0$, that is div $\vec{u} = 0$. Therefore from (10.5.1), we can write

$$\begin{aligned} \rho\frac{\partial^2 u}{\partial t^2} &= \mu\vec{\nabla}^2\vec{u} \\ \implies \vec{\nabla}^2\vec{u} &= \frac{1}{c_2^2}\frac{\partial^2 u}{\partial t^2} \end{aligned} \tag{10.5.7}$$

Hence the velocity c_2 arises again.

The interpretation of this result is that, equivoluminal wave propagate with the velocity c_2.

Next suppose that rotation $\vec{w} = \vec{0}$, that is, curl $\vec{u} = 0$ which shows that $\vec{u} =$ grad ϕ. Since div grad$= \nabla^2\phi$, in this case, equation (10.5.1) becomes,

$$\begin{aligned} \rho\frac{\partial^2}{\partial t^2}(\vec{\nabla}\phi) &= (\lambda+\mu)\vec{\nabla}(\nabla^2\phi) + \mu\nabla^2(\nabla\phi) \\ &= \vec{\nabla}[(\lambda+\mu)\nabla^2\phi + \mu\nabla^2\phi] \\ &= \vec{\nabla}[(\lambda+2\mu)\vec{\nabla}^2\phi] \\ &= (\lambda+2\mu)\vec{\nabla}^2(\vec{\nabla}\phi) \end{aligned}$$

Therefore,

$$\begin{aligned} \rho\frac{\partial^2\vec{u}}{\partial t^2} &= (\lambda+2\mu)\vec{\nabla}^2 u \\ \implies \vec{\nabla}^2 u &= \frac{1}{c_1^2}\frac{\partial^2\vec{u}}{\partial t^2}, \text{ where } c_1^2 = \frac{\lambda+2\mu}{\rho} \end{aligned}$$

The interpretation of this result is that an irrotational disturbance propagate with velocity c_1.

We thus see that in the material of an elastic solid, wave may propagate with two different velocities. Waves involving no rotation travel with velocity

$$c_1 = \sqrt{\frac{\lambda + 2\mu}{\rho}}$$

and are dilational. While waves involving no dilation, travel with velocity

$$c_2 = \sqrt{\frac{\mu}{\rho}}$$

and are called *distortional wave or secondary wave (s-wave).*

Thus, in an isotropic elastic medium, there are two possible velocities:c_1 and c_2. We note that $c_1 > c_2$, since $\lambda, \mu > 0$.

10.6 Few Probable Questions

(a) An isotropic elastic solid is subjected to the following stress system under no body forces and is in equilibrium: $\tau_{33} = p$(constant), $\tau_{11} = \tau_{22} = \tau_{23} = \tau_{31} = \tau_{12} = 0$. Find the displacement components.

(b) Show that there are two possible plane waves propagating in an isotropic elastic medium.

Unit 11

Course Structure

- Introduction to dynamical system
- Phase portrait
- Fundamental theorem of linear systems

11.1 Introduction to dynamical system

What is a dynamical system? Roughly speaking, a dynamical system is a system that evolves in time. More precisely, a dynamical system has the following three components:

- Some variable that acts like time, i.e. an independent variable that increases monotonically and independently of the evolution of the system, with other variables being indexed by our time variable.
- Some variables that describe the state of a system. These variables define a state space.
- A rule according to which the state evolves in time. We can think of the time evolution either as a time-dependent state, say $x(t)$, or as a trajectory in the state space.

In mathematics, an autonomus system or autonomus differential equation is a system of ordinary differential equations which does not explicitly depend on the independent variable. When the variable is time, the system is also called time invariant systems. An autonomus system is a system of ordinary differential equations of the form

$$\frac{d}{dt}[x(t)] = f(x(t)), \tag{11.1.1}$$

where x can takes values in n-dimensional Euclidean space and t is usually time. It is distinguished from the systems of differential equations of the form

$$\frac{d}{dt}[x(t)] = g[x(t), t], \tag{11.1.2}$$

in which the law governing the rate of motion of a particle depends not only on the particle location, but also on time; such systems are called non-autonomus. In this unit, we will study linear autonomus systems of ordinary differential equations

$$\dot{X} = AX \tag{11.1.3}$$

where $X \in \mathbb{R}^n$, A is an $n \times n$ matrix and $\dot{X} = \dfrac{dX}{dt} = \left[\dfrac{dx_1}{dt}, \ldots, \dfrac{dx_n}{dt}\right]^T$. It can be shown that the solution of the linear system (11.1.3) together with initial condition $X(0) = X_0$ is given by

$$X(t) = X_0 e^{At}. \tag{11.1.4}$$

Theorem 11.1.1. Existance and Uniqueness Theorem: Consider the initial value problem $\dot{X} = f(X),\ X(0) = X_0$. Suppose that $f(x)$ and $f'(x)$ are continuous on an open interval R and suppose that x_0 is a point in R, then the initial value problem has a solution $x(t)$ on some time interval $(-t, t)$ about $t = 0$, and the solution is unique.

We have two types of linear autonomus systems, namely

- Uncoupled linear systems
- Coupled linear systems

For example,

$$\begin{aligned} \dot{x_1} &= -x_1 \\ \dot{x_2} &= 2x_2 \end{aligned}$$

is a two dimensional uncoupled linear system which can be solved by the method of separation of variables, whereas

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

is a two dimensional coupled linear system.

Definition 11.1.2. The motion of a system can be described geometrically by drawing the solution curve on the plane x_1x_2, which is known as phase plane.

Definition 11.1.3. The point at which $\dot{x}_1 = 0 = \dot{x}_2$, i.e., $\dot{X} = 0$, is known as equilibrium point.

Definition 11.1.4. Geometrical representation of all solution curve of the system (11.1.3) is known as phase portrait.

Example 11.1.5. Draw the phase portrait for the uncoupled system

$$\begin{aligned} \dot{x_1} &= -x_1 \\ \dot{x_2} &= 2x_2 \end{aligned}$$

Solution: Solving the given system, we obtain

$$x_1(t) = c_1 e^{-t} \tag{11.1.5}$$
$$x_2(t) = c_2 e^{2t} \tag{11.1.6}$$

where c_1 and c_2 are arbitrary constants. It can be easily observe that solutions starting on the x_1-axis approach the origin as $t \to \infty$ and that the solutions starting on the x_2-axis approach to the origin as $t \to -\infty$ (see Figure. 11.1.1).

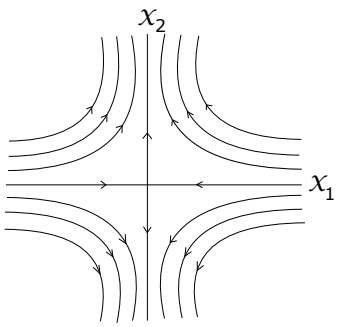


Figure 11.1.1: Phase portrait of the given dynamical system

Now eliminating t, from (11.1.5) and (11.1.6), we have

$$x_1^2 x_2 = c_1^2 c_2 = k, \quad \text{(say)}$$

This $f(x_1, x_2) = x_1^2 x_2 - k$ represent the solution curve for the above system.

Alternatively, the system may be expressed as

$$\dot{X} = AX \quad \text{where} \quad A = \begin{bmatrix} -1 & 0 \\ 0 & 2 \end{bmatrix}$$

Let us now calculate the eigenvalues of A

$$\begin{aligned} |A - \lambda I| = 0 \;\Rightarrow \begin{vmatrix} -1-\lambda & 0 \\ 0 & 2-\lambda \end{vmatrix} &= 0 \\ \Rightarrow (\lambda + 1)(\lambda - 2) &= 0 \\ \Rightarrow \lambda &= -1, 2 \end{aligned}$$

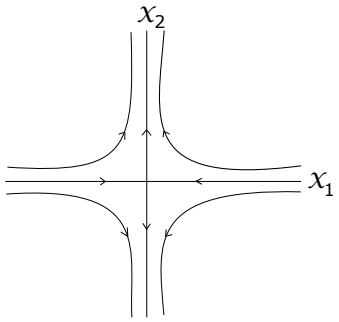


Figure 11.1.2: Phase portrait of the given dynamical system

Hence the eigenvalues are $\lambda_1 = -1$ and $\lambda_2 = 2$. Let us calculate the eigenvector corresponding to the eigenvalue $\lambda_1 = -1$. For this, we consider

$$\begin{aligned} & A \cdot v = -1 \cdot v \\ \Rightarrow & \begin{bmatrix} -1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} -v_1 \\ -v_2 \end{bmatrix} \\ \Rightarrow & \begin{bmatrix} 0 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \end{aligned}$$

Thus the solution set is $\left\{ v_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}$. Letting $v_1 = 1$, we have the eigen vector $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ corresponding to $\lambda_1 = 1$. Similarly, we can calculate eigen vector $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ corresponding to eigen value $\lambda_2 = 2$.

Now let us plot eigen vector in x_1x_2-plane. $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ correspond to x_1- axis while $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ correspond to x_2- axis. Since $\lambda_1 = -1 < 0$, thus arrows are towards origin, while $\lambda_2 = 2 > 0$ thus arrows are away from origin (see Figure 11.1.2).

Now since $e^{2t} > e^{-t}$, hence if we start near x_1- axis the trajectory will have the tendency to move towards the origin but it will change the direction along x_2- axis at the neighbourhood of x_2- axis at the neighbourhood of x_2-axis. Proceeding similarly we can have the phase portrait as figure shown above.

From the previous example, it is clear that we can easily draw phase portrait for a diagonal matrix. If the coefficient matrix is not diagonal the system is known as coupled system. We can use the algebraic technique of diagonalizing a square matrix A which can be used to reduce the linear system $\dot{X} = AX$ to an uncoupled linear system.

First we consider the case of real and distinct eigenvalues of A.

Theorem 11.1.6. If the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of an $n \times n$ matrix A are real and distint, then any set of corresponding eigenvectors $\{v_1, v_2, \ldots, v_n\}$ forms a basis of $\mathbb{R}^n$, the matrix $P = [v_1 \ v_2 \ \ldots \ v_n]$ is invertible and $P^{-1}AP = \text{diag}[\lambda_1, \lambda_2, \ldots, \lambda_n]$.

In order to reduce the system (11.1.3) to an uncoupled linear system using the above theorm, define the linear transformation of coordinates

$$y = P^{-1}x \tag{11.1.7}$$

where P is the invertible matrix defined in the theorem. Then

$$x = Py. \tag{11.1.8}$$

Now differentiating with respect to time t, we have

$$\dot{y} = P^{-1}\dot{x} = P^{-1}Ax = P^{-1}APy \tag{11.1.9}$$

and accordingly to the above theorem, we obtain the uncoupled linear system $\dot{y} = \text{diag}[\lambda_1, \ldots, \lambda_n]y$. Then we can easily obtain the solution of the uncoupled system and will able to draw the phase portrait.

Now since $y(0) = P^{-1}x(0)$ and $x(t) = Py(t)$, it follows that the original system has the solution

$$x(t) = PE(t)P^{-1}x(0), \tag{11.1.10}$$

where $E(t) = \text{diag}[e^{\lambda_1 t}, \ldots, e^{\lambda_n t}]$.

Example 11.1.7. Solve the following linear system and draw the phase portrait.

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

Solution: Here the coefficient matrix A=$\begin{bmatrix} -1 & -3 \\ 0 & 2 \end{bmatrix}$ The eigenvalues are $\lambda_1 = -1$ and $\lambda_2 = 2$ while the corresponding eigenvectors are $v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $v_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$. Hence, the matrix P and its inverse are given by

$$P = \begin{bmatrix} 1 & -1 \\ 0 & 2 \end{bmatrix} \quad \text{and} \quad P^{-1} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

Now,

$$P^{-1}AP = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -1 & -3 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & 2 \end{bmatrix} \tag{11.1.11}$$

Then under the coordinate transformation $y = P^{-1}x$, we obtain the following uncoupled system

$$\begin{aligned} \dot{y_1} &= -y_1 \\ \dot{y_2} &= 2y_2 \end{aligned}$$

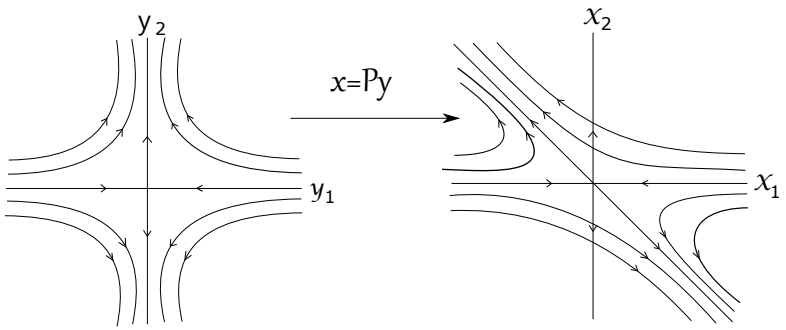


Figure 11.1.3: Transformation of the phase portrait from the uncoupled to coupled dynamical system

which has the general soluton $y_1(t) = c_1e^{-t}$ and $y_2(t) = c_2e^{2t}$. The corresponding phase portrait is shown in Fig.

Now the general solution of the original system is given by

$$x(t) = P\begin{bmatrix} e^{-t} & 0 \\ 0 & e^{2t} \end{bmatrix} P^{-1}C, \quad \text{where} \quad C = X(0)$$

which gives

$$\begin{aligned} x_1(t) &= c_1e^{-t} + c_2(e^{-t} - e^{2t} \\ x_2(t) &= c_2(e^{2t} \end{aligned} \tag{11.1.8}$$

The phase portrait of the original system are shown in Figure 11.1.3. which is nothing but the sketching of the solution curve (11.1.8).

Alternative way to draw the phase portrait:

i) Find eigenvalues

ii) Find corresponding eigenvectors

iii) Draw eigenvectors

iv) Draw the appropriate arrows using the sign of eigenvalues

Theorem 11.1.8. Fundamental theorem of linear systems: Let A be a $n \times n$ matrix. Then for a given $x_0 \in \mathbb{R}^n$, the initial value problem

$$\dot{x} = Ax; \quad x(0) = x_0 \tag{11.1.9}$$

has a unique solution given by

$$x(t) = e^{At}x_0.$$

Proof: Let A be a $n \times n$ matrix. Now

$$\begin{aligned}
\frac{d}{dt}[e^{At}] &= \lim_{h\to 0} \frac{e^{A(t+h)} - e^{At}}{h} = \lim_{h\to 0} e^{At}\frac{e^{Ah} - I}{h} \\
&= e^{At} \lim_{h\to 0} \frac{1}{h}\left[Ah + \frac{A^2h^2}{2!} + \cdots + \frac{A^kh^k}{k!} + \cdots\right] \\
&= e^{At} \lim_{h\to 0} \left[A + \frac{A^2h}{2!} + \cdots + \frac{A^kh^{k-1}}{k!} + \cdots\right] \\
&= Ae^{At}
\end{aligned}$$

Now, if $x(t) = e^{At}x_0$, then differentiating with respect to time t, we have

$$\dot{x}(t) = \frac{d}{dt}[x(t)] = \frac{d}{dt}[e^{At}x_0] = x_0\frac{d}{dt}[e^{At}] = Ax_0e^{At} = Ax(t) \quad \text{for all} \quad t \in \mathbb{R}$$

Also, $x(0) = Ix_0 = x_0$. Thus $x(t) = e^{At}x_0$ is a solution. Now in order to prove the uniqueness, let $x(t)$ be any solution of the given initial value problem and set

$$y(t) = e^{-At}x(t).$$

Now differentiating with respect to time t,

$$\begin{aligned}
\dot{y}(t) &= -Ae^{-At}x(t) + e^{-At}\dot{x}(t) \\
&= -Ae^{-At}x(t) + e^{-At}Ax(t) \\
&= 0
\end{aligned}$$

Thus $y(t)$ is a constant. Setting $t = 0$, we have $y(0) = x(0) = x_0$. Therefore, any solution of the initial value problem is given by

$$x(t) = e^{At}y(t) = c\, e^{At}, \quad \text{where } c\text{is constant}$$

Now, at $t = 0$, $x_0 = x(0) = c$, and hence $x(t) = x_0e^{At}$.

Unit 12

Course Structure

- Phase portrait of linear system

12.1 Phase portrait of linear systems

We have seen that the phase portrait depends on the eigen values of the matrix. The following cases may arrive.

- Eigenvalues are distinct and of opposite sign,
- Eigenvalues are distinct and both are negative,
- Eigenvalues are distinct and both are positive,
- Eigenvalues are complex number with positive real part,
- Eigenvalues are complex number with negative real part,
- Eigenvalues are purely imaginary number,
- Eigenvalues are real and equal.

Let us consider the following linear system

$$\begin{aligned} \frac{dx}{dt} &= ax + by \\ \frac{dy}{dt} &= cx + dy \end{aligned} \tag{12.1.1}$$

A point (x_0, y_0) is said to be a critical point if $\dfrac{dx}{dt} = 0 = \dfrac{dy}{dt}$ at (x_0, y_0). It can be easily seen that $(0, 0)$ is the critical point of the above system.

The characteristic equation is given by

$$\lambda^2 - (a + d)\lambda + (ad - bc) = 0. \tag{12.1.2}$$

Case 1. If the roots λ_1 and λ_2 of the characteristic equation are real unequal and of the same sign, then the critial point $(0, 0)$ of the linear system (12.1.1) is termed as *node*.

Sub-case 1. Both the eigenvalues are positive $\lambda_1 > 0$ and $\lambda_2 > 0$.

For example, let us consider the following system

$$\begin{aligned} \dot{x} &= -x + 4y \\ \dot{y} &= -2x + 5y \end{aligned}$$

whose coefficient matrix is $A = \begin{bmatrix} -1 & 4 \\ -2 & 5 \end{bmatrix}$. Computing the eigenvalues and eigenvectors we have $\lambda_1 = 1,\ v_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ and $\lambda_2 = 3,\ v_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. The phase portrait is shown in Fig. 12.1.1

Since both the $\lambda_1 > 0,\ \lambda_2 > 0$, therefore the critical point is unstable node.

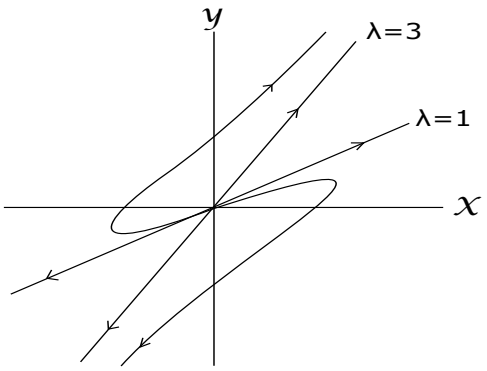


Figure 12.1.1: Phase portrait for sub-case 1

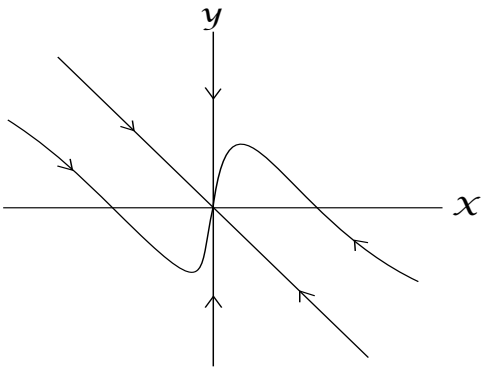


Figure 12.1.2: Phase portrait for sub-case 2

Sub-case 2. Both the eigenvalues are negative $\lambda_1 < 0$ and $\lambda_2 < 0$.

For example, let us consider the following system

$$\begin{aligned} \dot{x} &= -3x \\ \dot{y} &= 3x - 2y \end{aligned}$$

whose coefficient matrix is $A = \begin{bmatrix} -3 & 0 \\ 3 & -2 \end{bmatrix}$. Computing the eigenvalues and eigenvectors we have $\lambda_1 = -3,\ v_1 = \begin{bmatrix} 1 \\ -3 \end{bmatrix}$ and $\lambda_2 = -2,\ v_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. The phase portrait is shown in Fig. 12.1.2

Since both the $\lambda_1 < 0,\ \lambda_2 < 0$, therefore the critical point is stable node.

Case 2. If the roots λ_1 and λ_2 of the characteristic equation are real, unequal and of opposite sign, then the critical point $(0, 0)$ of the linear system (12.1.1) is called *saddle point.*

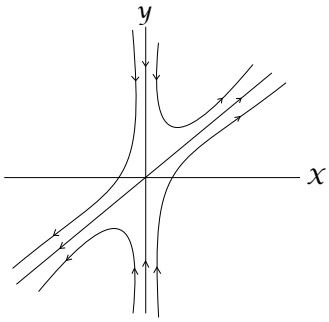


Figure 12.1.3: Phase portrait of the given dynamical system

For example, let us consider the following system

$$\begin{aligned} \dot{x} &= 4x \\ \dot{y} &= 2x - y \end{aligned}$$

whose coefficient matrix is $A = \begin{bmatrix} 4 & 0 \\ 2 & -1 \end{bmatrix}$. Computing the eigenvalues and eigenvectors we have $\lambda_1 = 4,\ v_1 = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$ and $\lambda_2 = -1,\ v_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. The phase portrait is shown in Fig. 12.1.3.

Since $\lambda_1 > 0,\ \lambda_2 < 0$, therefore the critical point is a saddle point.

Case 3. If the roots λ_1 and λ_2 of the characteristic equation are real and equal, then the critical point $(0, 0)$ of (12.1.1) is called *node.*

Sub-case 1. Both the eigenvalues are positive, i.e., $\lambda_1 > 0$ and $\lambda_2 > 0$ and equal $\lambda_1 = \lambda_2$.

For example, let us consider the following system

$$\begin{aligned} \dot{x} &= 2x - 3y \\ \dot{y} &= \frac{x}{3} + 4y \end{aligned}$$

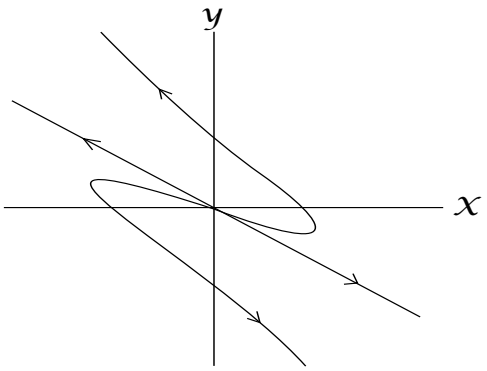


Figure 12.1.4: Phase portrait for sub-case 1

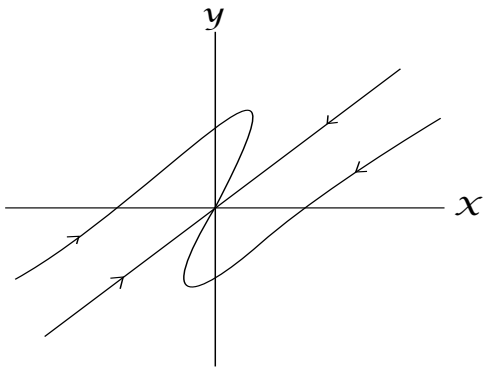


Figure 12.1.5: Phase portrait for sub-case 2

whose coefficient matrix is $A = \begin{bmatrix} 2 & -3 \\ \frac{1}{3} & 4 \end{bmatrix}$. Computing the eigenvalues and eigenvectors we have $\lambda_1 = 3 = \lambda_2,\ v_1 = \begin{bmatrix} 3 \\ -1 \end{bmatrix} = v_2$. The phase portrait is shown in Fig. 12.1.4.

Since $\lambda_1 = \lambda_2 = 3 > 0$, therefore the critical point $(0, 0)$ is unstable node.

Sub-case 2. Both the eigenvalues are negative, i.e., $\lambda_1 < 0$ and $\lambda_2 < 0$ and equal, i.e., $\lambda_1 = \lambda_2$.

For example, let us consider the following system

$$\begin{aligned} \dot{x} &= -7x + y \\ \dot{y} &= -4x - 3y \end{aligned}$$

whose coefficient matrix is $A = \begin{bmatrix} -7 & 1 \\ -4 & -3 \end{bmatrix}$. Computing the eigenvalues and eigenvectors we have $\lambda_1 = -5 = \lambda_2,\ v_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} = v_2$. The phase portrait is shown in Fig. 12.1.5.

Since $\lambda_1 < 0,\ \lambda_2 < 0$, therefore the critical point $(0, 0)$ is stable node.

Case 4. The roots λ_1 and λ_2 of the characteristic equation are conjugate complex with real part not zero (i.e., is not purely imaginary) the critical point $(0, 0)$ of the linear (12.1.1) is termed as *spiral point.*

Sub-case 1. Eigenvalues are complex number with negative real part.

For example, let us consider the following system

$$\begin{aligned} \dot{x} &= -2x + 3y \\ \dot{y} &= -3x - 2y \end{aligned}$$

whose coefficient matrix is $A = \begin{bmatrix} -2 & 3 \\ -3 & -2 \end{bmatrix}$. Computing the eigenvalues we have $\lambda_{1,2} = -2 \pm 3i$ with real$(\lambda) = -2 < 0$. The phase portrait is shown in Fig. 12.1.6.

Thus, the critical point $(0, 0)$ is stable spiral.

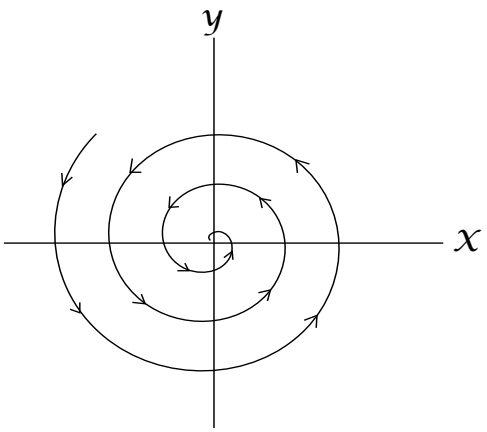


Figure 12.1.6: Phase portrait for sub-case 1

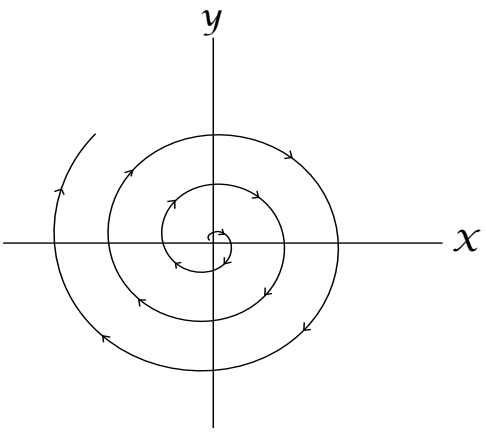


Figure 12.1.7: Phase portrait for sub-case 2

Sub-case 2. Eigenvalues are complex numbers with positive real part.

For example, let us consider the following system

$$\begin{aligned} \dot{x} &= 2x + 3y \\ \dot{y} &= -3x + 2y \end{aligned}$$

whose coefficient matrix is $A = \begin{bmatrix} 2 & 3 \\ -3 & 2 \end{bmatrix}$. Computing the eigenvalues we have $\lambda_{1,2} = 2 \pm 3i$, with real$(\lambda) = 2 > 0$. The phase portrait is shown in Fig. 12.1.7.

Therefore, the critical point $(0, 0)$ is unstable spiral.

Case 5. The roots λ_1 and λ_2 of the characteristic equation are purely imaginary, then the critical point $(0, 0)$ of the linear system is a centre.

For example, let us consider the following system

$$\begin{aligned} \dot{x} &= y \\ \dot{y} &= -5x \end{aligned}$$

whose coefficient matrix is $A = \begin{bmatrix} 0 & 1 \\ -5 & 0 \end{bmatrix}$. Computing the eigenvalues, we have $\lambda_{1,2} = \pm 5i$, with real$(\lambda) = 0$. The phase portrait is shown in Fig. 13.1.2.

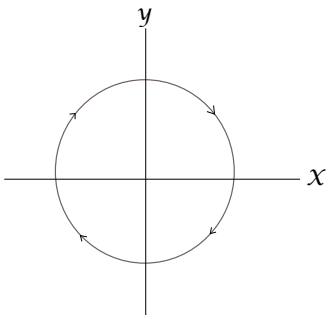


Figure 12.1.8: Phase portrait of the given dynamical system

Therefore the critical point $(0, 0)$ is a centre.

Unit 13

Course Structure

- Criteria for critical points and Stability
- Dynamical system with complex and multiple eigenvalues

13.1 Criteria for Critical Points: Stability

We continue our discussion of homogeneous linear systems with constant coefficients given by

$$\mathbf{y}' = \mathbf{A}\mathbf{y} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \mathbf{y}, \tag{13.1.1}$$

in components,

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

From the examples in the last unit, we have seen that we can obtain an overview of families of solution curves if we represent them parametrically as and graph them as curves in the y_1y_2-plane, called the phase plane. Such a curve is called a trajectory of (13.1.1), and their totality is known as the phase portrait of (13.1.1).

Now we have seen that solutions are of the form

$$\mathbf{y}(t) = \mathbf{x}e^{\lambda t}.$$

Substitution into (13.1.1) gives

$$\mathbf{y}'(t) = \lambda\mathbf{x}e^{\lambda t} = \mathbf{A}\mathbf{y} = \mathbf{A}\mathbf{x}e^{\lambda t}.$$

Dropping the common factor $e^{\lambda t}$, we have

$$\mathbf{Ax} = \lambda \mathbf{x}. \tag{13.1.2}$$

Hence $\mathbf{y}(t)$ is a (nonzero) solution of (13.1.1) if λ is an eigenvalue of $\mathbf{A}$ and $\mathbf{x}$ a corresponding eigenvector.

Our examples in the last section show that the general form of the phase portrait is determined to a large extent by the type of critical point of the system (13.1.1) defined as a point at which dy_2/dy_1 becomes undetermined, $0/0$; here

$$\frac{dy_2}{dy_1} = \frac{y_2' dt}{y_1' dt} = \frac{a_{21}y_1 + a_{22}y_2}{a_{11}y_1 + a_{12}y_2}. \tag{13.1.3}$$

We also recall that there are various types of critical points. What is now new, is that we shall see how these types of critical points are related to the eigenvalues. The latter are solutions $\lambda = \lambda_1$ and λ_2 of the characteristic equation

$$\det(\mathbf{A} - \lambda \mathbf{I}) = \begin{vmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{vmatrix} = \lambda^2 - (a_{11} + a_{12})\lambda + \det\mathbf{A} = 0. \tag{13.1.4}$$

This is a quadratic equation $\lambda^2 - p\lambda + q = 0$ with coefficients $p,\ q$ and discriminant Δ given by

$$p = a_{11} + a_{22}, \quad q = \det\mathbf{A} = a_{11}a_{22} - a_{12}a_{21}, \quad \Delta = p^2 - 4q \tag{13.1.5}$$

From algebra we know that the solutions of this equation are

$$\lambda_1 = \frac{1}{2}(p + \sqrt{\Delta}), \quad \lambda_2 = \frac{1}{2}(p - \sqrt{\Delta}). \tag{13.1.6}$$

Furthermore, the product representation of the equation gives

$$\lambda^2 - p\lambda + q = (\lambda - \lambda_1)(\lambda - \lambda_2) = \lambda^2 - (\lambda_1 + \lambda_2)\lambda + \lambda_1\lambda_2.$$

Hence p is the sum and q the product of the eigenvalues. Also $\lambda_1 - \lambda_2 = \sqrt{\Delta}$ from (13.1.6). Together,

$$p = \lambda_1 + \lambda_2, \quad q = \lambda_1\lambda_2, \quad \Delta = (\lambda_1 - \lambda_2)^2.$$

This gives the criteria in Table 13.1 for classifying critical points.

If $q = \lambda_1\lambda_2 > 0$, both of the eigenvalues are positive or both are negative or complex conjugates. If also $p = \lambda_1 + \lambda_2 < 0$, both are negative or have a negative real part. Hence P_0 is stable and attractive. The reasoning for the other two lines in Table 13.2 is similar.

If $\Delta < 0$, the eigenvalues are complex conjugates, say, $\lambda_1 = \alpha + i\beta$ and $\lambda_2 = \alpha - i\beta$. If also $p = \lambda_1 + \lambda_2 = 2\alpha < 0$, this gives a spiral point that is stable and attractive. If $p = 2\alpha > 0$, this gives an unstable spiral point.

If $p = 0$, then $\lambda_2 = -\lambda_1$ and $q = \lambda_1\lambda_2 = -\lambda^2$. If also $q > 0$, then $\lambda_1^2 = -q < 0$, so that λ_1, and thus λ_2, be pure imaginary. This gives periodic solutions, their trajectories being closed around P_0, which is a center.

Table 13.1: Eigenvalue Criteria for Critical Points

Name	$p = \lambda_1 + \lambda_2$	$q = \lambda_1\lambda_2$	$\Delta = (\lambda_1 - \lambda_2)^2$	Comments on $\lambda_1,\ \lambda_2$
Node		$q > 0$	$\Delta \geq 0$	Real, same sign
Saddle point		$q > 0$		Real, opposite signs
Center	$p = 0$	$q > 0$		Purely imaginary
Spiral point	$p \neq 0$		$\Delta < 0$	Complex, not pure imaginary

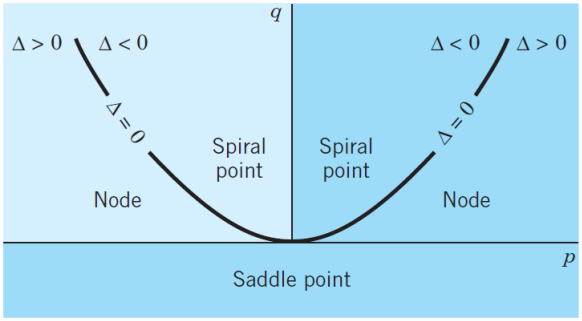


Figure 13.1.1: Stability chart of the dynamical system with $p,\ q,\ \Delta$

Definition 13.1.1. A steady state $P_0 = (x_1^*, y_1^*)$ is called stable if a solution which starts nearby stays nearby. More precisely, (x_1^*, y_1^*) is stable if for all $\epsilon > 0$, there exists a $\delta > 0$ such that solutions to initial data (x_1^0, y_1^0) with $||(x_1^0, y_1^0) - (x_1^*, y_1^*)|| < \delta$ satisfy

$$||(x_1(t), x_2(t)) - (x_1^*, y_1^*)|| < \epsilon$$

for all time $t > 0$. Here $|| \cdot ||$ denotes the Euclidean vector norm.

Definition 13.1.2. A steady state $P_0 = (x_1^*, y_1^*)$ which is not stable is called unstable. For a dynamical system, there is atleast one solution which diverges from (x_1^*, y_1^*).

Definition 13.1.3. A steady state $P_0 = (x_1^*, y_1^*)$ is called asymptotically stable if $P_0 = (x_1^*, y_1^*)$ is stable and all solutions near $P_0 = (x_1^*, y_1^*)$ converges to (x_1^*, y_1^*). More precisely, $P_0 = (x_1^*, y_1^*)$ is asymptotically stable if $P_0 = (x_1^*, y_1^*)$ is stable and there exists a $\delta > 0$ such that all solutions with initial data (x_1^0, y_1^0), with $||(x_1^0, y_1^0) - (x_1^*, y_1^*)|| < \delta$ satisfy

$$\lim_{t \to \infty} ||(x_1^0, y_1^0) - (x_1^*, y_1^*)|| = 0.$$

Exercise 13.1.4. Given the linear system

$$\begin{aligned} \dot{x_1} &= ax_1 - bx_2 \\ \dot{x_2} &= bx_1 + ax_2 \end{aligned}$$

Table 13.2: Stability Criteria for Critical Points

Type of Stability	$p = \lambda_1 + \lambda_2$	$q = \lambda_1\lambda_2$
Stable and attractive	$p < 0$	$q > 0$
Stable	$p \leq 0$	$q > 0$
Unstable	$p > 0$	$q < 0$

Differentiate the equations $r^2 = x_1^2 + x_2^2$ and $\theta = \tan^{-1}\left(\frac{x_2}{x_1}\right)$ with respect to time t in order to obtain

$$\dot{r} = \frac{x_1\dot{x_1} + x_2\dot{x_2}}{r} \quad \text{and} \quad \dot{\theta} = \frac{x_1\dot{x_2} - x_2\dot{x_1}}{r^2}; \quad r \neq 0$$

For the linear system given above, show that these equations reduces to

$$\dot{r} = ar \quad \text{and} \quad \dot{\theta} = b.$$

Solve these equations with the initial conditions $r(0) = r_0$ and $\theta(0) = \theta_0$ and draw the phase portrait for all the possible values of a and b.

13.1.1 Complex Eigenvalues

If the $2n \times 2n$ matrix A has $2n$ distinct complex eigenvalues $\lambda_j = a_j + ib_j$ and $\overline{\lambda}_j = a_j - ib_j$ and the corresponding eigen vectors are $w_j = u_j + iv_j$ and $\overline{w}_j = u_j - iv_j,\ j = 1, \cdots, n$, then $\{u_1, v_1, \cdots, u_n, v_n\}$ is a basis for $\mathbb{R}^{2n}$, the matrix

$$P = [v_1 \ \ u_1 \ \ v_2 \ \ u_2 \ \ \cdots \ \ v_n \ \ u_n]$$

is invertiable and $P^{-1}AP = \text{diag} \begin{bmatrix} a_j & -b_j \\ b_j & a_j \end{bmatrix}$ a real $2n \times 2n$ matrix with 2×2 blocks along the diagonal.

Remark: Note that if instead of the matrix P, we use the invertiable matrix $Q = [u_1 \ v_1 \ u_2 \ v_2 \ \cdots \ u_n \ v_n]$ then $Q^{-1}AQ = \text{diag} \begin{bmatrix} a_j & b_j \\ -b_j & a_j \end{bmatrix}$.

Under this hypotheses, the solution of the initial value problem $\dot{X} = AX; X(0) = X_0$ is given by

$$X(t) = P \ \text{ diag } \ e^{a_j t} \begin{bmatrix} \cos(b_j t) & -\sin(b_j t) \\ \sin(b_j t) & \cos(b_j t) \end{bmatrix} \ P^{-1} \ X_0.$$

Example 13.1.5. Solve the initial value problem $\dot{X} = AX;\ X(0) = X_0$ where

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

and roughly draw the phase portrait.

Solution: The matrix A has the eigenvalues

$$\lambda_1 = -3,\ \lambda_2 = 2 + i \ \text{ and } \ \overline{\lambda}_2 = 2 - i.$$

The corresponding eigen vectors are

$$v_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \ \text{ and } \ w_2 = u_2 + iv_2 = \begin{bmatrix} 0 \\ 1+i \\ i \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} + i \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

Thus

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad P^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad P^{-1}AP = \begin{bmatrix} -3 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & 1 & 2 \end{bmatrix}$$

Therefore, the solution of the IVP is given by

$$\begin{aligned} X(t) &= P \begin{bmatrix} e^{-3t} & 0 & 0 \\ 0 & e^{2t}\cos(t) & -e^{2t}\sin(t) \\ 0 & e^{2t}\sin(t) & e^{2t}\cos(t) \end{bmatrix} P^{-1}\, X_0 \\ &= \begin{bmatrix} e^{-3t} & 0 & 0 \\ 0 & e^{2t}[\cos(t) + \sin(t)] & -2e^{2t}\sin(t) \\ 0 & e^{2t}\sin(t) & e^{2t}[\cos(t) - \sin(t)]X_0 \end{bmatrix} \end{aligned}$$

The stable subspace is the x_1-axis and the unstable subspace is the x_2x_3 plane. Thus, the phase portrait is shown in Fig. 13.1.2.

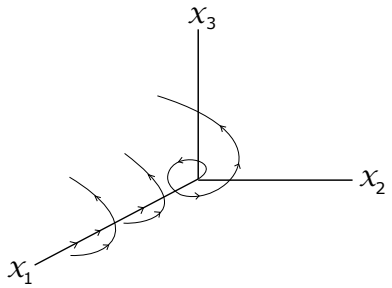


Figure 13.1.2: Phase portrait of the given dynamical system

Exercise 13.1.6. Solve the initial value problem $\dot{X} = AX$ for

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

13.1.2 Multiple eigenvalues

Let A be a real $n \times n$ matrix with real eigenvalues $\lambda_1,\ \lambda_2, \cdots, \lambda_n$ repeated according to their multiplicity. Then there exists a basis of generalised eigen vectors for $\mathbb{R}^n$ and if $\{v_1, \cdots, v_n\}$ is any basis of generalised eigen vectors for $\mathbb{R}^n$, then the matrix $P = [v_1\ \cdots\ v_n]$ is invertiable, so that

$$A = S + N \quad \text{where} \quad P^{-1}SP = \text{diag}[\lambda_j].$$

The matrix $N = A - S$ is nilpotent of order $k \leq n$, also S and N commute, i.e., $SN = NS$.

Under this hypothesis, the linear system $\dot{X} = AX, X(0) = X_0$ has the solution

$$X(t) = P\text{diag}[e^{\lambda_j t}]P^{-1}\left[I + Nt + \cdots + \frac{N^{k-1}t^{k-1}}{(k-1)!}\right]X_0.$$

Note: If λ is an eigenvalue of multiplicity n of an $n \times n$ matrix A, then the above results are particularly easy to apply, since in this case $S = \text{diag}[\lambda]$ with respect to the usual basis and $N = A - S$. The solution to the IVP $\dot{X} = AX;\ X(0) = X_0$ is therefore given by

$$X(t) = e^{\lambda t}\left[I + Nt + \cdots + \frac{N^k t^k}{k!}\right].$$

Example 13.1.7. Solve the IVP $\dot{X} = AX;\ X(0) = X_0$ with $A = \begin{bmatrix} 3 & 1 \\ -1 & 1 \end{bmatrix}$.

Solution: The eigenvalues of A are given by

$$\begin{aligned} & |A - \lambda I| = 0 \\ \Rightarrow & \begin{vmatrix} 3-\lambda & 1 \\ -1 & 1-\lambda \end{vmatrix} = 0 \\ \Rightarrow & (3-\lambda)(1-\lambda) + 1 = 0 \\ \Rightarrow & \lambda^2 - 4\lambda + 4 = 0 \\ \Rightarrow & (\lambda - 2)^2 = 0 \end{aligned}$$

Thus $\lambda_1 = 2$ and $\lambda_2 = 2$. Therefore,

$$S = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \quad \text{and} \quad N = A - S = \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}.$$

Now,

$$N^2 = \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Thus the solution of the IVP is given by

$$\begin{aligned} X(t) &= e^{2t}[I + Nt]X_0 \\ &= e^{2t}\left[\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}t\right]X_0 \\ &= e^{2t}\begin{bmatrix} 1+t & t \\ -1 & 1-t \end{bmatrix}X_0. \end{aligned}$$

Exercise 13.1.8. Solve the IVP $\dot{X} = AX;\ X(0) = X_0$ with $A = \begin{bmatrix} 0 & -2 & -1 & -1 \\ 1 & 2 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$.

Unit 14

Course Structure

- Conversion of an n-th Order ODE to a System
- Linearization of a dynmaical system
- Minimum Variance Unbiased Estimator
- Method of Maximum Likelihood for Estimation of a parameter

14.1 Conversion of an n-th Order ODE to a System

We show that an nth-order ODE can be converted to a system of n first-order ODEs. This is practically and theoretically important - practically because it permits the study and solution of single ODEs by methods for systems, and theoretically because it opens a way of including the theory of higher order ODEs into that of first-order systems. This conversion is another reason for the importance of systems, in addition to their use as models in various basic applications. The idea of the conversion is simple and straightforward, as follows.

Theorem 14.1.1. An n-th order ODE

$$y^{(n)} = F(t, y, y', \cdots, y^{(n-1)} \tag{14.1.1}$$

can be converted to a system of n first-order ODEs by setting

$$y_1 = y, \quad y_2 = y', \quad y_3 = y'', \quad \cdots, \quad y_n = y^{(n-1)} \tag{14.1.2}$$

This system is of the form

$$\begin{aligned} y_1' &= y_2 \\ y_2' &= y_3 \\ &\vdots \\ y_{n-1}' &= y_n \\ y_n' &= F(t, y_1, y_2, \cdots, y_n) \end{aligned} \tag{14.1.3}$$

Proof: The first $n-1$ of these n ODEs follows immediately from (14.1.2) by differentiation. Also, $y_n' = y^{(n)}$ by (14.1.2), so that the last equation in (14.1.3) results from the given ODE (14.1.1).

Example 14.1.2. To gain confidence in the conversion method, let us apply it to the modeling the free motions of a mass on a spring governed by the differential equation

$$my'' + cy' + ky = 0 \quad \text{or} \quad y'' = -\frac{c}{m}y' - \frac{k}{m}y.$$

For this ODE (14.1.1) the system (14.1.3) is linear and homogeneous,

$$\begin{aligned} y_1' &= y_2 \\ y_2' &= -\frac{k}{m}y_1 - \frac{c}{m}y_2. \end{aligned}$$

Setting $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$, we get in matrix form

$$\mathbf{y'} = \mathbf{Ay} = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{c}{m} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

The characteristic equation is

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} -\lambda & 1 \\ -\frac{k}{m} & -\frac{c}{m} - \lambda \end{vmatrix} = \lambda^2 + \frac{c}{m}\lambda + \frac{k}{m} = 0.$$

For an illustrative computation, let $m = 1$, $c = 2$, and $k = 0.75$. Then

$$\lambda^2 + 2\lambda + 0.75 = (\lambda + 0.5)(\lambda + 1.5) = 0.$$

This gives the eigenvalues $\lambda_1 = -0.5$ and $\lambda_2 = -1.5$. Eigenvectors follow from the first equation in $\mathbf{A} - \lambda\mathbf{I} = 0$, which is $-\lambda x_1 + x_2 = 0$. For λ_1 this gives $0.5x_1 + x_2 = 0$, say, $x_1 = 2,\ x_2 = -1$. For $\lambda_2 = -1.5$ it gives $1.5x_1 + x_2 = 0$, say, $x_1 = 1,\ x_2 = -1.5$. These eigenvectors

$$\mathbf{x}^{(1)} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}, \quad \mathbf{x}^{(2)} = \begin{bmatrix} 1 \\ -1.5 \end{bmatrix} \quad \text{give} \quad \mathbf{y} = c_1 \begin{bmatrix} 2 \\ -1 \end{bmatrix} e^{-0.5t} + c_2 \begin{bmatrix} 1 \\ -1.5 \end{bmatrix} e^{-1.5t}.$$

This vector solution has the first component

$$y = y_1 = 2c_1 e^{-0.5t} + c_2 e^{-1.5t}$$

which is the expected solution. The second component is its derivative

$$y_2 = y_1' = y' = -c_1 e^{-0.5t} - 1.5c_2 e^{-1.5t}.$$

14.2 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

$$\begin{bmatrix} \dot{u} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} + \text{Quadratic terms} \tag{14.2.1}$$

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. (14.2.1) 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.

Example 14.2.1. 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.

Example 14.2.2. 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

$$\begin{aligned} x &= r\cos\theta \\ y &= r\sin\theta. \end{aligned}$$

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

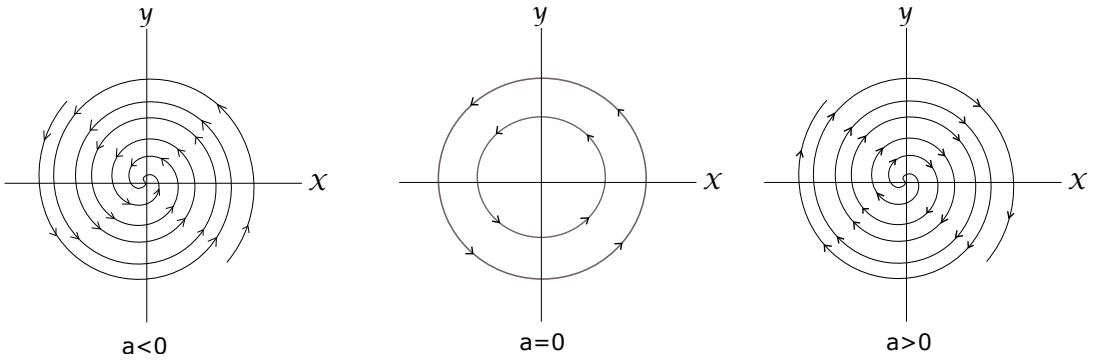
$$\begin{aligned} \dot{r} &= ar^3 \\ \dot{\theta} &= 1 \end{aligned}$$

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.



Unit 15

Course Structure

- Hyperbolic fixed point
- Hartman-Grobman theorem
- Interaction model for two population

Definition 15.0.1. Hyperbolic fixed point: A fixed point of an n-th order dynamical system is hyperbolic if all the eigenvalues of the linearization lie off the imaginary axis, i.e., $\text{Re}(\lambda_i) \neq 0$ for $i = 1, \cdots, n$.

The **Hartman-Grobman theorem** is another important result in the local qualitative theory of ODE. The theorem shows that $x' = f(x)$ with $f(0) = 0$ and its linearized system $x' = Df(0)x$ have the same qualitative structures near a hyperbolic equilibrium point.

Consider the system

$$\dot{x} = f(x) \tag{15.0.1}$$

where $x = 0$ is a hyperbolic equilibrium. The corresponding linarized system is

$$\dot{x} = Ax \quad \text{where} \quad A = Df(0) \tag{15.0.2}$$

Two autonomus system of differential equation such as (15.0.1) and (15.0.2) are said to be topolofically equivalent in a neighbourhood of the origin or to have the same qualitative structure near the origin if there is a homeomorphism H mapping an open set U containing the origin onto an open set V containing the origin which maps trajectories of (15.0.1) in U onto trajectories of (15.0.2) in V and preserves their orientation by time in the sense that if a trajectory is directed from x_1 to x_2 in U, then its image is directed from $H(x_1)$ to $H(x_2)$ in V. If the homeomorphism H preserves the parametrization by time, then the system (15.0.1) and

(15.0.2) are said to be topologically conjugate in a neighbourhood of the origin.

As illustrative example for topologically conjugate consider two linear systems $\dot{x} = Ax$ and $\dot{x} = Bx$ with

$$A = \begin{bmatrix} -1 & -3 \\ -3 & -1 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 2 & 0 \\ 0 & -4 \end{bmatrix}$$

Let $H(x) = Rx$, where $R = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$ and $R^{-1} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$ Then $B = RAR^{-1}$ and letting $y = H(x) = Rx$ which gives

$$\begin{aligned} & x = R^{-1}y \\ \Rightarrow\ & Ax = AR^{-1}y \\ \Rightarrow\ & \dot{x} = AR^{-1}y \quad [\text{since, } \dot{x} = Ax] \\ \Rightarrow\ & R\dot{x} = RAR^{-1}y \\ \Rightarrow\ & \dot{y} = By \quad [\text{since, } B = RAR^{-1} \ \& \ \Rightarrow \dot{y} = B\dot{x}] \end{aligned}$$

Thus if $x(t) = e^{At}x_0$ is the solution of $\dot{x} = Ax$ through x_0, then

$$y(t) = H(x(t)) = Rx(t) = Re^{At}x_0 = e^{Bt}Rx_0$$

is the solution of $\dot{x} = Bx$; i.e., H maps trajectories of $\dot{x} = Ax$ onto trajectories of $\dot{x} = Bx$ and it preserves the parametrization by t since $He^{At} = e^{Bt}H$. Therefore, H is a homomorphism from A onto B.

Theorem 15.0.2. Hartman-Grobman Theorem: If $x = 0$ is a hyperbolic equilibrium point of (15.0.1) and (15.0.2), then there exists a homeomorphism H of an open set U containing the origin onto an open set V containing the origin such that for each $x_0 \in U$, there exists an open interval $I_0 \subset R$ containing the origin such that for all $t \in I_0$

$$H \circ \phi_t(x_0) = e^{At}H(x_0).$$

Example 15.0.3. Consider the dynamical system $\dot{y} = -y;\ \dot{z} = z + y^2$. The solution with the conditions $y(0) = y_0$ and $z(0) = z_0$ is obtained as

$$y(t) = y_0e^{-t}, \quad z(t) = z_0e^t + \frac{y_0^2}{3}(e^t - e^{-2t})$$

The linearized system is given by

$$\begin{aligned} \dot{y} &= -y \\ \dot{z} &= z \end{aligned}$$

Its solution with $y(0) = y_0$ and $z(0) = z_0$ can be easily solved as

$$y(t) = y_0e^{-t}, \quad z(t) = z_0e^t$$

The homeomorphism H is defined as $H(x,y) = \begin{bmatrix} y \\ z + \frac{y^2}{3} \end{bmatrix}$. Then we can verify the result of the Hartman-Grobman theorem as follows:

Let the solution of the original system as

$$\phi_t(y_0, z_0) = \begin{bmatrix} y_0 e^{-t} \\ z_0 e^t + \frac{y_0^2}{3}(e^t - e^{-2t}) \end{bmatrix}$$

and e^{At} of the linarized system is given by $e^{At} = \begin{bmatrix} e^{-t} & 0 \\ 0 & e^t \end{bmatrix}$. Since,

$$e^{At} H(y_0, z_0) = \begin{bmatrix} e^{-t} & 0 \\ 0 & e^t \end{bmatrix} \begin{bmatrix} y_0 \\ z_0 + \frac{y_0^2}{3} \end{bmatrix} = \begin{bmatrix} e^{-t} y_0 \\ e^t \left(z_0 + \frac{y_0^2}{3} \right) \end{bmatrix}$$

and

$$\begin{aligned} H \circ \phi_t(y_0, z_0) &= H \circ \begin{bmatrix} y_0 e^{-t} \\ z_0 e^t + \frac{y_0^2}{3}(e^t - e^{-2t}) \end{bmatrix} = \begin{bmatrix} y \\ z + \frac{y^2}{3} \end{bmatrix}_{y = y_0 e^{-t}; z = z_0 e^t + \frac{y_0^2}{3}(e^t - e^{-2t}} \\ &= \begin{bmatrix} y_0 e^{-t} \\ z_0 e^t + \frac{y_0^2}{3}(e^t - e^{-2t}) + \frac{(y_0 e^{-t})^2}{3} \end{bmatrix} = \begin{bmatrix} e^{-t} y_0 \\ e^t \left(z_0 + \frac{y_0^2}{3} \right) \end{bmatrix} \end{aligned}$$

Hence we have

$$H \circ \phi_t(y_0, z_0) = e^{At} H(y_0, z_0) \quad \text{for all} \quad t \geq 0.$$

Note: Finding a homeomorphism H such that $H \circ \phi_t(x_0) = e^{At} H(x_0)$ is difficult. In fact, the Hartman-Grobman theorem only assures the existence of H. It does not tell us any information on how to find H. Moreover, it is a qualitative property.

15.0.1 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:

$$\begin{aligned} \dot{x} &= \alpha x + \beta xy \\ \dot{y} &= \gamma y + \delta xy \end{aligned} \tag{15.0.3}$$

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 (15.0.3) 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 (15.0.3) 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.

Example 15.0.4. 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

$$\begin{aligned} \dot{S} &= -\beta IS + \gamma R \\ \dot{I} &= \beta IS - \alpha I \\ \dot{R} &= \alpha I - \gamma R \end{aligned} \tag{15.0.4}$$

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

$$\begin{aligned} \dot{S} &= -\beta IS \\ \dot{I} &= \beta IS - \alpha I \end{aligned} \tag{15.0.5}$$

Qualitative Analysis of the epidemic model:

Let us analyse the epidemic model given in (15.0.5). 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 (15.0.5) 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

$$\begin{aligned} I(S) - I(S_0) &= -(S - S_0) + \frac{\alpha}{\beta}(\ln S - \ln S_0) \\ \Rightarrow I(S) &= \frac{\alpha}{\beta}\ln S - S + c_1 \end{aligned}$$

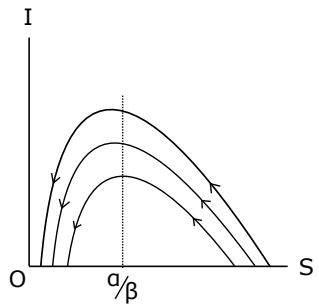


Figure 15.0.1

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. 15.0.1 the steady state to the right of $\frac{\alpha}{\beta}$, namely, $\overline{S} > \frac{\alpha}{\beta}$ are unstable in the direction away from the S-axis, and those to the left of $\frac{\alpha}{\beta}$ are stable.

Biologically, $\frac{\alpha}{\beta}$ represents the critical population size to sustain an epidemic. If the initial susceptible population is below $\frac{\alpha}{\beta}$, then no epidemic is possible and the number of infections decreases, whereas if $S_0 > \frac{\alpha}{\beta}$, then the number of infection initially increases, reaching its maximum when $S = \frac{\alpha}{\beta}$ and then declines.

Exercise 15.0.5. Find all the critical points of the logistic model

$$\begin{aligned} \dot{x} &= x - ax^2 \\ \dot{y} &= y - by^2 \end{aligned}$$

and discuss the stability of each critical points for all possible values of a and b.

Unit 16

Course Structure

- Stability and Liapunov Functions

16.1 Stability and Liapunov Functions

Here we discuss the stability of the equilibrium points of the non-linear system

$$\dot{x} = f(x). \tag{16.1.1}$$

The stability of any hyperbolic equilibrium point x_0 of (16.1.1) 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

$$\frac{dx}{dt} = P(x, y) \tag{16.1.2}$$

$$\frac{dy}{dt} = Q(x, y). \tag{16.1.3}$$

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.

Example 16.1.1. Show that $E(x, y) = x^2 + y^2$ is a Liapunov function for the non-linear system

$$\begin{aligned} \frac{dx}{dt} &= -x + y^2 \\ \frac{dy}{dt} &= -y + x^2. \end{aligned}$$

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 16.1.2. 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.

Note 16.1.3. (a) If $\dot{E} < 0$ for all $x \neq 0$, then $(0, 0)$ is asymptotically stable.

(b) If $\dot{E} > 0$ for all $x \neq 0$, then $(0, 0)$ is unstable.

(c) 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.

Example 16.1.4. Use the Liapunov function $v(x) = x_1^2 + x_2^2$ to establish the following results.

(a) The origin is an asymptotically stable equilibrium point of

$$\dot{X} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} X + \begin{bmatrix} -x_1^3 - x_1 x_2^2 \\ -x_2^3 - x_2 x_1^2 \end{bmatrix}.$$

(b) The origin is an unstable equilibrium point of

$$\dot{X} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} X + \begin{bmatrix} x_1^3 + x_1 x_2^2 \\ x_2^3 + x_2 x_1^2 \end{bmatrix}.$$

(c) The origin is a stable equilibrium point which is not asymptotically stable for

$$\dot{X} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} X + \begin{bmatrix} -x_1 x_2 \\ x_1^2 \end{bmatrix}.$$

Here, $v(x) = x_1^2 + x_2^2$. Differentiating with respect to time t, we have

$$\dot{v}(x_1, x_2) = 2x_1\dot{x}_1 + 2x_2\dot{x}_2. \tag{16.1.4}$$

(a) The system is given by

$$\begin{aligned} \dot{x}_1 &= -x_2 - x_1^3 - x_1 x_2^2 \\ \dot{x}_2 &= x_1 - x_2^3 - x_2 x_1^2. \end{aligned}$$

From (16.1.4),

$$\begin{aligned} \dot{v}(x_1, x_2) &= 2x_1[-x_2 - x_1^3 - x_1 x_2^2] + 2x_2[x_1 - x_2^3 - x_2 x_1^2] \\ &= -2x_1 x_2 - 2x_1^4 - 2x_1^2 x_2^2 + 2x_1 x_2 - 2x_2^4 - 2x_2^2 x_1^2 \\ &= -2[x_1^4 + x_2^4 + 2x_1^2 x_2^2] \\ &= -2(x_1^2 + x_2^2). \end{aligned}$$

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.

(b) The system is equivalent to

$$\begin{aligned} \dot{x}_1 &= -x_2 + x_1^3 + x_1 x_2^2 \\ \dot{x}_2 &= x_1 + x_2^3 + x_2 x_1^2. \end{aligned}$$

Now from (16.1.4),

$$\begin{aligned} \dot{v}(x_1, x_2) &= 2x_1\dot{x}_1 + 2x_2\dot{x}_2 \\ &= 2x_1(-x_2 + x_1^3 + x_1 x_2^2) + 2x_2(x_1 + x_2^3 + x_2 x_1^2) \\ &= -2x_1 x_2 + 2x_1^4 + 2x_1^2 x_2^2 + 2x_1 x_2 + 2x_2^4 + 2x_1^2 x_2^2 \\ &= 2x_1^4 + 2x_2^4 + 4x_1^2 x_2^2 \\ &= 2(x_1^2 x_2^2)^2. \end{aligned}$$

Hence $\dot{v}(0, 0) > 0$ for all $(x_1, x_2) \in \mathbb{R}^2$. Thus, $(0, 0)$ is unstable critical point.

(c) The system is given by

$$\begin{aligned} \dot{x}_1 &= -x_2 - x_1 x_2 \\ \dot{x}_2 &= x_1 + x_1^2. \end{aligned}$$

Now from (16.1.4),

$$\begin{aligned} \dot{v}(x_1, x_2) &= 2x_1\dot{x}_1 + 2x_2\dot{x}_2 \\ &= 2x_1(-x_2 - x_1x_2) + 2x_2(x_1 + x_1^2) \\ &= -2x_1x_2 - 2x_1^2x_2 + 2x_1x_2 + 2x_1^2x_2 \\ &= 0. \end{aligned}$$

Thus the origin is stable equilibrium point which is not asymptotically stable.

Example 16.1.5. 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 $v = I + S - 1 + \ln S$.

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 17

Course Structure

- Limit cycles and periodic solutions
- Existence and Non-existence of limit cycles
- Bendixon's Non-existence criterion, Dulac's criterion

17.1 Limit Cycles and Periodic solutions

Given an autonomous system

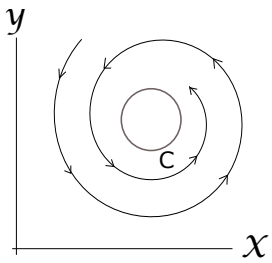
$$\begin{aligned} \frac{dx}{dt} &= P(x,y) \\ \frac{dy}{dt} &= Q(x,y). \end{aligned} \tag{17.1.1}$$

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 (17.1.1). 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.

Definition 17.1.1. A closed path C of the system (17.1.1) which is approached spirally from either the inside or the outside by a non-closed path C_1 of (17.1.1) either as $t \to +\infty$ or $t \to -\infty$ is called a limit cycle of (17.1.1).



The following example of a system having a limit cycle will illustrate the above discussion and definition.

Example 17.1.2. Consider the following system.

$$\begin{aligned} \frac{dx}{dt} &= y + x(1 - x^2 - y^2) \\ \frac{dy}{dt} &= -x + y(1 - x^2 - y^2). \end{aligned} \tag{17.1.2}$$

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

$$\begin{aligned} x\frac{dx}{dt} + y\frac{dy}{dt} &= r\cos\theta\left[-r\sin\theta\frac{d\theta}{dt} + \frac{dr}{dt}\cos\theta\right] + r\sin\theta\left[r\cos\theta\frac{d\theta}{dt} + \frac{dr}{dt}\sin\theta\right] \\ &= r\frac{dr}{dt}. \end{aligned} \tag{17.1.3}$$

Similarly,

$$x\frac{dy}{dt} - y\frac{dx}{dt} = r^2\frac{d\theta}{dt}. \tag{17.1.4}$$

Now, from (17.1.2),

$$\begin{aligned} & x\frac{dx}{dt} + y\frac{dy}{dt} = (x^2 + y^2)(1 - x^2 - y^2) \\ \Rightarrow\quad & r\frac{dr}{dt} = r^2(1 - r^2) \quad \text{[Using (17.1.3)]} \\ \Rightarrow\quad & \frac{dr}{dt} = r(1 - r^2). \end{aligned}$$

Again, from (17.1.2),

$$\begin{aligned} & y\frac{dx}{dt} - x\frac{dy}{dt} = y^2 + x^2 \\ \Rightarrow \quad & -r^2\frac{d\theta}{dt} = r^2 \quad \text{[Using (17.1.4)]} \\ \Rightarrow \quad & \frac{d\theta}{dt} = -1. \end{aligned}$$

Thus in polar coordinate system, we have

$$\frac{dr}{dt} = r(1-r^2) \tag{17.1.5}$$

$$\frac{d\theta}{dt} = -1. \tag{17.1.6}$$

Integrating (17.1.6), we have, $\theta = -t + t_0$, t_0 is constant. From (17.1.5),

$$\begin{aligned} & \frac{dr}{r(1-r^2)} = dt \\ \Rightarrow \quad & \frac{r^2 + (1-r^2)}{r(1-r^2)}dr = dt \\ \Rightarrow \quad & \frac{rdr}{1-r^2} + \frac{dr}{r} = dt \\ \Rightarrow \quad & \frac{2rdr}{1-r^2} + 2\frac{dr}{r} = 2dt. \end{aligned}$$

Integrating, we get

$$\begin{aligned} & \ln r^2 - \ln|1-r^2| = 2t + \ln|C_0| \\ \Rightarrow \quad & \frac{r^2}{1-r^2} = C_0\,\mathrm{e}^{2t} \\ \Rightarrow \quad & r^2 = (1-r^2)C_0\,\mathrm{e}^{2t} \\ \Rightarrow \quad & (1 + C_0\,\mathrm{e}^{2t})r^2 = C_0\,\mathrm{e}^{2t} \\ \Rightarrow \quad & r^2 = \frac{C_0\,\mathrm{e}^{2t}}{1 + C_0\,\mathrm{e}^{2t}} \\ \Rightarrow \quad & r = \frac{1}{\sqrt{1 + C\,\mathrm{e}^{-2t}}}, \end{aligned}$$

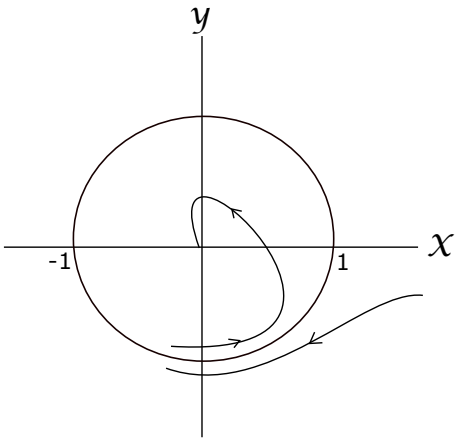
where $C = \dfrac{1}{C_0}$. Thus the solution of the system may be written as

$$\begin{aligned} r &= \frac{1}{\sqrt{1 + C\,\mathrm{e}^{-2t}}}, \\ \theta &= -t + t_0, \end{aligned}$$

where C and t_0 are arbitrary constants. We may choose $t_0 = 0$. Then $\theta = -t$, and hence

$$x = \frac{\cos t}{\sqrt{1 + C\,\mathrm{e}^{-2t}}}, \quad \text{and} \quad y = \frac{\sin t}{\sqrt{1 + C\,\mathrm{e}^{-2t}}}. \tag{17.1.7}$$

If $C = 0$, the path defined by (17.1.7) is the circle $x^2 + y^2 = 1$. If $C \neq 0$, the path defined by (17.1.7) 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.

17.1.1 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

$$\begin{aligned} \frac{dx}{dt} &= P(x, y) \\ \frac{dy}{dt} &= Q(x, y) \end{aligned} \tag{17.1.8}$$

where P and Q have continuous first order partial derivatives in D. Suppose that $\dfrac{\partial P(x,y)}{\partial x} + \dfrac{\partial Q(x,y)}{\partial y}$ has the same sign throughout D. Then the system (17.1.8) 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 (17.1.8). Let $x = f(t)$, $y = g(t)$ be an arbitrary solution of (17.1.8), 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 $\dfrac{\partial P}{\partial x} + \dfrac{\partial Q}{\partial y}$ changes sign. This is a contradiction. Thus C is not a path of (17.1.8) and hence (17.1.8) possesses no closed path in D. □

Example 17.1.3. Show that the following system has no closed path

$$\begin{aligned}\frac{dx}{dt} &= 2x + y + x^3 \\ \frac{dy}{dt} &= 3x - y + y^3.\end{aligned}$$

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.

Example 17.1.4. 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.

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 $t \to \infty$.

Exercise 17.1.5. (a) 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.

(b) Show that $v = ax^2 + 2bxy + cy^2$ is positive definite if and only if $a > 0$ and $ac - b^2 > 0$.

(c) 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]

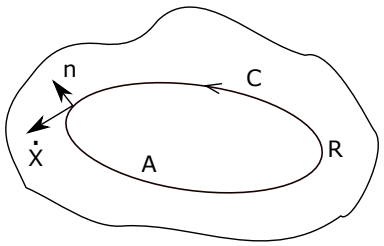
17.1.2 Dulac's Criterion

This is a method for ruling out closed orbits is based on Green's theorem, and is known as Dulac's criterion.

Theorem 17.1.6. 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. □

Note 17.1.7. 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

$$g = 1,\ \frac{1}{x^\alpha y^\beta},\ \mathrm{e}^{ax},\ \text{and}\ \mathrm{e}^{ay}.$$

Example 17.1.8. 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 $x, y > 0$.

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.

Example 17.1.9. Show that the system $\dot{x} = y$, $\dot{y} = -x - y - x^2 + y^2$ has no closed orbits.

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.

Exercise 17.1.10. (a) 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 $N_1, N_2 > 0$.

(b) 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 $g(x, y) = \dfrac{1}{y}$.

(c) 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 18

Course Structure

- Bifurcation
- Saddle-Node bifurcation

18.1 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.

Definition 18.1.1. 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.

18.1.1 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.

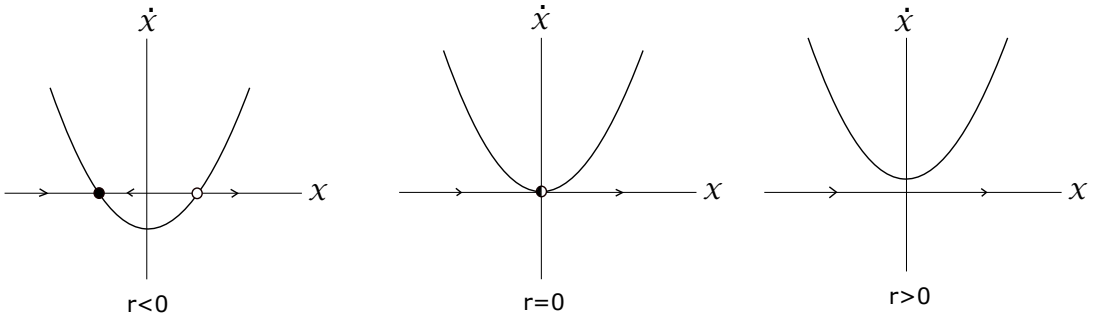


Figure 18.1.1: Saddle-Node Bifurcation

- 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 18.1.2. 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.

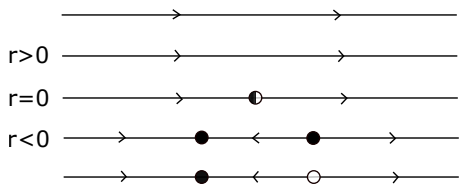


Figure 18.1.2

The most common way to depict the bifurcation is to invert the axis of Figure 18.1.3. The rationale is that r plays the role of an independent variable, and so should be plotted horizontally (Figure 18.1.4). 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.

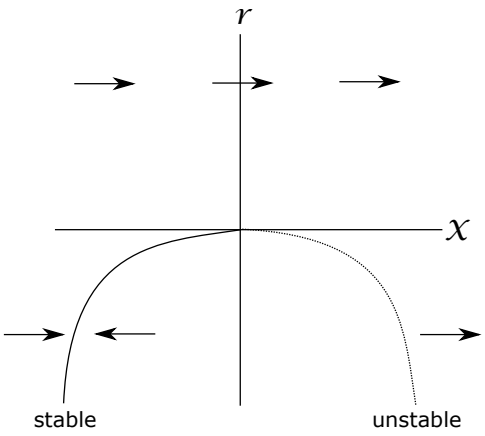


Figure 18.1.3

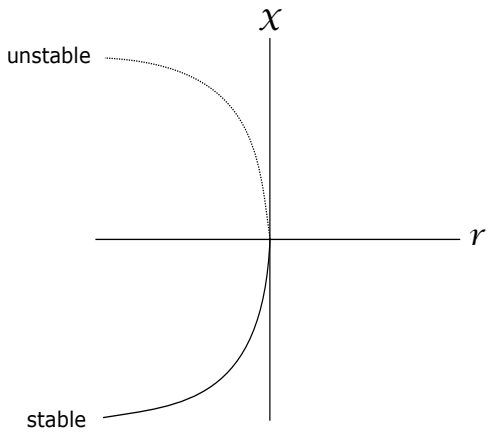


Figure 18.1.4

Example 18.1.2. 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.

Using the Taylor series expansion for e^{-x} about $x = 0$, we have

$$\mathrm{e}^{-x} = 1 - x + \frac{x^2}{2!} - \dots$$

Now,

$$\begin{aligned}\dot{x} &= r - x - \mathrm{e}^{-x} \\ &= r - x - \left[1 - x + \frac{x^2}{2!} - \ldots\right] \\ &= (r-1) - \frac{x^2}{2!} + \ldots\end{aligned}$$

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 $r^* = 1$.

Unit 19

Course Structure

- Transcritical bifurcation
- Pitchfork bifurcation

19.0.1 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.

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.

Note 19.0.1. 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 19.0.2 shows the bifurcation diagram for the transcritical bifurcation.

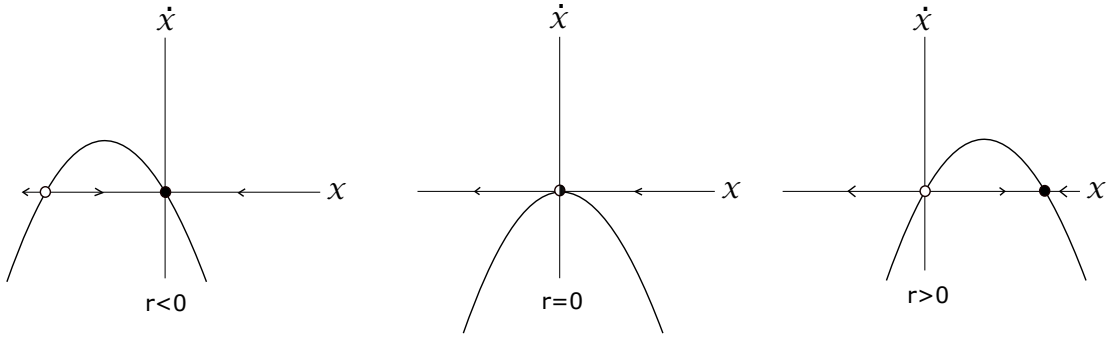


Figure 19.0.1: Transcritical Bifurcation

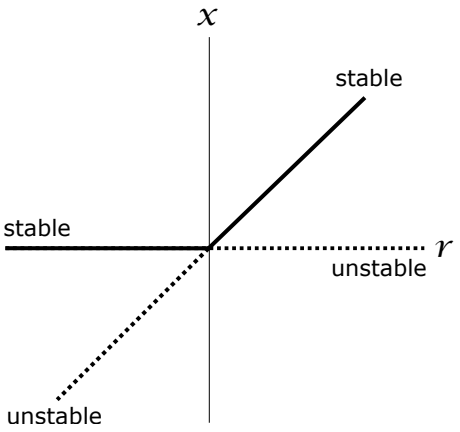


Figure 19.0.2: Bifurcation Diagram

Example 19.0.2. 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.

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

$$(1 - ab) + \frac{1}{2}b^2x^* \simeq 0 \Rightarrow x^* \simeq \frac{2(ab-1)}{ab^2}.$$

Example 19.0.3. Show that the system

$$\dot{x} = r\ln(x) + x - 1$$

undergoes a transcritical bifurcation at a certain value of r.

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 $r + 1 = 0 \Rightarrow r = -1$.

19.0.2 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

$$\dot{x} = rx - x^3. \tag{19.0.1}$$

Note 19.0.4. 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 (19.0.1) again. This invariance is the mathematical expression of the left right symmetry.

The following figure shows the vector field for different values of r.

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 $x^* = \pm\sqrt{r}$.

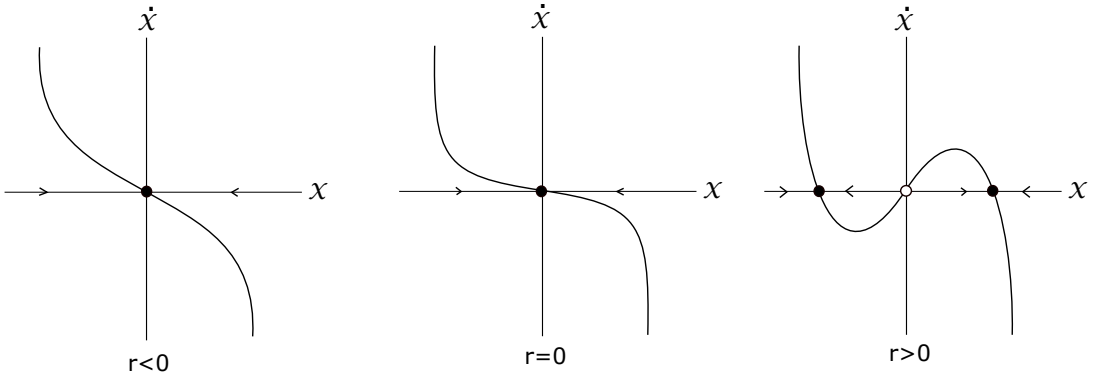


Figure 19.0.3: Supercritical Pitchfork Bifurcation

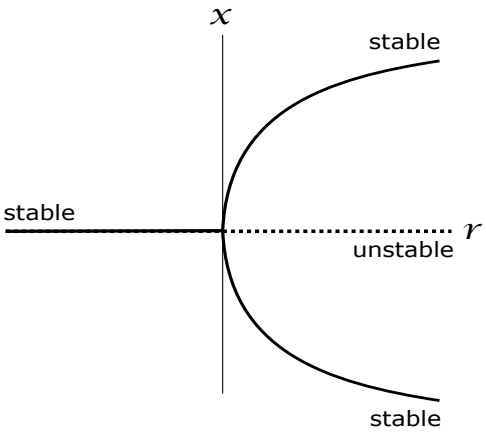


Figure 19.0.4: Bifurcation diagram for pitchfork bifurcation

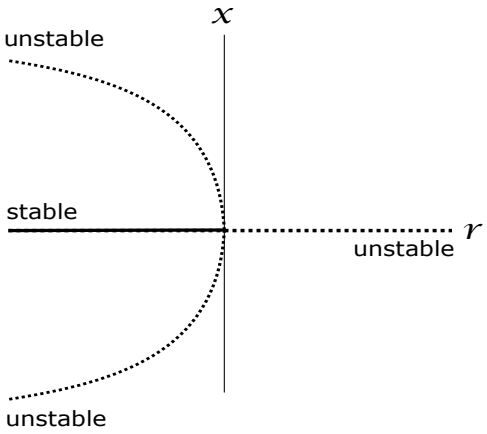


Figure 19.0.5: Bifurcation diagram for subcritical pitchfork bifurcation

Subcritical pitchfork bifurcation

The normal form of subcritical pitchfork bifurcation is given by

$$\dot{x} = rx + x^3.$$

Unit 20

Course Structure

- Hopf Bifurcation

20.1 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 20.1.1. Hopf Bifurcatoin Theorem: Consider the planar system

$$\begin{aligned} \dot{x} &= f_\mu(x, y), \\ \dot{y} &= g_\mu(x, y), \end{aligned} \tag{20.1.1}$$

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:

(a) $\alpha(0) = 0,\ \beta(0) = \omega \neq 0$, where $\text{sgn}(\omega) = \text{sgn}\left[\left(\frac{\partial g_\mu}{\partial x}\right)\Big|_{\mu=0}(0, 0)\right]$ (non-hyperbolicity condition: conjugate pair of imaginary eigenvalues)

(b) $\left.\frac{d\alpha(\mu)}{d\mu}\right|_{\mu=0} = d \neq 0$ (transversality condition: the eigenvalues cross the imaginary axis with non-zero speed)

(c) $a \neq 0$, where

$$a = \frac{1}{16}\left(f_{xxx} + f_{xyy} + g_{xxy} + g_{yyy}\right) + \frac{1}{16\omega}\left[f_{xy}(f_{xx} + f_{yy}) - g_{xy}(g_{xx} + g_{yy}) - f_{xx}g_{xx} + f_{yy}g_{yy}\right]$$

with $\left[\left(\frac{\partial^2 f_\mu}{\partial_x \partial_y}\right)\Big|_{\mu=0}(0,0)\right]$, 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

$$\begin{aligned} x_1' &= -x_2 + x_1(\mu - x_1^2 - x_2^2) \\ x_2' &= x_1 + x_2(\mu - x_1^2 - x_2^2). \end{aligned} \tag{20.1.2}$$

Using polar coordinates,

$$x_1 = r\cos\theta \quad \text{and} \quad x_2 = r\sin\theta.$$

We can rewrite the system (20.1.2) as

$$\begin{aligned} r' &= r(\mu - r^2) \\ \theta' &= 1. \end{aligned} \tag{20.1.3}$$

Note that the equation for r in (20.1.3) is the normal form for a pitchfork bifurcation. Thus as μ passes through the bifurcation value 0, the system (20.1.3) 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.

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.

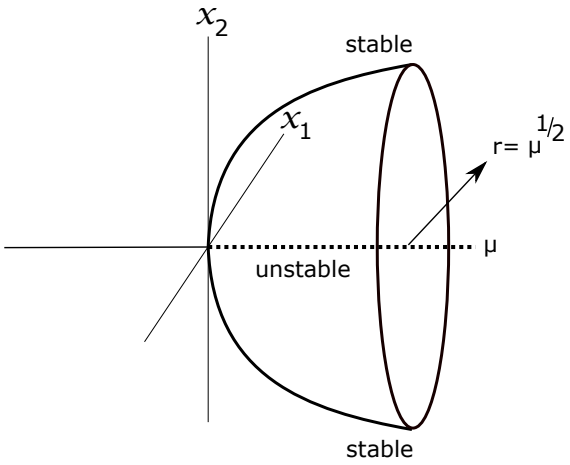


Figure 20.1.1: Hopf bifurcation diagram

Example 20.1.2. Perform a bifurcation analysis for the Liénard 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, $a = -\dfrac{1}{8} \neq 0$. Hence, all the conditions of the Hopf Bifurcation Theorem are satisfied. Since $ad = -\dfrac{1}{16} < 0$, the origin is stable for $\mu < 0$ (see Fig. 20.1.2) and unstable for $\mu > 0$, where there is a stable periodic orbit (see Fig. 20.1.3). The system has a supercritical Hopf bifurcaton at $\mu = 0$.

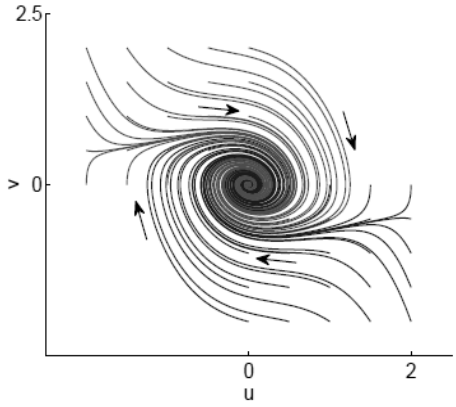


Figure 20.1.2: The origin is stable focus $\mu = -0.3$

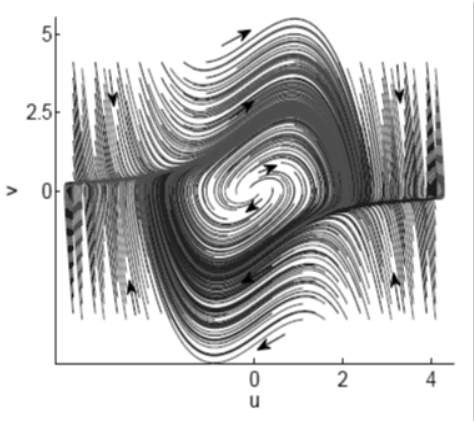


Figure 20.1.3: The origin is unstable focus $\mu = 1$

Example 20.1.3. 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 $\mu = \dfrac{4h}{r}$, 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.

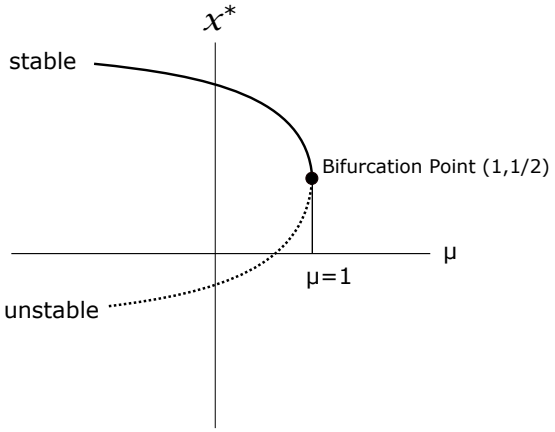


Figure 20.1.4: Bifurcation diagram

Exercise 20.1.4. 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

(a) S. Sokolnikoff: Mathematical Theory of Elasticity.

(b) A. E. H. Love: A Treatise on the Mathematical Theory of Elasticity.

(c) Y. C. Fung: Foundations of Solid Mechanics.

(d) R.N. Chatterjee: Mathematical Theory of Continuum Mechanics.

(e) D. W. Jordan and P. Smith (1998): Nonlinear Ordinary Equations- An Introduction to Dynamical Systems (Third Edition), Oxford Univ. Press.

(f) L. Perko (1991): Differential Equations and Dynamical Systems, Springer Verlag.

(g) F. Verhulust (1996): Nonlinear Differential Equations and Dynamical Systems, Springer Verlag.

(h) H. I. Freedman - Deterministic Mathematical Models in Population Ecology.

(i) Mark Kot (2001): Elements of Mathematical Ecology, Cambridge Univ. Press.

(j) W. G. Kelley and A. C. Peterson, Difference Equations- An Introduction with Applications, Academic Press.

(k) S. Elaydi. An Introduction of Difference Equation, Springer.

POST GRADUATE DEGREE PROGRAMME (CBCS)

M.SC. IN MATHEMATICS

SEMESTER I

SELF LEARNING MATERIAL

PAPER: AECC 1.5
(Pure and Applied Streams)

Computer Programming in C (Theory)

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

Course Preparation Team	
Dr. Biswajit Mallick, Assistant Professor (Cont), DODL, University of Kalyani	Ms. Audrija Choudhury, Assistant Professor (Cont), DODL, University of Kalyani

Dec 2021

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

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

Director
Directorate of Open and Distance Learning
University of Kalyani

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, 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

SYLLABUS

AECC 1.5

Marks: 50 (SEE: 40; IA: 10); Credits: 2

Counselling Duration: 6 Hours

Fundamentals of 'C' Language : Basic structure of a 'C' program, Basic Data type, Constants and Variables, Identifier, Keywords, Constants, Basic data type, Variables, Declaration and Initialization, Statements and Symbolic constants. Compilation and Execution of a 'C' program.

Operators and Expressions : Arithmetic, Relational, Logical operators. Increment, Decrement, Control, Assignment, Bitwise, and Special operators. Precedence rules of operators, Type Conversion (casting), Modes of arithmetic expressions, Conditional expressions.

Input / Output Operations : Formatted I/O - Single character I/O (getchar(), putchar()), Data I/O (scanf(), printf()), String I/O (gets(), puts()). Programming problems. Decision Making Statements : Branching – if Statement, if-else Statement, Nested if-else Statement. else-if and switch Statements. Loop Control: for Statement, while Statement, do while statement, break, continue and exit Statements. Programming problems.

Functions : Function declaration, Library functions, User defined function, Passing argument to a function, Recursion. Programming problems. Arrays: Array declaration and static memory allocation. One dimensional, two dimensional and multidimensional arrays. Passing arrays to functions. Sparse matrix.

Pointers : Basic concepts of pointer, Functions and Pointers. Pointers and Arrays, Memory allocation, Passing arrays to functions, Pointer type casting. Programming problems. Structures and Unions : Declaring a Structure, Accessing a structure element, Storing methods of structure elements, Array of structures, Nested structure, Self –referential structure, Dynamic memory allocation, Passing arrays to function. Union and rules of Union. Programming problems.

File Operations: File Input / Output operations – Opening and Closing a file, Reading and Writing a file. Character counting, Tab space counting, File-Copy program, Text and Binary files.

Computer Programming in C

(Theory)

Unit 1

Structure

History of C

Importance of C

Sample Programs

Basic Structure Of C Programs

Programming style

Executing A C Program

Variables, Identifier, Keywords, Constants

Basic data type, Variables, Declaration and Initialization

1.1 History of C

C is a structured general purpose machine Independent high level programming language developed by Dennis Ritchie at AT & T's Bell Labs of USA in the mid 1970s for the Unix based operating system. Many of the important concepts of C are borrowed from the language BCPL (Basic Combined Programming Language), developed by Martin Richards in 1967. Although originally designed as a systems programming language, C has proved to be a powerful and flexible language that is used for a variety of applications for nearly every available platform. The merit of C lay in the fact that it is easier to read, more flexible and more efficient at using memory. It is particularly popular for personal computer programmers because it requires less memory than other languages. C is the archetype or original model for many modern languages as when we find Language constructs in C, such as "if" statements, "for" and "while" loops, and types of variables, can be found in many later languages. Today, there are very few platforms that do not have a C compiler

In the late, seventies C began to replace the more familiar languages of that time like, ALGOL, PL/I, etc. The drawback of the B language was that it did not know data-types. Both BCPL and B are " type less" system programming languages. By Contrast, C Provides a variety of data types with powerful features. The fundamental data types are integers, characters and floating point numbers of various sizes. In addition there is a hierarchy of derived data types created with arrays, pointers, structures and union.

Since C was developed along with the UNIX operating system, it is has close association with UNIX. Major parts of the popular operating systems like windows, Linux and Unix are coded in C. This is because when it comes to performance nothing beats C. Although C is technically a high-level language, it is one of the "lowest-level" high-level programming languages in the sense; it is much closer to assembly language than are most other high-level languages. This closeness to the underlying machine language allows C programmers to write very efficient code. More over if one is

to extend the operating system to work with new devices one needs to write device driver programs. These programmes are exclusively written in C.

For many years, C was the reference manual, but eventually with the appearance of many C compilers coupled with the wide popularity of UNIX operating system, it gained wide popularity among computer professionals. Today, C is the language of choice while building a variety of hardware and operating system platforms.

The American National Standards Institute (ANSI) constituted a committee in 1983, to provide an updated definition of C. The resulting definition "**ANSI C** "was completed in late 1988, and modern compilers are already supporting most of the features of this standard .The standard is based on the original reference Manual in the first edition, the classic book **"The C Programming Language"** , with little or no changes on the original design of the C language . They ensured that old programs still worked with the new standard, failing that, the compiler would produce warnings of new behavior.

One of the significant contributions of the standard is the definition of a new syntax for the defining and declaration of the function. This extra information makes it easier for compilers to detect errors caused by mismatched arguments. A second significant contribution of the standard is the definition of a library to accompany C. These library functions specifies functions for accessing the operating system, formatted input and output, memory allocation, string manipulation, and the like. A collection of standard headers provides uniform.

1.2 Importance of C

C is an immensely popular language widely used and well understood. Some of the versatile features of C language are**:** reliability, portability, flexibility, interactivity, modularity and finally efficiency and effectiveness. It is a great tool for expressing programming ideas in a way it is easily understood, regardless of the language users are most familiar with. It is in fact the original or archetypal building block for many other currently known languages and it is very close to assembly language. C is a robust language whose rich set of built in functions, and operators can be used to write any complex programs. In C large programs are divided into small programs called functions and data moves freely around the systems from one function to another. Moreover, the C compiler combines the capabilities of an assembly language with the attributes of a high level language and therefore it is useful for writing both system software and business packages without worrying about the hardware platforms where they will be implemented..The great thing about C is that it can be used to write high performance code for both application and system software. Further it can interact with hardware at quite low level. In fact, many of the compilers available in market are written in C. It is the language used for developing system applications that forms major portion of operating systems such as Windows, UNIX and Linux. C is increasingly being used in Database systems, Graphics, Spread sheets, word processors, Compilers /Assemblers, Network drivers and interpreters.

The variety of data types and powerful operators available in C makes C programs very efficient and fast. In C there are only 32 key words and its strength lies in its built-in functions. Some standard functions are available which can be used for developing programs. C Being highly portable, programs written for one computer can be made to run on another system with little or no modification.

C is at once one of the pillars of modern information technology (IT) and computer science (CS). C is a high level language that lets us to write very low level stuff like device drivers that runs as fast as assembly written programs. C's power and fast program execution come from its ability to access low level commands, similar to assembly language, but with high level syntax. It allows low level access to information and commands while still retaining the portability and syntax of a high level language. In this process C imposes few constraints on the programmer. Further it is tailor- made for structured programming, thus requiring the user to think a problem in terms of function modules or blocks. A collection of these modules make a program debugging and testing easier..Thus, C meets the requirements, where speed, space and portability are important.

Another prime feature of C is its ability to extend itself. A program in C is basically a collection of functions that are supported by the C library. We can add our own functions to the C library .With the availability of large number of functions , the programming burden becomes simple. C being simple and easy to understand, most of the operating systems and game software are written in C .

Before discussing some distinct features of C, we shall look at some sample programs in C, and as we proceed, can learn more about the language.

1.3 Sample Programs

Printing A Message: Sample program 1

The only way to learn a new programming language is by writing programs in it. Let us begin by looking at the construction of a very simple program.

The following is the output of the above program code when it is executed:

```
hello, fine
```

```
main( )
  {
   /* ……Printing begins…….*/
        Printf(“ hello, fine ”);
    /* ……Printing ends…….*/
  }
```

The first C program to print a single line of text

In the above C program, the code begins executing at the beginning of **main**. **main( )** is a special function used by the C systems to tell the computer where the program begins. This means that every program must have a **main** somewhere. In this example, **main** is defined to be a function that expects no arguments, which is indicated by the empty list ( ). All the statements that belong to **main( )** are enclosed within a pair of braces { } as indicated above. The opening brace **"{"** indicates the beginning of the function **main** and the closing brace **"}"** marks the end of the program. All the statements between these two braces form the function body. The function body contains a set of instructions to perform the given task.

In our example, the function body contains three statements out of which only the **printf** line is an executable statement. A function is called by naming it, followed by parenthesized list of arguments, so this calls the function **printf** with the argument " hello, fine ". **printf** is a library function that prints output , in this case the string of characters (String constant or character string) between quotes.

The two lines

/* ……Printing begins…….*/

And

/* ……Printing ends…….*/

Are **comment lines** which in this program tells what the program does. Any characters between /* and .*/ are ignored by the compiler (comments are solely given for the understanding of the programmer or the fellow programmers); they may be used freely to make a program easier to understand . Any number of comments can be written at any place in the program. The normal language rules do not apply to text written with in /* and .*/ . Thus we can type this text in small case, capital, or a combination. Moreover, comment can be split over more than one line, as in,

```
/*  printing
    begins.*/
```

Such a comment is often called a multi-line comment. Comments cannot be nested. For example,

```
/* Printing begins /*Printing ends.*/*/
```

Is invalid and therefore results in an error.

Let us come back to the **printf** function, the only executable statement of the program .

```
printf(" hello, fine ");
```

The above quotation can be printed in two lines, by adding another **printf** function, as in,

```
printf("hello,\n");
printf("fine");
```

The information contained between the parentheses is called the **augment** (which are simply strings of character to be printed out) of the function. The argument of the first **printf** contains a combination of two characters \ and **n** at the end of the string. The combination sequence" \n " is called **newline** and it takes the character to the next line. Therefore, you will get the output split over two lines. **\n** is one of the several Escape Sequence (similar in concept to the carriage return key on a type writer, which when printed advances the output to the left margin on the next line) available in C. if you try something like

```
printf("hello, fine
");
```

The C compiler will produce an error message.

No space is allowed between \ and n. **printf** never supplies a new line automatically, so several function calls may be used to build up an output line in stages, as in,

.

```
main( )
  {
  /* ……printing begins…….*/
      printf(" hello,");
     printf(" fine,");
     printf(" \n");
   /* ……printing ends…….*/
  }
```

To produce identical output. Here \n represents only a single character. An escape sequence like \n provides a general and extensible mechanism for representing hard to type or invisible characters. It is also possible to produce multi line output by one printf statement with the use of newline character at appropriate places, as in,

```
printf ("hello\n….fine,\n……I\n……..am ok!");
```

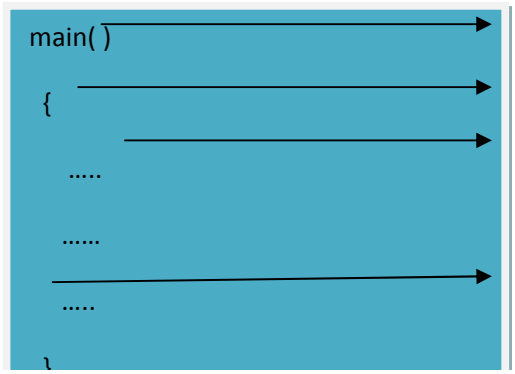
Where the output is

```
hello
…..fine,
…….I
…….am ok !
```

The inclusion of the preprocessor directive **# include < stdio.h >** at the beginning of all programs that use any input/output library functions should not be insisted for functions like, **printf** and **scanf**, **Printf** is a pre defined standard C function (predefined in the sense that it is function that has already been written, compiled, and linked together with the program at the time of linking).

Note that the print line ends with a **semi colon**. Thus the mark **;** acts as a statement terminator. That is, every C statement must end with a **;** mark. In C , everything is written in lowercase letters. However, uppercase letters are used for symbolic names representing constants. we may also use uppercase letters in output strings like "HELLO" and "FINE".

The General format of simple C programs is shown below.



Simple C program Format

The main Function

The main ( ) is a function and is part of every program. There are different forms of main statement in C. viz.,

main ( )

int main ( )

main (void)

void main (void)

int main (void)

The empty pair of parenthesis indicates that the function has no arguments This may be explicitly indicated by using the keyword **void** inside the parenthesis. Just like the way functions in a calculator returns a value, functions in C also return a value to the operating system. That is, It is also possible to specify the keyword **int or void** before the word **main**. Some compilers permit us to return nothing or no information to the operating system from **main ( ).** In such a case we should precede it with the key word **void.** The key word **void** means that the function does not return any value to the operating system and **int** means that the function s returns an integer value to operating system. When **int** is specified, the last statement in the program must be **"return 0".**

Addition of Two numbers: Sample program 2

Consider another program, which performs addition on two numbers. This program explains the need for the use of declaration of variables, and use of operators.

/Program to add two numbers:/

```
/* addition of  two numbers */
main ( )
 {
      int  num;
      float amount;
      num = 10;
      amount = 20.25+29.85;
      printf ( " % d\n",num);
      printf ("%5.2f",amount);
 }
```

On execution of this program we will get the following output:

10

50.10

The first line of the program is a **comment line**. Comment line in the beginning give information such as name of the program, author, date etc. To indicate line numbers comment characters can also be used. in other lines. The words **num** and **amount** are variable names used to store numeric data. The numeric data may be either in **real or integer** form. In C, all variables must be declared before they are used, usually at the beginning of the function before any executable statement. The type declaration statement is written at the beginning of main ( ) function. In lines 4 and 5, the declarations

int num;

float amount;

tells the compiler that num is an integer (**int)** and amount is a floating (**float**) point (numbers with fractional part) numbers. All declaration statements ends with a **semicolon.** The words such as **int** and **float** are called keywords and cannot be used as variable names .The range of both **int** and **float** depends on the machine you are using; 16- bit i**nts**, which lie between -32768 and +32768 , are common, as are 32-bit **ints**. A float number is typically 32-bit quantity, with at least six significant digits and magnitude generally between about 10^{-38} and 10^{+38}. While declaring the type of variable one can also initialize it as shown in line 7 and 9.That is , the statements

num = 10;

amount = 20.25+29.85;

are called the assignment statement. **Every assignment statement must have a semicolon at the end.**

The order in which we define the variables is sometimes important sometimes and sometimes not. For example,

int i =10, j =25;

is same as

int j= 25, i=10;

However,

float a= 1.5, b = a + 3.2;

Is alright. But

float b= a+3.2, a = 1.5 ;

Is not, because we are trying to use **a** even before defining it.

Moreover, the following statements would work

int a,b,c,d

a = b = c = d = 10;

However the following statement would not work

Int a= b= c= d =10;

The next statement of the program is an output statement that prints the value of **number.** The print statement

printf (" % d\n", num);

contains two arguments..The first argument **"%d'** tells the compiler that the value of the second argument **num** should be printed as a *decimal integer*. These arguments are separated by **comma. T**he newline character "\n " causes the next output to appear on a new line.

The last statement of the program

printf ("%5.2f", amount);

print out the value of **amoun**t in floating point format. The format specification **"%5.2f "** tells the compiler that the output must be floating type , with five places in all and two places to the right of the decimal point.

Calculation of Interest: Sample Program 3

C supports the basic four arithmetic operators (-, +, * . /) along with various others. The use of such operators along with other variable declarations, the while loop construct and # define preprocessor directive are illustrated in the program below. The program calculates the value of money at the end of each year of investment, assuming the interest rate at 11 percent with an initial investment of 50 000 for 10 years .In this program, the variable **value** represents the value of money at the end of the year and the **amount** represents the value of the money at the start of the year. The statement

amount = value ;

makes the value at the end of the current year as the value at the beginning of the *next* year .

The preprocessor compiler directive **#define,** defines a symbolic constant. Whenever a symbolic name is encountered, the compiler automatically substitutes the value associated with the name. If you want to change the value you have to simply change the definition. **#define** line should not end with a semicolon and are usually written in upper case letters(so that they can be readily distinguished from the lower case variable names), usually placed at the beginning before the **main ( )** function. They are not declared in the declaration section. The declaration section of the program declares **year** as integer and **amount ,value and rate** as floating point numbers. When two or more variables are declared in one statement, they are separated by commas. It is also possible to declare the floating point variables as multiple statements as in,

float amount;

float value;

float rate;

```
/*  ..... INVESTMENT PROBLEM …….. */
# define PERIOD       10
#define   PRINCIPAL  50000.00
/* ..... MAIN PROGRAM BEGINS …….. */
main ( )
 {  /*  .....DECLARATION STATEMENTS …….. */
        int year;
        float amount, value, rate;
 /*  ..... ASSIGNMENT STATEMENTS …….. */
       amount = PRINCIPAL ;
       rate = 0.11;
       year = 0;
/*  ..... ..... COMPUTATION STATEMENTS… …….. */

 /*  .....  COMPUTATION  USING while LOOP …….. */
       While (year < = PERIOD )
          {
            printf ( " % 2d     % 8.2 f \n" , year, amount );
            value = amount + rate * amount;
            year = year +1;
             amount = value;
          }
  /*  .....  .....  while LOOP ENDS… …….. */
 }
 /*  ..... ..... PROGRAM  ENDS              … …….. */
```

The Investment Program

In the **while** loop all computation and printing are accomplished. The body of a **while** loop can be one or more statements enclosed in braces . The parenthesis after the **while** contain a condition that is tested. So long as this condition remains true all , all statements within body of the while loop keep getting executed repeatedly. When the condition becomes false , the control passes to the first statement that follows the body of the **while** loop..In this case as long as the value of the **year** is less than or equal to the **PERIOD**, the four statements grouped by braces that follows the **while** are executed. The loop ends when year becomes greater than **PERIOD.**

Sample Program 4: Use of Sub routines:

A very simple program that explains the use of **mul ( )** function is shown below. It uses a user defined

```
/*  ..... PROGRAM USING FUNCTION …….. */
int mul  ( int a, int b);          /*  DECLARATION….. */
/*  ..... MAIN PROGRAM STARTS…….. */
      main ( )
        {
            int a, b,c;
            a =7;
            b =10;
            c = mul (a,b);
            printf ( “multiplication of %d and % d is  % d”, a,b,c);
         }
/*  ..... MAIN PROGRAM ENDS
                    MUL FUNCTION STARTS…….. */
     int mul (int x, int y)
     int p;
     {
          p = x * y;
          return ( p);
      }
/*                  MUL  (  )  FUNCTION ENDS                    . */
```

function equivalent to subroutine in **FORTRAN** or Sub program in **BASIC**. The Execution of the program will print the output

Multiplication of 7 and 10 is 70

The **mul ( )** function multiplies the value of variables x and y and the result is returned to the **main ( )** function when it is called in the statement

c = **mul (a,b);**

The **mul ( )** function has two arguments x and y (declared as integers) and when called the values of a and b are passed onto x and y respectively. This example also shows a bit more of how **printf** works.

Sample Program 5: Use of Math Functions:

There are many occasions where we often use standard mathematical functions like cos, sin, exp, etc.

```
/*  ... PROGRAM USING COSINE FUNCTION …….. */
# include < math.h >
# define PI  3.1416
# define MAX 180
main ( )
  {
      int angle;
     float x,y;
     angle = 0;
     Printf ( “Angle      Cos(angle) \n\n “);
     While (angle < = MAX)
      {
            x = ( PI/MAX) * angle;
            y = cos (x);
            printf ( “% 15 d % 13.4 f\ n “, angle, y);
            angle = angle +10;
       }
  }
```

The standard mathematical functions are defined and kept as a part of **C math library** for use in programs. The use of any of these mathematical functions in the program can be accomplished by means of **# include** instruction in the program. The **#include** directive tells the preprocessor to treat the contents of a specified file as if those contents had appeared in the source program at the point where the directive appears Like **# define**, it is also a compiler directive and tells the compiler to link the specified mathematical functions from the library. The instruction is of the form

 # include < math.h >

math.h is the file name containing the required information. Program code,(Figure 3.1) explains the use of cosine function. Another # include instruction that is often used is

include <stdio.h>

<stdio.h> refers to the standard I/O header file containing standard Input output functions. That is, it adds the contents of the file named **stdio.h** to the source program and the ankle brackets cause the preprocessor to search the directories specified by the Include environment variable for stdio.h, after searching directories specified by the / I compiler option. For example, to use the function printf( ) in a program, the line

#include <stdio.h>

Should be at the beginning of the source file, because the definition for printf() is found in the file stdio.h.

As explained earlier, C programs are divided into modules or functions. To use any of the standard functions, the appropriate header file should be included...Header files contain definitions of functions and variables which can be incorporated into any C program by using the pre-processor *#include* statement. This is done at the beginning of the C source file . To access the functions stored in the C library, it is necessary to tell the compiler about the files to be accessed. This is achieved by the use of pre processor directive

#include <filename>

Placed at the beginning of the program. Note here that **filename** is the name of the library file that contains the required function definition.

1.4 Basic Structure Of C Programs

The programs in C so far discussed illustrates that it can be viewed as a group of building blocks called functions. A function is a segment that groups a number of program statements to perform specific task. To write a c program , we must first create functions and then put them together.

The different sections of a C program as shown in figure 1.1 ..The documentation section consists of a set of comment lines giving the name of a program, author, date and other details, which the programmer would like to use later .The link section provides instructions to the compiler to link functions from the system library. All symbolic constants are defined in the definition section. Global

variables (variables that are used in more than one function) and all the user defined functions are declared in the global declaration section that is out side of all the functions.

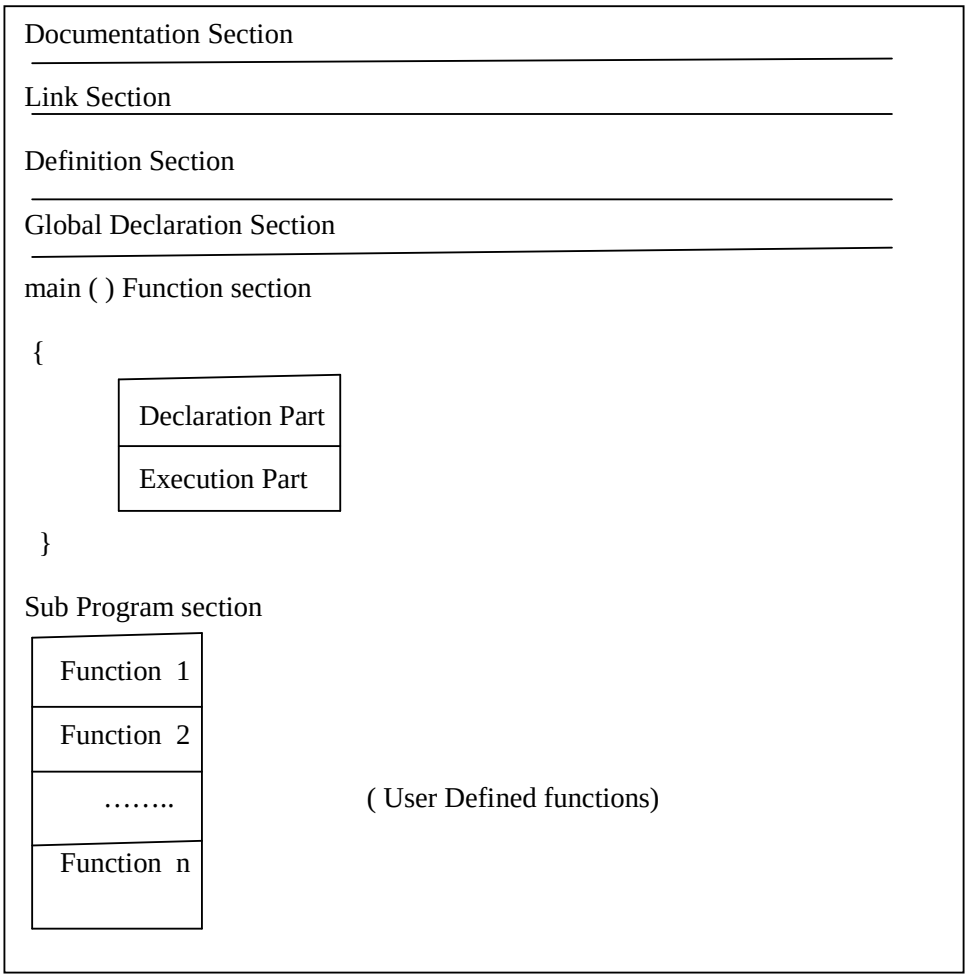


Fig.1.1 An over view of C program

Every C program must have one main ( ) function section that contains two parts, the declaration and executable part, appearing between the opening and closing braces. In the declaration part all those variables used in the executable part are declared..There is at least one statement in the executable part. The program execution begins at the opening brace and ends at the closing brace which marks the logical end of the program. Every statements in the declaration and executable parts end with a semi colon **(;)**.\

The sub program section contains all the user defined functions that are called in the **main** function. User defined functions are generally placed immediately after the **main** function, although they may appear in any order. All sections , except the main function may be absent when they are not required.

1.5 Programming Style

Programming style is a set of rules or guidelines used when writing the source code for a computer program. It is often claimed that following a particular programming style will help programmers to read and understand source code conforming to the style, and help to avoid introducing errors.

C has no specific rules for the position at which a statement is to be written. That's why it is often called a free –form language. First of all, all statements are entered in small case letters. Upper case letters are used only for symbolic constants. The statements in the program must appear in the same order in which we wish to be executed.; unless of course the logic of the problem demands a deliberate "jump", which is out of sequence. These statements are terminated with a semi-colon **(;)**, and are collected in sections known as functions. By convention, a statement should be kept on its own line. Blank spaces may be inserted between two words to improve the readability of the statement. However , no blank spaces are allowed with in a variable, constant or key word.

Since C is a free-form language, we can group statements together on one line. The statements

```
a = b;
x = y-1;
z = a-1;
```

can be written on one line as

```
a = b;  x = y-1; z = a-1;
```

The program

```
main ( )
{
    Print f ("hello");
}
```

May be written in one line like

```
main ( )  { Print f ("hello")};
```

However, this style makes the program more difficult to understand. Rather than putting everything on one line, it is much more readable to break up long lines so that each statement and declaration goes on its own line.

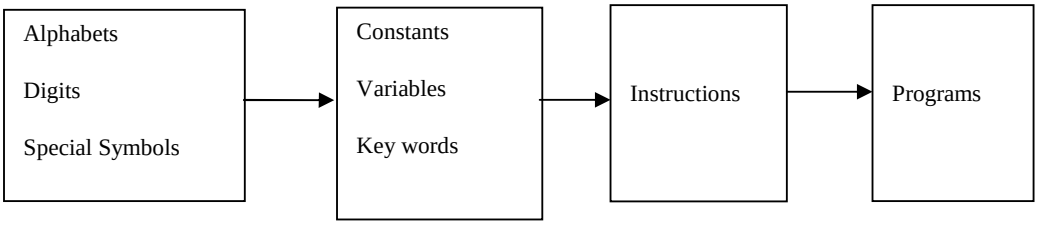
Comments in code can be useful and they provide the easiest way to set off specific parts of code (and their purpose); as well as providing a visual "split" between various parts of your code. Having good comments throughout your code will make it much easier to remember what specific parts of your code do. Care should be taken not to over emphasize generous use of comments inside the code. For debugging as well as testing of the code Judiciously inserted comments is very helpful and it improves the code readability as well as the understandability of the code logic.

1.6 Executing A C Program

C program Execution involves the following steps

1. **Creating the program**
2. **Compiling the program**
3. **Linking the program with functions that are needed from the C library**
4. **Executing the program.**

1.7 Alphabet in C language



As in any language, C language contains letters, alphabet and grammar*(or syntax rules)* and each program instruction must coıform precisely to the syntax of the C language. In this section we will discuss the concepts of constants, variables and their types.

1.7.1 The C Character Set

A C character set denotes any valid alphabet, digit or special symbol, to represent an information. The set of characters that can be used to write a source program is called source character set and the set of characters available during program execution is called execution character set. Very often, in most implementations of C, both character sets are taken as identical. Generally, a character data type holds a single character(or one byte), enclosed with in single quotes, to represent a character constant. For e.g., the expressions 'a' , 'b',and '0' represent character constants. Remember that "a" is used to represent a string of characters(or sequence of characters enclosed with in double quotes) and is different from 'a'. Further, '\n' is used to represent a new line character, that is used to move the cursor to a new line on the screen. Figure 1.2 shows the entire character set (i.e., the valid alphabets, numbers, special characters and white spaces) allowed in C. The compiler ignores white spaces unless they are part of a string constant. White spaces may be used to separate words, and are prohibited between characters of key words and identifiers.

Trigraph Characters

Some characters from the C character set are not available in all environments, because keyboard may not have keys to cover the entire characters set of the language. A Trigraph, is a three character replacement for a special character in the C character set. ANSI C introduces the concept of "**Trigraph**" Sequences to provide a way to enter certain characters that are not available on some keyboards. Actually, each T**rigraph** sequence contains three characters (i.e., two question marks followed by another character) as in Figure 1.3 i.e., Each trigraph sequence is introduced by two question marks followed by a third character that indicates the character to be represented. For eg., , if a key board does not support square brackets , we can still use them in a program using the **Trigraphs ?**? (and ??).

Alphabets	Upper case letters A,B,……., Z Lower case letters a,b,…… .., z
Digits	All decimal digits 0,1,2,…….9
Special Characters	; semicolon , comma & ampersand . period * asterisk + plus sign ‘ apostrophe ? question mark < opening bracket or less than sign > closing bracket or greater than sign ^ caret ~ tilde ! exclamation mark | vertical bar (left parenthesis) right parenthesis \ backlash [left bracket] right bracket $ dollar sign } right brace _ under score { left brace = equal sign % percent sign # number sign / slash @ commercial at - hyphen or minus sign “ quotation mark
	White Spaces Blank spaces Horizontal Tab Carriage Return New Line

Figure 1.2 : The C Character Set

1.7.2 C Tokens

A **token** is a source program text that the compiler does not break down into atomic units. They are the basic building blocks/elements of the C language, constructed together to make a C program. That is, each and every smallest individual units in a C program are called **Tokens.** The **Tokens** in C language include:

1. Key words (eg: float, double etc.,)
2. Constants (eg: 100, -10.0 etc.,)
3. Strings (eg: “ABC”, “month” etc.,)
4. Operators (eg: +, - etc.,)
5. Identifiers (eg: main, total etc.,)
6. Special Symbols (eg: [ ],( ) etc.,)

C Programs are written using these **tokens** and the syntax of the language.

Trigraph Sequence	Translation
??=	# number sign
??(	[left bracket
??)	] right bracket
??<	{ left brace
?? >	} right brace
??!	| vertical bar
??/	\ back slash
??|	^ caret
??~	~ tilde

Fig. 1.3 ANSI C Trigraph Sequences

1.7.4 Key Words and Identifiers

Every C word fall under two categories, viz,. either a **key word** or an **Identifier**. C **Key words** (also called Reserved words) are the words that convey a special meaning to the C Compiler. They are the system defined **identifiers** that do have a fixed meaning (i.e., it does not change) and cannot be used as variable names. They are the basic building blocks for program statements and are written in lowercase letters. C language supports **32 (Thirty Two) keywords** and are listed in Figure 1.4 below.

auto	float	double	long
short	signed	unsigned	const
goto	else	switch	break
if	do	while	for
typedef	extern	static	struct
default	enum	return	sizeof
register	union	int	case
void	char	continue	volatile

Figure 1.4: Keywords in C

An Identifier reters to the names of variables (i.e., the one which changes during program execution), names of functions, arrays, and structures. They are user defined names consisting of a combinations of alphabets, digits with a letter as the first character and underscore. The under score symbol is treated as a letter in the C character set and it helps in the readability of long variable names. That is, they are the names given to C entities such as , variables, types, functions, structures and labels in the program. However, the lengths of identifiers in C, vary from one implementation to another. In general, Identifier are created to give a unique name to C entities so as to identify it during the execution of the program. For example: **int** apple; Here apple is an identifier which denote a variable of ***integer*** type. In fact, **Keywords** (either C or Microsoft) are not used as **identifiers.**(i.e., they are reserved for special use). **Identifiers** are in general, used to name constants, functions, files and the like, apart from variables.

Rules for Identifiers

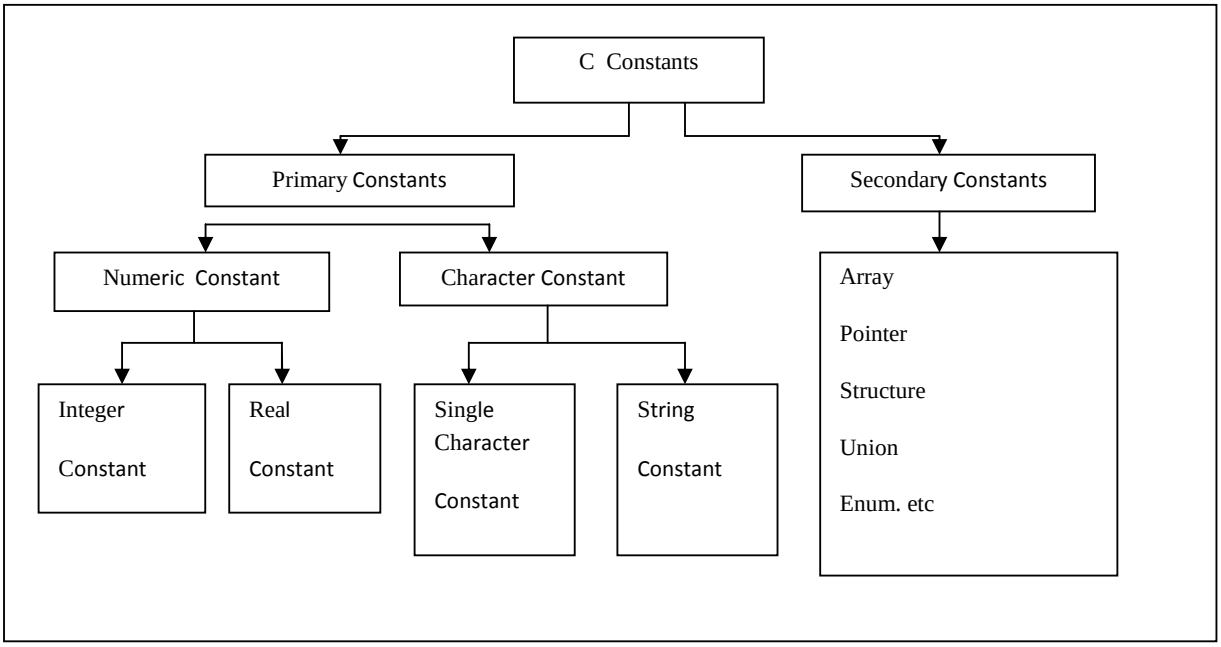
1. The first character must be an alphabet(uppercase or lowercase) or an under score.
2. All succeeding characters must be letters or digits.
3. Key words should not be used as identifiers.
4. Name of identifier is case sensitive i.e. num and Num are two different variables.
5. Identifier name cannot be exactly same as constant name which have been declared in the header file of C and you have included that header file.
6. Name of identifier cannot be exactly same as of name of function with in the scope of the function.
7. Name of function cannot be global identifier.
8. No two successive underscores are allowed.
9. Only first 31 characters are significant.
10. No special characters or punctuation symbols are used except the under score.

1.7.5 Constants

A **constant in C** refers to a piece of data that does not change throughout the execution of the program**. That is, C**onstants in C are expressions with a fixed value that are not changed during the execution of the program and are declared with the *define* keyword .In general, C constants can be divided into two major categories

1. Primary constants
2. Secondary constants.

These constants are further categorized as shown below



At this stage, we would restrict our discussion to only primary constants(or basic constants) namely, Integer, Real and character constant.

Integer Constants

Integer constants are the numeric constants (Constants associated with number) without any fractional or exponential part. Integer constants take one of the following forms:

1. A **decimal integer.** , *e.g., 1 , 134, 10005 (* Decimal integers are a set of digits, 0 through 9, preceded by an optional – or + sign). Embedded spaces, commas, and non digit characters are not allowed between digits.
2. An **Octal integer constant** (base 8), e.g., *0 1 , 134, 0303242* . An octal constant is introduced

by a leading 0 and digits, the digits are 0 through 7 .

3. A **Hexa decimal** (base 16) Number. *e.g., 1 , 0x1, 0X186A2.* A hex constant is preceded by a leading 0X or 0x and the digits are 0 through 9 followed by A through F (Note that upper and lower case Letters are allowed) .

4. A character Constant.

Integer constants can also be suffixed with an identifier U (or u) or L (or l), which is used to indicate that the constant is unsigned or long, respectively. For e.g., 567U or 567u These suffixes may be combined *as in .e.g.,* 989712343UL or 989712343ul . The largest integer value that can be stored is machine dependent. It is 32767 on 16-bit and 2147483647 on 32-bit machines. For constructing the integer constants, certain rules have been laid down. These rules are as under:

Rules for constructing Integer constants

1.An integer constant must have at least one digit

2.It must not have a decimal point.

3.It can be either + ve or - ve.(If no sign precedes, it is assumed to be + ve.).

4. No Commas or Blanks are allowed within an integer constant.

5. The allowable range is between -32768 to 32767(For 16 bit compiler).

Real Constant

Certain quantities that vary continuously, such as prices, distances, temps, and so on, are represented by numbers containing fractional parts like 10.246. Such numbers are called **Real or Floating** point constants. That is, a real constant is one of :

- A fractional constant followed by an optional exponent
- A digit sequence followed by an exponent.

In either case followed by an optional of f, l (for single precision) , F. L(For double Precision), where:

- An optional digit sequence followed by a decimal point followed by a digit sequence.
- A digit sequence followed by a decimal point.

Further, an exponent is one of :

- E or e followed by an optional + or – followed by a **digit sequence** (A digit sequence is an arbitrary combination of one or more digits).

Floating point constants are normally represented as double precision quantities. Following rules must be observed while constructing real constants in fractional form:

1. A real constant must have at least one digit
2. It must have a decimal point
3. It could be either positive or negative
4. If no sign precedes an integer constant, it is assumed to be positive.
5. No commas or blanks are allowed within the real constant.

The exponential form of representation of real constants is usually used if the value of the constant is either too small or too large . In this form of representation, the real constant is represented in two parts. The part appearing before 'e' is called mantissa, whereas the part following 'e' is called exponent. Thus 0.000213 is represented in exponential form as 2.13e-4 . The General form is

mantissa e exponent

Following rules must be observed while constructing real constants expressed in exponential form:

1. The mantissa and exponential part should be separated by a letter e or E.

2. The mantissa part may have + ve or –ve sign.(default sign is positive).

3. The exponent must have at least one digit , which must be a +ve or _ve integer. Default sign is +ve.

4 .Range of real constants expressed in exponential form is -3.4e38 to 3.4e38.

Character Constant

Character constants are the constant which use single quotation around characters. example, `b`, `k`, `l` etc. In general, A character constant is a single alphabet, a single digit, or a single special symbol enclosed with in single quotes(or inverted commas). For both the inverted commas(single quotes) should point to the left. For example, 'C' is a valid character constant while ' C' is not. In C, characters are small integers, so you can use a character constant anywhere you can use an integer constant and *vice versa.* More over, the maximum length of a character constant can be 1 character.

String Constants

It is a collection of characters enclosed in double quotes. It may contain letters, digits, special characters and blank space. Examples are:

"Hello!" "How Are You " " ? " "X "

Note that a character constant (e.g., 'X') is not equal to the single character string constant(e.g., "X") . Further, a single character string constant does not have an equivalent integer value while a character constant has an integer value. More over, character strings are often used in programs to build meaningful programs. Moreover, the entity having two consecutive double quotes without any characters in between them, i.e., " ", is called a null string. Here, the quotes acts as delimiters and are not part of the string.

Backlash character constants

Sometimes, it is necessary to use newline(enter), tab, quotation mark etc. in the program which either cannot be typed or has special meaning in C programming. Such characters with special meaning should be preceded by a backlash symbol to make use of special function of them.. The backlash (\ ) causes "escape" from the normal way the characters are interpreted by the compiler. Each backlash character constant represents one character, although they consist of two characters. These character combinations are called escape sequences. Given below (Table 1.1)is the list of special characters and their purpose .

1.8 Variables

Every language should support the basic data objects namely, variables and constants. **Variables** are memory location in computers memory to store data. To indicate the memory location, each variable should be given a unique name called **identifier**. Variable names are just the symbolic representation of a memory location. These memory locations can contain integer, real or character constants. Unlike constants that remain unchanged during the execution of program , a variable may take different values at different times during execution. Examples of variable names are : sum, count, bike, interest etc. A variable name can be chosen by the programmer in a meaningful manner so as to reflect its function. Variables are to be declared before using it in the program.

Rules for writing Variable names in C

1. Variable names can be composed of letters(upper & lower case) , digits, and underscore. There is no rule for the length of a variable. A variable name is any combination of 1 to 31 alphabets.
2. The first letter of a variable should be either a letter or an under score. Note that upper and lower case are significant
3. No commas or blanks are allowed with in a variable name.
4. No special symbol other than underscore can be used in the variable name.
5. It should not be a key word.
6. White spaces are not allowed.

Constant	*Meaning*
‘\a’	audible alarm
‘\b’	back space
‘\f’	form feed
‘\n’	new line
‘\r’	carriage return
‘\t’	horizontal Tab
‘\v’	vertical tab
‘\”’	double quote
‘\’’	single quote
‘\?’	question mark
‘\\’	backlash
‘\0’	null

Table 1.1

1.9 Data Types.

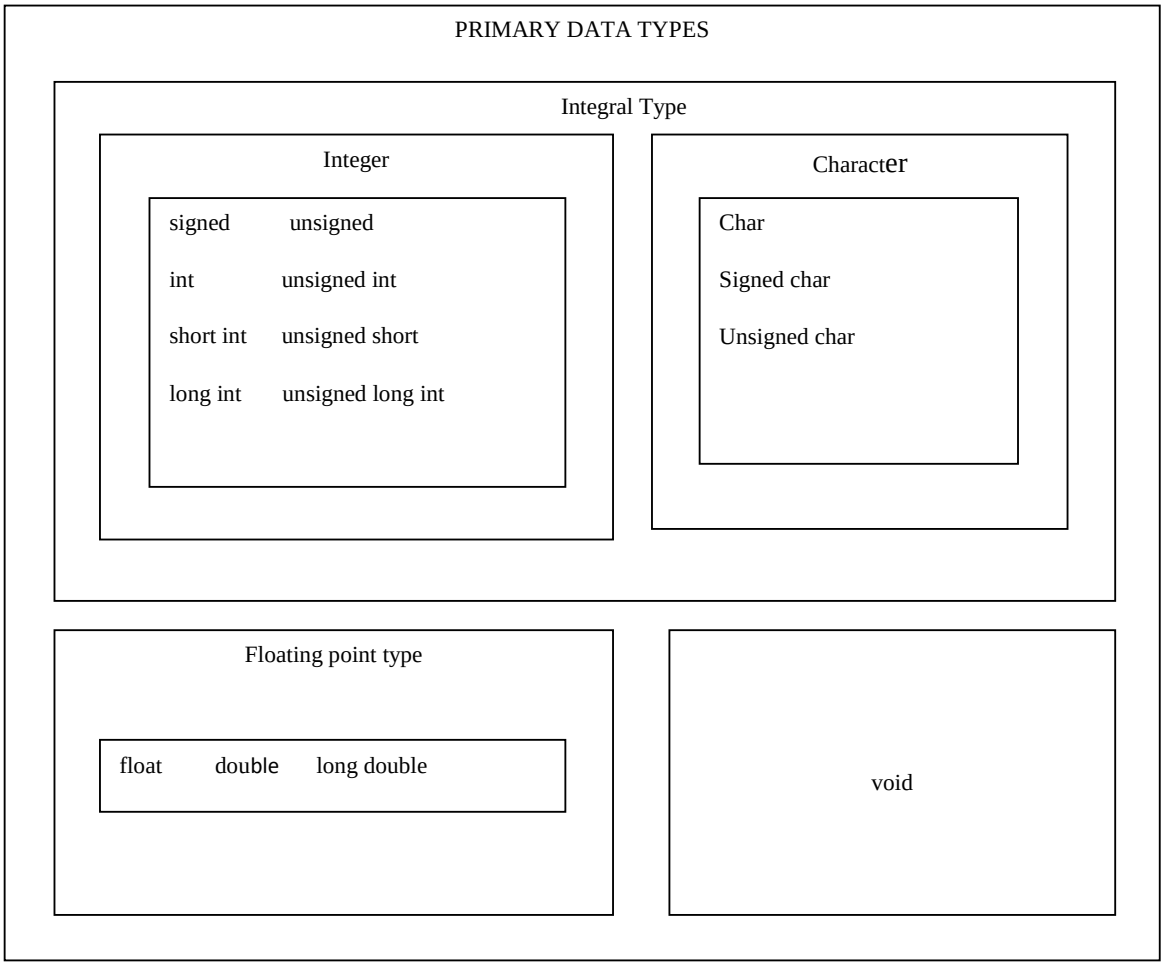
Like other computer languages, **C** supports data types namely, of **integer, character and** of **float** type. In C, all variables must be declared before they are used, usually at the beginning of the function before an executable statements. A declaration announces the properties of variables; it consists of a type name and a list of variables such as

int Celsius;

int count;

The type **int** means that the variables listed are integers. ANSI C supports three classes of data types:

1. Primary data types
2. Derived data Types
3. User defined data Types.



All **C** Compilers support five fundamental data types, namely integer(**int**) , character(**char)**, Floating point(**float**), double precision floating point(**double)** and **void**. Extended data types like **long int ,long double** are also in use in C. The above figure gives an overview of primary data types in C.

Integer Types

This data type allows a variable to store numeric values. **int** keyword is used to refer integer data type. The. integers are whole numbers with a range of values supported by a particular machine (that is, the storage size of **int** data type is 2 or 4 or 8 byte. It varies with the processor in the CPU that we use). Generally, the C integer types were intended to allow code to be portable among machines with different inherent data sizes (word sizes), so each type may have different ranges on different machines. The problem with this is that a program often needs to be written for a particular

range of integers, and sometimes must be written for a particular size of storage, regardless of what machine the program runs on. In fact, integers occupy one word of storage, and since the word size of machines vary, the size of integer that can be stored depends on the computer. For a 16 bit word length, the size of the integer value is limited to the range -2^{15} to $2^{15}-1$. A signed integer uses one bit for sign and 15 bits for the magnitude of the number.

In order to provide control over the range of numbers and storage space, the C language defines several integer data types: **integer, short integer, long integer, and character, all both in signed and unsigned varieties.** For eg., **Short int** represents fairly small integer values and requires half the amount of storage space as a regular **int** number uses. Unlike **signed integers**, unsigned integers use all the bits for the magnitude of the number and are always positive. To increase the range of values we declare long and unsigned integers

Floating point types

C uses the key word **float** to define floating point numbers . Floating point numbers are stored in 32-bit, with six digits precision. Key word **double i**s used to define **big** floating point numbers. It reserves twice the storage for the number. A **doubl**e data type number uses 64 bits giving a precision of 14 digits. On PC's this is likely to be 8 bytes. The **double** type represents the same data type that **float** represents, but with a greater precision. To extend the precision further, the key word **long double** with 80 bits are used.

Void types

Void is an empty data type normally used as a return type in C to declare that no value will be returned by the function. It can also play the role of generic type, meaning that it can represent any of the other standard types.

Character type

A single character of the character set of C, can be defined as a **character (or char)** type data . Key word **char** is used for declaring the variable of character type. Usually, a character enclosed between a pair of single quotes denotes a character constant. The size of **char** is 1 byte(or 8 bits of internal storage)..The qualifier s**igne**d or **unsigned** may explicitly applied to **char.**

1.10 Declaration of Variables

In order to use a variable in C, we must first declare it before they are used in the program. Declaration does two things:

1. It tells the compiler what the variable name (type name) is
2. It specifies what type of data (or properties) the variable will hold

The type declaration statement is written at the beginning of **main ( )** function.

Primary type instruction

A variable can be used to hold a value of any data type in a memory location. After assigning variable names, we have to declare them. The syntax for declaring a variable is:

data-type v1,v2,….vn;

Here **v1,v2,….vn** are the variable names and are separated by commas A declaration statement must end with a semicolon. For example,

i**nt** num, sum**;**

int code**;**

double ratio;

are valid declarations. Here, Keywords **int** and **double** are used to represent integer and real type data respectively. When qualifier is applied to the data type then it changes its size (The size qualifiers are :**shor**t and **long**) or its sign (sign qualifiers are: **signed** and **unsigned).** While using qualifiers like, **short, long, unsigned** without specifying the basic data type , the **C** compiler will treat the data type as **int** . Moreover, if we want to declare a character variable as **unsigned**, then we must do so by using both the terms like **unsigned char**

User Defined Declaration

In C language, a user can define an identifier that represents an existing data type. The user defined data type identifier can later be used to declare variables. The General syntax is:

typedef **type identifier;**

Here *type* represents existing data type and "identifier" refers to the row name given to the data type.

Example:

typedef int amount;

typedef float sum;

Here amount symbolizes **int** and sum symbolizes **float**. They can be later used to declare variables as follows:

amount dep1,dept2**;**

sum section1[20],section2[20];

Therefore dept1 and dept2 are indirectly declared as integer data type and section1 and section 2 are indirectly float data type.

Another user defined data type is enumerated data type provided by ANSI C standard which is defined as follows:.

enum identifier { value1,value2,…..valuen};

The "identifier " here , is a user- defined enumerated data type which can be used to declare variables that can have one of the values enclosed with in the braces . After the definition we can declare variables to be of this 'new' type as below.

enum identifier v1,v2,…..vn;

The enumerated variables v1,v2,…vn can have only one of the values value1, value2 ….. value n.

Th assignments of the following type:

v1 = value3;

v5 = value1;

are valid.

For example:

enum day { Monday, Tuesday,…….,Sunday};

enum day week_ st,week_end;

week_ st = Monday;

week_end = Friday;

If (week_st = = Tuesday)

week_end = Saturday;

The C compiler automatically assign integer digits beginning with 0 to all the enumeration constants. That is, the enumeration constant value 1 is assigned 0, value 2 is assigned 1, and so on. The automatic assignment can be overridden if we assign enumeration constant values explicitly as;

enum day { Monday = 1 , Tuesday,…….,Sunday};

Here Monday is assigned the value 1.The remaining constants are assigned values that increase successively by 1.

The definition and declaration of enumerated variables can be combined in one statement as in :

enum day { Monday, Tuesday,…….,Sunday} week _st, week_end;

Unit 2

Structure : Operators and Expressions

2.1 Introduction:

C language has a wide range of built –in operators to perform various operations. The symbols which are used to perform logical and mathematical operations in a C program are called operators. These C operators are used to join individual constants and variables to frame expressions. Moreover, operators, functions, constants and variables are combined to shape expressions. That is, operators are used with operands to build expressions. For example , the following is an expression containing two operands and one operator ' +' (an operator to perform addition).

8 + 5

whose value is 13. The value can be any type other than void. C offers the following operator Groups.

- Arithmetic
- Assignment
- Logical/relational
- Incremental and decrement operators
- Conditional
- Special Operators
- Bit wise operators.

2.2 Arithmetic Operators

The C arithmetic operators are the +, -, /, * and the modulo operator % . These C arithmetic operators are used to carry out mathematical calculations like addition, multiplication, division and modulus in c programs. Unlike /, which returns quotient, the % returns the reminder, the integer division truncates any fractional part. That is, the expression

x % y

produces the remainder when x is divided by y, and thus is zero when y divides x exactly. Note that the **operator ' % '** cannot be applied on floating point or double type data. Further, C does not have an operator for exponentiation. The operators in C with their meaning are listed in **Table** 2.1 below.

Integer Arithmetic

When both the operands in a single arithmetic expression are integers, the expression is called an integer expression, and the operation is called integer arithmetic. Integer arithmetic always yields an integer value. For example, for integer operands such as **a** and **b** with assigned values respectively, 15 and 5, we have:

a + b = 20

a - b = 10

a * b = 75

a / b = 3

a % b = 0

During integer division , if both operands are of the same sign, the result is truncated to zero. If one of them is negative, the direction of truncation is machine dependant. .That is , 6/7 = 0 and -6/-7 = 0 but -6/7 may be zero or -1(that is , machine dependent).

Similarly, **during modulo operation, the sign of the result is sign of the first operand**., as in:

-16 % 3 = - 1

-16 % -3 = - 1

16 % - 3 = 1

Operator	Meaning
+	Addition(unary plus)
-	Subtraction(Unary minus)
*	Multiplication
/	Division
%	Modulo division (reminder after division)

Table 2.1 Arithmetic Operators

The Precedence to the operations associated with the operators are listed as:

Operator type	Precedence	priority
Unary Minus	1	Highest
*, / , %	2	Second
+, -	3	Third

That is, when an expression is given for evaluation, they are evaluated from Left to Right, based on the precedence associated with the operators. On the other hand, if the precedence's associated with the operators are to be overridden, it is necessary to use parenthesis in the expression. However, the expression within the parenthesis is evaluated on the basis of the precedence rule , with parentheses again evaluated from left to right. For expressions with nested parentheses, we evaluate the innermost one first, then the one immediately outside and so on.

Real Arithmetic

The C language contains the basic real arithmetic operators. An arithmetic operation involving only real operands is called real arithmetic. A real operand may accept values either in decimal or exponential form. An arithmetic operation between an integer and integer gives an integer result, while , the result of applying the real operators to real is another real. For floating point values, it is rounded to the number of significant digits permissible, and the final value is an approximation of the corrected result. For example, if operands x, y ,z are floats, then we will have,

x = 6.0/ 7.0 = 0.857143

y = 1.0/ 3.0 = 0.333333

z = -2.0 /3.0 = -0. 666667.

The operator % cannot be used with real operands

Mixed Mode Arithmetic

If operands in an expression contains both integer and real constants or variables then it is a mixed mode arithmetic expression. That is, When one of the operands is real, an operation between an integer and real always gives a **real** result. In this operation, the integer is first promoted to a real one and then operation is performed. The expression thus obtained is called a Mixed mode arithmetic expression. For e.g., 25/ 10.0 = 2.5 while, 25/10= 2.

2.3 Relational Operators

Relational operators are used to check relationship between two operands. If the relation is true, it returns value 1 and if the relation is false, it returns value zero. The relational operators are

>, > = , < , < =

They all have the same precedence. C offers six relational operators in all. These operators and their meanings are listed in the table below.

Operator	**Meaning**
<	is less than
<=	is less than or equal to
>	is greater than
>=	is greater than or equal to
= =	is equal to
!=	is not equal to

A simple relational expression contains only one relational operator . When arithmetic operations are used on either side of a relational operator, arithmetic expressions will be evaluated first and then the results are compared. Relational operators have lower precedence than arithmetic operators and are used in decision making and loops(i.e., in statements like If and while) in C programming..The Syntax Is:

ae-1 relational operator ae-2

with **ae-1 and ae-2** representing arithmetic expressions**.**

For e.g., 4.6 < = 10 TRUE

4.6 < - 10 FALSE

x+y = y+z TRUE only if sum of values of x and y are equal to the sum of values of y and z

Relational operator complements

Among the six relational operators, each one is complement of another operator. They are as:

- > is complement of < =
- < is complement of > =
- = = is complement of ! =

We can simplify an expression involving the not and less than operators using the complements as :

! (x < y) simplified to x > = y

! (x > y) simplified to x < = y

! (x ! = y) simplified to x= = y

! (x < = y) simplified to x > y

! (x > = y) simplified to x < y

! (x = = y) simplified to x !> = y

2.4 Logical Operators.

Logical operators are used to combine expressions containing relational operators. These operators perform logical operations on the given expressions .In C there are 3 logical operators (Table 2.2) and are:

Operator	Meaning of operator
&&	logical AND
||	logical OR
!	logical NOT

. **Table 2.2**

Logical operators perform logical-AND (&&) and logical –OR (| |) operations. Its Syntax is:

logical-AND-expression:

inclusive-OR- expression

logical –AND- expression & & inclusive- OR- expression

logical-OR-expression:

logical –AND- expression

logical -OR- expression | | logical - AND- expression

some example of usage of logical expression is:

1. If (age > 60 & & salary < 300 000)

2.If (number < 0 | | number > 1000) .

Logical operators & & and | | are used when we want to test more than one condition and to make decisions. They do not perform the usual arithmetic conversions. Instead, they evaluate each operand in terms of its equivalence to 0.The result of logical operation is either 0 or 1 and is of **int** type**.** The operands of logical-AND and logical-OR are evaluated from left to right. If the value of the first operand is sufficient to determine the result of the operation, the second operand is not evaluated . The C logical operators are described below.

Operator	**Description**
&&	If both operand are non zero logical AND produces the value 1.If either operand is equal to zero, the result is zero and if the first operand is equal to zero, the second operand is not evaluated.
| |	The logical-OR performs an inclusive - OR operation on its operands. The result is 0 if both operands have 0 values. If either operands has a non zero value, the second operand is not evaluated.

While using compound expressions, care should be taken in using the precedence of relational and logical operators. The relative precedence are listed as:

! Highest

> > = < < =

= = ! =

& &

| | Lowest.

2.5 Assignment Operators.

The assignment operators perform an arithmetic operation on the 1value and assign the result to the 1value.The usual assignment operator is the '=' ,. In addition, C has a set of less frequent *shorthand* assignment operators of the form (+ +, - =, * =, / =, % =). The syntax s;

v op = exp;

where ***v*** is a variable, ***exp*** is an expression and ***op*** is a C binary arithmetic operator.(or *short hand* binary operator). For e.g., consider the statement x + = y +1 ; this is same as x= x+(y+1). Here the operator + = means add ' y + 1 to x ' (or increment x by y + 1) . Some of the commonly used *short hand* assignment operators with their description is shown in Table 2.3 In all expressions involving these operators, the type of an assignment expression is the type of its left operand, and the value is the value after the assignment.

Statement with simple assignment operator	**Statement with assignment operator**
a = a + 1	a + = 1
a = a - 1	a - = 1
a = a* (n+1)	a* = n+1
a = a/(n+1)	a / = n+1
a = a % b	a % = b

Table 2.3 Short hand assignment operators

2.6 Increment and Decrement operators.

C provides two operators ++ and - - called increment and decrement operators and these operators are useful in controlling the loops through an index variable. The + + operator adds 1 to its operand while the decrement operator - - subtracts 1. Both of these operators are unary operators. (That is, used on single operand. ++ adds 1 to operand and - - subtracts 1 to operand respectively). For example:

Let a = 3 and b = 7

a ++ ; becomes 4 and a - - becomes 6

The unusual aspect is that ++ and - - may be used either as prefix (before the variable as in ++a) or post fix (after the variable as in a ++) . In both case effect is to increment a. But the expression ++a increments a before its value is used, while a ++ increments a after its value has been used. This means that in a context where the value is being used, not just the effect, + + a and a++ are different. For e.g., in the assignment statement x = i ++, if i =5, then x = i++ sets x= 5 , but x = ++ i sets x to 6. In both case i becomes 6. The increment and decrement operators can only be applied to variables, an expression like (i +j) ++ is illegal. In general, a prefix operator first adds 1 to the operand and then the result is assigned to the variable on the left. On the other hand, a post fix operator first assigns the value to the variable on left and then increments the operand.

Similar is the case, when we use ++ or - - in subscripted variable. That is, the statement

```
a[ i++ ] = 5;
```

Is equivalent to

```
a[i] =5;
i = i+1;
```

Rules for increment (++) and decrement (- -) operators.

1.They are unary operators and require variable as their operands.

2.A postfix ++ or - - operator used with a variable in an expression, the expression is evaluated first using the original value of the variable and then the variable is incremented(or decremented by one).

3 When prefix ++ or - - is used in an expression, the variable is incremented (or decremented) first and then the expression is evaluated using the new value of the variable.

4.The precedence and associativity of ++ and - - operators are the same as those of unary + and unary -

2.7 Conditional Operator

Conditional operator (? :) is a ternary operator (that demands three operands) consisting of symbols " **?"** and "**:** " and are used for decision making in C. The operator works by evaluating test expression, returning a value if that expression is TRUE and different one if the expression is evaluated as FALSE. The general syntax is:

identifier = (test expression) ? expression1 : expression2;

This is an expression, not a statement, so it represents a value. If the condition (or test expression) is true , it evaluates and returns expression1, otherwise it evaluates and returns expression2 .Conditional operator can be used as a short hand for some **if-else** statements. For example, consider the statements,

```
a = 10;
b = 20;
x = ( a > b) ? a : b;
```

Here in this example, x will be assigned the value of b. This can be achieved using the **if…..else** statement as follows:

```
If ( a > b)
    x = a ;
else
   x = b;
```

2.8 Bitwise Operators

Bit wise operations in C are carried out by using operations on bits(or lowest form of data that can be accessed in digital hardware) at individual level. That means , Bit wise operators are used to perform bit operations on given two variables. Four commonly used bit wise operators in C are ~ , & ,| , and ^. Generally, Bitwise operators manipulate the value of individual bits(i.e., 1 or 0). Further, to understand "<< "and ">>" , there are two shift operators which are used to shift the position of a bit (or a set of bits) to another location, in a multi-bit value. Moreover, these operators work only on a limited number of types: **int** and **char.** That means**,** they may not be applied to data types : **float** and **double**. Bit wise operators supported by C are listed in the following table.

Operator	Description of the operator
&	Binary AND operator copies a bit to the result if it exists in both operands(or Bitwise AND)
|	Binary OR operator copies a bit if it exists in either operand(or Bitwise Inclusive OR).
^	Binary XOR operator copies the bit if it is set in one operand but not both (or Bitwise Exclusive OR).
~	Binary Ones complement operator is unary and it has the effect of flipping bits(or Bitwise ones complement).
<<	Binary left shift operator(or bitwise left shift). The left operands value is moved left by the number of bits specified by the right operand.
>>	Binary right shift operator (or bitwise right shift). The left operands value is moved right by the number of bits specified by the right operand

Table 2.4 Bit wise operators

2.9 Special Operators

C language provides a number of special operators which have no counter parts in other languages. These operators include **comma** operator, **sizeof** operator, **pointer** operators(& and *) and member selection operator (. and -- >) . Pointer operators will be discussed while introducing pointers and member selection operators will be discussed with structures and union. The c**omma** and **sizeof** operators are discussed in this section.

The Comma Operator

This operator is used to link the related expressions together. A coma -linked list of expressions are evaluated left to right and the value of right most expression is the value of the combined expression. For example, the statement

```
int x, y,z;
z =  ( x =10, y = 20, x * y);
```

Here the 1st statement will create three integer type variables : x, y,z . In the 2nd statement, R.H.S will be evaluated first. As a result, 10 will be stored in variable x, then 20 will be stored in variable y and then values in x and y will be multiplied, result of which will be stored in the variable z as 200 at the end of the execution. Since comma operator has the lowest precedence of all operators, the use of parentheses are necessary.

The size of Operator

The **sizeof** operator works on variables, constants and even on data types. It returns the number of bytes, the operand occupies in the memory. It is a compile time operator and when used with an operand, it returns the number of bytes occupied by its operand on that particular machine.

Examples include:

```
m = sizeof (sum);
n =  sizeof( long int);
o =  sizeof ( 235L) ;
```

The **sizeof** operator is normally used to determine the lengths of arrays and structures when their sizes are not known to the programmer and is also used during program execution, for dynamic memory space allocation of variables.

2.10 Arithmetic Expressions

Arithmetic expressions have numbers and variables combined with the regular numeric operators (+ , - , *, /) , as per syntax of the language and simplify to a single number .Some of the examples of C expressions are given in the table below.

Algebraic Expression	**C Expression**
a×b-c	a*b-c
ab/c	a*b/c
ax^2+bx+c	a*x*x+b*x+c

2.11 Evaluation of Expression

Every expression is formed out of operands and operators. Expressions in C, are evaluated using an assignment statement of the form:

variable = expression;

Usually when a statement is encountered, the expression (on the RHS) is first evaluated and the result obtained thus, is used to replaces the previous value of the variable on the LHS. All variables used in the expression must be assigned values before evaluation is attempted. An example of a valid evaluation expression is;

x = a* b-c;

Remember that blank space around an operator is optional and adds only to improve the readability..

2.12 Precedence of Arithmetic Operators

The two distinct priority levels of arithmetic operators in C are:

* / % High priority

+ - Low priority

An arithmetic operation without parentheses will be evaluated from *left to right*, using the rules of operator precedence. The basic evaluation procedure involves *two left to right pass* through the expression..During the 1st pass, high priority operators (if any) are applied. and during the 2nd pass low priority operators, if any , are applied as they are encountered. For example, consider the statement,

x = a-b/3 + c*2-1

when a= 9, b=12, and c =3 , the statement becomes

x = 9- 12/3 + 3*2 -1

1st pass

Step 1: x = 9- 4 + 3*2 -1

Step 2: x = 9-4+6-1

Second pass

Step 3: 5+6-1

Step 4: 11-1

Step 5: 10

However, one can change the order if evaluation, by introducing parentheses into the expression. The same above expression in parentheses reads as:

x = 9- 12/(3 + 3)*(2 -1)

Whenever parentheses are used, the expression contained in the left most set is evaluated first and the expression on the right most the last. The steps are as follows:

First pass:

Step 1: 9-12/6*(2-1)

Step 2: 9-12/6*1

Second Pass

Step 3: 9-2*1

Step 4: 9-2

Third pass

Step 5: 7

Though the procedure here, involves three left to right passes, number of evaluation steps is equal to the number of arithmetic operators. That is, the number of evaluation steps is same (equal to 5) for evaluation without and with parentheses

It may happen that parentheses may be nested, in which case evaluation will proceed outward from the inner most set of parentheses as in eg;, x = 9- (12/(3 + 3)*2) -1 = 4 .

Rules for evaluation of Expression

1. The arithmetic expressions are evaluated from left to right using the rules of precedence.
2. When parentheses are used , the expression with in the parentheses assume highest priority
3. First parenthesized sub expressions from left to right are evaluated.
4. The precedence rule is applied in determining the order of application of operators in evaluating sub expressions.
5. The associativity rule is applied when two or more operators of the same precedence level appear in a sub expression.
6. If parentheses are nested, the evaluation begins at the inner most sub expression

2.13 Some computational problems

On most computers, any attempt to divide a number by zero will result in an abnormal termination of the program. In such instances, care should be taken to test the denominator that is likely to assume zero value so that the division by zero error may be avoided. Further, one must specify the correct type of operands and it should be of the correct range, so that any error due to over flow / under flow may be eliminated.

2.14 Type conversion in expressions

C lets mixing of constants and variables of different types in an expression. It automatically, converts any intermediate values to the proper type so that expressions can be evaluated without loosing any significance. This automatic conversion is called *implicit type conversion*. If the operands are of different types, the lower type is automatically converted to the higher type before the operation proceeds. The result is of higher type. The sequence of rules to be followed while evaluating an expression are given below.

Rules for evaluating expressions

All s**hort** and **char** are automatically converted to **int:** then

1. If one of the operand is **long double**, the other will be converted to **long double** and the result will be **long double**.
2. else, if one of the operands is **double**, the other will be converted to **double** and the result will be **double**.
3. else, if the operand is **float**, the other will be converted to **float** and the result will be **float**;
4. else if one of the operand is **unsigned long int**, the other will be converted to u**nsigned long int** and the result will **be unsigned long int**.
5. else, if one of the operands is **long int** and the other is **unsigned int**, then

(a) If **unsigned int** can be converted to **long int**, the **unsigned int** operand will be converted as such and the result will be **long int;**

(b) else, both operands will be converted to **unsigned long int** and the result will be **unsigned long int;**

6. else, one of the operands is **long int**, the other will be converted **to long int** and the result will be **long int;**
7. else, if one of the operands is **unsigned int**, the other will be converted to **unsigned int** and the result will be **unsigned int**.

Explicit conversion

Explicit conversion is used to tell the compiler to treat a variable as of a different type in a specific context. The compiler will automatically change one type of data in to another (or locally convert) to make it sense. For instance, if you assign an integer value to a floating point variable, the compiler will insert code to convert the **int** to a **float**. The **general syntax** is:

(type-name)expression

Where type-name is one of the standard C data types. The expression may be a constant, variable or an expression. Casting allows you to make this type conversion explicit, or to force it when it would not normally happen. To perform casting, put the desired type including modifiers like unsigned inside parentheses to the left of the variable or constant you want to cast. For Example

```
float a = 5.25;
int b = (int)a;     /*Explicit casting from float to int */
```

The value of b here is 5.

2.15 Operator Precedence and associativity

Two operator characteristics (or precedence and associatively of operators) determines how operators group with operators. Precedence is the priority for grouping different types of operators with their operands. Associativity is the left to right or right to left order for grouping operand to operators that have the same precedence. An operator's precedence is meaningful only if other operators with higher to lower precedence are present. Expressions with higher-precedence operators are evaluated first. The grouping of operands can be forced by using parentheses Operators that have the same rank have the same precedence.

For example, in the following statements, the value of `1` is assigned to both `a` and `b` because of the right-to-left associativity of the `=` operator. The value of `c` is assigned to `b` first, and then the value of `b` is assigned to `a`.

```
b = 2;
c = 1;
a = b = c;
```

Because the order of sub expression evaluation is not specified, you can explicitly force the grouping of operands with operators by using parentheses.

In the expression

```
a + b * c / d
```

the `*` and `/` operations are performed before `+` because of precedence. `b` is multiplied by `c` before it is divided by `d` because of associativity. Table 5.8 gives a complete list of C operators, their precedence levels , and their rules of association.

Operator	Description	Associativity
( ) [ ]	Function call Array element reference	Left to right Right to Left
+	Unary plus	Right to left
-	Unary minus	
++	increment	
- -	decrement	
!	Logical negation	
~	Ones complement	
*	Pointer reference	
&	address	
Sizeof (type)	Size of an object Type cast	
*	multiplication	Left to right
/	division	
%	Modulo	

+	addition	Left to right
-	subtraction	
<<	Left shift	Left to right
>>	Right shift	
<	Less than	Left to right
< =	Less than or equal	
>	Greater than	
> =	Greater than or equal to	
= =	equality	Left to right
! =	In equality	
&	Bitwise AND	Left to right
^	Bitwise XOr	Left to right
|	Bitwise OR	Left to right
&&	Logical AND	Left to right
||	Logical Or	Left to right
?:	Conditional expression	Right to left
=	Assignment operators	Right to left
* = /= % =		
+ = - = & =		
^ = | =		
< < = > > =		
,	Comma operator	Left to right

Table 2.5 Precedence and Associativity of operators

Unit 3

Structure

Introduction

Reading a Character

Writing a Character

Formatted Input

Formatted output

3.1 Introduction

In order to learn a program effectively in C language, one should know, how to manage input and output of data to and from the screen and the key board. Most programs take some data as input and display the processed data, often as results, on a suitable medium. The two methods so far used, for providing data to program variables, rely on : (1) Assigning values to variables through assignment statements and (2) using the input function **scanf (**to read data from a key board). For getting the output results, usually the **printf** function that sends results out to a terminal, is used.

The Input and output operations are convenient for program that interact with the user, takes input from the user and print the message. Unlike, other higher level languages, C does not provide any input-output (I/O) statements as part of its syntax. Instead , a set of library functions provided by the operating system for input and output operations are borrowed and used by C. The standard library for I/O operations used in C is **stdlib**. That is , Standard input (or **stdin**) is a data stream used to receive input from user / collects characters typed at the keyboard and **stdout, is** the data stream for sending output to a device such as monitor etc., . In otherwords, to include input and output functionality in C programs, the **stdio** header is needed. Each program that uses a standard I/O function must contain the statement

 # include < stdio.h >

at the beginning. This instruction tells the compiler, 'to search for a file named **stdio.h** and place its contents at the appropriate place in the program . Indeed, the contents of the header file become part of the **source code** when it is compiled. In fact, this statement can be avoided in situations, where the functions **printf** and **scanf** have been defined as part of the C language. Here, in this chapter, a brief introduction of some common I/O function that can be used in many machines without much change is discussed.

3.2 Reading a Character

The simplest of all I/O operations is reading a character from the standard input unit(or key board) and writing it to the standard output unit(or the screen). The most basic way of reading input is by calling the function **getchar**. The C library function **getchar** gets a character from **stdin**, regardless of what it is, and

returns it to the program. That is, it is used to get a character from console, and echoes to the screen. It is the most basic input function in C, included in the **stdio.h** header file. The **getchar** takes the following form:

variable_name = getchar( );

Variable name is a valid C name that has been declared as of **char** type. When this statement is encountered, the computer waits until a key is pressed and then assigns this character as a value to g**etchar** function. Since **getchar** is used on the RHS of an assignment statement, the character value of **getcha**r is in turn assigned to the variable name on the left. For example,

char = name;

name **= getchar ( );**

Will assign the character "**a"** to the variable name when we press the key **a** on the keyboard. Since **getchar** is a function, it requires a set of parentheses as shown. The use of **getchar** function is illustrated in the program below.

Program **Output**

```
#include <stdio.h>
#include<conio.h>
int main( )
{
    char a;
    clrscr( );
    printf("Enter a character\n");
    a=getchar( );
    printf("The character entered is %c \n",a);
    getchar( );
    return 0;
}
```

```
Enter a character
b
The character entered is b
```

The **getchar** function may be called successively to read the characters contained in a line of text..The following program me segment , for example, reads characters from key board one after another until the 'return key' is pressed

```
------------

------------

call  character;

character = '  ';

while ( character ! = '\n' )

   {

         character = getchar ( );

    }

-------------

-------------
```

The **getchar** returns the character it reads, or, if there are no more characters accessible, it will return the special value EOF ("end of file") .That is, The **getchar** function accepts any character keyed in, This includes TAB and RETURN . In other words, when we enter single character input, the newline character is waiting in the input queue after **getchar( )** returns. A dummy **getchar or fflush function (**to flush out unwanted functio**n)** may be used to get away the unwanted new line character , when we use **getchar** in a loop interactively. However, **getc** is used to accept a character from standard input.

3.3 Writing a Character

Often there do occur circumstances, where we want to solve computational problems and to display the characters therein on the console. The two special functions in C, that is designed to handle the output of character to monitor is ***putch*** and ***putchar* . That is,** Like **getchar**, there is an analogous companion C library function **putchar** that writes a single character to the standard output stream, (or console), specified by the argument **char** to s**tdout(i.e., it** is same as calling **putc(c,stdout).** The **putchar** function displays a single character on the screen. The syntax is:

putchar (variable_name);

where variable_name is a type **char** variable containing a character. **For e.g.,** the statement

```
answer = 'N'
putchar (answer);
```

will display the character N on the screen. The statement

putchar ('\n');

would cause the cursor on the screen to move to the beginning of the next line. The following example (Fig. 3.1) explains the use of **putchar( )** function. ***Putch( )*** function, on the other hand is useful in writing the output, character by character, on the display.

The puts Function

The ***puts*** function stands for put string (or a bit of text) to the screen and this function works inside the main function. That means, the function puts( ) writes ***str*** to ***stdou***t, then writes a new line character. The general form of the function is:

int puts (char A [ ]);

A ***puts()*** function automatically appends a new line character at the end of any text it display and it uses a character array as parameter which is displayed on the screen. The ***puts()*** function performs a function that is similar to printf( ) with a %s conversion specifier (or formatted text display). However, **putc** is used for sending a single character to standard output.

```
# include <stdio.h>

int main ( )

 {

   char ch ;

   for (ch = 'A'; ch < = 'Z' ; ch++)  {

        putchar (ch);

   }

  return (0);

}
```

Output

ABCDEFGHIJKLMNOPQRSTUVWXYZ

Fig 3.1: Program to read and write all letters in English alphabet

3.4 Formatted Input

The standard formatted input function in C is **scanf (**that supply input in a fixed format**)** and is the input analog of **printf,** providing many of the conversion facilities in the opposite direction**..** The **scanf** contains two important things –the **format string** and the **address list** and it reads characters from the input file and converts them to internal form.. That is, **scanf** reads characters from the standard input, interprets them according to the specifications in format, and stores the results through the remaining arguments. Very often, This is the function used to read an input from the command line. The general format of an input statement is:

scanf(" format string", arg1,arg2,……, arg n);

Here the format string gives information to the computer on the type of data stored in the list of arguments arg1, arg2,….arg n and in how many columns (or address of locations) they are found. That is, format string specifies, how each input is read(.i.e., as a decimal integer, a decimal float, a character, a string and so on in matching arguments). The argument must be a pointer to a data type that is being read. In fact, format string and arguments are separated by commas.

scanf stops when it exhausts its format string, or when some input fails to match the control specification. It returns as its value the number of successfully matched and assigned input items. This can be used to decide how many items were found. On end of file, EOF is returned; note that this is different ' from 0, which means that the next input character does not match the first specification in the format string. The next call to **scanf** resumes searching immediately after the last character already converted. The format string usually contains conversion specifications, which are used to control conversion of input. The format string may contain:

- Blanks or tabs, which are ignored.
- Ordinary characters (not %), which are expected to match the next non-white space
- character of the input stream.
- Conversion specifications, consisting of the character %, an optional assignment suppression
- character *, an optional number specifying a maximum field width, an optional h, 1, or L indicating the width of the target, and a conversion character

A conversion specification directs the conversion of the next input field. Normally the result is placed in the variable pointed to by the corresponding argument. If assignment suppression is indicated by the * character, however, the input field is skipped; no assignment is made. An input field is defined as a string of non-white space characters; it extends either to the next white space character or until the field width, if specified, is exhausted. This implies that **scanf** will read across line boundaries to find its input, since newlines are white space

Inputting Integer l numbers

The field specification for reading an integer number is

% w sd

The percentage sign (%) indicates that a conversion specification follows.. **w** is an integer number specifying the field width of the number to be read and **d** the data type. For example, in the statement

scanf("%3d %5d", &num1,&num2);

the two variables in which numbers are to be stored are num1 and num2 and are of integer type. The input data items must be separated by spaces, tabs or new lines. A sample data line may thus be;

500 31246

The value 500 is assigned to num1 and 31246 to num2**.** Observe that the symbol & (ampersand) should precede each variable name, that is used to indicate the address of the variable name.

The **scanf** statement causes data to be read from one or more lines till numbers are stored in all the specified variable names. Also no blanks are permitted between characters in the format-string. The data type character d may be preceded by l to read long integers and h to read short integers.

Inputting real numbers

The **scanf** reads real numbers using the specification %f for both decimal and exponential notation. The input field specification may be separated by any arbitrary blank spaces. If the number to be read is of double type, then

Program	**Output**
main( )	values for x and y is : 12.3456 17.5e-2
{	x=12.345600
float x,y;	y=0.175000
double p,q;	
printf(“values for x and y is :\n”);	values of p and q is :4.142857142857
scanf(“%f %e” , &x ,&y);	18.5678901234567890
printf(“\n”);	p= 4.142857142857
printf(“x= %f\n y= %f\n\n”, x, y);	q= 1.8567890123456e+001
printf(“values of p and q is: ”);	
scanf(“%lf %lf ”, &p, &q);	
printf(“\n\np = % .12lf \np = %.12e”, p, q);	
}	

Table 3.2: Reading of Real Numbers

the specification should be %**lf.** Consider the statement

scanf(“%f %f %f”, &p,&q, ,&r) ;

with the data line

462.85 41.23E-1 543

It will assign the value 462.85 to p, 41.23E-1 to q and 543.0 to r. T he program (Table 3.2) below shows how to read real numbers in both decimal and exponential notation

Inputting character strings

A **scanf** function can input strings containing more than one character. The syntax is:

%ws or %wc

The corresponding arguments should be a pointer to character array. When the argument is a pointer to a ***char*** variable, then %c may be used to read a single character**.** Some **scanf** versions support the following string conversion specification.:

% [characters]

% [^ characters]

The specification **% [characters**] imply that only the characters within brackets are permissible in the input string. Any encounter of other string characters, will terminate the string. The specification % **[^characters]** does exactly the reverse. That is , characters after the ^ are not permitted in the input string, The reading of the string will be terminated at the encounter of one of these characters.

Reading Mixed data types

scanf can be used to input data containing mixed mode type. When one attempts to read an item that does not match the type , the **scanf** function does not read any further and immediately returns the value read. For e.g.,

scanf("%d %c %f", c %s " , &count, &code, &ratio, &name) ;

will read the data line

15 p 1.453 coffee

Correctly and assign values in the order in which they appear.

Rules for scanf

- Each variable to be read need a filed specification and a variable address of proper type.
- For any non -white space character used in the format string there must be a matching character in the user input.
- Ending the format string with white space will result in error.
- The **scanf** reads until:

1. A whitespace character is found in the numeric specification or
2. Maximum number of characters have been read
3. An error is detected.
4 .The EOF is reached

3.5 Formatted output

Formatted output refers to an output data that has been arranged in a particular format, using certain features, that are effectively exploited to control the alignment and spacing of print-outs on the terminals.. The main output routine is **printf** , which writes a formatted string to the **stdout** stream. The **printf( )** function is used to print the character, string, float, integer, octal and hexa decimal values on to the output screen and it returns the number of characters that was written if an error occurs, it will return a negative value. The required header for the **printf function** is:

#include <stdio.h>

The general form of **printf** statement is :

printf (" control string"' arg1,arg2,…., arg n);

Control string consists of three types:

1.character that will be printed on the screen as they appear.

2.format specification

3.escape sequence characters like, \n,\t, and \n.

The control string specifies the number of arguments (or variables whose values are formatted and printed according to the specification of control string) that follow with their types. The arguments should match in number, order and type with the format specification. A simple format specification is as:

% w. p type-specifier

Where w , is an integer specifying the total number of columns for output value and p is another integer that specifies the total number of digits to the right of the decimal point or the number of characters to be printed from a string.

Printf formatting is controlled by 'format identifiers' which in the simplest form are listed below:

%d % i	decimal signed integer.
% o	octal integer
%x % X	Hex integer
%u	unsigned integer
% c	character
%s	string
%f	double
%e %E	double
%p	pointer
%n	number of characters written by this printf, no argument expected
%%	% .No argument expected.

Output of Integer Numbers

The format specification for printing an integer number is:

% w d

Where **w** specifies the minimum field width for the output and d , the value to be printed as an integer. However, if a number (right justified in the given field width with leading blanks) is greater than the specified field width, it will be printed in full, over riding the minimum specification. It is possible to force the printing to be left- justified by placing a minus sign directly after the % character. More over, it is possible to pad with zeros the leading blanks by placing a zero before the field width specifier. Here, The minus (-) and zero (0) are named as flags. For printing short integers we may specify **hd** . And for printing long integers the specifier **ld** is used in place of **d** in the format specifier. Some examples of different format are:

Format	output
Printf(%d"', 1076)	1076
Printf(%6d"', 1076)	1076
Printf(%-6d"', 1076)	1076
Printf(%06d"', 1076)	001076

Output of Real Numbers:

Using the following form specification, the output of a real number may be displayed in decimal form:

% w.p f

The integer **w** indicates the number of positions that are to be used for the display of the value and the integer p represents the number of digits to be displayed after the decimal point. That is, the values when displayed, is rounded to p decimal places with right justification in the field of w columns, with leading trails and blanks. The default precision is actually 6 decimal places. The negative numbers will be printed with the minus sign and of the form [-] mmm-nnn.

A real number can be displayed in exponential form using the specification:

% w. p e

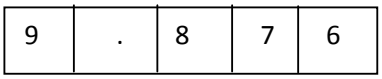
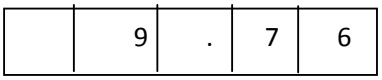
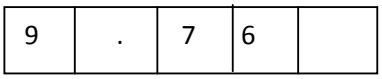
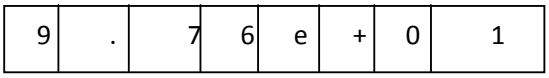
The display is of the form

[-] m.nnnne[±]xx

Where the length of the string n 's is specified by the precision p with the default precision being 6..Moreover, the field width w should satisfy the condition

$w \geq p + 7$

and will be rounded off and printed right justified in the field of w columns. Further, padding the leading blanks with zeros and printing with left justification using flags 0 or – before the field specifier is also possible. Following are some examples:

Format	output
Printf("%5.3f",x)	
Printf(%5.2f"',x)	
Printf(%-5.2f"',x)	
Printf(% -8.2e"',x)	

For dynamic format specification during run time (i.e., with field width and precision given as arguments for w and p) we have the special field specification:

printf("%*.*f" , width, precision, number);

For e.g.,

printf('%*.*f", 7,2, number);

Is equivalent to

printf('%7.2f", number);

Printing of a single character

A single character can be displayed in the keyboard at the desired position , right justified in the field of w column (with default value for w being 1) using the format

% wc

Printing of strings

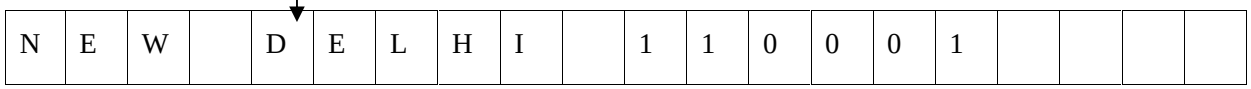
The format specification for outputting strings is similar to that of real numbers.. The format being:

% w. ps

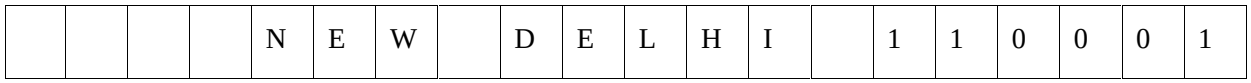
With w the field width for display and p indicates that only first p characters of the string are to be displayed with right justification..Some examples are:

Table showing specification and out put

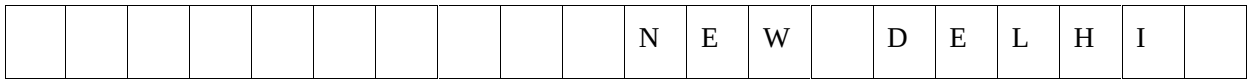
%s (specification) output



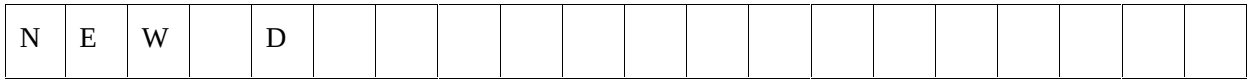
%20s(specification) output



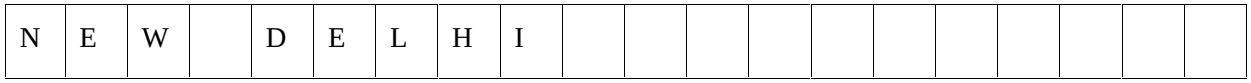
% 20.10s(specification) output



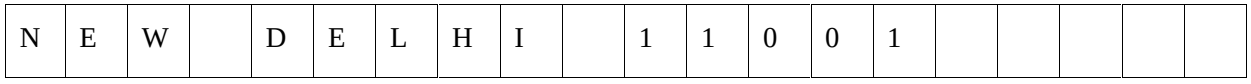
%.5s(specification) output



%-20.10s(specification) output



%5s(specification) output



Mixed data output

Mixed data types in one printf statement is permitted in C. For e.g.,

p**rintf("%d % f % s %c ,a,b,c,d);** is a valid one.

Code	Meaning
%c	Print a single character
%d	Print a decimal number
%e	Print a floating point number in exponent form
%f	Print a floating point number Without exponent form
%g	
%i	Print a floating point number Either e-
% o	type
%s	or f-
%u	type
%x	Print a signed decimal integer
	Print an octal integer without leading zero.
	Print a string
	Print an unsigned decimal integer
	Print a hexagonal integer, without leading 0.s

Table 3.3: Printf format codes

Remember that, the format specification should match the variables in number, order and type. Table 3.3 below shows commonly used **printf** format codes

The letters used as prefix for certain conversion characters are:

h short integer

l long or double

L for long double .

Unit 4

Structure : Decision making and Branching

4.1 Introduction.

Decision making is one of the most important concepts in C programming. That is, the programs should be able to make logical decisions based on the conditions they are in. **C** language has three major decision making instructions- the **if** statement, the **if else** statement, and the **switch** statement. These statements 'control' the flow of program execution (or they specify the order in which computations are performed), and are known as **control statements**. Here we will learn each of these, and discuss their features, capabilities and applications in more detail.

4.2 Decision Making with if statement

The key word, **if** statement, is a conditional branching statement. **It,** instructs the compiler that, what follows is a decision control instruction. That is, it allows the program to select an action (i.e., a condition is evaluated, and if it is true the statement is executed, and, the program skips past it if it is found false) based upon the user's input. The condition following the keyword if is always enclosed within a pair of parenthesis. It takes the form:

If (test expression)

A decision control instruction can be implemented in C using (1) The simple if statement, (2) The if – else statement (3) nested if-else statement and (4) else if ladder.

4.2.1 Simple if statement

The general form of **if** statement looks as:

```
if (test expression)
{
    statement block;
 }
 statement –x;
```

Here the expression can be any valid expression including a relational expression. We can even use arithmetic expressions in the **if** statement. In fact a compound statement composed of several statements enclosed with in braces (braces are used to group declarations and statements together into a compound statement or block), can replace the single statement. Remember, there is no semicolon after the right brace that ends a block. If the test expression evaluates to true, then the compound statement is executed. Otherwise the control jumps to the statement following the right brace ignoring the compound statement.. ***Please do remember that in C, a non zero value is considered to be true, where as a zero is considered to be false.*** Here is a simple program (Figure 4.1) using simple ***if statement***:

```
/* Demonstration of  if  statement*/
# include < stdio.h >
# include < conio.h>
int main ( )
 {
   int number;
   clrscr ( );
   printf ( " enter a number\n");
   scanf(" %d", &number);
   If (number > 0)
        printf(" The given number is positive\n");
   getch( );
   return 0;
}
```

output

```
enter a number
5
The given number is positive
```

Fig. 4.1: Program illustrating simple if statement

On execution of this program, if you type a number greater than zero, you will get a message on the

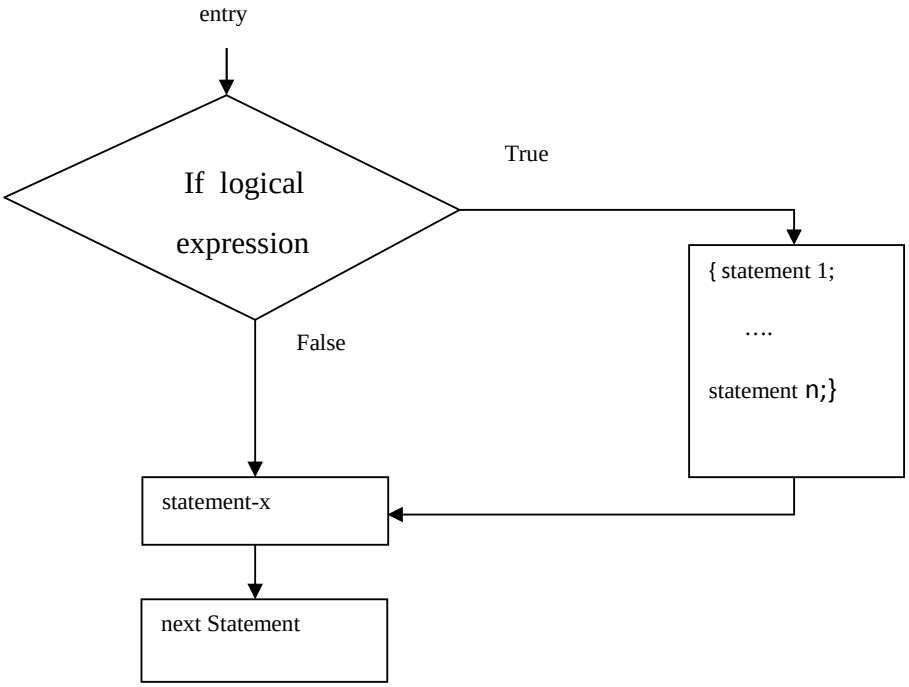


Fig 4.2: Flowchart illustrating simple if statement

screen through **printf( ).** If you type some other number(i.e., a number less than 0, the program does not do anything. The Flow chart given in Fig. 4.2 help you understand the flow of control in **simple if** statement.

4.2.2: If-Else statement

The **if** statement by itself will execute a group of statements or a single statement, when the expression following it evaluates to **true** and it does nothing when it evaluates to f**als**e .In fact, the **if –else** statement is an extension of the simple **if** statement and is used to express decisions. It permits the programmer to write a single comparison, and then execute one of the two statements depending on whether the test expression (in parentheses) is true or false. That is, the **if…else** statement is used, the intention of the programmer is- to execute the group of statements denoted as true (.i.e., the true block of statements immediately following the **if** statements), or else the test expression statements denoted as false are executed..In either case, either a true or a false block of codes/statements, are executed not both .In both cases, control is transferred to the subsequent statement-x. This is interpreted in the flow chart of Fig.4.3.

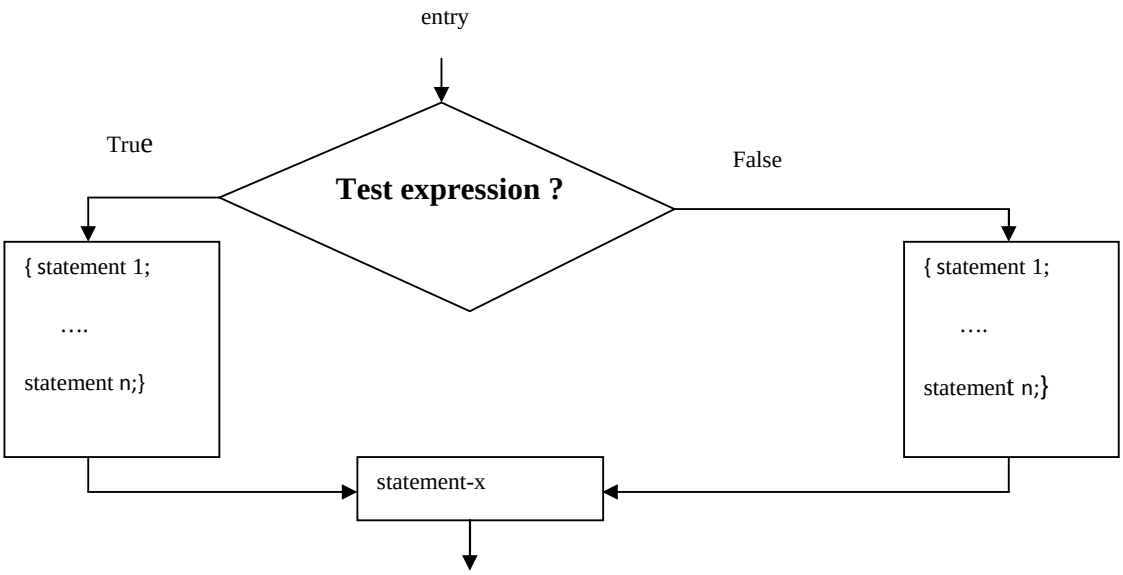


Fig. 4.3: Flowchart illustrating if-else statement

Example 4.1: A program to check whether the number is odd or even?

```
# include < stdio.h >

int main ( ) {

      int number;

      printf(“ Enter a number.\n”);

      scanf(“%d”, &number);

      if ((number  % 2) = = 0)

            printf(“%d is even,” , number);

     else

            printf(“%d is odd..” , number);

           return 0;

}
```

Output

```
Enter  a number
22
22 is even.
```

Fig 4.4: Program to illustrate if-else statement

There are a few points worth mentioning.

1.The group of statements after the if up to and not including the else is the ‘ if block’. Similarly, the statements after the else form the ‘else block’.

2.The statements in the if and those in the else block have been indented to the right.

3. As with the if statement, the default scope of else is also the statement immediately after the else. In order to override this default scope, a pair of braces must be used.

4.2.3: Nested If-else statement

The if….else statement can be used in nested form when a serious decision are involved. In nested if ..else construct, we write an entire if-else construct with in either the body of the if statement or the body of an else statement. The logic of execution is shown in Fig. 4.5.The syntax is:

```
if  (test condition-1)
     {
       if  (test condition-2);
       {
         statement-1;
       }
       else
        {
         statement-2;
        }
   }
 else
  {
    statement-3;
  }
  statement-x;
```

Here, if the test expression -1 is false, the statement -3 will be executed; otherwise control of the program jumps to perform the second test condition. If the condition- 2 is true, the statement-1 will be evaluated, otherwise the statement-2 will be evaluated and then the control is transferred to the statement-x.

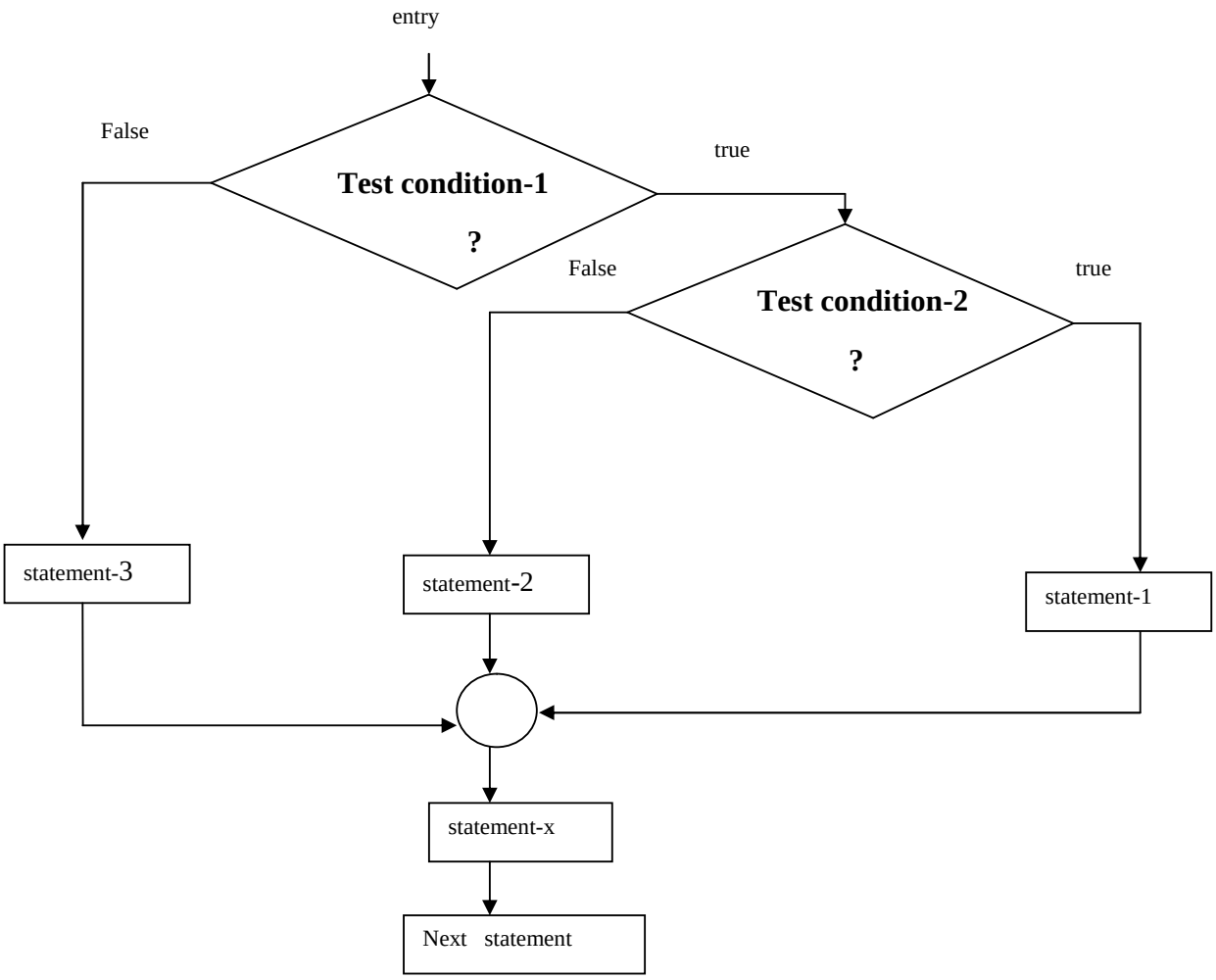


Fig 4.5: Flowchart illustrating nested if-else statements

Example 4.2: Program to check the validity of Law of Trichotomy for two real numbers.

```
# include < stdio.h >

int main ( ) {

      int num1, num2;

      printf(" Enter two integers.",\n);

      scanf("%d %d"; & num1, &num2);

      if (num1= = num2)

          printf( result: %d=%d", num1,num2);

     else

         if(num1> num2)

             printf("result:%d > %d", num1,num2);

          else

             print("result: %d >%d ",num2,num1);

 return 0;

}
```

Output

Enter two integers

4

2

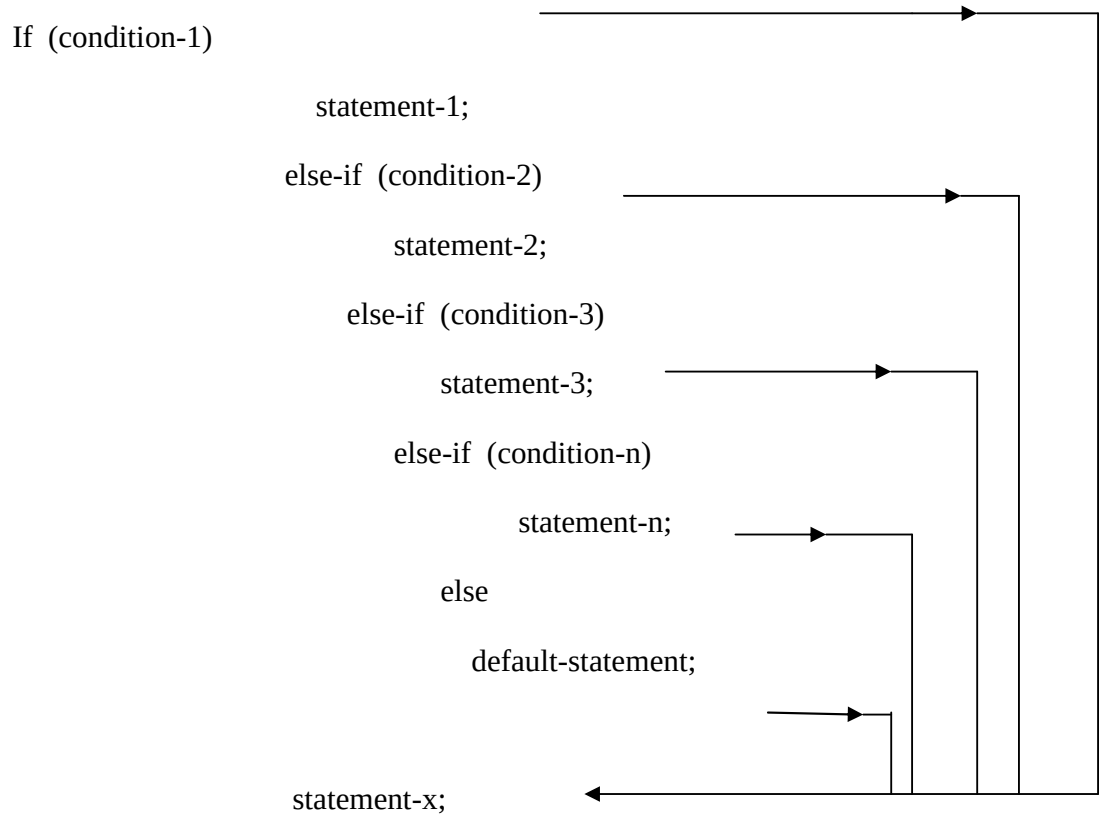
Result:4>2

Fig 4.6: Program illustrating nested if-else statement

4.2.4 The else -If Ladder

Another way of describing the nested **if-else** is the **else-if** ladder, where, every **else** is associated with an **if** statement. That is, **else-if**, is a combination of **if** and **else**. Like **else**, it extends an **if** statement to execute a different statement in case the original **if** expression is evaluated as False.

The syntax is:



This construct is called the **else-if ladder** and is useful where two or more alternatives are available for selection. In **else-if** ladder various conditions are evaluated one by one starting from top to bottom, on reaching a condition evaluating to TRUE the statement group associated with it are executed and skip other statements. If none of the expressions is evaluated to true, then the statement or group of statements associated with the final **else** is executed. In this construct nesting is allowed only in the **else** part . In fact, In **else……if** ladder, we do not have to pair **if** statements with **else** statements. That is, there is no need to remember the number of braces opened as in nested **if….else.** Moreover, **else….if** ladder produces the same effect as **nested if-else** with the benefit that it is easy to code. The flow chart corresponding to **else-if** ladder is shown in fig.4.7.

In this construct, the conditions are checked, starting from the top of the **else-if** ladder, moving downwards. That is, firstly, condition-1 is checked, and if it is true, statement-1 is executed and control is transferred to statement-x. On the other hand, **If** condition-1 is false, condition-2 is checked and if true, statement -2 is executed and control is transferred to statement-x skipping the rest of the ladder .When all the n conditions are false, then the final default-statement is executed followed by the execution of statement-x. The following program(Fig.4.8) explains the **else-if** construct.

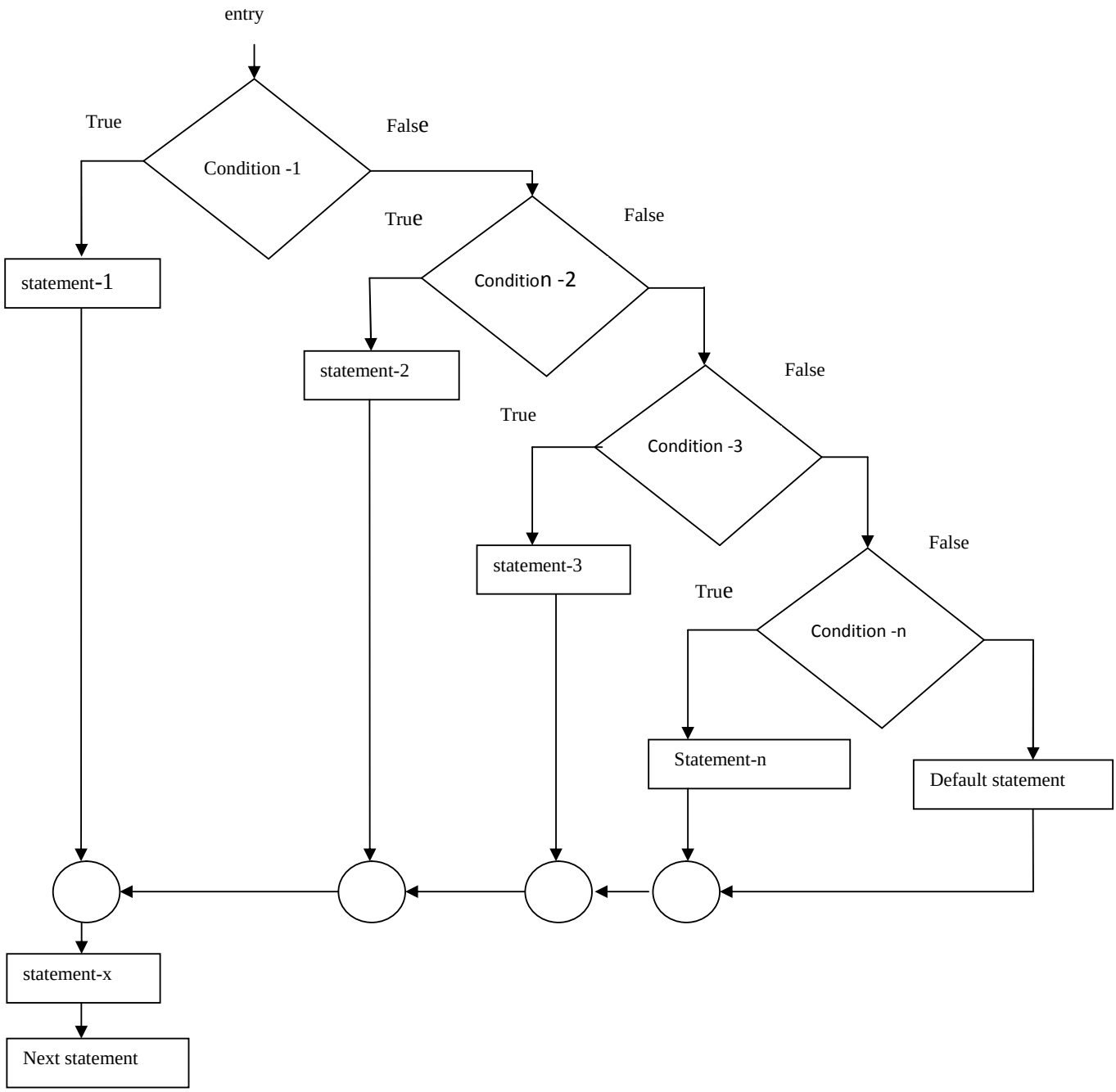


Fig 4.7: Flowchart illustrating the else-if ladder

```
#include < stdio.h>
#include <conio.h >
void main ( )
 {
   int num;
   clrscr( );
   printf("enter a number.\n");
   scanf("%d", &num);
   If( num = =0)
     Printf("Given number is Zero.\n");
   else if (number > 0)
     printf("Given number is positive.\n");
   else
     printf("Given number is negative.\n");
   getch ( );
  }
```

Output

```
Enter a number.
5
Given number is positive.
```

Fig 4.8: Program illustrating the else-if ladder

Rules for indentation:

The sections of this page cover the guidelines of acceptable code indentation. Indentation is important for clarity and sticking to standard. The guidelines that are to be followed while using indentation , for control statements are listed below:

1. Indent statements that are dependent on the previous statements; provide at least three spaces of indentation.
2. Align vertically else clause with their matching **if** clause.
3. Use braces on separate lines to identify a block of elements.
4. Indent the statements in the block by at least three spaces to the right of the braces.
5. Align the opening and closing braces.
6. Indent the nested statements as per the above rules.
7. Code only one statement/clause on each line.

4.2.5 The Switch Statement

The switch statement is much like a nested **if** statement and it allows us to make a decision from a number of choices. In fact, it is a powerful decision making statement that allows a variable to be tested for equality against a list of values. The condition of a **switch** statement is a value. The **case** says that if it has the value of whatever is after that **case** then do whatever follows the colon. That is,.each value is called a **case,** and the variable being switched on is checked for each **switch case**. More correctly, a **switch-case default** (since these keywords go together to make up the control statement) accepts single input from the user and based on that input executes a particular block of statements. The break is used to **break** out of the case statements, and is usually surrounded by braces, which it is in. The syntax is:

```
switch   (integer expression)
 {
    case value-1;
                block-1
                break;
   case value-2;
                block-2
                break;
   ……
   …….
   default:
                default-block
                break;
    }
statement-x;
```

The integer expression following the key word **switch i**s any C expression that yields an integer value. It could be an integer constant or an expression that evaluates to an integer. The keyword **case** is followed by an integer or a character constant. Each constant in each **case** must be different from all the others. When the **switch** is executed, the value of the expression is compared against the values value-1,value-2,…When a match is found, the program executes the statements following that case, and all subsequent case and default statements as well .If no match is found, with any of the

case statements, only the statements following the default are executed. Moreover, the **switch** statement transfers control to a statement within its body. Control passes to the statement whose **case** constant-expression matches the value of **switch (expression).** Further, execution of the statement body begins at the selected statement and proceeds until the end of the body or until a break statement transfers control out of the body. A default is optional. When present, it will be executed if the value of the expression does not match any of these **case** values .if not present, no action takes place if all matches fail and the control goes to the statement-x.

The selection process of **switch** statement is explained by the following flow diagram below.

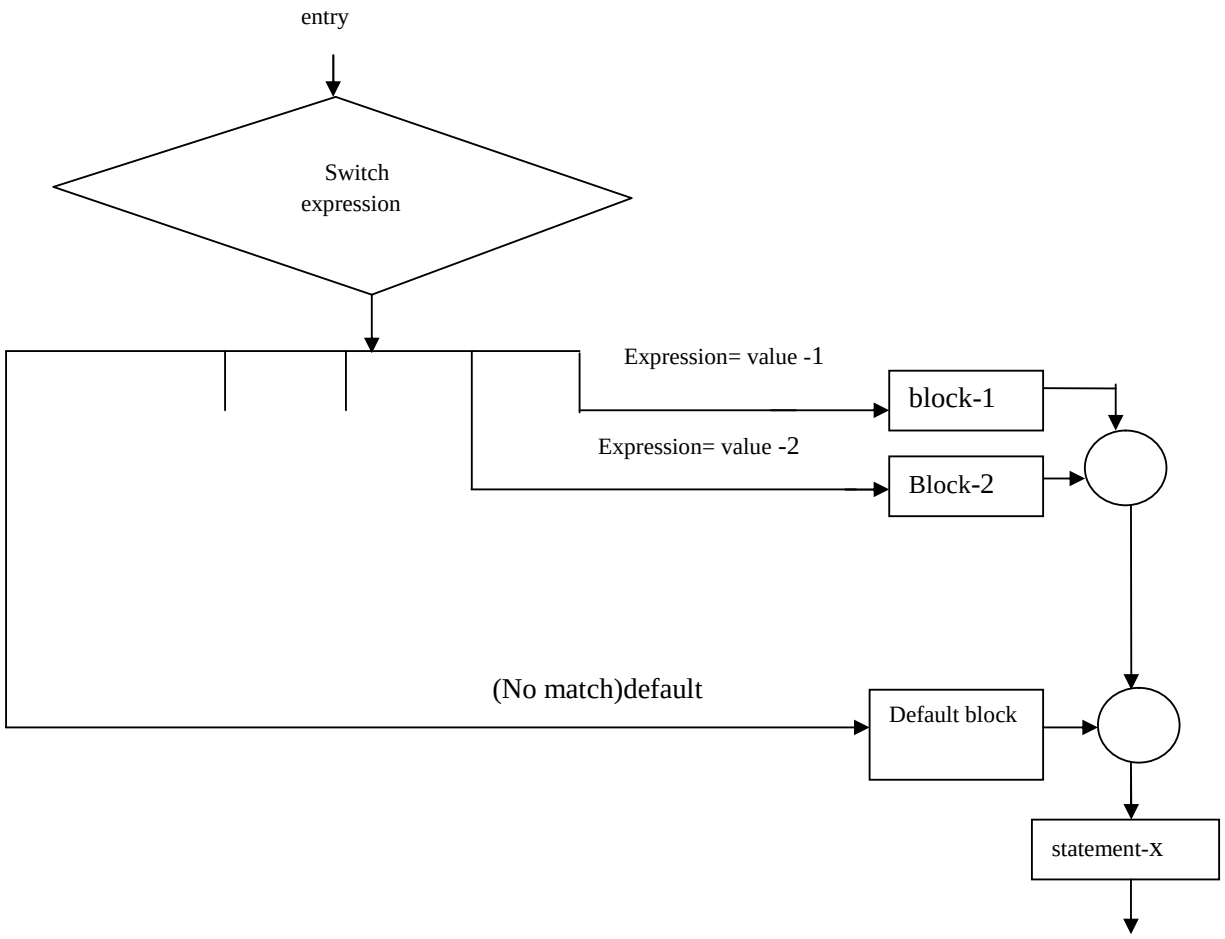


Fig 4.9: Flowchart illustrating switch statement

The following program explain how this control structure works. Here is a program (Fig.4.10)using switch statement:

```
#include <stdio.h>
int main ( )
 {
    char grade = 'B';
    switch (grade)
     {
     case 'A' :
            Printf( "very good!\n" );
            Break;
      case 'B':
      case'C' :
            Printf("good\n");
            Break;
     case 'D':
            Printf("passed\n");
            Break;
    case 'F':
           Printf("pl try again\n");
           Break;
    default :
           Printf("grade invalid\n");
    }
     Printf("grade is %c\n", grade);
     Return 0;
}
```

Fig 4.10: Program using the switch statement

This program on execution gives the following output:

Output

Good

Your grade is B.

Rules for using switch case :

1.The expression used in a **switch** statement must be an integral or enumerated type.

2.With in a **switch** statement one can have any number of **case** statements, with each **case** followed by the

v**alu**e to be compared to and a colon.

3.**case** label must be unique , and must be constants or constant expressions. case labels must end with

semicolon

4.**case** label must of integral type and should not be of floating point type.

5.When the variable being switched on is equal to a **case**, the statements following that **cas**e will execute

until a **break** statement is reached.

6.**switch** case should have at most one **default** label and can be placed anywhere in the **switch,** usually

placed at the end . d**efault** label is optional. No **break** is needed in the **default** case.

7.**break** statements takes control out of the **switc**h (or **switch** terminates and the flow of control jumps to

the next line following **switch** statement) and it is possible to share two or more case statement to

have one **break** statement.

8.Nesting(switch within switch) is permitted for **switch** statement.

9.It is not necessary that every case needs a break statement. If no break appears, the flow of control will

fall through to subsequent cases until a break is reached.

10 relational operators are not allowed in switch case statement .

4.2.6 The ?: Operator

The operator ?: is just like an **if..else** statement except that because it is an operator one can use it within expressions. This is a ternary operator in that it takes three values. The general form of use of this operator is:

conditional expression ? expression 1 : expression 2

Here, the conditional expression is evaluated first and the result if it is non zero, then expression 1 is evaluated and its value is returned as the value of the conditional expression. Otherwise, expression 2 is evaluated and its value is returned. For example the code segment,

```
If (x < 0)
flag = 0;
else
flag = 1;
```

can be written as

```
flag = (x< 0) ? 0 : 1;
```

consider evaluation of yet another function

y = 1.5x+3 for x≤ 2

2x +4 for x >2.

This can be done using the conditional operator ? : as:

```
y = ( x >2)  ? (2*x+4) : (1.5 *x+3);
```

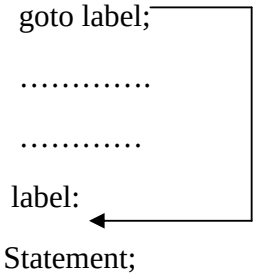
```
# include < stdio.h >

# include < conio.h >

Void main ( )

{

int a,b.c, maxm;

printf(“ program to find maxm value of three numbers:\n”);

printf(“enter the first number:\n”);

scanf(“%d”, &a);

printf(“enter the second number:\n”);

scanf(“%d”, &b);

printf(“enter the third number:\n”);

scanf(“%d”, &c);

max= a>b? (a>c?a: (b > c?b:c )) : (b>c? b:c);

printf(“the maximum  number is %d:”, maxm\n”);

}
```

Fig 4.11: Program illustrating the conditional operator

.On execution of the program, the maximum variable gives the maximum value of the three numbers .

4.2.7 The GOTO statement

In C, GO TO statement is used for altering the normal sequence of program execution by transferring control to some other part of the program. That is ,A **goto** statement provides an unconditional jump from the **go** to a labeled statement in the function. The general form of a go to statement is:



In this syntax label; is an identifier, to identify the place where the branch is to be made.? That is, when the control of program reaches to go to statement, it will jump to the label:, and execute the codes after it. Control may be transferred to anywhere within the current function. The **labe**l is placed immediately before the statement where the control is to be transferred. A **label:** is any valid variable name, followed by a colon and can be any where in the program either before or after the go to label; statement. During program execution when a statement like

go to begin;

Is met, the control flow will jump to the statement immediately following the label begin; This happens unconditionally.

Note that though, using **got**o statement give power to jump to any part of program, using **goto** makes the logic of the program complex and tangled .It breaks the normal sequential execution of the program. If the label: is used before the statement **goto** label; a loop will be formed and some statements will be executed repeatedly. Such a jump is called as a forward jump. On the other hand, if the label: is placed after the **goto** label; some statements will be skipped and the jump is called a forward jump.

A **goto i**s often used at the end of a program to direct the control to go to the input statement, to read further data, in fact, such goto statements puts one to enter in a permanent loop called infinite loop, until one take some special steps to terminate the program. Such infinite loops are to be avoided. Another use of goto is to transfer control out of a loop 9or nested loop) when certain peculiar conditions are encountered. Use of **goto** statement is highly discouraged in any programming language because it makes difficult to trace the control flow of a program, making the program hard to understand and hard to modify. An example to explain the control flow of goto statement is shown in fig. 4.12.

we want to display the numbers from 0 to 9. For this, we have defined the label statement **loop** above the **goto** statement. The given program declares a variable n initialized to 0. The **n++** increments the value of n till the loop reaches 10. Then on declaring the **goto** statement, it will jumps to the label statement and prints the value of n.

```
#include< stdio.h>

#include< conio.h>

int main( )

 {

  int n =0;

  loop: ;

  printf(“ \n%d”,  n);

  n++;

 if(n <10)

  {

 goto loop;

  }

getch( );

return 0

}
```

Fig 4.12: Use of the go to statement

Unit 5

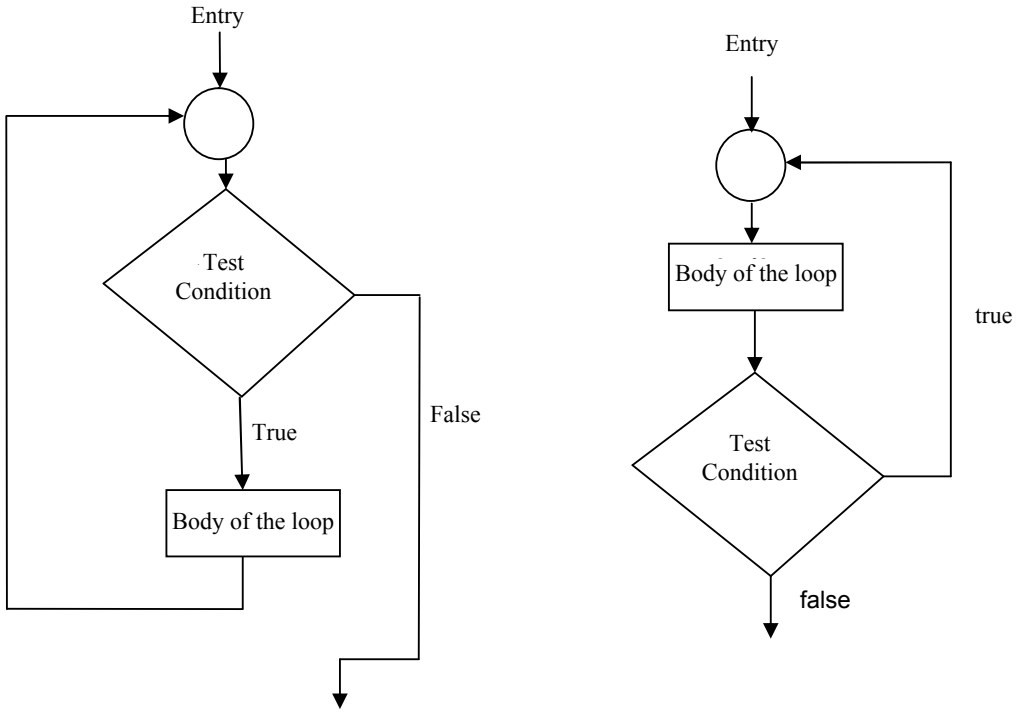
Structure

Introduction
While statement
Do while statement
For statement
Jumps in loops
Continue statement

5.1 Introduction

The multifunctional ability of the computer lies in its adaptability to perform a set of instructions repeatedly. This involves repeating some portion of the program either a specified number of times or until a particular condition is being satisfied. This repetitive operation is done through a loop control instruction. During looping, a set of statements are executed until some conditions for termination of the loop is encountered .A program loop consists of two segments, one is the **body of the loop** and the other known as the **control statement**. The control is tested always for execution of body of the loop.

Depending on the position of the control statement in the loop, a control may be classified as the **entry controlled loop** or as the **exit controlled one** (Fig.5.1). In the **entry** controlled loop, the control condition is tested first and if satisfied then only body of the loop is executed. In the **exit controlled** loop, the test is made at the end of the body, so the body is executed unconditionally first time.



(a) Entry controlled loop **(b) Exit controlled loop**

Fig 5.1: Loop control structure

A looping process, in general, would include the following four steps:

1. Setting and initialization of a counter.
2. Execution of the statement in the loop
3. Test for a specified condition for execution of the loop.
4. Incrementing the counter.

The three loop constructs in C language for performing loop operations are:

1. The ***while*** statement
2. The ***do-while*** statement
3. The ***for*** statement.

Sentinel loops

Based on the nature of control variable, and the type of value assigned to it, for testing the control expression, there are two types of loops:

1. counter controlled

2. sentinel controlled loops (repetition).

Counter controlled repetitions are the loops which the number of repetitions needed for the loop is known before the loop begins; these loops have control variables to count repetitions. Counter controlled repetitions need initialized control variable (loop counter), an increment (or decrement) statement and a condition used to terminate the loop (continuation condition).

Sentinel controlled repetitions are loops with an indefinite repetitions; this type of loop use a special value, called **sentinel** value, to change the loop control expression from true to false(i.e., to indicate end of iteration) .

5.2 The *While* statement.

While statement is a s**entinel c**ontrolled repetition which can be iterated infinite number of times. Number of iterations is controlled using the **sentinel** variable (test expression). It is one of the simplest looping structures. The basic format of the **while** statement is:

```
While (test condition)
  {
      body of the loop
  }
```

The ***while*** is an **entry-controlled** loop statement. The test condition is evaluated and only if the condition is true the body is executed. After execution of the body, the test-condition is once again evaluated and if it is true, the body is executed once again. This process of repeated execution of the body continues until the test-condition finally becomes false and the control is transferred out of the loop. On exit, the program continues with the statement immediately after the body of the loop. If the body contains only one statement it is not necessary to put the braces, but placing them is a good programming practice. Let us look at a simple example, which uses a ***while*** loop.

```
# include< stdio.h>
int main( )
 {
    int p,n,count;
    float r,si;
    count =1;
    while(count <= 4)
     {
            printf (”enter values for p,n,r\n”);
            scanf ( “%d %d %f “, &p,&n,&r”);
            si = p*n*r/100;
            printf(“Simple interest is: Rs. %\n f”, si);
            count = count +1;
       }
       return 0;
  }
```

Fig 5.2: Program to illustrate the while statement

Here, the program executes all the statements after ***while*** 4 times. The logic for calculating the simple interest is written within a pair of braces (i.e., the statements form body of while loop) immediately after the keyword while. The parentheses after the while contain a condition. So long as this condition remains true, all statements within the body of the ***while*** loop keeps getting executed repeatedly. .Also, to start with, the variable **count** is initialized to 1 and every time the logic of simple interest is executed, the value of count is incremented by one .The index variable **coun**t here, is called the loop counter .

The following points about ***while*** are worth noting.

1. The statements within *w**hile*** loop would keep on getting executed till the condition being tested remains true. When the condition becomes false, the control passes to the first statement that follows the body of the ***while*** loop.
2. In the place of condition there can be any other valid expression. So long as the expression evaluates to a non zero value, the statements within the loop would get executed.
3. The condition being tested may be relational or logical operators as in the example below.

   while (i < = 4)

   while (i > = 4 && j < = 5)

   while (i >. = 4 && (j < 5 || c< 10))

4. The statements within the loop may be a single line(i.e., here braces optional) or a block of Statements as in example shown below.

```
while( i < =5)
  i = i+1;
```

is same as,

```
while( i < =5)
{
   i = i+1;
}
```

5. Almost always, the ***while*** must test a condition that will eventually become false, otherwise the loop Will be executed for ever.

6. Instead of incrementing a loop counter (not necessarily integer it can be a float), one can Decrement it and can still manage the body of the loop to be executed repeatedly.

5.3 Do while statement

The ***do while*** loop is also a kind of loop, which is similar to the ***while*** loop, in contrast to while loop, the ***do while*** loop tests at the bottom of the loop after executing the body of the loop. Since the body of the loop is executed first and then the loop condition is checked we can be assured that the body of the loop is executed at

least once. The ***while*** on the other hand, will not execute its statements if the condition fails for the first time. That is, the *w**hile*** tests the condition before executing any of the statements within the ***while*** loop. As against this, the ***do-while*** tests the condition after having executed the statements within the loop. Since the test condition is evaluated at the bottom of the loop, *the* ***do-while*** statement is

```
do
{
        body of the loop
}
while (test condition);
```

an **exit controlled** loop statement. The ***do-while*** loop looks like this: Here the statement is executed first, and next the expression is evaluated. If the condition in the expression is true then the body is executed again and this process continues till the conditional expression becomes false. When the expression becomes false the loop terminates. This difference is brought about more clearly by the following program.

```
#include<stdio.h>
int main ( )
 {
    while ( 4<1)
          printf("hello\n");
    return 0;
 }
```

Here the, since the condition fails the first time itself, the printf ( ) will not get executed at all. The same program using **the *do-while*** construct is

```
#include<stdio.h>
int main ( )
{
   do
    {
         printf("hello\n");
      } while ( 4<1);
      return 0
  }
```

In this program, the printf ( ) would be executed once, since first the body of the loop is executed and then the condition is tested. ***Break*** and ***continue*** are used with ***do while*** just as they would be in a ***while***. A ***break*** takes one out of the ***do-while*** by passing the conditional test. A ***continue*** sends you straight to the test at the end of the loop.

5.4 For statement

The for loop is another entry-controlled loop that provides a more concise loop control structure. It is a counter controlled repetition. Therefore the number of iterations **must** be known before the loop starts (or predetermined). The body of a **for** statement is executed zero or more times until an optional condition becomes false. Also one can use optional expressions with in the **for** statement to initialize and change values during the for statements execution. **The** general form of the **for** loop is:

```
for (initialization; test condition; increment;)
{
      body of the loop
}
```

That is, in the control block of the **for** loop statement there are three expressions separated by semicolon (;).The execution of the for loop is as :

1. **The initialization**: Initialization of the control variables is done first using assignment statements .It is typically used to initialize a loop counter variable.
2. The value of the control variable is tested using the **test condition**. The test condition is a relational expression, such as i <5 that determines when the loop will exit. That is, the loop condition expression is evaluated at the beginning of each iteration. The execution of the loop continues until the loop condition evaluates to false.
3. **Increment**: The increment expression is evaluated at the end of each iteration. It is used to increase or decrease the loop counter variable.

Let us write down the simple interest program(which we have written earlier using **while** statement) using **for (Fig.5.3).** If this program is compared with the one written using **while** construct, we can see that , the three steps of f**o**r loop construct have now been incorporated in the **for** statement. Here in this program (fig 5.3), when the **for** statement is executed for the first time, the value of c**ount** is set to an initial value 1. Next the condition count <=3 is tested. Since the count was set to 1, the condition is satisfied and the body of the loop is executed for the first time. Up On reaching the closing brace of for**, control** is sent back to the **for** statement, where the value of count

is incremented by 1. Again the test is performed to check whether the new value of count exceeds 3. If the value of count is less than or equal to 3, the statements within braces of for are executed again,. The body of the for loop continues to get executed till count does not exceed the final value 3.The control exits from the loop , when count reaches the value 4.and the control is transferred to the statement(if any) immediately after the body of **for.**

```
#include<stdio.h>
int main( )
 {
      int p,n,si;
      float,si;
      for(count =1; count <=3; count= count+1)
       {
            printf(enter the values for p.n,r\n");
            scanf("%d %d %f",&p,&n,&r);
            si = p*n*r/100;
            printf(" simple interest + rs. %f\n", si);
            return 0
       }
 }
```

Fig 5.3 Program using for loop

Additional Features of **for** loop

1. More than one variable can be initialized at a time in the **for** statement as in :

for (p =1, n =6; n <11; $^{++}$n)

Statement. That is, initialization section has two parts p = 1 and n = 6 , separated by comma..Like initialization section, increment section too can have more than one part. The multiple arguments in the increment section too are separated by commas.

2. The test condition may have any compound relation and the testing need not be limited only to the Loop control variable. For eg:

```
sum = 0;
for ( i =1 ;i<10 && sum< 19; ++i )
{
  S = s+1;
 printf("%d %d \n",i,sum):
}
```

Here the loop uses a compound test condition with the counter variable i and variable sum .The loop is executed as long as both the conditions i<10 && sum < 19 are true. The sum is evaluated inside the loop.

3. It is also permissible to use expressions in the assignment statements of initialization and increment Sections. For eg. A statement of the type

```
for( x= (m + n)/2; x > 0; x = x/2)
```

is valid.

4. One or more sections can be omitted if necessary as in eg.,

```
--------------------
m=5;
for (; m ! = 100 ;)
 {
        printf( " %d\n", m);
        m = m+3;
 }
```

Here, both initialization and increment sections are omitted in the **for** statement. The initialization has been done before the **for** statement and the control variable is incremented inside the loop. Though the sections remains blank, the semicolons separating the sections must remain. If the test condition is not present, the **for** statement sets up an infinite loop. Such loops can be broken using **break or goto** statements in the loop..

5.Time delay loops in **for** loop can be set up using the null statement as:

```
for ( i = 100; i > 0; i = i-1)
;
```

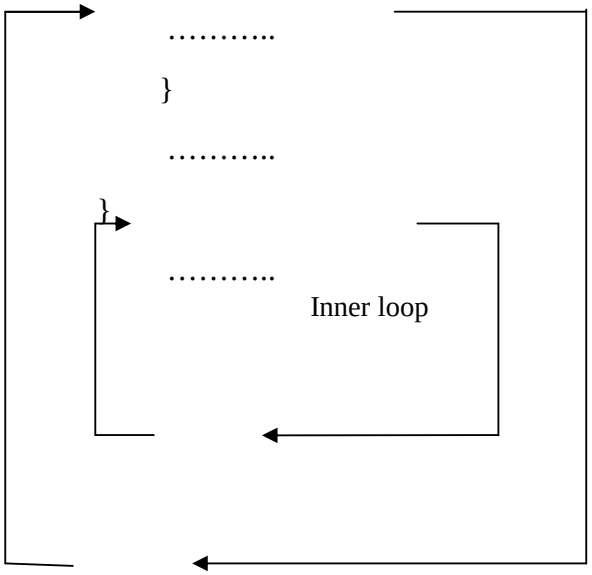
Here this loop is executed 100 times without any output. The body of the loop contains only a semicolon.

Known as null statement.

Nesting of For Loops

The way if statements can be nested, similarly whiles and **for**s can also be nested; two loops can be nested as follows:

```
…….
……..
for ( i =1; i <10; ++ i)
{
      ……
    for ( j= 1; j!  = 5: ++ j)
    {
      ……..
     }
      ……..
}
      ……..
```



The nesting may continue up to any desired level. To understand how nested loops work, we look at the program below.

```
# include< stdio.h>
int main ( )
 {
    int r,c,sum;
    for ( r =1; r < =3; r ++)
     {
          for( c=1; c<=2; c++)
           {
               sum = r+c;
               printf("r= %d sum = %d \n", r,c,sum);
            }
     }
       return 0;
 }
```

output

```
r =1 c=1 sum=2
 r =1 c=2 sum=3
r =2 c=1 sum=3
r =2 c=2 sum=4
r =3 c=1 sum=4
r =3 c=2 sum=5
```

Fig 5.4 Program to illustrate nested for loop

Here for each value of r, the inner loop cycles through twice, with variable c taking values 1and 2.The inner loop terminates when c exceeds 2 and the outer loop terminates when r exceeds 3.

5.5 Jumps in loops

We often come across situations, where we want to jump out of a loop instantly, without waiting to get back to the conditional test. The keyword **break** allows to do this. When **break** is encountered in a loop , control automatically passes to the first statement after the loop. A break is usually associated with an **if.** The key word **break, b**reaks the control only from the **while** in which it is placed. As an example we have :

```
# include < stdio.h>
int main( )
 {
    int num, i;
    printf(" enter a number");
    scanf("%d", & num);
    i =2;
    while( i < = num-1)
     {
       if (num% ! = = 0)
        {
            printf( "not a prime number\n");
            break;
        }
        i++;
     }
     if ( i = = num)
          printf("prime number\n");
  }
```

Fig 5.5: Use of break statement

5.6 The continue statement

The keyword **continu**e, allows us to take the control to the beginning of the loop, by passing the statements inside the loop, which have not yet been executed. That is , when the key word **continue** is encountered inside any loop, control automatically passes to the beginning of the loop .A **continue** is usually associated with an **if.** The **syntax** is:

Continue;

```
#include < stdio.h >

main()
{
    int i;
    int j = 10;

    for( i = 0; i <= j; i ++ )

    {

       if( i == 5 Goods 1

)
       {
          continue; Goods 1

       }
       printf("goods %d\n", i );
    }
}

Output

Goods 1

Goods 2

Goods 3

Goods 4

Goods 5

Goods 6

Goods 7

Goods 8

Goods 9

Goods 10
```

Fig 5.6: Use of continue statement

As an example consider the program of Fig 5.6. The use of **continue s**tatement in loops is illustrated below. In **while** and **do while** loops, **continue**, causes the control to go directly to the test condition and then to continue the iteration process. In the case of **for** loop, , the increment section of the loop is executed before the test condition is evaluated.

```
While (test condition)        do                          for(initialization; test condition; increment)
{                             {                           {
  ……..                        ……                          ……..
  If (……)                     if(……)                      if(……..)
    Continue;                   continue;                   continue;
  ……                          ……..                        ……..
  ……                          ……                          ……..
  }                           } (while test condition );  }
```

Jumping out of loops

We have seen that we can jump out of a loop using either the break or goto statement. In the same way we can jump out of a program by using the library function exit( ).. The use of exit( ) function is shown below.

```
……..
…….
If (test condition) exit (0);
……
……
```

Unit 6

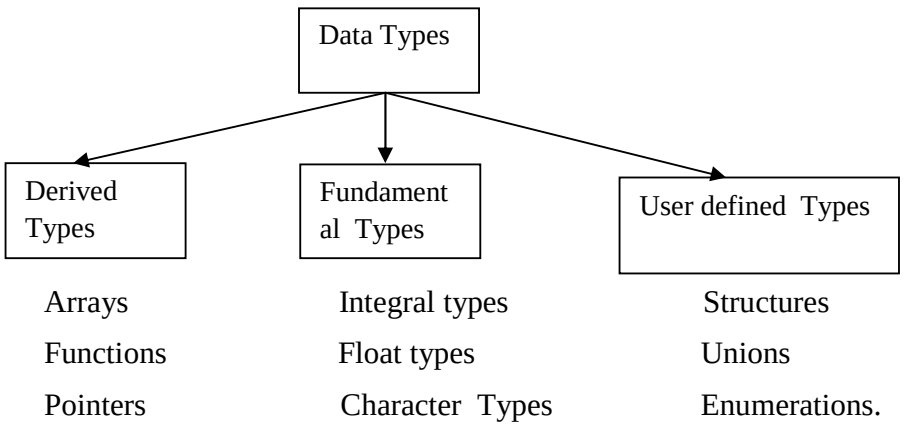
Structure

Introduction
One dimensional arrays
Declaration of one dimensional arrays
Initialization of one dimensional arrays
Two dimensional arrays
Initialization of two dimensional arrays
Multi-dimensional arrays
Dynamic Arrays

6.1 Introduction

An array is a collection of similar elements. These similar elements could be all integers, or all floats, or all characters, etc. Usually, an array of characters is called a 'string', whereas an array of integers or floats is simply called an array. All elements of any given array must be of the same type. That is, we cannot have an array of 10 numbers, of which five are integrs and five are float.

C supports a rich set of derived and user defined data types, in addition to a variety of fundamental data types a detailed in the figure below:



Arrays and structures are referred to as **structured data types** because they can be used to represent data values that have a structure of some sort. Structured data types provide an organizational scheme that shows the relationship among the individual elements and facilitate efficient data manipulation. In programming language such data types are known as **data structures**.

6.2 One dimensional Arrays.

As already discussed**,** an array is a collective name given to a group of similar variables .The values in an array is called as **elements** of array, and are accessed by numbers called **subscripts**. The array which is used to represent and store data in a linear form (or accessing its elements involve only a single subscript) is called as single or **one dimensional array**. As an example consider the C declaration:

```
int number [5];
```

Here in this declaration, the array variable **number** contain 5 elements of any value available to the i**nt** type .and the computer reserves 5 storage locations. The values to the array elements can be assigned as:

```
number [0]= 12;
number [1]=13;
number [2]=15;
number [3]=20;
number [4]=25;
```

This would cause the array number to store the values as shown below:

number [0]	12
number [1]	13
number [2]	15
number [3]	20
number [4]	25

These elements may be used in programs just like any C variable

6.3 Declaration of one dimensional Arrays

To begin with, like other variables an array needs to be declared before they are used so that the compiler will know what kind of an array and how large an array we want. The general form of array declaration is:

type variable-name [size];

The **type** specifies the type of element that will be contained in the array, such as int, float or char and the **siz**e indicates the maximum number of elements that can be stored inside the array .For example,

int marks [10];

Declares the marks as an array to contain a maximum of 10 integer constants. This number is often called the **dimension** of the array .The bracket ([ ]) tells the compiler that we are dealing with an array.

The C treats character strings simply as array of characters. The size in a character string represents the maximum number of characters that the string can hold. For instance,

char name[13];

Declares the name as a character array(string) variable that can hold a maximum of 13 characters. Suppose we read the following string constant in to the string variable name

"GOOD MORNING"

In this, each character of the string is treated as an element of the array name and is stored in the memory as:

'G'
'O'
'O'
'D'
' '
'M'
'O'
'R'
'N'
'I'
'N'
'G'
'\o'

When the compiler sees a character string , it terminates with an additional null character \o. Thus the element name[13] holds the null character '\o'. Remember that, while declaring character arrays, we must allow one extra space for the null terminator.

6.4 Initialization of one dimensional Array.

After an array is declared, its elements must be initialized. If they are not given any specific value, they are supposed to contain garbage values. An array can be initialized at either of the following stages:

- at compile time
- at run time

Compile time initialization

Whenever we declare an array we can initialize it directly at compile time. In this type of initialization, we assign certain set of values to array elements before executing program The general form of initialization of arrays is:

type array-name[size] = [list of values];

the values in the list are separated by commas. The type size can be specified directly as :

int num [5] = { 2.3,4,5,6};

Here the size of the array is specified directly as 5 in the initialization statement. The compiler will assign values to the particular elements of the array. i.e., At the time of compilation all, the elements are at specified positions as shown below.

num [0] = 2

num [1] = 3

num [2] = 4

num [3] = 5

num [4] = 6

Also the type size can be specified indirectly as in:

int num [ ] = { 2.3,4,5,6};

The compiler counts the number of elements written with in the braces and determines the size of the array.

Character arrays may be initialized in the same manner. Thus the statement

char name [ ] = { 'j', 'o', 'h', 'n', '\o'};

Declares the name to be an array of five characters, initialized with the string 'john' ending with the null character. Alternatively, we can assign the string literal directly as :

char name [ ] = 'john';

Run time initialization

An array can also be explicitly initialized at run time usually; .this approach is applied for initialization of large arrays. For example, consider the following program segment;

```
for (i  = 0;  i < 5; i++)
{
 scanf ( "% d "'  & x [ i ] );
}
```

The above segment will initialize the array elements with the values entered through the keyword .In this type of initialization (run time initialization) of the arrays. looping elements are almost compulsory. Looping statements are used to initialize the values of the arrays one by one by using assignment operator or through the keyboard by the user. we can also use read function such as **scanf** to initialize an array as in example below.

```
int x [2] ;
```

```
# include < stdio.h >
void  main  ( )
 {
    int array [3], i;
    printf( " enter 3 numbers to store them in an array\n" );
    for ( i =0; i < 3; i ++)
     {
        scanf ( " % d ", & array [ i] ) ;
     }
     printf ( " elements  in the array are: \ n");
     for  i =0; i < 3; i ++)
      {
        printf (" elements stored at a [ %d] = %d\n",i, array [ i]);
      }
      getch ( );
}
```

output

```
enter  3 elements in the array : 2 3 4
elememts in the array are :
element stored at a[ 0] = 2
element stored at a[ 1] = 3
element stored at a[ 2] = 4
```

Fig 6.1: Program using an array

scanf (" %d % d", & x[0], & x[1]);

will initialize the array elements with the values entered through the key word. Here is a sample program (Fig.6.1) to store the elements in the array and to print them from this array.

Searching and sorting are two operations performed on arrays. Searching is the process of arranging elements in the list according to their values, in ascending or descending order. An ordered list is a sorted one. The three simple and important sorting methods are:

Bubble sort

Selection sort

Insertion sort.

Other sorting methods include, Merge sort, quick sort and Shell sort.

Searching is the process of finding the location of the specified element in a list. The specified element is often called the **search key**. If the process of searching finds a match of the search key with a list element value, then the search is sad to be successful. Otherwise it is unsuccessful. Two most commonly used searching methods are ;

Sequential search

Binary Search.

6.5 Two dimensional Arrays.

So far, we have explored arrays with only one dimension. It is also possible to have two or more dimensions. The 2-D array is also called a matrix. The 2-D arrays are declared as :

type array-name [size of row] [column size];

2-D arrays are stored in memory as shown below. In memory, whether, it is single or two dimensional array, the array elements are stored in one continuous chain .Each dimension of the array is indexed from zero to its maximum size minus one: the first index selects the row and the second index selects the column within that row,

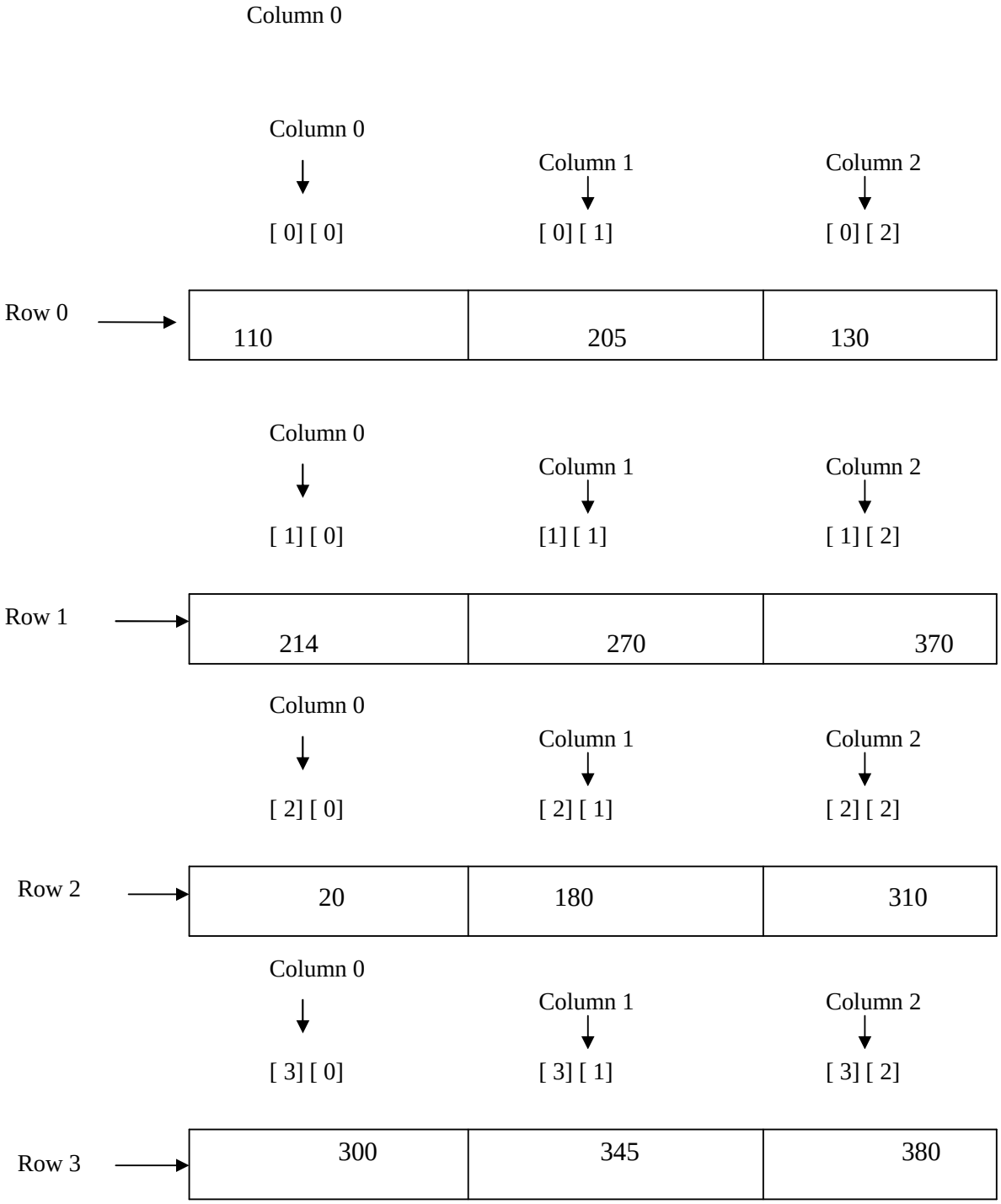


Fig 6.2: Representation of 2D array in memory

Here is a sample program:

```
# include< stdio.h>
int main( )
{
    int students [4] [2];
    int i,j;
    for (i = 0; i < = 3; i ++)
     {
          printf ( “enter the roll no of student and marks\n”);
          scanf(“ %d %d”,  &student [i] [0], &student[i][1]);
       }
    for (i =0; i< = 3; i ++)
           printf( “ %d %d “, student [i] [0],student [i] [1]);
    return 0;
  }
```

Fig 6.3: Program to illustrate 2D array

This program stores the roll number and marks obtained by a student side by side in a matrix. In the first part of the program, i.e., in the first f**or** loop, we read in the values of roll number and marks, where as in the second **for** loop, we print out these values. Also, in the first **scanf** , the first subscript of the variable student is row number which changes for every student. The second subscript tells which of the two columns are we talking about- the zeroth column which contains the roll number or the first column which contains the mark. The counting of rows and columns begins with zero. Remember that two dimensional array is a collection of a number of one dimensional arrays placed one below the other .In this program, the array elements have been stored row wise and accessed row wise. Although it is possible to access the elements column wise, row-wise strategy is accepted widely.

6.6 Initializing 2-D arrays

Like 1-D arrays, 2-D arrays could be initialized by following their declaration with a list of initial values enclosed in braces as in ,

```
int table [2][3] = { 0,0,0,1,1,1};
```

which initializes the first row to zero and second row to one. Equivalently one can write the above statement as:

```
int table [2][3] = {{ 0,0,0} ,{ 1,1,1}};
```

We can also initialize a 2-D array in matrix form as:

```
int table [2][3] = {
                    {0,0,0},
                    {1,1,1}
                   };
```

More over, the declaration

```
int table [ ][3] = {
                    { 0,0,0},
                    {1,1,1}
                   };
```

Is perfectly valid.

If the values are missing in the initatializer, they are automatically set to zero. For instance, the statement

```
int table [2][3] = {
                    {1,1}
                    {2}
                   };
```

will initialize the first two elements of the first row to one, the first element of the second row to 2 and all other elements to zero.

In situations where we have to initialize all the elements to zero, a short cut method as in,

```
int m [3] [5] = { { 0}, { 0},{0} };
```

may be used. Here the first element of each row is explicitly initialized to zero, while all other elements are automatically initialized to zero. the following statement would also work.

int m [3] [5] =- { 0,0};

6.7 Multi dimensional Arrays

The general form of a multidimensional Array is:

Type array-name [s1] [s2] [s3] …….[sm] ;

Where si is the size of the ith dimension. A 3-D array can be thought of as an array of arrays of array. The outer array has three elements, each of which is 2-d array of four 1-D arrays., each of which contains two integers. That is, a 1-D array of two elements is constructed first, followed by placing four 1-D arrays placed one below the other. So that a 2-d array containing four rows is obtained. Thereafter, three 2-D arrays are placed one behind the other to yield a 3-D array containing three 2-D arrays.

6.8 Dynamic Arrays

In C it is possible to allocate memory to arrays at run time. The arrays created at run time are called dynamic arrays .Dynamic arrays are created using memory management functions like malloc, calloc, realloc, that are included in the header file< stdlib.h > The concept of dynamic arrays is used in creating and manipulating data structures like lists, stack and queues.