POST GRADUATE DEGREE PROGRAMME (CBCS) IN

MATHEMATICS

SEMESTER IV

SELF LEARNING MATERIAL

PAPER : MATC 4.1
(Applied and Pure Streams)

Discrete Mathematics
Probability and Statistical Methods

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

Course Preparation Team

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

May, 2020

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, 2017 had been our endeavour. We are happy to have achieved our goal.

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

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

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

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

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

Director

Directorate of Open and Distance Learning

University of Kalyani

Core Paper

MATC 4.1
Marks : 100 (SEE : 80; IA : 20)

Discrete Mathematics(Marks : 60 (SEE: 50; IA: 10))
Probability and Statistical Methods (Marks : 40 (SEE: 30; IA: 10))
(Applied and Pure Streams)

Syllabus

- **Unit 1:** Definition of graphs, circuits, cycles, Subgraphs, induced subgraphs, degree of a vertex, Connectivity.
- **Unit 2:** Trees, Euler's formula for connected graphs, Spanning trees, Complete and complete bipartite graphs.
- **Unit 3:** Planar graphs and their properties, Fundamental cut set and cycles. Matrix representation of graphs, Kuratowski's theorem (statement only) and its use, Chromatic index, chromatic numbers and stability numbers.
- **Unit 4:** Lattices as partial ordered sets. Their properties, Lattices as algebraic system. Sublattices. Direct products and Homomorphism. Some special Lattices e.g. complete complemented and distributed lattices.
- **Unit 5:** Boolean Algebra: Basic Definitions, Duality, Basic theorems, Boolean algebra as lattices.
- **Unit 6:** Boolean Algebra: Boolean functions, Sum and Product of Boolean algebra, Minimal Boolean Expressions, Prime implicants Propositions and Truth tables.
- **Unit 7:** Boolean Algebra: Logic gates and circuits, Applications of Boolean Algebra to Switching theory (using AND, OR, & NOT gates), Karnaugh Map method.
- **Unit 8:** Combinatorics : Introduction, Basic counting principles, Permutation and combination, pigeonhole principle, Recurrence relations and generating functions.
- **Unit 9:** Grammar and Language : Introduction, Alphabets, Words, Free semi group, Languages, Regular expression and regular languages. Finite Automata (FA). Grammars.
- **Unit 10:** Finite State Machine. Non-deterministic and deterministic FA. Push Down Automation (PDA). Equivalence of PDAs and Context Free Languages (CFLs), Computable Functions.
- **Unit 11:** Fields and σ-fields of events. Probability as a measure. Random variables. Probability distribution.

- **Unit 12:** Expectation. Moments. Moment inequalities, Characteristic function. Convergence of sequence of random variables-weak convergence, strong convergence and convergence in distribution, continuity theorem for characteristic functions. Weak and strong law of large numbers. Central Limit Theorem.
- **Unit 13:** Definition and classification of stochastic processes. Markov chains with finite and countable state space, classification of states.
- **Unit 14:** Statistical Inference, Estimation of Parameters, Minimum Variance Unbiased Estimator, Method of Maximum Likelihood for Estimation of a parameter.
- **Unit 15:** Interval estimation, Method for finding confidence intervals, Statistical hypothesis, Level of significance; Power of the test.
- **Unit 16:** Analysis of variance, One factor experiments, Linear mathematical model for ANOVA.

Contents

Unit 1

Course Structure

- Definition of graphs, circuits, cycles
- Subgraphs, induced subgraphs, degree of a vertex
- Connectivity

1.1 Introduction

In the time of Euler, in the town of Königsberg in Prussia, there was a river containing two islands. The islands were connected to the banks of the river by seven bridges (1.1.1). The bridges were very beautiful, and on their days off, townspeople would spend time walking over the bridges (see figure below). As time passed, a question arose: was it possible to plan a walk so that you cross each bridge once and only once? This is known as the Knigsberg seven bridge problem. In the year 1736, Euler represented the problem as a graph and answered the question in negative. This marked the birth of graph theory.

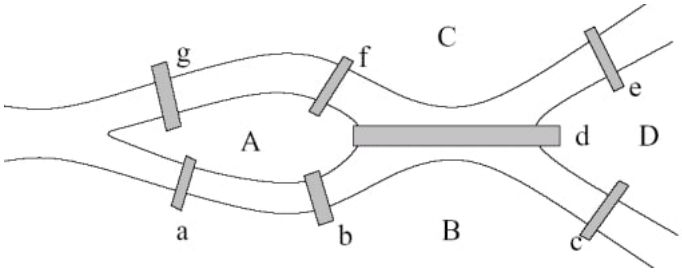


Figure 1.1.1: Königsberg Seven Bridge Problem

Since then it has blossomed in to a powerful tool used in nearly every branch of science and is currently an active area of mathematics research. Over the past 200 years, graph theory has been used in a variety of applications. Graphs are used to model electric circuits, chemical compounds, highway maps, and many more. They are also used in the analysis of electrical circuits, finding the shortest route, project planning, linguistics, genetics and social science.

Objectives

After reading this unit, you will be able to

- define graphs, vertex, edges
- define circuits, cycles and learn their properties
- define subgraphs
- define connected and disconnected graphs and learn their properties
- define planar graphs and trees
- deduce the Euler's formula for connected graphs

1.2 Graphs

Definition 1.2.1. A graph G is a triple (V, E, g), where

1. V is a finite non-empty set, called the set of vertices;
2. E is a finite set (may be empty), called the set of edges;
3. g is a function, called the incidence function, that assigns to each edge, $e \in E$ a one element subset $\{v\}$, or a two-element subset $\{u, v\}$, where u, v are vertices.

For convenience, we will write $g(e) = \{u, v\}$, where v and u may be same in which case we write $g(e) = \{v\}$.

Let $G = (V, E, g)$ be a graph. Suppose e be an edge of this graph. Then there are vertices u and v such that $g(e) = \{u, v\}$; the vertices u and v are called the end vertices or the endpoints of the vertex e. When a vertex v is an endpoint of an edge e, we say that e is incident with vertex v and v is incident with the edge e. Two vertices are said to be adjacent if there exists an edge $e \in E$ such that $g(e) = \{u, v\}$. Two edges e and f are said to be adjacent if they have a common endpoint, that is, if $g(e) = \{u, v\}$ and $g(f) = \{v, w\}$. If e is an edge such that $g(e) = \{u, v\}$ such that $u = v$, that is, $g(e) = \{u\}$, then e is an edge from u to itself, or u is adjacent to itself and such an edge e is called a loop on the vertex u.

From now on, we will simply write the graph $G = (V, E, g)$ as G.

Example 1.2.2. Let $V = \{a, b, c, d\}$ and $E = \{e, f, h, i, j\}$ and g is defined as

$$g(e) = \{a, b\},\ \ g(f) = \{b, c\},\ \ g(h) = \{c, d\},\ \ g(i) = \{d, a\},\ \ g(j) = \{d, b\}.$$

Thus, $G = (V, E, g)$ is a graph. We can also write the above definition of g as follows:

e	f	h	i	j	k
$\{a, b\}$	$\{b, c\}$	$\{c, d\}$	$\{d, a\}$	$\{d, b\}$	$\{d, b\}$

Such a representation of the incidence function g is the incidence table whose columns are indexed by the edges. The vertices adjacent to an edge are placed in the second row below the edge.

In this example, we see that the edge e is incident on the two vertices a and b. Thus, the vertex a and b are adjacent. Similarly, we see that the edges e and f are adjacent since the vertex b is common for both.

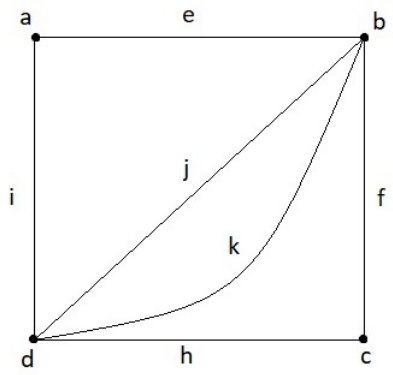


Figure 1.2.1: G as in example 1.2.2

The set of vertices and the set of edges of a graph are finite. Thus, one of the features that make the study of graphs easy and interesting is that they can be represented pictorially. That is, the corresponding diagram for a graph helps us to visualize the facts easily. If we represent the graph in the above example pictorially, then we get something as depicted in the figure 1.2.1.

The incidence function need not be one-to-one. There may be more than one edge having the same end-points. Such edges are called parallel edges. We formally define parallel edges as follows.

Definition 1.2.3. Let $G = (V, E, g)$ be a graph. Two edges e and f are said to be parallel if $g(e) = g(f) = \{u, v\}$ for $u, v \in V$.

In the previous example, the edges j and k are parallel edges since $g(j) = g(k) = \{d, b\}$. This can easily be seen from the figure.

Definition 1.2.4. Let G be a graph and v be a vertex in G. We call v as isolated vertex if it is not incident with any edge, or, v is not an endpoint of any edge.

Definition 1.2.5. Let G be a graph and v be a vertex in G. Then the degree of v is defined as the number of edges incident with v. It is written as $\deg(v)$ or $d(v)$. By convention, it is considered that each loop contributes 2 to the degree of a vertex.

Note that for an isolated vertex v, we will always have $d(v) = 0$. In fact, this is a necessary and sufficient condition for a vertex to be isolated.

Example 1.2.6. $G = (V, E, g)$ is a graph (see figure 1.2.2 where $V = \{A, B, C, D\}$ and $E = \{e, f, h, i, j\}$, where g is defined as

e	f	h	i	j
$\{A, B\}$	$\{B, C\}$	$\{C, B\}$	$\{B, A\}$	$\{A, A\}$

Then, we can see that D is an isolated vertex. Also, $d(A) = 4$, $d(B) = 4$, $d(C) = 2$ and $d(D) = 0$. e and i are parallel edges and f and h are also so. Notice that $g(A) = g(i)$ irrespective of the order in which A and B are written in the incidence table of g. But it is not the case always (as we will study in case of the directed graphs). These graphs that we are studying now are also called undirected graphs (or simply, graphs).

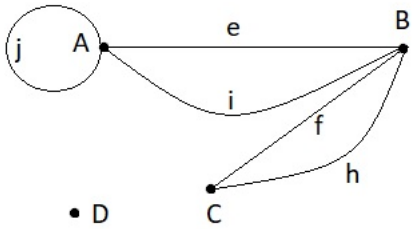


Figure 1.2.2: G in example 1.2.6

Exercise 1.2.7. Represent the following graphs pictorially and find the degree of each of its vertices. Also state the parallel vertices and loops, if any.

1. $V = \{v_1, v_2, \ldots, v_7\}$ and $E = \{e_1, e_2, \ldots, e_7\}$ where g is defined as

e_1	e_2	e_3	e_4	e_5	e_6	e_7
$\{v_1, v_2\}$	$\{v_1, v_2\}$	$\{v_4, v_3\}$	$\{v_6, v_3\}$	$\{v_2, v_4\}$	$\{v_6, v_3\}$	$\{v_6, v_3\}$

2. $V = \{v_1, v_2, v_3, v_4\}$ and $E = \{e_1, e_2, e_3\}$ where g is defined as

e_1	e_2	e_3
$\{v_1, v_2\}$	$\{v_3, v_3\}$	$\{v_4, v_3\}$

The graphs in which all the vertices are of the same degree are called the regular graphs. The two examples of the graphs we have seen so far are not regular graphs (verify it for example 1.2.2).

Definition 1.2.8. Let G be a graph and k be a non-negative integer. Then G is called a k-regular graph if the degree of each vertex of G is k.

An interesting k regular graph is the Petersen 3-regular graph as shown in the figure.

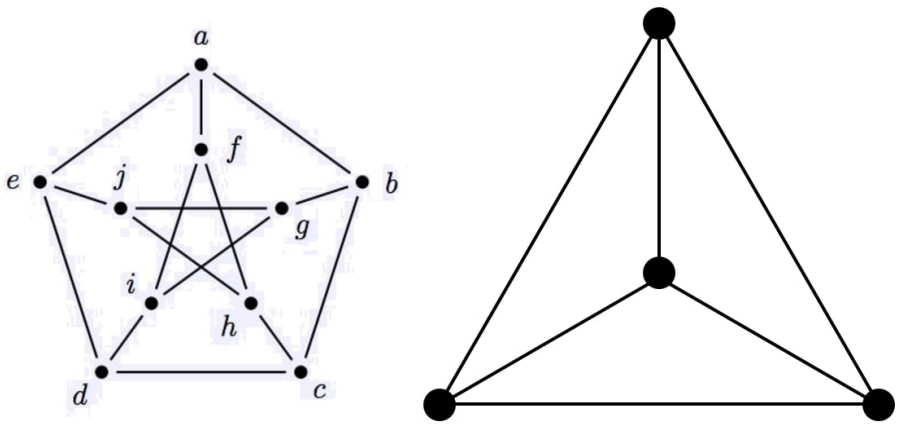


Figure 1.2.3: 3-regular graphs (Petersen 3 regular graph on the left)

Definition 1.2.9. Let G be a graph and v be a vertex of G. v is called an even degree vertex if $d(v)$ is an even number. Similarly, v is odd degree vertex if $d(v)$ is odd.

The Petersen 3-regular graph has every vertex an odd vertex.

Definition 1.2.10. Let $n_1, n_2, \ldots, n_k$ be the degrees of vertices of a graph G such that $n_1 \leq n_2 \leq \cdots \leq n_k$. Then the finite sequence $n_1, n_2, \ldots, n_k$ is called the degree sequence of the graph.

Clearly, every graph has a unique degree sequence. But, we can construct completely different graphs having the same degree sequence.

Exercise 1.2.11.
1. State the even and odd vertices of the graphs in exercise 1.2.7. Also find the degree sequence of them.
2. Construct a 1-regular graph having 3 vertices.
3. Construct two different graphs having the same degree sequence.

Consider the degree sequence of the graph in 1.2.2 which is, 2, 2, 4, 4 and adding them gives $(2+2+4+4 =)12$, which is an even number. In fact, the sum of the degrees of all the vertices is always an even number which is given in the following theorem due to Euler.

Theorem 1.2.12. The sum of the degrees of all the vertices of a graph is twice the number of edges.

Proof. Let G be a graph with n edges and m vertices, say $v_1, v_2, \ldots, v_m$. We want to determine

$$d(v_1) + d(v_2) + \cdots + d(v_m).$$

Now the degree, $d(v_i)$, of v_i is the number of edges incident with v_i. Each edge e is either a loop or incident with two distinct vertices. If e is a loop on a vertex v, then e contributes 2 to the degree of v. On the other hand, if e is incident with two distinct vertices v and w, then e contributes 1 to the degree of each vertex. Thus we find that when we compute the sum of the degrees, each edge contributes 2 to the sum. Because there are n edges, the total contribution to the above sum is $2n$. Hence

$$d(v_1) + d(v_2) + \cdots + d(v_m) = 2n.$$

□

Corollary 1.2.13. The sum of the degrees of all the vertices of a graph is an even integer.

Proof. Since $2n$ is an even integer, the corollary follows from the previous theorem. □

Corollary 1.2.14. In a graph, the number of odd degree vertices is even.

Proof. Suppose a graph G has k odd vertices, $v_1, v_2, \ldots, v_k$, and t even degree vertices, $u_1, u_2, \ldots, u_t$. Thus, by the above corollary,

$$d(v_1) + d(v_2) + \cdots + d(v_k) + d(u_1) + d(u_2) + \cdots + d(u_t) = 2n,$$

where n is the number of edges. Because each $d(u_j)$ is even, it follows that $d(u_1) + d(u_2) + \cdots + d(u_t)$ is an even integer. Also, $2n$ is even. Hence, $d(v_1) + d(v_2) + \cdots + d(v_k)$ must also be even. Now, the sum of odd number of odd integers is an odd integer. Because each number $d(v_i)$ is an odd number and $d(v_1) + d(v_2) + \cdots + d(v_k)$ is even, it follows that the number k cannot be odd. So k is even and this completes the proof. □

1.3 Directed Graphs

Definition 1.3.1. A directed graph, or digraph G is a triple (V, E, g) such that

1. V is a finite non-empty set of vertices;

2. E is a finite set (may be empty) of directed edges or arcs;

3. $g : E \to V \times V$ is a function, that assigns to each edge, $e \in E$ an ordered pair (u, v), where u, v are vertices (u and v may be same).

We can represent a digraph pictorially. The only difference between the representation of graph and digraph is in the directed edges which are drawn with arrows representing the starting and terminating vertices.

If $g(e) = (u, v)$, then u is called the starting vertex and v is called the terminating vertex of the arc e. The in-degree of a vertex v is the number of arcs with v as the terminating vertex and the out-degree of v is the number of arcs with v as the starting vertex. In computing in-degree and out-degree of a vertex, we assume that each loop contributes 1 to the in-degree and 1 to the out-degree of v.

Theorem 1.3.2. In any digraph $G = (V, E, g)$, the following three numbers are equal:

1. The sum of the in-degrees of all the vertices;

2. The sum of the out-degrees of all the vertices;

3. The number of arcs.

Proof. The proof is similar to that of theorem 1.2.12. We just consider the fact that each arc e with starting vertex u and terminating vertex u contributes 1 to the out-degree and 1 to the in-vertex of v. The details are left as exercise. □

Example 1.3.3. Let G be a digraph such that $V = \{a, b, c, d\}$, $E = \{e, f, h\}$ and $g(e) = (a, a)$, $g(f) = (b, c)$ and $g(h) = (b, d)$. The diagram is as follows:

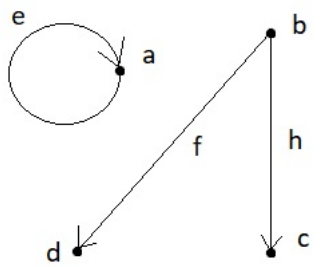


Figure 1.3.1: G in example 1.3.3

The in-degrees of a, b, c and d are 1, 0, 1 and 1 respectively and the out-degrees are 1, 2, 0 and 0 respectively. Then, sum of the in-degrees of all the vertices=sum of the out-degrees of all the vertices=number of arcs=3.

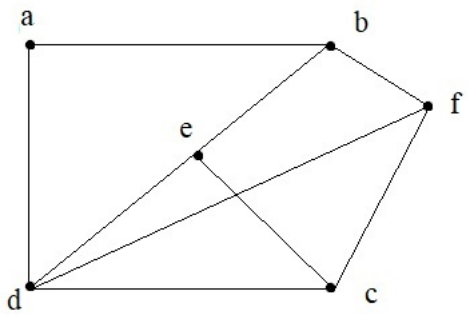


Figure 1.4.1: A simple graph

1.4 Simple Graphs

Definition 1.4.1. Let G be a graph. Then it is called a simple graph if it does not contain any parallel edges or loop.

The graph in the figure 1.4.1 has no loop or parallel edges.

Theorem 1.4.2. Let G be a simple graph with at least two vertices. Then G has at least two vertices of same degree.

Proof. Let G be a simple graph with $n \geq 2$ vertices. G has no loops or parallel edges. Thus, the degree of a vertex v is the same as the number of vertices adjacent to it. The graph G has n vertices. Thus, a vertex v has at most $n-1$ adjacent vertices, because v is not adjacent to itself. Hence, for any vertex v, the degree of v is one of integers: $0, 1, 2, \ldots, n-1$.

We now show that if there exists a vertex v such that $d(v) = 0$, then for each vertex u of G, $d(u) < n-1$. On the contrary, suppose that in G, v is a vertex with degree 0 and u is a vertex with degree $n-1$. Then v is an isolated vertex and u has $n-1$ adjacent vertices. Because G is a simple graph, u is not adjacent to itself. From this and the fact that G is simple and $d(u) = n-1$, it follows that every vertex of G other than u is adjacent to u. This implies that v is adjacent to u, which is a contradiction since v is an isolated vertex. This proves our claim.

In a similar manner, we can prove that if there exists a vertex v in G such that the degree of v is $n-1$, then for each vertex u in G, $d(u) > 0$.

We now conclude that the degree of all the vertices of G are either in the set $\{0, 1, 2, \ldots, n-2\}$ or in the set $\{1, 2, \ldots, n-1\}$.

Let $v_1, v_2, \ldots, v_n$ be the n vertices of G. Then, either for all of $i = 1, 2, \ldots, n$, $d(v_i) \in \{0, 1, 2, \ldots, n-2\}$ or $d(v_i) \in \{1, 2, \ldots, n-1\}$. Thus, by the pigeonhole principle, there exists i and j, $1 \leq i \leq n$ $1 \leq j \leq n$, $i \neq j$, such that $d(v_i) = d(v_j)$. Hence there are atleast two vertices of same degree. □

Remark 1.4.3. The converse of the above theorem is not true in general. For example, a and c have equal degree in example 1.2.2, but the graph G is not simple.

Definition 1.4.4. A simple graph with n vertices in which there is an edge between every pair of distinct vertices is called a complete graph on n vertices. This is denoted by K_n.

A complete graph of three vertices is a triangle.

Theorem 1.4.5. The number of edges in a complete graph with n vertices is $\frac{n(n-1)}{2}$.

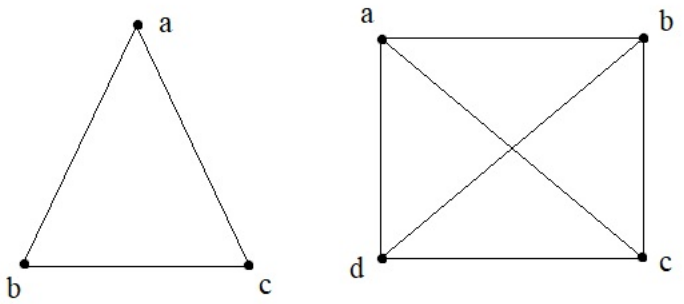


Figure 1.4.2: K_3 and K_4

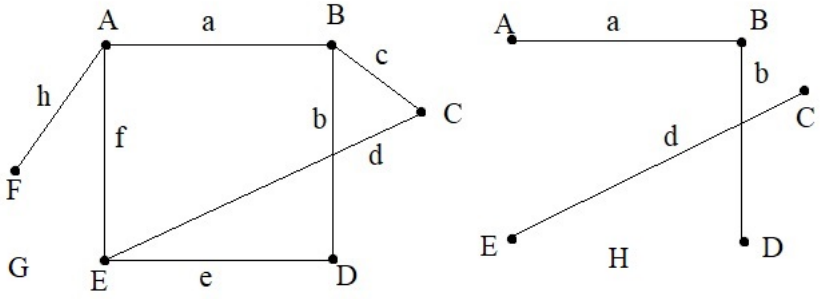
Proof. Let G be a complete graph with n vertices. Then G is a simple graph such that there exists an edge between any two distinct vertices. Hence, for any vertex v of G, each of the remaining $n-1$ vertices is adjacent to v. Hence the degree of each vertex is $n-1$. Also, since G has n vertices, so the sum of the degree of all the vertices is $n(n-1)$. We know that the sum of the degree of all the vertices is 2 times the number of edges. Let the number of edges be m. So, we have, $n(n-1) = 2m$ and thus, we get,

$$m = \frac{n(n-1)}{2}.$$

□

1.5 Subgraph

Consider the graph $G = (V, E, g)$ and $H = (V_1, E_1, g_1)$ such that $V = \{A, B, C, D, E, F\}$, $E = \{a, b, c, d, e, f, h\}$ and $V_1 = \{A, B, C, D, E\}$, $E_1 = \{a, b, d\}$ and g and g_1 are as shown in the figure. It is worthy to note



that $V_1 \subset V$ and $E_1 \subset E$. Also, g_1 is the function g restricted over E_1. Such a graph H is called a subgraph of G. We will now formally define subgraphs.

Definition 1.5.1. Let $G = (V, E, g)$ be a graphs. A graph $H = (V_1, E_1, g_1)$ is called a subgraph of G if V_1 is a non-empty subset of V and E_1 is a subset of E and g_1 is a restriction of g on E_1 such that for all $e \in E_1$, we have $g_1(e) = g(e) = \{u, v\}$ for $u, v \in V_1$.

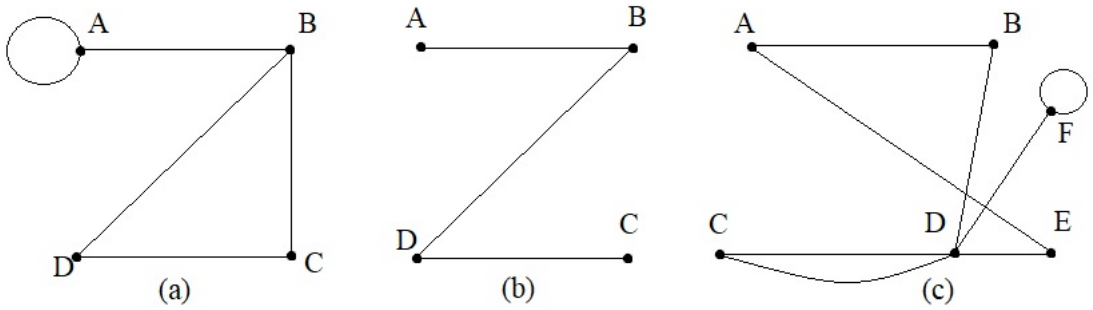
Remark 1.5.2. Let $G = (V, E)$ be a graph and $H = (V_1, E_1)$ be a subgraph of G. From the previous definition, it follows that if $e \in E_1$, and u, v are the end vertices of e in G, then $u, v \in V_1$.

Let G be a graph with vertex set V and edge set E. Suppose that V contains more than one vertex. Then for any vertex $v \in V$, $G \setminus \{v\}$ denotes the subgraph whose vertex set is $V_1 = V \setminus \{v\}$ and the edge set is $E_1 = \{e \in E |\ v \text{ is not an end vertex of } e\}$. Then $G \setminus \{v\}$ is called a subgraph obtained from G by deleting the vertex v.

Let $e \in E$, and $G \setminus \{e\}$ denote the subgraph whose edge set is $E \setminus \{e\}$ and the vertex set is $V_1 = V$. Then $G \setminus \{e\}$ is the subgraph obtained by deleting the edge e.

Remark 1.5.3. $G \setminus \{v\}$ is obtained by deleting the vertex v and at the same time deleting all the edges incident with v. However, the graph $G \setminus \{e\}$ is obtained from G by deleting only the edge e without deleting any of the vertices of G.

Exercise 1.5.4. 1. Determine which of the following graphs are simple:



2. Draw a graph having the following properties and explain why no such graph exists:
   (a) Simple graph, five vertices, each of degree 2
   (b) Simple graph having degree sequence $3, 3, 3, 3, 4$
   (c) Six edges and having the degree sequence $1, 2, 3, 4, 6$
3. Find three subgraphs of G in the figure 1.5.1 with at least four vertices and six edges:
4. How many vertices are there in a graph with 20 edges if each vertex is of degree 5?
5. Does there exist a simple graph with degree sequence $1, 2, 3, 4, 5$? Justify.
6. Does there exist a graph with five edges and degree sequence $1, 2, 3, 4$?

1.6 Walks, Path, Cycles, Circuits

Definition 1.6.1. Let u and v be two vertices in a graph G. A walk from u to v in G, is an alternating sequence of $n + 1$ vertices and n edges of G

$$(u = v_1, e_1, v_2, e_2, \ldots, v_{n-1}, e_{n-1}, v_n, e_n, v_{n+1} = v)$$

beginning with vertex u, called the initial vertex, and ending with vertex v, called the terminal vertex, in which v_i and v_{i+1} are endpoints of edge e_i, for $i = 1, 2, \ldots, n$.

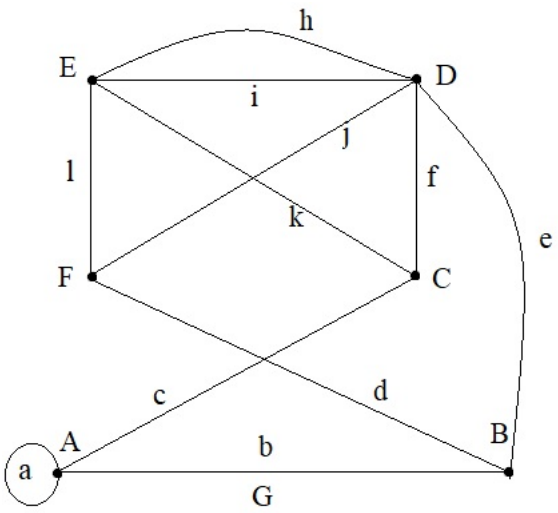


Figure 1.5.1

Definition 1.6.2. Let u and v be two vertices in a digraph G. A directed walk from u to v in G, is an alternating sequence of $n+1$ vertices and n arcs of G

$$(u = v_1, e_1, v_2, e_2, \ldots, v_{n-1}, e_{n-1}, v_n, e_n, v_{n+1} = v)$$

beginning with vertex u and ending with vertex v, in which each e_i is an arc from v_i to v_{i+1} for $i = 1, 2, \ldots, n$.

Definition 1.6.3. The length of a walk(or a directed walk) is the total number of occurrences of edges(or, arcs) in the walk(or, directed walk). A walk or directed walk of length zero is only a vertex.

A walk (or, directed walk) from a vertex u to v in a graph (or, digraph) G is also called a $u - v$ walk (or, directed walk). If u and v are the same, then $u - v$ walk (or, directed walk) is called a closed walk (or, directed walk). Otherwise, it is called an open walk (or, directed walk).

Definition 1.6.4. A walk with no repeated edges is called a trail, and a walk with no repeated vertices except possibly the initial and terminal vertices is called a path.

Thus, from the previous definitions, it is clear that in a path, no edge can be repeated. Hence, every path is a trail, but not conversely.

Definition 1.6.5. A walk, path, or trail is called trivial if it has only one vertex and no edges. A walk, path, or trail that is not trivial is called nontrivial.

Definition 1.6.6. A nontrivial closed trail from a vertex u to itself is called a circuit.

Hence, a circuit is a closed walk of nonzero length from a vertex u to itself with no repeated edges.

Example 1.6.7. Consider the graph in figure 1.6.1. In this graph

$$(A, a, B, b, C, f, E, e, B, d, D)$$

is a walk of length 5. It is an open walk from A to D. This is a walk with no repeated edges. Hence this walk is a trail since B appears twice. But

$$(B, b, C, f, E, i, D, j, G)$$

is a path of length 4 from vertex B to G.

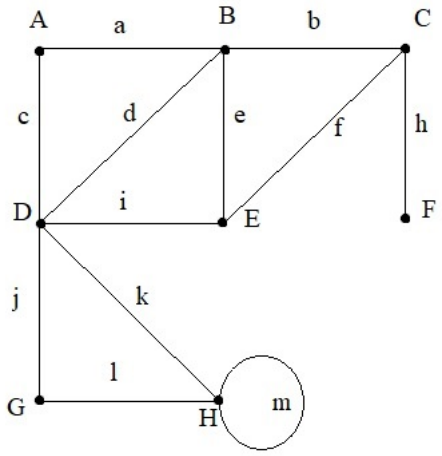


Figure 1.6.1

Definition 1.6.8. A circuit that does not contain any repetition of vertices except the starting and terminal vertices is called a cycle. A cycle of length k is called a k-cycle. A cycle is called even (odd) if it contains an even (odd) number of edges.

It follows from definition that a 3-cycle is a triangle.
Directed walks, trails, paths, circuits, cycles are defined analogously.

Definition 1.6.9. Let $P = (v_1, e_1, v_2, e_2, \ldots, v_{n-1}, e_{n-1}, v_n)$ be a walk in a graph G. A subwalk of P is a subsequence of **consecutive** entries $Q = (v_i, e_i, v_{i+1}, e_{i+1}, \ldots, v_{k-1}, e_{k-1}, v_k)$, $1 \leq i \leq k \leq n$, that begins at a vertex and ends at a vertex.

From the definition, it follows that every subwalk is a walk.
Let $P = (v_1, e_1, v_2, e_2, \ldots, v_{n-1}, e_{n-1}, v_n)$ be a walk in a graph G and $Q = (v_i, e_i, v_{i+1}, e_{i+1}, \ldots, v_{k-1}, e_{k-1}, v_k = v_i)$ be a closed subwalk of P. If we delete this subwalk Q from P except for the vertex v_i, then we obtain a new walk. This walk is denoted by $P - Q$ and is called the reduction of P by Q.

Theorem 1.6.10. Let G be a graph and u, v be two vertices of G. If there is a walk from u to v, there is a path from u to v.

Proof. Let $P = (u = v_1, e_1, v_2, e_2, \ldots, v_{n-1}, e_{n-1}, v_n = v)$ be a walk. If $u = v$, then this is a closed walk. In this case, (u) from u to u consisting of a single vertex and no edge. Suppose $P = (u = v_1, e_1, v_2, e_2, \ldots, v_{n-1}, e_{n-1}, v_n = v)$ is an open walk. If this is not a path, then $v_i = v_j$ for some $1 \leq i < j \leq n$. This shows that there is a closed subwalk Q from v_i to v_j. We reduce P to $P - Q$. Now, $P - Q$ is a new walk from u to v. If this walk is not a path, we repeat this deletion process of subwalks. Because the number of closed subwalks in P is finite, we eventually obtain a path from u to v. □

We can also follow the proof of the above theorem and deduce an analogous result for circuit.

Theorem 1.6.11. Every circuit contains a subwalk that is a cycle.

Proof. Let T be a circuit. Let S be the collection of all closed nontrivial subwalks of T. Because $T \in S$, S is nonempty. Now S is a finite set. Thus we can find a member of S of minimum length. Let T_1 be a nontrivial closed subwalk $(u = v_1, e_1, v_2, e_2, \ldots, v_{n-1}, e_{n-1}, v_n = u)$ of T of minimum length. Since T_1 is of minimum length, T_1 cannot contain a nontrivial closed subwalk other than T_1. This implies that T_1 has no repeated vertices except the vertex u. Hence T_1 is a cycle. □

Definition 1.6.12. Let G be a graph. A vertex u is said to be connected to a vertex v of G if there is a $u-v$ walk in graph G. And G is said to be connected if for any two vertices u and v of G, there is a $u-v$ walk in G, otherwise G is called a disconnected graph.

We can show that a graph G is connected if and only if for any two vertices $u, v \in G$, there is a $u-v$ path in G. We assume that a graph with only a single vertex and no edges is connected.

We now define a relation R on the vertex set V of a graph G as

$$R = \{(u, v) \in V \times V : \text{ there is a } u - v \text{ walk in } G\}.$$

Since the trivial walk (u) is a $u-u$ walk in G, R is reflexive. Suppose there is a $u-v$ walk $(u = v_1, e_1, v_2, e_2, \ldots, v_{n-1}, e_{n-1}, v_n = v)$. Then, $(v = v_n, e_{n-1}, v_{n-1}, \ldots, v_2, e_1, v_1 = u)$ is a $v-u$ walk in G. Thus, R is symmetric. Again, suppose there is a $u-v$ walk

$$(u = v_1, e_1, v_2, e_2, \ldots, v_{n-1}, e_{n-1}, v_n = v)$$

from a vertex u to v. Also, suppose there is a $v-w$ walk $(v = u_1, f_1, u_2, f_2, \ldots, v_m = w)$ from vertex v to another vertex w. Then clearly,

$$(u = v_1, e_1, v_2, e_2, \ldots, v_{n-1}, e_{n-1}, v_n = v = u_1, f_1, u_2, f_2, \ldots, v_m = w)$$

is a walk from vertex u to vertex w. Thus, the relation R is transitive. Hence R partitions the vertex set V into disjoint equivalence classes. Let V_1 be an equivalence class of R and E_1 be the set of edges joining the vertices in V_1 in the graph G. Then $G_1 = (V_1, E_1)$ is a subgraph of G. In this subgraph, we see that any two vertices are connected. This subgraph is called a component of G.

Definition 1.6.13. A subgraph H of a graph G is called a component of G if

1. any two vertices of H are connected in H, and
2. H is not properly contained in any connected subgraph of G.

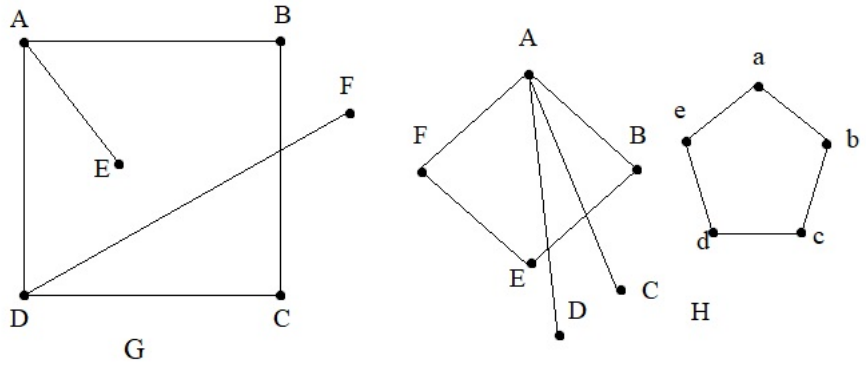


Figure 1.6.2: Connected and Disconnected Graphs

Graph G in the above figure has only one component, which is G itself. Graph H, on the other hand, has two components with vertices $\{A, B, C, D, E, F\}$ and $\{a, b, c, d, e\}$.

From the definition, it follows that any component of a graph is always connected. Now, every equivalence class of the equivalence relation R gives a component of G. Hence, every graph can be partitioned into finite number of components. It follows that a graph G is connected if and only if G has only one component.

Theorem 1.6.14. A connected graph with n vertices has at least $n-1$ edges.

Proof. We prove the result by induction on n. If $n = 1$, then the result is trivially true. Assume that any connected graph with n vertices has at least $n-1$ edges. Consider a connected graph G with $n+1$ vertices. Because G is a connected graph, the degree of each vertex of G is ≥ 1. Suppose the degree of each vertex of G is ≥ 2. Then the sum of the degree of the vertices of G is $\geq 2(n+1) > 2n$. Thus, the number of edges of G is $> n$. Suppose now that G has a vertex v of degree 1. We construct a graph G_1 by deleting the vertex v and the edge incident with v. The graph G_1 is a connected graph with n vertices. By the induction hypothesis, the number of edges of G_1 is at least $n-1$. Therefore, the number of edges of G is at least n. Thus, the result is true for a graph with $n+1$ vertices. Hence, by induction for any connected graph with n vertices, the number of edges is at least $n-1$. □

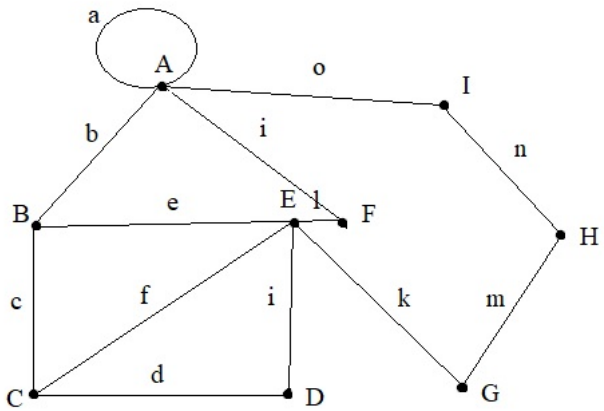
We prove another interesting theorem for connected graphs.

Theorem 1.6.15. Let G be a simple graph with at most $2n$ vertices. If the degree of each vertex is at least n, then the graph is connected.

Proof. Suppose that G is not connected. Then G can be partitioned into components $C_1, C_2, \ldots, C_m$, $m \geq 2$. Since the degree of each vertex of G is at least n and the graph is simple, we find that each vertex has at least n adjacent vertices. Then each component contains at least $n+1$ vertices. This implies that the number of vertices of G is at least $m(n+1) \geq 2(n+1) > 2n$. This contradiction implies that the given graph is connected. □

1.7 Few Probable Questions

1. Consider the graph below:



   (a) Find an open walk of length 4. Is is a trail? Is your walk a path?

   (b) Find a closed walk of length 5. Is your walk a circuit?

2. Does there exist a graph with 20 edges if each vertex is of degree 3?

3. Draw s simple graph such that every vertex is adjacent to two vertices and every edge is adjacent to two edges.

4. Define a path of a graph G. If G has exactly two vertices of odd degree, then show that there exists a path between these two vertices.

5. Define simple graph. If there is a trail from a vertex u to another vertex v of a graph G, then show that there is a path from u to v.

6. Define connected graph. Show that a simple graph with n vertices and m components can have at most $\frac{(n-m)(n-m+1)}{2}$ edges.

7. Let G be a connected graph with at least two vertices. If the number of edges in G is less than the number of vertices, then prove that G has a vertex of degree 1.

Unit 2

Course Structure

- Trees, Euler's formula for connected graphs, Spanning trees
- Complete and complete bipartite graphs

2.1 Introduction

In the previous unit, we learnt about the basic definitions of graph theory and certain properties related to them. This unit is a continuation of the previous unit.

Objectives

After reading this unit, you will be able to

- define complete graphs, bipartite graphs, and complete bipartite graphs
- define trees and spanning trees
- learn various properties of connected graphs due to Euler

2.2 Bipartite graphs

Definition 2.2.1. A simple graph G is called a bipartite graph if the vertex set V of G can be partitioned into nonempty subsets V_1 and V_2 such that each edge of G is incident with one vertex in V_1 and one vertex in V_2. $V_1 \cup V_2$ is called a bipartition of G.

In the figure 2.2.1, the graph in (a) is a bipartite graph with partition $\{A\}$ and $\{B, C, D\}$. Whereas, the second graph is not bipartite as we can easily verify. (Verify!)

Definition 2.2.2. A bipartite graph G with bipartition $V_1 \cup V_2$ is called a complete bipartite graph on m and n vertices if the subsets V_1 and V_2 contain m and n vertices, respectively, such that there is an edge between each pair of vertices $v_1 \in V_1$ and $v_2 \in V_2$. A complete bipartite graph with m and n vertices is denoted by $K_{m,n}$.

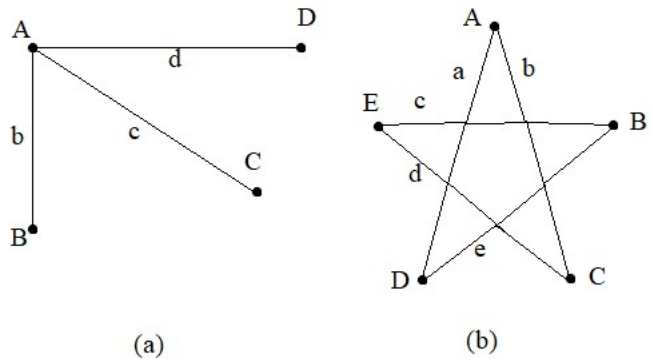
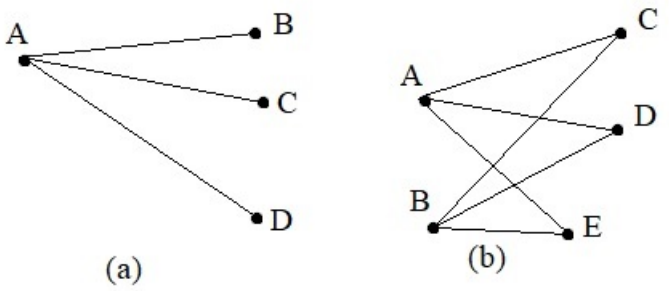


Figure 2.2.1



The two graphs in the above figure represents two complete bipartite graphs. (a) is $K_{1,3}$ while (b) is $K_{2,3}$. Note that the number of edges in the graph $K_{m,n}$ is mn.

Definition 2.2.3. Let G be a graph. Then the distance between two vertices u, v of G, written as $d(u, v)$, is the length of a shortest path, if any exists, from u to v.

If G is a connected graph, then we can prove that

1. $d(u, v) \geq 0$, and equality holds if and only if $u = v$;
2. $d(u, v) = d(v, u)$;
3. $d(u, v) + d(v, w) \geq d(u, w)$, for all vertices $u, v, w \in G$.

We will now deduce a necessary and sufficient condition for a graph to be bipartite.

Theorem 2.2.4. A graph is bipartite if and only if it does not contain any cycle of odd length.

Proof. Let $G = (V, E)$ be a bipartite graph with bipartition $V = V_1 \cup V_2$. Now, each edge of G is incident with one vertex in V_1 and one vertex in V_2. Let $(v_1, e_1, v_2, e_2, \ldots, v_k, e_k, v_1)$ be a cycle in G. Because v_i and v_{i+1} are end vertices of e_i, for $i = 1, 2, \ldots, k$ (assuming $v_{k+1} = v_1$), it follows that for $i = 1, 2, \ldots, k$, if $v_i \in V_1$, then $v_{i+1} \in V_2$. Suppose $v_1 \in V_1$. This implies that $v_k \in V_2$. Also it follows that $v_i \in V_1$ if and only if i is odd. Now $v_k \in V_2$, which implies that k is even and hence the length of this cycle is even. This implies that the length of each cycle is even.

Conversely, let G be a graph such that G has no odd cycle. Suppose G is partitioned into components $C_1, C_2, \ldots, C_m$, $m \geq 1$. If we can show that each C_i is a bipartite graph, then G will be also so. We therefore assume that G is connected. Let u be an arbitrary but fixed vector of G. Define the subsets V_1 and V_2 by

$$V_1 = \{v \in V |\ d(u, v) \text{ is even}\}\ \ V_2 = \{w \in V |\ d(u, w) \text{ is odd}\}.$$

From our assumption that G is a connected graph, it follows that every vertex of G is either in V_1 or in V_2. Then $\{V_1, V_2\}$ is a partition of V. Because $d(u, u) = 0$, it follows that $u \in V_1$. Let v be an adjacent vertex of u. Then $d(u, v) = 1$. Hence, $v \in V_2$.

Suppose there are two distinct vertices v and w in V_1 and suppose there exists an edge e with v, w as end vertices. Then there is a walk from u to v in G and hence there is a shortest path, say P_1, from u to v. Similarly, we have a shortest path P_2, from u to w. Because v and w belong to V_1, these two shortest paths are of even length. Paths P_1 and P_2 may have several vertices and edges in common.

Now starting from u, let x be the last vertex common to both P_1 and P_2. Let P_1^* be the section of the path of P_1 from u to x and let P_2^* be the section of the path of P_2 from u to x. Because P_1 and P_2 are the shortest paths, P_1^* and P_2^* have equal lengths, which are either both even or both odd. Let $P_1^{'}$ be the part of P_1 from x to v and $P_2^{'}$ be the part of P_2 from x to w. It follows that the lengths of $P_1^{'}$ and $P_2^{'}$ are both either even or odd. Now the walk $P_1^{'}$ followed by e followed by $P_2^{'}$ forms a closed walk C from x to x. Moreover, C does not contain any repetitions of the vertices. Hence C is a cycle. Because the lengths of paths $P_1^{'}$ and $P_2^{'}$ are both even or odd, C must be of odd length, which is a contradiction. Thus, v and w cannot be both in V_1. Similarly, we can show that v and w cannot both belong to V_2. Hence each edge of G connects one vertex of V_1 with one vertex of V_2. Consequently, G is bipartite. □

Exercise 2.2.5. 1. Draw a complete bipartite graph on 3 and 4 vertices.

2. How many edges are there in each of the following graphs

$$(a)\ K_{2,3} \qquad (b)\ K_{4,3} \qquad (c)\ K_{4,4} \qquad (d)\ K_{n,n}$$

3. Prove that a simple graph with a cycle of length 3 can't be a bipartite graph.

2.3 Special Circuits

2.3.1 Euler Circuits

Let us consider a connected graph with more than one vertex such that every vertex has odd degree. For example consider the graph in the figure 2.3.1. It is a connected graph whose every vertex is of odd degree. This graph has no circuit, so it has no circuit that contains all the edges. Also the graph K_4 contains 4 vertices and 6 edges. The degree of each vertex is 3. And this graph also has no circuit consisting of all the edges. But there are circuits consisting of all the edges for some graphs which are of special interest. Let us write the following

Definition 2.3.1. A circuit in a graph that includes all the edges of the graph is called an **Euler Circuit**. And a graph G is called **Eulerian** if either G is trivial graph or G has an Euler circuit.

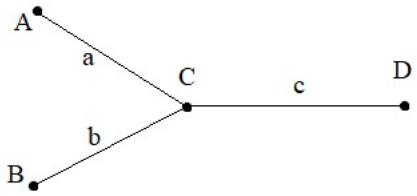


Figure 2.3.1: G with odd vertices

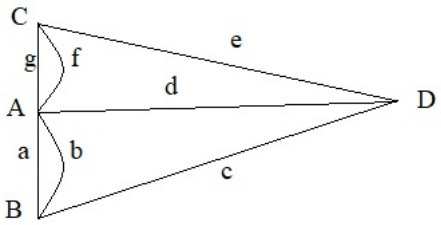
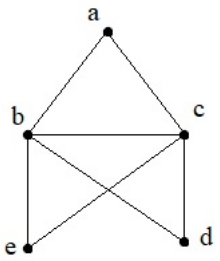


Figure 2.3.2: K*ö*nigsberg Bridge problem

Recall the K*ö*nigsberg bridge problem at the beginning of unit 1. The problem was to determine whether it is possible to take a walk that crosses each bridge exactly once before returning to the starting point. Euler converted this into a problem of graph theory as follows : Each of the islands A, B, C and D are considered as the vertices of a graph and the seven bridges a the seven vertices of the graph. Now the problem reduces to finding a circuit in the graph such that it contains all the edges, or, to find an Euler circuit, or to show that the graph is Eulerian. It is evident from the figure that there does not exist any Euler circuits of the graph.

Example 2.3.2. Consider the graph below.



Each vertex of the above graph are even vertices. In fact, this is a feature of Eulerian graphs as we will soon show.

Theorem 2.3.3. If a connected graph G is Eulerian, then every vertex of G has even degree.

Proof. Suppose that G is Eulerian.

First suppose that G is the trivial graph. Then G has only one vertex v and no edges. Hence the degree of v is 0 which is even.

Next suppose that G contains more than one vertex. SInce G is Eulerian, it has an Euler circuit, say

$$C : (v_1, e_1, v_2, e_2, v_3, \ldots, e_{n-1}, v_n = v_1)$$

from a vertex v_1 to $v_n = v_1$. Now, C contains all the vertices (since G is connected) and all the edges of G. However, there are no repeated edges in C, though in C a vertex may appear more than once. Let u be a vertex of G. Since G is connected, u is not an isolated vertex. So u is the end vertex of some edge. Since C contains all the edges, it follows that u is a member of C.

Suppose u is v_1. Let us say that this is the first appearance of u in C. Now, if u is also v_n, we say that v_n is the last appearance of u in C. For each of these two appearances of u, the edge e_1 and the edge e_{n-1} together contribute 2 to the degree of u.

Suppose now u is v_i in C for some i, $1 < i < n$. Then u is an end vertex of the edges e_{i-1} and e_i. These edges together contribute 2 to the degree of u. It now follows that the degree of any vertex in C is even. Hence the degree of any vertex in G is even. □

Suppose G is connected in which every vertex is of even degree. We shall show that G contains an Euler circuit. To do so, we first prove the following lemma.

Lemma 2.3.4. Let G be a connected graph with one or two vertices. If every vertex of G is of even degree, then G has an Euler circuit.

Proof. Suppose G is a graph with only one vertex, say u. Now there may exist zero or more loops at u. However, the number of loops at u must be finite. If there is no loop at u, then (u) is an Euler circuit of G. Also suppose that there are loops $e_1, e_2, \ldots, e_n$, $n \geq 1$, at u. Then $(u, e_1, u, e_2, \ldots, e_n, u)$ is an Euler circuit of G. Hence, G contains an Euler circuit.

Suppose now that G has two vertices u and v such that both are of even degree. Because G is connected, u and v are connected. So there exists an even number of parallel edges between u and v. Let $\{f_1, f_2, \ldots, f_{2k}\}$, $k \geq 1$ be the set of all edges between u and v. Let $e_1, e_2, \ldots, e_n$, $n \geq 0$, be the loops at u and let $g_1, g_2, \ldots, g_m$, $m \geq 0$, be the loops at v. (If $n = 0$, then there are no loops at u. Similarly, if $m = 0$, there are no loops at v). Now,

$$(u, e_1, u, e_2, \ldots, u, e_n, u, f_1, v, g_1, v, g_2, v, \ldots, g_m, v, f_2, u, f_3, v, , f_4, \ldots, f_{2k-1}, v, f_{2k}, u)$$

is a trail that begins at u, traverses all the loops incident with u, traverses one edge from u to v, traverses all the loops at v, then traverses one edge from v to u, and then traverses all the edges between u and v. This trail does not contain any repeated edges. Hence, it is a circuit from u to u. Because this circuit contains all the edges of G, it follows that the graph G has an Euler circuit. □

Theorem 2.3.5. Let G be a connected graph such that every vertex of G is of even degree. Then G has an Euler circuit.

Proof. Suppose G has n edges. We prove by induction on the number of edges of G to show that G has an Euler circuit.

Basic Step: Suppose $n = 0$. Because G has no edges, it follows that G has a single vertex, say u. Then (u) is an Euler circuit.

Inductive hypothesis: Let n be a positive integer. Assume that any connected graph with k edges, $0 \leq k < n$, in which every vertex has even degree has an Euler circuit.

Inductive step: Let $G = (V, E)$ be a connected graph with n edges and the degree of each vertex of G is even. If the number of vertices of G is 1 or 2, then by previous lemma, it follows that G has an Euler circuit. So assume that G has at least three vertices.

Since G is connected, there are vertices v_1, v_2, v_3 and edges e_1, e_2 such that v_1, v_2 are the end vertices of e_1, and v_2, v_3 are the end vertices of e_2. Now consider the subgraph $G_1 = (V_1, E_1)$, where $V_1 = V$ and $E_1 = E - \{e_1, e_2\}$. Next we add a new edge e with v_1, v_3 as end vertices to the subgraph and obtain a new graph $G_2 = (V_2, E_2)$, where $V_2 = V$, $E_2 = E_1 \cup \{e\}$.

Notice that the graph G_2 is obtained from G by deleting edges e_1, e_2, but not removing any vertices, and adding a new edge e with end vertices v_1 and v_3.

In G, suppose $\deg(v_1) = r$, $\deg(v_2) = m$, and $\deg(v_3) = t$. Because we deleted edges e_1, e_2 in G, $\deg(v_1) = r - 1$, $\deg(v_2) = m - 2$, and $\deg(v_3) = t - 1$. Now in graph G_2, we add a new edge e with end vertices v_1 and v_3. Hence, in graph G_2, we have $\deg(v_1) = r$, $\deg(v_2) = m - 2$, $\deg(v_3) = t$. While constructing G_1 from G and G_2 from G_1, the other vertices of G were not disturbed; i.e., their degree in G_2 is the same as their degree in G. Thus, it follows that every vertex of G_2 is of even degree.

Now graph G_2 may not be a connected graph. We show that the number of components of G_2 is less than or equal to two.

Since v_1 and v_3 are the end vertices of the edge e in G_2, it follows that v_1 and v_3 belong to the same component of G_2, say C_1. Now, vertex v_2 may not be in C_1. Let C_2 be the component of G_2 that contains v_2. Let v be a vertex of G_2. Then v is also a vertex of G. Since G is a connected graph, there is a path P from v to v_1 in G.

If P contains one of the edges e_1 or e_2, then P cannot be a path from v to v_1 in G_2. Let P_1 be the path in G_2 that is a portion of the path P starting at v whose edges are also in G_2. Path P_1 may terminate at v_1, v_2, or v_3. If P_1 is a path from v to v_1 in G_1, then v and v_1 belong to the same component, C_1. If P_1 ends at v_3, then (P_1, e, v_1) is a path from v to v_1. Hence in this case, v also belongs to the same component, C_1. Suppose P_1 ends at v_2. Then v belongs to component C_2. Thus, any vertex v of G_2 belongs to either C_1 or C_2. Hence, C_2 has one(if $C_1 = C_2$) or two components.

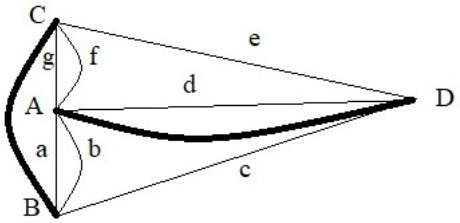
Suppose G_2 has only one component, C_1. Then G_2 is a connected graph with $n - 1$ edges. Thus, by the inductive hypothesis G_2 has an Euler circuit, say T_1. From circuit T_1, we can construct an Euler circuit T in G by simply replacing the subpath (v_1, e, v_3) by the path $(v_1, e_1, v_2, e_2, v_3)$. Hence in this case, we find that G is Eulerian.

Suppose G_2 has two components, C_1 and C_2. Now, each component C_i, $i = 1, 2$ is a connected graph such that each vertex has even degree and the number of edges in C_i is $n_i < n$. Hence, by the inductive hypothesis, C_i has an Euler circuit T_i, $i = 1, 2$. Now, T_1 contains v_1, v_3 and T_2 contains v_2. Hence (v_1, e, v_3) is a subpath of T_1. Moreover, we can assume that T_2 is a circuit from v_2 to v_2.

We now construct an Euler circuit in G by modifying T_1 as follows: In T_1, replace (v_1, e, v_3) by (v_1, e_2, v_2), followed by T_2, followed by (v_2, e_2, v_3). Thus, we find that G has an Euler circuit. The result now follows by induction. □

The above theorem is an effective way of determining when a connected graph is Eulerian.

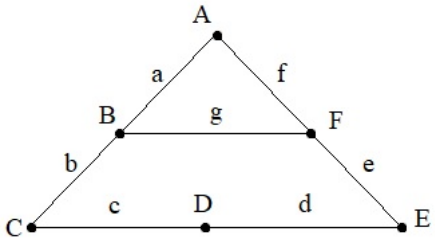
Example 2.3.6. Consider the Königsberg bridge problem. All the vertices in the graph are of odd degree. Then by the preceding two theorems, we can say that there does not exist an Euler circuit for the problem.



But if we add two more edges as shown in the figure, then the resulting graph is Eulerian since every vertex is of even degree.

Definition 2.3.7. An open trail in a graph is called an Euler trail if it contains all the edges.

Example 2.3.8. Consider the following graph. It is a connected graph having two vertices of odd degree. So



it does not have an Euler circuit but the trail $(B, g, F, e, E, d, D, c, C, b, B, a, A, f, F)$ contains all the edges of G. Hence this is an Euler trail.

Theorem 2.3.9. A connected graph G has an open Euler trail if and only if G has only two vertices of odd degree.

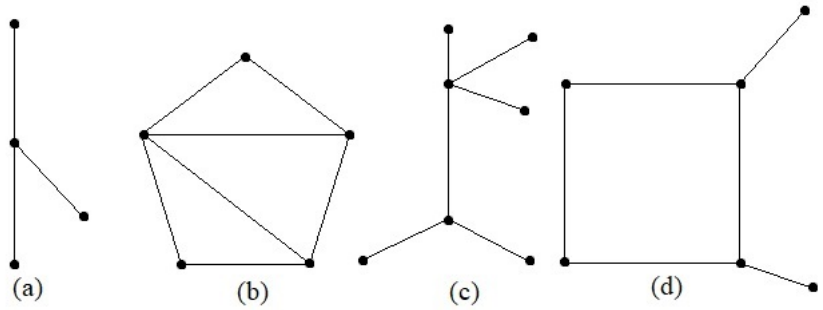
Proof. Suppose G has an open Euler trail P from a vertex u to a vertex v of G. Construct a new graph G_1 by adding a new edge e to G with u and v as the end vertices. In G_1, the trail P with e forms an Euler circuit. Hence every vertex of G_1 is of even degree. In graph G_1, e contributes 1 each to the degree of the vertices u and v. Since G does not contain the edge e, it follows that u and v are the only vertices of odd degree in G.

Conversely assume that a connected graph G has only two vertices u and v of odd degree. Construct a new graph, G_1, by adding a new edge, e, to G with u and v as the end vertices. Then G_1 is a connected graph where every vertex is of even degree. Then G_1 contains an Euler circuit, say P. Now, (u, e, v) is a subpath of P. This subpath is not present in G. Hence, if we delete (u, e, v) from P, then we obtain an open Euler trail P_1 from u to V in G. Hence the theorem is done. □

2.4 Trees

Definition 2.4.1. A graph that is connected and has no cycles is called a tree. Generally, a graph that does not contain any cycles is called an acyclic graph.

Example 2.4.2. Consider the graphs below.



All the graphs are connected. The graphs a and c clearly contains no cycle and hence are trees. Also, the graphs b and d contains cycles and hence are not trees.

Let T be a tree. Then T is a simple connected graph, so T does not have any parallel edges or loops. Let u and v be two vertices in T. It follows that there is at most one edge connecting u and v. Since G is connected,

there is a path from u to v. Let $P = (u, e_1, u_1, e_2, \ldots, u_k, e_k, v)$. If no confusion arises, then we write the path P as $(u, u_1, \ldots, u_k, v)$, that is, when listing the vertices of the path, we will omit the edges.

Theorem 2.4.3. Let u and v be two vertices of a tree T. Then there exists only one path from u to v.

Proof. If $u = v$, then the result is trivial.

Suppose $u \neq v$. Because T is connected, there is at least one path from u to v. Suppose there are distinct paths $P_1 = (u, u_1, \ldots, u_k, v)$ and $P_2 = (u, v_1, \ldots, v_t, v)$ from u to v. Since P_1 and P_2 are distinct, we have the following two cases.

Case 1: $\{u_1, \ldots, u_k\} \cap \{v_1, \ldots, v_t\} = \emptyset$. Then the path P_1 followed by P_2, that is,

$$(u, u_1, \ldots, u_k, v, v_t, \ldots, v_1, u),$$

forms a cycle from u to u, which is a contradiction.

Case 2: $\{u_1, \ldots, u_k\} \cap \{v_1, \ldots, v_t\} \neq \emptyset$. Hence $u_i = v_j$ for some i and j.

Let w_1 be the first common vertex other than u and v, on paths P_1 and P_2. Next, we follow path P_1 until we come to the first vertex w_s, which is again on both P_1 and P_2. This vertex w_s is different from w_1. We must get such a vertex w_s, because P_1 and P_2 meet again at v. Let P_1^* be the portion of the path P_1 from w_1 to w_s and P_2^* be the portion of path P_2 from w_s to w_1. Then, P_1^* followed by P_2^* forms a cycle from w_1 to w_1 in graph T. But this contradicts our assumption that T is a tree, so it has no cycles.

Hence T does not contain two distinct paths between any two distinct vertices u and v.

□

Theorem 2.4.4. In a tree with more than one vertex, there are at least two vertices of degree 1.

Proof. Let T be a tree with more than one vertex. Since T is a connected graph with at least two vertices, there is a path with at least two distinct vertices. Because the number of vertices and the number of edges is finite, the number of paths in T is also finite. Thus we can find a path P of maximal length. Suppose path P is from vertex u to vertex v. We show that $\deg(u) = \deg(v) = 1$.

Suppose $\deg(v) \neq 1$. Let P be the path $(u = v_1, e_1, v_2, e_2, v_3, \ldots, v_{k-1}, e_{k-1}, v)$. Since $\deg(v) \neq 1$, there exists an edge e_k with v as an end vertex such that $e_k \neq e_{k-1}$. Since T has no loops, the other end of e_k can't be v. Suppose the other end is v_k. Suppose $v_k = v_i$ for some i such that $1 \leq i \leq k-1$. Then $(v, e_k, v_i, e_{i+1}, v_{i+1}, \ldots, v_{k-1}, e_{k-1}, v)$ is a cycle from v to v, which contradicts the fact T is a tree. If $v_k \neq v_i$, $1 \leq i \leq k-1$, then we get the path $(v_1, e_1, v_2, e_2, v_3, \ldots, v_{k-1}, e_{k-1}, v, e_k, v_k)$ whose length is greater than that of P. This contradicts the fact that path P is of maximal length in T. It now follows that $\deg(v) = 1$. Similarly, we can show that $\deg(u) = 1$. □

The converse of the above theorem is not true as shown by the following example.

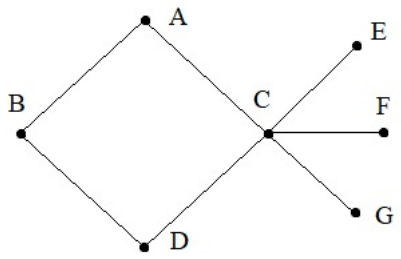
Example 2.4.5. Consider the graph in the given figure.

This is a connected graph and it has at least two vertices of degree 1. But it contains a cycle. Hence it is not a tree.

Theorem 2.4.6. Let T be a tree with n vertices, $n \geq 1$. Then T has exactly $n-1$ edges.

Proof. We prove the result by induction on n.

Basic Step: Let $n = 1$. Since T is a simple graph, it does not contain any loop. Therefore it does not contain any edge and hence the number of edges in T is $0 = 1 - 1$. Hence the theorem is true for $n = 1$.

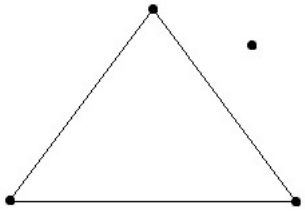


Inductive hypothesis: Let $k \geq 1$ be a positive integer. We assume that the theorem holds for any tree with k vertices.

Inductive step: Let T be a tree with $k+1$ vertices. Since $k+1 \geq 2$, it follows from theorem 2.4.4 that T has at least two vertices of degree 1. Let u be a vertex of degree 1 in T. We construct a new graph G by deleting u from T and also the edge e, which is incident on u. Now, G is still a connected graph and does not contain any cycle. Hence G is a tree with k vertices. By inductive hypothesis, we find that G has exactly $k-1$ edges. This implies that T has k edges. Hence by induction, the theorem holds for any integer n. □

The converse of the above theorem is not true in general. This is proved by the following example.

Example 2.4.7. The given graph is clearly a graph containing 4 vertices and $3 = 4-1$ edges. But this is clearly not a tree. Also, it is not connected.



Theorem 2.4.8. Let T be a graph with n vertices. Then the following are equivalent:

1. T is a tree.
2. T has no loops and if u and v are two distinct vertices in T, then there exists only one path from u to v.
3. T is a connected graph and has $n-1$ edges.
4. T has no cycles and has $n-1$ edges.

2.5 Spanning Tree

We begin with the following definition.

Definition 2.5.1. A tree T is called a spanning tree of a graph G if T is a subgraph of G and T contains all the vertices of G.

Note that the spanning tree of a graph need not be unique. The following theorem gives a necessary and sufficient condition for a graph to have a spanning tree.

Theorem 2.5.2. A graph G has a spanning tree if and only if G is connected.

Proof. Suppose G has a spanning tree, G_1. G_1 contains all the vertices of G. Then between any two vertices, there exists a path in G_1, which is also a path of G. Hence, G is a connected graph.

Conversely, suppose G is a connected graph. If G has no cycles, then G is a tree. Suppose G has cycles. Let C_1 be a cycle in G and e_1 be an edge in C_1. Now construct the graph $G_1 = G \setminus \{e_1\}$, which is obtained by deleting the edge e_1 from G but not removing any vertex from G. Clearly, G_1 is a subgraph of G and it contains all the vertices of G. Because e_1 is an edge of a cycle, G_1 is still a connected graph. If G_1 is acyclic, then G_1 is a tree. If G_1 contains a cycle C_2, then we delete an edge e_2 from C_2 and construct a connected subgraph G_2 that contains all the vertices. If G_2 contains cycles, then we continue this process. Since G has a finite number of edges, it contains only a finite number of cycles. Hence, continuing the process of deleting an edge from a cycle, we eventually obtain a connected subgraph G_k that contains all the vertices of G and is also acyclic. It follows that G_k is a spanning tree of G. □

Exercise 2.5.3.

1. Draw a tree with 9 vertices such that three vertices are of degree 3.
2. How many edges are there in a tree with 16 vertices?
3. How many vertices are there in a tree with 16 edges?
4. Suppose there exists a simple connected graph with 16 vertices that has 15 edges. Does it contain a vertex of degree 1? Justify your answer.

2.6 Few Probable Questions

1. Define bipartite graphs. Show that a graph is bipartite if and only if it does not contain any cycle of odd length.
2. Define Euler circuit. Deduce a necessary condition for a connected graph to be Eulerian.
3. Deduce a necessary and sufficient condition for a connected graph G to have an Euler trail.
4. Define a tree. Show that in a tree T, there exits only one path between two vertices of T.
5. Show that in a tree with more than one vertex, there exits at least two vertices of degree 1. Is the converse true? Justify.
6. Show that a tree with n vertices has $n - 1$ edges. Is the converse true? Justify.

Unit 3

Course Structure

- Planar graphs and their properties
- Fundamental cut set and cycles. Matrix representation of graphs
- Kuratowski's theorem (statement only) and its use
- Chromatic index, chromatic numbers and stability numbers

3.1 Introduction

The present unit starts with the matrix representation of graphs. We have dealt with two types of matrix representations of graphs, viz., the adjacency matrix and the incidence matrix. The matrix representations are compact and say everything about the graph in a very simple manner as we shall see.

The next topic that is covered is the graph isomorphisms. A graph can exist in different forms having the same number of vertices, edges, and also the same edge connectivity. Such graphs are called isomorphic graphs or "equal" or "same" graphs.

Next we have covered the planar graphs. Such graphs in which the can be drawn in a plane of paper can be thought of as a planar graph and such graphs that don't satisfy this property is called a non-planar graph. Of particular importance are the connected simple planar graphs from which we can deduce the Kuratowski's theorem that characterises simple non-planar graphs. The proof of this is however excluded.

Lastly, we have dealt with the graph coloring. It is nothing but a simple way of labelling graph components such as vertices, edges, and regions under some constraints. Vertex coloring and edge coloring are two common graph coloring problems. The graph coloring problem has huge number of applications, like making schedule or time tables, sudoku, map coloring, etc.

We have now given a brief idea about all that we are about to study. Let's explore!

Objectives

After reading this unit, you will be able to

- define incidence matrix and adjacency matrix of a graph

- say when two graphs are said to be same (or, isomorphic)
- define planar graphs and learn related properties
- define planar graphs and related terms like faces, boundaries, etc.
- deduce important results related to planar graphs
- get to know the Kuratowski's theorem
- define vertex and edge coloring and related terms

3.2 Matrix Representation of a Graph

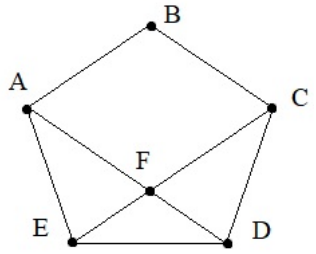
Definition 3.2.1. Let G be a graph with n vertices, where $n > 0$. Let $V(G) = \{v_1, v_2, \ldots, v_n\}$. The **adjacency matrix** A_G with respect to the particular listing, $v_1, v_2, \ldots, v_n$ of n vertices of G is an $n \times n$ matrix $[a_{ij}]$ such that the (i, j)th entry of A_G is the number of edges from v_i to v_j. That is,

$$a_{ij} = \text{number of edges from } v_i \text{ to } v_j.$$

Since a_{ij} is the number of edges from v_i to v_j, the adjacency matrix A_G is a square matrix over the set of non-negative integers.

If G is a digraph, then the adjacency matrix A_G with respect to the particular listing $v_1, v_2, \ldots, v_n$ of n vertices of G is an $n \times n$ matrix $[a_{ij}]$ such that the (i, j)th entry is the number of arcs from a_i to a_j.

Example 3.2.2. Consider the graph G below. The vertices of the graph are $\{A, B, C, D, E, F\}$. Then the



adjacency matrix with respect to this ordering of the vertices is

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

Example 3.2.3. Consider another graph 3.2.1. The vertices of the graph is $\{A, B, C, D\}$. The adjacency matrix of the graph with respect to the listing is

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

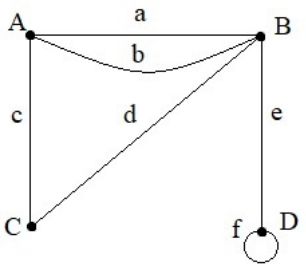


Figure 3.2.1

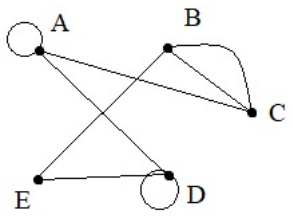
Notice that the matrix A_G is symmetric symmetric since $a_{ij} = a_{ji}$. But, if G is a digraph, then the adjacency matrix need not be symmetric. The adjacency matrix has the following properties:

1. If G does not contain any loops and parallel edges, then each element of A_G is either 0 or 1.
2. If G does not contain any loops, then all the diagonal elements of A_G are 0.

Example 3.2.4. Let A denote 5×5 matrix

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

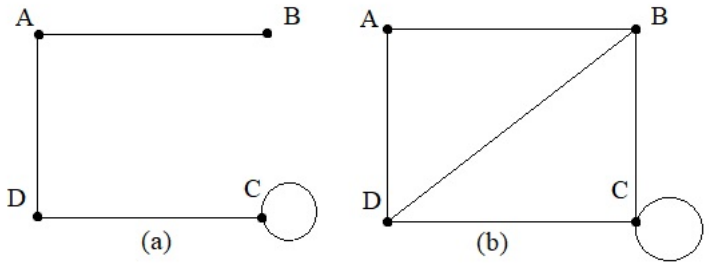
We construct a graph G such that $A_G = A$. For this, we denote the rows by A, B, C, D, E and the columns by A, B, C, D, E. Now we draw a graph with vertices A, B, C, D, E. Since $(1, 1)$ and $(4, 4)$ are the only diagonal elements with entries equal to one, we draw one loop each at the vertices A and D only. Now, we see that $(1, 2)$th element= $(2, 1)$th element $= 0 \Rightarrow$ there is no edge between A and B. Again, $(1, 3)$th element $= (3, 1)$th element $= 1 \Rightarrow$ there exists one edge between A and C. Continuing in this way, we find the following graph



Definition 3.2.5. Let G be a graph with n vertices $v_1, v_2, \ldots, v_n$, where $n > 0$ and m edges $e_1, e_2, \ldots, e_m$. The incidence matrix I_G with respect to the ordering $v_1, v_2, \ldots, v_n$ of vertices and $e_1, e_2, \ldots, e_m$ edges is an $n \times m$ matrix $[a_{ij}]$ such that

$$\begin{aligned} a_{ij} &= 0; \text{ if } v_i \text{ is not an end vertex of } e_j, \\ &= 1; \text{ if } v_i \text{ is an end vertex of } e_j \text{ but } e_j \text{ is not a loop}, \\ &= 2; \text{ if } e_j \text{ is a loop at } v_i. \end{aligned}$$

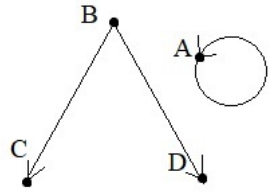
Exercise 3.2.6. 1. Find the adjacency matrix of the following graphs with respect to the listing A, B, C, D of the vertices:



2. Draw the graph of G represented by the given adjacency matrix

$$(a)\ A_G = \begin{bmatrix} 0 & 2 & 2 & 0 \\ 2 & 0 & 0 & 1 \\ 2 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix} \quad (b)\ A_G = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 0 \\ 2 & 1 & 0 \end{bmatrix} \quad (c)\ A_G = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

3. Find the adjacency matrix of the digraph with respect to the listing A, B, C, D:



4. Draw the digraph represented by the given adjacency matrices:

$$(a) \begin{bmatrix} 1 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix} \quad (b) \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

5. Find the adjacency matrices of the graphs K_3 and $K_{2,3}$.

3.3 Isomorphism

Definition 3.3.1. Let $G_1 = (V_1, E_1, g_1)$ and $G_2 = (V_2, E_2, g_2)$ be two graphs. G_1 is said to be isomorphic to G_2 if there exists a one-to-one correspondence $f : V_1 \to V_2$ and a one-to-one correspondence $h : E_1 \to E_2$ in such a way that for any edge $e_k \in E_1$, $g_1(e_k) = \{v_i, v_j\}$ in G_1 if and only if $g_2(h(e_k)) = \{f(v_i), f(v_j)\}$ in G_2.

In other words, if $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ be two graphs, then G_1 is said to be isomorphic to G_2 if there exist a one-to-one correspondence $f : V_1 \to V_2$ and a one-to-one correspondence $h : E_1 \to E_2$ such that for any edge e_k in E_1, vertices v_i, v_j are end vertices of e_k in G_1 if and only if $f(v_i)$, $f(v_j)$ are end vertices of $h(e_k)$ in G_2. When we say two graphs are same, we mean they are isomorphic to each other.

Example 3.3.2. Let G and H be graphs as in figure 3.3.1 Both these graphs have six vertices and six edges.

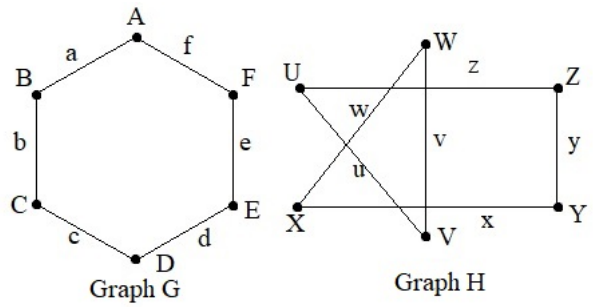


Figure 3.3.1

Moreover, both the graphs are simple. The degree sequence of both the graphs is $2, 2, 2, 2, 2, 2$.

Let us define $f : V_1 \to V_2$ and $h : E_1 \to E_2$ by

$$f : \quad A \mapsto U,\ \ B \mapsto V,\ \ C \mapsto W,\ \ D \mapsto X,\ \ E \mapsto Y,\ \ F \mapsto Z;$$
$$h : \quad a \mapsto u,\ \ b \mapsto v,\ \ c \mapsto w,\ \ d \mapsto x,\ \ e \mapsto y,\ \ f \mapsto z.$$

Then we can check that these maps f and h serve as the one-to-one correspondence maps between the vertex sets and edge sets of the two graphs that satisfies the isomorphism conditions. Thus G and H are isomorphic.

If two graphs G_1 and G_2 are isomorphic, then it is written as $G_1 \simeq G_2$.

The following theorem is evident

Theorem 3.3.3. Let G, G_1, G_2 and G_3 be graphs. Then the following assertions hold:

(i) $G \simeq G$;

(ii) If $G_1 \simeq G_2$, then $G_2 \simeq G_1$;

(iii) If $G_1 \simeq G_2$, and $G_2 \simeq G_3$, then $G_1 \simeq G_3$.

Proof. Left as an exercise. □

Definition 3.3.4. Two graph G_1 and G_2 are said to be different if G_1 is not isomorphic to G_2.

Let us write few properties of isomorphic graphs.

1. Two graphs G_1 and G_2 are isomorphic if and only if there exists a one-to-one correspondence f between the vertex sets of them such that if v_1, v_2 are adjacent vertices in G_1, then $f(v_1)$ and $f(v_2)$ are adjacent vertices in G_2.

2. Two graphs G_1 and G_2 are isomorphic. Then G_1 has a vertex of degree k if and only if G_2 has a vertex of degree k.

3. Two graphs G_1 and G_2 are isomorphic. Then G_1 has a cycle of length k if and only if G_2 has a cycle of length k.

3.4 Planar Graphs

Consider the graph in figure 3.4.1a. It can be redrawn as in the figure 3.4.1b.

(a) (b)

Figure 3.4.1: Planar Graphs

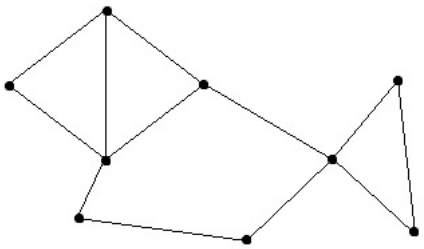
We can also say that the above two graphs are isomorphic or *equal*. In the latter graph, notice that no two edges intersect except at the vertices. Such graphs are called planar graphs as we will formally define now.

Definition 3.4.1. A graph G is called a planar graph if it can be drawn in the plane such that no two edges intersect except at the vertices, which may the common end point of the edges. We can also say that a graph is planar if it is isomorphic to a graph having the property said above.

Definition 3.4.2. A graph drawn in the plane (on paper or a chalkboard) is called a plane graph if no two edges meet at any point except the common vertex, if they meet at all.

From the preceding two definitions, it is clear that a graph is a planar graph if and only if it has a pictorial representation in a plane which is a plane graph. The pictorial representation of a planar graph G as a plane graph is called the planar representation of G.

Consider the planar representation of a planar graph given below

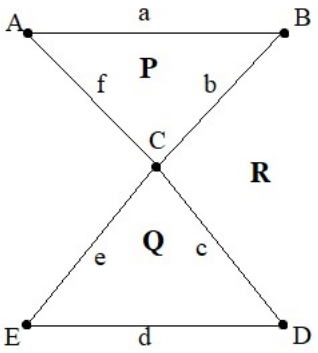


Let G denotes the graph in the above figure. Then G divides the plane into different *regions*, called the **faces** of G. Suppose x is a point in the plane that is not a vertex of G or a point on any edge of G. Then a face of G containing x is the set of all points on the plane that can be reached from x by a straight line or a curved line that does not cross any edge of G or pass through any vertex of G. Thus, it follows that a face is a region produced by a planar graph that is an area of the plane bounded by the edges and that is not further subdivided into sub-areas.

The set of edges that *bound* a region is called its **boundary**. Of course, there exists a region of infinite area in any plane graph G. This is the part of the plane that lies outside the planar representation of G. This region

is called the **exterior face**. A face that is not exterior is called an **interior face**. We illustrate these concepts by the following example.

Example 3.4.3. Consider the graph below



This plane graph divides the plane into three regions:

1: Bounded by the cycle (A, a, B, b, C, f, A). The boundary of P consists of the edges a, b, f.

2: Bounded by the cycle (D, d, E, e, C, c, D). The boundary of Q consists of the edges d, e, c.

3: The part of the plane outside this plane graph. The boundary of the region consists of the edges a, b, c, d, e and f.

It follows that this plane graph contains three faces, namely P, Q and R.

For this plane graph, the number of edges $n_e = 6$, the number of vertices $n_v = 5$, the number of faces $n_f = 3$, and we see that

$$n_v - n_e + n_f = 2.$$

Theorem 3.4.4. Let G be a connected planar graph with n_v vertices, n_e edges and n_f faces. Then $n_v - n_e + n_f = 2$.

Proof. We prove the theorem by induction on n_e.

Basic Step: Let $n_e = 0$. Then it has only one vertex and one region, which is the exterior region. Then, $n_v - n_e + n_f = 1 - 0 + 1 = 2$.

Inductive hypothesis: Let k be a positive integer and assume that $n_v - n_e + n_f = 2$ for any connected planar graph with $n_e = k - 1$.

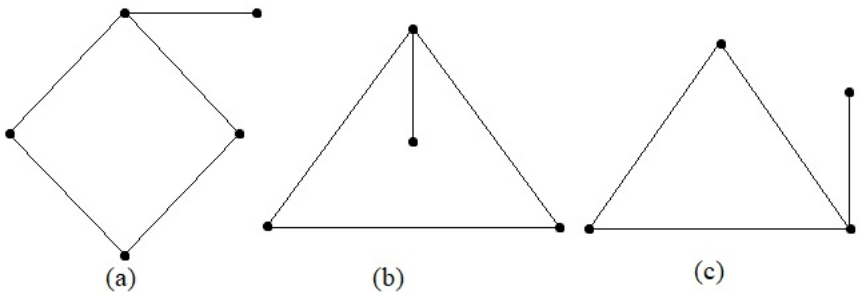
Inductive Step: Let G be a connected planar graph with $n_e = k$ edges and $n_f = t$ vertices. Suppose G has no cycles. Then G has no interior region, which implies that the exterior region is the only region of the graph. Thus, $n_f = 1$. We now show that G contains a vertex of degree 1. Choose a vertex v in G. If $\deg(v) = 1$, we are done. Suppose $\deg(v) > 1$. Let v_1 be an adjacent vertex of v. Since G has no cycles, G is loop free and hence $v_1 \neq v$. If $\deg(v_1) = 1$, we are done. Suppose $\deg(v_1) > 1$. Let v_2 be an adjacent vertex of v_1. Since G has no cycles, G is loop free and hence v_2 is different from v and v_1. If $\deg(v_2) \neq 1$, we find an adjacent vertex v_3 of v_2 different from v, v_1 and v_2. Because G has finite number of vertices, it follows that G has a vertex u of degree 1. We now delete this vertex and thus from a new connected planar graph H with $k - 1$ edges and $t - 1$ vertices. By the inductive hypothesis, for this graph H, we have, $n_v - n_e + n_f = 2$. Hence, $(t - 1) - (k - 1) + n_f = 2$, which implies that $t - k + n_f = 2$, that is, $n_v - n_e + n_f = 2$ holds for G.

Suppose now that G has a cycle C. Let e be an edge in C. Now construct a new graph $G_1 = G \setminus \{e\}$. This is still a connected planar graph. For this planar graph G_1, we compute n_v, n_e and n_f. Let $n_f = m$. In the

construction of G_1, we delete only the edge without deleting any vertex. Therefore, $n_v = t$, $n_e = k-1$. Now, $C \setminus \{e\}$ will not form a boundary in G_1. Thus in G_1, $n_f = m - 1$. Hence G_1 is a connected planar graph with $n_v = t$ vertices, $n_e = k - 1$ edges, and $n_f = m - 1$ faces. By the inductive hypothesis, it follows that $t - (k-1) + (m-1) = 2$. This implies that $t - k + m = 2$. Hence, $n_v - n_e + n_f = 2$.

The result now follows by induction. □

Exercise 3.4.5. Verify the above theorem for the following graphs:



Corollary 3.4.6. The graph $K_{3,3}$ is not a planar graph.

Theorem 3.4.7. Let G be a connected simple planar graph with $n_v \geq 3$ vertices and n_e edges. Then

$$n_e \leq 3n_v - 6.$$

Proof. Since G is a planar graph, it has a planar representation. Consider a planar representation of G. Suppose $n_v = 3$. Because G is a simple connected graph with 3 vertices, it follows that $n_e \leq 3$. Then $n_e \leq 3 \cdot 3 - 6 = 3$, which implies that $n_e \leq 3n_v - 6$.

Suppose now $n_v \geq 3$. If G does not contain any cycles then we can show that $n_e = n_v - 1$. Now, $3n_v - 6 = (n_v - 1) + (n_v - 2) + (n_v - 3) > (n_v - 1) = n_e$.

Suppose G contains a cycle. Because G is simple, it may contain a cycle with 3 edges. Thus, the number of edges in the boundary of a face is ≥ 3. Now, there are n_f faces and every edge is a member of some boundary of the planar representation. Hence, the total number of appearances of the edges in boundaries of n_f faces is $\geq n_f \cdot 3$. In counting these appearances, an edge may be counted atmost two times. Thus, the total number of appearances of the n_e edges in boundaries is $\leq 2n_e$. Hence, $n_f \cdot 3 \leq 2n_e$. Now, by Euler's theorem,

$$\begin{aligned} & n_v - n_e + n_f = 2 \\ \Rightarrow \quad & 3n_v - 3n_e + 3n_f = 6 \\ \Rightarrow \quad & 3n_e = 3n_v + 3n_f - 6 \\ \Rightarrow \quad & 3n_e \leq 3n_v + 2n_e - 6 \\ \Rightarrow \quad & n_e \leq 3n_v - 6. \end{aligned}$$

□

Corollary 3.4.8. The graph K_5 is not a planar graph.

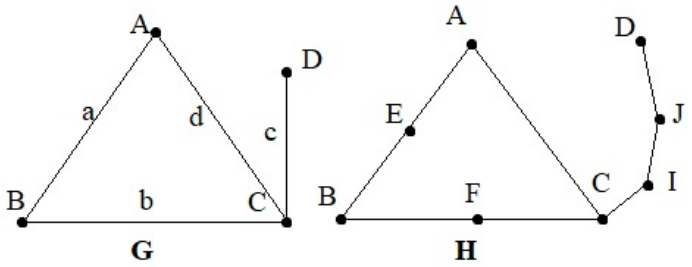
Proof. Left as an exercise. □

Let $G = (V, E)$ be a graph. Suppose that e is an edge with v_1, v_2 as end vertices. Construct the subgraph $G_1 = G \setminus \{e\}$. To construct G_1, we have deleted edge e without deleting any vertices from G. We now construct a new graph, $G_2 = (V_2, E_2)$, by taking $V_2 = V \cup \{w\}$, $E_2 = (E \setminus \{e\}) \cup \{f_1, f_2\}$ such that $w \notin V$, $f_1, f_2 \notin E$, v_1, w are end vertices of f_1 and v_2, w are end vertices of f_2. The process of obtaining G_2 from G is called a one-step subdivision of an edge of G.

Definition 3.4.9. A graph H is said to be a subdivision of a graph G if there exist graphs $H_0, H_1, H_2, \ldots, H_n$ such that $H_0 = G$, $H_n = H$, and H_i is obtained from H_{i-1} be a one-step subdivision of an edge of H_{i-1} for $i =, 1, 2, \ldots, n$.

If a graph H is a subdivision of a graph G, then we say that H is obtained from G by subdividing the edges of G.

Example 3.4.10. Consider graphs G and H below.

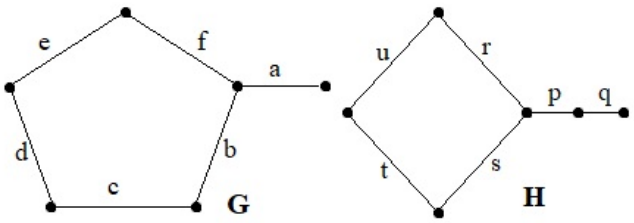


We see that H is obtained from H by a finite sequence of subdivisions of edges. H is obtained from G by dividing the edge a one time, b one time and c twice.

Definition 3.4.11. Two graphs G and H are said to be homeomorphic graphs if there is an isomorphism from a subdivision of G to a subdivision of H.

Consider the following example.

Example 3.4.12. Consider the graphs G and H below. We see that G contains a cycle of length 5, and H



contains a cycle of length 4. Hence these two graphs are not isomorphic. But we find a subdivision G' of G and H' of H such that G' and H' are isomorphic (see fig. 3.4.2). Hence G and H are homeomorphic.

In 1930, Kuratowski proved the following famous theorem, characterising simple planar graphs in terms of K_5 and $K_{3,3}$.

Theorem 3.4.13. Kuratowski. A simple graph is planar if and only if it does not contain a subgraph homeomorphic to K_5 or $K_{3,3}$.

The proof of the above theorem is omitted.

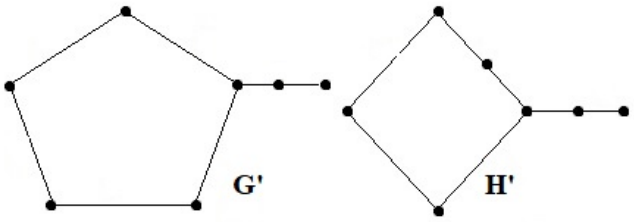


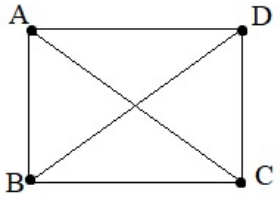
Figure 3.4.2

3.5 Graph Coloring

Definition 3.5.1. Let $G = (V, E)$ be a simple graph and $C = \{c_1, c_2, \ldots, c_n\}$ be a set of n colors. A **vertex coloring** of G using the colors of C is a function $f : V \to C$. Let $f : V \to C$ be a vertex coloring of G.If for every adjacent vertices $u, v \in V$, $f(u) \neq f(v)$, then f is called a **proper vertex coloring**.

For each vertex v, its image $f(v)$ is called the **color** of v. It follows that a vertex coloring of a graph G is an assignment of the colors $c_1, c_2, \ldots, c_n$ to the vertices of graph G. Similarly, a proper vertex coloring of G is an assignment of the colors $c_1, c_2, \ldots, c_n$ to the vertices of G such that adjacent vertices have different colors. The following graph is an illustration.

Example 3.5.2. Consider the following graph:



This is a graph with 4 vertices A, B, C and D. Suppose $C = \{r, b, y, g\}$, where r denotes red, b denotes blue, y denotes yellow and g denotes green. Define $f : V \to C$ by

$$\begin{aligned} A &\mapsto r \\ B &\mapsto g \\ C &\mapsto y \\ D &\mapsto b. \end{aligned}$$

Then f is a proper vertex coloring with four colors.

Definition 3.5.3. The smallest number of colors needed to make a proper vertex coloring of a simple graph G is called the **chromatic number** of G and is denoted by $\chi(G)$.

Next we determine the chromatic number of bipartite graphs.

Theorem 3.5.4. Let G be a nontrivial simple graph. Then $\chi(G) = 2$ if and only if G is a bipartite graph.

Proof. Let $G = (V, E)$ be a bipartite graph. Then vertex set V can be partitioned into two non-empty subsets V_1 and V_2 such that each edge of G is incident with one vertex of V_1 and one vertex of V_2. Let $C = \{c_1, c_2\}$ be a set of two colors.

Define a function $f : V \to C$ such that

$$\begin{aligned} f(v) &= c_1; \text{ if } v \in V_1 \\ &= c_2; \text{ if } v \in V_2. \end{aligned}$$

Since $V_1 \cap V_2 = \emptyset$, it follows that f is well-defined. Now, no two vertices of V_1 are adjacent. Therefore, all the vertices can have the same color. Similarly, all the vertices of V_2 can have the same color. From the definition of f, it follows that two adjacent vertices of G have different colors. Thus, $\chi(G) \leq 2$. Also, since G has at least one edge, $\chi(G) < 1$. Hence combining, we get $\chi(G) = 2$.

Conversely suppose that $\chi(G) = 2$. This implies that the graph contains at least one edge. Also, there exists a function $f : V \to C = \{c_1, c_2\}$ such that no two adjacent vertices have the same image.

Let $V_1 = \{v \in V : \ f(v) = c_1\}$ and $V_2 = \{v \in V : \ f(v) = c_2\}$. It follows that $V_1 \cap V_2 = \emptyset$ and $V_1 \cup V_2 = V$. Let e be an edge with end vertices v_1 and v_2. Because v_1 and v_2 can't have the same color, $v_1 \in V_1$ and $v_2 \in V_2$. Thus, G is a bipartite graph. □

Definition 3.5.5. Let G be a graph with vertices $v_1, v_2, \ldots, v_{n-1}, v_n$. The maximum of the integers $\deg(v_i)$, for $i = 1, 2, \ldots, n$ is denoted by $\Delta(G)$.

Theorem 3.5.6. For any simple graph G, $\chi(G) \leq \Delta(G) + 1$.

Proof. We prove this theorem by induction on n, where n is the number of vertices of G.

Basic Step: Let $n = 1$. Then G is a graph with only one vertex and G has no edge. Hence $\chi(G) = 1$ and $\Delta(G) = 0$. This implies that $\chi(G) \leq \Delta(G) + 1$ for $n = 1$.

Inductive hypothesis: Suppose that $k > 1$ is an integer such that for any simple graph G, with $k-1$ vertices, $\chi(G) \leq \Delta(G) + 1$.

Inductive step: Let G be a simple graph with k vertices. Consider a vertex v of G and construct the graph $G_1 = G \setminus \{v\}$. The graph G_1 is obtained by deleting the vertex v and also all the edges incident on v. Clearly, $\Delta(G_1) \leq \Delta(G)$. This is a simple graph with $k - 1$ vertices. Thus, by the inductive hypothesis, $\chi(G_1) \leq \Delta(G_1) + 1$. Then, $\chi(G_1) \leq \Delta(G) + 1$. This implies that G_1 can be properly colored by atmost $\Delta(G_1) + 1$ colors. Now, v has atmost $\Delta(G)$ adjacent vertices. Because $\Delta(G) < \Delta(G) + 1$, it follows that not all the $\Delta(G) + 1$ colors are needed to color these $\Delta(G)$ adjacent vertices. Thus, from these $\Delta(G) + 1$ colors one unused color is definitely available to color vertex v. Hence, $\chi(G) \leq \Delta(G) + 1$. □

Definition 3.5.7. Let $G = (V, E)$ be a simple graph and $C = \{c_1, c_2, \ldots, c_n\}$ be a set of n colors. An **edge coloring** of G using the colors of C is a function $f : E \to C$. Let $f : E \to C$ be an edge coloring of G. If, for any two edges e_1 and e_2 meeting at a common vertex, $f(e_1) \neq f(e_2)$, then f is called a **proper edge coloring**.

For each edge e, its image $f(e)$ is called the color of e. It follows that a proper edge coloring of a graph G is an assignment of the colors $c_1, c_2, \ldots, c_n$ to the edges of graph G such that any two edges meeting at a common vertex have different colors. The following graph is an illustration.

Example 3.5.8. Consider the graph G in fig 3.5.1.

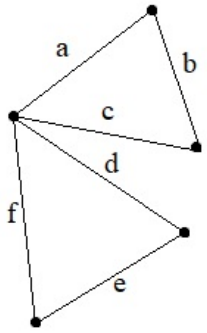


Figure 3.5.1

The graph G has six edges a, b, c, d, e, f. Suppose $C = \{R, B, Y, G\}$, where R denotes red, B denotes blue, Y denotes yellow, and G denotes green. Define $f : E \to C$ by

$$\begin{aligned} a &\mapsto R \\ c &\mapsto G \\ d &\mapsto B \\ f &\mapsto Y \\ b &\mapsto B \\ e &\mapsto R. \end{aligned}$$

Then f is a proper edge coloring of the graph G.

Definition 3.5.9. The smallest number of colors needed to make a proper coloring of the edges of a simple graph G is called the chromatic index of G, and is denoted by $\chi'(G)$.

For a simple graph, we have the following theorem.

Theorem 3.5.10. For any simple graph G, we have, $\chi'(G) = \Delta(G)$ or $\chi'(G) = \Delta(G) + 1$.

Let us see a few examples.

Example 3.5.11. In a connected simple planar graph, there exists a vertex v such that $\deg(v) \leq 5$.

We know that in a connected simple planar graph, $n_e \leq 3n_v - 6$. Suppose $\deg(v) \geq 6$ for all vertices v. Now, $\sum \deg(v) = 2n_e$. Hence $2n_e \geq 6n_v$. Again, $2n_e \leq 6n_v - 12$. This implies that $6n_v \leq 6n_v - 12$. Thus, we find that $0 \leq -12$, which is absurd. Hence the result.

Example 3.5.12. For the graph $K_{2,3}$, we find $\chi(K_{2,3})$. Let us first draw the graph (fig. 3.5.2). We see that p, q, r are adjacent vertices of both a and b. Let $C = \{G, R\}$ be the set of two colors. Let us define $f : V \to C$ as follows:

$$\begin{aligned} p &\mapsto R \\ a &\mapsto G \\ b &\mapsto G \\ q &\mapsto R \\ r &\mapsto R. \end{aligned}$$

This is a proper coloring of G. Hence $\chi(K_{2,3}) = 2$.

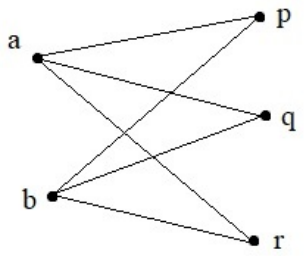
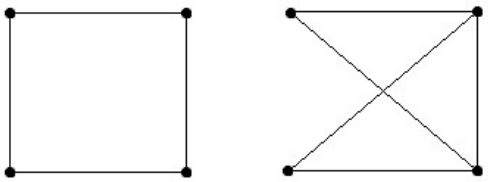


Figure 3.5.2

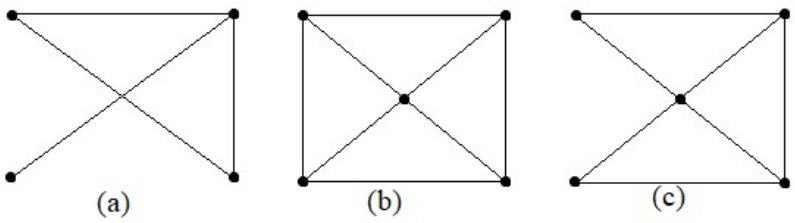
Example 3.5.13. For the graph K_n, we find $\chi(K_n)$. K_n is a complete graph with n vertices. For any vertex v of K_n, each of the remaining $n-1$ vertices is an adjacent vertex of v. Hence we need n distinct colors for proper coloring of K_n. Then, $\chi(K_n) \geq n$. But K_n has n vertices. So, $\chi(K_n) = n$.

3.6 Few Probable Questions

1. Define isomorphism of graphs. Determine whether the following graphs are isomorphic:



2. Define planar graphs. Show that for a connected simple planar graph G with $n_v \geq 3$, $n_e \leq 3n_v - 6$.
3. Show that for a connected planar graph, $n_v - n_e + n_f = 2$.
4. Define chromatic number of a graph G. Show that a simple nontrivial graph G has chromatic number 2 if and only if G is bipartite.
5. Show that for any simple graph G, $\chi(G) \leq \Delta(G) + 1$.
6. Find $\chi(G)$ for each of the following graphs:



Unit 4

Course Structure

- Lattices as partial ordered sets. Their properties.
- Lattices as algebraic system. Sublattices.
- Direct products and Homomorphism.
- Some special Lattices e.g. complete complemented and distributed lattices.

4.1 Introduction

A lattice is an abstract structure studied in the mathematical subdisciplines of order theory and abstract algebra. It consists of a partially ordered set in which every two elements have a unique supremum (also called a least upper bound or join) and a unique infimum (also called a greatest lower bound or meet). An example is given by the natural numbers, partially ordered by divisibility, for which the unique supremum is the least common multiple and the unique infimum is the greatest common divisor.

Lattices can also be characterized as algebraic structures satisfying certain axiomatic identities. Since the two definitions are equivalent, lattice theory draws on both order theory and universal algebra. Semilattices include lattices, which in turn include Heyting and Boolean algebras. These "lattice-like" structures all admit order-theoretic as well as algebraic descriptions.

Objectives

After reading this unit, you will be able to

- define partial ordered sets and see its examples
- upper and lower bounds of a poset
- define lattice and deduce the algebra of join and meet
- draw the Hasse diagram for posets
- define sublattice and direct product of lattice

4.2 Partially Ordered Sets

Definition 4.2.1. A relation R on a set S is called **antisymmetric** if for all $a, b \in S$, $aRb \in R$ and $bRa \in R$ implies $a = b$.

On the set of all integers, the usual "less than or equal to" relation is an antisymmetric relation since $a \leq b$ and $b \leq a$ implies $a = b$.

Similarly if T is the set of all subsets of a set A, then the inclusion relation "$\subseteq$" is an antisymmetric relation since for any two subsets X and Y of A, we always have $X \subseteq Y$ and $Y \subseteq X$ implies $X = Y$.

Definition 4.2.2. A relation R on a set A is called a **partial order** on A if R is reflexive, antisymmetric and transitive. In other words, if R satisfies the following conditions:

1. aRa for all $a \in A$;

2. For all $a, b \in A$ if aRb and bRa, then $a = b$;

3. For all $a, b, c \in A$, if aRb and bRc, then aRc.

A set A together with a partial order relation R is called a **partially ordered set**, or **poset**, and we denote this poset by (A, R).

Let (A, R) be a poset. If there is no confusion about the partial order, we may refer to the poset simply by A.

Example 4.2.3. The set $\mathbb{Z}$, together with the usual "less than or equal to", $\leq$ relation is a poset. Note that the relation '$<$' is not a partial order relation on $\mathbb{Z}$ since the relation is not reflexive.

Example 4.2.4. Consider $\mathbb{N}$, the set of all natural numbers, and the divisibility relation R on $\mathbb{N}$. That is, for all $a, b \in \mathbb{N}$, aRb if $a|b$, that is, there exists a positive integer c such that $b = ac$. Check that this relation R is partial ordered. Thus, $\mathbb{N}$ with the divisibility relation is a poset.

Though the divisibility relation is a partial order relation on the set of all positive integers, it is not so on the set of all nonzero integers. For example, $5 = (-1)(-5)$ and also, $-5 = (-1)(5)$ and thus, $5|(-5)$ and $(-5)|5$ but $5 \neq -5$.

Let R be a partial order on a set A, that is, (A, R) is a poset. We usually denote R by $\leq_A$, or simple $\leq$. If A is a partially ordered set with a partial order $\leq$, then we denote it has $(A, \leq_A)$ or $(A, \leq)$.

Definition 4.2.5. Let $(S, \leq)$ be a poset and $a, b \in S$. If either $a \leq b$ or $b \leq a$ holds, then we say that a and b are **comparable**. The poset $(S, \leq)$ is called a linearly set, or totally ordered set, or a chain. if for all $a, b \in S$, either $a \leq b$ or $b \leq a$.

Example 4.2.6. Consider the poset $(\mathbb{Z}, \leq)$ with the less equal to relation. For any two integers a and b, either $a < b$, or $a = b$, or $b < a$. Thus, any two integers with respect to the partial order $\leq$ are comparable. Hence $(\mathbb{Z}, \leq)$ is a chain.

Example 4.2.7. Consider the poset $(\mathbb{N}, \leq)$ with respect to the divisibility relation. Notice here that 2 does not divide 5 and 5 does not divide 2. Thus, 2 and 5 are not comparable and hence $(\mathbb{N}, \leq)$ is not a chain.

Theorem 4.2.8. Any subset T of a poset S is itself a poset under the same relation (restricted to T). Any subset of a chain is a chain.

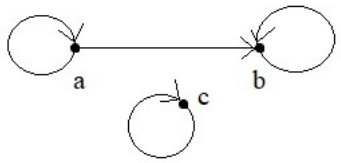
4.2.1 Digraphs of Posets

Because any partial order is also a relation, we can give a digraph representation of partial order.

Example 4.2.9. On the set $S = \{a, b, c\}$, consider the relation

$$R = \{(a, a), (b, b), (c, c), (a, b)\}.$$

The digraph of R is shown below.



From the directed graph it follows that the given relation is reflexive and transitive. This relation is also asymmetric because there is a directed edge from a to b, but there is no directed edge from b to a. Again, in the graph we notice that there are two distinct vertices a and c such that there are no directed edges from a to c and from c to a.

In a digraph of a partial order, one can see that if there is a directed edge from a vertex a to a different vertex b, then there is no directed edge from b to a.

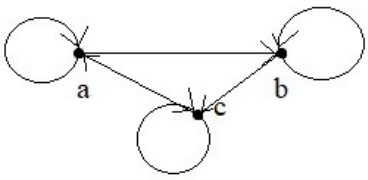
Theorem 4.2.10. A digraph of a partial order relation R cannot contain a closed directed path other than loops. (A path $a_1, a_2, \ldots, a_n$ in a digraph is closed if $a_1Ra_2, a_2Ra_3, \ldots, a_nRa_1$.)

By the above theorem, it follows that if a digraph of a relation contains a closed path other than loops, then the corresponding relation is not a partial order.

Example 4.2.11. On the set $S = \{a, b, c\}$, consider the relation

$$R = \{(a, a), (b, b), (c, c), (a, b), (b, c), (c, a)\}.$$

The digraph of the above relation is given by



In this digraph, a, b, c, a forms a closed path. Hence, the given relation is not a partial order relation.

Hasse Diagram

Posets can also be represented visually by Hasse diagram. First we define a few terms that we will need in the sequel.

Let $(S, \leq)$ be a poset and $x, y \in S$. We say that y **covers** x, if $x \leq y$, $x \neq y$, and there are no element $x \in S$ such that $x < z < y$.

We draw a diagram using the elements of S as follows: We represent the elements of S in the diagram by the elements themselves such that if $x \leq y$, then y is placed above x. We connect x with y by a line segment if and only if y covers x. The resulting diagram is called the **Hasse diagram** of $(S, \leq)$. We see the illustration below.

Example 4.2.12. Let $S = \{1, 2, 3\}$. Then $\mathcal{P}(S) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}, \{1, 3\}, S\}$. Now, $(\mathcal{P}(S), \leq$) is a poset, where $\leq$ denotes the set inclusion relation. The poset diagram of $(\mathcal{P}(S), \leq)$ is shown in fig. 4.2.1.

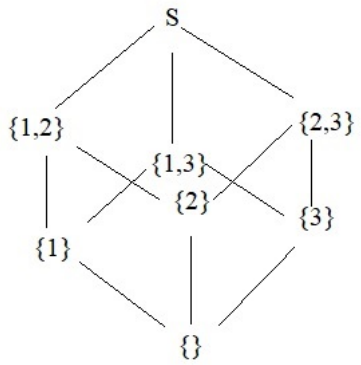


Figure 4.2.1

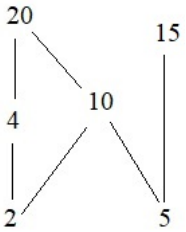
Minimal and Maximal Elements

Definition 4.2.13. Let $(S, \leq)$ be a poset. An element $a \in S$ is called

1. a **minimal element** if there is no element $b \in S$ such that $b < a$,
2. a **maximal element** if there is no element $b \in S$ such that $a < b$,
3. a **greatest element** if $b \leq a$ for all $b \in S$,
4. a **least element** if $a \leq b$ for all $b \in S$.

Let us illustrate this with the following example.

Example 4.2.14. Let $S = \{2, 4, 5, 10, 15, 20\}$. Let $(S, \leq)$ be a poset where $\leq$ denotes the divisibility relation. Then the Hasse diagram becomes



Now, it is clear from the Hasse diagram that there exists no greatest or least element of the poset since no element a satisfies $b \leq a$ for all $b \in S$ (for example, $2 \leq 15$ is not satisfied and also, $15 \leq 20$ is not satisfied),

and also, no element a exists that satisfy $a \leq b$ for every $b \in S$ (if we consider 2 or 5 as the least element then we see that $2 \leq 5$ does not hold and also $5 \leq 2$ does not hold). Now, 5 and 2 are definitely minimal elements since there exist no element $b \in S$ such that $b < 2$ or $b < 5$ (in other words, there is no line segment in the Hasse diagram extending below 2 or 5). Also, 20 and 15 are maximal elements of the poset (verify).

The following lemma ensures the existence of minimal element for every finite poset.

Lemma 4.2.15. Let $(S, \leq)$ be a poset such that S is a finite non-empty set. Then this poset has a minimal element.

Proof. Let a_1 be an element of S. If a_1 is a minimal element, then we are done. Suppose a_1 is not a minimal element. Then there exists $a_2 \in S$ such that $a_2 < a_1$. If a_2 is a minimal element, then we are done, otherwise there exists $a_3 \in S$ such that $a_3 < a_2$. If a_3 is not a minimal element, then we repeat this process. Now, $a_3 < a_2 < a_1$ shows that a_3, a_2, a_1 are distinct elements in S. Since S is finite, after a finite number of steps, we get an element $a_n \in S$, such that a_n is a minimal element. □

We must note that, a poset $(S, \leq)$, where S is a finite non-empty set, has minimal and maximal elements but may not have least or greatest elements. You can take the previous example as a confirmation of this fact.

Definition 4.2.16. Let S be a set and let $\leq_1$ and $\leq_2$ be two partial orders on S. The relation $\leq_2$ is said to be **compatible** with the relation $\leq_1$ if $a \leq_1 b$ implies $a \leq_2 b$.

It should be noted that given a finite non-empty set, say S, we can define a linear order in it as follows.

Since S is non-empty, S has at least one element. Choose an element S, and call it the first element, a_1. Let $S_1 = S \setminus \{a_1\}$. If S_1 is not empty, then from S_1, choose an element a_2. Let $S_2 = S \setminus \{a_1, a_2\}$. If S_2 is not empty, then from S_2, choose an element a_3. Let $S_3 = S\setminus\{a_1, a_2, a_3\}$. If S_3 is not empty, continue this process. Since S is a finite set, this process must stop after a finite number of steps. Hence, there exists a positive integer n such that $S_n = S \setminus \{a_1, \ldots, a_n\}$ is empty, where a_n is the element of $S_{n-1} = S \setminus \{a_1, \ldots, a_{n-1}\}$. We now define a partial order $\leq_1$ on S by $a_1 \leq_1 a_2 \leq_1 a_3 \cdots \leq_1 a_n$. This means that $a_i \leq_1 a_j$ if and only if either $i = j$ or $i < j$, where $i, j \in \{1, 2, \ldots, n\}$. It follows that this is a linear order.

Next suppose that not only S is a finite non-empty set, but S also has a partial order $\leq$. Can we define a linear order $\leq_1$ on S that is compatible with the partial order $\leq$? This following theorem is all about answering this question.

Theorem 4.2.17. Let $(S, \leq)$ be a finite poset. There exists a linear order $\leq_1$ on S which is compatible with the relation $\leq$.

We omit the proof of this theorem and go on to define lattices.

4.3 Lattice

Definition 4.3.1. Let $(S, \leq)$ be a poset and let $\{a, b\}$ be a subset of S. An element $c \in S$ is called an **upper bound** of $\{a, b\}$ if $a \leq c$ and $b \leq c$. Also, if T is any subset of S, then $c \in S$ is called an upper bound of T if $t \leq c$ for all $t \in T$.

An element $d \in S$ is called **least upper bound (lub)** of $\{a, b\}$ if,

1. d is an upper bound of $\{a, b\}$; and
2. if $c \in S$ is an upper bound of $\{a, b\}$, then $d \leq c$.

We can also define the lub of any general subset T of S and denote it by $\sup T$.

Example 4.3.2. Consider the set $\mathbb{N}$ together with the divisibility relation. Consider the subset $\{12, 8\}$. We see that $24, 48, 72$ are all common divisors of 12 and 8. Hence $12 \leq 24$ and $8 \leq 24$; $12 \leq 48$ and $8 \leq 48$; $12 \leq 48$ and $8 \leq 48$; $12 \leq 72$ and $8 \leq 72$. Thus, $24, 48, 72$ are upper bounds of $\{12, 8\}$ and hence 24 is the least upper bound of $\{12, 8\}$. Notice that $24 \notin \{12, 8\}$.

Theorem 4.3.3. In a poset $(S, \leq)$, if a subset $\{a, b\}$ of S has a lub, then it is unique.

Proof. Let $a, b \in S$ and a lub of $\{a, b\}$ exists. Suppose $c, d \in S$ are two lubs of $\{a, b\}$. Then c and d are upper bounds of $\{a, b\}$. Since c is a lub of $\{a, b\}$ and d is an upper bound, so $c \leq d$. Similarly, $d \leq c$. Then we have $c \leq d$ and $d \leq c$. By antisymmetry, we can say that $c = d$. Hence the result. □

The lub of $\{a, b\}$ in $(S, \leq)$, if it exists, is denoted by $a \vee b$, or the "join" of a and b.

Definition 4.3.4. Let $(S, \leq)$ be a poset and let $\{a, b\}$ be a subset of S. An element $c \in S$ is called a lower bound of $\{a, b\}$ if $c \leq a$ and $c \leq b$. Also, if T is any subset of S, then $c \in S$ is called an lower bound of T if $c \leq t$ for all $t \in T$.

An element $d \in S$ is called **greatest lower bound (glb)** of $\{a, b\}$ if,

1. d is a lower bound of $\{a, b\}$; and
2. if $c \in S$ is a lower bound of $\{a, b\}$, then $c \leq d$.

We can also define the glb of any general subset T of S and denote it by $\inf T$.

Then similar to the previous theorem, we can prove the following

Theorem 4.3.5. In a poset $(S, \leq)$, if a subset $\{a, b\}$ of S has a glb, then it is unique.

Proof. Left as an exercise. □

The glb of $\{a, b\}$ in $(S, \leq)$, if it exists, is denoted by $a \wedge b$, or the "meet" of a and b.

Definition 4.3.6. A poset $(L, \leq)$ is called a **lattice** if both $a \vee b$ and $a \wedge b$ exist for all $a, b \in L$. A lattice L is called **complete** if each of its subsets has a lub and glb in L.

Example 4.3.7. Any chain is a lattice in which $a \wedge b$ is simply the smaller of a and b and $a \vee b$ is simply the bigger of the two. Not every lattice is complete; the rational numbers are not complete with respect to the "usual less than or equal to relation", and the real numbers (in their natural order) are also not complete unless $-\infty$ and ∞ are adjoined to it.

Example 4.3.8. Let L be the set of all nonnegative real numbers. Then $(L, \leq)$ is a poset, where $\leq$ denotes the usual "less than or equal to" relation. Let $a, b \in L$. Now, $\max\{a, b\} \in L$ and $\min\{a, b\} \in L$. It is easy to see that $\max\{a, b\}$ is the lub of $\{a, b\}$ and $\min\{a, b\}$ is the glb of $\{a, b\}$. For example, $\max\{2, 5\} = 5 = 2 \vee 5$ and $\min\{2, 5\} = 2 = 2 \wedge 5$. Hence $(L, \leq)$ is a lattice. But it is not complete as we have discussed in the previous example.

Example 4.3.9. Let S be a set. Then $(\mathcal{P}(S), \leq)$ is a poset, where $\leq$ is the set inclusion relation. For $A, B \in \mathcal{P}(S)$, we can show that $A \vee B = A \cup B$ and $A \wedge B = A \cap B$. Hence $(\mathcal{P}(S), \leq)$ is a lattice. This lattice is however, complete and the glb of any family A of subsets of S is simply $\bigcap_A A_\alpha$ and the lub is $\bigcup_A A_\alpha$, both of which belong to $\mathcal{P}(S)$.

Theorem 4.3.10. Let $(L, \leq)$ be a lattice and let $a, b \in L$. Then

L1. $a \vee b = b \vee a$, $a \wedge b = b \wedge a$ (commutative laws),

L2. $a \vee (b \vee c) = (a \vee b) \vee c$, $a \wedge (b \wedge c) = (a \wedge b) \wedge c$ (associative laws),

L3. $a \vee a = a$, $a \wedge a = a$ (idempotent laws),

L4. $a \vee (a \wedge b) = a$, $a \wedge (a \vee b) = a$ (absorption laws).

Proof. Left as an exercise. □

Theorem 4.3.11. Let $(S, \leq)$ be a poset and $a, b \in S$. Then the following conditions are equivalent:

1. $a \leq b$;
2. $a \vee b = b$;
3. $a \wedge b = a$.

This is known as the **consistency** of the poset.

Proof. Left as exercise. □

Theorem 4.3.12. In any lattice $(L, \leq)$, the operations of join and meet are **isotonic**, that is, if $b \leq c$, then

$$a \wedge b \leq a \wedge c \quad \text{and} \quad a \vee b \leq a \vee c.$$

Proof. Let $b \leq c$. Then

$$a \wedge b = (a \wedge a) \wedge (b \wedge c) = (a \wedge b) \wedge (a \wedge c),$$

whence $a \wedge b \leq a \wedge c$ by consistency. Similarly, the other inequality can be shown. □

Theorem 4.3.13. In any lattice $(L, \leq)$, we have the distributive inequalities

$$\begin{array}{lrcl} \text{D} & (a \wedge b) \vee (a \wedge c) & \leq & a \wedge (b \vee c), \\ \text{D'} & a \vee (b \wedge c) & \leq & (a \vee b) \wedge (a \vee c), \end{array}$$

for all $a, b, c \in L$.

Proof. Clearly, $a \wedge b \leq a$ and $a \wedge b \leq b \leq b \vee c$. Hence, $a \wedge b \leq a \wedge (b \vee c)$. Also, $a \wedge c \leq a$, $a \wedge c \leq c \leq b \vee c$. Hence, $a \wedge c \leq a \wedge (b \vee c)$. That is, $a \wedge (b \vee c)$ is an upper bound of $a \wedge b$ and $a \wedge c$, from which, D follows. Similarly, we can prove D'. □

Definition 4.3.14. A lattice $(L, \leq)$ is called **distributive** if it satisfies

D1. $a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c)$,

for all $a, b, c \in L$.

The two previous examples of lattices that we discussed earlier, were both distributive lattices. However, it is worth mentioning that all lattices are not distributive as we see in the following example.

Example 4.3.15. Consider the lattice in the figure 4.3.1.
Since $a \wedge (b \vee c) = a \wedge 1 = a \neq 0 = 0 \vee 0 = (a \wedge b) \vee (a \wedge c)$, so this lattice is not distributive.

The next theorem gives a necessary and sufficient condition for a lattice to be distributive.

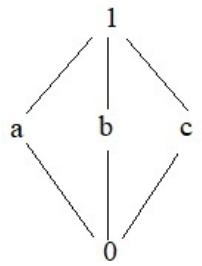


Figure 4.3.1

Theorem 4.3.16. A lattice $(L, \leq)$ is distributive if and only if

$$\text{D2.}\quad a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c),$$

for all $a, b, c \in L$.

Proof. Suppose $(L, \leq)$ is distributive. Let $a, b, c \in L$. Then

$$\begin{aligned}
(a \vee b) \wedge (a \vee c) &= ((a \vee b) \wedge a) \vee ((a \vee b) \wedge c) && \text{by D1}\\
&= (a \wedge (a \vee b)) \vee ((a \vee b) \wedge c) && \text{by L1}\\
&= a \vee ((a \vee b) \wedge c) && \text{by L4}\\
&= a \vee (c \wedge (a \vee b)) && \text{by L1}\\
&= a \vee ((c \wedge a) \vee (c \wedge b)) && \text{by D1}\\
&= (a \vee (c \wedge a)) \vee (c \wedge b) && \text{by L2}\\
&= (a \vee (c \wedge a)) \vee (b \wedge c) && \text{by L1}\\
&= a \vee (b \wedge c) && \text{by L4.}
\end{aligned}$$

Hence, $a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c)$. Similarly, we can show that D2⇒D1. □

Theorem 4.3.17. In a distributive lattice $(L, \leq)$,

$$a \wedge b = a \wedge c \quad \text{and} \quad a \vee b = a \vee c \Rightarrow b = c,$$

for all $a, b, c \in L$.

Proof. Let $(L, \leq)$ be a distributive lattice. Now,

$$\begin{aligned}
b &= b \wedge (a \vee b)\\
&= b \wedge (a \vee c)\\
&= (b \wedge a) \vee (b \wedge c)\\
&= (a \wedge c) \vee (b \wedge c)\\
&= (c \wedge a) \vee (c \wedge b)\\
&= c \wedge (a \vee b)\\
&= c \wedge (a \vee c)\\
&= c.
\end{aligned}$$

□

Note that a poset $(L, \leq)$ may not contain a greatest element, but from the antisymmetric property of $\leq$, it can be shown that if there exists a greatest element in a poset, then it is unique, for if, a and b are two such elements, then $a \leq b$ and by the same argument, $b \leq a$, which implies that $a = b$. Similarly, a poset may contain at most one least element. We denote the greatest element of L by I and the least element by O. The elements O and I, when they exist, are called the **universal bounds** of L, since then $O \leq x$ and $x \leq I$ for all $x \in L$.

Theorem 4.3.18. If $(L, \leq)$ is a poset having O and I, then

$$\begin{aligned} O \wedge x &= O \quad \text{and} \quad O \vee x = x, \\ x \wedge I &= x \quad \text{and} \quad x \vee I = I, \end{aligned}$$

for all $x \in L$.

Proof. Left as exercise. □

Theorem 4.3.19. Let $(L, \leq)$ be a lattice. Then for all $a, b, c \in L$,

$$a \leq c \Rightarrow a \vee (b \wedge c) \leq (a \vee b) \wedge c.$$

This is called modular inequality.

Proof. We have, $a \leq a \vee b$ and $a \leq c$. Hence, $a \leq (a \vee b) \wedge c$. Also, $b \wedge c \leq b \leq a \vee b$ and $b \wedge c \leq c$. Thus, $b \wedge c \leq (a \vee b) \wedge c$. Thus, combining, we get the desired result. □

Definition 4.3.20. Let $(L, \leq)$ be a lattice with I and O. If $a \in L$, then an element $b \in L$ is said to be a **complement** of a if $a \vee b = I$ and $a \wedge b = O$.

Example 4.3.21. Let D_{30} denote the set of all positive divisors of 30. Then

$$D_{30} = \{1, 2, 3, 5, 6, 10, 15, 30\}.$$

Now, $(D_{30}, \leq)$ is a poset, where $a \leq b$ if and only if a divides b. Since 1 divides all the elements of D_{30}, it follows that $1 \leq m$ for all $m \in D_{30}$. Thus, 1 is the least element of this poset. Again, every member of D_{30} divides 30. Thus, $m \leq 30$. Hence, 30 is the greatest element of this poset. The Hasse diagram of this poset is given by fig. 4.3.2.

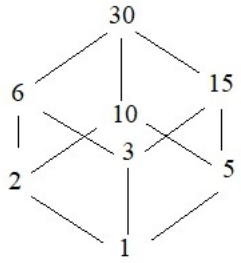


Figure 4.3.2

Let $a, b \in D_{30}$. Let $d = \gcd(a, b)$ and $m = \text{lcm}(a, b)$. Now, $d|a$ and $d|b$. Hence, $d \leq a$ and $d \leq b$. This shows that d is a lower bound of $\{a, b\}$. Let $c \in D_{30}$ and $c \leq a$ and $c \leq b$. Then $c|a$ and $c|b$ and since d is the gcd of a and b, so $c|d$, and hence $c \leq d$. Thus, $d = \gcd(a, b) = \text{glb}\{a, b\}$. Since all the positive divisors

of a, b are also divisors of 30, $d \in D_{30}$, so $d = a \wedge b$. Similarly we can show that $m \in D_{30}$ and $m = a \vee b$. Hence D_{30} is a complete lattice with least element 1 and greatest element 30.

Now, for any $a \in D_{30}$, $\frac{30}{a} \in D_{30}$. Using properties of gcd and lcm, we can show that for any $a \in D_{30}$,

$$a \wedge \frac{30}{a} = 1 \quad \text{and} \quad a \vee \frac{30}{a} = 30.$$

Hence, every element a has a complement $\frac{30}{a}$ in D_{30}.

Note that for any positive integer n, we can construct the lattice $(D_n, \leq)$, where $\leq$ denotes the usual divisibility relation in a similar way as shown in the preceding example.

Theorem 4.3.22. In a distributive lattice $(L, \leq)$ with I and O, every element has at most one complement.

Proof. Let $a \in L$. Suppose b, c are two complements of a in L. Then $a \vee b = I$ and $a \wedge b = O$; $a \vee c = I$ and $a \wedge c = O$. Hence $a \vee b = a \vee c$ and $a \wedge b = a \wedge c$. Then by theorem 4.3.17, it follows that $b = c$. Hence the result. □

A special type of distributive lattice is the **Boolean Algebra**. We will read about it in the next unit.

4.4 Sublattice

Definition 4.4.1. A **sublattice** of a lattice L is a subset X of L such that $a, b \in X$ imply $a \wedge b \in X$ and $a \vee b \in X$.

A sublattice is a lattice in its own right with the same join and meet operations. The empty set is a sublattice; so is any one-element subset. More generally, given $a \leq b$ in a lattice L, the interval $[a, b]$ of all elements $x \in L$ such that $a \leq x$ and $x \leq b$ is a sublattice.

A subset of a lattice L can be itself under the same (relative) order without being a lattice. Let us check the following example.

Example 4.4.2. Let Σ consist of the subgroups of a group G and let $\leq$ be the usual set inclusion relation. Then Σ is a complete lattice with $H \wedge K = H \cap K$ and $H \vee K$ the least subgroup in Σ containing H and K (which is not their set-theoretic union). Here, the set-union of two non-comparable subgroups is never a subgroup (since we know that the union of two subgroups H and K is a subgroup if and only if either $H \leq K$ or $K \leq H$). Hence this lattice is not a sublattice of the lattice of all subsets of G.

Definition 4.4.3. A property of subsets of a set I is a **closure property** when

1. I has the property, and
2. any intersection of subsets having the given property itself has this property.

Theorem 4.4.4. Let L be any complete lattice and let S be any subset of L such that

1. $I \in S$, and
2. $T \subset S$ implies $\inf T \in S$.

Then S is a complete lattice.

Proof. For any non-empty subset T of S, evidently $\inf T \in L$ is a member of S by 2, and it is the glb of T in S. Also, let U be the set of all upper bounds of T in S. It is non-empty since $I \in S$. Then, $\inf U \in S$ is also an upper bound of T. Moreover, it is the least upper bound since $\inf U \leq u$ for all $u \in U$. This proves that S is a complete lattice. □

Corollary 4.4.5. Those subsets of any set which have a given closure property form a complete lattice, in which the lattice meet of any family of subsets S_α is their intersection, and their lattice join is the intersection of all subsets T_β which contain every S_α.

4.5 Direct Products

Besides occurring naturally, new lattices can also be constructed from given ones by various processes. One such process consists in forming direct products.

Definition 4.5.1. The direct product PQ of two posets P and Q is the set of all ordered pairs (a, b) with $a \in P$ and $b \in Q$, partially ordered by the rule $(a_1, b_1) \leq (a_2, b_2)$ if and only if $a_1 \leq_P a_2$ in P and $b_1 \leq_Q b_2$ in Q.

Theorem 4.5.2. The direct product LM of any two lattices is a lattice.

Proof. For any two elements (a_1, b_1) and (a_2, b_2) in LM, the element $(a_1 \vee a_2, b_1 \vee b_2)$ (here we have taken the join operations in all L, M and LM as $\vee$) contains both (a_1, b_1) and (a_2, b_2), hence is an upper bound for the pair. Moreover every other upper bound (u, v) of the two satisfies $a_1 \leq u$ and $a_2 \leq u$ and hence by the definition of lub, $a_1 \vee a_2 \leq u$. Similarly, $b_1 \vee b_2 \leq v$, and so, $(a_1 \vee a_2, b_1 \vee b_2) \leq (u, v)$. This shows that

$$(a_1 \vee a_2, b_1 \vee b_2) = (a_1, b_1) \vee (a_2, b_2),$$

if the latter exists. By a similar argument for lower bound, we can show that

$$(a_1 \wedge a_2, b_1 \wedge b_2) = (a_1, b_1) \wedge (a_2, b_2),$$

if the latter exists. This shows that LM is a lattice. □

4.6 Few Probable Questions

1. Define poset. Show that every non-empty finite set has a minimal element.
2. Define lattice. Deduce the modular inequality of a lattice.
3. Deduce the distributive inequality of a lattice.
4. Deduce the necessary and sufficient condition for a lattice to be distributive.
5. Define direct product of lattice. Show that the direct product is two lattice is a lattice.
6. Draw the Hasse diagram of D_{36} with respect to the usual divisibility relation and show that it is a lattice. Also, find the complement of each of the elements of D_{36}, if it exists.

Unit 5

Course Structure

- Boolean Algebra: Basic Definitions, Duality, Basic theorems,
- Boolean algebra as lattices.

5.1 Introduction

In mathematics and mathematical logic, Boolean algebra is the branch of algebra in which the values of the variables are the truth values true and false, usually denoted 1 and 0 respectively. Instead of elementary algebra where the values of the variables are numbers, and the prime operations are addition and multiplication, the main operations of Boolean algebra are the meet (and) denoted as $\wedge$, the join (or) denoted as $\vee$, and the negation (not) denoted as $\neg$. It is thus a formalism for describing logical operations in the same way that elementary algebra describes numerical operations.

Boolean algebra was introduced by George Boole in his first book The Mathematical Analysis of Logic (1847), and set forth more fully in his An Investigation of the Laws of Thought (1854). According to Huntington, the term "Boolean algebra" was first suggested by Sheffer in 1913, although Charles Sanders Peirce in 1880 gave the title "A Boolian Algebra with One Constant" to the first chapter of his "The Simplest Mathematics". Boolean algebra has been fundamental in the development of digital electronics, and is provided for in all modern programming languages. It is also used in set theory and statistics.

Objectives

After reading this unit, you will be able to:

- define Boolean algebra and derive some useful properties of it
- establish a partial order relation on a Boolean algebra
- deduce that an Boolean algebra is a lattice with respect to the partial order defined

5.2 Boolean Algebra

Though we gave a rough idea about Boolean Algebra in the previous unit, we start afresh in this unit to define Boolean Algebra.

Definition 5.2.1. A class of elements B together with two binary operations $(+)$ and $(\cdot)$ (where $a \cdot b$ will be written as ab) is a Boolean algebra if and only if it satisfies the following postulates:

B1. The operations $(+)$ and $(\cdot)$ are commutative.

B2. There exist in B distinct identity elements 0 and 1 relative to the operations $(+)$ and $(\cdot)$ respectively.

B3. Each operation is distributive over the other.

B4. For every $a \in B$, there exists an element a' in B such that

$$a + a' = 1 \quad \text{and} \quad aa' = 0.$$

The symbols " $+$ " and " $\cdot$ " is just a convention. We could use any other symbols in place of these two.

Example 5.2.2. Let S be any set and $\mathcal{P}(S)$ be the set of all subsets of S. Then $\mathcal{P}(S)$ forms a Boolean algebra where the binary operations $(+)$ and $(\cdot)$ are the set-theoretic union and intersections respectively. The corresponding identity elements are S and $\emptyset$ respectively. For every element $T \in \mathcal{P}(S)$, the complement is given by $S \setminus T$.

Theorem 5.2.3. Every statement or algebraic identity deducible from the postulates of a Boolean algebra remains valid if the operations $(+)$ and $(\cdot)$, and the identity elements 0 and 1 are interchanged throughout.

This theorem is called the principle of duality.

Proof. The proof of this theorem follows at once from the symmetry of the postulates with respect to the two operations and the two identities. □

It should be noted that the steps in one proof are dual statements to those in the other, and the justification for each step is the same postulate or theorem in one case as in the other.

Theorem 5.2.4. For every element a in a Boolean algebra B,

$$a + a = a \quad \text{and} \quad aa = a.$$

Proof.

$$\begin{aligned} a &= a + 0 && \text{by B2} \\ &= a + aa' && \text{by B4} \\ &= (a + a)(a + a') && \text{by B3} \\ &= (a + a)(1) && \text{by B4} \\ &= a + a, && \text{by B2} \end{aligned}$$

and similarly,

$$\begin{aligned} a &= a(1) && \text{by B2} \\ &= a(a + a') && \text{by B4} \\ &= aa + aa' && \text{by B3} \\ &= aa + 0 && \text{by B4} \\ &= aa. && \text{by B2} \end{aligned}$$

□

Thus, we can say that $(+)$ and $(\cdot)$ operations are idempotent.

Theorem 5.2.5. For every element a in a Boolean algebra B,

$$a + 1 = 1 \quad \text{and} \quad a0 = 0.$$

Proof.

$$\begin{aligned} 1 &= a + a' && \text{by B4} \\ &= a + a'(1) && \text{by B2} \\ &= (a + a')(a + 1) && \text{by B3} \\ &= 1(a + 1) && \text{by B4} \\ &= a + 1. && \text{by B2} \end{aligned}$$

The other part is left as an exercise. □

Theorem 5.2.6. For each pair of elements a and b in a Boolean algebra B,

$$a + ab = a \quad \text{and} \quad a(a + b) = a.$$

Proof.

$$\begin{aligned} a &= 1a && \text{by B2} \\ &= (1 + b)a && \text{by Theorem 5.2.5} \\ &= 1a + ba && \text{by B3 and B1} \\ &= a + ba && \text{by B2} \\ &= a + ab. && \text{by B1} \end{aligned}$$

The other part is left as exercise. □

Theorem 5.2.7. In every Boolean algebra B, each of the binary operations $(+)$ and $(\cdot)$ is associative. That is, for every a, b, and c in B,

$$a + (b + c) = (a + b) + c \quad \text{and} \quad a(bc) = (ab)c.$$

Proof. First we will show that $a + a(bc) = a + (ab)c$, as follows:

$$\begin{aligned} a + a(bc) &= a && \text{by Theorem 5.2.6} \\ &= a(a + c) && \text{by Theorem 5.2.6} \\ &= (a + ab)(a + c) && \text{by Theorem 5.2.6} \\ &= a + (ab)c. && \text{by B3} \end{aligned}$$

Next we will show that $a' + a(bc) = a' + (ab)c$, as follows:

$$\begin{aligned} a' + a(bc) &= (a' + a)(a' + bc) && \text{by B3} \\ &= 1(a' + bc) && \text{by B4} \\ &= a' + bc && \text{by B2} \\ &= (a' + b)(a' + c) && \text{by B3} \\ &= [1(a' + b)](a' + c) && \text{by B2} \\ &= [(a' + a)(a' + b)](a' + c) && \text{by B4} \\ &= (a' + ab)(a' + c) && \text{by B3} \\ &= a' + (ab)c. && \text{by B3} \end{aligned}$$

Now if we multiply these two equations, we obtain

$$[a + a(bc)][a' + a(bc)] = [a + (ab)c][a' + (ab)c]. \tag{5.2.1}$$

The left side of the above equation may be reduced as follows

$$\begin{aligned} [a + a(bc)][a' + a(bc)] &= [a(bc) + a][a(bc) + a'] && \text{by B1} \\ &= a(bc) + aa' && \text{by B3} \\ &= a(bc) + 0 && \text{by B4} \\ &= a(bc). && \text{by B2} \end{aligned}$$

Similarly, the right side of equation (5.2.1) reduces as follows:

$$\begin{aligned} [a + (ab)c][a' + (ab)c] &= [(ab)c + a][(ab)c + a'] && \text{by B1} \\ &= (ab)c + aa' && \text{by B3} \\ &= (ab)c + 0 && \text{by B4} \\ &= (ab)c. && \text{by B2} \end{aligned}$$

Thus, equation (5.2.1) reduces to

$$a(bc) = (ab)c,$$

which is the required associative law we were to prove. By duality principle, the analogous part for $(+)$ follows. □

From now on, we shall write both $a(bc)$ and $(ab)c$ as abc, and similarly, we shall write both $(a+b)+c$ and $a+(b+c)$ as $a+b+c$.

Theorem 5.2.8. The element a' associated with the element a in a Boolean algebra is unique.

Proof. Suppose that $a + x = 1$, $ax = 0$, and also that $a + y = 1$, $ay = 0$. Then,

$$\begin{aligned} x &= 1.x && \text{by B2} \\ &= (a + y)x && \text{by assumption} \\ &= (ax + yx) && \text{by B3 and B1} \\ &= 0 + yx && \text{by assumption} \\ &= yx && \text{by B2} \\ &= xy && \text{by B1} \\ &= xy + 0 && \text{by B2} \\ &= xy + ay && \text{by assumption} \\ &= (x + a)y && \text{by B3 and B1} \\ &= 1y && \text{by assumption} \\ &= y. && \text{by B2} \end{aligned}$$

□

Thus any two elements associated with a as specified in B4 are equal. In other words, a' is uniquely determined by a. We will refer to a' as the complement of a.

Theorem 5.2.9. For every a in a Boolean algebra B, $(a')' = a$.

Proof. By B4, $a + a' = 1$ and $aa' = 0$. But this is exactly the necessary condition that $(a')'$ is equal to a. By the previous theorem, this is unique and hence the result. □

Theorem 5.2.10. S. In any Boolean algebra, $0' = 1$ and $1' = 0$.

Proof. By theorem 5.2.5, $1 + 0 = 1$, and $1 \cdot 0 = 0$. Since theorem 5.2.8 shows that for each a there is only one element a', these equations imply that $0' = 1$, and $1' = 0$. □

Theorem 5.2.11. For every a and b in a Boolean algebra B,

$$(ab)' = a' + b' \quad \text{and} \quad (a + b)' = a'b'.$$

Proof. First,

$$\begin{aligned} (ab)(a' + b') &= aba' + abb' && \text{by B3} \\ &= 0b + a0 && \text{by B1, B2, B4} \\ &= 0 + 0 = 0. && \text{by theorem 5.2.5} \end{aligned}$$

Further,

$$\begin{aligned} ab + a' + b' &= a' + b' + ab && \text{by B1} \\ &= (a' + b' + a)(a' + b' + b) && \text{by B3} \\ &= (1 + b')(1 + a') && \text{by B4 and B1} \\ &= 1. && \text{by theorem 5.2.5 and B2} \end{aligned}$$

Now, by B4 and theorem 5.2.8, we can show that $(ab)' = a' + b'$. The part can be shown by duality principle. □

This is known as D'Morgan's law.

5.3 Boolean Algebra as Lattices

We now define an order relation on a Boolean algebra B by the following.

Definition 5.3.1. The "order" relation $a \leq b$ is defined by the statement:
For every a and b in a Boolean algebra B, $a \leq b$ if and only if $ab' = 0$.

Let us see certain properties of the relation as follows:

Theorem 5.3.2. The following four properties of $\leq$ are valid in every Boolean algebra for arbitrary elements x, y, and z:

1. $x \leq x$ (reflexive);
2. if $x \leq y$ and $y \leq x$, then $x = y$ (antisymmetry);
3. if $x \leq y$ and $y \leq z$, then $x \leq z$ (transitive);
4. if $x \leq y$ and $x \leq z$, then $x \leq yz$;
5. if $x \leq y$, then $x \leq y + z$, for any z;
6. $x \leq y$ if and only if $y' \leq x'$.

Proof. 1. Left for reader.

2. $x \leq y$ and $y \leq z$ are equivalent to $xy' = 0$ and $yx' = 0$ respectively. Now,

$$\begin{aligned} x &= x(1) && \text{by B2} \\ &= x(y + y') && \text{by B4} \\ &= xy + xy' && \text{by B3} \\ &= xy && \text{by assumption} \\ &= yx && \text{by B1} \\ &= yx + yx' && \text{by B2 and assumption} \\ &= y(x + x') && \text{by B3} \\ &= y(1) = y. && \text{by B4} \end{aligned}$$

3. $x \leq y$ is equivalent to $xy' = 0$. Also, $y \leq z$ is equivalent to $yz' = 0$. Now,

$$\begin{aligned} xz' &= xz'(1) && \text{by B2} \\ &= xz'(y + y') && \text{by B4} \\ &= xyz' + xy'z' && \text{by B1 and associativity} \\ &= 0 + 0. && \text{by assumption} \end{aligned}$$

Thus, $x \leq z$.

4. $x \leq y$ and $x \leq z$ are equivalent to $xy' = 0$ and $xz' = 0$ respectively. Now,

$$\begin{aligned} x(yz)' &= x(y' + z') && \text{by theorem 5.2.11} \\ &= xy' + xz' && \text{by B3} \\ &= 0. && \text{by assumption} \end{aligned}$$

Hence $x \leq yz$.

5. $x \leq y$ is equivalent to $xy' = 0$. Let $z \in B$ be arbitrary. Then

$$\begin{aligned} x(y + z)' &= x(y'z') && \text{by theorem 5.2.11} \\ &= 0. && \text{by associativity and assumption} \end{aligned}$$

Thus, $x \leq y + z$ for any $z \in B$.

6. $x \leq y$ is equivalent to $xy' = 0$. Thus,

$$\begin{aligned} y'(x')' &= y'x && \text{by theorem 5.2.9} \\ &= xy' && \text{by B1} \\ &= 0. && \text{by assumption} \end{aligned}$$

Hence, $y' \leq x'$.

□

The first three points of the above theorem show that B forms a poset with respect to the relation $\leq$ defined above. We will show that Boolean algebra forms lattice with respect to the defined partial order.

Theorem 5.3.3. Let B be a Boolean algebra with respect to the partial order $\leq$ defined as $x \leq y$ if and only if $xy' = 0$. Then B is a lattice with respect to $\leq$.

Proof. We will be done if we show that $\{x, y\}$ has lub and glb in B. We show that $x + y$ is the lub and xy is the glb of the set. Since $x(x + y)' = x(x'y') = xx'y' = 0$ and similarly, $y(x + y)' = 0$ so $x \leq (x + y)$ and $y \leq (x + y)$. Thus, $x + y$ is an upper bound of $\{x, y\}$. Let z be any other upper bound of $\{x, y\}$. Then $x \leq z$ and $y \leq z$ which imply $xz' = 0$ and $yz' = 0$. Now,

$$(x + y)z' = xz' + yz' = 0$$

which shows that $x + y \leq z$. Thus, $x + y$ is the lub of $\{x, y\}$. We can similarly show that xy is the glb of $\{x, y\}$. Thus, $(B, \leq)$ forms a lattice. □

The join and meet are defined as $x \vee y = x + y$ and $x \wedge y = xy$, for any arbitrary $x, y \in B$.

Also, note from the previous theorems that a Boolean algebra is distributive, and each element of it has a complement. Thus, a Boolean algebra is a distributive complemented lattice. Let us see certain examples.

Example 5.3.4. Let B be a Boolean algebra. We simplify the expression $x + (yx)'$, where $x, y \in B$.

We have,

$$\begin{aligned} x + (yx)' &= x + (y' + x') && \text{by theorem 5.2.11} \\ &= (x + x') + y' && \text{by B1} \\ &= 1 + y' && \text{by B4} \\ &= (0y)' = 0' = 1. && \text{by theorems 5.2.11 and 5.2.10} \end{aligned}$$

Example 5.3.5. In a Boolean algebra B, we simplify $(xy)'(x' + y)(y' + y)$, for $x, y \in B$.

We have,

$$\begin{aligned} (xy)'(x' + y)(y' + y) &= (xy)'(x' + y) && \text{by B4 and B2} \\ &= (x' + y')(x' + y) && \text{by theorem 5.2.11} \\ &= x' + y'y && \text{by theorem 5.2.7} \\ &= x'. && \text{by B4} \end{aligned}$$

Example 5.3.6. In a Boolean algebra B, we simplify $(x + z)(xt + xt') + xz + z$, for $x, z, t \in B$.

We have,

$$\begin{aligned} (x + z)(xt + xt') + xz + z &= (x + z)x(t + t') + xz + z \\ &= (x + z)x + xz + z \\ &= x((x + z) + z) + z \\ &= x(x + z) \\ &= xx + xz + z \\ &= x + (x + 1)z \\ &= x + z. \end{aligned}$$

Example 5.3.7. In a Boolean algebra B, we simplify $x'(x + y) + (y + xx)(x + y')$, for $x, y \in B$.

We have,

$$\begin{aligned} x'(x+y)+(y+xx)(x+y') &= x'x+x'y+(y+x)x+(y+x)y' \\ &= x'y+(y+x)x+(y+x)y' \\ &= x'y+yx+xx+yy'+xy' \\ &= x'y+yx+x+xy' \\ &= x'y+x(y+1+y') \\ &= x'y+x \\ &= x+x'y \\ &= (x+x')(x+y) \\ &= x+y. \end{aligned}$$

5.4 Few Probable Questions

1. Establish the distributive property of Boolean algebra.
2. Define a partial order relation on a Boolean algebra B. Hence show that it is a lattice with respect to the defined partial order.
3. Deduce the De'Morgan's law for Boolean algebra.
4. Show that the complement of an element in a Boolean algebra is always unique.
5. Show that $0' = 1$ and $1' = 0$ in a Boolean algebra.
6. In a Boolean algebra B, show that for any $a, b \in B$, $a(a+b) = a$.
7. Deduce the idempotent property of both the binary operators $(+)$ and $(\cdot)$ in a Boolean algebra.
8. In a Boolean algebra B, simplify the following:
   (a) $y(x'z+xz')+x(yz+yz')$;
   (b) $xyz+x'+xy'z$;
   (c) $(xy'+x'y)'(x+y)$;
   (d) $(xy)'(x'+y)(y'+y)$.

Unit 6

Course Structure

- Boolean Algebra: Boolean functions, Sum and Product of Boolean algebra,
- Minimal Boolean Expressions, Prime implicants
- Propositions and Truth tables

6.1 Introduction

This unit starts with the dnf and cnf which are normals forms and continuation of the previous unit. Next we move on to logic gates.

Logic is an extensive field of study with many special areas of inquiry. In general, logic is concerned with the study and analysis of methods of reasoning or argumentation. Symbolic logic is not precisely defined as distinct from logic in general, but might be described as a study of logic which employs an extensive use of symbols. In any discussion of logic, the treatment centers around the concept of a proposition (statement). The principal tool for treatment of propositions is the algebra of propositions, a Boolean algebra. In talking about propositions, we will also investigate certain logical forms which represent acceptable techniques for constructing precise proofs of theorems. Since statements are formed from words, it is apparent that some consideration must be given to words and their meanings. No logical argument can be based on words that are not precisely described. That part of logic which is concerned with the structure of statements is much more difficult than the areas mentioned previously, and in fact, has not been satisfactorily formalized.

Objectives

After reading this unit, you will be able to

- define disjunctive normal forms and deduce related results
- define conjunctive normal forms and deduce related results
- solve problems related to dnf and cnf
- define propositions and learn to form complex propositions by conjunction, disjunction and negation

- show that the set of all propositions form a Boolean algebra with respect to the conjunction, disjunction and negation so defined
- draw truth tables for complex propositions

6.2 Disjunctive Normal Form

We start assuming that the reader is familiar with the terms monomial, polynomial, terms, factor, variable constants. By a Boolean function we will mean any expression which represents the combination of a finite set of symbols, each representing a constant or a variable, by the operations of $(+)$, $(\cdot)$, or complement. Thus, $(a'+b)'c+ab'x+0$ is a Boolean function provided that each of the symbols a, b, c, x represents an element of a Boolean algebra. Further example such as the equation $x + x' = 1$ represents the statement that a function $x + x'$ of the variable x equals the constant 1.

Among the functions of n variables $x_1, x_2, \ldots, x_n$ which can be written, a particular class of functions is of special interest, namely, those written as a sum of terms in which each term is a product involving all n variables either with or without a prime. Examples of such functions are $x + x'$, xy', $xyz' + x'yzxy'z$ in one, two, and three variables, respectively. The following definition gives a name to such functions.

Definition 6.2.1. A Boolean function is said to be in disjunctive normal form in n variables $x_1x_2, \ldots, x_n$, for $n > 0$, if the function is a sum of terms of the type $f_1(x_1)f_2(x_2)\cdots f_n(x_n)$, where $f_i(x_i)$ is x_i, or x_i' for each $i = 1, 2, \ldots, n$, and no two terms are identical. In addition, 0 and 1 are said to be in disjunctive normal form in n variables for any $n \geq 0$.

Some important properties of the disjunctive normal form are given in the following theorems.

Theorem 6.2.2. Every function in a Boolean algebra which contains no constants is equal to a function in disjunctive normal form.

Proof. Let an arbitrary function (without constants) of the n variables $x_1, x_2, \ldots, x_n$ denoted by f. If f contains an expression of the form $(A+B)'$ or $(AB)'$ for some functions A and B, then D'Morgan's law may be applied to yield $A'B'$ and $A' + B'$ respectively. This process may be continued until each prime which appears applies only to a single variable x_i.

Next, by applying the distributive law of $(\cdot)$ over $(+)$, f can be reduced to a polynomial.

Now suppose some term t does not contain either x_i or x_i' for some variable x_i. This term may be multiplied by $x_i + x_i'$ without changing the function. Continuing this process for each missing variable in each of the terms in f will give an equivalent function whose terms contain x_j or x_j' for each $j = 1, 2, \ldots, n$.

Finally by idempotent property, duplicate terms are eliminated and this completes the proof. □

The following is an illustration.

Example 6.2.3. Write the function $f = (xy' + xz)' + x'$ in disjunctive normal form.

We have,

$$\begin{aligned}
(xy' + xz)' + x' &= (xy')'(xz)' + x' \\
&= (x' + y)(x' + z') + x' \\
&= x' + x'y + yz' + x' \\
&= x'(x + y')(z + z') + yz'(x + x') \\
&= x'yz + x'yz' + x'y'z + x'y'z' + xyz' + x'yz' \\
&= x'y'z + xyz' + x'yz' + x'y'z + x'y'z'.
\end{aligned}$$

The usefulness of the normal form lies primarily in the fact that each function uniquely determines a normal form in a given number of variables, as we shall see in later theorems. However, any function may be placed in normal form in more than one way by changing the number of variables. For example, $f = xy$ is in normal form in x and y, but if xy is multiplied by $z + z'$, then $f = xyz + xyz'$ also in normal form in the variables x, y, and z. Similarly, $g = x'yz + xyz + x'yz' - xyz'$ is in normal form in x, y, and z, but reduces, on factoring, to $g = x'y - xy$, which is in normal form in x and y. From now on we shall assume that unless stated otherwise, disjunctive normal form refers to that disjunctive normal form which contains the smallest possible number of variables. With this exception, we will be able to show that the normal form of a function is uniquely determined by the function.

Suppose that we desire to select a single term out of the possible terms in a disjunctive normal form in n variables. This corresponds to selecting either x_i or x_i', for each of the n variables x_i, $i = 1, 2, \ldots, n$. Thus there are exactly 2^n distinct terms which may occur in a normal form in n variables.

Theorem 6.2.4. That disjunctive normal form in n variables which contains 2^n terms is called the complete disjunctive normal form in n variables.

It will be a consequence of the following theorems that the complete disjunctive normal form is identically 1. A simple argument to prove this directly is to note that for any variable x_j, the coefficients of x_j and x_j' must be identical in a complete normal form, namely, these coefficients are each the complete normal form in the remaining $n - 1$ variables. Factoring serves to eliminate x_j, and this process may be repeated to eliminate each variable in succession, thus reducing the expression to 1.

Theorem 6.2.5. If each of n variables is assigned the value 0 or 1 in an arbitrary, but fixed manner, then exactly one term of the complete disjunctive normal form in the n variables will have the value 1 and all other terms will have the value 0.

Proof. Let $a_1, a_2, \ldots, a_n$ represent the values assigned to $x_1, x_2, \ldots, x_n$ in that order, where each a_i is 0 or 1. Select a term from the complete normal form as follows: use x_i if $a_i = 1$, and use x_i' if $a_i = 0$ for each x_i, $i = 1, 2, \ldots, n$. The term so selected is then a product of n ones, and hence is 1. All other terms in the complete normal form will contain at least one factor 0 and hence will be 0. □

Corollary 6.2.6. Two functions are equal if and only if their respective disjunctive normal forms contain the same terms.

Proof. Two functions with the same terms are obviously equal. Conversely, if two functions are equal, then they must have the same value for every choice of value for each variable. In particular, they assume the same value for each set of values 0 and 1 which may be assigned to the variables. By idempotent property, the combinations of values of 0 and 1 which, when assigned to the variables, make the function assume the value 1 uniquely determine the terms which are present in the normal form for the function. Hence both normal forms contain the same terms. □

Corollary 6.2.7. To establish any identity in Boolean algebra, it is sufficient to check the value of each function for all combinations of 0 and 1 which may be assigned to the variables.

We have seen in the preceding theorems that a function is completely determined by the values it assumes for each possible assignment of 0 and 1 to the respective variables. This suggests that functions could be conveniently specified by giving a table to represent such properties. In applications, particularly to the design of circuits, this is precisely the way in which Boolean functions are constructed. If such a table has been given, then the function, in disjunctive normal form, may be written down by inspection. For each set of conditions for which the function is to be 1, a corresponding term is included in the disjunctive normal form selected, as indicated in the proof of the idempotent property in the previous unit. The sum of these terms gives the function, although not necessarily in simplest form. The following example indicates this method.

Row	x	y	z	$f(x,y,z)$
1	1	1	1	0
2	1	1	0	1
3	1	0	1	1
4	1	0	0	0
5	0	1	1	0
6	0	1	0	0
7	0	0	1	1
8	0	0	0	0

Table 6.1

Example 6.2.8. Find and simplify the function $f(x, y, z)$ specified by table 6.1

Note that the table shows the value of f for each of the $2^3 = 8$ possible assignments of 0 and 1 to x, y, and z.

We observe that for the combinations represented by rows 2, 3 and 7 of the table, the function will have the value 1. Thus the disjunctive normal form of f will contain three terms. For 2, since the x variable is 1, y variable is 1 and z variable is zero, the term in f corresponding to this combination will be xyz' (note that the value if 1 by idempotent property). Similarly, for the terms in 3 and 7th rows, we get $xy'z$ and $x'y'z$ respectively (each giving values 1). Thus, summing these terms over, we get $f(x, y, z) = xyz' + xy'z + x'y'z$. We have

$$\begin{aligned} f(x,y,z) &= xyz' + xy'z + x'y'z \\ &= xyz' + (x + x')y'z \\ &= xyz' + y'z. \end{aligned}$$

Exercise 6.2.9. 1. Express the following in disjunctive normal form in the smallest possible number of variables:

(a) $x'yz + xy'z' + x'y'z + x'yz' + xy'z + x'y'z'$

(b) $(x + y')(y + z')(z + x')(x' + y')$

(c) $(u + v + w)(uv + u'w)'$

(d) $xy' + xz + xy$

(e) $xyz + (x + y)(x + z)$

(f) $x + x'y$

(g) $(x + y)(x + y')(x' + z)$

2. Write separately, and simplify, the three functions f_1, f_2 and f_3 as given in the table 6.3.

6.3 Conjunctive Normal Form

There are other normal forms, besides the disjunctive normal form, which are equally useful. One of these represents each function as a product of sums, rather than as a sum of products. If each statement in the preceding section were replaced by its dual, the resulting discussion would be a corresponding treatment of this second form called the conjunctive normal form. To make this clear, the definition and theorems are repeated here in their dual forms. No proofs are needed, of course, because of the principle of duality.

Row	x	y	z	f_1	f_2	f_3
1	1	1	1	0	0	1
2	1	1	0	1	1	1
3	1	0	1	0	1	0
4	1	0	0	1	0	0
5	0	1	1	0	0	0
6	0	1	0	0	1	0
7	0	0	1	0	1	1
8	0	0	0	0	0	1

Table 6.2

Definition 6.3.1. A Boolean function is said to be in conjunctive normal form in n variables $x_1, x_2, \ldots, x_n$ for $n > 0$, if the function is a product of factors of the type $f_1(x_1) + f_2(x_2) + \cdots + f_n(x_n)$, where $f + i(x_i)$ is x_i or x_i' for each $i = 1, 2, \ldots, n$, and no two factors are identical. In addition, 0 and 1 are said to be in conjunctive normal form in n variables for $n \geq 0$.

Theorem 6.3.2. Every function in a Boolean algebra which contains no constants is equal to a function in conjunctive normal form.

Example 6.3.3. Write the function $(xy' + xz)' + x'$ in conjunctive normal form.

The procedure is essentially dual to that of the disjunctive normal form that we saw in the previous section, although, depending on the initial form of the function, it may require more steps to perform the reduction in one case than in another. Here, after primes are removed from parentheses, the function is factored into linear factors and then extra variables are introduced as needed by adding, within each factor, products of the form ww'. The final step is to expand into linear factors again and remove like factors. The solution for this example is given by the steps below.

$$\begin{aligned}
(xy' + xz)' + x' &= (x' + y)(x' + z') + x' \\
&= (x' + x' + y)(x' + x' + z') \\
&= (x' + y)(x' + z') \\
&= (x' + y + zz')(x + z' + yy') \\
&= (x' + y + z)(x' + y + z')(x' + y + z')(x' + y' + z') \\
&= (x' + y + z)(x' + y + z')(x' + y' + z').
\end{aligned}$$

Definition 6.3.4. That conjunctive normal form in n variables which contains 2^n factors is called the complete conjunctive normal form in n variables.

Theorem 6.3.5. If each of n variables is assigned the value 0 or 1 in an arbitrary, but fixed manner, then exactly one factor of the complete conjunctive normal form in the n variables will have the value 0 and all other factors will have the value 1.

Note that to select the factor which will be 0 when a set of values $a_1, a_2, \ldots, a_n$ are assigned to $x_1, x_2, \ldots, x_n$ in that order, where each a_i is 0 or 1, we simply apply duality principle of that described in the previous section. x_i is selected if $a_i = 0$ and x_i' is selected if $a_i = 1$ for each $i = 1, 2, \ldots, n$. The proper factor is then the sum of these letters, each of which has value 0. All other factors have the value 1.

Corollary 6.3.6. Two functions, each expressed in conjunctive normal form in n variables, are equal if and only if they contain identical factors.

Row	x	y	z	$f(x,y,z)$
1	1	1	1	1
2	1	1	0	1
3	1	0	1	0
4	1	0	0	1
5	0	1	1	1
6	0	1	0	1
7	0	0	1	0
8	0	0	0	1

Example 6.3.7. Find and simplify the function $f(x,y,z)$ specified in the table above.

Observe that only two rows of the table show the value 0 for f. Corresponding to the third row, we see that x is 1, y is 0 and z is 1. So the corresponding factor will be $x' + y + z'$. Similarly, for the 7th row, we have the term as $x + y + z'$. Thus, we would have $f(x,y,z) = (x' + y + z')(x + y + z')$ which gives us,

$$\begin{aligned} f(x,y,z) &= (x' + y + z')(x + y + z') \\ &= x'y + x'z' + y + yz' + z'x + z'y \\ &= (x' + 1)y + z'(x' + y + x + y) \\ &= y + z'(x' + x + y) \\ &= y + z'(1 + y) \\ &= y + z'. \end{aligned}$$

In problems of this type, the disjunctive normal form would normally be used if the number of 1's is were less than the number of 0's in the f column, and the conjunctive normal form would be used if the number of 0's were less than the number of l's.

Again, as in the previous section, we can use the conjunctive normal form to find complements of functions written in this form by inspection. The complement of any function written in conjunctive normal form is that function whose factors are exactly those factors of the complete conjunctive normal form which are missing from the given function. For example, the complement of $(x + y')(x' + y)$ is $(x + y)(x' + y')$.

It may be desirable to change a function from one normal form to the other. This can be done more readily than by following the general procedure for converting a function to a particular form. An example will illustrate the method, which is based on the fact that $(f')' = f$.

Example 6.3.8. Find the conjunctive normal form for the function

$$f = xyz + x'yz + xy'z' + x'yz'.$$

We have,

$$\begin{aligned} f &= xyz + x'yz + xy'z' + x'yz' \\ &= [(xyz + x'yz + xy'z' + x'yz')']' \\ &= [(x' + y' + z')(x + y' +, z')(x' + y + z)(x + y' + z)]' \\ &= (x + y + z)(x' + y + z')(x + y + z')(x' + y' + z). \end{aligned}$$

Here, the first complement was taken with the aid of D'Morgan's law and the second complement was taken by the method discussed above. These steps could have been reversed, with the same results. A similar procedure will change a function from conjunctive normal form to disjunctive normal form.

Exercise 6.3.9. 1. Express each of the following in conjunctive normal form in the smallest possible number of variables:

(a) $xyz + (x + y)(x + z)$

(b) $(x'y + xyz' + xy'z + x'y'z't + t')'$

(c) $x'yz + xy'z' + x'y'z + x'yz' + xy'z + x'y'z'$

(d) $(x + y')(y + z')(z + x')(x' + y')$

(e) $(u + v + w)(uv + u'w)'$

(f) $xy' + xz + xy$

(g) $xyz + (x + y)(x + z)$

2. Change each of the following from disjunctive normal form to conjunctive normal form:

(a) $uv + u'v + u'v'$

(b) $abc + ab'c' + a'bc' + a'b'c + a'b'c'$

3. Change each of the following from conjunctive normal form to disjunctive normal form:

(a) $(x + y')(x' + y)(x' + y')$

(b) $(u + v + w)(u + v + w')(u + v + w)(u' + v + w')(u' + v' + w)(u' + v' + w')$

4. Write separately, and simplify, the four functions f_1, f_2, f_3 and f_4 as given in the table below. Use whichever normal form seems easier.

Row	x	y	z	f_1	f_2	f_3	f_4
1	1	1	1	1	0	0	1
2	1	1	0	0	1	1	1
3	1	0	1	1	0	0	1
4	1	0	0	1	0	1	0
5	0	1	1	1	0	1	1
6	0	1	0	1	0	1	1
7	0	0	1	0	1	0	1
8	0	0	0	1	0	0	0

Table 6.3

6.4 Propositions and definitions of symbols

In the algebra of sets, it is necessary to start with certain primitive concepts in the form of undefined terms. This is typical of any formal system and is true of the algebra of propositions as well. The terms **true**, **false**, and proposition will be taken here as undefined. Without any attempt to investigate the philosophical meaning of truth and falsehood, we will assume that the words true and false are attributes which apply to propositions. By a proposition, we will infer the content of meaning of any declarative sentence which is free of ambiguity

and which has the property that it is either true or false, but not both. The following examples are typical propositions:

3 is a prime number;

living creatures exist on the planet Venus.

Note that of these propositions, the first is known to be true, while the second is either true or false. In contrast to these, the following is not a proposition:

this statement you are reading is false.

We shall use lower case italic letters to represent propositions. Where no specific proposition is given, these will be called propositional variables and used to represent arbitrary propositions.

From any proposition, or set of propositions, other propositions may be formed. The simplest example is that of forming from the proposition p, the negation of p, denoted by $\neg p$ or p'. or example, suppose that p is the proposition

sleeping is pleasant.

has negation

sleeping is unpleasant.

Any two propositions p and q may be combined in various ways to form new propositions. To illustrate, let p be the proposition

ice is cold,

and let q be the proposition

blood is green.

These propositions may be combined by the connective **and** to form the proposition

ice is cold and blood is green.

This proposition is referred to as the **conjunction** of p and q. We will denote the conjunction of p and q by pq, and we will require that the proposition be true in those cases in which both p and q are true, and false in cases in which either one or both of p and q are false.

Another way in which the propositions in the preceding paragraph may be combined is indicated in the proposition

either ice is cold or blood is green.

This proposition is referred to as the **disjunction** of p and q. We will denote the disjunction of p and q by $p+q$ and is the proposition "either p or q or both". We will require that this proposition be true whenever either one of p and q or both are true, and false only when both are false.

It follows from our definitions that the negation of "p or q" is the proposition "not p and not q," which can also be stated "neither p nor q." Likewise, the negation of "p and q" is "either not p or not q. " That is, the laws of D'Morgan hold for propositions just as they do for sets. In symbolic form we have the following laws for propositions:

$$\begin{aligned}(p+q)' &= p'q' \\ (pq)' &= p'+q'.\end{aligned}$$

Example 6.4.1. Let p be the proposition "missiles are costly" and q be the proposition "Grandpa chews gum". Write in English the propositions represented by the symbols

1. $p + q'$ 2. $p'q'$ 3. $pq' + p'q$

We have

$$\begin{aligned} p &: \text{missiles are costly;} \\ q &: \text{Grandpa chews gum;} \\ p' &: \text{missiles are not costly;} \\ q' &: \text{Grandpa does not chew gum;} \end{aligned}$$

Then,

1. $p + q'$: Either missiles are costly or Grandpa does not chew gum.
2. $p'q'$: Missiles are not costly and Grandpa does not chew gum.
3. $pq' + p'q$: Either missiles are costly and Grandpa does not chew gum, or missiles are not costly and Grandpa chews gum.

Exercise 6.4.2. 1. Which of the following sentences, or phrases, represent propositions?

(a) Grass is yellow.

(b) Beautiful white roses.

(c) All mathematics is difficult, and some mathematics is impossible.

2. Let p be the proposition "mathematics is easy," and let q be the proposition "two is less than three." Write out, in reasonable English, the propositions represented by

i. $p + q$ ii. $pq' + p'q$

6.5 Truth tables

To show that the set of propositions and the operations of conjunction, disjunction, and negation form a Boolean algebra, it is necessary first to define the concept of equality. Two propositional functions g and h, each functions of the n propositional variables $p_1, p_2, \ldots, p_n$, are said to be equal if and only if they have the same truth value for every possible way of assigning truth values to each of the n variables. To complete our algebra, we will create two new propositions represented by 0 and 1, respectively. We define 0 to be a proposition that is always false, and 1 to be a proposition that is always true. The equation $p = 0$ is equivalent to the statement that p is false. Similarly, $q = 1$ is equivalent to saying that q is true.

The definition we have given for equality makes it possible to represent a function with a table of values exactly as was done previously. The only difference is that now we have a special meaning attached to the symbols which appear in the table. These symbols stand for propositions rather than for abstract elements of an arbitrary Boolean algebra. Such a table will be termed a **truth table**. We give an example of such a table below.

This table represents the truth values of the propositions pq and $p + q$ for two simple propositions p and q according to their truth values.

The construction of a truth table for a complicated propositional function can best be carried out in steps, using at each step the basic truth table for one of the operations $(+)$, $(\cdot)$ or complement.

If it happens that the truth table for a function contains only 1's (in the function column), we call the corresponding proposition a **tautology**. Both $p + p'$ and $pq + pq' + p'q + p'q'$ are examples of tautologies for any propositions p and q.

Row	p	q	pq	$p+q$
1	1	1	1	1
2	1	0	0	1
3	0	1	0	1
4	0	0	0	0

We again attempt to draw the truth table of $(r' + pq)'$, for three propositions p, q and r. Since we have three propositions and two truth values, viz., 0 and 1, so we have $2^3 = 8$ possible combinations of truth values for the propositions. The table is given in table 6.4.

Row	p	q	r	r'	pq	$r' + pq$	$(r' + pq)'$
1	1	1	1	0	1	1	0
2	1	1	0	1	1	1	0
3	1	0	1	0	0	0	1
4	1	0	0	1	0	1	0
5	0	1	1	0	0	0	1
6	0	1	0	1	0	1	0
7	0	0	1	0	0	0	1
8	0	0	0	1	0	1	0

Table 6.4

An illustration of the usefulness of truth tables occurs in the proof of the following theorem. From the definition of equality, it follows that two functions are equal if and only if their truth tables are identical. This fact is used in the third part of the proof below.

Theorem 6.5.1. The algebra of propositions is a Boolean algebra.

Proof. In order to prove that the set of propositions forms a Boolean algebra, we will have to show that the four postulates hold which we stated in the beginning of the previous unit. We begin with them one by one.

(a) From the definition of disjunction and conjunction of propositions (denoted as $(\cdot)$ and $(+)$), it follows that they are commutative and hence the first postulate holds true.

(b) 0 is the identity element for the operation $(+)$ since $0 + p$ has the same truth value as p and hence equals p. Similarly, $(1)(q)$ has the same truth value as q and hence equals q, showing that 1 is the identity for the operation of conjunction.

(c) Each operation is distributive over the other as is shown by the table below (table 6.5).

From table 6.5, it can be seen that the truth values of $p + qr$ and $(p + q)(p + r)$ are same, and hence they are equal. Also, the truth values of $pq + pr$ and $p(q + r)$ are same and hence they are equal.

(d) For each proposition p, there is a second proposition p', the negation of p, which satisfies the relations $pp' = 0$ and $p + p' = 1$ as can be verified by the truth table 6.6.

Thus, p' is the complement of p.

Hence the theorem. □

p	q	r	pq	pr	qr	$p+qr$	$pq+pr$	$p+q$	$p+r$	$q+r$	$p(q+r)$	$(p+q)(p+r)$
1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	0	1	0	0	1	1	1	1	1	1	1
1	0	1	0	1	0	1	1	1	1	1	1	1
1	0	0	0	0	0	1	0	1	1	0	0	1
0	1	1	0	0	1	1	0	1	1	1	0	1
0	1	0	0	0	0	0	0	1	0	1	0	0
0	0	1	0	0	0	0	0	0	1	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0

Table 6.5

p	p'	pp'	$p+p'$
1	0	0	1
0	1	0	1

Table 6.6

Exercise 6.5.2. 1. Determine which of the following are tautologies by constructing the truth table for each.

(a) $pq+p'+q'$ 2. $p+q+p'$

2. Construct a truth table for each of the following functions.

(a) $pqr+p'qr'+p'q'r'$ 2. $(p'+qr)'(pq+q'r)$ 3. $pq'+p'(qr+q'r)'$

6.6 Few Probable Questions

1. Define disjunctive normal form. Find the disjunctive normal form for the function $f(x,y,z)=(x+y')(y+z')(z+x')(x'+y')$, in the smallest possible number of variables.

2. Define conjunctive normal form. Find the conjunctive normal form for the function $f(x,y,z)=xy'+xz+xy$, in the smallest possible number of variables.

3. Convert f from cnf to dnf where $f(x,y)=(x+y')(x'+y)(x'+y')$.

4. Let p be the proposition "x is an even number," and let q be the proposition "x is the product of two integers." Translate into symbols each of the following propositions.

   (a) Either x is an even number, or x is a product of two integers.

(b) Either x is an even number and a product of integers, or x is an odd number and is not a product of integers.

(c) x is neither an even number nor a product of integers.

5. Write, in reasonable English, the negation of each of the following propositions.

   (a) Either good health is desirable, or I have been misinformed.

   (b) Oranges are not suitable for use in vegetable salads.

   (c) There is a number which, when added to 6, gives a sum of 13.

6. Construct the truth table for $(p' + qr)'(pq + q'r)$.

Unit 7

Course Structure

- Boolean Algebra: Logic gates and circuits,
- Applications of Boolean Algebra to Switching theory (using AND, OR, & NOT gates),
- Karnaugh Map method.

7.1 Introduction

In this unit, we will introduce a third important application of Boolean algebra, the algebra of circuits, involving two-state (bistable) devices. The simplest example of such a device is a switch or contact. The theory introduced holds equally well for such two-state devices as rectifying diodes, magnetic cores, transistors, various types of electron tubes, etc. The nature of the two states varies with the device and includes conducting versus nonconducting, closed versus open, charged versus discharged, magnetized versus nonmagnetized, high-potential versus low-potential, and others. The algebra of circuits is receiving more attention at present, both from mathematicians and from engineers, than either of the two applications of Boolean algebra which we considered in the previous chapters. The importance of the subject is reflected in the use of Boolean algebra in the design and simplification of complex circuits involved in electronic computers, dial telephone switching systems, and many varied kinds of electronic control devices. The algebra of circuits fits into the general picture of Boolean algebra as an algebra with two elements 0 and 1. This means that except for the terminology and meaning connecting it with circuits, it is identical with the algebra of propositions considered as an abstract system. Either of these Boolean algebras is much more restricted than an algebra of sets.

Objectives

After reading this unit, you will be able to

- learn basic elements of a switching circuit
- learn to minimize a switching circuit using Boolean function
- define the logical circuit elements
- learn to simplify functions using Karnaugh maps

7.2 Switching Circuits

For the present, we will limit our discussion to the simplest kinds of circuits, those involving only switches. We will designate a switch by a single letter $a, b, c, x, y, \ldots$. If two switches operate so that they open and close simultaneously, we designate them by the same letter. If they operate so that the first is always open when the second is closed, and closed when the second is open, we denote the first by a letter, say x, and the second by x' (or, equally well, the first by x' and the second by x).

A circuit consisting of two switches x and y connected in parallel is denoted by $x+y$, and a circuit consisting of x and y connected in series is denoted by xy. Thus to each series-parallel circuit, there corresponds an algebraic expression; and conversely to each algebraic expression involving only $(+)$, $(\cdot)$ and negation, there corresponds a circuit (fig 7.2.1). We will speak of this relationship by saying that the function represents the circuit, and the circuit realizes the function. We will agree to assign the value 1 to a letter if it represents a

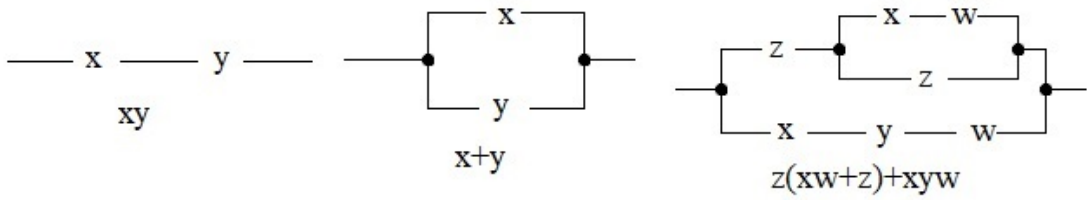


Figure 7.2.1

closed switch, and the value 0 if it represents an open switch. If a and a' both appear, then a is 1 if and only if a' is 0. A switch that is always closed is represented by 1, one that is always open by 0. Letters play the role of variables which take on the value 0 or 1, and we note the close analogy to proposition variables, which have the same possible values, although the meaning attached to these values has changed.

Two circuits involving switches $a, b, \ldots$ are said to be equivalent if the closure conditions of the two circuits are the same for any given position of the switches involved (values of the variables $a, b, \ldots$). That is, they are equivalent if for every position of the switches, current may either pass through both (both closed) or not pass through either (both open). Two algebraic expressions are defined to be equal if and only if they represent equivalent circuits.

It is now possible, by drawing the appropriate circuits and enumerating the possible positions of the switches involved, to check that each of the laws of Boolean algebra is valid when interpreted in terms of switching circuits. For example, consider the circuits that realize the functions on each side of the identity stating the distributive law for $(+)$ over $(\cdot)$, shown in figure 7.2.2. By inspection, it is apparent that the circuit

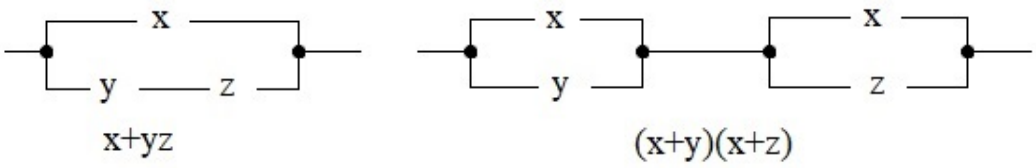


Figure 7.2.2

is closed (current can pass) if switch x is closed, or if both y and z are closed, and that the circuit is open (current cannot pass) if x and either y or z are open. Hence the circuits are equivalent, and this distributive law holds.

A simpler procedure for checking the validity of the fundamental laws is to note that numerical values of the switching functions a', ab, and $a + b$ are identical to the truth tables for the corresponding propositional functions (table 7.1).

Row	a	b	a'	ab	$a+b$
1	1	1	0	1	1
2	1	0	0	0	1
3	0	1	1	0	1
4	0	0	1	0	0

Table 7.1: Closure Properties of switching functions a', ab and $a + b$

Example 7.2.1. We want to find a circuit which realizes the Boolean function $xyz' + x'(y + z')$.

This expression indicates a series connection of x, y, and z' in parallel with a circuit corresponding to $x'(y' + z')$. This latter circuit consists of x' in series with a parallel connection of y and z. Hence the circuit diagram is that shown in fig. 7.2.3.

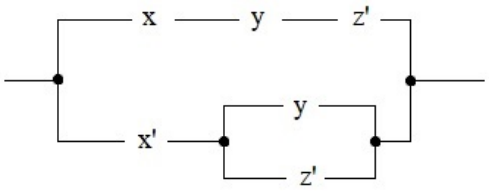


Figure 7.2.3

Example 7.2.2. We want to find the Boolean function which represents the circuit shown in fig. 7.2.4.

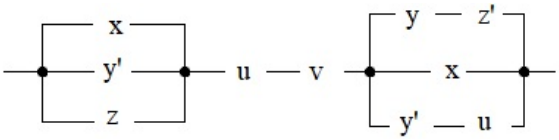


Figure 7.2.4

By inspection, the function is $(x + y' + z)uv(yz' + x + y'u)$.

Example 7.2.3. Construct the table of closure properties for the function $f(x, y, z) = x'y + z(x + y')$.

A table of closure properties for a function is identical, except for interpretation, to a truth table for a propositional function. This function has the closure properties listed in table 7.2.

Exercise 7.2.4. 1. Draw circuits which realize each of the following expressions, without first simplifying the expressions.

(a) $abc + ab(dc + ef)$

(b) $a + b(c + de) + fg$

2. Find the function which represents the circuits in the figure 7.2.5.

3. Find circuit which realize the function given in table 7.3.

Row	x	y	z	$x'y$	$x+y'$	$z(x+y')$	$x'y+z(x+y')$
1	1	1	1	0	1	1	1
2	1	1	0	0	1	0	0
3	1	0	1	0	1	1	1
4	1	0	0	0	1	0	0
5	0	1	1	1	0	0	1
6	0	1	0	1	0	0	1
7	0	0	1	0	1	1	1
8	0	0	0	0	1	0	0

Table 7.2

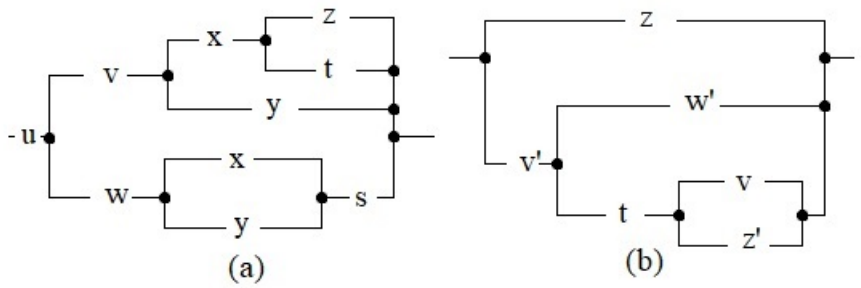


Figure 7.2.5

7.2.1 Simplification of circuits

In the previous section, we showed that the algebra of circuits is a Boolean algebra, and hence all the results proved earlier for Boolean algebras hold. In particular, theorems and rules relating to simplification of Boolean functions apply in the algebra of circuits.

Two basic problems that arise in connection with applications of Boolean algebra to switching circuits are (a) simplification of a given circuit which is known to have the desired closure properties, and (b) the design of circuits with given properties. The design problem will be discussed in later sections, and in this section we will consider the problem of simplifying a given circuit. This problem has often been solved in specific cases by trial-and-error methods. There are several known methods, based on the theory of Boolean functions, for writing schematic charts for simplifying functions. We will emphasize instead a straightforward approach using the properties of Boolean algebras directly to effect reasonable simplifications.

A general method of simplifying a circuit is first to find the Boolean function which represents the circuit, then to simplify the function as we have done repeatedly in earlier sections, and finally to draw a new circuit diagram realizing the simplified function. We give a simple illustration below.

Example 7.2.5. Simplify the circuit in fig. 7.2.6.

This circuit is represented by the Boolean function $(xy+abc)(xy+a'+b'+c')$, which simplifies to xy. Hence the given circuit is equivalent to the series connection of the two switches x and y, with the diagram given in fig. 7.2.7.

In using the basic laws of Boolean algebra, it often happens that a possible simplification is overlooked. It may happen that a certain step is easier to recognize if stated in terms of one of the dual laws rather than in terms of the other. This suggests another method of simplification which may help. To simplify a function f,

Row	x	y	z	f_1
1	1	1	1	0
2	1	1	0	1
3	1	0	1	1
4	1	0	0	0
5	0	1	1	0
6	0	1	0	0
7	0	0	1	0
8	0	0	0	1

Table 7.3

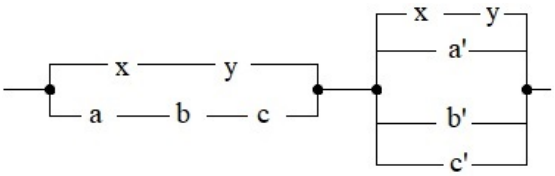


Figure 7.2.6

the dual of f may be taken and the resulting expression simplified. If the dual is taken again, the function f is obtained in a different form. This will usually be simpler than the original.

Example 7.2.6. Simplify the circuit in fig. 7.2.8. The circuit is represented by the function $f = cb + ab'cd + cd' + ac' + a'bc' + b'c'd'$. Consider the first three terms as the function g, and the last three terms as the function h. Then $g = cb + ab'cd + cd'$. The dual of g, which we write as $d(g)$ is then

$$d(g) = (c + b)(a + b' + c + d)(c + d') = c + abd'.$$

Taking the dual again, we find

$$g = c(a + b + d').$$

Similarly,

$$\begin{aligned} h &= ac' + a'bc' + b'c'd' \\ d(h) &= (a + c')(a' + b + c')(b' + c' + d') = c' + abd', \end{aligned}$$

Combining g and h yields

$$f = (c + c')(a + b + d') = a + b + d',$$

which corresponds to the circuit given in fig. 7.2.9.

Exercise 7.2.7. Simplify the circuits given in fig. 7.2.10.

7.3 Logical Circuit elements

Circuit elements involving diodes or vacuum tubes are very common. Rather than discuss the many types of electronic apparatus that may be used, we will introduce the idea of a logical circuit element. It will be

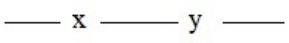


Figure 7.2.7

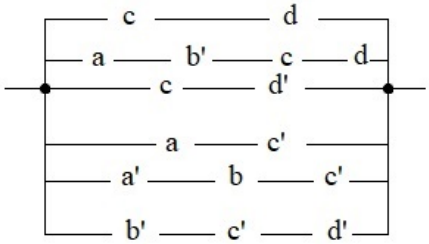


Figure 7.2.8

enough to know that these elements can be constructed; in fact, commercially packaged elements of these types, suitable for use in many types of equipment, can be purchased directly. We will conceive of a logical circuit element as a little box or package with one or more input leads (wire connections) and one or more output leads. These leads will carry signals in the form of positive voltage corresponding to a value 1, or zero voltage corresponding to a value 0. We will use a single letter, say x, to stand for the condition of the lead. When the lead carries a signal, we will say that x takes on the value 1. When the lead does not carry a signal, we say that x has the value 0. This represents only a slight modification of our earlier point of view, where 1 and 0 meant closed or open circuits, since we can think of a closed circuit as one carrying a signal, and of an open circuit as one incapable of carrying a signal. Other signals than that of a positive voltage could be used equally well, and in fact the signal used will in general depend on the type of components used in circuit construction. We will use just this one type of signal for simplicity, and we will adapt all our circuits to its use.

We will draw a circuit element as a circle with a letter inside to designate the type of element, and with lines indicating inputs and outputs. Arrows on these lines will indicate the difference between input and output, an arrow pointing toward the circle being used on each input.

The first logical circuit element we will consider has a single input and a single output. The function of this element is to obtain the complement of a given signal; that is, the output is 0 when the input is 1, and conversely. Fig. 7.3.1 shows the notation we will use, a circle with C in the center. The input is designated x, so the output is x'.

The next two logical circuit elements correspond to the logical connections "and" and "or." Each may have two or more inputs and only a single output. The "and" element is shown in diagrams as a circle with A in the center. This element produces an output signal (output has value 1) if and only if every input carries a signal (has value 1). If the inputs to an "and" element are x, y, and z, for example, the output function may be written as xyz, where the notation is that of Boolean algebra. The "or" element, represented graphically by a circle with O in the center, produces an output signal whenever one or more inputs carry a signal. If the inputs to an "or" element are x, y, and z, for example, the output is the Boolean function $x + y + z$. Fig. 7.3.2 shows the symbolic notations for these elements. Each is shown with only two inputs.

7.4 Karnaugh Maps

A Karnaugh map provides a pictorial method of grouping together expressions with common factors and therefore eliminating unwanted variables. The Karnaugh map can also be described as a special arrangement

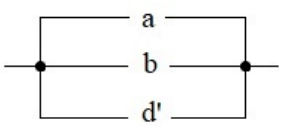


Figure 7.2.9

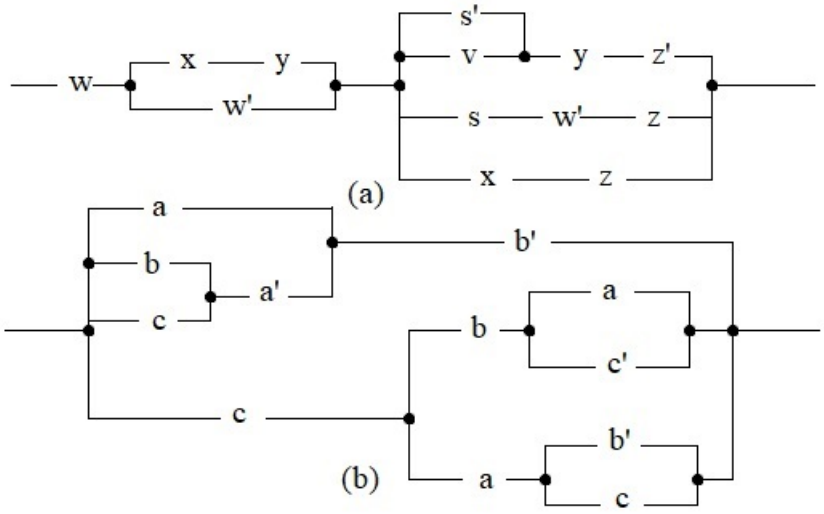


Figure 7.2.10

of a truth table.

The diagram below illustrates the correspondence between the Karnaugh map and the truth table for the general case of a two variable problem (fig. 7.4.1).

The values inside the squares are copied from the output column of the truth table, therefore there is one square in the map for every row in the truth table. Around the edge of the Karnaugh map are the values of the two input variable. x is along the top and y is down the left hand side. The diagram 7.4.2 explains this:

The values around the edge of the map can be thought of as coordinates. So as an example, the square on the top right hand corner of the map in the above diagram has coordinates $x = 1$ and $y = 0$. This square corresponds to the row in the truth table where $x = 1$ and $y = 0$ and $f = 1$. Note that the value in the f column represents a particular function to which the Karnaugh map corresponds.

Example 7.4.1. Consider the following map (fig. 7.4.3). The function plotted is:

$$f(x, y) = xy' + xy.$$

Note that values of the input variables form the rows and columns. That is the logic values of the variables x and y (with one denoting true form and zero denoting false form) form the head of the rows and columns respectively. Bear in mind that the above map is a one dimensional type which can be used to simplify an expression in two variables. There is a two-dimensional map that can be used for up to four variables, and a three-dimensional map for up to six variables.

Using algebraic simplification,

$$f = xy' + xy = x(y' + y) = x.$$

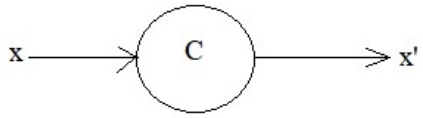


Figure 7.3.1

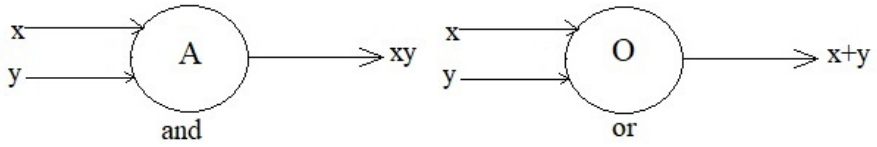


Figure 7.3.2

x	y	f
0	0	a
0	1	b
1	0	c
1	1	d

y \ x	0	1
0	a	b
1	c	d

Figure 7.4.1

x	y	f
0	0	0
0	1	1
1	0	**1**
1	1	1

y \ x	0	1
0	0	**1**
1	1	1

Figure 7.4.2

Variable B becomes redundant due to B4. Referring to the map 7.4.3, the two adjacent 1's are grouped together. Through inspection it can be seen that variable y has its true and false form within the group. This eliminates variable y leaving only variable x which only has its true form. The minimised answer therefore is f.

Example 7.4.2. Consider the expression $f(x, y) = x'y' + xy' + x'y$ plotted on the Karnaugh map 7.4.4. Pairs of 1's are grouped as shown in the figure, and the simplified answer is obtained by using the following steps:

Note that two groups can be formed for the example given above, bearing in mind that the largest rectangular clusters that can be made consist of two 1's. Notice that a 1 can belong to more than one group. The first group labelled I, consists of two 1's which correspond to $x = 0,\ \ y = 0$ and $x = 1,\ \ y = 0$. Put in another way, all squares in this example that correspond to the area of the map where $y = 0$ contains 1's, independent of the value of x. So when $y = 0$, the output is 1. The expression of the output will contain the term y'.

For group labelled II corresponds to the area of the map where $x = 0$. The group can therefore be defined as x'. This implies that when $x = 0$ the output is 1. The output is therefore 1 whenever $y = 0$ and $x = 0$. Hence the simplified answer is $f = x' + y'$.

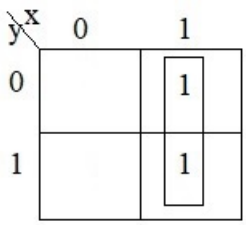


Figure 7.4.3

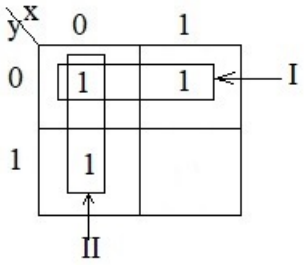


Figure 7.4.4

7.5 Few Probable Questions

1. Construct a table of closure properties and draw circuits which realize the function $(a + b' + c)(a + bc') + c'd + d(b' + c)$.

2. Find the function which represents the circuit in fig. 7.5.1.

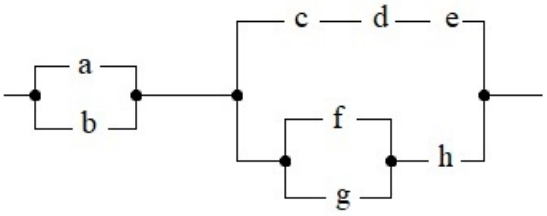


Figure 7.5.1

3. Minimise the following problems using the Karnaugh maps method.

   (a) $f = x'y'z' + x'y + xyz' + xz$;

   (b) $x'y + yz' + yz + xy'z'$.

4. Find circuits which realize each of the functions given in table 7.6.

Row	x	y	z	f_1	f_2
1	1	1	1	1	1
2	1	1	0	0	1
3	1	0	1	0	0
4	1	0	0	1	1
5	0	1	1	1	1
6	0	1	0	1	0
7	0	0	1	0	1
8	0	0	0	1	1

Table 7.6

5. Simplify each of the circuits given in fig. 7.5.2.

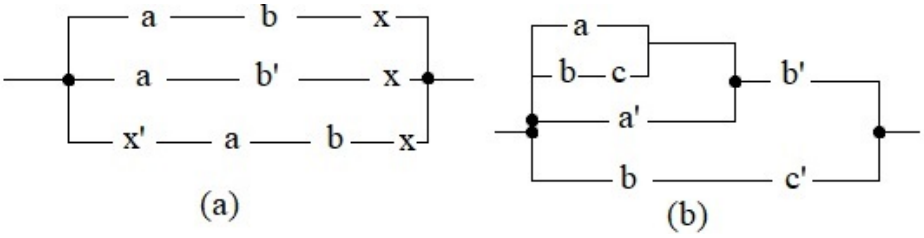


Figure 7.5.2

Unit 8

Course Structure

- Combinatorics : Introduction, Basic counting principles,
- Permutation and combination, pigeonhole principle,
- Recurrence relations and generating functions.

8.1 Introduction

Combinatorics studies the way in which discrete structures can be combined or arranged. Enumerative combinatorics concentrates on counting the number of certain combinatorial objects - e.g. the twelvefold way provides a unified framework for counting permutations, combinations and partitions. Analytic combinatorics concerns the enumeration (i.e., determining the number) of combinatorial structures using tools from complex analysis and probability theory. In contrast with enumerative combinatorics which uses explicit combinatorial formulae and generating functions to describe the results, analytic combinatorics aims at obtaining asymptotic formulae. Design theory is a study of combinatorial designs, which are collections of subsets with certain intersection properties. Partition theory studies various enumeration and asymptotic problems related to integer partitions, and is closely related to q-series, special functions and orthogonal polynomials. Originally a part of number theory and analysis, partition theory is now considered a part of combinatorics or an independent field. Order theory is the study of partially ordered sets, both finite and infinite.

Objectives

After reading this unit, you will be able to

- learn the sum rule and product rule principles and solve examples related to them
- learn various mathematical functions such as factorial function, and solve examples related to them
- define permutation and combination and solve related problems
- learn pigeonhole and generalized pigeonhole principle and solve related sums
- learn Inclusion-Exclusion principle and solve related sums

- define tree diagrams and solve sums related to these

8.2 Basic Counting principles

There are two basic counting principles used throughout this chapter. The first one involves addition and the second one multiplication.

1. **Sum Rule Principle:** Suppose some event E can occur in m ways and a second event F can occur in n ways, and suppose both events cannot occur simultaneously. Then E or F can occur in $m + n$ ways.

2. **Product Rule Principle:** Suppose there is an event E which can occur in m ways and, independent of this event, there is a second event F which can occur in n ways. Then combinations of E and F can occur in mn ways.

The above principles can be extended to three or more events. That is, suppose an event E_1 can occur in n_1 ways, a second event E_2 can occur in n_2 ways, and, following E_2; a third event E_3 can occur in n_3 ways, and so on.

Sum Rule: If no two events can occur at the same time, then one of the events can occur in:

$$n_1 + n_2 + \cdots \text{ ways.}$$

Product Rule: If the events occur one after the other, then all the events can occur in the order indicated in:

$$n_1 \cdot n_2 \cdots \text{ ways.}$$

Example 8.2.1. Suppose a college has 3 different history courses, 4 different literature courses, and 2 different sociology courses.

1. The number m of ways a student can choose one of each kind of courses is $m = 3(4)(2) = 24$.

2. The number n of ways a student can choose just one of the courses is $n = 3 + 4 + 2 = 9$.

There is a set theoretical interpretation of the above two principles. Specifically, suppose $n(A)$ denotes the number of elements in a set A. Then:

1. **Sum Rule Principle:** Suppose A and B are disjoint sets. Then

$$n(A \cup B) = n(A) + n(B).$$

2. **Product Rule Principle:** Let $A \times B$ be the Cartesian product of sets A and B. Then

$$n(A \times B) = n(A) \cdot n(B).$$

Example 8.2.2. There are four bus lines between A and B, and three bus lines between B and C. Find the number m of ways that a man can travel by bus: (a) from A to C by way of B; (b) roundtrip from A to C by way of B; (c) roundtrip from A to C by way of B but without using a bus line more than once.

(a) There are 4 ways to go from A to B and 3 ways from B to C; hence $n = 4 \cdot 3 = 12$.

(b) There are 12 ways to go from A to C by way of B, and 12 ways to return. Thus $n = 12 \cdot 12 = 144$.

(c) The man will travel from A to B to C to B to A. Enter these letters with connecting arrows as follows:

$$A \to B \to C \to B \to A.$$

The man can travel four ways from A to B and three ways from B to C, but he can only travel two ways from C to B and three ways from B to A since he does not want to use a bus line more than once. Enter these numbers above the corresponding arrows as follows:

$$A \overset{4}{\to} B \overset{3}{\to} C \overset{2}{\to} B \overset{3}{\to} A.$$

Thus, by the Product Rule, $n = 4 \cdot 3 \cdot 2 \cdot 3 = 72$.

8.3 Mathematical Functions

We discuss two important mathematical functions frequently used in combinatorics.

8.3.1 Factorial Function

The product of the positive integers from 1 to n inclusive is denoted by $n!$, read âĂIJn factorial.âĂİ Namely:

$$n! = 1 \cdot 2 \cdot 3 \cdots (n-2)(n-1)n = n(n-1)(n-2) \cdots 3 \cdot 2 \cdot 1.$$

Accordingly, $1! = 1$, $n! = n(n-1)!$. It is also convenient to define $0! = 1$.

8.3.2 Binomial Coefficients

The symbol $\binom{n}{r}$, read nCr, or "n Choose r", where r and n are positive integers with $r \leq b$, is defined as follows

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}.$$

Note that $n - (n-r) = r$. This yields the following lemma.

Lemma 8.3.1. $\binom{n}{n-r} = \binom{n}{r}$ or equivalently, $\binom{n}{a} = \binom{n}{b}$, where $a + b = n$.

Motivated by that fact that we defined $0! = 1$, we define:

$$\binom{n}{0} = \frac{n!}{0!n!} = 1 \quad \text{and} \quad \binom{0}{0} = \frac{0!}{0!0!} = 1.$$

Binomial Coefficients and Pascal's Triangle

The numbers $\binom{n}{r}$ are called binomial coefficients, since they appear as the coefficients in the expansion of $(a+b)^n$. Specifically:

Theorem 8.3.2. (Binomial Theorem)

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k.$$

The coefficients of the successive powers of $a + b$ can be arranged in a triangular array of numbers, called Pascal's triangle, as pictured in fig. 8.3.1. The numbers in Pascal's triangle have the following interesting properties:

$$(a+b)^0 = 1$$
$$(a+b)^1 = a + b$$
$$(a+b)^2 = a^2 + 2ab + b^2$$
$$(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$$
$$(a+b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$$
$$(a+b)^5 = a^5 + 5a^4b + 10a^3b^2 + 10a^2b^3 + 5ab^4 + b^5$$
$$(a+b)^6 = a^6 + 6a^5b + 15a^4b^2 + 20a^3b^3 + 15a^2b^4 + 6ab^5 + b^6$$

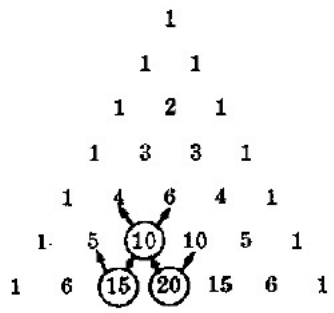


Figure 8.3.1

1. The first and last number in each row is 1.
2. Every other number can be obtained by adding the two numbers appearing above it.

Since these numbers are binomial coefficients, we state the above property formally.

Theorem 8.3.3.
$$\binom{n+1}{r} = \binom{n}{r-1} + \binom{n}{r}.$$

Exercise 8.3.4. 1. Compute: (a) $4!$, $5!$; (b) $6!, 7!, 8!, 9!$; (c) $50!$ [Hint: For large n, use Sterling's approximation: $n! = \sqrt{2\pi n} n^{\pi}\, \mathrm{e}^{-n}$, where $e \approx 2.718$].

2. Compute: (a) $\binom{18}{5}$, (b) $\binom{12}{4}$

3. Prove
$$\binom{17}{6} = \binom{16}{5} + \binom{16}{6}.$$

8.4 Permutations

Definition 8.4.1. Any arrangement of a set of n objects in a given order is called a permutation of the object (taken all at a time). Any arrangement of any $r \leq n$ of these objects in a given order is called an "r-permutation" or "a permutation of the n objects taken r at a time."

Consider, for example, the set of letters A, B, C, D. Then:

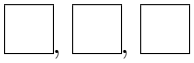
- BDCA, DCBA, and ACDB are permutations of the four letters (taken all at a time).
- BAD, ACB, DBC are permutations of the four letters taken three at a time.
- AD, BC, CA are permutations of the four letters taken two at a time.

We usually are interested in the number of such permutations without listing them. The number of permutations of n objects taken r at a time will be denoted by $P(n, r)$. We have the following theorem.

Theorem 8.4.2.
$$P(n, r) = \frac{n!}{(n-r)!}.$$

We emphasize that there are r factors in $n(n-1)(n-2)\cdots(n-r+1)$.

Example 8.4.3. Find the number m of permutations of six objects, say, A, B, C, D, E, F, taken three at a time. In other words, find the number of "three-letter words" using only the given six letters without repetition. Let us represent the general three-letter word by the following three positions:



The first letter can be chosen in 6 ways; following this the second letter can be chosen in 5 ways; and, finally, the third letter can be chosen in 4 ways. Write each number in its appropriate position as follows:

By the Product Rule there are $m = 6 \cdot 5 \cdot 4 = 120$ possible three-letter words without repetition from the six letters. Namely, there are 120 permutations of 6 objects taken 3 at a time. This agrees with the formula in the previous theorem.

$$P(6, 30) = 6 \cdot 5 \cdot 4 = 120.$$

Consider now the special case of $P(n, r)$ when $r = n$. We get the following result.

Corollary 8.4.4. There are $n!$ permutations of n objects (taken all at a time).

For example, there are $3! = 6$ permutations of the three letters A, B, C. These are:

$$ABC,\ ACB,\ BAC,\ BCA,\ CAB,\ CBA.$$

8.4.1 Permutations with Repetitions

Frequently we want to know the number of permutations of a multiset, that is, a set of objects some of which are alike. We will let

$$P(n; n_1, n_2, \ldots, n_t)$$

denote the number of permutations of n objects of which n_1 are alike, n_2 are alike, ..., n_t are alike.

Theorem 8.4.5. We have,

$$P(n; n_1, n_2, \ldots, n_t) = \frac{n!}{n_1!n_2!\cdots n_t!}.$$

We indicate the proof of the above theorem by a particular example. Suppose we want to form all possible five-letter "words" using the letters from the word "BABBY." Now there are $5! = 120$ permutations of the objects B_1, A, B_2, B_3, Y, where the three B's are distinguished. Observe that the following six permutations

$$B_1B_2B_3AY,\ \ B_2B_1B_3AY,\ \ B_3B_1B_2AY,\ \ B_1B_3B_2AY,\ \ B_2B_3B_1AY,\ \ B_3B_2B_1AY$$

produce the same word when the subscripts are removed. The 6 comes from the fact that there are $3! = 3 \cdot 2 \cdot 1 = 6$ different ways of placing the three B's in the first three positions in the permutation. This is true for each set of three positions in which the B's can appear. Accordingly, the number of different five-letter words that can be formed using the letters from the word "BABBY" is:

$$P(5; 3) = \frac{5!}{3!} = 20$$

Example 8.4.6. Find the number m of seven-letter words that can be formed using the letters of the word "BENZENE."

We seek the number of permutations of 7 objects of which 3 are alike (the three E's), and 2 are alike (the two N's). Thus,

$$m = P(7; 3, 2) = \frac{7!}{3!2!} = 420.$$

Ordered Samples

Definition 8.4.7. Many problems are concerned with choosing an element from a set S, say, with n elements. When we choose one element after another, say, r times, we call the choice an ordered sample of size r.

We consider two cases.

1. **Sampling with replacement:** Here the element is replaced in the set S before the next element is chosen. Thus, each time there are n ways to choose an element (repetitions are allowed). The Product rule tells us that the number of such samples is:

$$n \cdot n \cdot n \cdots n \ (r \ \text{ factors}) = n^r.$$

2. **Sampling without replacement:** Here the element is not replaced in the set S before the next element is chosen. Thus, there is no repetition in the ordered sample. Such a sample is simply an r-permutation. Thus the number of such samples is:

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

Example 8.4.8. Three cards are chosen one after the other from a 52-card deck. Find the number m of ways this can be done: (a) with replacement; (b) without replacement.

(a) Each card can be chosen in 52 ways. Thus $m = 52(52)(52) = 140608$.

(b) Here there is no replacement. Thus the first card can be chosen in 52 ways, the second in 51 ways, and the third in 50 ways. Therefore, $m = P(52, 3) = 52(51)(50) = 132600$.

Exercise 8.4.9. 1. Find the number n of distinct permutations that can be formed from all the letters of each word: (a) THOSE; (b) UNUSUAL; (c) SOCIOLOGICAL.

2. Find n if $P(n, 2) = 72$.

3. A class contains 8 students. Find the number n of samples of size 3: (a)With replacement; (b)Without replacement.

8.5 Combinations

Definition 8.5.1. Let S be a set with n elements. A combination of these n elements taken r at a time is any selection of r of the elements where order does not count. Such a selection is called an r-combination; it is simply a subset of S with r elements. The number of such combinations will be denoted by $C(n, r)$.

Before we give the general formula for $C(n, r)$, we consider a special case.

Example 8.5.2. Find the number of combinations of 4 objects, A, B, C, D, taken 3 at a time.
Each combination of three objects determines $3! = 6$ permutations of the objects as follows:

$$\begin{aligned} ABC &: \ ABC,\ ACB,\ BAC,\ BCA,\ CAB,\ CBA \\ ABD &: \ ABD,\ ADB,\ BAD,\ BDA,\ DAB,\ DBA \\ ACD &: \ ACD,\ ADC,\ CAD,\ CDA,\ DAC,\ DCA \\ BCD &: \ BDC,\ BDC,\ CBD,\ CDB,\ DBC,\ DCB. \end{aligned}$$

Thus the number of combinations multiplied by 3! gives us the number of permutations; that is,

$$C(4,3) \cdot 3! = P(4,3) \quad \text{or} \quad C(4,3) = \frac{P(4,3)}{3!}.$$

But $P(4,3) = 4 \cdot 3 \cdot 2 = 24$ and $3! = 6$; hence $C(4,3) = 4$ as noted above.

As indicated above, any combination of n objects taken r at a time determines $r!$ permutations of the objects in the combination; that is,

$$P(n,r) = r!C(n,r).$$

Accordingly, we obtain the following formula for $C(n,r)$ which we formally state as a theorem.

Theorem 8.5.3. We have,

$$C(n,r) = \frac{P(n,r)}{r!} = \frac{n!}{r!(n-r)!}.$$

Recall that the binomial coefficient $\binom{n}{r}$ was defined as

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}.$$

Hence,

$$C(n,r) = \binom{n}{r}.$$

We shall use $C(n,r)$ and $\binom{n}{r}$ interchangeably.

Example 8.5.4. A farmer buys 3 cows, 2 pigs, and 4 hens from a man who has 6 cows, 5 pigs, and 8 hens. Find the number m of choices that the farmer has.

The farmer can choose the cows in $C(6,3)$ ways, the pigs in $C(5,2)$ ways, and the hens in $C(8,4)$ ways. Thus the number m of choices follows:

$$m = \binom{6}{3}\binom{5}{2}\binom{8}{4} = 20 \cdot 10 \cdot 70 = 14000.$$

Example 8.5.5. A class contains 10 students with 6 men and 4 women. We want to find the number n of ways to:

(a) select a 4-member committee from the students. This concerns combinations, not permutations, since order does not count in a committee. There are "10 choose 4" such committees. That is:

$$n = C(10,4) = \binom{10}{4} = 210.$$

(b) select a 4-member committee with 2 men and 2 women. The 2 men can be chosen from the 6 men in $C(6,2)$ ways, and the 2 women can be chosen from the 4 women in $C(4,2)$ ways. Thus, by the Product Rule:

$$n = \binom{6}{2}\binom{4}{2} = 15 \cdot 6 = 90.$$

(c) elect a president, vice president, and treasurer. This concerns permutations, not combinations, since order does count. Thus, $n = P(6,3) = 6 \cdot 5 \cdot 4 = 120$.

Exercise 8.5.6. 1. A box contains 8 blue socks and 6 red socks. Find the number of ways two socks can be drawn from the box if: (a) They can be any color. (b) They must be the same color.

2. Find the number m of committees of 5 with a given chairperson that can be selected from 12 people.

8.6 Pigeonhole Principle

Many results in combinational theory come from the following almost obvious statement.

Theorem 8.6.1. (Pigeonhole Principle) If n pigeonholes are occupied by $n+1$ or more pigeons, then at least one pigeonhole is occupied by more than one pigeon.

This principle can be applied to many problems where we want to show that a given situation can occur.

Example 8.6.2.
1. Suppose a department contains 13 professors, then two of the professors (pigeons) were born in the same month (pigeonholes).
2. Find the minimum number of elements that one needs to take from the set $S = \{1, 2, \ldots, 9\}$ to be sure that two of the numbers add up to 10.

   Here the pigeonholes are the five sets $\{1, 9\}, \{2, 8\}, \{3, 7\}, \{4, 6\}, \{5\}$. Thus any choice of six elements (pigeons) of S will guarantee that two of the numbers add up to ten.

The Pigeonhole Principle is generalized as follows.

Theorem 8.6.3. (Generalized Pigeonhole Principle) If n pigeonholes are occupied by $kn + 1$ or more pigeons, where k is a positive integer, then at least one pigeonhole is occupied by $k + 1$ or more pigeons.

Example 8.6.4. Find the minimum number of students in a class to be sure that three of them are born in the same month.

Here $n = 12$ months are the pigeonholes, and $k+1 = 3$, so $k = 2$. Hence among any $kn+1 = 25$ students (pigeons), three of them are born in the same month.

Exercise 8.6.5.
1. Find the minimum number of students needed to guarantee that five of them belong to the same class (Freshman, Sophomore, Junior, Senior).
2. Let L be a list (not necessarily in alphabetical order) of the 26 letters in the English alphabet (which consists of 5 vowels, A, E, I, O, U, and 21 consonants).
   (a) Show that L has a sublist consisting of four or more consecutive consonants.
   (b) Assuming L begins with a vowel, say A, show that L has a sublist consisting of five or more consecutive consonants.

8.7 Inclusion-Exclusion Principle

Let A and B be any finite sets. Then we know that

$$n(A \cup B) = n(A) + n(B) - n(A \cap B).$$

In other words, to find the number $n(A \cup B)$ of elements in the union of A and B, we add $n(A)$ and $n(B)$ and then we subtract $n(A \cap B)$. This follows from the fact that, when we add $n(A)$ and $n(B)$, we have counted the elements of $n(A \cap B)$ twice. The principle in fact holds for any finite number of sets. We state it for three sets.

Theorem 8.7.1. For any finite sets A, B, C, we have

$$n(A \cup B \cup C) = n(A) + n(B) + n(C) - n(A \cap B) - n(A \cap C) - n(B \cap C) + n(A \cap B \cap C).$$

Example 8.7.2. Find the number of mathematics students at a college taking at least one of the languages French, German, and Russian, given the following data:

65 study French,	20 study French and German,	
45 study German,	25 study French and Russian,	8 study all three languages.
42 study Russian,	15 study German and Russian.	

We want to find $n(F \cup G \cup R)$, where F, G, and R denote the sets of students studying French, German, and Russian, respectively.

By the InclusionâĂŞExclusion Principle,

$$\begin{aligned} n(F \cup G \cup R) &= n(F) + n(G) + n(R) - n(F \cap G) - n(F \cap R) - n(G \cap R) + n(F \cap G \cap R) \\ &= 65 + 45 + 42 - 20 - 25 - 15 + 8 = 100. \end{aligned}$$

Namely, 100 students study at least one of the three languages.

Now, suppose we have any finite number of finite sets, say $A_1, A_2, \ldots, A_m$. Let s_k be the sum of the cardinalities

$$n(A_{i1} \cap A_{i2} \cap \ldots A_{ik})$$

of all possible k-tuple intersections of the given m sets. Then we have the following general InclusionâĂŞExclusion Principle.

Theorem 8.7.3. We have

$$n(A_1 \cup A_2 \cup \ldots \cup A_m) = s_1 - s_2 + s_3 - \cdots + (-1)^{m-1} s_m.$$

Exercise 8.7.4.

1. Suppose among 32 people who save paper or bottles (or both) for recycling, there are 30 who save paper and 14 who save bottles. Find the number m of people who: (a) save both; (b) save only paper; (c) save only bottles.

2. Let A, B, C, D denote, respectively, art, biology, chemistry, and drama courses. Find the number N of students in a dormitory given the data:

   2 take A, 5 take A and B, 4 take B and D, 2 take B, C, D,
   20 take B, 7 take A and C, 3 take C and D, 3 take A, C, D,
   20 take C, 4 take A and D, 3 take A, B, C, 2 take all four,
   8 take D, 16 take B and C, 2 take A, B, D, 71 take none.

8.8 Tree Diagrams

Definition 8.8.1. A tree diagram is a device used to enumerate all the possible outcomes of a sequence of events where each event can occur in a finite number of ways.

The construction of tree diagrams is illustrated in the following example

Example 8.8.2. (a) We want to find the product set $A \times B \times C$, where $A = \{1, 2\}$, $B = \{a, b, c\}$ and $C = \{x, y\}$. The tree diagram for $A \times B \times C$ is shown in fig. 8.8.1 (a). Here the tree is constructed from left to right, and the number of branches at each point corresponds to the possible outcomes of the next event. Each endpoint (leaf) of the tree is labelled by the corresponding element of $A \times B \times C$. As noted previously, $A \times B \times C$ has $n = 2 \cdot 3 \cdot 2 = 12$ elements.

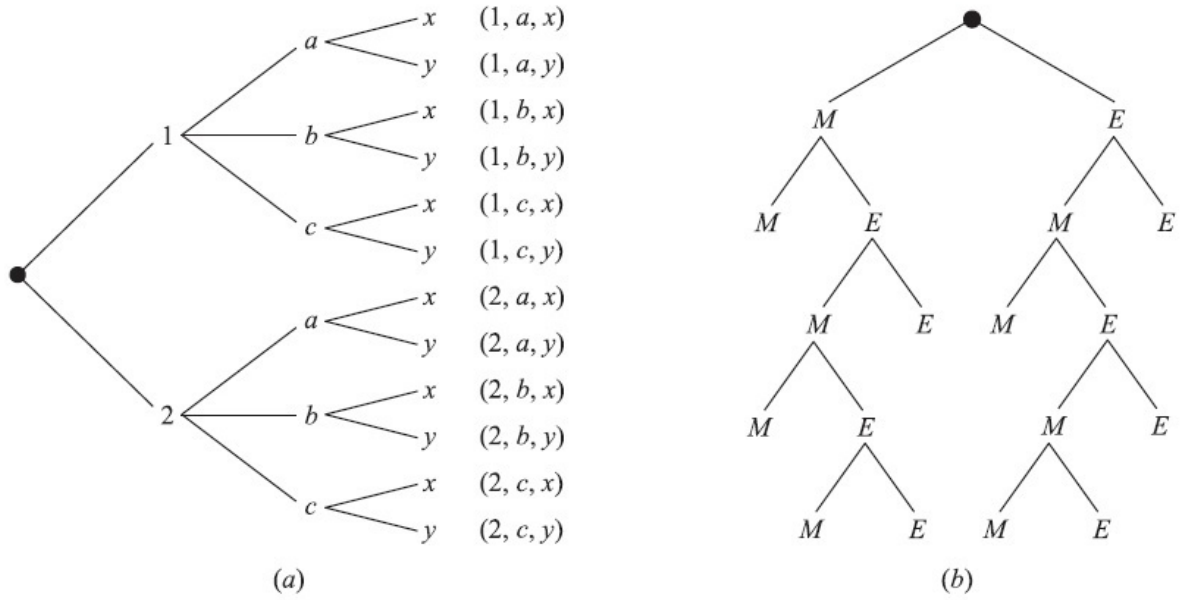


Figure 8.8.1

(b) Mark and Erik are to play a tennis tournament. The first person to win two games in a row or who wins a total of three games wins the tournament. We want to find the number of ways the tournament can occur.

The tree diagram showing the possible outcomes of the tournament appears in fig. 8.8.1 (b). Here the tree is constructed from top-down rather than from left-right. (That is, the âĂIJrootâĂİ is on the top of the tree.) Note that there are 10 endpoints, and the endpoints correspond to the following 10 ways the tournament can occur:

$$MM,\ MEMM,\ MEMEM,\ MEMEE,\ MEE,\ EMM,\ EMEMM,\ EMEME,\ EMEE,\ EE.$$

The path from the beginning (top) of the tree to the endpoint describes who won which game in the tournament.

Exercise 8.8.3. 1. Teams A and B play in a tournament. The first team to win three games wins the tournament. Find the number n of possible ways the tournament can occur. Construct the appropriate tree diagram.

2. Construct the tree diagram that gives the permutations of $\{a, b, c\}$.

8.9 Few Probable Questions

1. State sum rule principle. A store sells clothes for men. It has 3 kinds of jackets, 7 kinds of shirts, and 5 kinds of pants. Find the number of ways a person can buy: (a) one of the items; (b) one of each of the three kinds of clothes.

2. State product rule principle. Suppose a code consists of five characters, two letters followed by three digits. Find the number of: (a) codes; (b) codes with distinct letter; (c) codes with the same letters.

3. Find n if: (a) $P(n, 4) = 42P(n, 2)$; (b) $2P(n, 2) + 50 = P(2n, 2)$.

4. Consider all positive integers with three different digits. (Note that zero cannot be the first digit.) Find the number of them which are: (a) greater than 700; (b) odd; (c) divisible by 5.

5. A class contains l0 students. Find the number n of ordered samples of size 4: (a) with replacement; (b) without replacement.

6. A women student is to answer 10 out of 13 questions. Find the number of her choices where she must answer:

   (a) the first two questions;

   (b) the first or second question but not both;

   (c) exactly 3 out of the first 5 questions;

   (d) at least 3 of the first 5 questions.

7. Consider all integers from 1 up to and including 300. Find the number of them that are divisible by:

   (a) at least one of 3, 5, 7;

   (b) 3 and 5 but not by 7;

   (c) by 5, but by neither 3 nor 7;

   (d) by none of the numbers 3, 5, 7.

8. Find the number m of elements in the union of sets A, B, C, D where:

   (a) A, B, C, D have 50, 60, 70, 80 elements, respectively.

   (b) Each pair of sets has 20 elements in common.

   (c) Each three of the sets has 10 elements in common.

   (d) All four of the sets have 5 elements in common.

9. State pigeonhole principle. Suppose 5 points are chosen at random in the interior of an equilateral triangle T where each side has length two inches. Show that the distance between two of the points must be less than one inch.

Unit 9

Course Structure

- Grammar and Language : Introduction, Alphabets, Words, Free semi group,
- Languages, Regular expression and regular languages. Finite Automata (FA). Grammars.

9.1 Introduction

Automata theory is the study of abstract machines and automata, as well as the computational problems that can be solved using them. It is a theory in theoretical computer science and discrete mathematics (a subject of study in both mathematics and computer science). The word automata (the plural of automaton) comes from a Greek word, which means "self-making".

Automata theory is closely related to formal language theory. An automaton is a finite representation of a formal language that may be an infinite set. Automata are often classified by the class of formal languages they can recognize, typically illustrated by the Chomsky hierarchy, which describes the relations between various languages and kinds of formalized logics.

Automata play a major role in theory of computation, compiler construction, artificial intelligence, parsing and formal verification.

Objectives

After reading this unit, you will be able to

- define alphabets, words, concatenation of words, subwords
- see that the set of all words form a semigroup with respect to the concatenation of words
- define language, regular expression and regular languages
- define finite state automata and related terms and find its relation with languages
- define grammar find its relation with languages

9.2 Alphabet, Words, Free Semigroup

Definition 9.2.1. Consider a non-empty set A of symbols. A word or string w on the set A is a finite sequence of its elements.

For example, suppose $A = \{a, b, c\}$. Then the following sequences are words on A:

$$u = ababb, \quad \text{and} \quad v = accbaaa.$$

When discussing words on A, we frequently call A the alphabet, and its elements are called letters. We will also abbreviate our notation and write a^2 for aa, a^3 for aaa, and so on. Thus, for the above words, $u = abab^2$ and $v = ac^2ba^3$.

The empty sequence of letters, denoted by λ, or ϵ, or 1 is also considered to be a word on A, called the **empty word**. The set of all words on A is denoted by A^*.

Definition 9.2.2. The length of a word u, written $|u|$ or $l(u)$, is the number of elements in its sequence of letters.

For the above words u and v, we have $l(u) = 5$ and $l(v) = 7$. Also, $l(\lambda) = 0$.

Unless otherwise stated, the alphabet A will be finite, the symbols u, v, w will be reserved for words on A, and the elements of A will come from the letters a, b, c.

Definition 9.2.3. (**Concatenation**) Consider two words u and v on the alphabet A. The concatenation of u and v, written uv, is the word obtained by writing down the letters of u followed by the letters of v.

For the above words u and v, we have

$$uv = ababbaccbaaa = abab^2ac^2ba^3$$

As with letters, for any word u, we define $u^2 = uu$, $u^3 = uuu$, and in general, $u^{n+1} = uu^n$.

Clearly, for any words u, v, w, the words $(uv)w$ and $u(vw)$ are identical, they simply consist of the letters of u, v, w written down one after the other. Also, adjoining the empty word before or after a word u does not change the word u. That is:

Theorem 9.2.4. The concatenation operation for words on an alphabet A is associative. The empty word λ is an identity element for the operation.

(Generally speaking, the operation is not commutative, e.g., $uv \neq vu$ for the above words u and v.)

Definition 9.2.5. (Subwords, Initial Segments) Consider any word $u = a_1a_2 \dots a_n$ on an alphabet A. Any sequence $w = a_ja_{j+1} \dots a_k$ is called a subword of u. In particular, the subword $w = a_1a_2 \dots a_k$ beginning with the first letter of u, is called an initial segment of u. In other words, w is a subword of u if $u = v_1wv_2$ and w is an initial segment of u if $u = wv$. Observe that λ and u are both subwords or uv since $u = \lambda u$.

Consider the word $u = abca$. The subwords and initial segments of u are as follows:

1. Subwords: $\lambda, a, b, c, ab, bc, ca, abc, bca, abca = u$.
2. Initial segments: $\lambda, a, ab, abc, abca = u$.

Observe that the subword $w = a$ appears in two places in u. The word ac is not a subword of u even though all its letters belong to u.

Definition 9.2.6. Let F denote the set of all non-empty words from an alphabet A with the operation of concatenation. As noted above, the operation is associative. Thus F is a semigroup; it is called the **free semigroup** over A or the free semigroup generated by A.

One can easily show that F satisfies the right and left cancellation laws. However, F is not commutative when A has more than one element. We will write F_A for the free semigroup over A when we want to specify the set A.

Now let $M = A^*$ be the set of all words from A including the empty word λ. Since λ is an identity element for the operation of concatenation, M is a monoid, called the **free monoid** over A.

9.3 Languages

Definition 9.3.1. A **language** L over an alphabet A is a collection of words on A. Recall that A^* denotes the set of all words on A. Thus a language L is simply a subset of A^*.

Example 9.3.2. Let $A = \{a, b\}$. The following are languages on A.

1. $L_1 = \{a, ab, ab^2, \ldots\}$, consisting of all words beginning with an a and followed by zero or more b's.
2. $L_2 = \{b^m a b^n : \ m \geq 0, \ n \geq 0\}$, consisting of all words with exactly one a.

9.3.1 Operations on Languages

Suppose L and M are languages over an alphabet A. Then the "concatenation" of L and M, denoted by LM, is the language defined as follows:

$$LM = \{uv : \ u \in L, \ v \in V\}$$

That is, LM denotes the set of all words which come from the concatenation of a word from L with a word from M. For example, for the languages L_1 and L_2 described in the preceding example, we would have

$$L_1 L_2 = \{ab^m a b^n : \ m \geq 0, \ n \geq 0\}$$

Clearly, the concatenation of languages is associative since the concatenation of words is associative.

Powers of a language L are defined as follows:

$$L^0 = \lambda, \ L^1 = L, \ L^2 = LL, \ \ L^{m+1} = L^m L, \ m > 1.$$

The unary operation L^* of a language L, called the **Kleene closure** of L is defined as

$$L^* = \bigcup_{k=0}^{\infty} L^k.$$

9.4 Regular Expressions and Regular Languages

Let A be a (nonempty) alphabet. This section defines a regular expression r over A and a language $L(r)$ over A associated with the regular expression r. The expression r and its corresponding language $L(r)$ are defined inductively as follows.

Definition 9.4.1. Each of the following is a regular expression over an alphabet A.

1. The symbol λ and the pair "()"(empty expression) are regular expressions;

2. Each letter a in A is a regular expression;
3. If r is a regular expression, then r^* is a regular expression;
4. If r_1 and r_2 are regular expressions, then $(r_1 \vee r_2)$ is a regular expression;
5. If r_1 and r_2 are regular expressions, then $(r_1 r_2)$ is a regular expression.

All regular expressions are formed in this way.

Definition 9.4.2. The Language $L(r)$ over A defined by a regular expression r over A is as follows:

1. $L(\lambda) = \{\lambda\}$ and $L(()) = \emptyset$;
2. $L(a) = \{a\}$, where a is a letter in A;
3. $L(r^*) = (L(r))^*$ (the Kleene closure of $L(r)$);
4. $L(r_1 \vee r_2) = L(r_1) \cup L(r_2)$ (union of the languages);
5. $L(r_1 r_2) = L(r_1)L(r_2)$ (concatenation of the languages).

And finally,

Definition 9.4.3. Let L be a language over A. Then L is called a regular language over A if there exists a regular expression r over A such that $L = L(r)$.

Example 9.4.4. Let $A = \{a, b\}$. Each of the following is an expression r and its corresponding language $L(r)$:

1. Let $r = a^*$. Then $L(r)$ consists of all powers of a including the empty word.
2. Let $r = aa^*$. Then $L(r)$ consists of all positive powers of a excluding the empty word.
3. Let $r = a \vee b^*$. Then $L(r)$ consists of a or any word in b, that is, $L(r) = \{a, \lambda, b, b^2, \ldots\}$.
4. Let $r = (a \vee b)^*$. Then $L(r) = \{a\} \cup \{b\} = A$. Hence, $L(r) = A^*$.

Example 9.4.5. Consider the following languages over $A = \{a, b\}$. Find a regular expression r over A such that $L_i = L(r)$, for $i = 1, 2$.

1. $L_1 = \{a^m b^n : \ m > 0, \ n > 0\}$. L_1 consists of those words beginning with one or more a's followed by one or more b's. Thus we can set $r = aa^*bb^*$. Note that this r is not unique. We could also take $r = a^*abb^*$.
2. $L_2 = \{a^m b^m : \ m > 0\}$. L_2 consists of all words beginning with one or more a's followed by the same number of b's. There exists no regular expression r such that $L_2 = L(r)$; that is, L_2 is not a regular language.

Exercise 9.4.6.
1. Let $u = a^2b$ and $v = b^3ab$. Find (a) uvu; (b)λu, $u\lambda$.
2. State the difference between the free semigroup on an alphabet A and the free monoid on A.
3. Let $A = \{a, b, c\}$. Find where (a) $L = \{b^2\}$; (b) $L = \{a, b, c^3\}$
4. Let $A = \{a, b, c\}$. State whether w belongs to $L(r)$ or not, where (a) $r = a^* \vee (b \vee c)^*$; (b) $r = a^*(b \vee c)^*$.

9.5 Finite State Automata

Definition 9.5.1. A finite state automaton (FSA) or, simply, an automaton M, consists of five parts:

1. A finite set (alphabet) A of inputs.
2. A finite set S of (internal) states.
3. A subset Y of S (called accepting or "yes" states).
4. An initial state s_0 in S.
5. A next-state function $F : S \times A \to A$.

Such an automaton M is denoted by $M = (A, S, Y, s_0, F)$ (The plural of automaton is automata).

Example 9.5.2. The following defines an automaton M with two input symbols and three states:

1. $A = \{a, b\}$, input symbols.
2. $S = \{s_0, s_1, s_2\}$, internal states.
3. $Y = \{s_0, s_1\}$, "yes" states.
4. s_0, initial state,
5. Next state function $F : S \times A \to S$ defined explicitly as follows:

$$F(s_0, a) = s_0,\ F(s_1, a) = s_0,\ F(s_2, a) = s_2$$
$$F(s_0, b) = s_1,\ F(s_1, b) = s_2,\ F(s_2, b) = s_2.$$

9.5.1 State Diagram of an Automaton M

An automaton M is usually defined by means of its state diagram $D = D(M)$ rather than by listing its five parts. The state diagram $D = D(M)$ is a labelled directed graph as follows.

1. The vertices of $D(M)$ are the states in S and an accepting state is denoted by means of a double circle.
2. There is an arrow (directed edge) in $D(M)$ from state s_j to state s_k labelled by an input a if $F(s_j, a) = s_k$.
3. The initial state s_0 is indicated by means of a special arrow which terminates at s_0 but has no initial vertex.

For each vertex s_j and each letter a in the alphabet A, there will be an arrow leaving s_j, which is labelled by a; hence the outdegree of each vertex is equal to number of elements in A. For notational convenience, we label a single arrow by all the inputs which cause the same change of state rather than having an arrow for each such input.

The state diagram $D = D(M)$ of the automaton M in the preceding Example is shown in fig. 9.5.1.

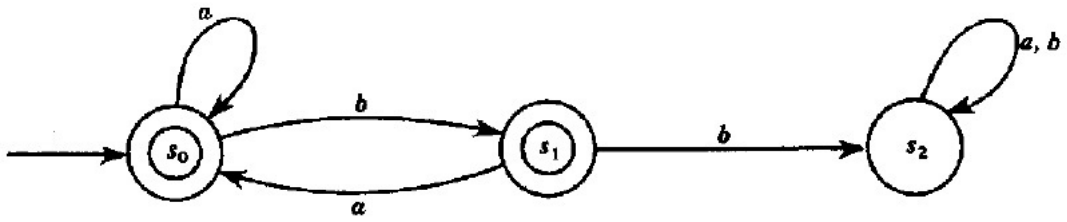


Figure 9.5.1

Language $L(M)$ Determined by an Automaton M

Each automaton M with input alphabet A defines a language over A, denoted by $L(M)$, as follows.

Let $w = a_1a_2\cdots a_m$ be a word on A. Then w determines the following path in the state diagram graph $D(M)$ where s_0 is the initial state and $F(s_{i-1}, a_i) = s_i$ for $i \geq 1$

$$P = (s_0, a_1, s_1, a_2, s_2, \ldots, a_m, s_m).$$

We say that M recognizes the word w if the final state s_m is an accepting state in Y. The language $L(M)$ of M is the collection of all words from A which are accepted by M.

Example 9.5.3. We determine whether or not the automaton M in fig. 9.5.1accepts the words

$$w_1 = ababba, \quad w_2 = baab, \quad w_3 = \lambda.$$

Using fig. 9.5.1 and the words w_1 and w_2, we obtain the respective paths:

$$P_1 = s_0 \xrightarrow{a} s_0 \xrightarrow{b} s_1 \xrightarrow{a} s_0 \xrightarrow{b} s_1 \xrightarrow{b} s_2 \xrightarrow{a} s_2, \quad \text{and} \quad P_2 = s_0 \xrightarrow{b} s_1 \xrightarrow{a} s_0 \xrightarrow{a} s_0 \xrightarrow{b} s_1$$

The final state in P_1 is s_2 which is not in Y. Hence w_1 is not accepted by M. Also, the final state of P_2 is s_1 which is in Y so w_2 is accepted by M. The final state determined by w_3 is the initial state s_0 which is in Y. Thus w_3 is also accepted by M.

We also describe the language $L(M)$. $L(M)$ will consist of all words w on A which do not have two successive b's. This comes from the following facts:

1. We can enter the state s_2 if and only if there are two successive b's.
2. We can never leave s_2.
3. The state s_2 is the only rejecting (non-accepting) state.

Example 9.5.4. Consider the automaton M in fig. 9.5.2. We want to find the words w in language L that are accepted by M. The system can reach the accepting state s_2 only when there exists an a in w which follows a b.

The fundamental relationship between regular languages and automata is contained in the following theorem.

Theorem 9.5.5. (Kleene): A language L over an alphabet A is regular if and only if there is a finite state automaton M such that $L = L(M)$.

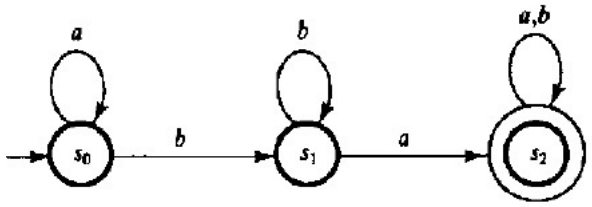


Figure 9.5.2

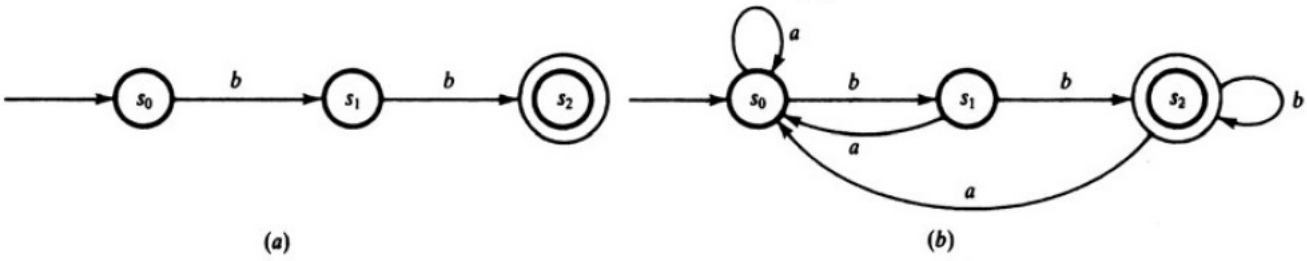


Figure 9.5.3

Example 9.5.6. Let $A = \{a, b\}$. We construct an automaton M which will accept precisely those words from A which end in two b's. Since b^2 is accepted, but not λ or b, we need three states, s_0, the initial state, and s_1 and s_2 with an arrow labelled b going from s_0 to s_1 and one from s_1 to s_2. Also, s_2 is an accepting state, but not s_0 nor s_1. This gives the graph in fig. 9.5.3(a). On the other hand, if there is an a, then we want to go back to s_0, and if we are in s_2 and there is a b, then we want to stay in s_2. These additional conditions give the required automaton M which is shown in fig. 9.5.3(b).

Example 9.5.7. Let $A = \{a, b\}$. We construct an automaton M which will accept those words from A which begin with an a followed by (zero or more) b's in fig. 9.5.4.

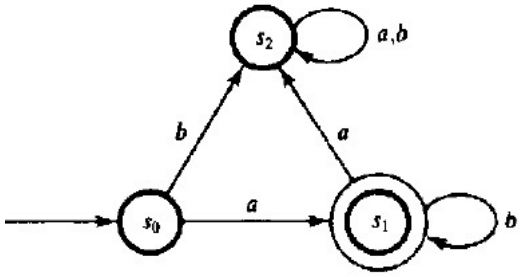


Figure 9.5.4

Pumping Lemma

Let M be an automaton over A with k states. Suppose $w = a_1 a_2 \cdots a_n$ is a word over A accepted by M and suppose $|w| = n > k$, the number of states. Let $P = (s_0, s_1, \ldots, s_n)$ be the corresponding sequence of states determined by the word w. Since $n > k$, two of the states in P must be equal, say $s_i = s_j$ where $i < j$. Let

w be divided into subwords x, y, z as follows:

$$x = a_1 a_2 \cdots a_i, \quad y = a_{i+1} \cdots a_j, \quad z = a_{j+1} \cdots a_n.$$

As shown in fig. 9.5.5, xy ends in $s_i = s_j$; hence xy^m also ends in s_i. Thus, for every m, $w_m = xy^m z$ ends in s_n, which is an accepting state.

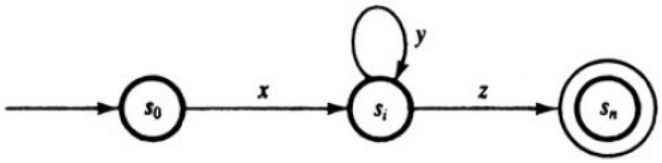


Figure 9.5.5

The above discussion proves the following important result.

Theorem 9.5.8. (Pumping Lemma): Suppose M is an automaton over A such that:

1. M has k states.
2. M accepts a word w from A where $|w| > k$.

Then $w = xyz$ where, for every positive m, $w_m = xy^m z$ is accepted by M.

The next example gives an application of the Pumping Lemma.

Example 9.5.9. We want to show that the language $L = \{a^m b^m : \ m > 0\}$ is not regular.

Suppose L is regular. Then by theorem 9.5.5, there exists a finite state automaton M which accepts L. Suppose M has k states. Let $w = a^k b^k$. Then $|w| > k$. By theorem 9.5.8, $w = xyz$ where y is not empty and $w_2 = xy^2 z$ is also accepted by M. If y consists of only a's or only b's, then w_2 will not have the same number of a's as b's. If y contains both a's and b's, then w_2 will have a's following b's. In either case, w_2 does not belong to L, which is a contradiction. Hence L is not regular.

9.6 Grammars

Fig. 9.6.1 shows the grammatical construction of a specific sentence. Observe that there are:

1. various variables, for example, (sentence), (noun phrase), etc.;
2. various terminal words, example, "The", "boy",etc.;
3. a beginning variable (sentence);
4. various substitutions or productions, for example,

$$\begin{aligned} \langle sentence \rangle &\rightarrow \langle noun\ phrase \rangle \langle verb\ phrase \rangle \\ \langle object\ phrase \rangle &\rightarrow \langle article \rangle \langle noun \rangle \\ \langle noun \rangle &\rightarrow apple \end{aligned}$$

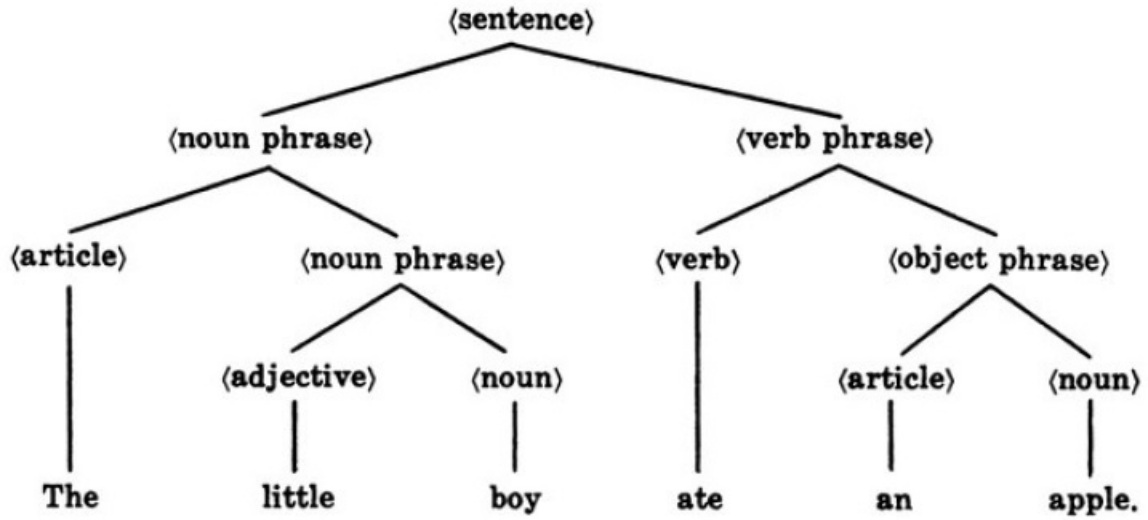


Figure 9.6.1

The final sentence only contains terminals, although both variables and terminals appear in its construction by the productions. This intuitive description is given in order to motivate the following definition of a grammar and the language it generates.

Definition 9.6.1. A phrase structure grammar or, simply, a grammar G consists of four parts:

1. A finite set (vocabulary) V;
2. A subset T of V whose elements are called **terminals**; the elements of $N = V \setminus T$ are called **non-terminals** or **variables**;
3. A non-terminal symbol S called the **start symbol**;
4. A finite set P of productions. (A production is an ordered pair (α, β), usually written $\alpha \to \beta$, where α and β are words in V, and the production must contain at least one non-terminal on its left side α.)

Such a grammar G is denoted by $G = G(V, T, S, P)$.

The following notation, unless otherwise stated or implied, will be used for our grammars. Terminals will be denoted by italic lower case Latin letters $a, b, c \ldots$, and non-terminals will be denoted by italic capital Latin letters $A, B, C, \ldots$, with S as the start symbol. Also, Greek letters, $\alpha, \beta, \ldots$, will denote words in V, that is, words in terminals and non-terminals. Furthermore, we will write $\alpha \to (\beta_1, \beta_2, \cdots, \beta_k)$.

9.6.1 Language $L(G)$ of a Grammar G

Suppose w and w' are words over the vocabulary set V of a grammar G. We write

$$w \Rightarrow w'$$

if w' can be obtained from w by using one of the productions; that is, if there exist words u and v such that $w = u\alpha v$ and $w' = u\beta v$ and there is a production $\alpha \to \beta$. Furthermore, we write

$$w \Rightarrow\Rightarrow w'$$

if w' can be obtained from w using a finite number of productions.

Now let G be a grammar with terminal set T . The language of G, denoted by $L(G)$, consists of all words in T that can be obtained from the start symbol S by the above process; that is,

$$L(G) = \{w \in T^* : S \Rightarrow\Rightarrow w\}$$

Example 9.6.2. The following defines a grammar G with S as the start symbol:

$$V = \{A, B, S, a, b\}, \quad T = \{a, b\}, \quad P = \{S \xrightarrow{1} AB, A \xrightarrow{2} Aa, B \xrightarrow{3} Bb, A \xrightarrow{4} a, B \xrightarrow{5} b\}$$

Now, $w = a^2b^4$ can be obtained from the start symbol S as follows:

$$S \Rightarrow AB \Rightarrow AaB \Rightarrow aaB \Rightarrow aaBb \Rightarrow aaBbb \Rightarrow aaBbbb \Rightarrow aabbbb = a^2b^4.$$

Here we used the productions 1, 2, 4, 3, 3, 3, 5, respectively. Thus, we write $S \Rightarrow\Rightarrow a^2b^4$ belongs to $L(G)$. More generally, the production sequence:

$$1,\ 2(r \text{ times}),\ 4,\ 3(s \text{ times}),\ 5$$

will produce the word $w = a^rab^sb$, where r and s are non-negative integers. On the other hand, no sequence of productions can produce an a after a b. Accordingly,

$$L(G) = \{a^m b^n :\ m > 0,\ n > 0\}$$

That is, the language $L(G)$ of the grammar G consists of all words which begin with one or more a's followed by one or more b's.

9.7 Few Probable Questions

1. Define Language. Let $A = \{a, b\}$. Find a regular expression r such that $L(r)$ consists of all words w where:

   (a) w begins with a^2 and ends with b^2;

   (b) w contains an even number of a's.

2. Define finite state automaton. Let M be the automaton with the following input set A, state set S with initial state s_0 and accepting set Y:

   $$A = \{a, b\},\ \ S = \{s_0, s_1, s_2\},\ \ Y = \{s_2\}.$$

   Also, the next-state function is given by

   $$F(s_0, a) = s_0,\ F(s_1, a) = s_1,\ F(s_2, a) = s_2$$
   $$F(s_0, b) = s_1,\ F(s_1, b) = s_2,\ F(s_2, b) = s_2.$$

   Draw the State diagram $D(M)$ of M. Also, describe the language $L = L(M)$ accepted by M.

3. Let $A = \{a, b\}$. Construct an automaton M which will accept precisely those words from A which have an even number of a's.

4. Find the language $L(G)$ generated by the grammar G with variables S, A, B, terminals a, b, and productions $S \to aB, B \to b, B \to bA, A \to aB$.

Unit 10

Course Structure

- Finite State Machine. Non-deterministic and deterministic FA.
- Push Down Automation (PDA).
- Equivalence of PDAs and Context Free Languages (CFLs),
- Computable Functions.

10.1 Introduction

This unit discusses two types of "machines." The first is a finite state machine (FSM) which is similar to a finite state automaton (FSA) except that the finite state machine "prints" an output using an output alphabet which may be distinct from the input alphabet. The second is the celebrated Turing machine which may be used to define computable functions.

Objectives

After reading this unit, you will be able to

- define finite state machines and draw their state tables and diagrams
- learn about Turing machines and how to work with them
- define computable functions and solve related problems

10.2 Finite State Machines

Definition 10.2.1. A finite state machine (or complete sequential machine) M consists of six parts:

1. A finite set A of input symbols;
2. A finite set S of "internal" states;

3. A finite set Z of output symbols;
4. An initial state s_0 in S;
5. A next-state function $f : S \times A \to S$;
6. An output function g from $S \times A$ into Z.

Such a machine M is denoted by $M = M(A, S, Z, s_0, f, g)$.

Example 10.2.2. The following defines a finite state machine M with two input symbols, three internal states, and three output symbols:

$$A = \{a, b\}, \;\; S = \{s_0, s_1, s_2\}, \;\; Z = \{x, y, z\}, \;\; \text{Initial state} \;\; s_0,$$

the next-state function $f : S \times A \to S$ defined by

$$f(s_0, a) = s_1, \;\; f(s_1, a) = s_2, \;\; f(s_2, a) = s_0$$
$$f(s_0, b) = s_2, \;\; f(s_1, b) = s_1, \;\; f(s_2, b) = s_1.$$

and the output function $g : S \times A \to Z$ defined by

$$g(s_0, a) = x, \;\; g(s_1, a) = x, \;\; g(s_2, a) = z$$
$$g(s_0, b) = y, \;\; g(s_1, b) = z, \;\; g(s_2, b) = y.$$

10.2.1 State Table and State Diagram of a Finite State Machine

There are two ways of representing a finite state machine M in compact form. One way is by a table called the state table of the machine M, and the other way is by a labelled directed graph called the state diagram of the machine M.

The state table combines the next-state function f and the output function g into a single table which represent the function $F : S \times A \to S \times Z$ defined as follows

$$F(s_j, a_j) = [f(s_i, a_j), g(s_i, a_j)]$$

For instance, the state table of the machine M in the preceding example is given in table 10.1. The states are

F	a	b
s_0	s_1, x	s_2, y
s_1	s_2, x	s_1, z
s_2	s_0, z	s_1, y

Table 10.1

listed on the left of the table with the initial state first, and the input symbols are listed on the top of the table. The entry in the table is a pair (s_k, z_r) where $s_k = f(s_i, a_j)$ is the next state and $z_r = g(s_i, a_j)$ is the output symbol. The corresponding state diagram is given in fig. 10.2.1.

The state diagram $D = D(M)$ of a finite state machine M is a labelled digraph the vertices of which are the states of M. Moreover, if

$$F(s_i, a_j) = (s_k, z_r) \quad \text{or equivalently,} \quad s_k = f(s_i, a_j), \;\; z_r = g(s_i, a_j)$$

then there is an arc (arrow) from s_i to s_k which is labelled with the pair a_j, z_r. We usually put the input symbol a_i near the base of the arrow (near s_i) and the output symbol z_r near the center of the arrow. We also label the initial state s_0 by drawing an extra arrow into s_0. See fig. 10.2.1.

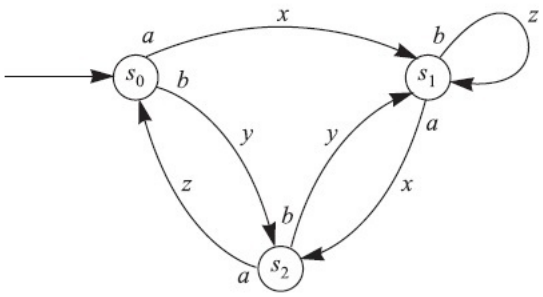


Figure 10.2.1

Input and Output Tapes

The above discussion of a finite state machine M does not show the dynamic quality of M. Suppose M is given a string (word) of input symbols, say

$$u = a_1 a_2 \cdots a_m$$

We visualize these symbols on an "input tape." The machine M "reads" these input symbols one by one and, simultaneously, changes through a sequence of states

$$v = s_0 s_1 s_2 \cdots s_m$$

where s_0 is the initial state, while printing a string(word) of output symbols

$$w = z_1 z_2 \cdots z_m$$

on an "output tape." Formally, the initial state s_0 and the input string u determine the strings v and w as follows, where $i = 1, 2, \ldots, m$:

$$s_i = f(s_{i-1} a_i), \quad \text{and} \quad z_i = g(s_{i-1}, a_i).$$

Example 10.2.3. Consider the machine M of fig. 10.2.1. Suppose the input is the word $u = abaab$. We calculate the sequence v of states and the output word w from the state diagram as follows. Beginning at the initial state s_0, we follow the arrows which are labelled by the input symbols as follows:

$$s_0 \xrightarrow{a,x} s_1 \xrightarrow{b,z} s_1 \xrightarrow{a,x} s_2 \xrightarrow{a,z} s_0 \xrightarrow{b,y} s_2$$

This yields the following sequence v of states and output word w:

$$v = s_0 s_1 s_1 s_2 s_0 s_2 \quad \text{and} \quad w = xzxy.$$

Binary Addition

This subsection describes a finite state machine M which can do binary addition. By adding 0's at the beginning of our numbers, we can assume that our numbers have the same number of digits. If the machine is given the input $1101011 + 0111011$ then we want the output to be the binary sum 10100110. Specifically, the input is the string of pairs of digits to be added:

$$11,\ 11,\ 00,\ 11,\ 01,\ 11,\ 10,\ b$$

where b denotes blank spaces, and the output should be the string:

$$0,\ 1,\ 1,\ 0,\ 0,\ 1,\ 0,\ 1$$

We also want the machine to enter a state called "stop" when the machine finishes the addition.

The input symbols and output symbols are, respectively, as follows:

$$A = \{00, 01, 10, 11, b\} \quad \text{and} \quad Z = \{0, 1, b\}.$$

The machine M that we "construct" will have three states:

$$S = \{\text{carry}(c), \text{no carry}(n), \text{stop}(s)\}$$

Here n is the initial state. The machine is shown in fig. 10.2.2. In order to show the limitations of our

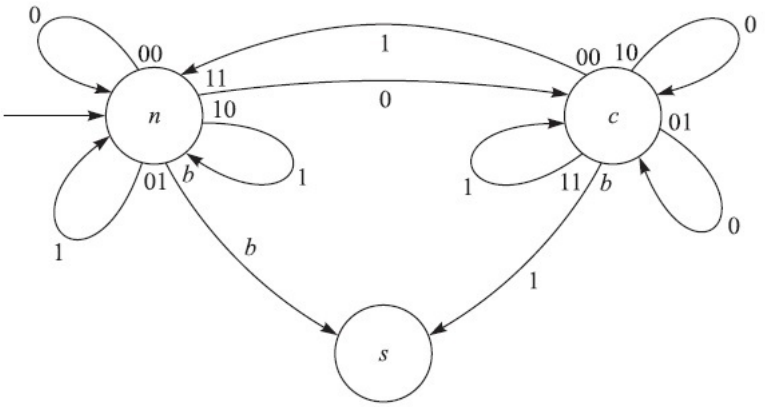


Figure 10.2.2

machines, we state the following theorem.

Theorem 10.2.4. There is no finite state machine M which can do binary multiplication.

If we limit the size of the numbers that we multiply, then such machines do exist. Computers are important examples of finite state machines which multiply numbers, but the numbers are limited as to their size.

10.3 Turing Machines

There are a number of equivalent ways to formally define a "computable" function. We do it by means of a Turing machine M. This section formally defines a Turing machine M, and the next section defines a computable function.

Our definition of a Turing machine uses an infinite two-way tape, quintuples, and three halt states. Other definitions use a one-way infinite tape and/or quadruples, and one halt state. However, all the definitions are equivalent.

A Turing machine M involves three disjoint non-empty sets:

1. A finite tape set where $B = a_0$ is the blank symbol:

$$A = \{a_1, a_2, \ldots, a_m\} \cup \{B\}$$

2. A finite state set where s_0 is the initial state:

$$S = \{s_1, s_2, \ldots, s_n\} \cup \{s_H, s_Y, s_N\}$$

where s_H(HALT) is the halting state, s_Y(YES) is the accepting state, and s_N(NO) is the non-accepting state.

3. A direction set where L denotes "left" and R denotes "right:"

$$d = \{L, R\}$$

Definition 10.3.1. An expression is a finite (possibly empty) sequence of elements from $A \cup S \cup d$. In other words, an expression is a word whose letters (symbols) come from the sets A, S, and d.

Definition 10.3.2. A tape expression is an expression using only elements from the tape set A.

The Turing machine M may be viewed as a read/write tape head which moves back and forth along an infinite tape. The tape is divided lengthwise into squares (cells), and each square may be blank or hold one tape symbol. At each step in time, the Turing machine M is in a certain internal state s_i scanning one of the tape symbols a_j on the tape. We assume that only a finite number of non-blank symbols appear on the tape.

Fig. 10.3.1(a) is a picture of a Turing machine M in state s_2 scanning the second symbol where $a_1a_3Ba_1a_1$ is printed on the tape. (Note again that B is the blank symbol.) This picture may be represented by the expression $\alpha = a_1s_2a_3Ba_1a_1$ where we write the state s_2 of M before the tape symbol a_3 that M is scanning. Observe that α is an expression using only the tape alphabet A except for the state symbol s_2 which is not at the end of the expression since it appears before the tape symbol a_3 that M is scanning. Fig. 10.3.1 shows two other informal pictures and their corresponding picture expressions.

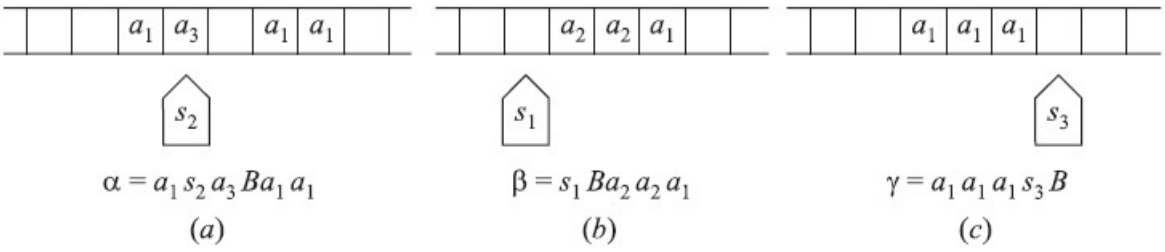


Figure 10.3.1

Definition 10.3.3. A picture α is an expression as follows where P and Q are tape expressions (possibly empty):

$$\alpha = Ps_ia_kQ$$

Definition 10.3.4. Let $\alpha = Ps_ia_kQ$ be a picture. We say that the Turing machine M is in state s_i scanning the letter a_k and that the expression on the tape is the expression Pa_kQ, that is, without its state symbol s_i.

As mentioned above, at each step in time the Turing machine M is in a certain state s_i and is scanning a tape symbol a_k. The Turing machine M is able to do the following three things simultaneously:

1. M erases the scanned symbol a_k and writes in its place a tape symbol a_l(where we permit $a_l = a_k$);

2. M changes its internal states s_i to a state s_j (where we permit $s_j = s_j$).

3. M moves one square to the left or moves one square to the right.

The above action by M may be described by a five-letter expression called a quintuple which we define below.

Definition 10.3.5. A quintuple q is a five-letter expression of the following form:

$$q = \left(s_i, a_k, a_l, s_j, \left\{ \begin{matrix} L \\ R \end{matrix} \right\} \right)$$

That is, the first letter of q is a state symbol, the second is a tape symbol, the third is a tape symbol, the fourth is a state symbol, and the last is a direction symbol L or R.

Next we give a formal definition of a Turing machine.

Definition 10.3.6. A Turing machine M is a finite set of quintuples such that:

1. No two quintuples begin with the same first two letters.

2. No quintuple begins with s_H, s_Y, or s_N.

First condition guarantees that the machine M cannot do more than one thing at any given step, and second condition guarantees that M halts in state s_H, s_Y, or s_N.

The following is an alternative equivalent definition.

Definition 10.3.7. Turing machine M is a partial function from

$$S \setminus \{s_H, s_Y, s_N\} \times A \text{ into } A \times S \times d$$

The term partial function simply means that domain of M is a subset of $S \setminus \{s_H, s_Y, s_N\} \times A$.

The action of the Turing machine described above can now be formally defined.

Definition 10.3.8. Let α and β be pictures. We write $\alpha \to \beta$ if one of the following holds where a, b, c are tape letters and P and Q are tape expressions (possibly empty):

1. $\alpha = Ps_iacQ$, $\beta = Pbs_jcQ$ and M contains the quintuple $q = s_iabs_jR$;

2. $\alpha = Pcs_iaQ$, $\beta = Ps_jcbQ$ and M contains the quintuple $q = s_iabs_jL$;

3. $\alpha = Ps_ia$, $\beta = Pbs_jB$ and M contains the quintuple $q = s_iabs_jR$;

4. $\alpha = s_iaQ$, $\beta = s_jBbQ$ and M contains the quintuple $q = s_iabs_jL$.

Observe that, in all four cases, M replaces a on the tape by b (where we permit $b = a$), and M changes its state from s_i to s_j (where we permit $s_j = s_i$). Furthermore:

1. Here M moves to the right.

2. Here M moves to the left.

3. Here M moves to the right; however, since M is scanning the rightmost letter, it must add the blank symbol B on the right.

4. Here M moves to the left; however, since M is scanning the leftmost letter, it must add the blank symbol B on the left.

Definition 10.3.9. A picture α is said to be terminal if there is no picture β such that $\alpha \to \beta$.

10.3.1 Computing with a Turing Machine

Definition 10.3.10. A computation of a Turing machine M is a sequence of pictures $\alpha_1, \alpha_2, \ldots, \alpha_m$ such that $\alpha_{i-1} \to \alpha_i$ for $i = 1, 2, \ldots, m$ and α_m is a terminal picture.

Turing Machines with Input

Definition 10.3.11. An input for a Turing machine M is a tape expression W. The initial picture for an input W is $\alpha(W)$, where $\alpha(W) = s_0(W)$.

Observe that the initial picture $\alpha(W)$ of the input W is obtained by placing the initial state s_0 in front of the input tape expression W. In other words, the Turing machine M begins in its initial state s_0 and it is scanning the first letter of W.

Definition 10.3.12. Let M be a Turing machine and let W be an input. We say M halts on W if there is a computation beginning with the initial picture $\alpha(W)$.

That is, given an input W, we can form the initial picture $\alpha(W) = s_0(W)$ and apply M to obtain the sequence

$$\alpha(W) \to \alpha_1 \to \alpha_2 \to \ldots$$

Two things can happen:

1. M halts on W. That is, the sequence ends with some terminal Picture α_r.
2. M does not halt on W. That is, the sequence never ends.

Grammars and Turing Machines

Turing machines may be used to recognize languages. Specifically, suppose M is a Turing machine with tape set A. Let L be the set of words W in A such that M halts in the accepting state s_Y when W is the input. We will then write $L = L(M)$, and we will say that M recognizes the language L. Thus an input W does not belong to $L(M)$ if M does not halt on W or if M halts on W but not in the accepting state s_Y.

Theorem 10.3.13. A language L is recognizable by a Turing machine M if and only if L is a type 0 language.

10.4 Computable Functions

Computable functions are defined on the set of non-negative integers. We denote the set of non-negative integers by N_0. Throughout this section, the terms number, integer, and nonnegative integer are used synonymously. The preceding section described the way a Turing machine M manipulates and recognizes character data. Here we show how M manipulates numerical data. First, however, we need to be able to represent our numbers by our tape set A. We will write 1 for the tape symbol a_1 and 1^n for $111\ldots1$, where 1 occur n times.

Definition 10.4.1. Each number n will be represented by the tape expression $\langle n \rangle$ where $\langle n \rangle = 1^{n+1}$.

Thus, $\langle 4 \rangle = 11111 = 1^5$, $\langle 0 \rangle = 1$.

Definition 10.4.2. Let E be an expression. Then $[E]$ will denote the number of times 1 occurs in E.

Then $[11Bs_2a_3111ba_4] = 5$.

Definition 10.4.3. A function $f : N_0 \to N_0$ is computable if there exists a Turing machine M such that, for every integer n, M halts on $\langle n \rangle$ and

$$f(n) = [\text{term}(\alpha(\langle n \rangle)].$$

We then say that M computes f.

That is, given a function f and an integer n, we input $\langle n \rangle$ and apply M. If M always halts on $\langle n \rangle$ and the number of 1's in the final picture is equal to $f(n)$, then f is a computable function and we say that M computes f.

Example 10.4.4. The function $f(n) = n + 3$ is computable. The input is $W = 1^{n+1}$. Thus we need only add two 1's to the input. A Turing machine M which computes f follows:

$$M = \{q_1, q_2, q_3\} = \{s_0 1 s_0 L, s_0 B 1 s_1 L, s_1 B 1 s_H L\}.$$

Observe that:

1. q_1 moves the machine M to the left.
2. q_2 writes 1 in the blank square B, and moves M to the left.
3. q_3 writes 1 in the blank square B, and halts M.

Accordingly, for any positive integer n,

$$s_0 1^{n+1} \to s_0 B 1^{n+1} \to s_1 B 1^{n+2} \to s_H B 1^{n+3}$$

Thus M computes $f(n) = n + 3$. It is clear that, for any positive integer k, the function $f(n) = n + k$ is computable.

Theorem 10.4.5. Suppose $f : N_0 \to N_0$ and $g : N_0 \to N_0$ are computable. Then the composition function $h = g \circ f$ is computable.

10.4.1 Functions of Several Variables

This subsection defines a computable function $f(n_1, n_2, \ldots, n_k)$ of k variables. First we need to represent the list $m = (n_1, n_2, \ldots, n_k)$ in our alphabet A.

Definition 10.4.6. Each list $m = (n_1, n_2, \ldots, n_k)$ of k integers is represented by the tape expression

$$\langle m \rangle = \langle n_1 \rangle B \langle n_2 \rangle B \cdots B \langle n_k \rangle$$

For example, $\langle (2, 0, 4) \rangle = 111B1B11111 = 1^3 B 1^1 B 1^5$.

Definition 10.4.7. The function $f(n_1, n_2, \ldots, n_k)$ of k variables is computable if there is a Turing machine M such that, for every list $m = (n_1, n_2, \ldots, n_k)$, M halts on $\langle m \rangle$ and

$$f(m) = [\text{term}(\alpha(\langle m \rangle))]$$

We then say that M computes f.

Example 10.4.8. The addition function $f(m, n) = m + n$ is computable. The input is $W = 1^{m+1} B 1^{n+1}$. Thus we need only erase two of the 1's. A Turing machine M which computes f follows:

$$M = \{q_1, q_2, q_3, q_4\} = \{s_0 1 B s_1 R, s_1 1 B s_H R, s_1 B B s_2 R, s_2 1 B s_H R\}$$

Observe that:

1. q_1 erases the first 1 and moves M to the right.
2. If $m \neq 0$, then q_2 erases the second 1 and halts M.
3. If $m = 0$, q_3 moves M to the right past the blank square B.
4. q_4 erases the 1 and halts M.

Accordingly, if $m \neq 0$, we have,

$$s_0 1^{m+1} B 1^{n+1} \rightarrow s_1 1^m B 1^{n+1} \rightarrow s_H 1^{m-1} B 1^{n+1}$$

but if $m = 0$ and $m + n = n$, we have

$$s_0 1 B 1^{n+1} \rightarrow s_1 B 1^{n+1} \rightarrow s_2 1^{n+1} \rightarrow s_H 1^n$$

Thus, M computes $f(m, n) = m + n$.

10.5 Few Probable Questions

1. Define finite state machine. Let M be a FSM with state table 10.2.

F	a	b
s_0	s_1, x	s_2, y
s_1	s_3, y	s_1, z
s_2	s_1, z	s_0, x
s_3	s_0, z	s_2, z

Table 10.2

   (a) Find the input set A, the state set S, the output set Z, and the initial state.
   (b) Draw the state diagram $D = D(M)$ of M
   (c) Suppose $w = aababaabbab$ is an input word (string). Find the corresponding output word v.

2. Define Turing machine. Suppose $\alpha = aas_2ba$ is a picture. Find β such that $\alpha \rightarrow \beta$ if the Turing machine M has the quintuple q where: (a) $q = s_2bas_1L$; (b) $q = s_2bbs_3R$.
3. Define computable functions. Show that the function f is computable where:
   (a) $f(n) = n - 1$, when $n > 0$, and $f(0) = 0$.
   (b) $f(x, y) = y$.

Unit 11

Course Structure

- Fields and σ-fields of events. Probability as a measure. Random variables. Probability distribution.

11.1 Introduction

The theory of probability had its origin in gambling and games of chance. It owes much to the curiosity of gamblers who pestered their friends in the mathematical world with all sorts of questions. A random (or statistical) experiment is an experiment in which

- All outcomes of the experiment are known in advance.
- Any performance of the experiment results in an outcome that is not known in advance.
- The experiment can be repeated under identical conditions.

In probability theory we study this uncertainty of a random experiment. It is convenient to associate with each such experiment a set Ω, the set of all possible outcomes of the experiment. To engage in any meaningful discussion about the experiment, we associate with Ω a σ-field S of subsets of Ω. We recall that a σ-field is a non-empty class of subsets of Ω that is closed under the formation of countable unions and complements and contains the null set ϕ.

The sample space of a statistical experiment is a pair (Ω, S), where

- Ω is the set of all possible outcomes of the experiment.
- S is a σ-field of subsets of Ω.

The elements of Ω are called *sample points.* Any set $A \in S$ is known as an *event*. Clearly, A is a collection of sample points. We say that an event A happens if the outcome of the experiment corresponds to a point in A. Each one point set is known as a *simple* or *elementary event*. If the set Ω contains only a finite number of points, we say that (Ω, S) is a *finite sample space*. If Ω contains at most a countable number of points, we call (Ω, S) a *discrete sample space*. If, however, Ω contains uncountably many points, we say that (Ω, S) is an *uncountable sample space*. In particular, if $\Omega = R_k$ or some rectangle in R_k, we call it a *continuous sample*

space.

Let us toss a coin. The set Ω is the set of symbols H and T, where H denotes head and T represents tail. Also, S is the class of all subsets of Ω, namely $\{\{H\}, \{T\}, \{H, T\}, \phi\}$. If the coin is tossed two times, then

$$\Omega = \{(H, H),\ (H, T),\ (T, H),\ (T, T)\},$$

and

$$\begin{aligned} S = \Big\{ & \phi,\ \{(H, H)\},\ \{(H, T)\},\ \{(T, H)\},\ \{(T, T)\},\ \{(H, H)\} \\ & \{(H, H), (H, T)\},\ \{(H, H), (T, H)\},\ \{(H, H), (T, T)\},\ \{(H, T), (T, H)\}, \\ & \{(T, T), (T, H)\},\ \{(T, T), (H, T)\},\ \{(H, H), (H, T), (T, H)\}, \\ & \{(H, H), (H, T), (T, T)\},\ \{(H, H), (T, H), (T, T)\},\ \{(H, T), (T, H), (T, T)\}, \Omega \Big\} \end{aligned} \quad (11.1.1)$$

where the first element of a pair denotes the outcome of the first toss, and the second element, the outcome of the second toss. The event *at least one head* consists of sample points $(H, H), (H, T), (T, H)$. The event *at most one head* is the collection of sample points $(H, T), (T, H), (T, T)$.

11.2 Random Variables

Suppose that to each point of a sample space we assign a number. We then have a function defined on the sample space. This function is called a random variable (or stochastic variable) or more precisely a random function (stochastic function). It is usually denoted by a capital letter such as X or Y. In general, a random variable has some specified physical, geometrical, or other significance.

Example 11.2.1. Suppose that a coin is tossed twice so that the sample space is S = {HH, HT, TH, TT}. Let X represent the number of heads that can come up. With each sample point we can associate a number for X as shown in Table 11.1. Thus, for example, in the case of HH (i.e., 2 heads), X = 2 while for TH (1 head), X = 1. It follows that X is a random variable.

Table 11.1

Sample Point	*HH*	*HT*	*TH*	*TT*
X	2	1	1	0

A random variable that takes on a finite or countably infinite number of values is called a discrete random variable while one which takes on a non-countably infinite number of values is called a non-discrete random variable.

11.3 Discrete Probability Distribution

Let X be a discrete random variable, and suppose that the possible values that it can assume are given by $x_1, x_2, x_3, \ldots$, arranged in some order. Suppose also that these values are assumed with probabilities given by

$$P(X = x_k) = f(x_k) \quad k = 1, 2, \ldots \quad (11.3.1)$$

It is convenient to introduce the probability function, also referred to as probability distribution, given by

$$P(X = x) = f(x) \tag{11.3.2}$$

For $x = x_k$, this reduces to (11.3.1) while for other values of x, $f(x) = 0$. In general, $f(x)$ is a probability function if

1. $f(x) \geq 0$
2. $\sum_x f(x) = 1$

where the sum in 2 is taken over all possible values of x.

11.4 Distribution Functions for Random Variables

The cumulative distribution function, or briefly the distribution function, for a random variable X is defined by

$$F(x) = P(X \leq x) \tag{11.4.1}$$

where x is any real number, i.e., $-\infty < x < \infty$. The distribution function F(x) has the following properties:

1. $F(x)$ is non-decreasing [i.e., $F(x) \leq F(y)$ if $x \leq y$].
2. $F(x)$ is continuous from the right [i.e., $\lim_{h \to 0^+} F(x + h) = F(x)$ for all x].

11.5 Continuous Random Variables

A non-discrete random variable X is said to be absolutely continuous, or simply continuous, if its distribution function may be represented as

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)\, du \quad (-\infty < x < \infty) \tag{11.5.1}$$

where the function $f(x)$ has the properties

1. $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)\, dx = 1$

It follows from the above that if X is a continuous random variable, then the probability that X takes on any one particular value is zero, whereas the interval probability that X lies between two different values, say, a and b, is given by

$$P(a < X < b) = \int_a^b f(x)\, dx \tag{11.5.2}$$

Example 11.5.1. Find the constant c such that the function

$$f(x) = \begin{cases} cx^2 & 0 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$

is a density function and compute $P(1 < X < 2)$.

Solution: Since $f(x)$ satisfies Property 1 if $c \geq 0$, it must satisfy Property 2 in order to be a density function. Now

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{0}^{3} cx^2\,dx = \left.\frac{cx^3}{3}\right|_0^3 = 9c$$

and since this must equal 1, we have $c = 1/9$. Now

$$P(1 < X < 2) = \int_{1}^{2} \frac{1}{9}x^2\,dx = \left.\frac{x^3}{27}\right|_1^2 = \frac{8}{27} - \frac{1}{27} = \frac{7}{27}$$

11.6 Joint Distributions

11.6.1 Discrete Case:

If X and Y are two discrete random variables, we define the joint probability function of X and Y by

$$P(X = x, Y = y) = f(x, y) \tag{11.6.1}$$

where

1. $f(x, y) \geq 0$
2. $\sum_x \sum_y f(x, y) = 1$

The joint distribution function of X and Y is defined by

$$F(x, y) = P(X \leq x, Y \leq y) = \sum_{u \leq x} \sum_{v \leq y} f(u, v) \tag{11.6.2}$$

Continuous case:

The case where both variables are continuous is obtained easily by analogy with the discrete case on replacing sums by integrals. Thus the joint probability function for the random variables X and Y (or, as it is more commonly called, the joint density function of X and Y) is defined by

1. $f(x, y) \geq 0$
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\,dx\,dy = 1$

The joint distribution function of X and Y in this case is defined by

$$F(x, y) = P(X \leq x, Y \leq y) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{y} f(u, v)\,du\,dv \tag{11.6.3}$$

From (11.6.3), we obtain

$$P(X \leq x) = F_1(x) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{\infty} f(u, v)\,du\,dv \tag{11.6.4}$$

$$P(Y \leq y) = F_2(y) = \int_{u=-\infty}^{\infty} \int_{v=-\infty}^{y} f(u, v)\,du\,dv \tag{11.6.5}$$

Eq. (11.6.4) and (11.6.5) are called the marginal distribution functions, or simply the distribution functions, of X and Y, respectively. The derivatives of (11.6.4) and (11.6.5) with respect to x and y are then called the marginal density functions, or simply the density functions, of X and Y and are given by

$$f_1(x) = \int_{v=-\infty}^{\infty} f(x,v)\,dv \qquad f_2(y) = \int_{u=-\infty}^{\infty} f(u,y)\,du \tag{11.6.6}$$

11.7 Change of Variables

Given the probability distributions of one or more random variables, we are often interested in finding distributions of other random variables that depend on them in some specified manner. Procedures for obtaining these distributions are presented in the following theorems for the case of discrete and continuous variables.

11.7.1 Discrete Variables

Theorem 11.7.1. Let X be a discrete random variable whose probability function is $f(x)$. Suppose that a discrete random variable U is defined in terms of X by $U = \phi(X)$, where to each value of X there corresponds one and only one value of U and conversely, so that $X = \psi(U)$. Then the probability function for U is given by

$$g(u) = f[\psi(u)] \tag{11.7.1}$$

Theorem 11.7.2. Let X and Y be discrete random variables having joint probability function $f(x,y)$. Suppose that two discrete random variables U and V are defined in terms of X and Y by $U = \phi_1(X,Y), V = \phi_2(X,Y)$, where to each pair of values of X and Y there corresponds one and only one pair of values of U and V and conversely, so that $X = \psi_1(U,V), Y = \psi_2(U,V)$. Then the joint probability function of U and V is given by

$$g(u,v) = f[\psi_1(u,v), \psi_2(u,v)] \tag{11.7.2}$$

Continuous variables

Theorem 11.7.3. Let X be a continuous random variable with probability density $f(x)$. Let us define $U = \phi(X)$ where $X = \psi(U)$ as in Theorem 11.7.1. Then the probability density of U is given by $g(u)$ where

$$g(u)|du| = f(x)|dx| \tag{11.7.3}$$

$$\text{or } g(u) = f(x)\left|\frac{dx}{du}\right| = f[\psi(u)]\left|\psi'(u)\right| \tag{11.7.4}$$

Theorem 11.7.4. Let X and Y be continuous random variables having joint density function $f(x,y)$. Let us define $U = \phi_1(X,Y)$, $V = \phi_2(X,Y)$ where $X = \psi_1(U,V)$, $Y = \psi_2(U,V)$ as in Theorem 11.7.2. Then the joint density function of U and V is given by $g(u,v)$ where

$$g(u,v)|du\,dv| = f(x,y)|dx\,dy| \tag{11.7.5}$$

$$\text{or } g(u,v) = f(x,y)\left|\frac{\partial(x,y)}{\partial(u,v)}\right| = f[\psi_1(u,v), \psi_2(u,v)]|J| \tag{11.7.6}$$

where Jacobian determinant or briefly Jacobian, is given by

$$J = \frac{\partial(x,y)}{\partial(u,v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} \tag{11.7.7}$$

Example 11.7.5. The probability function of a random variable X is

$$f(x) = \begin{cases} 2^{-x} & x = 1, 2, 3, \ldots \\ 0 & \text{otherwise} \end{cases}$$

Find the probability function for the random variable $U = x^4 + 1$.

Solution: Since $U = X^4 + 1$, the relationship between the values u and x of the random variables U and X is given by $u = x^4 + 1$ or $x = \sqrt[4]{u-1}$, where $u = 2, 17, 82, \ldots$ and the real positive root is taken. Then the required probability function for U is given by

$$g(u) = \begin{cases} 2^{-\sqrt[4]{u-1}} & u = 2, 17, 82, \ldots \\ 0 & \text{otherwise} \end{cases}$$

Example 11.7.6. If the random variables X and Y have joint density function

$$f(x) = \begin{cases} xy/96 & 0 < x < 4,\ 1 < y < 5 \\ 0 & \text{otherwise} \end{cases}$$

then, find the joint density function $U = XY^2,\ V = X^2Y$.

Solution: Consider $u = xy^2$, $v = x^2y$. Dividing these equations, we obtain $y/x = u/v$ so that $y = ux/v$. This leads to the simultaneous solution $x = v^{2/3}u^{-1/3},\ y = u^{2/3}v^{-1/3}$. The image of $0 < x < 4,\ 1 < y < 5$ in the uv-plane is given by

$$0 < v^{2/3}u^{-1/3} < 4 \qquad 1 < u^{2/3}v^{-1/3} < 5$$

which are equivalent to

$$v^2 < 64u \qquad v < u^2 < 125v$$

The Jacobian is given by

$$J = \begin{vmatrix} -\frac{1}{3}v^{2/3}u^{-4/3} & \frac{2}{3}v^{-1/3}u^{-1/3} \\ \frac{2}{3}u^{-1/3}v^{-1/3} & -\frac{1}{3}u^{2/3}v^{-4/3} \end{vmatrix} = -\frac{1}{3}u^{-2/3}v^{-2/3}$$

Thus the joint density function of U and V is

$$g(u,v) = \begin{cases} \frac{(v^{2/3}u^{-1/3})(u^{2/3}v^{-1/3})}{96}\left(\frac{1}{3}u^{-2/3}v^{-2/3}\right) & v^2 < 64u,\ v < u^2 < 125v \\ 0 & \text{otherwise} \end{cases}$$

$$\Rightarrow g(u,v) = \begin{cases} u^{-1/3}v^{-1/3}/288 & v^2 < 64u,\ v < u^2 < 125v \\ 0 & \text{otherwise} \end{cases}$$

11.8 Convolutions

As a particular consequence of the above theorems, we can show the density function of the sum of two continuous random variables X and Y, i.e., of $U = X + Y$, having joint density function $f(x, y)$ is given by

$$g(u) = \int_{-\infty}^{\infty} f(x, u - x)\, dx \tag{11.8.1}$$

In special case where the X and Y are independent, $f(x,y) = f_1(x)f_2(y)$, and (11.8.1) reduces to

$$g(u) = \int_{-\infty}^{\infty} f_1(x)f_2(u-x)\,dx \tag{11.8.2}$$

which is called the convolution of f_1 and f_2, abbreviated, $f_1 * f_2$. The following are some important properties of the convolution:

1. $f_1 * f_2 = f_2 * f_1$
2. $f_1 * (f_2 * f_3) = (f_1 * f_2) * f_3$
3. $f_1 * (f_2 + f_3) = (f_1 * f_2) + (f_1 * f_3)$

These results show that f_1, f_2, f_3 obey the commutative, associative and distributive laws of algebra with respect to the operation of convolution.

Theorem 11.8.1. Let X and Y be random variables having joint density function $f(x,y)$. Prove that the density function of $U = X + Y$ is

$$g(u) = \int_{-\infty}^{\infty} f(v, u-v)\,dv$$

Proof. Let $U = X+Y$, $V = X$, where we have arbitrary added the second equation. Corresponding to these we have $u = x + y,\ v = x$ or $x = v,\ y = u - v$. The Jacobian of the transformation is given by

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ 1 & -1 \end{vmatrix} = -1$$

Thus, the joint density function of U and V is

$$g(u,v) = f(v, u-v)$$

Therefore, the marginal density function of U is

$$g(u) = \int_{-\infty}^{\infty} f(v, u-v)\,dv$$

□

Example 11.8.2. If X and Y are independent random variables having density functions

$$f_1(x) = \begin{cases} 2e^{-2x} & x \geq 0 \\ 0 & x < 0 \end{cases} \qquad f_2(x) = \begin{cases} 3e^{-3y} & y \geq 0 \\ 0 & y < 0 \end{cases}$$

find the density function of their sum, $U = X + Y$.

Solution: The required density function is the the convolution of f_1 and f_2 is given by

$$g(u) = f_1 * f_2 = \int_{-\infty}^{\infty} f_1(v)f_2(u-v)\,dv$$

In the integrand f_1 vanish when $v < 0$ and f_2 vanishes when $v > u$. Hence

$$\begin{aligned} g(u) &= \int_0^u (2e^{-2v})(3e^{-3(u-v)}dv \\ &= 6e^{-3u}\int_0^u e^v dv = 6e^{-3u}(e^u - 1) = 6(e^{-2u} - e^{-3u}) \end{aligned}$$

if $u \geq 0$ and $g(u) = 0$ if $u < 0$. Check

$$\int_{-\infty}^{\infty} g(u)\, du = 6\int_0^{\infty}(e^{-2u} - e^{-3u})\, du = 6\left(\frac{1}{2} - \frac{1}{3}\right) = 1$$

Unit 12

Course Structure

- Expectation. Moments. Moment inequalities, Characteristic function. Convergence of sequence of random variables-weak convergence, strong convergence and convergence in distribution, continuity theorem for characteristic functions. Weak and strong law of large numbers. Central Limit Theorem.

12.1 Mathematical Expectation

A very important concept in probability and statistics is that of the mathematical expectation, expected value, or briefly the expectation, of a random variable. For a discrete random variable X having the possible values $x_1, \ldots, x_n$, the expectation of X is defined as

$$E(X) = x_1 P(X = x_1) + \ldots + x_n P(X = x_n) = \sum_{j=1}^{n} x_j P(X = x_j) \tag{12.1.1}$$

For a continuous random variable X having density function $f(x)$, the expectation of X is defined as

$$E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx \tag{12.1.2}$$

Another quantity of great importance in probability and statistics is called the variance and is defined by

$$\text{Var}(X) = E[(X - \mu)^2] \tag{12.1.3}$$

The variance is a non-negative number. The positive square root of the variance is called the standard deviation and is given by

$$\sigma_X = \sqrt{\text{Var}(X)} = \sqrt{E[(X - \mu)^2]} \tag{12.1.4}$$

If X is a continuous random variable having density function $f(x)$, then the variance is given by

$$\sigma_X^2 = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx \tag{12.1.5}$$

provided that the integral converges.

12.2 Moments

The r-th moment of a random variable X about the mean μ , also called the r-th central moment, is defined as

$$\mu_r = E[(X-\mu)^r] \tag{12.2.1}$$

where $r = 0, 1, 2, \ldots$. It follows that $\mu_0 = 1, \mu_1 = 0$ and $\mu_2 = \sigma^2$, i.e., the second central moment or second moment about the mean is the variance. We have, assuming absolute convergence,

$$\mu_r = \sum (x-\mu)^r f(x) \quad \text{(discrete variable)} \tag{12.2.2}$$

$$\mu_r = \int_{-\infty}^{\infty} (x-\mu)^r f(x)\, dx \quad \text{(continuous variable)} \tag{12.2.3}$$

The r-th moment of X about the origin, also called the r-th raw moment, is defined as

$$\mu'_r = E(X^r) \tag{12.2.4}$$

where $r = 0, 1, 2, \ldots$, and in this case there are formulas analogous to (12.2.2) and (12.2.3) in which $\mu = 0$. The relation between these moments is given by

$$\mu_r = \mu'_r - \binom{r}{1}\mu'_{r-1}\mu + \ldots + (-1)^j \binom{r}{j}\mu'_{r-j}\mu^j + \ldots + (-1)^r \mu_0^r \mu^r \tag{12.2.5}$$

Proof.

$$\begin{aligned}
\mu_r &= E(X-\mu)^r] \\
&= E\left[X^r - \binom{r}{1}X^{r-1}\mu + \cdots + (-1)^j\binom{r}{j}X^{r-j}\mu^j + \ldots + (-1)^{r-1}\binom{r}{r-1}X\mu^{r-1} + (-1)^r\mu^r\right] \\
&= E(X^r) - \binom{r}{1}E(X^{r-1})\mu + \ldots + (-1)^j\binom{r}{j}E(X^{r-j})\mu^j + \ldots + (-1)^{r-1}\binom{r}{r-1}E(X)\mu^{r-1} \\
&\qquad\qquad + (-1)^r\mu^r \\
&= \mu'_r - \binom{r}{1}\mu'_{r-1}\mu + \ldots + (-1)^j\binom{r}{j}\mu'_{r-j}\mu^j + \ldots + (-1)^{r-1}r\mu^r + (-1)^{-r}\mu^r
\end{aligned}$$

where the last two terms can be combined to give $(-1)^{r-1}(r-1)\mu^r$. □

12.3 Moment Generating Functions

The moment generating function of X is defined by

$$M_X(t) = E(e^{tX}) \tag{12.3.1}$$

Using the power series expansion, we have

$$\begin{aligned}
M_X(t) &= E(e^tX) = E\left(1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \ldots\right) \\
&= 1 + tE(X) + \frac{t^2}{2!}E(X^2) + \frac{t^3}{3!}E(X^3) + \ldots \\
&= 1 + \mu t + \mu'_2\frac{t^2}{2!} + \mu'_3\frac{t^3}{3!} + \ldots
\end{aligned} \tag{12.3.2}$$

Since the coefficients in this expansion enable us to find the moments, the reason for the name moment generating function is apparent. From the expansion we can show that

$$\mu'_r = \frac{d^r}{dt^r} M_X(t)\bigg|_{t=0} \tag{12.3.3}$$

i.e., μ'_r is the r-th derivative of $M_X(t)$ evaluated at $t = 0$.

Example 12.3.1. A random variable X has density function given by

$$f(x) = \begin{cases} 2e^{-2x} & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{12.3.4}$$

Find the moment generating function and the first four moments about origin.

Solution: We have

$$\begin{aligned} M(t) &= E(e^{tX}) = \int_{-\infty}^{\infty} e^{tX} f(x)\, dx \\ &= \int_0^{\infty} e^{tX}(2e^{-2x})\, dx = 2\int_0^{\infty} e^{(t-2x)} dx \\ &= \frac{2e^{(t-2)x}}{t-2}\bigg|_0^{\infty} = \frac{2}{2-t}, \quad \text{assuming } t < 2 \end{aligned}$$

If $|t| < 2$, we have

$$\frac{2}{2-t} = \frac{1}{1-\frac{t}{2}} = 1 + \frac{t}{2} + \frac{t^2}{4} + \frac{t^3}{8} + \frac{t^4}{16} + \cdots$$

But

$$M(t) = 1 + \mu t + \mu'_2 \frac{t^2}{2!} + \mu'_3 \frac{t^3}{3!} + \mu'_4 \frac{t^4}{4!} + \cdots$$

Therefore, on comparing terms, we have $\mu = \frac{1}{2},\ \mu'_2 = \frac{1}{2},\ \mu'_3 = \frac{3}{4},\ \mu'_4 = \frac{3}{2}$.

12.4 Characteristic Function

If we let $t = i\omega$, where i is the imaginary unit, in the moment generating function we obtain an important function called the characteristic function. We denote this by

$$\phi_X(\omega) = M_X(i\omega) = E(e^{i\omega X}) \tag{12.4.1}$$

It follows that

$$\phi_X(\omega) = \sum e^{i\omega x} f(x) \quad \text{(discrete variable)} \tag{12.4.2}$$

$$\phi_X(\omega) = \int_{-\infty}^{\infty} e^{i\omega x} f(x) \quad \text{(continuous variable)} \tag{12.4.3}$$

Since $|e^{i\omega t}| = 1$, the series and the integral always converge absolutely.
The corresponding results (12.3.2) and (12.3.3) becomes

$$\phi_X(\omega) = 1 + i\mu\omega - \mu_2' \frac{\omega^2}{2!} + \cdots + i^r \mu_r' \frac{\omega^r}{r!} + \cdots \tag{12.4.4}$$

where

$$\mu_r' = (-1)^r i^r \frac{d^r}{d\omega^r} \phi_X(\omega)\Big|_{\omega=0} \tag{12.4.5}$$

Theorem 12.4.1. If $\phi_X(\omega)$ is the characteristic function of the random variable X and a and b $(b \neq 0)$ are constants, then the characteristic function of $(X + a)/b$ is

$$\phi_{(X+a)/b}(\omega) = e^{ai\omega/b} \phi_X\left(\frac{\omega}{b}\right) \tag{12.4.6}$$

Theorem 12.4.2. If X and Y are independent random variables having characteristic functions $\phi_X(\omega)$ and $\phi_Y(\omega)$, respectively, then

$$\phi_{X+Y}(\omega) = \phi_X(\omega)\phi_Y(\omega) \tag{12.4.7}$$

Example 12.4.3. Find the characteristic function of the random variable X having density function given by

$$f(x) = \begin{cases} 1/2a & |x| < a \\ 0 & \text{otherwise} \end{cases}$$

Solution: The characteristic function is given by

$$\begin{aligned} E(e^{i\omega X} &= \int_{-\infty}^{\infty} e^{i\omega x} f(x)\, dx = \frac{1}{2a} \int_{-a}^{a} e^{i\omega x} dx \\ &= \frac{1}{2a} \frac{e^{i\omega x}}{i\omega}\Big|_{-a}^{a} = \frac{e^{ia\omega} - e^{-ia\omega}}{2ia\omega} = \frac{\sin a\omega}{a\omega} \end{aligned}$$

Example 12.4.4. Find the characteristic function of the random variable X having density function $f(x) = ce^{-a|x|}$, $-\infty < x < \infty$, where $a > 0$, and c is a suitable constant.

Solution: Since $f(x)$ is a density function, we must have

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

so that

$$\begin{aligned} c \int_{-\infty}^{\infty} e^{-a|x|} dx &= c \left[\int_{-\infty}^{0} e^{-a(-x)} dx + \int_{0}^{\infty} e^{-ax} dx \right] \\ &= c \frac{e^{ax}}{a}\Big|_{-\infty}^{0} + c \frac{e^{-ax}}{-a}\Big|_{0}^{\infty} = \frac{2c}{a} = 1 \end{aligned}$$

Then $c = a/2$. The characteristic function is therefore given by

$$\begin{aligned}
E(e^{i\omega X}) &= \int_{-\infty}^{\infty} e^{i\omega x} f(x)\, dx \\
&= \frac{a}{2}\left[\int_{-\infty}^{0} e^{i\omega x} e^{-a(-x)} dx + \int_{0}^{\infty} e^{i\omega x} e^{-ax} dx\right] \\
&= \frac{a}{2}\left[\int_{-\infty}^{0} e^{(a+i\omega)x} dx + \int_{0}^{\infty} e^{-(a-i\omega)x} e^{-ax} dx\right] \\
&= \frac{a}{2}\frac{e^{(a+i\omega)x}}{a+i\omega}\Bigg|_{-\infty}^{0} + a\frac{e^{-(a-i\omega)x}}{-(a-i\omega)}\Bigg|_{0}^{\infty} \\
&= \frac{a}{2(a+i\omega)} + \frac{a}{2(a-i\omega)} \\
&= \frac{a^2}{a^2+\omega^2}
\end{aligned}$$

12.5 Chebyshev's Inequality

Suppose that X is a random variable (discrete or continuous) having mean μ and variance σ^2, which are finite. Then if ϵ is any positive number,

$$P(|X-\mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2} \tag{12.5.1}$$

or, with $\epsilon = k\sigma$

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2} \tag{12.5.2}$$

Proof. We shall present the proof fro continuous random variables. A proof for discrete variables is similar if integrals are replaced by sums. If $f(x)$ is the density function of X, then

$$\sigma^2 = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$$

Since the integrand is non-negative, the value of the integral can only decrease when the range of integration is diminished. Therefore,

$$\sigma^2 \geq \int_{|x-\mu|\geq\epsilon} (x-\mu)^2 f(x)\, dx \geq \int_{|x-\mu|\geq\epsilon} \epsilon^2 f(x)\, dx = \epsilon^2 \int_{|x-\mu|\geq\epsilon} f(x)\, dx$$

But the last integral is equal to $P(|X-\mu| \geq \epsilon)$. Hence,

$$P(|X-\mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$$

□

12.6 Law of Large Numbers

Theorem 12.6.1. Let $X_1, X_2, \ldots, X_n$ be mutually independent random variables (discrete or continuous), each having finite mean μ and variance σ^2. Then if $S_n = X_1 + X_2 + \cdots + X_n (n = 1, 2, \ldots)$,

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) = 0 \tag{12.6.1}$$

Proof. We have

$$E(X_1) = E(X_2) = \cdots = E(X_n) = \mu$$
$$\text{Var}(X_1) = \text{Var}(X_2) = \cdots = \text{Var}(X_n) = \sigma^2$$

Then

$$E\left(\frac{S_n}{n}\right) = E\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{1}{n}[E(X_1) + \cdots + E(X_n)] = \frac{1}{n}(n\mu) = \mu$$
$$\text{Var}(S_n) = \text{Var}(X_1 + \cdots + X_n) = \text{Var}(X_1) + \cdots + \text{Var}(X_n) = n\sigma^2$$

so that

$$\text{Var}\left(\frac{S_n}{n}\right) = \frac{1}{n^2}\text{Var}(S_n) = \frac{\sigma^2}{n}$$

Therefore, by Chebyshev's inequality with $X = S_n/n$, we have

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}$$

Taking the limit as $n \to \infty$, this becomes, as required,

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) = 0$$

□

Note: Since S_n/n is the arithmetic mean of $X_1, \ldots, X_n$, this theorem states that the probability of the arithmetic mean S_n/n differing from its expected value μ by more than ϵ approaches zero as $n \to \infty$. A stronger result which we might expect to be true, is that

$$\lim_{n\to\infty} \frac{S_n}{n} = \mu,$$

but this is actually false. However, we can prove that $\lim_{n\to\infty} S_n/n = \mu$ with probability one. This result is often called the strong law of large numbers, and, by contrast Theorem 12.6.1 is called the weak law of large numbers.

12.7 Special Probability Distributions

12.7.1 The Binomial Distribution

Let p be the probability that an event will happen in any single Bernoulli trial (called the probability of success). Then $q = 1 - p$ is the probability that the event will fail to happen in any single trial (called the

probability of failure). The probability that the event will happen exactly x times in n trials (i.e., x successes and $n - x$ failures will occur) is given by the probability function

$$f(x) = P(X = x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x q^{n-x} \tag{12.7.1}$$

where the random variable X denotes the number of successes in n trials and $x = 0, 1, \ldots, n$.

1. Mean of Binomial Distribution, $\mu = np$
2. Variance of Binomial Distribution, $\sigma^2 = npq$
3. Moment generating function $M(t) = (q + pe^t)^n$
4. Characteristic function $\phi(\omega) = (q + pe^{i\omega})^n$

12.7.2 The Normal Distribution

One of the most important examples of a continuous probability distribution is the normal distribution, some times called the Gaussian distribution. The density function for this distribution is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty \tag{12.7.2}$$

where μ and σ are the mean and standard deviation, respectively. The corresponding distribution function is given by

$$F(x) = P(X \leq x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(v-\mu)^2}{2\sigma^2}} dv \tag{12.7.3}$$

If X has the distribution function given by (12.7.3), we say that the random variable X is normally distributed with mean μ and variance σ^2. If we let Z be the standardized variable corresponding to X, i.e., if we let

$$Z = \frac{X - \mu}{\sigma} \tag{12.7.4}$$

then the mean or expected value of Z is 0 and the variance is 1. In such case the density function for Z can be obtained from (12.7.2) by formally placing $\mu = 0$ and $\sigma = 1$, yielding

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \tag{12.7.5}$$

This is often referred to as the standard normal density function.

1. Mean of Normal Distribution, μ
2. Variance of Binomial Distribution, σ^2
3. Moment generating function $M(t) = e^{ut+(\sigma^2 t^2/2)}$
4. Characteristic function $\phi(\omega) = e^{i\mu\omega-(\sigma^2\omega^2/2)}$

12.7.3 Relation Between Binomial and Normal Distributions

If n is large and if neither p nor q is too close to zero, the binomial distribution can be closely approximated by a normal distribution with standardized random variable given by

$$Z = \frac{X - np}{\sqrt{npq}} \tag{12.7.6}$$

Here X is the random variable giving the number of successes in n Bernoulli trials and p is the probability of success. The approximation becomes better with increasing n and is exact in the limiting case. The fact that the binomial distribution approaches the normal distribution can be described by writing

$$\lim_{n\to\infty} P\left(a \leq \frac{X - np}{\sqrt{npq}} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-u^2/2} du \tag{12.7.7}$$

In words, we say that the standardized random variable $(X - np)/\sqrt{npq}$ is asymptotically normal.

12.7.4 The Poisson Distribution

Let X be a discrete random variable that can take on the values $0, 1, 2, \ldots$ such that the probability function of X is given by

$$f(x) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \ldots \tag{12.7.8}$$

where λ is a given positive constant. This distribution is called the Poisson distribution and random variable having this distribution is said to be Poisson distributed.

1. Mean of Poisson Distribution, $\mu = \lambda$
2. Variance of Poisson Distribution, $\sigma^2 = \lambda$
3. Moment generating function $M(t) = e^{\lambda(e^t - 1)}$
4. Characteristic function $\phi(\omega) = e^{\lambda(e^{i\omega} - 1}$

12.7.5 Relation Between the Poisson and Normal Distribution

We can show that if X is the Poisson random variable of (12.7.8) and $(X - \lambda)/\sqrt{\lambda}$ is the corresponding standardized random variable, then

$$\lim_{\lambda\to\infty} P\left(a \leq \frac{X - \lambda}{\sqrt{\lambda}} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-u^2/2} du \tag{12.7.9}$$

i.e., the Poisson distribution approaches the normal distribution as $\lambda \to \infty$ or $(X - \lambda)/\sqrt{\lambda}$ is asymptotically normal.

12.8 The Central Limit Theorem

The similarity between (12.7.7) and (12.7.9) naturally leads us to ask whether there are any other distribution besides the binomial and Poisson that have the normal distribution as the limiting case. The following remarkable theorem reveal that actually a large class of distribution have this property.

Theorem 12.8.1. Let $X_1, X_2, \ldots, X_n$ be independent random variables that are identically distributed (i.e., all have the same probability function in the discrete case or density function in the continuous case) and have finite mean μ and variance σ^2. Then is $S_n = X_1 + X_2 + \ldots + X_n$ $(n = 1, 2, \ldots)$,

$$\lim_{n\to\infty} P\left(a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-u^2/2} du \tag{12.8.1}$$

that is, the random variable $(S_n - n\mu)/\sigma\sqrt{n}$, which is the standardized variable corresponding to S_n, is asymptotically normal

Proof. For $n = 1, 2, \ldots$, we have $S_n = X_1 + X_2 + \ldots + X_n$. Now $X_1, X_2, \ldots, X_n$ each have mean μ and variance σ^2. Thus,

$$E(S_n) = E(X_1) + E(X_2) + \cdots + E(X_n) = n\mu$$

and, because the X_k are independent,

$$\text{Var}(S_n) = \text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n) = n\sigma^2$$

It follows that the standardized random variable corresponding to S_n is

$$S_n^* = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

The moment generating function for S_n^* is

$$\begin{aligned} E(e^{tS_n^*}) &= E[e^{t(S_n - n\mu)/\sigma\sqrt{n}}] \\ &= E[e^{t(X_1-\mu)/\sigma\sqrt{n}} e^{t(X_2-\mu)/\sigma\sqrt{n}} \cdots e^{t(X_n-\mu)/\sigma\sqrt{n}}] \\ &= E[e^{t(X_1-\mu)/\sigma\sqrt{n}}] \cdot E[e^{t(X_2-\mu)/\sigma\sqrt{n}}] \cdots E[e^{t(X_n-\mu)/\sigma\sqrt{n}}] \\ &= \{E[e^{t(X_1-\mu)/\sigma\sqrt{n}}]\}^n \end{aligned}$$

where, in the last two steps, we have respectively used the facts that the X_k are independent and are identically distributed. Now, by a Taylor series expansion,

$$\begin{aligned} E[e^{t(X_1-\mu)/\sigma\sqrt{n}}] &= E\left[1 + \frac{t(X_1-\mu)}{\sigma\sqrt{n}} + \frac{t^2(X_1-\mu)^2}{2\sigma^2 n} + \cdots\right] \\ &= E(1) + \frac{t}{\sigma\sqrt{n}} E(X_1 - \mu) + \frac{t^2}{2\sigma^2 n} E[(X_1-\mu)^2] + \cdots \\ &= 1 + \frac{t}{\sigma\sqrt{n}}(0) + \frac{t^2}{2\sigma^2 n}(\sigma^2) + \cdots \\ &= 1 + \frac{t^2}{2n} + \cdots \end{aligned}$$

so that

$$E(e^{tS_n^*}) = \left(1 + \frac{t^2}{2n} + \cdots\right)^n$$

But the limit of this as $n \to \infty$ is $e^{t^2/2}$, which is the moment generating function of the standardized normal distribution. Hence, the required result follows. □

Unit 13

Course Structure

- Definition and classification of stochastic processes
- Markov chains with finite and countable state space
- Classification of states.

13.1 Introduction

Since the last century there have been marked changes in the approach to scientific enquiries. There has been greater realisation that probability (or non-deterministic) models are more realistic than deterministic models in many situations. Observations taken at different time points rather than those taken at a fixed period of time began to engage the attention of probabilist. This lead to a new concept of indeterminism: indeterminism in dynamic studies. This has been called "dynamic indeterminism". Many phenomenon occurring in physical and life sciences are studied now not only as a random phenomenon but also as one changing with time or space. Similar considerations are also made in other areas, such as, social sciences, engineering and management and so on. The scope of applications of random variables which are functions of time or space or both has been on the increase.

Families of random variables which are functions of say, time, are known as stochastic processes (or random processes, or random functions). A few simple examples are given as illustrations.

Example 13.1.1. Consider a simple experiment like throwing a true die.

(i) Suppose that X_n is the outcome of the n-th throw, $n \geq 1$. Then $\{X_n,\ n \geq 1\}$ is a family of random variables such that for a distinct value of $n\ (= 1, 2, \ldots)$, one get a distinct random variable X_n; $\{X_n,\ n \geq 1\}$ constitutes a stochastic process, known as Bernoulli process.

(ii) Suppose that X_n is the number of sixes in the first n throws. For a distinct value of $n = 1, 2, \ldots,$ we get a distinct binomial variable X_n; $\{X_n,\ n \geq 1\}$ which gives a family of random variables is a stochastic process.

(iii) Suppose that X_n is the maximum number shown in the first n throws. Here $\{X_n,\ n \geq 1\}$ constitutes a stochastic process.

Example 13.1.2. Consider that there are r cells and an infinitely large number of identical balls and that balls are thrown at random, one by one, into the cells, the ball thrown being equally likely to go into any one of the cells. Suppose that X_n is the number of occupied cells after n throws. Then $\{X_n,\ n \geq 1\}$ constitutes a stochastic process.

13.2 Specification of Stochastic Processes

The set of possible values of a single random variable X_n of a stochastic process $\{X_n,\ n \geq 1\}$ is known as its *state space.* The state space is discrete if it contains a finite or a denumerable infinity of points; otherwise, it is continuous.

For example, if X_n is the toal number of sixes appearing in the first n throws of a die, the set of possible values of X_n is finite set of non-negative integers $0, 1, \ldots, n$. Here, the state space of X_n is discrete. We can write

$$X_n = Y_1 + Y_2 + \ldots + Y_n,$$

where Y_i is a discrete random variable denoting the outcome of the i-th throw and $Y_i = 1$ or 0 according as the i-th throw shows six or not. Secondly, consider

$$X_n = Z_1 + Z_2 + \ldots + Z_n,$$

where Z_i is a continuous random variable assuming values $[0, \infty)$. Here, the set of possible values of X_n is the interval $[0, \infty)$, and so the state space of X_n is continuous.

In the above two examples, we assume that the parameter n of X_n is restricted to the non-negative integers $n = 0, 1, 2, \ldots$ We consider the state of the system at distinct time points $n = 0, 1, 2, \ldots$, only. Here the word *time* is used in a wide sense. We note that in the first case considered above the "time n" implies throw number n.

On the other hand, one can visualise a family of random variables $\{X_t,\ t \in T\}$ (or $\{X(t),\ t \in T\}$) such that the state of the system is characterized at every instant over a finite or infinite interval. The system is then defined for a continuous range of time and we say that we have a family of random variable in *continuous* time. A stochastic process in continuous time may have either a discrete or a continuous state space. For example, suppose that $X(t)$ gives the number of incoming calls at a switchboard in an interval $(0, t)$. Here the state space of $X(t)$ is discrete through $X(t)$ is defined for a continuous range of time. We have a process in continuous time having a discrete state space. Suppose that $X(t)$ represents the maximum temperature at a particular place in $(0, t)$, then the set of possible values of $X(t)$ is continuous. Here we have a system in continuous time having a continuous state space.

So far we have assumed that the values assumed by the random variable X_n (or $X(t)$) are one-dimensional, but he process $\{X_n\}$ (or $\{X(t)\}$) may be multi-dimensional. Consider $X(t) = (X_1(t), X_2(t))$, where X_1 represents the maximum and X_2 the minimum temperature at a place in an interval of time $(0, t)$, We have here a two-dimensional stochastic process in continuous time having continuous state space. One can similarly have multi-dimensional processes. One-dimensional processes can be classified, in general, into the following four types of processes:

- Discrete time, discrete state space
- Discrete time, continuous state space

- Continuous time, discrete state space
- Continuous time, continuous state space.

All the four types may be represented by $\{X(t),\ t \in T\}$. In case of discrete time, the parameter generally used is n, i.e., the family is represented by $\{X_n,\ n = 0, 1, 2, \ldots\}$. In case of continuous time both the symbols $\{X_t,\ t \in T\}$ and $\{X(t),\ t \in T\}$ (where T is a finite or infinite interval) are used . The parameter t is usually interpreted as time, though it may represent such characters as distance, length, thickness and so on. We shall use the notation $\{X(t),\ t \in T\}$ both in the cases of discrete and continuous parameters and shall specify, whenever necessary.

Relationship

In some of the cases, the random variable X_n, i.e., members of the family $\{X_n,\ n \geq 1\}$ are mutually independent, but more often they are not. We generally come across processes whose members are mutually dependent. The relationship among them is often of great importance.

The nature of dependence could be infinitely varied. Here dependence of some special types, which occurs quite often and is of great importance, will be considered. We may broadly describe some stochastic processes according to the nature of dependence relationship existing among the members of the family.

Processes with independent increments

If for all $t_1, \ldots, t_n$, $t_1 < t_2 < \ldots < t_n$, the random variables

$$X(t_2) - X(t_1),\ \ X(t_3) - X(t_2),\ \ \ldots,\ \ X(t_n) - X(t_{n-1})$$

are independent, then $\{X(t),\ t \in T\}$ is said to be a process with independent increments.

Suppose that we wish to consider the discrete parameter case. Consider a process in discrete time with independent increments. Writing

$$T = \{0, 1, 2, \ldots\},\ \ t_i = i - 1,\ X(t_i) = X_{i-1},$$
$$Z_i = X_i - X_{i-1},\ \ i = 1, 2, \ldots \ \text{ and } \ Z_0 = X_0,$$

We have a sequence of independent random variables $\{Z_n,\ n \geq 0\}$.

Markov Process If $\{X(t), t \in T\}$ is a stochastic process such that, given the value $X(s)$, the values of $X(t), t > s$, do not depend on the values of $X(u),\ u < s$, then the process is said to be a Markov process. A definition of such a process is given below. If, for, $t_1 < t_2 < \ldots < t_n < t$,

$$P\Big\{a \leq X(t) \leq b | X(t_1) = x_1, \ldots, X(t_n) = x_n\Big\} = P\Big\{a \leq X(t) \leq b | X(t_n) = x_n\}$$

the process $\{X(t),\ t \in T\}$ is a *Makrov process*. A discrete parameter Markov process is known as a *Markov chain.*

13.3 Markov Chains

Consider a simple coin tossing experiment repeated for a number of times. The possible outcomes at each trial are two: head with probability, say, p and tail with probability q, $p + q = 1$. Let us denote head by 1 and

tail by 0 and the random variable denoting the result of the n-th toss by X_n. Then for $n = 1, 2, 3, \ldots$

$$P\{X_n = 1\} = p, \quad P\{X_n = 0\} = q.$$

Thus we have a sequence of random variables $X_1,\ X_2, \ldots$. The trials are independent and the result of the n-th trial does not depend in any way on the previous trials numbered $1,\ 2,\ \ldots,\ (n-1)$. The random variables are independent.

Consider now the random variable given by the partial sum $S_n = X_1 + \ldots + X_n$. The sum S_n gives the accumulated number of heads in the first n trials and its possible values are $0,\ 1,\ \ldots, n$.

We have $S_{n+1} = S_n + X_{n+1}$. Given that $S_n = j$ $(j = 0, 1, \ldots, n)$, the random variable S_{n+1} can assume only two possible values: $S_{n+1} = j$ with probability q and $S_{n+1} = j+1$ with probability p; these probabilities are not at all affected by the values of the variables $S_1, S_2, \ldots, S_{n-1}$. Thus

$$P\{S_{n+1} = j + 1 | S_n = j\} = p$$
$$P\{S_{n+1} = j | S_n = j\} = q.$$

We have an example of a Markov chain, a case of simple dependence that the outcome of $(n + 1)$-st trial depends directly on that of n-th trial and *only* on it. The conditional probability of S_{n+1} given S_n depends on the value of S_n and the manner in which the value of S_n was reached is of no consequence.

Definition 13.3.1. The stochastic process $\{X_n,\ n = 0, 1, 2, \ldots\}$ is called a Markov chain, if, for $j,\ k,\ j_1, \ldots, j_{n-1} \in N$ (or any subset of I),

$$P\{X_n = k | X_{n-1} = j,\ X_{n-1} = j_1, \ldots, X_0 = j_{n-1}\}$$
$$P\{X_n = k | X_{n-1} = j\} = p_{jk} \ \text{(say)}$$

whenever the first member is defined.

The outcomes are called the states of the Markov chain; if X_n has the outcome j (i.e., $X_n = j$), the process is said to be at state j at n-th trial. To a pair of states (j,k) at the two successive trials (say, n-th and (n+1)-th trials) there is an associated conditional probability p_{jk}. It is the probability of transition from the state j at n-th trial to the state k at (n+1)-th trial. The transition probabilities p_{jk} are basic to the study of the structure of the Markov chain.

The transition probability may or may not be independent of n. If the transition probability p_{jk} is independent of n, the Markov chain is said to be *homogeneous* (or to have *stationary transition probabilities*). If it is dependent on n, the chain is said to be non-homogeneous. Here we shall confine to *homogeneous chains*.

13.4 Transition Probabilities and Transition Matrix

For a finite Markov chain with m states $E_1, E_2, \ldots, E_m$, introduce the notation

$$p_{ij} = P\{X_n = j | X_{n-1} = i\} \tag{13.4.1}$$

where $i, j = 1, 2, \ldots, m$. The numbers p_{ij} are known as the transition probabilities of the chain, and must satisfy

$$p_{ij} \geq 0, \quad \sum_{j=1}^{m} p_{ij} = 1$$

for each $i = 1, 2, \ldots, m$.

Transition probabilities form an $m \times m$ array which can be assembled into a transition matrix T, where

$$T = [p_{ij}] = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{bmatrix} \tag{13.4.2}$$

Note that each row of T is a probability distribution. Any square matrix for which $p_{ij} \geq 0$ and $\sum_{j=1}^{m} p_{ij} = 1$ is said to be row-stochastic.

Example 13.4.1. The matrix $A = [a_{ij}]$ and $B = [b_{ij}]$ are $m \times n$ row-stochastic matrices. Show that $C = AB$ is also row-stochastic.

Solution: By the multiplication rule for the matrices

$$C = AB = [a_{ij}][b_{ij}] = \sum_{k=1}^{m} a_{ik} b_{kj}.$$

Since $a_{ij} \geq 0$ and $b_{ij} \geq 0$ for all $i, j = 1, 2, \ldots, m$, it follows that $c_{ij} \geq 0$. Also

$$\sum_{j=1}^{m} c_{ij} = \sum_{j=1}^{m} \sum_{k=1}^{m} a_{ik} b_{kj} = \sum_{k=1}^{m} a_{ik} \sum_{j=1}^{m} b_{kj} = \sum_{k=1}^{m} a_{ik} \cdot 1 = 1,$$

since, $\sum_{j=1}^{m} b_{kj} = 1$ and $\sum_{k=1}^{m} a_{ik} = 1$.

It follows from this example that any power T^n of the transition matrix T must also be row-stochastic.

13.5 Classification of States

Let us consider the general m-state chain with states $E_1, E_2, \ldots, E_m$ and transition matrix

$$T = [p_{ij}], \quad (1 \leq i, j \leq m)$$

For a homogeneous chain, recollect that p_{ij} is the probability that a transition occurs between E_i and E_j at any step or change of state in the chain. We intend to investigate and classify some of the more common types of states which can occur in Markov chains.

(a) **Absorbing state:** An absorbing state E_i is characterised by the probabilities

$$p_{ii} = 1, \quad p_{ij} = 0, \ \ (i \neq j, \ \ j = 1, 2, \ldots, m)$$

in the i-th row of T.

(a) **Periodic state:** The probability of a return to E_i at step n is $p_{ii}^{(n)}$. Let t be an integer greater than 1. Suppose that

$$\begin{aligned} p_{ii}^{(n)} &= 0 \quad \text{for} \quad n \neq t, \ 2t, \ 3t, \ldots \\ p_{ii}^{(n)} &\neq 0 \quad \text{for} \quad n = t, \ 2t, \ 3t, \ldots \end{aligned}$$

In this case, the state E_i is said to be periodic with period t. If, for a state, no such t exists with this property, then the state is described as aperiodic. Let

$$d(i) = \gcd\{n | p_{ii}^{(n)} > 0\}, \tag{13.5.1}$$

that is, the greatest common divisor of the set of integers n for which $p_{ii}^{(n)} > 0$. Then the state E_i is said to be periodic if $d(i) > 1$ and aperiodic if $d(i) = 1$.

Example 13.5.1. A four-state Markov chain has the transition matrix

$$T = \begin{bmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Show that all states have period 3.

Solution: The transition diagram is shown in Fig. 13.5.1, from which it is clear that all states are period 3. For example, if the chain start in E_1, then returns to E_1 are only possible at steps 3, 6, 9, ... either through E_2 or E_3.

The analysis of chains with periodic states can be complicated. However, one can check for a suspected

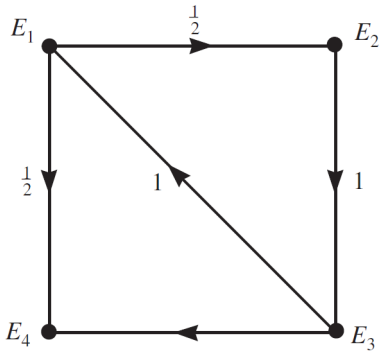


Figure 13.5.1: The transition diagram for Example 13.5.1

periodicity as follows. By direct computation

$$S = T^3 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}$$

In this example,

$$S^2 = T^6 = S \cdot S = S,$$

so that

$$S^r = T^{3r} = S, \quad (r = 1, 2, \ldots),$$

which always has non-zero elements on its diagonal. On the other hand,

$$S^{r+1} = S^r S = \begin{bmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad S^{r+2} = S^r S^2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 1 & 0 & 0 & 0 \end{bmatrix},$$

and both these matrices have zero diagonal elements for $r = 1, 2, 3 \ldots$. Hence, for $i = 1, 2, 3, 4$,

$$\begin{aligned} p_{ii}^{(n)} &= 0 \quad \text{for} \quad n \neq 3, 6, 9, \ldots, \\ p_{ii}^{(n)} &\neq 0 \quad \text{for} \quad n = 3, 6, 9, \ldots, \end{aligned}$$

which means that all the states are period 3.

(c) **Persistent State:** Let $f_j^{(n)}$ be the probability that the first return or visit to E_j occurs at the n-th step. This probability is not the same as $p_{jj}^{(n)}$ which is the probability that a return occurs at the n-th step, and includes possible returns at steps 1, 2, 3, ..., $n-1$ also. It follows that

$$p_{jj}^{(1)} (= p_{jj}) = f_j^{(1)}, \tag{13.5.2}$$

$$p_{jj}^{(2)} = f_j^{(2)} + f_j^{(1)} p_{jj}^{(1)}, \tag{13.5.3}$$

$$p_{jj}^{(3)} = f_j^{(3)} + f_j^{(1)} p_{jj}^{(2)} + f_j^{(2)} p_{jj}^{(1)}, \tag{13.5.4}$$

and, in general,

$$p_{jj}^{(n)} = f_j^{(n)} + \sum_{r=1}^{n-1} f_j^{(r)} p_{jj}^{(n-r)} \quad (n \geq 2). \tag{13.5.5}$$

The terms in Eqn.(13.5.4) imply that the probability of a return at the third step is the probability of a first return at the third step, or the probability of a first return at the first step and a return two steps later, or the probability of a first return at the second step and a return one step later.

Equations (13.5.2) and (13.5.5) become iterative formulas for the sequence of first returns $f_j^{(n)}$ which can be expressed as:

$$f_j^{(1)} = p_{jj}, \tag{13.5.6}$$

$$f_j^{(n)} = p_{jj}^{(n)} - \sum_{r=1}^{n-1} f_j^{(r)} p_{jj}^{(n-r)} \quad (n \geq 2). \tag{13.5.7}$$

The probability that a chain returns at some step to the state E_j is

$$f_j = \sum_{n=1}^{j} f_j^{(n)}.$$

If $f_j = 1$, then a return to E_j is certain, and E_j is called a persistent state.

Example 13.5.2. A three-state Markov chain has the transition matrix

$$T = \begin{bmatrix} p & 1-p & 0 \\ 0 & 0 & 1 \\ 1-q & 0 & q \end{bmatrix}$$

where $0 < p < 1,\ 0 < q < 1$. Show that the state E_1 is persistent.

Solution: For simple chains a direct approach using the transition diagram is often easier than the formula (13.5.7) for $f_j^{(n)}$. For this example the transition diagram is shown in Fig. 13.5.2.

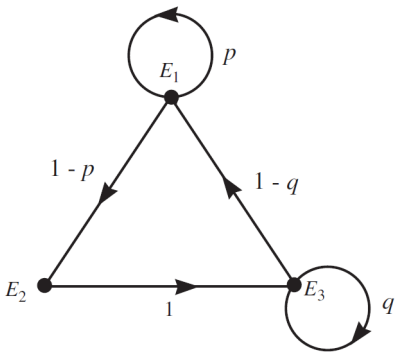


Figure 13.5.2: The transition diagram for Example 13.5.2

If a sequence starts in E_1, then it can be seen that first returns to E_1 can be made to E_1 at every step except for $n = 2$, since after two steps the chain must be in state E_3. From the figure it can be argued that

$$f_1^{(1)} = p, \quad f_1^{(2)} = 0, \quad f_1^{(3)} = (1-p)\cdot 1 \cdot (1-q),$$
$$f_1^{(n)} = (1-p)\cdot 1 \cdot q^{n-3} \cdot (1-q), \quad (n \geq 4).$$

The last result for $f_1^{(n)}$ for $n \geq 4$ follows from the following sequence of transitions:

$$E_1 E_2 \underbrace{E_3 E_3 \cdots E_3}_{(n-3) \text{ times}} E_1.$$

The probability f_1 that the system returns at least once to E_1 is

$$\begin{aligned} f_1 = \sum_{n=1}^{\infty} = f_1^{(n)} &= p + \sum_{n=3}^{\infty} (1-p)(1-q)q^{n-3} \\ &= p + (1-p)(1-q)\sum_{s=0}^{\infty} q^s \quad (s = n-3) \\ &= p + (1-p)\frac{(1-q)}{(1-q)} \\ &= 1 \end{aligned}$$

Hence, $f_1 = 1$, and consequently the state E_1 is persistent.

The mean recurrence time μ_j of a persistent state E_j, for which $\sum_{n=1}^{\infty} f_j^{(n)} = 1$, is given by

$$\mu_j = \sum_{n=1}^{\infty} n f_j^{(n)}. \tag{13.5.8}$$

In Example 13.5.2, the state E_1 is persistent and its mean recurrence time is given by

$$\begin{aligned}\mu_1 = \sum_{n=1}^{\infty} n f_1^{(n)} &= p + (1-p)(1-q)\sum_{n=3}^{\infty} nq^{n-3} \\ &= p + (1-p)(1-q)\left[\frac{3-2q}{(1-q)^2}\right] \\ &= \frac{3-2p-2q+pq}{1-q}\end{aligned}$$

which is finite. For some chains, however, the mean recurrence time can be infinite; in other words, the mean number of steps to a first return is unbounded.

A persistent state E_j is said to be null if $\mu_j = \infty$ and nonnull if $\mu_j < \infty$.

Example 13.5.3. A three-state inhomogeneous Markov chain has the transition matrix

$$T_n = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 \\ 1/(n+1) & 0 & n/(n+1) \end{bmatrix}$$

where T_n is transition matrix at step n. Show that E_1 is a persistent null state.

Solution: The transition diagram at a general step n is shown in Fig. 13.5.3 From the figure, we have

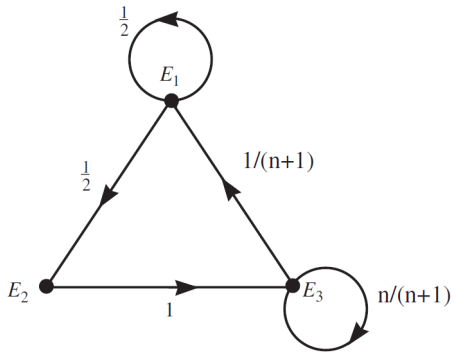


Figure 13.5.3: The transition diagram for Example 13.5.3

.

$$\begin{aligned}&f_1^{(1)} = \frac{1}{2}, \quad f_1^{(2)} = 0, \quad f_1^{(3)} = \frac{1}{2}\cdot 1 \cdot \frac{1}{4}, \\ &f_1^{(n)} = \frac{1}{2}\cdot 1 \cdot \frac{3}{4}\cdot\frac{4}{5}\cdots\frac{n-1}{n}\cdot\frac{1}{n+1} = \frac{3}{2n(n+1),} \quad (n \geq 4).\end{aligned}$$

Hence,

$$f_1 = \frac{1}{2} + \frac{1}{8} + \frac{3}{2}\sum_{n=4}^{\infty}\frac{1}{n(n+1)}.$$

Since

$$\frac{1}{n(n+1)} = \frac{1}{n} - \frac{1}{n+1},$$

it follows that

$$\sum_{n=4}^{\infty}\frac{1}{n(n+1)}=\lim_{N\to\infty}\sum_{n=4}^{N}\left(\frac{1}{n}-\frac{1}{n+1}\right)=\lim_{N\to\infty}\left(\frac{1}{4}-\frac{1}{N+1}\right)=\frac{1}{4}.$$

Hence

$$f_1=\frac{5}{8}+\frac{3}{8}=1,$$

which means E_1 is persistent. On the other hand, the mean recurrence time

$$\begin{aligned}\mu_j=\sum_{n=1}^{\infty}nf_1^{(n)} &= \frac{7}{8}+\frac{3}{2}\sum_{n=4}^{\infty}\frac{n}{n(n+1)}\\ &= \frac{7}{8}+\frac{3}{2}\left(\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\cdots\right)\\ &= \frac{7}{8}+\frac{3}{2}\sum_{n=5}^{\infty}\frac{1}{n}.\end{aligned}$$

The series in the previous equation is the harmonic series

$$\sum_{n=1}^{\infty}\frac{1}{n}=1+\frac{1}{2}+\frac{1}{3}+\cdots, \tag{13.5.9}$$

minus the first four terms. The harmonic series is a well-known divergent series, which means that $\mu_1=\infty$. Hence E_1 is persistent and null.

(d) **Transient state:** For a persistent state the probability of a first return at some step in the future is certain. For some states,

$$f_j=\sum_{n=1}^{\infty}f_j^{(n)}<1, \tag{13.5.10}$$

which means that the probability of a first return is not certain. Such states are described as transient.

Example 13.5.4. A four state Markov chain has the transition matrix

$$T=\begin{bmatrix}0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4}\\ \frac{1}{2} & \frac{1}{2} & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & \frac{1}{2} & \frac{1}{2}\end{bmatrix}$$

Show that E_1 is a transient state.

Solution: The transition diagram is shown in Fig. 13.5.4. From the figure

$$f_1^{(1)}=0,\ \ f_1^{(2)}=\frac{1}{2}\cdot\frac{1}{2}=\left(\frac{1}{2}\right)^2,\ \ f_1^{(3)}=\left(\frac{1}{2}\right)^3,\ \ f_1^{(n)}=\left(\frac{1}{2}\right)^n.$$

Hence

$$f_1=\sum_{n=1}^{\infty}f_1^{(n)}=\sum_{n=2}^{\infty}\left(\frac{1}{2}\right)^n=\frac{1}{2}<1$$

implying that E_1 is a transient state. The reason for the transience of E_1 can be seen from Fig. 13.5.4, where transitions from E_3 or E_4 to E_1 or E_2 are not possible.

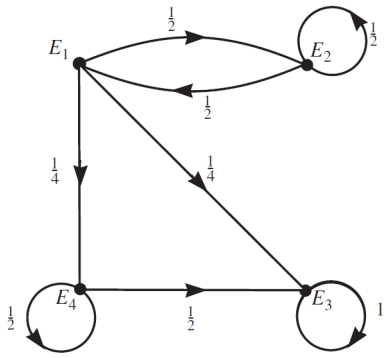


Figure 13.5.4: The transition diagram for Example 13.5.4

(d) **Ergodic states:** A state which is persistent, nonnull and aperiodic is called ergodic state.

Example 13.5.5. A three-state Markov chain has the transition matrix

$$T = \begin{bmatrix} p & 1-p & 0 \\ 0 & 0 & 1 \\ 1-q & 0 & q \end{bmatrix}$$

where $0 < p < 1,\ 0 < q < 1$. Show that the state E_1 is ergodic.

Solution: It was already shown in Example 13.5.2, that E_1 is persistent with

$$f_1^{(1)} = p, \quad f_1^{(2)} = 0, \quad f_1^{(n)} = (1-p)(1-q)q^{n-3}, \ \ (n \geq 3).$$

It follows that its mean recurrence time is

$$\mu_1 = \sum_{n=1}^{\infty} n f_1^{(1)} = p + (1-p)(1-q)\sum_{n=3}^{\infty} nq^{n-3} = \frac{3-2q}{(1-q)^2} < \infty.$$

The convergence of μ_1 implies that E_1 is nonnull. Also the diagonal elements $p_{ii}^{(n)} > 0$ for $n \geq 3$ and $i = 1, 2, 3$, which means that E_1 is aperiodic. Hence from the definition above E_1 (and E_2 and E_3 also) is ergodic.

Unit 14

Course Structure

- Statistical Inference
- Estimation of Parameters
- Minimum Variance Unbiased Estimator
- Method of Maximum Likelihood for Estimation of a parameter

14.1 Introduction

To study the features of any population we first select a sample from the population. A carefully selected sample may be expected to possess the characteristics of the population. A scientific theory developed to get an idea regarding the properties of a population on the basic of the knowledge of the properties of a sample drawn from it is known as *Statistical Inference*.

Statistical Inference may be classified into two main categories :

(i) Problems of Estimation.

(ii) Problems of Testing of Hypothesis or Testing of Significance.

14.2 Estimation of Parameters

Let the distribution function of a population contains one or more unknown parameters and let our task is to make a guess about them on the basis of a sample. The theory regarding this is called *theory of estimation.* In particular, let $x_1, x_2, \ldots, x_n$ be n samples drawn from a population whose distribution has an unknown parameter θ. The problem to replace this unknown θ by a suitable statistic (i.e., a function of the sample values) $\hat{\theta}(x_1, x_2, \ldots, x_n)$ is the problem of estimation.

There are two types of estimation:

(i) Point Estimation, and (ii) Interval Estimation.

In the case of point estimation, the value of θ may vary from sample to sample and this function is known as *'estimator'* of the parameter and its value for a particular sample is called and *'estimate'*.

In the case of interval estimation, two statistics $\hat{\theta}_1(x_1, x_2, \ldots, x_n)$ and $\hat{\theta}_2(x_1, x_2, \ldots, x_n)$ are selected within which the value of the parameter θ is expected to lie. This interval is known as Confidence Interval and the two quantities used to specify the interval are known as Confidence Limits.

According to R. A. Fisher, a good estimator must have the following characteristics:

(i) Unbiasedness,

(ii) Consistency,

(iii) Efficiency,

(iv) Sufficiency.

14.3 Unbiasedness

A statistic T is said to be an unbiased estimator of a parameter θ if the expected value of the statistic coincides with the actual value of the parameter, i.e., if

$$E(T) = \theta$$

Otherwise, the estimation will be called biased. $E(T) - \theta$ is called the bias of the statistic T in estimating θ. It will be called positively or negatively biased according as $E(T) - \theta$ is greater or less than zero.

Theorem 14.3.1. The sample mean is an unbiased estimate of the population mean.

Proof. Let $x_1, x_2, \ldots, x_n$ be n simple samples drawn from a finite population $X_1, X_2, \ldots, X_N$ with replacement. In this case, each x_i have equal chance to be selected from any of the N population values. Therefore,

$$\begin{aligned} E(x_i) &= \frac{1}{N}X_1 + \frac{1}{N}X_2 + \ldots + \frac{1}{N}X_N \\ &= \frac{1}{N}(X_1 + X_2 + \ldots + X_N) \\ &= m, \quad \text{the population mean,} \quad i = 1, 2, \ldots, n \end{aligned} \tag{14.3.1}$$

Again,

$$\begin{aligned} \overline{x} &= \text{sample mean} \\ &= \frac{x_1 + x_2 + \ldots + x_n}{n} \end{aligned}$$

Now,

$$\begin{aligned} E(\overline{x}) &= E\left(\frac{x_1+x_2+\ldots+x_n}{n}\right) \\ &= \frac{1}{n}\Big[E(x_1)+E(x_2)+\ldots+E(x_n)\Big] \\ &= \frac{1}{n}\Big[m+m+\ldots+m\Big] \\ &= \frac{nm}{n} \\ &= m = \text{ population mean} \end{aligned} \tag{14.3.2}$$

Therefore, the sample mean $\overline{x}$ is an unbiased estimate of the population mean m. □

Theorem 14.3.2. The sample variance is a biased estimator of the population variance.

Proof. Let m and σ^2 be the population mean and variance respectively and let $\overline{x}$ and S^2 be the corresponding sample mean and variance.

It is easy to note that $E(x_i) = m$ and $Var(x_i) = E\{(x_i - m)^2\} = \sigma^2$ for $i = 1, 2, \ldots, n$.

Again, Sample mean $\overline{x} = \dfrac{x_1+x_2+\ldots+x_n}{n}$ and sample variance

$$\begin{aligned} S^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-m+m-\overline{x})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\{(x_i-m)^2 - 2(x_i-m)(\overline{x}-m)+(\overline{x}-m)^2\} \\ &= \frac{1}{n}\left[\sum_{i=1}^{n}(x_i-m)^2 - 2(\overline{x}-m)\sum_{i=1}^{n}(x_i-m)+n(\overline{x}-m)^2\right] \\ &= \frac{1}{n}\left[\sum_{i=1}^{n}(x_i-m)^2 - 2(\overline{x}-m)\left(\sum_{i=1}^{n}x_i-nm\right)+n(\overline{x}-m)^2\right] \\ &= \frac{1}{n}\left[\sum_{i=1}^{n}(x_i-m)^2 - 2(\overline{x}-m)(n\overline{x}-nm)+n(\overline{x}-m)^2\right] \\ &= \frac{1}{n}\left[\sum_{i=1}^{n}(x_i-m)^2\right] - (\overline{x}-m)^2. \end{aligned}$$

Therefore,

$$\begin{aligned} E(S^2) &= E\left\{\frac{1}{n}\sum_{i=1}^{n}(x_i-m)^2\right\} - E\{(\overline{x}-m)^2\} \\ &= \frac{1}{n}\sum_{i=1}^{n} E\{(x_i-m)^2\} - Var(\overline{x}) \\ &= \frac{\sum_{i=1}^{n}\sigma^2}{n} - \frac{\sigma^2}{n} \\ &= \sigma^2 - \frac{\sigma^2}{n} \\ &= \frac{n-1}{n}\sigma^2 \end{aligned}$$

Since, $E(S^2) \neq \sigma^2$, so S^2 is not an unbiased estimate of σ^2. Again,

$$\begin{aligned} \text{bias} &= E(S^2) - \sigma^2 \\ &= \frac{n-1}{n}\sigma^2 - \sigma^2 \\ &= -\frac{\sigma^2}{n}. \end{aligned}$$

Again, if we write

$$s^2 = \frac{n}{n-1}S^2, \tag{14.3.3}$$

$$\begin{aligned} \text{then} \quad E(s^2) &= \frac{n}{n-1}E(S^2) \\ &= \frac{n}{n-1}\cdot\frac{n-1}{n}\sigma^2 \\ &= \sigma^2. \end{aligned}$$

Thus s^2 as defined by (14.3.3) is an unbiased estimate of σ^2. □

14.4 Minimum-Variance Unbiased (M.V.U.) Estimator

Among all the unbiased estimators the minimum-variance unbiased estimator will be that one which has the minimum variance. Thus, if T_m be the minimum-variance unbiased estimator of any parameter θ, then $E(T_m) = \theta$ and Var$(T_m) <$ Var(T), where T is any other unbiased estimator of θ, i.e., $E(T) = \theta$.

14.4.1 Consistent Estimator:

A statistic T_n computed from a sample of n observations is said to be a consistent estimator of a population parameter θ if

$$T_n \xrightarrow[\text{in P}]{} \theta \quad \text{as} \quad n \to \infty. \tag{14.4.1}$$

In other notation,

$$\lim_{n\to\infty} P(|T_n - \theta| < \epsilon) = 1 \tag{14.4.2}$$

or its equivalent,

$$\lim_{n\to\infty} P(|T_n - \theta| \geq \epsilon) = 0 \tag{14.4.3}$$

Thus, a consistent estimator is expected to come more closer to the parameter as the size of the sample becomes larger.

It may be shown that two sufficient conditions for an estimator T_n to be consistent estimator of θ are

$$\text{(i) } E(T_n) \to \theta \text{ and (ii) Var } (T_n) \to 0 \text{ as } n \to \infty.$$

14.5 Efficient Estimator

Among all consistent estimators that one which has minimum asymptotic variance is called the most efficient estimator. Thus a consistent estimator T_n' is said to be most efficient estimator if its sampling variance is less than that of any other consistent estimator T_n, i.e., in this case

$$\text{Var}(T_n') < \text{Var}(T_n).$$

If V_m be the variance of the most efficient estimator and V be the variance of another estimator for a parameter θ, then the efficiency of the estimator is defined as

$$\text{Efficiency} = \frac{V_m}{V}.$$

Since, $V_m \leq V$, so efficiency cannot exceed 1.

14.6 Sufficient Estimator

A statistic T is said to be a sufficient estimator for a parameter θ if it contains all information in the sample about θ. In this case, the joint distribution of the sample can be expressed as the product of two factors, one of which is the sampling distribution of T and contains θ, but the other factor is independent of θ.

Thus for a random sample $x_1, x_2, \ldots, x_n$ from a population whose probability density function (p.m.f) is $f(x, \theta)$ if T be a sufficient estimator of θ, then

$$f(x_1, \theta) \cdot f(x_2, \theta) \ldots f(x_{n_1}, \theta) = f_1(T, \theta) \cdot f_2(x_1, x_2, \ldots, x_n)$$

where $f_1(T, \theta)$ is the sampling of T and contains θ, but $f_2(x_1, x_2, \ldots, x_n)$ is independent of θ.

14.7 Method of Maximum Likelihood for Estimation of a parameters

There are many methods generally used for estimation of parameters of a distribution. Among these, the Method of Maximum Liklihood is one of the most familiar methods.

Let $x_1, x_2, \ldots, x_n$ be a random sample of size n drawn from a population and let $\theta_1, \theta_2, \ldots, \theta_k$ be k parameters of the distribution. This event can be denoted by $(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$ and the probability of this is clearly a function of sample values $x_1, x_2, \ldots, x_n$ and the parameters $\theta_1, \theta_2, \ldots, \theta_k$. This function is known as likelihood function of the sample and it is generally denoted by $L(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_k)$. Thus,

$$L(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_k) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

Since $X_1, X_2, \ldots, X_n$ are mutually independent random variables each having the distribution of the population, then in the discrete case

$$P(X = x_i) = f_{x_i}(\theta_1, \theta_2, \ldots, \theta_k)$$

and in the continuous case

$$P(X = x_i) = f(x_i, \theta_1, \theta_2, \ldots, \theta_k).$$

Then the liklihood function L in the two cases are given as

$$\begin{aligned} L(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_n) &= P(X_1 = x_1)P(X_2 = x_2)\ldots P(X_n = x_n) \\ \text{(In discrete case)} &= f_{x_1}(\theta_1, \theta_2, \ldots, \theta_k) f_{x_2}(\theta_1, \theta_2, \ldots, \theta_k)\ldots f_{x_n}(\theta_1, \theta_2, \ldots, \theta_k) \\ \text{(In continuous case)} &= f(x_1, \theta_1, \theta_2, \ldots, \theta_k) f(x_2, \theta_1, \theta_2, \ldots, \theta_k)\ldots f(x_n, \theta_1, \theta_2, \ldots, \theta_k) \end{aligned}$$

Now, this method states that regarding the sample values as fixed, we shall try to find the values of $\theta_1, \theta_2, \ldots, \theta_k$ such that for these values the likelihood function L will be maximised. Since $L > 0$, so when L is maximum, then $\log L$ is also maximum. The corresponding equations for determining $\theta_1, \theta_2, \ldots, \theta_k$ are

$$\frac{\partial \log L}{\partial \theta_1} = 0, \quad \frac{\partial \log L}{\partial \theta_2} = 0, \quad \ldots \quad , \frac{\partial \log L}{\partial \theta_k} = 0,$$

which are called *likelihood equations.* Solving these k equations we get likelihood estimates of $\theta_1, \theta_2, \ldots, \theta_k$ and they are generally denoted by

$$\theta_1 = \hat{\theta_1}(x_1, x_2, \ldots, x_n), \theta_2 = \hat{\theta_2}(x_1, x_2, \ldots, x_n), \ldots, \theta_k = \hat{\theta_k}(x_1, x_2, \ldots, x_n).$$

Also it may tested that for these values of $\hat{\theta_1}, \hat{\theta_2}, \ldots, \hat{\theta_k}$, L is maximum.

Example 14.7.1. Let T_1 and T_2 be two estimators of the parameter θ. Under what condition $aT_1 + bT_2$ will be an unbiased estimator of θ?

Solution: Since T_1 and T_2 are two unbiased estimator of θ, so $E(T_1) = E(T_2) = \theta$. Again, if $(aT_1 + bT_2)$ be an unbiased estimator of θ, then

$$\begin{aligned} & E(aT_1 + bT_2) = \theta \\ \Rightarrow\ & aE(T_1) + bE(T_2) = \theta \\ \Rightarrow\ & a\theta + b\theta = \theta \\ \Rightarrow\ & a + b = 1, \quad \text{which is the required condition.} \end{aligned}$$

Example 14.7.2. If $X_1, X_2, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$ population, show that the estimator

$$T = \frac{1}{n+1}\sum_{i=1}^{n} X_i$$

is a biased but consistent for μ. Hence obtain the unbiased estimator for μ.

Solution: We have

$$\begin{aligned} T &= \frac{1}{n+1}\sum_{i=1}^{n} X_i \\ &= \frac{n}{n+1} \cdot \frac{1}{n}\sum_{i=1}^{n} X_i \\ &= \frac{n}{n+1}\overline{X} \end{aligned}$$

We know, $X \xrightarrow{\text{in p}}$ as $n \to \infty$ and $\dfrac{n}{n+1} \to 1$ as $n \to \infty$. So

$$T \xrightarrow{\text{in p}} \mu \quad \text{as} \quad n \to \infty.$$

Thus, T is a consistent estimator of μ. Again

$$E(T) = \frac{1}{n+1}\sum_{i=1}^{n} E(X_i) = \frac{1}{n+1} n\mu = \frac{n}{n+1}\mu \quad (\neq \mu)$$

So T is a biased estimator of μ. If we put $T_1 = \dfrac{n+1}{n}$, then, $E(T_1) = \dfrac{n+1}{n}E(T) = \mu$.
Thus, $T_1 = \dfrac{n+1}{n}T$ is the unbiased estimator for μ.

Example 14.7.3. Maximum likelihood estimate of the parameter p of the Binomial (N, p) population for n sample values.

Solution: For Binomial (N, P) population, the density function is given by

$$f_{x_i} = {}^{N}C_{x_i} p^{x_i}(1-p)^{N-x_i}, \quad i = 0, 1, 2, \ldots, n.$$

Now, the likelihood function L is given by

$$\begin{aligned} L &= f_{x_1} \cdot f_{x_2} \cdots f_{x_n} \\ &= {}^{N}C_{x_1} p^{x_1}(1-p)^{N-x_1} \cdot {}^{N}C_{x_2} p^{x_2}(1-p)^{N-x_2} \cdots {}^{N}C_{x_n} p^{x_n}(1-p)^{N-x_n} \\ &= {}^{N}C_{x_1}\, {}^{N}C_{x_2} \cdots {}^{N}C_{x_n} p^{x_1+x_2+\cdots+x_n}(1-p)^{nN-(x_1+x_2+\ldots+x_n)} \end{aligned}$$

So,

$$\log L = (x_1 + x_2 + \ldots + x_n)\log p + [nN - (x_1 + x_2 + \ldots + x_n)]\log(1-p) + \text{terms independent of } p.$$

Now,

$$\frac{\partial \log L}{\partial p} = \frac{n\overline{x}}{p} = \frac{nN - n\overline{x}}{1-p} \quad \left[\because \overline{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)\right]$$

Thus, $\dfrac{\partial \log L}{\partial p} = 0$ gives

$$\begin{aligned} \frac{n\overline{x}}{p} &= \frac{nN - n\overline{x}}{1-p} \\ \Rightarrow p &= \frac{\overline{x}}{N}. \end{aligned}$$

Thus, $\hat{p} = \dfrac{\overline{x}}{N}$ is the likelihood estimate of p.

It can be verified that

$$\left[\frac{\partial^2 L}{\partial p}\right]_{p=\hat{p}} < 0.$$

Example 14.7.4. Maximum likelihood estimator of the parameter of a Poisson distribution.

Solution: Let $x_1, x_2, \ldots, x_n$ be n sample values drawn from a Poisson distribution having parameter μ. Then

$$f(x, \mu) = \frac{e^{-\mu}\mu^x}{x!} \quad (x = 0, 1, 2, \ldots, \infty)$$

The likelihood function L of the sample observations is given by

$$\begin{aligned} L &= f(x_1, \mu) \cdot f(x_2, \mu) \cdots f(x_n, \mu) \\ &= \frac{e^{-\mu}\mu^{x_1}}{x_1!} \cdot \frac{e^{-\mu}\mu^{x_2}}{x_2!} \cdots \frac{e^{-\mu}\mu^{x_n}}{x_n!} \\ &= \frac{e^{-n\mu}\mu^{x_1+x_2+\ldots+x_n}}{(x_1!)(x_2!)\ldots(x_n!)} \end{aligned} \tag{14.7.1}$$

So

$$\begin{aligned} \log L &= \log(e^{-n\mu}) + \log(\mu^{x_1+x_2+\ldots+x_n}) - \log(x_1!\, x_2!\, \cdots x_n!) \\ &= -n\mu + \left(\sum_{i=1}^{n} x_i\right) \log \mu - \sum_{i=1}^{n} \log(x_i!). \end{aligned}$$

If $\hat{\mu}$ be the likelihood estimator of μ, then it will be given by

$$\left[\frac{\partial \log L}{\partial \mu}\right]_{\mu=\hat{\mu}} = 0 \quad \text{and} \quad \left[\frac{\partial^2 \log L}{\partial \mu^2}\right]_{\mu=\hat{\mu}} < 0.$$

From above, we have

$$\frac{\partial \log L}{\partial \mu} = -n + \frac{1}{\mu} + \sum_{i=1}^{n} x_i \quad \text{and} \quad \frac{\partial^2 \log L}{\partial \mu^2} = -\frac{1}{\mu^2} + \sum_{i=1}^{n} x_i$$

Now, $\dfrac{\partial \log L}{\partial \mu} = 0$ gives,

$$\begin{aligned} -n + \frac{1}{\mu}\sum_{i=1}^{n} x_i &= 0 \\ \Rightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} x_i &= \overline{x}. \end{aligned}$$

Therefore, $\overline{\mu} = \overline{x}$. Again

$$\left[\frac{\partial^2 \log L}{\partial \mu^2}\right]_{\mu=\hat{\mu}} = -\frac{1}{\hat{\mu}^2}\sum_{i=1}^{n} x_i = -\frac{n\overline{x}}{(\overline{x})^2} = -\frac{n}{\overline{x}} < 0.$$

Thus, $\hat{\mu} = \overline{x}$, the sample means is the likelihood estimate of the parameter μ of a Poisson distribution.

Example 14.7.5. Maximum likelihood estimates of the parameter m and σ in Normal (m, σ) population for a sample of size n.

Solution: We know for a Normal (m, σ) distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2}\sigma^2}, \quad -\infty < x < \infty.$$

So, the likelihood function L is given by

$$\begin{aligned} L &= f(x_1)f(x_2)\cdots f(x_n) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{\sum_{i=1}^{n}(x_i-m)^2}{2\sigma^2}}. \end{aligned}$$

Then,

$$\log L = -n\log(\sqrt{2\pi}) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-m)^2.$$

Then, $\dfrac{\partial L}{\partial m} = 0$ gives $\displaystyle\sum_{i=1}^{n}(x_i-m) = 0 \Rightarrow m = \frac{\sum_{i=1}^{n} x_i}{n} \Rightarrow \hat{m} = \overline{x}.$

Also, $\dfrac{\partial \log L}{\partial \sigma} = 0$ gives

$$\begin{aligned} -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-m)^2 &= 0 \\ \Rightarrow \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-m)^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 = S^2. \end{aligned}$$

Thus, $\hat{\sigma}^2 = S^2$.

Example 14.7.6. Find the maximum likelihood estimate of the parameter λ for the Weibuzl distribution

$$f(x) = \lambda\alpha x^{\alpha-1}e^{-\lambda x^\alpha}, \quad (x > 0)$$

using a sample of size n assuming that α is known.

Solution: If $x_1, x_2, \ldots, x_n$ be n sample values, then the maximum likelihood function L is given by

$$L = \lambda^n\alpha^n(x_1, x_2, \ldots, x_n)^{n-1}e^{-\lambda(x_1^\alpha + x_2^\alpha + \ldots + x_n^\alpha)}$$

Then,

$$\log L = n\log\lambda - \lambda(x_1^\alpha + x_2^\alpha + \ldots + x_n^\alpha) \ + \text{ terms independent of } \lambda.$$

So, the likelihood equation $\dfrac{\partial \log L}{\partial \lambda} = 0$ gives

$$\begin{aligned} \frac{n}{\lambda} - (x_1^\alpha + x_2^\alpha + \ldots + x_n^\alpha) &= 0 \\ \Rightarrow \hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i^\alpha} & \end{aligned}$$

Again $\left[\dfrac{\partial^2 \log L}{\partial \lambda^2}\right]_{\lambda=\hat{\lambda}} = -\dfrac{n}{\lambda^2}$. So for this $\hat{\lambda}$, L is maximum.

Example 14.7.7. Prove that the maximum likelihood estimate of the parameter α of a population having density function

$$\frac{2}{\alpha^2}(\alpha - x), \quad 0 < x < \alpha,$$

for a sample of unit size is $2x$, x being the sample value. Show also that the estimate is biased.

Solution: Since the sample is of unit size, so the likelihood function L is given by

$$\begin{aligned} L &= \frac{2}{\alpha^2}(\alpha - x) \\ \Rightarrow \log L &= \log 2 - 2\log\alpha + \log(\alpha - x) \end{aligned}$$

Now, the likelihood equation $\dfrac{\partial \log L}{\partial \alpha} = 0$ gives

$$\begin{aligned} -\frac{2}{\alpha} + \frac{1}{\alpha - x} &= 0 \\ \Rightarrow \alpha &= 2x \end{aligned}$$

So, $\hat{\alpha} = 2x$. Also it can be shown that

$$\frac{\partial^2 \log L}{\partial \alpha^2} = \frac{2}{\alpha^2} - \frac{1}{(\alpha - x)^2} < 0 \quad \text{for} \quad \hat{\alpha} = 2x.$$

Thus, the maximum likelihood estimate of α is $\hat{\alpha} = 2x$. Again,

$$\begin{aligned} E(2x) &= \int_0^\alpha 2x \cdot \frac{2}{\alpha^2}(\alpha - x)\, dx = \frac{4}{\alpha^2}\int_0^\alpha (\alpha x - x^2)\, dx \\ &= \frac{4}{\alpha^2}\left[\alpha\frac{x^2}{2} - \frac{x^3}{3}\right]_0^\alpha = \frac{4}{\alpha^2}\left[\frac{\alpha^3}{2} - \frac{\alpha^3}{3}\right] \\ &= \frac{2\alpha}{3} \neq \alpha \end{aligned}$$

Thus, $\hat{\alpha}$ is a biased estimate of α.

Unit 15

Course Structure

- Interval estimation
- Method for finding confidence intervals
- Statistical hypothesis
- Level of significance; Power of the test

15.1 Introduction

We have studied the problem of estimation of a parameter occurring in a distribution such an estimate is called parameter estimate and the corresponding problem is known as the problem of estimation. Such and estimate always associated with random error. For this reason, it is sometime desirable to find a $\delta > 0$ for a given small ϵ where $0 < \epsilon < 1$ such that an estimate $\hat{\theta}$ for the parameter θ satisfies

$$P(\hat{\theta} - \delta < \theta < \hat{\theta} + \delta) = 1 - \epsilon$$

15.2 Interval Estimation

Let θ be a population parameter and let T_1 and T_2 be two functions based on sample observations such that

$$P(T_1 \leq \theta \leq T_2) = 1 - \epsilon \tag{15.2.1}$$

where $\epsilon(0 < \epsilon < 1)$ is a parameter. Then the interval (T_1, T_2) is called an interval estimate or a confidence interval for the parameter θ with confidence coefficient $1 - \epsilon$; the statistics T_1 and T_2 are respectively called the lower and upper confidence limits for θ.

A practical interpretation of this result is that if a long sequence of random samples, are drawn from a population under uniform conditions and the statistics T_1 and T_2 are computed in each time, then

$$\frac{\text{The number of times the interval } (T_1, T_2) \text{ includes the true parameter } \theta}{\text{The total number of samples drawn}} = 1 - \epsilon$$

The number ϵ is usually chosen to be very small, like 0.05, 0.01, 0.001 etc. and the corresponding confidence coefficients are 0.95, 0.99, 0.999 etc. and then the corresponding confidence intervals will be called 95%, 99%, 99.9% etc. confidence intervals.

The length of the interval $(T_2 - T_1)$ is used as an inverse measure of precision of the interval estimate.

15.3 Method for finding confidence intervals

To find the confidence interval for a parameter θ, the following steps to be followed.

1. We choose, if possible, a suitable statistic $z = z(x_1, x_2, \ldots, x_n, \theta)$ whose sampling distribution is independent of the parameter θ but which itself depends on θ.

2. Now we choose two numbers $\alpha_\epsilon, \beta_\epsilon (> \alpha_\epsilon)$ such that

$$P(\alpha_\epsilon < z < \beta_\epsilon) = 1 - \epsilon \tag{15.3.1}$$

3. We rewrite the above equation (15.3.1) as

$$P(T_1 < \theta < T_2) = 1 - \epsilon \tag{15.3.2}$$

Then (T_1, T_2) is the desired confidence interval for the population parameter θ.

15.4 Confidence interval for some special cases

(a) The confidence interval for m for a Normal (m, σ) population.
Case 1. σ known: The suitable statistic for this case will be chosen as

$$z = \frac{\overline{x} - m}{\sigma/\sqrt{n}}$$

whose sampling distribution is normal $(0, 1)$ and which depends on the parameter m.

Since normal curve is symmetrical curve, so we take two points $\pm u_\epsilon$ symmetrically about the origin, Fig. 15.4.1, such that

$$\begin{aligned} &P(-u_\epsilon < z < u_\epsilon) = 1 - \epsilon \\ \Rightarrow\ &P\left(-u_\epsilon < \frac{\overline{x} - m}{\sigma/\sqrt{n}} < u_\epsilon\right) = 1 - \epsilon \end{aligned}$$

which can be rewritten as

$$P\left(\overline{x} - \frac{\sigma u_\epsilon}{\sqrt{n}} < m < \overline{x} + \frac{\sigma u_\epsilon}{\sqrt{n}}\right) = 1 - \epsilon.$$

Hence a confidence interval for m having confidence coefficient $1 - \epsilon$ is

$$\left(\overline{x} - \frac{\sigma u_\epsilon}{\sqrt{n}}, \overline{x} + \frac{\sigma u_\epsilon}{\sqrt{n}}\right) \tag{15.4.1}$$

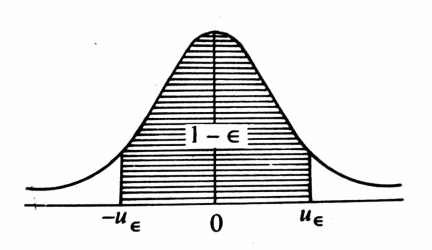


Figure 15.4.1

where u_ϵ is given by $P(-u_\epsilon < z < u_\epsilon) = 1 - \epsilon$ or from symmetry $P(z > u_\epsilon) = \frac{1}{2}\epsilon$. For 95% confidence interval, $1 - \epsilon = 0.95$ and $u_\epsilon = 1.96$, then the corresponding confidence interval for the population mean m will be

$$\left(\overline{x} - 1.96\frac{\sigma}{\sqrt{m}}, \overline{x} + 1.96\frac{\sigma}{\sqrt{m}}\right) \tag{15.4.2}$$

Case II: σ unknown: In this case, the suitable statistic will be

$$t = \frac{\overline{x} - m}{s/\sqrt{n}}, \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

whose sampling distribution has t-distribution with $n - 1$ degrees of freedom.

Now procedding exactly as in the Case I, we can calculate two numbers $\pm t_\epsilon$, Fig. 15.4.2, by

$$P(-t_\epsilon < t < t_\epsilon) = 1 - \epsilon$$

which gives the confidence interval of m as

$$\left(\overline{x} - \frac{st_\epsilon}{\sqrt{n}}, \overline{x} + \frac{st_\epsilon}{\sqrt{n}}\right) \tag{15.4.3}$$

Here t_ϵ is given by $P(-t_\epsilon < t < t_\epsilon) = 1 - \epsilon$ or by $P(t > t_\epsilon) = \frac{1}{2}\epsilon$. In case of large samples, if σ is unknown, then the approximate call interval for m may be obtained by replacing σ by s or S in (15.4.2).

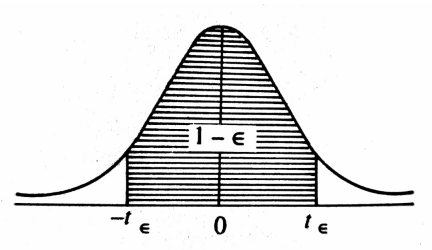


Figure 15.4.2

(b) Confidence interval for σ

It is known that the statistic

$$\chi^2 = \frac{nS^2}{\sigma^2}$$

is a χ^2-distributed with $(n-1)$ degrees of freedom, where S^2 is the sample variance, σ is the population variance and n is the size of the sample.

We choose any positive number $\chi^2_{\epsilon_1}$ and determine $\chi^2_{\epsilon_2}$ such that

$$\begin{aligned} & P(\chi^2_{\epsilon_1} < \chi^2 < \chi^2_{\epsilon_2} = 1-\epsilon \\ \Rightarrow\ & P\left(\chi^2_{\epsilon_1} < \frac{nS^2}{\sigma^2} < \chi^2_{\epsilon_2}\right) = 1-\epsilon \\ \Rightarrow\ & P\left(S\sqrt{\frac{n}{\chi^2_{\epsilon_2}}} < \sigma < S\sqrt{\frac{n}{\chi^2_{\epsilon_1}}}\right) = 1-\epsilon \end{aligned}$$

Therefore, $\left(S\sqrt{\frac{n}{\chi^2_{\epsilon_2}}}, S\sqrt{\frac{n}{\chi^2_{\epsilon_1}}}\right)$ is the confidence interval for σ having confidence coefficient $1-\epsilon$.

In practice, $\chi^2_{\epsilon_1}$ and $\chi^2_{\epsilon_2}$ are given by

$$P(\chi^2 > \chi^2_{\epsilon_1}) = 1 - \frac{1}{2}\epsilon \quad \text{and} \quad P(\chi^2 > \chi^2_{\epsilon_2}) = \frac{1}{2}\epsilon. \tag{15.4.4}$$

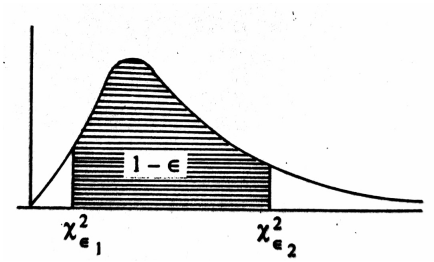


Figure 15.4.3

Example 15.4.1. A sample 2.3, -0.2, -0.4, -0.9 is taken from a normal population with variance 9. Find a 95% confidence interval for the population mean. (Given $P(U) > 1.960) = 0.025$, where U is a normal $(0,1)$ variate.

Solution: With usual notation, we have

$$\overline{x} = \frac{2.3 + (-0.2) + (-0.4) + (-0.9)}{4} = 0.2$$

Also, $n = 4$ and $\sigma^2 = 9,\ \epsilon = 0.05,\ u_\epsilon = 1.96$. Hence, the confidence interval for mean when σ is known is

$$\begin{aligned} & \left(\overline{x} - 1.96\frac{\sigma}{\sqrt{m}}, \overline{x} + 1.96\frac{\sigma}{\sqrt{m}}\right) \\ &= \left(0.2 - 1.96 \times \frac{3}{2},\ \ 0.2 + 1.96 \times \frac{3}{2}\right) \\ &= (-2.74,\ 3.14) \end{aligned}$$

Example 15.4.2. The mean and variance of a sample of size 400 from a normal population are found to be 18.35 and 3.25 respectively. Given $P(U > 1.96) = 0.025$, U being a standard normal variate, find 95% confidence interval for the population mean.

Solution: From the given data

$$\overline{x} = 18.35,\ S^2 = 3.25,\ n = 400,\ \epsilon = 0.05$$

Now

$$s^2 = \frac{n}{n=1} S^2 = \frac{400}{399} \times 3.25 = 3.26]$$

Hence the confidence interval for mean when σ is unknown is

$$\begin{aligned}
&\left(\overline{x} - 1.96\frac{s}{\sqrt{n}}, \overline{x} + 1.96\frac{s}{\sqrt{n}}\right) \\
&= \left(18.35 - 1.96 \times \frac{1.80}{20},\ 18.35 + 1.96 \times \frac{1.80}{20}\right) \\
&= (18.17,\ 18.53)
\end{aligned}$$

Example 15.4.3. Obtain 99% confidence interval of the population standard deviation (σ) on the basis of the data $\sum_{i=1}^{10} x_i = 620$ and $\sum_{i=1}^{10} x_i^2 = 39016$. (It is given that $\chi^2_{0.005,9} = 23.59$ and $\chi^2_{0.995,9} = 1.74$)

Solution:

$$\begin{aligned}
\text{Sample variance } S^2 &= \frac{1}{10}\sum_{i=1}^{10} x_i^2 - \left(\frac{1}{10}\sum_{i=1}^{10} x_i\right)^2 \\
&= \frac{39016}{10} - \left(\frac{620}{10}\right)^2 \\
&= 3901.6 - 3844 \\
&= 57.6
\end{aligned}$$

For 99% confidence interval

$$1 - \epsilon = 0.99,\ i.e.,\ \epsilon = 0.01 \text{ and } \frac{1}{2}\epsilon = 0.005$$

Here, $n = 10$. We know that confidence interval of σ is

$$\begin{aligned}
\left(S\sqrt{\frac{n}{\chi^2_{\epsilon_2}}}, S\sqrt{\frac{n}{\chi^2_{\epsilon_1}}}\right) &= \left(\sqrt{\frac{57.6 \times 10}{23.59}},\ \sqrt{\frac{57.6 \times 10}{1.74}}\right) \\
&= (4.94,\ 18.19)
\end{aligned}$$

15.5 Statistical Hypothesis

To make decision regarding a statistical population on the basis of sample observation is called a *Statistical Hypothesis*. It is an assertion or conjecture about the distribution of one or more random variables.

There are two types of hypothesis, viz. simple and composite. When a statistical hypothesis completely specifies the population distribution, it will be called a simple hypothesis and when it will not completely specify the population distribution, it will be called a composite hypothesis. In the case of composite hypothesis the number of unspecified parameters is called the degrees of freedom of the composite hypothesis.

As an illustration, let us consider a Normal (m, σ) distribution and let m_0 and σ_0 be taken to be two given values of m and σ respectively. Then

(i) Hypothesis $m = m_0$ is simple if σ is known and m is unknown.

(ii) Hypothesis $\sigma = \sigma_0$ is simple if m is known and σ is unknown.

(iii) Hypothesis $m = m_0$ is composite if both m and σ are unknown and its degrees of freedom is 1.

15.6 Null Hypothesis and Alternative Hypothesis

Let a population has only one parameter θ. Then a hypothesis about the parameter θ which we want to test is called Null Hypothesis. This is generally written as

$$H_0 : \theta = \theta_0$$

Any other hypothesis about the parameter θ against which we wish to test the null hypothesis is called *Alternative Hypothesis* and this is written as

$$H_1 : \theta = \theta_1$$

Generally, the hypothesis wishing to be rejected by the test is taken as null hypothesis. Say we have two alternatives i.e., either $\theta = \theta_0$ or $\theta = \theta_1$ and we have a priori reason to be more inclined to believe the second hypothesis, then we take the hypothesis $H_0 : \theta = \theta_0$ as null hypothesis.

15.7 Critical Region

Any sample $x_1, x_2, \ldots, x_n$ of size n may be considered to be a point in n-dimensional space and it will be called a *sample point.* All such sample points corresponding to various random samples of size n constitute a sample space S, Fig. 15.7.1, so every sample is a point in S.

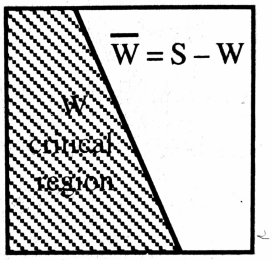


Figure 15.7.1

Let us divide the sample space S into two disjoint parts W and $\overline{W}(= S - W)$. Let us assume that we reject the null hypothesis $H_0 : \theta = \theta_0$ if the observed sample point falls in W and in this case we accept $H_1 : \theta = \theta_1$. On the other hand we accept H_0 if the point fall in $\overline{W}$. Technically the region W, i.e., the region of rejection of the null hypothesis H_0 is called the critical region or region of rejection.

15.8 Two Types of Errors

The decision whether the null hypothesis to be reject or accepted is taken on the basis of the information supplied by the observed sample observations. The conclusion drawn on the basis of a particular sample may

not be always true in respect of the population. The following two cases are called Type I and Type II errors.

Type I Error: When the null hypothesis H_0 is rejected i.e., H_1 is accepted but H_0 is true, the error arising in this situation is called Type I Error. If α be the probability of Type I Error, then

$$\begin{aligned} \alpha &= \text{Probability of Type I Error} \\ &= \text{Probability of rejecting } H_0 \text{ where } H_0 \text{ is true} \\ &= P(x \in W/H_0 = \theta_0), \text{ where } x = (x_1, x_2, \ldots, x_n) \end{aligned} \tag{15.8.1}$$

Type II Error: When the null hypothesis H_0 is accepted i.e., H_1 is rejected but H_0 is false, the error arising in this situation will be called Type II error. If β be the probability of Type II Error, then

$$\begin{aligned} \beta &= \text{Probability of Type II Error} \\ &= \text{Probability of accepting } H_0 \text{ where } H_0 \text{ is false} \\ &= P(x \in \overline{W}/H_1 = \theta_1), \text{ where } x = (x_1, x_2, \ldots, x_n) \end{aligned} \tag{15.8.2}$$

15.9 Level of Significance

The probability of Type I Error, α, is called *level of significance* of the test. It is also called the *size of the critical region.*

15.10 Power of the test

If β be the probability of Type II Error, then $1 - \beta$ is defined as the *power function* of the test hypothesis. The graph obtained by plotting power on the y-axis against various values of the parameter θ on the x-axis on a graph paper is called a *power curve.* The value of the power function at a parameter point is called the *power of the test* at that point.

Example 15.10.1. A random sample of size 10 is taken from a normal population and the following values were calculated for the variable (x) under study:

$$\sum_{i=1}^{10} x_i = 620, \quad \sum_{i=1}^{10} x_i^2 = 39016.$$

Test the null hypothesis $H_0 : \sigma = 8$ against $H_1 : \sigma > 8$ on the basis of the above data. Use $\alpha = 0.05$ as level of significance. (Given $\chi^2_{0.05}$ for 9 degrees of freedom = 16.92)

Solution: Here

$$n = 10, \quad \sum_{i=1}^{10} x_i = 620, \quad \sum_{i=1}^{10} x_i^2 = 39016.$$

Then

$$\begin{aligned} S^2 &= \frac{\sum_{i=1}^{10} x_i^2}{n} - \left(\frac{\sum_{i=1}^{10} x_i}{n} \right)^2 \\ &= \frac{39016}{10} - \left(\frac{620}{10} \right) \\ &= 3901.6 - 3844 \\ &= 57.6 \end{aligned}$$

We make the null hypothesis $H_0 : \sigma = 8$ against $H_1 : \sigma > 8$. Again

$$\begin{aligned}\chi^2 &= \frac{nS^2}{\sigma^2} \\ &= \frac{10 \times 57.6}{(8)^2} \\ &= 9\end{aligned}$$

Since $\chi^2_{observed} = 9 < \chi^2_{0.05,9} = 16.92$, so, H_0 is accepted and thus we conclude that the value of σ may be taken as 8 at 95% level of significance.

Example 15.10.2. For a large lot of freshly minted coins a random sample of size 50 is taken. The mean weight of coins in the sample is found to be 28.57 gm. Assuming that the population standard deviation of weight is 1.25 gm., will it it be reasonable to suppose that the population mean is 28 gm ?

Solution: The size of the sample is 50 and so $n = 50$. Population mean $m = 28$ gm and population s.d. $\sigma = 1.25$ gm. Let the null hypothesis H_0 and the alternative hypothesis H_1 be given by

$$\begin{aligned}&H_0 : m = 28 \\ &H_1 : m \neq 28 \\ &\text{S.E. of } \overline{x} = \frac{\sigma}{\sqrt{n}} = \frac{1.25}{\sqrt{50}} = \frac{1.25}{7.071} = 0.18\end{aligned}$$

Let us consider the statistic

$$z = \frac{\overline{x} - m}{\text{S.E. of } (\overline{x})}$$

which is standard normal. Therefore,

$$z = \frac{28.57 - 28}{0.18} = 3.17$$

Since the observed value of z exceeds 1.64, thus z falls in the critical region at 5% level of significance and so the null hypothesis H_0 is rejected at 5% level of significance. So it will not be reasonable to suppose that the population mean is 28 gm. at 5% level of significance.

Example 15.10.3. The mean life time of a sample of 100 electric bulbs produced by a manufacturing company is estimated to be 1570 hours with a standard deviation of 120 hours. If μ be the mean life time of all the bulbs produced by the company, test the hypothesis $\mu = 1600$ hours against the alternative hypothesis $\mu \neq 1600$ hours, using level of significance 0.05.

Solution: Here $n =$ size of the sample = 100, population mean $\mu = 1570$ and population S.D. $\sigma = 120$. We test the null hypothesis $H_0 = \mu = 1600$ against the alternative hypothesis $H_1 : \mu \neq 1600$ at 5% level of significance.

$$\text{Here } \overline{x} = 1570 \text{ and S.E. of } \overline{x} = \frac{\sigma}{\sqrt{n}} = \frac{120}{\sqrt{100}} = 12$$

Therefore,

$$\begin{aligned}z &= \frac{\overline{x} - \mu}{\text{S.E. of } (\overline{x})} \\ &= \frac{1570 - 1600}{12} \\ &= -2.5\end{aligned}$$

Hence z falls in the critical region at 5% level of significance and so we reject the null hypothesis.

Thus at 5% level of significance it will not be reasonable to suppose that the mean life of the bulb will be 1600 hours.

Example 15.10.4. In a sample of 600 students of a certain college, 400 are found to use dot pens. In another college from a sample of 900 students 450 were found to use dot pens. Test whether the colleges are significantly different with respect to the habit of using dot pens. (Null and alternative hypothesis should be stated clearly.)

Solution: With usual notations, null hypothesis will be that the the population proportions of the two colleges regarding the habit of using dot pen are equal. So $H_0 : (P_1 = P_2)$ and alternative hypothesis is $H_1 : (P_1 \neq P_2)$.

Here,

$$n_1 = 600, \quad p_1 = \frac{400}{600} = 0.667$$
$$n_2 = 900, \quad p_2 = \frac{450}{900} = 0.5$$

If for the null hypothesis $P_1 = P_2 = P$, then sample estimate of P is

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{600 \times 0.667 + 900 \times 0.5}{600 + 900} = 0.567$$

Now,

$$\begin{aligned} \text{S.E. of } (p_1 - p_2) &= \sqrt{pq\left(\frac{1}{p_1} + \frac{1}{p_2}\right)} \\ &= \sqrt{0.567 \times (1 - 0.567) \times \left(\frac{1}{600} + \frac{1}{900}\right)} \\ &= \sqrt{0.567 \times 0.433 \times (0.0017 + 0.0011)} \\ &= 0.026 \end{aligned}$$

Now,

$$z = \frac{p_1 - p_2}{\text{S.E.}} = \frac{0.667 - 0.5}{0.026} = 6.42$$

At 1% level the critical region is $|z| > 2.58$. So this z falls in the critical region and hence H_0 is rejected. So the two colleges are significantly different with respect to the habit of using dot pens.

Unit 16

Course Structure

- Analysis of variance
- One factor experiments
- Linear mathematical model for ANOVA

16.1 Introduction

Suppose that in an agricultural experiment, four different chemical treatments of soil produced mean wheat yields of 28, 22, 18 and 24 bushels per acre, respectively. Is there a significant difference in these means, or is the observed spread simply due to chance?

Such problem can be solved by using an important technique known as the *analysis of variance,* developed by Fisher. It makes use of the F distribution already considered in previous unit. Basically, in many situations there is a need to test the significance of differences among three or more sample means, or equivalently to test the null hypothesis that the sample means are all equal.

16.2 One-Way Classification or One-Factor Experiments

In a *one-factor experiment* measurements or observations are obtained for a independent groups of samples, where the number of measurements in each group is b. We speak of *a treatments,* each of which has *b repetitions* or *replications.* In the above example, $a = 4$.

The results of a one-factor experiment can be presented in a table having a rows and b columns (Table. 16.1). Here x_{jk} denotes the measurement in the j-th row and k-th column, where $j = 1, 2, \ldots, a$ and $k = 1, 2, \ldots, b$. For example, x_{35} refers to the fifth measurement for the third treatment.

Table 16.1

Treatment 1	x_{11}	x_{12}	$\cdots$	x_{1b}	$\overline{x}_1$
Treatment 2	x_{21}	x_{22}	$\cdots$	x_{2b}	$\overline{x}_2$
$\vdots$			$\vdots$		
Treatment a	x_{a1}	x_{a2}	$\cdots$	x_{ab}	$\overline{x}_a$

We shall denote by $\overline{x}_{j\cdot}$ the mean of the measurements in the j-th row. We have

$$\overline{x}_{j\cdot} = \frac{1}{b}\sum_{k=1}^{b} x_{jk}, \quad j = 1, 2, \ldots, a \tag{16.2.1}$$

The dot in $\overline{x}_{j\cdot}$ is used to show that the index k has been summed out. The values $\overline{x}_{j\cdot}$ are called group means or treatment means or row means. The grand mean or overall mean is the mean of all the measurement in all the groups and is denoted by $\overline{x}$, i.e.,

$$\overline{x} = \frac{1}{ab}\sum_{j,k} x_{jk} = \frac{1}{ab}\sum_{j=1}^{a}\sum_{k=1}^{b} x_{jk}. \tag{16.2.2}$$

16.3 Total Variation, Variation Within Treatments, Variation Between Treatments

We define the total variation, denoted by v, as the sum of the squares of the deviations of each measurement from the grand mean $\overline{x}$, i.e.,

$$\text{Total variation } = v = \sum_{j,k}(x_{jk} - \overline{x}^2). \tag{16.3.1}$$

By writing the identity,

$$x_{jk} - \overline{x} = (x_{jk} - \overline{x}_j) + (\overline{x}_j - \overline{x}) \tag{16.3.2}$$

and then squaring and summing over j and k, we can show that

$$\sum_{j,k}(x_{jk} - \overline{x})^2 = \sum_{j,k}(x_{jk} - \overline{x}_j)^2 + \sum_{j,k}(\overline{x}_j - \overline{x})^2 \tag{16.3.3}$$

$$\Rightarrow \sum_{j,k}(x_{jk} - \overline{x})^2 = \sum_{j,k}(x_{jk} - \overline{x}_j)^2 + b\sum_{j}(\overline{x}_j - \overline{x})^2 \tag{16.3.4}$$

We call the first summation on the right side of (16.3.4) the variation within the treatments (since it involves the squares of the deviations of x_{jk} from the treatment means $\overline{x}_j$) and denoted it by v_w. Therefore,

$$v_w = \sum_{j,k}(x_{jk} - \overline{x}_j)^2 \tag{16.3.5}$$

The second summation on the right side of (16.3.4) is called the variation between treatments (since it involves the squares of the deviation of the various treatment means $\overline{x}_j$ from the grand mean $\overline{x}$ and is denoted by v_b). Therefore,

$$v_b = \sum_{j,k}(\overline{x}_j - \overline{x})^2 = b\sum_{j}(\overline{x}_j - \overline{x})^2 \tag{16.3.6}$$

Equation (16.3.4) can thus be written as

$$v = v_w + v_b. \tag{16.3.7}$$

16.4 Shortcut Methods for Obtaining Variations

To minimize the labour in computing the above variations, the following forms are convenient:

$$v = \sum_{j,k} x_{jk}^2 - \frac{\tau^2}{ab} \tag{16.4.1}$$

$$v_b = \frac{1}{b}\sum_{j} \tau_j^2 - \frac{\tau^2}{ab} \tag{16.4.2}$$

$$v_w = v - v_b \tag{16.4.3}$$

where τ is the total of all values x_{jk} and $\tau_{j\cdot}$ is the total of all values in the j-th treatment, i.e.,

$$\tau = \sum_{j,k} x_{jk} \quad \tau_{j\cdot} = \sum_{k} x_{jk} \tag{16.4.4}$$

In practice it is convenient to subtract some fixed value from all the data in the table; this has no effect on the final results.

16.5 Linear Mathematical Model for Analysis of Variance

We can consider each row of Table 16.1 as a random sample of size b from the population fro that particular treatment. Therefore, for treatment j we have the independent, identically distributed random variables $X_{j1}, X_{j2}, \ldots, X_{jb}$, which respectively, take on the values $x_{j1}, x_{j2}, \ldots, x_{jb}$. Each of the $X_{jk}(k = 1, 2, \ldots, b)$ can be expressed as the sum of its expected value and a "chance" or "error" term:

$$X_{jk} = \mu_j + \Delta_{jk} \tag{16.5.1}$$

The Δ_{jk} can be taken as independent (relative to j as well as to k), normally distributed random variables with mean zero and variance σ^2. This is equivalent to assuming the the $X_{jk}(j = 1, 2, \ldots, a;\ k = 1, 2, \ldots, b)$ are mutually independent, normal variables with means μ_j and common variance σ^2. Let us define the constant μ by

$$\mu = \frac{1}{a}\sum_{j} \mu_j$$

We can think of μ as the mean for a sort of grand population comprising all the treatment populations. Then (16.5.1) can be rewritten as

$$X_{jk} = \mu + \alpha_j + \Delta_{jk} \quad \text{where} \quad \sum_{j} \alpha_j = 0 \tag{16.5.2}$$

The constant α_j can be viewed as the special effect of the j-th treatment.
The null hypothesis that all treatment means are equal is given by $(H_0 : \alpha_j = 0; j = 1, 2, \ldots, a)$ or equivalently by $(H_0 = \mu_j = \mu;\ j = 1, 2, \ldots, a)$. If H_0 is true, the treatment populations, which by assumption are normal, have a common mean as well as a common variance. Then there is just one treatment population, and all treatments are statistically identical.

16.6 Expected Values of the Variations

The between-treatments variation V_b, the within-treatments variation V_w, and the total variation V are random variables that, respectively, assume the values v_b, v_w, and v as defined in (16.3.6), (16.3.5) and (16.3.1), we can show that

$$E(V_b) = (a-1)\sigma^2 + b\sum_j \alpha_j^2 \tag{16.6.1}$$

$$E(V_w) = a(b-1)\sigma^2 \tag{16.6.2}$$

$$E(V) = (ab-1)\sigma^2 + b\sum_j \alpha_j^2 \tag{16.6.3}$$

From (16.6.2) it follows that

$$E\left[\frac{V_w}{a(b-1)}\right] = \sigma^2 \tag{16.6.4}$$

so that

$$\hat{S}_w^2 = \frac{V_w}{a(b-1)} \tag{16.6.5}$$

is always a best (unbiased estimate of σ^2 regardless of whether H_0 is true or not. On the other hand, from (16.6.1) and (16.6.3), we see that only if H_0 is true will we have have

$$E\left[\frac{V_b}{a-1}\right] = \sigma^2 \qquad E\left[\frac{V}{ab-1}\right] = \sigma^2 \tag{16.6.6}$$

so that only in such case will

$$\hat{S}_b^2 = \frac{V_b}{a-1} \qquad \hat{S}^2 = \frac{V}{ab-1} \tag{16.6.7}$$

provide unbiased estimates of σ^2. If H_0 is not true, however, then we have from (16.6.1)

$$E[\hat{S}_b^2] = \sigma^2 + \frac{b}{a-1}\sum_j \alpha_j^2 \tag{16.6.8}$$

16.7 Distributions of the Variations

Theorem 16.7.1. $\dfrac{V_w}{\sigma^2}$ is chi-square distributed with $a(b-1)$ degrees of freedom.

Theorem 16.7.2. Under the null hypothesis H_0, $\dfrac{V_b}{\sigma^2}$ and $\dfrac{V}{\sigma^2}$ are chi-square distributed with $a-1$ and $ab-1$ degrees of freedom, respectively.

16.8 The F Test for the Null Hypothesis of Equal Means

If the null hypothesis H_0 is not true, i.e., if the treatment means are not equal, we see from (16.6.8) that we can expect $\hat{S}_b^2$ to be greater than σ^2, with the effect becoming more pronounced as the discrepancy between means increases. On the other hand, from (16.6.4) and (16.6.5) we can expect $\hat{S}_w^2$ to be equal to σ^2 regardless of whether the means are equal or not. It follows that a good statistic for testing the hypothesis H_0 is provided by $\dfrac{\hat{S}_b^2}{\hat{S}_w^2}$. If this is significantly large, we can conclude that there is a significant difference between treatment means and thus reject H_0. Otherwise, we can either accept H_0 or reserve judgement pending further analysis.

Theorem 16.8.1. The statistic $F = \dfrac{\hat{S}_b^2}{\hat{S}_w^2}$ has the F distribution with $a-1$ and $a(b-1)$ degrees of freedom.

16.9 Analysis of Variance Tables

The calculations required for the above test are summarized in Table 16.2, which is called an *analysis of variance table*. In practice we would compute v and v_b using either the long method, (16.3.1) and (16.3.6), or the short method, (16.4.1) and (16.4.2), and then compute $v_w = v - v_b$. It should be noted that the degrees of freedom for the total variation, i.e., $ab - 1$, is equal to the sum of the degrees of freedom for the between-treatment and within-treatments variations.

Table 16.2

Variation	Degrees of Freedom	Mean Square	F
Between Treatments, $v_b = b\sum_j (\bar{x}_{j\cdot} - \bar{x})^2$	$a - 1$	$\hat{s}_b^2 = \dfrac{v_b}{a-1}$	$\dfrac{\hat{s}_b^2}{\hat{s}_w^2}$ with $a-1$, $a(b-1)$ degrees of freedom
Within Treatments, $v_w = v - v_b$	$a(b-1)$	$\hat{s}_w^2 = \dfrac{v_w}{a(b-1)}$	
Total, $v = v_b + v_w = \sum_{j,k}(x_{jk} - \bar{x})^2$	$ab - 1$		

Example 16.9.1. Table 16.3 shows the yields in bushels per acre of a certain variety of wheat grown in a particular type of soil treated with chemicals A, B, or C.

Table 16.3

A	48	49	50	49
B	47	49	48	48
C	49	51	50	50

Find (a) the mean yields for the different treatments, (b) the grand mean for all treatments, (c) the total variation, (d) the variation between treatments, (e) the variation within treatments. Use the long method.

Solution: To simplify the arithmetic, we may subtract some suitable number, say, 45, from all the data without affecting the values of the variations. We then obtain the data of Table 16.4

Table 16.4

3	4	5	4
2	4	3	3
4	6	5	5

(a) The treatment (row) means for Table 16.4 are given, respectively, by

$$\begin{aligned}\overline{x}_1 &= \frac{1}{4}(3+4+5+4) = 4,\\ \overline{x}_2 &= \frac{1}{4}(2+4+3+3) = 3,\\ \overline{x}_3 &= \frac{1}{4}(4+6+5+5) = 5,\end{aligned}$$

Therefore, the mean yields, obtained by adding 45 to these, are 49, 48 and 50 bushels per acre for A, B and C respectively.

$$\text{(b)} \quad \overline{x} = \frac{1}{12}(3+4+5+4+2+4+3+3+4+6+5+5) = 4$$

Therefore, the grand mean for the original set of data is $45 + 4 = 46$ bushels per acre.

(c)

$$\begin{aligned}\text{Total variation} = v &= \sum_{j,k}(x_{jk} - \overline{x})^2\\ &= (3-4)^2 + (4-4)^2 + (5-4)^2 + (4-4)^2\\ &\quad +(2-4)^2 + (4-4)^2 + (3-4)^2 + (3-4)^2\\ &\quad +(4-4)^2 + (6-4)^2 + (5-4)^2 + (5-4)^2\\ &= 14\end{aligned}$$

(d)

$$\begin{aligned}\text{Variation between treatments} = v_b &= b\sum_j(\overline{x}_j - \overline{x})^2\\ &= 4[(4-4)^2 + (3-4)^2 + (5-4)^2] = 8\end{aligned}$$

(e)

$$\text{Variation within treatments} = v_w = v - v_b = 14 - 8 = 6$$

Example 16.9.2. Referring to Example 16.9.1, find an unbiased estimate of the population variance σ^2 from (a) the variation between treatments under the null hypothesis of equal treatment means, (b) the variation within treatments.

Solution:

(a)

$$\hat{s}_b^2 = \frac{v_b}{a-1} = \frac{8}{3-1} = 4$$

(b)

$$\hat{s}_w^2 = \frac{v_w}{a(b-1)} = \frac{6}{3(4-1)} = \frac{2}{3}$$

Example 16.9.3. Referring to Example 16.9.1, can we reject the null hypothesis of equal means at (a) the 0.05 significance level? (b) the 0.01 significance level? (Given that $F_{0.95,2,9} = 4.26$ and $F_{0.99,2,9} = 8.02$).

Solution: We have

$$F = \frac{\hat{s}_b^2}{\hat{s}_w^2} = \frac{4}{2/3} = 6$$

with $a - 1 = 3 - 1 = 2$ and $a(b-1) = 3(4-1) = 9$ degrees of freedom.

(a) Since $F = 6 > F_{0.95,2,9} = 4.26$, we can reject the null hypothesis of equal means at the 0.05 level.

(b) Since $F = 6 > F_{0.99,2,9} = 8.02$, we cannot reject the null hypothesis of equal means at the 0.01 level.

The analysis of variance table for Examples 16.9.1 - 16.9.3 is shown in Table 16.5.

Table 16.5

Variation	Degrees of Freedom	Mean Square	F
Between Treatments, $v_b = 8$	$a - 1 = 2$	$\hat{s}_b^2 = \frac{8}{2} = 4$	$F = \frac{\hat{s}_b^2}{\hat{s}_w^2} = \frac{4}{2/3} = 6$ with 2, 9 degrees of freedom
Within Treatments, $v_w = v - v_b = 14 - 8 = 6$	$a(b-1) = (3)(3) = 9$	$\hat{s}_w^2 = \frac{6}{9} = \frac{2}{3}$	
Total, $v = 14$	$ab - 1 = (3)(4) - 1 = 11$		

Exercise 16.9.4. Use the shortcut formulas (16.4.1) through (16.4.3) to obtain the results of Example 16.9.1.

16.10 Modifications for Unequal Number of Observations

In case the treatments $1, \ldots, a$ have different numbers of observations equal to $n_1, \ldots, n_a$, respectively, the above results are easily modified. We therefore obtain

$$v = \sum_{j,k}(x_{jk} - \overline{x})^2 = \sum_{jk} x_{jk}^2 - \frac{\tau^2}{n} \tag{16.10.1}$$

$$v_b = \sum_{j,k}(\overline{x}_{j\cdot} - \overline{x})^2 = \sum_{j} n_j(\overline{x}_{j\cdot} - \overline{x})^2 = \sum_{j} \frac{\tau_{j\cdot}^2}{n_j} - \frac{\tau^2}{n} \tag{16.10.2}$$

$$v_w = v - v_b \tag{16.10.3}$$

where $\sum_{j,k}$ denotes the summation over k from 1 to n_j and then over j from 1 to a, $n = \sum_j n_j$ is the total number of observations in all treatments, τ is the sum of all observations, $\tau j\cdot$ is the sum of all values in the j-th treatment, and $\sum_j$ is the sum from $j = 1$ to a. The analysis of variance table for this case is given in Table 16.6.

Example 16.10.1. Table 16.7 shows the lifetimes in hours of samples from three different types of television tubes manufactured by a company. Using the long method, test at (a) the 0.05, (b) the 0.01 significance level whether there is a difference in the three types. (Given that $F_{0.95,2,9} = 4.26$ and $F_{0.99,2,9} = 8.02$).

Solution. It is convenient to subtract a suitable number, say, 400, obtaining Table 16.8. In this table we have indicated the row total, the sample or group means, and the grand mean. We then have

$$v = \sum_{j,k}(x_{jk} - \overline{x})^2 = (7-7)^2 + (11-7)^2 + \cdots + (8-7)^2 = 72$$

$$v_b = \sum_{j,k}(\overline{x}_{j\cdot} - \overline{x})^2 = \sum_{j} n_j(\overline{x}_{j\cdot} - \overline{x})^2 = 3(9-7)^2 + 5(7-5)^2 + 4(8-7)^2 = 36$$

$$v_w = v - v_b = 72 - 36 = 36$$

The data can be summarized in the analysis of variance table, Table 16.9. Now, for 2 and 9 degrees of freedom

Table 16.6

Variation	Degrees of Freedom	Mean Square	F
Between Treatments, $v_b = \sum_j n_j(\bar{x}_{j.} - \bar{x})^2$	$a - 1$	$\hat{s}_b^2 = \dfrac{v_b}{a-1}$	$\dfrac{\hat{s}_b^2}{\hat{s}_w^2}$ with $a-1, n-a$ degrees of freedom
Within Treatments, $v_w = v - v_b$	$n - a$	$\hat{s}_w^2 = \dfrac{v_w}{n-a}$	
Total, $v = v_b + v_w = \sum_{j,k}(x_{jk} - \bar{x})^2$	$n - 1$		

Table 16.7

Sample 1	407	411	409		
Sample 2	404	406	408	405	402
Sample 3	410	408	406	408	

we have $F_{0.95,2,9} = 4.26$ and $F_{0.99,2,9} = 8.02$. Therefore, we can reject the hypothesis of equal means (i.e., there is no difference in the tree types of tubes) at the 0.05 level but not at the 0.01 level.

Table 16.8

						Total	Mean
Sample 1	7	11	9			27	9
Sample 2	4	6	8	5	2	25	5
Sample 3	10	8	6	8		32	8
	$\bar{x}$ = grand mean = $\frac{84}{12} = 7$						

Table 16.9

Variation	Degrees of Freedom	Mean Square	F
$v_b = 36$	$a - 1 = 2$	$\hat{s}_b^2 = \frac{36}{2} = 18$	$\frac{\hat{s}_b^2}{\hat{s}_w^2} = \frac{18}{4}$
$v_w = 36$	$n - a = 9$	$\hat{s}_w^2 = \frac{36}{9} = 4$	$= 4.5$

Exercise 16.10.2. Use the shortcut formulas (16.10.1) through (16.10.3) to obtain the results of Example 16.10.1.

References

1. J. P Tremblay and R. Manohar : Discrete Mathematical Structures with Applications to Computers.
2. J. L. Gersting : Mathematical Structures for Computer Sciences.
3. S. Lepschutz : Finite Mathematics.
4. F. Harary : Graph Theory.
5. S. Sahani : Concept of Discrete Mathematics.
6. J. E. Whitesitt : Boolean Algebra and its Applications.
7. K. L. P. Mishra and N. Chandrasekaran : Automata, Languages, and Computation.
8. G. E. Revesz : Introduction to Formal Languages.
9. K . D. Joshi : Foundation of Discrete Mathematics.
10. Modern Probability Theory: B. R. Bhat.
11. Elementary Probability Theory and Stochastic Processes: K. L. Chung.
12. Linear Statistical Inference and its Applications: C. R. Rao.
13. Life-testing and Reliability Estimation: S. K. Sinha & B. K. Kale.
14. An Outline of Statistical Theory (Vol 1 and 2): A. M. Goon, M. K. Gupta & B. Dasgupta.

POST GRADUATE DEGREE PROGRAMME (CBCS) IN

MATHEMATICS

SEMESTER IV

SELF LEARNING MATERIAL

PAPER : MATO 4.2
(Applied and Pure Streams)

Advanced Operations Research II

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

Course Preparation Team

1. Dr. Sahidul Islam
Assistant Professor
Department of Mathematics
University of Kalyani

2. Mr. Biswajit Mallick
Assistant Professor (Cont.)
DODL, University of Kalyani

3. Ms. Audrija Choudhury
Assistant Professor (Cont.)
DODL, University of Kalyani

May, 2020

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, 2017 had been our endeavour. We are happy to have achieved our goal.

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

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

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

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

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

Director

Directorate of Open and Distance Learning

University of Kalyani

Optional Paper

MATO 4.2
Marks : 100 (SEE : 80; IA : 20)

Advanced Operations Research II
(Applied and Pure Streams)

Syllabus

- **Unit 1:** Reliability: Elements of Reliability theory, failure rate, extreme value distribution.
- **Unit 2:** Analysis of stochastically falling equipments including the reliability function, reliability and growth model.
- **Unit 3:** Information Theory: Information concept, expected information, Entropy and properties of entropy function.
- **Unit 4:** Bivariate Information theory,
- **Unit 5:** Economic relations involving conditional probabilities,
- **Unit 6:** Coding theory: Communication system, encoding and decoding.
- **Unit 7:** Shannon-Fano encoding procedure, Haffman encoding, noiseless coding theory, noisy coding.
- **Unit 8:** Family of codes, Hammimg code, Golay code, BCH codes, Reed-Muller code, Perfect code, codes and design, Linear codes and their dual, weight distribution.
- **Unit 9:** Markovian Decision Process : Ergodic matrices, regular matrices.
- **Unit 10:** Imbedded Markov Chain method for Steady State solution.
- **Unit 11:** Posynomial, Signomial, Degree of difficulty, Unconstrained minimization problems, Solution using Differential Calculus, Solution seeking Arithmetic-Geometric inequality, Primal dual relationship & sufficiency conditions in the unconstrained case,
- **Unit 12:** Constrained minimization, Solution of a constrained Geometric Programming problem, Geometric programming with mixed inequality constrains, Complementary Geometric programming.
- **Unit 13:** A brief introduction to Inventory Control, Single-item deterministic models without shortages.
- **Unit 14:** Single-item deterministic models with shortages Dynamic Demand Inventory Models.
- **Unit 15:** Multi-item inventory models with the limitations on warehouse capacity
- **Unit 16:** Models with price breaks, single-item stochastic models without Set-up cost and with Set-up cost, Average inventory capacity, Capital investment.

Contents

Unit 1

Course Structure

- Reliability Theory
- MTTF in terms of failure density

1.1 Introduction

Reliability is the probability of a device performing its purpose adequately for the period of time intended under the operating conditions encountered. The definition brings into focus, four important factors, namely,

- the reliability of a device is expressed as a probability;
- the device is required to give adequate performance;
- the duration of adequate performance is specified;
- the environment or operating conditions are prescribed.

Some of the important aspects of reliability are:

a) Reliability is a function of time. We could not expect an almost wornout light bulb to be as reliable as one recently put into service.

b) Reliability is a function of conditions to use. In very severe environments, we expect to encounter frequent system breakdowns than in normal environments.

c) Reliability is expected as a probability which helps us to quantify it and think of optimizing system reliability.

1.2 Reliability

Definition 1.2.1. Hazard Rate/Failure Rate: Failure rate is the ratio of the number of failures during a particular unit interval to the average population during that interval. Thus the failure rate for the ith interval is

$$\frac{n_i}{\frac{1}{2}\left[\left(N-\sum_{k=1}^{i-1} n_k\right)+\left(N-\sum_{k=1}^{i} n_k\right)\right]},$$

where n_i is the number of failures during the ith interval and N is the total number of components.

Definition 1.2.2. Failure Density: The failure density in a particular unit interval is the ratio of the number of failures during that interval to the number of components. So the failure density during the ith interval is

$$\frac{n_i}{N} = f_{d_i}.$$

Let l be the last interval after which there are no intervals. Then

$$f_{d_l} = \frac{n_l}{N}.$$

Thus,

$$f_{d_1} + f_{d_2} + \cdots + f_{d_l} = \frac{1}{N}(n_1 + n_2 + \cdots + n_l) = \frac{N}{N} = 1.$$

Hence the sum of values entered in column 5 is 1 (Table 1.1).

Definition 1.2.3. Reliability: Reliability(R), is the ratio of the number of survivals at any given time to the total initial population. That is, reliability at ith time is

$$R(i) = \frac{s_i}{N},$$

s_i is the number of survivals during the ith interval.

Definition 1.2.4. Probability of failure: The concept of probability of failure is similar to that of the concept of probability of survival. This is the ratio of the number of units failed within a certain time to the total population.

Hence, the probability of failure within ith time is

$$\frac{n_1 + n_2 + \cdots + n_i}{N} \quad \text{or} \quad \frac{F_i}{N},$$

so that the probability of failure at ith time plus reliability at ith time is

$$\frac{F_i}{N} + \frac{s_i}{N} = 1$$

(since $F_i + s_i = N$), that is, probability of failure and reliability at the same time is always 1.

Definition 1.2.5. Mean Failure Rate(h)**:** If Z_1 is the failure rate for the first unit of time, Z_2 is the failure rate for the second unit of time, ..., Z_T is the failure rate for the Tth unit of time, then the mean failure rate for T times will be

$$h(T) = \frac{Z_1 + Z_2 + \cdots + Z_T}{T}.$$

The mean failure rate is also obtained from the formula

$$\frac{1}{T}\left[\frac{N(0) - N(T)}{N(0)}\right],$$

where $N(0)$ is the total population at $t = 0$ and $N(T)$ is the total population remaining at time $t = T$.

Time(t)	Number of failures(n)	Cumulative failures(F)	Number of Survivals(S)	Failure density(f_d)	Failure/ Hazard rate(Z)	Reliability
0	0	0	1000	0	0	1
1	130	130	870	$\frac{130}{1000} = 0.130$	$\frac{130}{\frac{1000+870}{2}} = 0.139$	$1 - 0.130 = 0.870$
2	83	213	787	0.083	$\frac{83}{\frac{870+787}{2}} = 0.100$	$1 - (0.130 + 0.083) = 0.787$
3	75	288	712	0.075	$\frac{75}{\frac{787+712}{2}} = 0.100$	0.712
4	68	356	644	0.068	$\frac{68}{\frac{712+644}{2}} = 0.100$	0.644
5	62	418	582	0.062	$\frac{62}{\frac{644+582}{2}} = 0.101$	0.582
6	56	474	526	0.056	$\frac{56}{\frac{582+526}{2}} = 0.101$	0.526
7	51	525	475	0.051	$\frac{51}{\frac{526+475}{2}} = 0.101$	0.475
8	46	571	429	0.046	$\frac{46}{\frac{475+429}{2}} = 0.102$	0.429
9	41	612	388	0.041	$\frac{41}{\frac{429+388}{2}} = 0.100$	0.388
10	37	659	341	0.037	$\frac{37}{\frac{388+341}{2}} = 0.101$	0.341
11	34	683	317	0.034	$\frac{34}{\frac{341+317}{2}} = 0.103$	0.317
12	31	714	286	0.031	$\frac{31}{\frac{317+286}{2}} = 0.102$	0.286
13	28	742	258	0.028	$\frac{28}{\frac{286+258}{2}} = 0.102$	0.258
14	64	806	194	0.064	$\frac{64}{\frac{258+194}{2}} = 0.283$	0.194
15	76	882	118	0.076	$\frac{76}{\frac{194+118}{2}} = 0.487$	0.118
16	62	944	56	0.062	$\frac{62}{\frac{118+56}{2}} = 0.713$	0.056
17	40	984	16	0.040	$\frac{40}{\frac{56+16}{2}} = 1.111$	0.016
18	12	996	4	0.012	$\frac{12}{\frac{16+4}{2}} = 1.2$	0.004
19	4	1000	0	0.004	$\frac{4}{\frac{4+0}{2}} = 2$	0.000

Table 1.1

Definition 1.2.6. Mean time to failure (MTTF): In general, if t_1 is the time to failure for the first specimen, t_2 is the time to failure for the second specimen, ..., t_N is the time to failure for the Nth specimen, then the MTTF for N specimens is

$$\frac{t_1 + t_2 + \cdots + t_N}{N}.$$

If n_1 is the number of specimens that failed during first unit of time, n_2 be that during second unit of time, ..., n_l be that during the last (lth) unit of time, then the MTTF for the N specimens will be

$$MTTF = \frac{n_1 + 2n_2 + \cdots + ln_l}{N},$$

where $N = n_1 + n_2 + \cdots + n_l$. If the time interval is δt unit instead of 1 unit, then

$$\begin{aligned} MTTF &= \frac{n_1 + 2n_2 + \cdots + ln_l}{N}\delta t \\ &= \frac{\sum_{k=1}^{l} kn_k}{N}\delta t. \end{aligned}$$

Example 1.2.7. In the life testing of 100 specimens of a particular device, the number of failures during each time interval of 20 hours is shown in the following table:

Time Interval (T)(in hours)	**Number of failures during the interval**
$T \leq 100$	0
$1000 < T \leq 1020$	25
$1020 < T \leq 1040$	40
$1040 < T \leq 1060$	20
$1060 < T \leq 1080$	10

Estimate the MTTF for these specimens.

Solution. As the number of specimens tested is large, it is tedious to record the time of failure for each specimen. So we note the number of specimen that fail during each 20 hours interval. Thus

$$MTTF = \frac{(0 \times 1000) + (25 \times 1020) + (40 \times 1040) + (20 \times 1060) + (10 \times 1080)}{100} = 1040 \text{ hrs.}$$

■

Example 1.2.8. The following table gives the results of tests conducted under severe adverse conditions on 1000 safety valves. Calculate the failure density $f_d(t)$ and the hazard rates $Z(t)$ where the time interval is 4 hours instead of 1 hour.

Time Interval (in hours)	**Number of failures (h)**
$t = 0$	0
$0 < t \leq 4$	267
$4 < t \leq 8$	59
$8 < t \leq 12$	36
$12 < t \leq 16$	24
$16 < t \leq 20$	23
$20 < t \leq 24$	11

Solution.

Time interval	Number of failures	Cumulative frequency	Number of Survivals(S)	Failure density(f_d)	Failure/ Hazard rate($Z(t)$)	Reliability (R)
$t = 0$	0	0	1000	0	0	1
$0 < t \leq 4$	267	267	733	0.067	$\frac{267}{\frac{4(1000+733)}{2}} = 0.077$	$1 - 0.067 = 0.933$
$4 < t \leq 8$	59	326	674	0.0148	$\frac{56}{\frac{4(733+674)}{2}} = 0.021$	$1 - (0.067 + 0.0148)$ $= 0.9182$
$8 < t \leq 12$	36	362	638	0.009	$\frac{36}{\frac{4(674+638)}{2}} = 0.014$	0.9092
$12 < t \leq 16$	24	386	614	0.006	$\frac{24}{\frac{4(638+614)}{2}} = 0.009$	0.9032
$16 < t \leq 20$	23	409	591	0.0057	$\frac{23}{\frac{4(614+591)}{2}} = 0.009$	0.8975
$20 < t \leq 24$	11	420	580	0.0027	$\frac{11}{\frac{4(591+580)}{2}} = 0.0047$	0.8948

■

Four Important Points

(i) Sum of the failure densities is 1, that is,

$$\begin{aligned} f_{d_1} + f_{d_2} + \cdots + f_{d_l} = \sum_{i=1}^{l} f_{d_i} &= 1 \text{ (For discrete case)} \\ \int_0^T f_d(\xi) d\xi &= 1, \end{aligned}$$

where the limits of the integration are taken from the beginning of the first at $t = 0$ till the end where all the specimens failed at time $t = T$.

(ii) The reliability $R(i)$ for the ith hour is given by

$$\begin{aligned} R(i) &= 1 - (f_{d_1} + f_{d_2} + \cdots + f_{d_i}) \\ &= 1 - \sum_{k=1}^{i} f_{d_k} \text{ [For discrete case]} \end{aligned}$$

Hence the reliability $R(t)$, for the tth hour for continuous case is given by

$$\begin{aligned} R(t) &= 1 - \int_0^t f_d(\xi) d\xi \\ &= \int_t^T f_d(\xi) d\xi \text{[For continuous case].} \end{aligned}$$

(iii) The probability of failure in hours, $F(i)$ is given by

$$F(i) = f_{d_1} + f_{d_2} + \cdots + f_{d_i} = \sum_{k=1}^{i} f_{d_k},$$

that is, $R(i) + F(i) = 1$.

Since the reliability and probability of failure are complementary so, $R(t) + F(t) = 1$. Thus for continuous case,

$$F(t) = 1 - R(t) = \int_0^t f_d(\xi) d\xi.$$

(iv) The failure rate or hazard rate for the ith hour is

$$Z(i) = \frac{n_i}{\frac{1}{2}\left(N - \sum_{k=1}^{i-1} n_k\right)} = \frac{2[R(i-1) - R(i)]}{R(i-1) + R(i)}.$$

(For one hour interval between $t = (i-1)$ hr to $t = i$ hr)

If the interval is δt, instead of 1 hour, then for continuous case,

$$\begin{aligned} Z(t) &= \frac{2[R(t - \delta t) - R(t)]}{[R(t - \delta t) + R(t)]\delta t} \\ \text{i.e, } Z(t + \delta t) &= \frac{2[R(t) - R(t + \delta t)]}{[R(t) + R(t + \delta t)]\delta t} \end{aligned}$$

For continuous case, when $\delta t \to 0$, we have

$$\begin{aligned}
\lim_{\delta t \to 0} Z(t+\delta t) &= \lim_{\delta t \to 0} \frac{2[R(t)-R(t+\delta t)]}{[R(t)+R(t+\delta t)]\delta t} \\
\Rightarrow Z(t) &= \lim_{\delta t \to 0} \frac{R(t)-R(t+\delta t)}{R(t)\delta t} \\
&= -\frac{1}{R(t)} \lim_{\delta t \to 0} \frac{R(t+\delta t)-R(t)}{\delta t} \\
&= -\frac{1}{R(t)} \frac{d}{dt}(R(t)) \\
&= -\frac{R'(t)}{R(t)} \qquad (1.2.1)
\end{aligned}$$

Thus,

$$\begin{aligned}
\int_0^t Z(t)dt &= -[\log R(t)]_0^t \\
\Rightarrow \log R(t) = \log R(0) - \int_0^t Z(t)dt &
\end{aligned}$$

Since at $t = 0$, $R(0) = 1$, that is, $\log R(0) = 0$, thus,

$$R(t) = \mathrm{e}^{-\int_0^t Z(\xi)d\xi} \qquad (1.2.2)$$

Finally we shall get an expression for $f_d(t)$ for continuous case.

By definition, we have

$$\begin{aligned}
f_d(t+\delta t) &= \frac{(\text{no. of survivals at time } t=t) - (\text{no. of survivals at time } t=t+\delta t)}{\delta t.(\text{total number of survivals})} \\
\text{or, } f_d(t+\delta t) &= \left[\left(\frac{\text{no. of survivals at } t=t}{\text{total no. of survivals}}\right) - \left(\frac{\text{no. of survivals at } t=t+\delta t}{\text{total no. of survivals}}\right)\right]\frac{1}{\delta t} \\
&= \frac{1}{\delta t}[R(t)-R(t+\delta t)]
\end{aligned}$$

Letting $\delta t \to 0$, we get for continuous case,

$$f_d(t) = -\lim_{\delta t \to 0} \frac{R(t+\delta t)-R(t)}{\delta t} = -R'(t) \qquad (1.2.3)$$

From equations (1.2.1) and (1.2.3), we get

$$\begin{aligned}
Z(t) &= \frac{f_d(t)}{R(t)} \\
\Rightarrow f_d(t) &= Z(t)R(t) \\
&= Z(t)\,\mathrm{e}^{-\int_0^t Z(\xi)d\xi} \qquad (1.2.4)
\end{aligned}$$

1.3 MTTF in terms of failure density

The mean time to failure is given by

$$\text{MTTF} = \frac{\left(\sum_{k=1}^{l} k n_k\right)\delta t}{N},$$

where N is the initial total survivals; n_1 is the total no. of specimens that failed during the first δt time interval, n_2 is the total no. of specimens that failed during the second δt time interval, ... , n_k is the total no. of specimens failed during the kth δt interval. Now, by definition

$$\begin{aligned} f_{d_k} &= \frac{n_k}{N.\delta t} \\ \Rightarrow \frac{n_k}{N} &= f_{d_k}\delta t. \end{aligned}$$

Further, $f\delta t$ is the elapsed time t. Hence the expression for MTTF can be written as

$$\text{MTTF} = \sum_{k=1}^{l}(k.f_{d_k}.\delta t)\delta t = \sum_{k=1}^{l} f_{d_k}(k\delta t)\delta t \tag{1.3.1}$$

where the summation is for the period from the first δt time interval to lth δt interval.

For continuous case, when $\delta t \to 0$, and $f\delta t$ is the elapsed time t and f_{d_k} will be the failure density $f_d(t)$ at time t, then

$$\text{MTTF} = \int_0^T t f_d(t)dt, \tag{1.3.2}$$

where T is the number of hours after which there are no survivals.

Now we have, $F(t) + R(t) = 1$. Thus,

$$F(t) = 1 - R(t) = \int_0^t f_d(\xi)d\xi$$

Thus,

$$\frac{d}{dt}(F(t)) = -\frac{d}{dt}(R(t)) = f_d(t).$$

Thus,

$$\begin{aligned} \text{MTTF} &= \int_0^\infty t f_d(t)dt \quad [\text{For } t > T, \text{ there are no survivals, so the values of the integration is } 0, \text{ for } t > T] \\ &= \int_0^\infty -t\frac{d}{dt}(R(t))dt \\ &= -[t.R(t)]_0^\infty + \int_0^\infty 1.R(t)dt \\ &= \int_0^\infty R(t)dt \quad [\text{Since } R(0) = 1 \text{ and } R(\infty) = 0 \text{ as } t \to \infty, \text{ there are no survivals}] \end{aligned} \tag{1.3.3}$$

Also, when $t_1 \le t \le t_2$, we have,

$$F(t_2) - F(t_1) = \int_{t_1}^{t_2} f_d(\xi)d\xi.$$

For continuous case, when the hazard rate is constant, that is, $Z(t) = \lambda$, a constant, say, then

$$\int_0^t Z(\xi)d\xi = \int_0^t \lambda d\xi = \lambda t.$$

Thus,

$$R(t) = \mathrm{e}^{-\int_0^t Z(\xi)d\xi} = \mathrm{e}^{-\lambda t}$$

and $F(t) = 1 - \mathrm{e}^{-\lambda t}$. Similarly,

$$f_d(t) = Z(t) \times R(t) = \lambda\,\mathrm{e}^{-\lambda t}\,.$$

Thus,

$$\begin{aligned}\text{MTTF} &= \int_0^\infty R(t)dt \\ &= \int_0^\infty \mathrm{e}^{-\lambda t}\,dt \\ &= -\left[\frac{\mathrm{e}^{-\lambda t}}{\lambda}\right]_0^\infty = \frac{1}{\lambda}.\end{aligned}$$

Thus, for a constant hazard model, the MTTF is simply the reciprocal of the hazard rate.

The constant hazard rate is also known as the exponential reliability rate.

$$\begin{aligned}\text{MTTF} = \int_0^\infty t f_d(t)dt &= \int_0^\infty t\,\mathrm{e}^{-\lambda t}\,dt \\ &= \lambda\left[\frac{t\,\mathrm{e}^{-\lambda t}}{-\lambda}\right]_0^\infty + \int_0^\infty \frac{\lambda}{\lambda}\,\mathrm{e}^{-\lambda t}\,dt \\ &= 0 + \int_0^\infty \mathrm{e}^{-\lambda t}\,dt = \frac{1}{\lambda}.\end{aligned}$$

The mean of $\lambda\,\mathrm{e}^{-\lambda t}$ is

$$\int_0^\infty \lambda t\,\mathrm{e}^{-\lambda t}\,dt = \frac{1}{\lambda}.$$

Example 1.3.1. It is found that the random variations with respect to time in the output voltage of a particular system are exponentially distributed with a mean value 100V. What is the probability that the output voltage will be found at any time to lie in the range $90 - 100$V?

Solution. For an exponential distribution, the MTTF is the reciprocal of the hazard rate λ(say), where λ is a constant, that is, MTTF$= \frac{1}{\lambda}$.

Here, we identify the MTTF with a mean value 100V. Thus,

$$\frac{1}{\lambda} = 100 \Rightarrow \lambda = 0.01.$$

Hence the p.d.f $f_d(t)$ for the voltage distribution is $= \lambda\,\mathrm{e}^{-\lambda t} = 0.01 \times \mathrm{e}^{-0.01t}$.

Now, the probability that the voltage lies between V_1 and V_2 is given by

$$\begin{aligned}F(V_2) - F(V_1) &= \int_{V_1}^{V_2} f_d(t)dt \\ &= \int_{V_1}^{V_2} \lambda\,\mathrm{e}^{-\lambda t}\,dt \\ &= 1 - \mathrm{e}^{-\lambda(V_2 - V_1)}.\end{aligned}$$

Here, $V_2 = 100$V, $V_1 = 90$V.

Hence, $F(100) - F(90) = 1 - \mathrm{e}^{-0.01(100-90)} = 1 - \mathrm{e}^{-0.1} \simeq 0.095$. ■

Example 1.3.2. It is observed that the failure pattern of an electronic system follows an exponential distribution with mean time to failure of 100 hours. What is the probability that the system failure occurs within 750 hours?

Solution. MTTF $= \frac{1}{\lambda} = 1000$, where λ is the constant hazard rate. Thus,

$$\lambda = \frac{1}{1000}.$$

Thus,

$$f_d(t) = \lambda \, \mathrm{e}^{-\lambda t}$$

Hence the probability that the system failure occurs within a period V is

$$F(V) = \int_0^V f_d(t)dt = \int_0^V \lambda \, \mathrm{e}^{-\lambda t} \, dt = 1 - \mathrm{e}^{-\lambda V} \, .$$

Here, $V = 750$ hrs. and $\lambda = 0.001$. Thus,

$$F(750) = 1 - \mathrm{e}^{-0.750} \simeq 0.528.$$

■

Unit 2

Course Structure

- Linearly Increasing Hazard
- System Reliability
- Redundancy

2.1 Linearly Increasing Hazard

Here the hazard increases linearly with time, that is, $Z(t) = kt$, where k is a constant. Thus the time integral of $Z(t)$ is given by

$$\int_0^t ktdt = \frac{k}{2}t^2$$

Therefore,

$$R(t) = \mathrm{e}^{-\int_0^t Z(\xi)d\xi} = \mathrm{e}^{-\frac{k}{2}\,t^2}$$

And thus,

$$f_d(t) = Z(t) \times R(t) = kt\,\mathrm{e}^{-\frac{k}{2}t^2}$$

This function $f_d(t) = kt\,\mathrm{e}^{-\frac{k}{2}t^2}$ is known as the **Rayleigh density function**.

Now,

$$\begin{aligned}\frac{d}{dt}(f_d(t)) &= k\,\mathrm{e}^{-\frac{k}{2}t^2} + kt\left(-\frac{k}{2}.2t\,\mathrm{e}^{-\frac{k}{2}t^2}\right)\\ &= k\,\mathrm{e}^{-\frac{k}{2}t^2}[1-kt^2]\end{aligned}$$

and

$$\begin{aligned}\frac{d^2}{dt^2}(f_d(t)) &= k\left(-\frac{k}{2}.2t\right)\mathrm{e}^{-\frac{k}{2}t^2}[1-kt^2] + k\,\mathrm{e}^{-\frac{k}{2}t^2}[-2kt]\\ &= -k^2t\,\mathrm{e}^{-\frac{k}{2}t^2}[1-kt^2] - 2k^2t\,\mathrm{e}^{-\frac{k}{2}t^2}\\ &= -3k^2t\,\mathrm{e}^{-\frac{k}{2}t^2} + k^3t^3\,\mathrm{e}^{-\frac{k}{2}t^2}\end{aligned}$$

Now,

$$\frac{d}{dt}(f_d(t)) = 0 \Rightarrow t = \frac{1}{\sqrt{k}} \text{ [since } t > 0]$$

At $t = \frac{1}{\sqrt{k}}$,

$$\begin{aligned}
\frac{d^2}{dt^2}(f_d(t)) &= \frac{-3k^2}{\sqrt{k}}\, e^{-\frac{k}{2}t^2} + \frac{k^3}{k\sqrt{k}}\, e^{-\frac{k}{2}t^2} \\
&= -\frac{2k^2}{\sqrt{k}}\, e^{-\frac{k}{2}t^2} \\
&= -2k\sqrt{k}\, e^{-\frac{k}{2}t^2} \\
&= = -2k\sqrt{k}\, e^{-\frac{k}{2}\frac{1}{k}} \\
&= -\frac{2k\sqrt{k}}{\sqrt{e}} < 0
\end{aligned}$$

Thus, $f_d(t)$ is maximum at $t = \frac{1}{\sqrt{k}}$. Thus

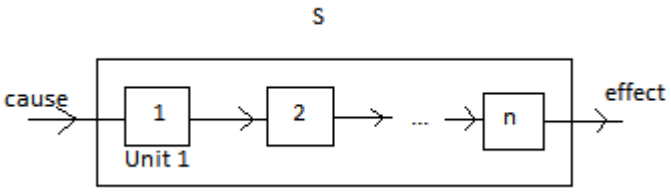
$$f_d(t)\big|_{t=\frac{1}{\sqrt{k}}} = k.\frac{1}{\sqrt{k}}\, e^{-1/2} = \sqrt{k}\, e^{-1/2} = \sqrt{\frac{k}{e}}.$$

Hence $f_d(t)$ reaches a maximum value $\sqrt{\frac{k}{e}}$ at $t = \frac{1}{\sqrt{k}}$ and tends to zero as t becomes larger. Now, we calculate the MTTF when the hazard rate increases linearly.

$$\begin{aligned}
\text{MTTF} = \int_0^\infty R(t)dt &= \int_0^\infty e^{-k/2t^2}\, dt \\
&= \sqrt{\frac{2}{k}} \int_0^\infty e^{-z^2}\, dz \qquad [\text{Put } \sqrt{\frac{k}{2}}t = z] \\
&= \sqrt{\frac{2}{k}}.\frac{\sqrt{\pi}}{2} \\
&= \sqrt{\frac{\pi}{2k}}.
\end{aligned}$$

2.2 System Reliability

A. **Series Configuration:** The simplest combination of units that form a system is a series combination. This is one of the most commonly used structures and is shown in the following figure: The system S



consists of n units which are connected in series as shown. Let the successful operation of these individual units be represented by $X_1, X_2, \ldots, X_n$ and their respective probabilities by $P(X_1), P(X_2), \ldots, P(X_n)$.

For the successful operation of the system, it is necessary that all n units function satisfactorily. Hence the probability of the successful operation of all the units is $P(X_1 \text{ and } X_2 \text{ and } \cdots \text{and } X_n)$.

We shall assume that these units are not independent of one another, that is, the successful operation of unit 1 might affect the successful operation of all other units and so on.

The system reliability is given by

$$\begin{aligned} P(S) &= P(X_1 \text{ and } X_2 \text{ and } \cdots \text{and } X_n) \\ &= P(X_1).P(X_2|X_1).P(X_3|X_1X_2)\ldots P(X_n|X_1X_2\ldots X_{n-1}). \end{aligned}$$

If they are independent, then

$$P(S) = P(X_1)P(X_2)\ldots P(X_n).$$

Example 2.2.1. In a hydraulic control system, the connecting linkage has a reliability factor 0.98 and the valve which has a reliability factor 0.92. Also the pressure sensor which activates the linkage, has a reliability factor 0.90. Assume that all the three elements namely the activator, the linkage and the hydraulic valve are connected in series with independent reliability factors. What is the reliability of the control system?

Solution. Let the successful operation of the elements namely the activator, the linkage and the hydraulic valve be denoted by X_1, X_2 and X_3 respectively. Thus,

$$P(X_1) = 0.98, \;\; P(X_2) = 0.91, \;\; P(X_3) = 0.90 \;\; \text{(given)}$$

Since these elements are connected in series with independent reliability factors, hence the reliability of the control system, S(say) is

$$P(S) = P(X_1)P(X_2)P(X_3) = 0.98 \times 0.92 \times 0.90 = 0.81144.$$

■

Note 2.2.2. There is an important point that the reliability of a series system is always worse than the poorest component of the system.

Example 2.2.3. If the system consists of n identical units in series and if each unit has a reliability factor p, determine the system reliability under the assumption that all units function independently.

Solution. $P(S) = p.p\ldots p(\text{n times}) = p^n$. Now, if q is the probability of failure of each unit, then $p = 1 - q$.

Hence the system reliability

$$P(S) = p^n = (1-q)^n = 1 - nq + \cdots$$

If q is very small, this expression can be approximated to $1 - nq$. Thus,

$$P(S) \simeq 1 - nq.$$

■

Example 2.2.4. A system has 10 identical equipments. It is desired that the system reliability be 0.95. Determine how good each component should be?

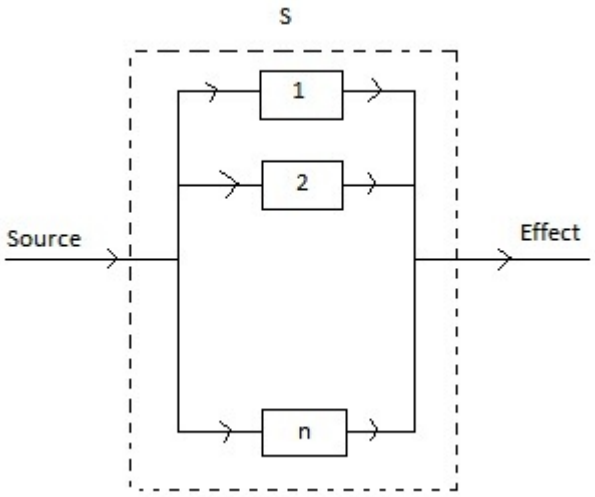
Solution. Let p be the reliability factor of each equipment. Then

$$P(S) = 0.95 = p^{10} \Rightarrow p = \sqrt[10]{0.95} = 0.99488.$$

■

B. **Parallel Configuration:** Several systems exist in which successful operation depends on the satisfactory functioning of any one of their n subsystems or elements. These are said to be connected in parallel. We can also ass a system in which several signal paths perform the same operation and the satisfactory performance of any one of these paths is sufficient to ensure the successful operation of the system. The elements of such a system are said to be connected in parallel.

A block diagram representing a parallel configuration is shown in the figure below The reliability of the



system can be calculated very easily by considering the conditions for system failure.

Let $X_1, X_2, \ldots, X_n$ represent successful operation of units $1, 2, \ldots, n$ respectively. Similarly, let $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_n$ respectively represent their successful operation, that is, the failure of the units.

If $P(X_1)$ is the probability of successful operation of unit 1, then $P(\overline{X}_1) = 1 - P(X_1)$, and so on. For the complete failure of the system S, all the n units have to fail simultaneously. If $P(\overline{S})$ is the probability of failure of the system, then

$$\begin{aligned} P(\overline{S}) &= P(\overline{X}_1 \text{ and } \overline{X}_2 \text{ and } \cdots \text{ and } \overline{X}_n) \\ &= P(\overline{X}_1)P(\overline{X}_2|\overline{X}_1)P(\overline{X}_3|\overline{X}_1\overline{X}_2)\ldots P(\overline{X}_n|\overline{X}_1\overline{X}_2\ldots\overline{X}_{n-1}) \end{aligned}$$

The expression $P(\overline{X}_3|\overline{X}_1\overline{X}_2)$ represents the probability of failure of unit 3 under the condition that units 1 and 2 have failed.

The other terms can also be interpreted in the same manner. If the unit failures are independent of one another, then

$$\begin{aligned} P(\overline{S}) &= P(\overline{X}_1)P(\overline{X}_2)\ldots P(\overline{X}_n) \\ &= [1 - P(X_1)][1 - P(X_2)]\ldots[1 - P(X_n)]. \end{aligned}$$

Since if any one of them does not fail, then the problem of successful configuration of the system is

$$P(S) = 1 - P(\overline{S}).$$

For independent cases, $P(S) = 1 - [1 - P(X_1)][1 - P(X_2)] \dots [1 - P(X_n)]$. If the n elements are identical and the unit failures are independent of one another, then

$$P(S) = 1 - (1 - P(X))^n$$

where, $P(X) = P(X_1) = P(X_2) = \cdots = P(X_n)$.

Example 2.2.5. Consider a system consisting of three identical units connected in parallel. The unit reliability factor is 0.10. If the unit failures are independent of one another and if the successful operation of the system depends on the satisfactory performance of any one unit, then determine the system reliability.

Solution. $P(S) = 1 - (1 - 0.10)^3 = 1 - 0.729 = 0.271$. This reveals the important fact that a parallel configuration can greatly increase system reliability with just three elements connected in parallel. ■

Example 2.2.6. A parallel system is composed of 10 independent identical components. If the system reliability $P(S)$, is to be 0.95, how poor can the components be?

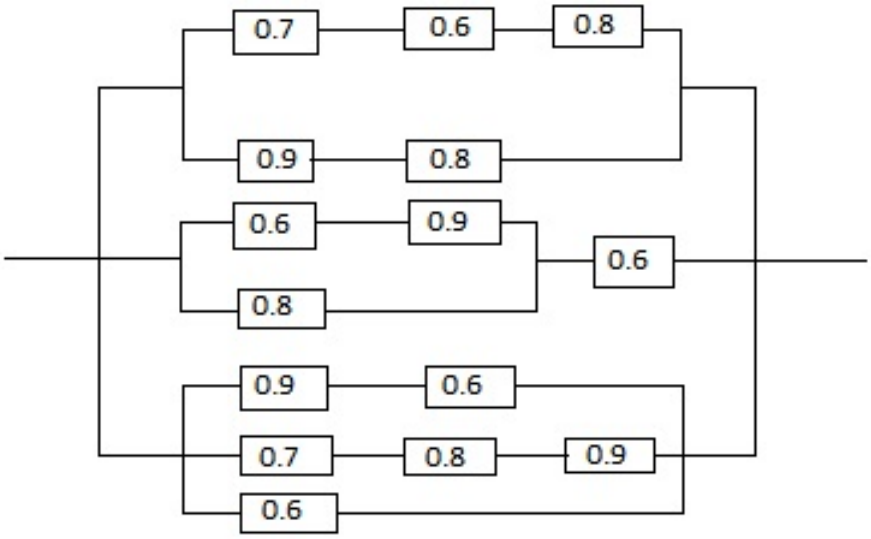
Solution. Let $P(X)$ be the probability of successful operation of each component. Thus,

$$\begin{aligned} P(S) &= 1 - (1 - P(X))^1 0 = 0.95 \\ \Rightarrow (1 - P(X))^1 0 &= 1 - 0.95 \\ &= 0.05 \\ \Rightarrow 1 - P(X) &= \sqrt[10]{0.05} = 0.74113 \\ \Rightarrow P(X) &= 1 - 0.74113 = 0.25887. \end{aligned}$$

Each component can have a very low reliability factor of 0.2589 but still gives the system a reliability factor as high as 0.95. ■

C. **Mixed Configuration:** Consider the following example:

Example 2.2.7. Find the reliability of the above system:



(KU 2011)

Solution. The complete system is composed of the following subsystems:

S_1: $\boxed{0.7} \longrightarrow \boxed{0.6} \longrightarrow \boxed{0.8}$
S_2: $\boxed{0.9} \longrightarrow \boxed{0.8}$
S_3: $\boxed{0.6} \longrightarrow \boxed{0.9}$
S_4: $\boxed{0.9} \longrightarrow \boxed{0.6}$
S_5: $\boxed{0.7} \longrightarrow \boxed{0.8} \longrightarrow \boxed{0.9}$
S_6: $S_1||S_2$
S_7: $S_3||\boxed{0.8}$
S_8: $S_4||S_5||\boxed{0.6}$
S_9: $S_7 \longrightarrow \boxed{0.6}$
S_{10}: $S_6||S_9||S_8$

Now,

$$\begin{aligned}
P(S_1) &= 0.7 \times 0.6 \times 0.8 = 0.336 \\
P(S_2) &= 0.9 \times 0.8 = 0.72 \\
P(S_3) &= 0.6 \times 0.9 = 0.54 \\
P(S_4) &= 0.9 \times 0.6 = 0.54 \\
P(S_5) &= 0.7 \times 0.8 \times 0.9 = 0.504 \\
P(S_6) &= 1 - [(1 - P(S_1))(1 - P(S_2))] \\
&= 1 - [(1 - 0.336)(1 - 0.72)] = 0.81408 \\
P(S_7) &= 1 - [(1 - P(S_3))(1 - 0.8)] \\
&= 1 - [(1 - 0.54)(1 - 0.8)] = 0.908 \\
P(S_8) &= 1 - [(1 - P(S_4))(1 - P(S_5))(1 - 0.6)] \\
&= 1 - [(1 - 0.54)(1 - 0.504)(1 - 0.6)] = 0.908736 \\
P(S_9) &= P(S_7) \times 0.6 = 0.908 \times 0.6 = 0.5448 \\
P(S_{10}) &= 1 - [(1 - P(S_6))(1 - P(S_9))(1 - P(S_8))] \\
&= 1 - [(1 - 0.81408)(1 - 0.5448)(1 - 0.908736)] \simeq 0.99228.
\end{aligned}$$

Hence the system reliability is 0.99228. ■

2.3 Redundancy

If the state of art is such that either it is not possible to produce highly reliable components or the cost of producing such components is very high, then we can improve the system reliability by the technique of introducing redundancies.

This involves the deliberate creation of new parallel path in a system. If two elements A, B with probability of success $P(A)$ and $P(B)$ are connected in parallel, then the probability of the successful operation of the system,

$$\begin{aligned}
P(A \text{ or } B) &= P(A) + P(B) - P(A \text{ and } B) \\
&= P(A) + P(B) - P(A)P(B),
\end{aligned}$$

assuming that the elements are independent.

Since both $P(A)$ and $P(B)$ are less than 1, excluding the condition where $P(A) = P(B) = 1$, then their product is always less than both $P(A)$ and $P(B)$.

This illustrates a simple method of improving the reliability of a system when the element reliability cannot be increased. Although either one of the elements is sufficient for the successful operation of the system, we deliberately use both elements so as to increase the reliability causing the system to become redundant.

Unit 3

Course Structure

- Information Theory: Fundamentals of Information theory
- Measures of information and characterisation
- Entropy and its properties

3.1 Introduction

In everyday life we observe that there are numerous means for the transmission of information. For example, the information is usually transmitted by means of a human voice, i.e., as in telephone, radio, television etc., by means of letters, newspapers, books etc. We often come across sentences like

- We have received a lot of information about the postponement of examination.
- We have a bit of information that he will be appointed as a professor.

But few people have suspected that it is really possible to measure information quantitatively. An amount of information has a useful numeric value just like an amount of sugar or an amount of bank balance. For example, suppose a man goes to a new community to rent a house and asks an unreliable agent “is this house cool in summer season?" If the agent answers ‘yes’, the man has received very little information, because more than likely that agent would have answered ‘yes’ regardless of the facts. If on the other hand, the man has a friend who lives in a neighbouring house, he can get more information by asking his friend the same question because the answer will be more reliable.

In general way it would appear that the amount of information in the message should be measured by extent of the change in probability produced by the message. There will be atleast three essential parts of simplest communication system:

- Transmitter or Source,
- Communication channel or transmission network which carries the message from the transmitter to the receiver,

- Receiver or Sink

3.2 Fundamental theorem of information theory

It is possible to transmit information through a noisy channel at any rate less than the channel capacity with an arbitrarily small probability of error.

3.2.1 Origination

The information theory is an appealing name assigned to a scientific discipline which deals with the mathematical theory of communication. The origin of information theory dates back to the work of R.V. Hartley ("Transmission of informations", Bellsys technical journal vol. 7, 1928), who tried to develop a quantitative measure of information in the telecommunication system. The field of information theory grown considerably often the publication of C.E. Shannon's ("A mathematical theory of communication", Bellsys technical journal, vol. 27, 1948). Information theory answers two fundamental questions in communication system.

a) What is the ultimate data compression?

b) What is the ultimate data transmission rate?

For this reason, some consider information theory as a subset of communication theory. Indeed it has fundamental contribution in statistical physics, computer science, probability and statistics, Biology, Economics etc. We see information only when we are in doubt which arises when there are number of alternatives and we are uncertain about the outcome of the event. On the other hand, if the event can occur in just one way, there is no uncertainty about it and no information is called for we get some information by the occurrence of the event when there was some uncertainty before its occurrence. Therefore, the amount of information received must be equal to the amount of uncertainty may be before the occurrence of the event.

3.3 Measure of information and characterisation

Let E be an event and p be its probability of occurrence. If we are told that the event E has occurred, then the question is "what is the amount of information conveyed by this message?" If p is close to 1, then it is nearly certain to occur and hence it conveys very little information. On the other hand, if p is close 0, then it is almost certain that E will not occur and consequently the message starting with its occurrence is quite unexpected. In general, let E_1 and E_2 are two events with p_1 and p_2 as their probability of occurrence respectively and let $p_1 < p_2$.

Then the event E_2 is more likely to occur and so the message conveying the occurrence of E_2 contains low information (bit information) than that conveying the occurrence of E_1. Further if p_2 continually decreased to p_1, the uncertainty associated with the occurrence of E_2 increases continually corresponding to the event E_1.

The above intuitive idea suggested that the measure of information conveyed by the message stating the occurrence of event with the probability p must be a function of p only, say $h(p)$, which is non-negative, strictly decreasing, continuous and $h(1) = 0$. Also $h(p)$ is very large when p is nearly equal to 0.

Next consider two events E_1 and E_2 with probability of occurrence p_1 and p_2 respectively. If we are told that the event E_1 has occurred, then we have received an amount of information $h(p_1)$. Giving this message, the probability that E_2 will occur is

$$p_{21} = p(E_2|E_1).$$

Suppose now we are told that the event E_2 has also occurred. Then the additional amount of information received is $h(p_{21})$.

Therefore the total amount of information received form their two successive messages is

$$h(p_1) + h(p_{21}).$$

Assume that the events E_1 and E_2 are independent. Then

$$p_{21} = p_2.$$

So the total amount of information received in this case is $h(p_1) + h(p_2)$.

Again, the probability of both the events E_1 and E_2 is p_1p_2 and the amount of information conveyed by the message stating that both the events E_1 and E_2 have occurred is $h(p_1p_2)$.

So from the above considerations we have

$$h(p_1p_2) = h(p_1) + h(p_2).$$

Thus from the above discussion we see that the amount of information received from the message stating that the event E with probability p has occurred is a function of p only, say $h(p)$ and has the following characterisations.

(i) $h(p)$ is non-negative, continuous and strictly decreasing function in p in $(0, 1]$.

(ii) $h(1) = 0$ and $h(p)$ is very large when p is very close to 0, i.e., $h(p) \to \infty$ as $p \to 0$.

(iii) if E_1 and E_2 are independent events with probability of occurrence p_1 and p_2 respectively, then the amount of information conveyed by the message stating that the occurrence of both events E_1 and E_2 is equal to the amount of information conveyed by the massage dealing with the event E_1 plus the amount of information dealing with the event E_2, i.e.,

$$h(p_1p_2) = h(p_1) + h(p_2).$$

Theorem 3.3.1. Let $h(p)$ denote the amount of information received form the message stating the event E with probability p has occurred. Then

$$h(p) = -k \log p,$$

where, k is a positive constant.

Proof. The function $h(p)$ has the following properties:

(i) $h(p)$ is non-negative, continuous, strictly decreasing in $(0, 1]$.

(ii) $h(1) = 0$ and $h(p) \to \infty$ as $p \to 0$.

(iii) $h(p_1p_2) = h(p_1) + h(p_2)$.

Take any $p \in (0, 1]$ and let n be a positive integer. We first show that

$$h(p^n) = nh(p) \tag{3.3.1}$$

Clearly, (3.3.1) holds for $n = 1$.

Assume that (3.3.1) holds for the positive integer n. Then

$$\begin{aligned} h(p^{n+1}) &= h(p^n \cdot p) \\ &= h(p^n) + h(p) \quad \text{[using property (iii)]} \\ &= n\,h(p) + h(p) \\ &= (n+1)\,h(p) \end{aligned}$$

Therefore, (3.3.1) holds for the positive integer $(n + 1)$.

Hence by the principle of finite induction, (3.3.1) holds for all $n \in \mathbb{N}$.

Let $p \in (0, 1]$ and $n \in \mathbb{N}$. Consider $q = p^{1/n}$ and $q \in (0, 1]$.

$$\begin{aligned} &\therefore p = q^n \text{ and } h(p) = h(q^n) = n\,h(q) \text{ (By (3.3.1)).} \\ &\Rightarrow h(q) = \frac{1}{n}h(p) \\ &\Rightarrow h(p^{1/n}) = \frac{1}{n}h(p) \end{aligned} \tag{3.3.2}$$

Let r be a positive rational number and $r = \dfrac{m}{n}$, where $m, n \in \mathbb{N}$.

$$\begin{aligned} \text{Then } h(p^r) &= h(p^{m/n}) \\ &= h\Big(\big(p^{1/n}\big)^m\Big) \\ &= m\,h(p^{1/n}) \\ &= \frac{m}{n}h(p) \\ &= r\,h(p) \end{aligned}$$

Let r be any positive number. Then choose any sequence $\{r_n\}$ of positive rational numbers such that $r_n \to r$ as $n \to \infty$. For such n, we get

$$h(p^{r_n}) = r_n\,h(p).$$

Since h is a constant function, letting $n \to \infty$, we get,

$$h(p^r) = r\,h(p) \tag{3.3.3}$$

Putting $p = \frac{1}{2}$ in (3.3.3) we get

$$h\left(\left(\frac{1}{2}\right)^r\right) = r\,h\left(\frac{1}{2}\right) \tag{3.3.4}$$

Let $p \in (0, 1]$. We write $r = \dfrac{\log p}{\log 1/2}$ so that $r > 0$ and $\left(\dfrac{1}{2}\right)^r = p$. Substituting in (3.3.4), we get

$$h(p) = -\frac{h(1/2)}{\log 2}\log p = -k \log p \quad \text{where} \quad k = \frac{h(1/2)}{\log 2}$$

Since h is strictly decreasing and $h(1) = 0$, therefore

$$h(1/2) > 0 \quad \text{and so} \quad k > 0.$$

□

3.3.1 Units of information

Taking $k = 1$, we have $h(p) = -\log p$. The choice of the base of the logarithmic amounts to the choice of the units of information,

(i) when base 2, i.e., $h(p) = -\log_2 p$, the unit is 'bits'.

(ii) when base is natural 'e', unit is 'nats'

(iii) when base is 10, unit is 'Hartley'

Note 3.3.2. 1 Har =3.32 bits and 1 nat =1.44 bits

3.4 Entropy (Shannon's Definition)

Let X be the random variable with range $\{x_1, x_2, \ldots, x_n\}$ and probability mass function (p.m.f)

$$\rho_X(x) = \begin{cases} p_i & \text{for } x = x_i \ (i = 1, 2, \ldots, n) \\ 0 & \text{otherwise.} \end{cases}$$

Then the quantity $-\sum_{i=1}^{n} p_i \log p_i$ is called the *entropy* of the random variable X and is denoted by $H(x)$ or $H_n(p_1, p_2, \ldots, p_n)$.

$$\therefore \text{ We have } H(X) = H_n(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log p_i.$$

Clearly, $H(X) \geq 0$.

Note 3.4.1. $x \log x \to 0$ as $x \to 0$ and we have used the conversion that $0 \log 0 = 0$.

3.4.1 Units of entropy

(i) when base is 2, unit of entropy is bits

(ii) when base is e, unit of entropy is nats

(iii) when base is 10, unit of entropy is Hartley

3.4.2 Properties of entropy function

(Shannon's characterization of entropy function)

1. For a fixed n, $H_n(p_1, p_2, \ldots, p_n)$ is a continuous function of $p_1, p_2, \ldots, p_n$ $(0 \leq p_i \leq 1, \ i = 1, 2, \ldots, n)$. It is obvious from the definition of H_n.

2. If $p_i = \frac{1}{n},\ i = 1, 2, \ldots, n$, then

$$\begin{aligned} H_n(p_1, p_2, \ldots, p_n) &= H_n\left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right) \\ &= -\sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{n} \\ &= \log n. \end{aligned}$$

So, $H_n\left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right)$ is strictly increasing function of n.

3.

$$\begin{aligned} \text{Let } s_1 &= p_1 + p_2 + \ldots + p_{n_1} \\ s_2 &= p_{n_1+1} + p_{n_1+2} + \ldots + p_{n_2} \\ \ldots &\quad \ldots \quad \ldots\ldots\ldots\ldots \\ s_k &= p_{n_{k-1}+1} + p_{n_{k-1}+2} + \ldots + p_{n_k}, \quad (n_k = n) \\ \text{and } m_1 &= n_1,\ m_2 = n_2 - n_1, \ldots, m_k = n_k - n_{k-1} \end{aligned}$$

Then,

$$\begin{aligned} H_n(p_1, p_2, \ldots, p_n) &= H_k(s_1, s_2, \ldots, s_k) + s_1 H_{m_1}\left(\frac{p_1}{s_1}, \frac{p_2}{s_1}, \cdots, \frac{p_{n_1}}{s_1}\right) \\ &+ s_2 H_{m_2}\left(\frac{p_{n_1+1}}{s_2}, \cdots, \frac{p_{n_2}}{s_2}\right) + \cdots + s_k H_{m_k}\left(\frac{p_{n_{k-1}+1}}{s_k}, \cdots, \frac{p_{n_k}}{s_k}\right). \end{aligned}$$

The above relation may be expressed as follows:

If a random experiment is decomposed into several successive ones, then the original value of H is equal to the weighted sum of the corresponding values of H with weights $1, s_1, s_2, \ldots, s_k$.

Now, we have $H_k(s_1, s_2, \ldots, s_k) = -\sum_{i=1}^{k} s_i \log s_i$.

$$\begin{aligned} \therefore s_1 H_{m_1}\left(\frac{p_1}{s_1}, \frac{p_2}{s_1}, \ldots, \frac{p_{n_1}}{s_1}\right) &= -s_1 \sum_{i=1}^{n_1} \frac{p_i}{s_1} \log \frac{p_i}{s_1} \\ &= -\sum_{i=1}^{n_1} p_i \log p_i + \sum_{i=1}^{n_1} p_i \log s_1 \\ &= -\sum_{i=1}^{n_1} p_i \log p_i + s_1 \log s_1. \end{aligned}$$

Similarly, we find

$$s_2 H_{m_2}\left(\frac{p_{n_1+1}}{s_2}, \ldots, \frac{p_{n_2}}{s_2}\right) = -\sum_{i=n_1+1}^{n_2} p_i \log p_i + s_2 \log s_2$$

$$\vdots$$

$$s_k H_{m_k}\left(\frac{p_{n_{k-1}} + 1}{s_k}, \ldots, \frac{p_{n_k}}{s_k}\right) = -\sum_{i=n_{k-1}+1}^{n_k} p_i \log p_i + s_k \log s_k$$

Adding the above expressions, we get

$$\begin{aligned} H_n(p_1, p_2, \ldots, p_n) &= -\sum_{i=1}^{n_k} p_i \log p_i \\ &= -\sum_{i=1}^{n} p_i \log p_i, \quad \text{where } n_k = n. \end{aligned}$$

Theorem 3.4.2. For a fixed n, the entropy function $H_n(p_1, p_2, \ldots, p_n)$ is maximum when $p_1 = p_2 = \ldots = p_n = \frac{1}{n}$ and $H_n(\max) = \log n$.

Proof. We first show that $\log x \leq x - 1$ for all $x > 0$ and the equality holds for $x = 1$.

Let $\phi(x) = x - 1 - \log x$ for all $x > 0$.

$\therefore \phi'(x) = 1 - \frac{1}{x}$.

If $x > 1$, then $\phi'(x) > 0$ and if $0 < x < 1$, then $\phi'(x) < 0$.

So $\phi(x)$ is a strictly increasing function in $(1, \infty)$ and strictly decreasing in $(0, 1)$.

Therefore, $\phi(x) \geq \phi(1) = 0$ for all $x > 0$.

$$\therefore \log x \leq x - 1 \text{ for all } x > 0 \tag{3.4.1}$$

Let us take $x = \dfrac{1}{np_i}$ in (3.4.1) and we get

$$\begin{aligned} &\log \frac{1}{np_i} \leq \frac{1}{np_i} - 1 \\ \Rightarrow\ & p_i \log \frac{1}{np_i} \leq \frac{1}{n} - p_i \\ \Rightarrow\ & -\sum_{i=1}^{n} p_i \log p_i - \sum_{i=1}^{n} p_i \log n \leq 1 - \sum_{i=1}^{n} p_i \\ \Rightarrow\ & -\sum_{i=1}^{n} p_i \log p_i \leq \log n \quad \left(\because \sum_{i=1}^{n} p_i = 1\right) \\ \Rightarrow\ & H_n(p_1, p_2, \ldots, p_n) \leq \log n \end{aligned} \tag{3.4.2}$$

When $p_1 = p_2 = \ldots = p_n = \frac{1}{n}$, then

$$\begin{aligned} H_n(p_1, p_2, \ldots, p_n) &= H_n\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right) \\ &= -\sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{n} \\ &= \log n \end{aligned} \tag{3.4.3}$$

From (3.4.2) and (3.4.3) we see that when the events are equally likely, H_n is maximum and its maximum value is $\log n$ i.e.,

$$H_n(\max) = \log n$$

□

Note 3.4.3. In this case units are taken as 'nats', since

$$\log_e x = \log_D x \log_e D \quad \text{for any} \quad D \geq 2.$$

Note 3.4.4. The entropy of X may be interpreted as the expected value of the function $\log \frac{1}{p_i}$ where p_i is the p.m.f of X. Thus

$$E\left[\log \frac{1}{p_i}\right] = \sum_{i=1}^{n} p_i \log \frac{1}{p_i} = -\sum_{i=1}^{n} p_i \log p_i = H(X).$$

Unit 4

Course Structure

- Bivariate Information Theory
- Joint, conditional and relative entropies
- Mutual Information

4.1 Joint, conditional and relative entropies

Let X, Y be two discrete random variables with ranges $\{x_1, x_2, \ldots, x_m\}$ and $\{y_1, y_2, \ldots, y_n\}$ respectively and probability mass functions $p(x)$ and $q(y)$ and joint p.m.f $p(x, y) = P(X = x; Y = y)$.

i) The joint entropy, $H(X, Y)$ of the pair of random variables X, Y is defined as

$$\begin{aligned} H(X,Y) &= -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log p(x_i, y_i) \\ &= E\left[\log \frac{1}{p(x,y)}\right] \end{aligned} \tag{4.1.1}$$

(ii) The conditional entropy, $H(X|Y)$ is defined by

$$\begin{aligned} H(X|Y) &= \sum_{j=1}^{n} q(y_j) H(X|Y = y_j) \\ &= -\sum_{j=1}^{n} q(y_j) \sum_{i=1}^{m} p(x_i|y_j) \log p(x_i|y_j) \\ &= -\sum_{i=1}^{m}\sum_{j=1}^{n} q(y_j) p(x_i|y_j) \log p(x_i|y_j) \\ &= -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log p(x_i|y_j) \\ &= E_{p(x,y)}\left[\log \frac{1}{p(x|y)}\right] \end{aligned}$$

Similarly, we can show that $H(Y|X) = E_{p(x,y)}\left[\log \frac{1}{p(y|x)}\right]$.

(iii) The relative entropy or Kullback-leibler distance between two probability mass functions $p(x)$ and $q(x)$ with $X = \{x_1, x_2, \ldots, x_m\}$ is defined as

$$\begin{aligned} D(p||q) &= \sum_{i=1}^{m} p(x_i) \log \frac{p(x_i)}{q(x_i)} \\ &= E_{p(x)}\left[\log \frac{p(x)}{q(x)}\right] \end{aligned}$$

4.2 Mutual information

Let X and Y be two discrete random variables with ranges $X = \{x_1, x_2, \ldots, x_m\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$ respectively and probability mass functions $p(x)$ and $q(y)$ with joint p.m.f $p(x, y) = p(X = x; Y = y)$. Then the mutual information of the random variables X and Y is denoted by $I(X; Y)$ and is defined by

$$\begin{aligned} I(X, Y) &= \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)q(y_j)} \\ &= D\big(p(x, y)||p(x)q(y)\big) \\ &= E_{p(x,y)}\left[\log \frac{p(x, y)}{p(x)q(y)}\right]. \end{aligned}$$

Theorem 4.2.1. Let $p_1, p_2, \ldots, p_n$ and $q_1, q_2, \ldots, q_n$ be two sets of non-negative numbers and $\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i$, then

$$\sum_{i=1}^{n} p_i \log_D q_i \leq \sum_{i=1}^{n} p_i \log_D p_i,$$

where D is any positive number greater than 1. Equality holds if and only if $p_i = q_i$ for all i.

Proof. We use the convention $0 \log 0 = 0$. First consider the case when $D = e$ and $p_i > 0,\ q_i > 0$ for all $i = 1, 2, \ldots, n$.

For any positive number x, we have

$$\log x \leq (x - 1) \tag{4.2.1}$$

equality holds if and only if $x = 1$. Taking $x = \frac{q_i}{p_i}$ in (4.2.1), we get

$$\log \frac{q_i}{p_i} \leq \frac{q_i}{p_i} - 1$$

Multiplying by p_i and taking summation we get

$$\begin{aligned} &\sum_{i=1}^{n} p_i \log \frac{q_i}{p_i} \leq \sum_{i=1}^{n} (q_i - p_i) = 0 \\ \Rightarrow &\sum_{i=1}^{n} p_i \log q_i \leq \sum_{i=1}^{n} p_i \log p_i \end{aligned} \tag{4.2.2}$$

Now, let $p_k = 0$ for some k and $q_k \neq 0$, but $p_i > 0,\ q_i > 0$ for $i \neq k$. Then clearly (4.2.2) holds because $p_k \log p_k = 0$ and $p_k \log q_k = 0$ if $q_k = 0$ for some k but $q_k \neq 0$. $p_k \log q_k = -\infty$ and so (4.2.2) holds.

Suppose that the equality holds in (4.2.2). Also assume that $p_k \neq q_k$ for some k. Then $\dfrac{q_k}{p_k} \neq 1$ and so

$$\log \frac{q_k}{p_k} < \frac{q_k}{p_k} - 1.$$

This gives

$$\begin{aligned} \sum_{i=1}^{n} p_i \log \frac{q_i}{p_i} &< \sum_{i=1}^{n} (q_i - p_i) = 0 \\ \Rightarrow \sum_{i=1}^{n} p_i \log q_i &< \sum_{i=1}^{n} p_i \log p_i \end{aligned}$$

which contradicts (4.2.2) since here equality does not hold because $\dfrac{q_k}{p_k} \neq 1$.

$$\therefore p_i = q_i \quad \text{for all } i.$$

Now, let $D \neq e$ for any $x > 0$. Then $\log_D x = \log_D e \cdot \log_e x$ and $\log_D e > 0$. So, multiplying (4.2.2) by $\log_D e$ we get

$$\sum_{i=1}^{n} p_i \log_D q_i \leq \sum_{i=1}^{n} p_i \log_D p_i.$$

□

Theorem 4.2.2. For any two discrete random variables X and Y

$$H(X,Y) \leq H(X) + H(Y)$$

Equality holds if and only if X, Y are independent.

Proof. Let X, Y be two discrete random variables with ranges $X = \{x_1, x_2, \ldots, x_m\}$, $Y = \{y_1, y_2, \ldots, y_n\}$ and probability mass functions (p.m.f) $p(x)$ and $q(y)$ with the joint p.m.f $p(x,y) = p(X = x; Y = y)$. We have

$$\begin{aligned} H(X) + H(Y) &= -\sum_{i=1}^{m} p(x_i) \log p(x_i) - \sum_{j=1}^{n} q(y_j) \log q(y_j) \\ &= -\sum_{i=1}^{m} \sum_{j=1}^{n} p(x_i, y_j) \log p(x_i) - \sum_{i=1}^{m} \sum_{j=1}^{n} p(x_i, y_j) \log q(y_j) \\ &= -\sum_{i=1}^{m} \sum_{j=1}^{n} p(x_i, y_j) \log \Big(p(x_i) q(y_j) \Big) \\ \text{Also, } H(X,Y) &= -\sum_{i=1}^{m} \sum_{j=1}^{n} p(x_i, y_j) \log p(x_i, y_j) \\ \text{Now, } \sum_{i=1}^{m} \sum_{j=1}^{n} p(x_i) q(y_j) &= \sum_{i=1}^{m} p(x_i) \sum_{j=1}^{n} q(y_j) = 1 \quad \text{and} \quad \sum_{i=1}^{m} \sum_{j=1}^{n} p(x_i, y_j) = 1 \end{aligned}$$

By Theorem 4.2.1, we have

$$\begin{aligned}
\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log p(x_i,y_j) &\geq \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log\Big(p(x_i)q(y_j)\Big)\\
\Rightarrow -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log p(x_i,y_j) &\leq -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log\Big(p(x_i)q(y_j)\Big)\\
\Rightarrow H(X,Y) &\leq H(X)+H(Y)
\end{aligned}$$

Equality holds if and only if $p(x_i,y_j)=p(x_i)q(y_j)$, i.e., if and if $X,\ Y$ are independent random variables. □

Theorem 4.2.3. For any two discrete random variables X and Y

$$H(X,Y)=H(X)+H(Y|X)=H(Y)+H(X|Y)$$

Proof. Let $X,\ Y$ be two discrete random variables with ranges $X=\{x_1,x_2,\ldots,x_m\}$ and $Y=\{y_1,y_2,\ldots,y_n\}$ and p.m.f $p(x)$ and $q(y)$ with joint p.m.f $p(x,y)=p(X=x;Y=y)$. Then

$$\begin{aligned}
H(X)+H(Y|X) &= -\sum_{i=1}^{m} p(x_i)\log p(x_i) - \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log p(y_j|x_i)\\
&= -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log p(x_i) - -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log p(y_j|x_i)\\
&= -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log\Big(p(x_i)\cdot p(y_j|x_i)\Big)\\
&= -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log p(x_i,y_j) \quad [\because p(x_i)p(y_j|x_i)=p(x_i,y_j)]\\
&= H(X,Y)
\end{aligned}$$

In a similar way, we can show that

$$H(Y)+H(X|Y)=H(X,Y)$$

□

Theorem 4.2.4. For any two discrete random variables X and Y,

$$I(X,Y)=H(X)-H(X|Y)=H(Y)-H(Y|X)$$

Proof. We have

$$\begin{aligned}
H(X)-H(X|Y) &= -\sum_{i=1}^{m} p(x_i)\log p(x_i) + \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log p(x_i|y_j)\\
&= -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log p(x_i) + \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log p(x_i|y_j)\\
&= \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log\frac{p(x_i|y_j)}{p(x_i)}\\
&= \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i,y_j)\log\frac{p(x_i,y_j)}{p(x_i)q(y_j)}\\
&= I(X,Y)
\end{aligned}$$

Similarly, we can show that

$$H(Y) - H(Y|X) = I(X;Y)$$

□

Note 4.2.5. I(X;Y)=I(Y,X)

Theorem 4.2.6. For any three discrete random variables $X,\ Y,\ Z$,

$$H\big((X,Y)|Z\big) = H(X|Z) + H\big(Y|(X,Z)\big)$$

Proof. Let X, Y, Z be three discrete random variables with ranges $X = \{x_1, x_2, \ldots, x_m\}$, $Y = \{y_1, y_2, \ldots, y_n\}$ and $Z = \{z_1, z_2, \ldots, z_k\}$ respectively and probability mass functions are $p(x),\ q(y)$ and $r(z)$ with joint p.m.f $p(x,y,z) = p(X = x; Y = y; Z = z)$. Then

$$\begin{aligned}
H(X|Z) + H\big(Y|(X,Z)\big) &= -\sum_{i=1}^{m}\sum_{l=1}^{k} p(x_i,z_l)\log p(x_i|z_l) - \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{k} p(x_i,y_j,z_l)\log p(y_j|(x_i,z_l)) \\
&= -\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{k} p(x_i,y_j,z_l)\log p(x_j|z_l) - \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{k} p(x_i,y_j,z_l)\log p(y_j|(x_i,z_l)) \\
&= -\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{k} p(x_i,y_j,z_l)\log p(x_i|z_l)p(y_j|(x_i,z_l)) \\
&= -\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{k} p(x_i,y_j,z_l)\log\left(\frac{p(x_i,z_l)}{p(z_l)}\cdot\frac{p(x_i,y_j,z_l)}{p(x_i,z_l)}\right) \\
&= -\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{k} p(x_i,y_j,z_l)\log\left(\frac{p(x_i,y_j,z_l)}{p(z_l)}\right) \\
&= -\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{k} p(x_i,y_j,z_l)\log p\big((x_i,y_j)|z_l\big) \\
&= H\Big((X,Y)|Z\Big)
\end{aligned}$$

□

Note 4.2.7. For n random variables

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, X_{i-2}, \ldots, X_1)$$

4.2.1 Conditional mutual information

i) The conditional mutual information of random variables $X,\ Y$ given Z is defined by

$$\begin{aligned}
I(X;Y|Z) &= H(X|Z) - H(X|(Y,Z)) \\
&= E_{p(x,y,z)}\left[\log\frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}\right]
\end{aligned}$$

ii) The conditional mutual information of random variables X and Y given $Z_1, Z_2, \ldots, Z_n$ is defined by

$$\begin{aligned} I(X;Y|Z_1,Z_2,\ldots,Z_n) &= H(X|Z_1,Z_2,\ldots,Z_n) - H(X|(Y,Z_1,Z_2,\ldots,Z_n)) \\ &= E_{p(x,y,z_1,\ldots,z_n)}\left[\log \frac{p(X,Y|Z_1,Z_2,\ldots,Z_n)}{p(X|Z_1,Z_2,\ldots,Z_n)p(Y|Z_1,Z_2,\ldots,Z_n)}\right] \end{aligned}$$

Theorem 4.2.8. (i) For the random variables X, Y, Z

$$I(X;Y,Z) = I(X;Y) + I(X;Z|Y) = I(X;Z) + I(X;Y|Z)$$

Proof.

$$\begin{aligned} I(X;Y) + I(X;Z|Y) &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} + \sum_x \sum_y \sum_z p(x,y,z) \log \frac{p(x,z|y)}{p(x|y)p(z|y)} \\ &= \sum_x \sum_y \sum_z p(x,y,z) \log \frac{p(x,y)}{p(x)p(y)} + \sum_x \sum_y \sum_z p(x,y,z) \log \frac{p(x,z|y)}{p(x|y)p(z|y)} \\ &= \sum_x \sum_y \sum_z p(x,y,z) \log \left\{ \frac{p(x,y)}{p(x)p(y)} \cdot \frac{p(x,y,z)}{p(y)} \cdot \frac{p(y)}{p(x,y)} \cdot \frac{p(y)}{p(y,z)} \right\} \\ &= \sum_x \sum_y \sum_z p(x,y,z) \log \frac{p(x,y,z)}{p(x)p(y,z)} \\ &= I(X;Y,Z) \end{aligned}$$

Similarly, we can show that

$$I(X;Z) + I(X;Y|Z) = I(X;Y,Z)$$

□

Theorem 4.2.9. (ii) For the random variables $X_1, X_2, \ldots, X_n, Y$

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, X_{i-2}, \ldots, X_1)$$

Proof. Follows from induction on n. □

Theorem 4.2.10. (Information inequality): Let $p(x)$ and $q(x)$ for $x \in X$ be two probability mass functions. Then

$$D(p||q) \geq 0$$

Proof. Let $X = \{x_1, x_2, \ldots, x_n\}$ and let $p_i = p(x_i),\ q_i = q(x_i),\ i = 1, 2, \ldots, n$. Then by definition,

$$\begin{aligned} D(p||q) &= \sum_{i=1}^{n} p(x_i) \log \frac{p(x_i)}{q(x_i)} \\ &= \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \end{aligned}$$

Also we have $\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i = 1$. So, by Theorem 4.2.1,

$$\begin{aligned} &\sum_{i=1}^{n} p_i \log q_i \leq \sum_{i=1}^{n} p_i \log p_i \\ \Rightarrow &\sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \geq 0 \\ \Rightarrow &D(p||q) \geq 0. \end{aligned}$$

□

Theorem 4.2.11. (Non-negativity of mutual information) For any two random variables X and Y, $I(X;Y) \geq 0$.

Proof. Let X and Y be two discrete random variables with range $\{x_1, x_2, \ldots, x_m\}$ and $\{y_1, y_2, \ldots, y_n\}$ respectively and the p.m.f $p(x)$ and $q(y)$, joint p.m.f $p(x,y) = P(X = x, Y = y)$. Then the mutual information $I(X;Y)$ between X and Y is given by

$$I(X;Y) = \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)q(y_j)}.$$

Now, we have

$$\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) = 1 \quad \text{and} \quad \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i)q(y_j) = \sum_{i=1}^{m} p(x_i) \sum_{j=1}^{n} q(y_j) = 1$$

So by Theorem 4.2.1

$$\begin{aligned} &\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log p(x_i, y_j) \geq \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log \Big(p(x_i)q(y_j)\Big) \\ i.e., \quad &\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)q(y_j)} \geq 0 \\ i.e., \quad &I(X;Y) \geq 0. \end{aligned}$$

□

Theorem 4.2.12. (Non-negativity of conditional mutual information) For any two random variables X and Y given Z, the conditional mutual information $I(X;Y|Z) \geq 0$.

Proof. Let X, Y, Z be three discrete random variables with ranges $\{x_1, x_2, \ldots, x_m\}$, $\{y_1, y_2, \ldots, y_n\}$, $\{z_1, z_2, \ldots, z_k\}$ respectively and probability mass functions $p(x)$, $p(y)$, $p(z)$ with joint p.m.f $p(x,y,z) = P(X = x, Y = y, Z = z)$.

Then by definition,

$$I(X;Y|Z) = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{k} p(x_i, y_j, z_l) \log \frac{p(x_i, y_j|z_l)}{p(x_i|z_l)p(y_j|z_l)}$$

$$\text{Now, } \frac{p(x_i, y_j|z_l)}{p(x_i|z_l)p(y_j|z_l)} = \frac{p(x_i, y_j, z_l)}{p(z_l)} \frac{p(z_l)}{p(x_i, z_l)} \frac{p(z_l)}{p(y_j, z_l)}$$

$$= \frac{p(x_i, y_j, z_l)}{\frac{p(x_i, z_l)p(y_j, z_l)}{p(z_l)}}$$

$$\therefore I(X, Y|Z) = \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{l=1}^{k} p(x_i, y_j, z_l) \log \frac{p(x_i, y_j, z_l)}{\frac{p(x_i, z_l)p(y_j, z_l)}{p(z_l)}}$$

$$\text{Now, } \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{l=1}^{k} p(x_i, y_j, z_l) = 1$$

$$\text{and, } \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{l=1}^{k} \frac{p(x_i, z_l)p(y_j, z_l)}{p(z_l)} = \sum_{i=1}^{m} \sum_{l=1}^{k} p(x_i, z_l) \cdot \sum_{j=1}^{n} \frac{p(y_j, z_l)}{p(z_l)}$$

$$= \sum_{i=1}^{m} \sum_{l=1}^{k} p(x_i, z_l) \frac{p(z_l)}{p(z_l)} \quad \left(\because \sum_{j=1}^{n} p(y_j, z_l) = p(z_l) \sum_{j=1}^{n} p(y_j) = p(z_l) \right)$$

$$= \sum_{i=1}^{m} \sum_{l=1}^{k} p(x_i, z_l)$$

$$= 1$$

Therefore, by Theorem 4.2.1, $I(X; Y|Z) \geq 0$. □

Unit 5

Course Structure

- Conditional Relative Entropy
- Channel Capacity
- Redundancy

5.1 Conditional relative entropy

The conditional relative entropy $D(p(y|x)||q(y|x))$ is the average of the relative entropies between the conditional probability mass functions $p(y|x)$ and $q(y|x)$ averaged over the p.m.f $p(x, y)$.

$$\begin{aligned}
\therefore D(p(y|x)||q(y|x)) &= \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)} \\
&= \sum_x \sum_y p(x, y) \log \frac{p(y|x)}{q(y|x)} \\
&= E_{p(x,y)} \left[\log \frac{p(Y|X)}{q(Y|X)} \right]
\end{aligned}$$

5.1.1 Convex and Concave functions

Let I be an interval and $f : I \to \mathbb{R}$ be a function. The function f is said to be convex if for any two points $x_1,\ x_2\ (x_1 \neq x_2)$ in I and $\lambda,\ \mu \geq 0$ with $\lambda + \mu = 1$, the relation

$$f(\lambda x_1 + \mu x_2) \leq \lambda f(x_1) + \mu f(x_2)$$

holds. A function $g : I \to \mathbb{R}$ is said to be concave if $-g$ is convex.

5.1.2 Jensen's Inequality

If $f : \mathbb{R} \to \mathbb{R}$ is a convex function and X is a random variable, then $f(E \cdot X) \leq Ef(X)$, where E is a constant. Moreover, if f strictly convex, then the equality implies that X is constant.

Theorem 5.1.1. (Log-Sum Inequality) Let $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ be two sets of n non-negative numbers. Then

$$\sum_{i=1}^{n} a_i \log_D \frac{a_i}{b_i} \geq \sum_{i=1}^{n} a_i \log_D \left(\frac{\sum a_i}{\sum b_i} \right)$$

where D is any positive number and $D > 1$. Equality holds if and only if $\frac{a_i}{b_i}$ is constant.

Proof. We use the conventions $0 \log 0 = 0,\ a \log \frac{a}{0} = +\infty,\ 0 \log \frac{0}{0} = 0$. Without loss of generality we may assume that $a_i > 0,\ b_i > 0,\ i = 1, 2, \ldots, n$.

Consider the function $f(t) = t \log_D t,\ t > 0$. Therefore, we have

$$f'(t) = (1 + \log_e t) \log_D e$$
$$f''(t) = \frac{1}{t} \log_D e > 0 \text{ for all } t > 0$$

So, $f(t)$ is strictly convex for $t > 0$.

Now consider

$$\lambda = \sum_{i=1}^{n} b_i, \quad \alpha_i = \frac{b_i}{\lambda}, \quad t_i = \frac{a_i}{b_i}$$

Then $\sum_{i=1}^{n} \alpha_i = 1$ and $\alpha_i > 0$ for all i.

So, by Jensen's inequality, we have

$$\sum_{i=1}^{n} \alpha_i f(t_i) \geq f\left(\sum_{i=1}^{n} \alpha_i t_i \right) \tag{5.1.1}$$
$$\Rightarrow \sum_{i=1}^{n} \frac{b_i}{\lambda} \frac{a_i}{b_i} \log_D \left(\frac{a_i}{b_i} \right) \geq \left(\sum_{i=1}^{n} \frac{a_i}{\lambda} \right) \log_D \left(\sum_{i=1}^{n} \frac{b_i}{\lambda} \frac{a_i}{b_i} \right)$$
$$\Rightarrow \sum_{i=1}^{n} a_i \log_D \left(\frac{a_i}{b_i} \right) \geq \sum_{i=1}^{n} a_i \log_D \left(\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \right) \tag{5.1.2}$$

If $\frac{a_i}{b_i}$ = constant = k (say), for $i = 1, 2, \ldots, n$.
Then clearly equality in (5.1.2) holds.

Suppose that equality holds in (5.1.2) i.e., in (5.1.1). Then,

$$t_1 = t_2 = \ldots = t_n$$
$$\Rightarrow \frac{a_1}{b_1} = \frac{a_2}{b_2} = \ldots = \frac{a_n}{b_n}$$
$$i.e., \ \frac{a_i}{b_i} = \text{ constant; } \ i = 1, 2, \ldots, n$$

□

Theorem 5.1.2. $D(p||q)$ is convex in pair (p, q) i.e., if $(p_1, q_1),\ (p_2, q_2)$ be two pairs of probability mass functions and $\lambda > 0,\ \mu > 0$ with $\lambda + \mu = 1$, then

$$D((\lambda p_1 + \mu p_2)||(\lambda q_1 + \mu q_2)) \leq \lambda D(p_1||q_1) + \mu D(p_2||q_2).$$

Proof. Let (p_1, q_1) and (p_2, q_2) be two pairs of probability mass functions and $\lambda > 0,\ \mu > 0$ with $\lambda + \mu = 1$. Then by Log-Sum inequality, we have

$$\begin{aligned}(\lambda p_1(x) + \mu p_2(x)) \log \frac{\lambda p_1(x) + \mu p_2(x)}{\lambda q_1(x) + \mu q_2(x)} &\leq \lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + \mu p_2(x) \log \frac{\mu p_2(x)}{\mu q_2(x)} \\ &= \lambda p_1(x) \log \frac{p_1(x)}{q_1(x)} + \mu p_2(x) \log \frac{p_2(x)}{q_2(x)}\end{aligned}$$

Now, taking summation we get

$$\begin{aligned}&\sum_x \{\lambda p_1(x) + \mu p_2(x)\} \log \frac{\lambda p_1(x) + \mu p_2(x)}{\lambda q_1(x) + \mu q_2(x)} \leq \sum_x \lambda p_1(x) \log \frac{p_1(x)}{q_1(x)} + \sum_x \mu p_2(x) \log \frac{p_2(x)}{q_2(x)} \\ &\Rightarrow D\Big((\lambda p_1 + \mu p_2)||(\lambda q_1 + \mu q_2)\Big) \leq \lambda D(p_1||q_1) + \mu D(p_2||q_2) \\ &i.e.,\ \ D(p||q) \ \text{ is convex in } \ (p, q).\end{aligned}$$

□

Theorem 5.1.3. The entropy function $H(p)$ is a concave function of p.

Proof. The entropy function $H(p)$ is defined by

$$H(p) = -\sum_{i=1}^{n} p_i \log_D p_i$$

$$\begin{aligned}\text{Now, } \frac{\partial H}{\partial p_i} &= -\{1 + \log_e p_i\} \log_D e \\ \frac{\partial^2 H}{\partial p_i^2} &= -\frac{1}{p_i} \log_D e \\ \frac{\partial^2 H}{\partial p_i \partial p_j} &= 0, \quad i \neq j\end{aligned}$$

The Hessian matrix is given by

$$\nabla^2 H(p) = \begin{bmatrix} \frac{\partial^2 H}{\partial p_1^2} & \frac{\partial^2 H}{\partial p_1 \partial p_2} & \cdots & \frac{\partial^2 H}{\partial p_1 \partial p_n} \\ \frac{\partial^2 H}{\partial p_2 \partial p_1} & \frac{\partial^2 H}{\partial p_2^2} & \cdots & \frac{\partial^2 H}{\partial p_2 \partial p_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 H}{\partial p_n \partial p_1} & \frac{\partial^2 H}{\partial p_n \partial p_2} & \cdots & \frac{\partial^2 H}{\partial p_n^2} \end{bmatrix} = \begin{bmatrix} -\frac{1}{p_1} & 0 & \cdots & 0 \\ 0 & -\frac{1}{p_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & -\frac{1}{p_n} \end{bmatrix} \log_D e$$

Clearly, $\nabla^2 H(p)$ is negative definite for $p_i > 0,\ \ (\because \log_D e > 0)$.

Hence, $H(p)$ is a concave function of p. □

Theorem 5.1.4. Non-negativity of conditional relative entropy

$$D\Big(p(y|x)||q(y|x)\Big) \geq 0.$$

Proof.

$$\begin{aligned}
\text{We have, } D(p(y|x)||q(y|x)) &= \sum_x \sum_y p(x,y) \log \frac{p(y|x)}{q(y|x)} \\
&= \sum_x \sum_y p(x,y) \log \frac{p(x,y)\, q(x)}{q(x,y)\, p(x)} \\
\text{Now, } \sum_x \sum_y p(x,y) \cdot q(x) &= \sum_x \sum_y q(x,y)\, p(x) = 1
\end{aligned}$$

$$\therefore \text{By Theorem } 4.2.1, \quad D(p(y|x)||q(y|x)) \geq 0.$$

□

Example 5.1.5. In a certain community, 25% of all girls are blondes, and 75% of all blondes are blue eyed. Also, 50% of all girls in the community have blue eyes. If you know that a girl has blue eyes, how much additional information do you being informed that she is blond?

Solution. Let $p_1 =$ probability of a girl being blonde $= 0.25$.

$p_2 =$ probability of a girl to have blue eyes if she is blonde = p_{blonde} (blue eyes) = 0.75

$p_3 = p$ (blue eyes) = 0.50

$p_4 = p$ (blonde, blue eyes) = probability that a girl is blonde and has blue eyes

and $p_x = p_{\text{ blue eyes}}$ (blonde) = probability that a blue eyed girl is blonde = ?

Then

$$p_4 = p_1 p_2 = p_3 p_x \Rightarrow p_x = \frac{p_1 p_2}{p_3} = \frac{0.25 \times 0.75}{0.50}$$

If a girl has blue eyes, the additional information obtained by being informed that she is blonde is

$$\begin{aligned}
\log_2 \frac{1}{p_x} &= \log_2 \frac{p_3}{p_1 p_2} \\
&= \log_2 p_3 - \log_2 p_1 - \log_2 p_2 \\
&= \log_2 \frac{1}{2} - \log_2 \frac{1}{4} - \log_2 \frac{3}{4} \\
&= \log_2 4 + \log_2 \frac{4}{3} - \log_2 2 \\
&= 1.41503 \\
&\approx 1.42 \text{ bits}
\end{aligned}$$

■

Example 5.1.6. Evaluate the average uncertainty associated with the probability of events A, B, C, D with probability of events $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}$ respectively.

Solution.

$$\begin{aligned}\text{We have, } H\left(\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{8}\right) &= -\frac{1}{2}\log\frac{1}{2}-\frac{1}{4}\log\frac{1}{4}-\frac{1}{8}\log\frac{1}{8}-\frac{1}{8}\log\frac{1}{8}\\ &= \frac{1}{2}\log_2 2+\frac{1}{2}\log_2 2+\frac{3}{4}\log_2 2\\ &= \left(\frac{1}{2}+\frac{1}{2}+\frac{3}{4}\right)\log_2 2\\ &= \frac{7}{4}\text{ bits}\end{aligned}$$

which is the averaged uncertainty associated with the probability of events A, B, C, D. ■

Example 5.1.7. A transmitter has an alphabet consisting of 5 letters $\{x_1, x_2, x_3, x_4, x_5\}$ and the receiver has an alphabet consisting of 4 letters $\{y_1, y_2, y_3, y_4\}$. The joint probabilities for communication are given below

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array}\begin{array}{c} \begin{array}{cccc} y_1 & y_2 & y_3 & y_4 \end{array} \\ \left[\begin{array}{cccc} 0.25 & 0.00 & 0.00 & 0.00 \\ 0.10 & 0.30 & 0.00 & 0.00 \\ 0.00 & 0.05 & 0.10 & 0.00 \\ 0.00 & 0.00 & 0.05 & 0.10 \\ 0.00 & 0.00 & 0.05 & 0.00 \end{array}\right] \end{array}$$

Determine the marginal, conditional and joint entropies for this channel. (Assume $0\log 0 \equiv 0$)

Solution. The channel is described here by joint probabilities $p_{ij},\ i = 1,2,3,4,5;\ j = 1,2,3,4$. Then the conditional and marginal probabilities are easily obtained from p_{ij}'s as follows:

$$\begin{aligned} p_{10} &= 0.25+0.00+0.00+0.00 = 0.25\\ p_{20} &= 0.10+0.30+0.00+0.00 = 0.40\\ p_{30} &= 0.00+0.05+0.10+0.00 = 0.15\\ p_{40} &= 0.00+0.00+0.05+0.10 = 0.15\\ p_{50} &= 0.00+0.00+0.05+0.00 = 0.05\\ p_{01} &= 0.25+0.10+0.00+0.00+0.00 = 0.35\\ p_{02} &= 0.00+0.30+0.05+0.00+0.00 = 0.35\\ p_{03} &= 0.00+0.00+0.10+0.05+0.05 = 0.20\\ p_{04} &= 0.00+0.00+0.00+0.10+0.00 = 0.10\end{aligned}$$

By using the result, $p_{j|i} = \dfrac{p_{ij}}{p_{i0}}$, the conditional probabilities are given in the following channel matrix

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array}\begin{array}{c} \begin{array}{cccc} y_1 & y_2 & y_3 & y_4 \end{array} \\ \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{3}{4} & 0 & 0 \\ 0 & \frac{1}{3} & \frac{2}{3} & 0 \\ 0 & 0 & \frac{1}{3} & \frac{2}{3} \\ 0 & 0 & 1 & 0 \end{array}\right] \end{array}$$

Marginal entropies:

$$\begin{aligned}
\therefore H(X) &= -\sum_{i=1}^{5} p_{i0} \log_2 p_{i0} \\
&= -(0.25 \log_2 0.25 + 0.40 \log_2 0.40 + \ldots + 0.05 \log_2 0.05) \\
&= 1.326 \text{ bits} \\
\therefore H(Y) &= -\sum_{j=1}^{4} p_{0j} \log_2 p_{0j} \\
&= 1.8556 \text{ bits}
\end{aligned}$$

Conditional entropies

$$\begin{aligned}
H(Y|X) &= -\sum_{i=1}^{5} \sum_{j=1}^{4} p_{ij} \log_2 p_{j|i} = 0.6 \text{ bits} \\
\text{Similarly, } H(X|Y) &= H(X) + H(Y|X) - H(Y) \\
&= 1.3260 + 0.6 - 1.8336 = 0.0704 \text{ bits}
\end{aligned}$$

Joint Entropy

$$\begin{aligned}
H(X,Y) &= H(X) + H(Y|X) \\
&= 1.3260 + 0.6 \\
&= 1.9260 \text{ bits}
\end{aligned}$$

■

5.2 Channel Capacity

Definition 5.2.1. Mutual information $I(X;Y)$ indicates a measure of the average information per symbol transmitted in the system. According to Shannon, in a discrete communication system, the channel capacity is the maximum of the mutual information, i.e.,

$$C = \max I(X;Y) = \max\{H(X) - H(X|Y)\}$$

For noise free channel, $I(X;Y) = H(X) = H(Y) = H(X,Y)$. Thus

$$C = \max I(X;Y) = \max\{H(X)\} = \max\left\{-\sum_{i=1}^{n} p_i \log p_i\right\}$$

Since $\max\{H(X)\}$ occurs when all symbols have equal probabilities, hence the channel capacity for a noise free channel is

$$C = -\log\left(\frac{1}{n}\right) = \log_2 n \text{ bits/symbol.}$$

5.3 Redundancy

i)

$$\begin{aligned}
\text{Absolute redundancy} &= C - I(X;Y) \\
&= C - H(X) \\
&= \log n - H(X) \quad \text{(For noise free channel)}
\end{aligned}$$

ii)

$$\begin{aligned}\text{Relative redundancy} &= \frac{C - I(X;Y)}{C} \\ &= \frac{\log n - H(X)}{\log n} = 1 - \frac{H(X)}{\log n}\end{aligned}$$

iii)

$$\begin{aligned}\text{Efficiency of a noise free system} &= \frac{H(X)}{\log n} \\ &= 1 - \text{Relative redundancy.}\end{aligned}$$

Example 5.3.1. Find the capacity of the memory less channel specified by the channel matrix

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

Solution. The capacity of the memoryless channel is given by

$$\begin{aligned} C &= \max I(X,Y) \\ &= \max\{H(X) + H(Y) - H(X,Y)\} \\ &= -\sum_{i=1}^{4} p_{ij} \log p_{ij}, \quad j = 1,2,3,4 \\ &= -\sum_{i=1}^{4} p_{i1} \log p_{i1} - \sum_{i=1}^{4} p_{i2} \log p_{i2} - \sum_{i=1}^{4} p_{i3} \log p_{i3} - \sum_{i=1}^{4} p_{i4} \log p_{i4} \end{aligned}$$

where

$$\begin{aligned} p_{i1} &= \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}, 0\right) \\ p_{i2} &= \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) \\ p_{i3} &= (0,0,1,0) \\ p_{i1} &= \left(\frac{1}{2}, 0, 0, \frac{1}{2}\right) \end{aligned}$$

$$\begin{aligned} \text{Thus, } C &= \frac{1}{2}\log_2 \frac{1}{2} + 2\left(\frac{1}{4}\log_2 \frac{1}{4}\right) + 4\left(\frac{1}{4}\log_2 \frac{1}{4}\right) + 1\log_2 1 + 2\left(\frac{1}{2}\log_2 \frac{1}{2}\right) \\ &= \frac{3}{2}\log_2 2 + 3\log_2 2 \\ &= \frac{9}{2} \text{ bits/symbol} \end{aligned}$$

■

Example 5.3.2. Show that the entropy of the following probability distribution is $2 - \left(\frac{1}{2}\right)^{n-2}$.

Events	x_1	x_2	$\ldots$	x_i	$\ldots$	x_{n-1}	x_n	x_{n+1}
Probabilities	$\frac{1}{2}$	$\frac{1}{2^2}$	$\ldots$	$\frac{1}{2^i}$	$\ldots$	$\frac{1}{2^{n-1}}$	$\frac{1}{2^{n-1}}$	$\frac{1}{2^n}$

Solution. From the given data of the problem, we have

$$\begin{aligned} p_i &= \frac{1}{2^i},\ \ i = 1, 2, \ldots, n-1 \ \text{ and } \ p_n = \frac{1}{2^{n-1}} \\ \text{and} \quad \sum_{i=1}^{n} p_i &= \left[\frac{1}{2} + \frac{1}{2^2} + \ldots + \frac{1}{2^{n-1}}\right] + \frac{1}{2^{n-1}} \\ &= \frac{1}{2}\frac{1 - \frac{1}{2^{n-1}}}{1 - \frac{1}{2}} + \frac{1}{2^{n-1}} \\ &= 1 - \frac{1}{2^{n-1}} + \frac{1}{2^{n-1}} \\ &= 1 \end{aligned}$$

The entropy function H is defined as

$$\begin{aligned} H(p_1, p_2, \ldots, p_n) &= -\sum_{i=1}^{n} p_i \log p_i \\ \Rightarrow H(p_1, p_2, \ldots, p_n) &= -\sum_{i=1}^{n-1} p_i \log p_i - p_n \log p_n \\ \Rightarrow H(p_1, p_2, \ldots, p_n) &= -\sum_{i=1}^{n-1} \left(\frac{1}{2^i}\right) \log_2 \left(\frac{1}{2^i}\right) - \frac{1}{2^{n-1}} \log_2 \left(\frac{1}{2^{n-1}}\right) \\ \Rightarrow H(p_1, p_2, \ldots, p_n) &= \sum_{i=1}^{n-1} \left(\frac{1}{2^i}\right) \log_2 (2^i) + \frac{1}{2^{n-1}} \log_2 (2^{n-1}) \\ \Rightarrow H(p_1, p_2, \ldots, p_n) &= \sum_{i=1}^{n-1} i \cdot \frac{1}{2^i} + (n-1)\frac{1}{2^{n-1}} \\ \Rightarrow H(p_1, p_2, \ldots, p_n) &= \left\{\frac{1}{2} + \frac{2}{2^2} + \frac{3}{2^3} + \ldots + \frac{n-1}{2^{n-1}}\right\} + \frac{n-1}{2^{n-1}} \qquad (5.3.1) \\ \Rightarrow \frac{1}{2} H(p_1, p_2, \ldots, p_n) &= \left\{\frac{1}{2^2} + \frac{2}{2^3} + \frac{3}{2^4} + \ldots + \frac{n-1}{2^n}\right\} + \frac{n-1}{2^n} \qquad (5.3.2) \end{aligned}$$

Subtracting (5.3.2) from (5.3.1) we get,

$$\begin{aligned}
\frac{1}{2}H(p_1,p_2,\ldots,p_n) &= \left(\frac{1}{2}-\frac{1}{2^2}\right)+\left(\frac{2}{2^2}-\frac{2}{2^3}\right)+\left(\frac{3}{2^3}-\frac{3}{2^4}\right)+\ldots \\
&+ \left(\frac{n-1}{2^{n-1}}-\frac{n-1}{2^n}\right)+\left(\frac{n-1}{2^{n-1}}-\frac{n-1}{2^n}\right) \\
&= \frac{1}{2}+\left(\frac{2}{2^2}-\frac{1}{2^2}\right)+\left(\frac{3}{2^3}-\frac{2}{2^3}\right)+\left(\frac{4}{2^4}-\frac{3}{2^4}\right)+\ldots \\
&+ \left(\frac{n-1}{2^{n-1}}-\frac{n-2}{2^{n-1}}\right)-\frac{n-1}{2^n}+\frac{n-1}{2^n} \\
&= \frac{1}{2}+\frac{1}{2^2}+\frac{1}{2^3}+\ldots++\frac{1}{2^{n-1}} \\
&= 1-\left(\frac{1}{2}\right)^{n-1}
\end{aligned}$$

$$\therefore H(p_1,p_2,\ldots,p_n) = 2-\left(\frac{1}{2}\right)^{n-2}$$

■

Example 5.3.3. If the probability distribution $P=\{p_1,p_2,\ldots\},\ p_i\geq 0, \sum_{i=1}^{\infty} p_i = 1$ is such that the entropy function, $H(P) = -\sum_{i=1}^{\infty} p_i \log p_i < \infty$, then show that $\sum_{i=1}^{\infty} p_i \log i < \infty$.

Solution. Let us assume that $\{p_i\}$ are decreasing in i, which is quite possible because reordering of the $\{p_i\}$ does not affect the value of entropy. Then

$$1 = \sum_{j=1}^{\infty} p_j \geq \sum_{j=1}^{i} p_j \geq ip_i$$

Thus we have $-\log p_i > \log i$ and consequently

$$\sum_{i=1}^{\infty} p_i \log i \leq -\sum_{i=1}^{\infty} p_i \log p_i = H(P) < \infty.$$

Hence, $\sum_{i=1}^{\infty} p_i \log i < \infty$.

■

The following example is similar.

Example 5.3.4. If the probability distribution $\Phi=(p_1,p_2,\ldots),\ p_i\geq 0,\ \sum_{i=1}^{\infty} p_i = 1$ is such that $\sum_{i=1}^{\infty} p_i \log i < \infty$, then show that $H(\Phi) = -\sum_{i=1}^{\infty} p_i \log p_i < \infty$.

Example 5.3.5. Let H be the entropy of the probability distribution $p_1, p_2, \ldots, p_n$. If H_1 be the entropy of the probability distribution $p_1 + p_2, p_3, \ldots, p_n$, then show that

$$H - H_1 = P_s H_s \text{ where } P_s = p_1 + p_2 \text{ and } H_s = \left[\frac{p_1}{P_s}\log\frac{P_s}{p_1} + \frac{p_2}{P_s}\log\frac{P_s}{p_2}\right]$$

Solution. We have

$$\begin{aligned} H &= -p_1\log p_1 - p_2\log p_2 - p_3\log p_3 \ldots - p_n\log p_n \quad (5.3.3)\\ H_1 &= -(p_1+p_2)\log(p_1+p_2) - p_3\log p_3 - \ldots - p_n\log p_n \\ &= -P_s\log P_s - p_3\log p_3 - \ldots - p_n\log p_n \quad (5.3.4)\end{aligned}$$

Subtracting (5.3.4) from (5.3.3), we get

$$\begin{aligned} H - H_1 &= -p_1\log p_1 - p_2\log p_2 + P_s\log P_s \\ &= P_s\cdot\frac{1}{P_s}\Big[-p_1\log p_1 - p_2\log p_2 + P_s\log P_s\Big] \\ &= P_s\left[-\frac{p_1}{P_s}\log p_1 - \frac{p_2}{P_s}\log p_2 + \frac{p_1+p_2}{P_s}\log P_s\right] \\ &= P_s\left[\frac{p_1}{P_s}\log P_s - \frac{p_1}{P_s}\log p_1 + \frac{p_2}{P_s}\log P_s - \frac{p_2}{P_s}\log p_2\right] \\ &= P_s\left[\frac{p_1}{P_s}\log\frac{P_s}{p_1} + \frac{p_2}{P_s}\log\frac{P_s}{p_2}\right] \\ &= H_s\,P_s\end{aligned}$$

where $P_s = p_1 + p_2$, $H_s = \left[\frac{p_1}{P_s}\log\frac{P_s}{p_1} + \frac{p_2}{P_s}\log\frac{P_s}{p_2}\right]$. ■

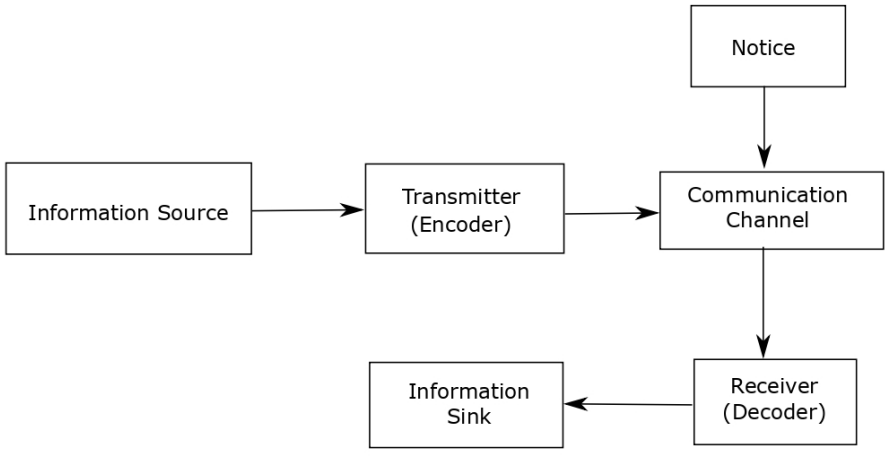
Unit 6

Course Structure

- Coding Theory
- Expected or average length of a code
- Uniquely decodable code

6.1 Introduction

Coding theory is the study of the method for efficient transfer of information from source; the physical medium through which the information transmitted for the channel, the telephone line and atmosphere are examples of channel. The undesirable disturbances are called noises. The following diagram provides a rough idea of the general information system:



Definition 6.1.1. Code: Let X be a random variable with range $S = \{x_1, x_2, \ldots, x_q\}$ and let $\mathfrak{D}$ be the D-ary alphabet, i.e., the set of all finite strings of symbols $\{0, 1, 2, \ldots, D-1\}$. A mapping $C : S \to \mathfrak{D}$ will be

called a code for the random variable X and S is called the source alphabet and $\mathfrak{D}$ is called the code alphabet.

If $x_i \in S$, then $C(x_i)$ is called codeword. Corresponding to x_i, the number of symbols in codeword $C(x_i)$ is called the length of the codeword and it is denoted by $l(x_i)$.

Example 6.1.2. Let X be a random variable with range $S = \{x_1, x_2, x_3, x_4\}$, $\mathfrak{D} = \{0, 1\}$ be the code alphabet. Define $C : S \to \mathfrak{D}$ as follows

$$x_1 \to 0, \quad x_2 \to 00, \quad x_3 \to 01, \quad x_4 \to 11$$

Then C is a code for the random variable X.

Definition 6.1.3. A code with code alphabet $\mathfrak{D} = \{0, 1\}$ is called a binary code. A code with code alphabet $\mathfrak{D} = \{0, 1, 2\}$ is called a ternary code.

Definition 6.1.4. A code C is said to be non-singular code if the mapping C is one-to-one, i.e., if $C(x_i) \neq C(x_j)$ for $x_i \neq x_j$. Clearly the code C in Example 6.1.2 is a non-singular code.

Definition 6.1.5. Extension of code: Let X be a random variable with range $S = \{x_1, x_2, \ldots, x_q\}$ and $\mathfrak{D} = \{0, 1, 2, \ldots, D-1\}$ as the code alphabet and C be a code for the random variable X. The n-th extension of C is a mapping $C^* : S^n (= S \times S \times \ldots \times S(n \text{ times})) \to \mathfrak{D}$ defined by

$$C^*(x_{i1}, x_{i2}, \ldots, x_{in}) = C(x_{i1})\, C(x_{i2})\, \ldots C(x_{in})$$

Example 6.1.6. Let X be a random variable with range $S = \{x_1, x_2, x_3, x_4\}$, $\mathfrak{D} = \{0, 1\}$ as the code alphabet and $C : S \to \mathfrak{D}$ be a code defined by

$$x_1 \to 0, \quad x_2 \to 00, \quad x_3 \to 01, \quad x_4 \to 11$$

Then the 2^{nd} extension of the above code C is given by

$$\begin{array}{llll} x_1x_1 \to 00, & x_1x_2 \to 000, & x_1x_3 \to 001, & x_1x_4 \to 001, \\ x_2x_1 \to 000, & x_2x_2 \to 0000, & x_2x_3 \to 0001, & x_2x_4 \to 0011, \\ x_3x_1 \to 010, & x_3x_2 \to 0100, & x_3x_3 \to 0101, & x_3x_4 \to 0111, \\ x_4x_1 \to 110, & x_4x_2 \to 1100, & x_4x_3 \to 1101, & x_4x_4 \to 1111. \end{array}$$

The 3^{rd} extension is

$$x_1x_2x_3 \to 000001; \quad x_1x_2x_4 \to 00011, \quad \ldots \quad \text{so on.}$$

6.1.1 Expected or average length of a code

Let X be a random variable with range $S = \{x_1, x_2, \ldots, x_q\}$ and p.m.f $p(x)$. Let $\mathfrak{D} = \{0, 1, 2, \ldots, D-1\}$ be the code alphabet. Then the expected length of the code C for the random variable X is denoted by $L(C)$ and is defined by

$$L(C) = \sum_{i=1}^{q} p(x_i) l(x_i) = \sum_{i=1}^{q} p_i l_i$$

6.1.2 Uniquely decodable (separable) code

A code is said to be uniquely decodable if all its extensions including itself are non-singular. For example, the code C in Example 6.1.6 is non-singular but its second extension is not singular. So it is not uniquely decodable ($\because x_1x_2 \neq x_2x_1$ but $C(x_1, x_2) = C(x_2, x_1) = 000$).

Examples of uniquely decodable codes are given below:

$$(a)\ \ x_1 \to 0, \quad x_2 \to 10, \quad x_3 \to 110, \quad x_4 \to 111$$
$$(b)\ \ x_1 \to 0, \quad x_2 \to 01, \quad x_3 \to 011, \quad x_4 \to 0111$$

Example 6.1.7. Let X be a random variable with range $S = \{x_1, x_2, x_3, x_4\}$ and code alphabet $\mathfrak{D} = \{0, 1\}$ with p.m.f $p(x)$ defined by

$$p(x_1) = \frac{1}{2},\ \ p(x_2) = \frac{1}{4},\ \ p(x_3) = \frac{1}{8} = p(x_4)$$

Let the code C be defined as follows:

$$x_1 \to 0, \quad x_2 \to 10, \quad x_3 \to 110, \quad x_4 \to 111$$
$$\therefore l(x_1) = 0, \quad l(x_2) = 2, \quad l(x_3) = 3, \quad l(x_4) = 3$$
$$\therefore \text{Expected length of } C,\ \ L(C) = 0 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{8} + 3 \cdot \frac{1}{8} = 1.25$$

Definition 6.1.8. Prefix: Let $i_1i_2 \ldots i_m$ be a codeword for some code C. Then $i_1i_2 \ldots i_\nu,\ \ \nu \leq m$ is called the prefix of the codeword $i_1i_2 \ldots i_m$. From definition it follows that every codeword is a prefix of itself.

Definition 6.1.9. Prefix code or instantaneous code: This is a code in which no codeword is a prefix of any other codeword. For example, the code in Example 6.1.7 is an instantaneous code whereas the code in Example 6.1.2 is not an instantaneous code. Another example of instantaneous code is the code defined by

$$x_1 \to 00, \quad x_2 \to 01, \quad x_3 \to 10, \quad x_4 \to 110$$

Theorem 6.1.10. An instantaneous code is uniquely decodable.

Proof. Let $S = \{x_1, x_2, \ldots, x_q\}$ be the source alphabet and $\mathfrak{D} = \{0, 1, 2, \ldots, D-1\}$ be the code alphabet for a random variable X.

Let $C : S \to \mathfrak{D}$ be an instantaneous code of the random variable X. The codewords are $C(x_1),\ C(x_2),\ \ldots,\ C(x_q)$. Since no codeword is a prefix of any other codeword, we have $C(x_i) \neq C(x_j)$ for $x_i \neq x_j$.

So C is one-to-one. Assuming C is not uniquely decodable, then there is a positive integer $n > 1$ such that $2^{nd},\ 3^{rd},\ \ldots,\ (n+1)^{th}$ extension of C are one to one. But the n^{th} extension is not one-to-one.

So, there are two elements

$$x = x_{i_1}x_{i_2} \ldots x_{i_n} \ \text{ and } \ y = y_{\nu_1}y_{\nu_2} \ldots y_{\nu_n} \ \text{ in } \ S \ \text{ such that } \ x \neq y \tag{6.1.1}$$

But

$$C^n(x) = C^n(y) \tag{6.1.2}$$

Write $x' = x_{i_2}x_{i_3} \ldots x_{i_n}$ and $y' = y_{\nu_2}y_{\nu_3} \ldots y_{\nu_n}$, then

$$x = x_{i_1}x', \quad y = y_{\nu_1}y'$$

$$\therefore \text{ We have } C^n(x) = C(x_{i_1})C(x_{i_2})\ldots C(x_{i_n}) = C(x_{i_1})C^{n-1}(x') \tag{6.1.3}$$

$$\text{Similarly, } C^n(y) = C(y_{\nu_1})C^{n-1}(y')$$

Without loss of generality, we may suppose that

$$l(x_{i_1}) \leq l(y_{\nu_1}) \tag{6.1.4}$$

where $l(x_{i_1})$ is the length of the codeword $C(x_{i_1})$ and $l(y_{\nu_1})$ be that of $C(y_{\mu_1})$. From (6.1.2), and (6.1.3) (6.1.4), it follows that the codeword $C(x_{i_1})$ is a prefix of the codeword $C(y_{\nu_1})$. Since C is an instantaneous code, it follows that

$$x_{i_1} = y_{\nu_1} \Rightarrow C^{n-1}(x') = C^{n-1}(y')$$

Since C^{n-1} is one-to-one, we have $x' = y'$. So, we have $x = y$ $[\because x_{i_1} = y_{\nu_1}]$ which contradicts (6.1.1).

Hence C is uniquely decodable code. □

Theorem 6.1.11. Kraft inequality for instantaneous code: Let $S = \{x_1, x_2, \ldots, x_q\}$ be the source alphabet and $\mathfrak{D} = \{0, 1, 2, \ldots, D-1\}$ be a code alphabet for a random variable X. Then a necessary and sufficient condition for the existence of an instantaneous code for the random variable X with codeword lengths $l_1, l_2, \ldots, l_q$ formed by the elements of $\mathfrak{D}$ is that

$$\sum_{i=1}^{q} D^{-l_i} \leq 1.$$

Proof. We first show that the condition is sufficient assuming that we have given codeword lengths $l_1, l_2, \ldots, l_q$ satisfying the condition

$$\sum_{i=1}^{q} D^{-l_i} \leq 1. \tag{6.1.5}$$

We show that there exists an instantaneous code for the random variable X with these codeword lengths. The lengths $l_1, l_2, \ldots, l_q$ may or may not be distinct. We shall find it useful to consider all codewords of the same length at a time.

Let $l = \max\{l_1, l_2, \ldots, l_q\}$. We denote by n_1, the number of codewords of length 1; by n_2, the number of codewords of length 2, and so on.

$$\therefore n_1 + n_2 + \ldots + n_l = q. \tag{6.1.6}$$

The inequality (6.1.5) may be written as

$$\sum_{i=1}^{l} n_i D^{-i} \leq 1 \tag{6.1.7}$$

Multiplying (6.1.7) by D^l, we have

$$\sum_{i=1}^{l} n_i D^{l-i} \leq D^l$$

$$n_l \leq D^l - n_1 D^{l-1} - n_2 D^{l-2} - \ldots - n_{l-1} D \tag{6.1.8}$$

From (6.1.8), we have,

$$n_{l-1} \leq D^{l-1} - n_1 D^{l-2} - n_2 D^{l-3} - \ldots - n_{l-2} D \tag{6.1.9}$$

Proceeding in this way we obtain,

$$\begin{aligned} n_{l-2} &\leq D^{l-2} - n_1 D^{l-3} - n_2 D^{l-4} - \ldots - n_{l-3} D \\ \cdots &\quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \\ n_3 &\leq D^3 - n_1 D^2 - n_2 D \\ n_2 &\leq D^2 - n_1 D \end{aligned} \tag{6.1.10}$$

We form n_1 codewords of length 1. Then there are $(D - n_1)$ unused codewords of length 1 which may be used as prefixes. By adding one symbol to the end of these permissible prefixes we may form as many as $(D - n_1)D = D^2 - n_1 D$ codewords of length 2. The inequalities (6.1.10) assures that we need no more than these number of (i.e., $D^2 - n_1 D$) codewords of length 2. As before, we chose n_2 codewords arbitrarily from $(D^2 - n_1 D)$ choices and we are left with $(D^2 - n_1 D - n_2)$ unused prefixes of length 2 with which we may form $(D^2 - n_1 D - n_2)D = D^3 - n_1 D^2 - n_2 D$ codewords of length 3. We select arbitrarily n_3 codewords from them and left with $D^3 - n_1 D^2 - n_2 D - n_3$ unused prefixes of length 3.

Continuing this process we obtain a code in which no codeword is prefix of any other codeword. So the code constructed is an instantaneous code.

We now show that the condition is necessary. Suppose that the codewords $C(x_1),\ C(x_2), \ldots, C(x_q)$ of lengths $l_1, l_2, \ldots, l_q$ for an instantaneous code for a random variable X.

Let $l = \max\{l_1, l_2, \ldots, l_q\}$ and let $n_i (i = 1, 2, \ldots, l)$ denote the number of codewords of length i.

There are all together D codewords of length 1 of which only n_1 codewords have been used. So $(D - n_1)$ codewords of length 1 are left unused. By adding one symbol to the end of these $(D - n_1)$ permissible prefixes we may form as $(D - n_1)D = D^2 - n_1 D$ codewords of length 2. Of these $(D^2 - n_1 D)$ codewords of length 2, n_2 are used.

$$\therefore n_1 \leq D, \quad n_2 \leq D^2 - n_1 D$$

Similarly,

$$\begin{aligned} & n_3 \leq D^3 - n_1 D^2 - n_2 D \\ & n_4 \leq D^4 - n_1 D^3 - n_2 D^2 - n_3 D \\ & \cdots \quad \cdots \quad \cdots \quad \cdots \\ & n_l \leq D^l - n_1 D^{l-1} - n_2 D^{l-2} - \ldots - n_{l-1} D \\ \Rightarrow\ & n_l + n_{l-1} D + n_{l-2} D^2 + \ldots + n_1 D^{l-1} \leq D^l \\ \Rightarrow\ & \sum_{i=1}^{l} n_i D^{-i} \leq 1 \\ \Rightarrow\ & \sum_{i=1}^{q} D^{-l_i} \leq 1 \end{aligned}$$

□

Definition 6.1.12. Optimal code: An instantaneous code is said to be optimal if the expected length of the code is less than or equal to the expected length of all other instantaneous codes for the same source alphabet and the same code alphabet.

Theorem 6.1.13. Let $S = \{x_1, x_2, \ldots, x_q\}$ be the source alphabet and $\mathfrak{D} = \{0, 1, 2, \ldots, D-1\}$ be the code alphabet for a random variable X. Then the expected length L^* of an optimal instantaneous code for the random variable X is given by

$$L^* = \frac{H(X)}{\log D},$$

where $H(X)$ is the entropy of the random variable X.

Theorem 6.1.14. Let $S = \{x_1, x_2, \ldots, x_q\}$ be the source alphabet and $\mathfrak{D} = \{0, 1, 2, \ldots, D-1\}$ be the code alphabet for the random variable X with p.m.f $p(X)$. Then the expected length $L(C)$ of any instantaneous code C for X satisfies the inequality

$$L(C) \geq \frac{H(X)}{\log D}.$$

Proof. Let $p_i = p(x_i) = P(X = x_i)$ and $l_i = l(x_i)$. Since C is an instantaneous code, by Kraft inequality

$$\sum_{i=1}^{q} D^{-l_i} \leq 1. \tag{6.1.11}$$

For any $x > 0$, we have

$$\log x \leq x - 1. \tag{6.1.12}$$

Write $\mu = \sum_{i=1}^{q} D^{-l_i}$, $0 < \mu \leq 1$. Taking $x = \frac{D^{-l_i}}{\mu p_i}$ in inequality (6.1.12), we get

$$\log \frac{D^{-l_i}}{\mu p_i} \leq \frac{D^{-l_i}}{\mu p_i} - 1$$

$$-l_i \log D - \log \mu - \log p_i \leq \frac{D^{-l_i}}{\mu p_i} - 1$$

Multiplying by p_i and taking sum we get

$$-\sum_{i=1}^{q} p_i l_i \log D - \sum_{i=1}^{q} p_i \log \mu - \sum_{i=1}^{q} p_i \log p_i \leq \sum_{i=1}^{q} \frac{D^{-l_i}}{\mu} - \sum_{i=1}^{q} p_i$$

$$\Rightarrow -L(C) \log D - \log \mu + H(X) \leq \frac{1}{\mu} \cdot \mu - 1$$

$$\Rightarrow H(X) - L(C) \log D \leq \log \mu \qquad \left[\because \mu = \sum_{i=1}^{q} D^{-l_i},\ \sum_{i=1}^{q} p_i = 1,\ L(C) = \sum_{i=1}^{q} p_i l_i\right]$$

$$\Rightarrow H(X) - L(C) \log D \leq 0 \quad [\because 0 < \mu \leq 1,\ \log \mu \leq 0]$$

$$\Rightarrow L(C) \geq \frac{H(X)}{\log D}$$

□

Theorem 6.1.15. Let L^* be the expected length of an instantaneous optimal code for the random variable X with code alphabet $\mathfrak{D} = \{0, 1, 2, \ldots, D-1\}$. Then

$$\frac{H(X)}{\log D} \leq L^* \leq \frac{H(X)}{\log D} + 1,$$

where $H(X)$ is the entropy function of the random variable X.

Proof. Let $S = \{x_1, x_2, \ldots, x_q\}$ be the source alphabet and $\mathfrak{D} = \{0, 1, \ldots, D-1\}$ be the code alphabet of the random variable X with p.m.f $p(x)$.

Let us define $p_i = p(x_i) = P(X = x_i),\ l_i = l(x_i)$. Now, L^* be the minimum value of $\sum_{i=1}^{q} p_i l_i$ subject to the constraint

$$\sum_{i=1}^{q} D^{-l_i} \leq 1 \tag{6.1.13}$$

We neglect the integer constraint on $l_1, l_2, \ldots, l_q$ and assume the inequality (6.1.13) hold.

The choice of the codeword length $l_i = -\dfrac{\log p_i}{\log D},\ \ (i = 1, 2, \ldots, q)$ gives

$$L = \sum_{i=1}^{q} p_i l_i = \sum_{i=1}^{q} \frac{-p_i \log p_i}{\log D} = \frac{H(X)}{\log D}.$$

$\because -\dfrac{\log p_i}{\log D}$ may not equal to an integer

Therefore, we round it upto the even integer. So we take $l_i = \left[-\frac{\log p_i}{\log D}\right]$, where for any real $x > 0$, $[x]$ denote the greatest positive integer not greater than x. Then

$$-\frac{\log p_i}{\log D} \leq l_i \leq -\frac{\log p_i}{\log D} + 1 \tag{6.1.14}$$

From (6.1.14), we have

$$\begin{aligned} &-\log p_i \leq l_i \log D \\ &\Rightarrow \log p_i \geq \log D^{-l_i} \\ &\Rightarrow p_i \geq D^{-l_i} \end{aligned}$$

Therefore,

$$\sum_{i=1}^{q} D^{-l_i} \leq \sum_{i=1}^{q} p_i = 1$$

Thus the codeword lengths $l_1, l_2, \ldots, l_q$ satisfies the Kraft inequality.
Hence the code with word lengths $l_1, l_2, \ldots, l_q$ as chosen is an instantaneous code.

Multiplying (6.1.14) by p_i and taking sum, we get

$$\begin{aligned} &-\sum_{i=1}^{q} \frac{p_i \log p_i}{\log D} \leq \sum_{i=1}^{q} p_i l_i \leq -\sum_{i=1}^{q} \frac{p_i \log p_i}{\log D} + \sum_{i=1}^{q} p_i \\ &\Rightarrow \frac{H(X)}{\log D} \leq L \leq \frac{H(X)}{\log D} + 1 \end{aligned} \tag{6.1.15}$$

Since L^* is the expected length of the optimal code, hence we have

$$\frac{H(X)}{\log D} \leq L^* \leq L \tag{6.1.16}$$

From (6.1.15) and (6.1.16) the result follows.

$$\therefore \frac{H(X)}{\log D} \leq L^* \leq \frac{H(X)}{\log D} + 1.$$

□

Example 6.1.16. Let $S = \{x_1, x_2, \ldots, x_q\}$ be the source alphabet and $\mathfrak{D} = \{0, 1, \ldots, D-1\}$ be the code alphabet of the random variable X with p.m.f $p(X_i) = D^{-\alpha_i}$, where $\alpha_1, \alpha_2, \ldots, \alpha_q$ are positive integers. Show that any code $C : S \to \mathfrak{D}$ for X with codeword lengths $\alpha_1, \alpha_2, \ldots, \alpha_q$ is an instantaneous optimal code.

Solution. Let C be any code for the random variable X with codeword lengths $l_i = l(x_i) = \alpha_i, \quad i = 1, 2, \ldots, q$. Then

$$\sum_{i=1}^{q} D^{-l_i} = \sum_{i=1}^{q} D^{-\alpha_i} = \sum_{i=1}^{q} p_i = 1$$

Thus the codeword lengths $l_1, l_2, \ldots, l_q$ of the code C satisfy Kraft inequality. Hence C is an instantaneous code. Again

$$\begin{aligned}
&p_i = D^{-\alpha_i} = D^{-l_i} \\
&\therefore \log p_i = -l_i \log D \\
&\Rightarrow -\sum_{i=1}^{q} p_i \log p_i = \sum_{i=1}^{q} l_i p_i \log D \\
&\Rightarrow H(X) = L(C) \log D \\
&\Rightarrow L(C) = \frac{H(X)}{\log D}
\end{aligned}$$

Therefore, the expected length $L(C)$ of the code C is minimum. Hence C is an instantaneous optimal code. ■

Definition 6.1.17. Efficiency of a code: Let C be a uniquely decodable D-ary code for the random variable X and $L(C)$ be its expected length. Then the efficiency η of the code C is defined by

$$\eta = \frac{H(X)}{L(C) \log D}$$

Redundancy of a code = $\beta = 1 - \eta$.

Theorem 6.1.18. Let C^* be a code of the random variable X of the following distribution

X :	x_1	x_2	...	x_n
p_i :	p_1	p_1	...	p_n

where $p_1 \geq p_2 \geq \ldots \geq p_n$. If $L(C^*) \leq L(C)$ for any code C of X, then $l_1^* \leq l_2^* \leq \ldots \leq l_n^*$ where l_i^* is the length of the code $C^*(x_i)$.

If $p_i = p_{i+1}$, it is assumed that $l_i^* \leq l_i^* + 1$.

Proof. Let $E = \{1, 2, \ldots, n\}$. We take any two elements i and j of E with $i < j$. Denote by α, the permutation of the set E such that $\alpha(i) = j$ and $\alpha(j) = i$ but all other elements of E remain unchanged.

Let C be a code of the random variable X such that $l_k = l^*_{\alpha(k)}$, where l_k is the length of the codeword $C(x_k)$. Then

$$\begin{aligned} l_i &= l^*_{\alpha(i)} = l^*_j \quad [\because \alpha(i) = j] \\ \text{and } l_j &= l^*_{\alpha(j)} = l^*_i \quad [\because \alpha(j) = i] \end{aligned}$$

and $l_k = l^*_k$ for all other elements k of E.

$$\begin{aligned} L(C) - L^*(C) &= p_i l_i + p_j l_j - p_i l^*_i - p_j l^*_j \\ &= p_i l^*_j + p_j l^*_i - p_i l^*_i - p_j l^*_j \\ &= (p_i - p_j)(l^*_j - l^*_i) \quad [\because l_i = l^*_j;\ l_j = l^*_i] \end{aligned}$$

Since $p_i \geq p_j$, we must have

$$\begin{aligned} l^*_i &\leq l^*_j. \\ \therefore l^*_1 &\leq l^*_2 \leq \ldots \leq l^*_n \end{aligned}$$

Hence the result follows. □

Unit 7

Course Structure

- Shannon-Fano Encoding Procedure for Binary code
- Construction of Haffman binary code
- Construction of Haffman D-ary code

7.1 Shannon-Fano Encoding Procedure for Binary code:

Let $S = \{x_1, x_2, \ldots, x_q\}$ be the source alphabet and $\mathfrak{D} = \{0, 1\}$ be the code alphabet of a random variable X with p.m.f $p(x)$. We shall give here an encoding procedure of assigning an efficient uniquely decodable binary code for the random variable X. This is known as Shannon-Fano encoding procedure.

Let $p_i = p(x_i) = P(X = x_i),\ i = 1, 2, \ldots, q.$

The two necessary requirements are

(i) No complete codeword can be prefix of some other codeword.

(ii) The binary digit in each codeword appeared independent with equal probabilities.

The encoding procedure follows the following steps.

Step 1: We arrange source symbols in descending order of their probabilities.

Step 2: Partition the set S of source symbols into two equiprobable groups S_0 and S_1 as

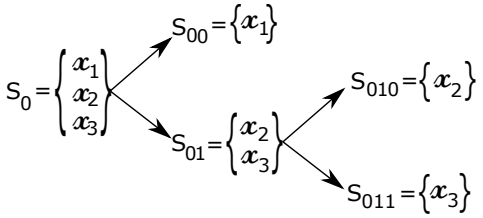
$$S_0 = \{x_1, x_2, \ldots, x_r\}, \quad S_1 = \{x_{r+1}, \ldots, x_q\}$$
$$\text{i.e., } P(S_0) \equiv P(S_1)$$

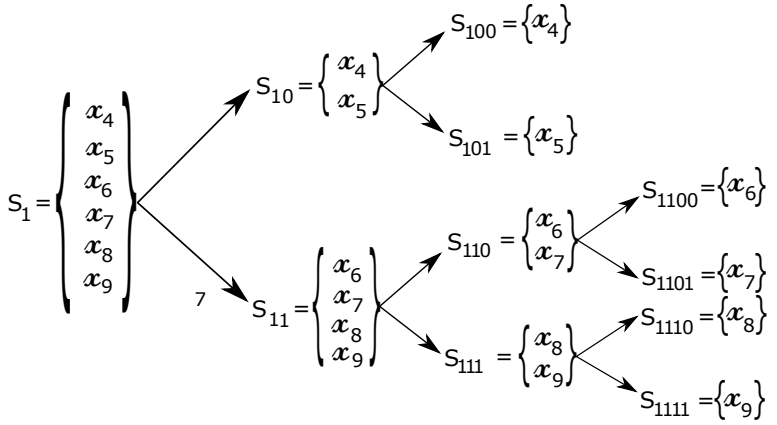
where $P(S_0) = p_1 + p_2 + \cdots + p_r$ and $P(S_1) = p_{r+1} + p_{r+2} + \cdots + p_q$.

Step 3: We further partition each of the subgroups S_0 and S_1 into two most equiprobable subgroups $S_{00},\ S_{01}$ and $S_{10},\ S_{11}$ respectively.

Step 4: We continue partitioning each of the resulting subgroups into two most equiprobable subgroups till each subgroup contain only one source symbol.

For example, let $S = \{x_1, x_2, \ldots, x_9\}$.





Therefore, the codes are

$$x_1 \to 00, \quad x_2 \to 010, \quad x_3 \to 011, \quad x_4 \to 100, \quad x_5 \to 101$$
$$x_6 \to 1100, \quad x_7 \to 1101, \quad x_8 \to 1110, \quad x_9 \to 1111$$

Clearly no codeword is a prefix of any other codeword. So it is an instantaneous code and hence it is uniquely decodable.

Advantages:

(1) Efficiency is nearly 100%.

(2) Expected length of the code is minimum.

(3) Entropy per digit of the encoded message is maximum.

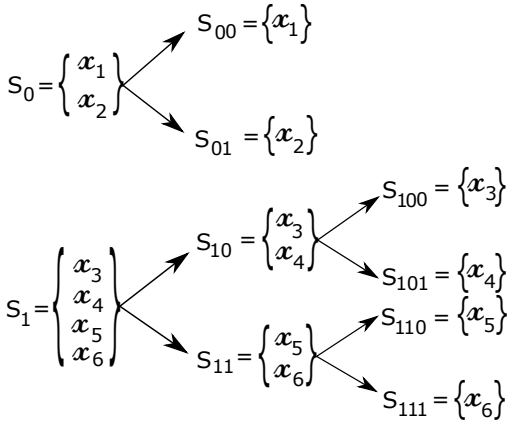
Example 7.1.1. Construct Shannon Fanno binary code for the random variable X with the following distribution.

Source symbols :	x_1	x_2	x_3	x_4	x_5	x_6
Probability :	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{12}$	$\frac{1}{12}$

Calculate the expected length and the efficiency of the code.

Solution. We have

$$p_1 + p_2 = \frac{1}{3} + \frac{1}{4} = \frac{7}{12}$$
$$p_3 + p_4 + p_5 + p_6 = \frac{1}{4} + \frac{1}{6} = \frac{5}{12}$$
$\therefore \frac{7}{12}$ and $\frac{5}{12}$ are close to each other
$\therefore$ We consider the equiprobable groups as follows.



So the Shannon-Fano binary code will be as follows:

$$x_1 \to 00, \quad x_2 \to 01, \quad x_3 \to 100, \quad x_4 \to 101, \quad x_5 \to 110, \quad x_6 \to 111$$

$$\begin{aligned} L(C) = \text{Expected length} &= 2 \cdot \frac{1}{3} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{8} + 3 \cdot \frac{1}{8} + 3 \cdot \frac{1}{12} + 3 \cdot \frac{1}{12} \\ &= \frac{2}{3} + \frac{1}{2} + \frac{3}{4} + \frac{1}{2} \\ &= 1 + \frac{2}{3} + \frac{3}{4} \\ &= \frac{12 + 8 + 9}{12} = \frac{29}{12} \text{ bits/symbol} \\ \text{Entropy, } H(X) &= -\sum_{i=1}^{6} p_i \log_2 p_i = 2.3758 \text{ bits} \\ \text{Efficiency, } \eta &= \frac{H(X)}{L(C) \log_2 2} = \frac{2.3758}{29/12} = 98.30\% \end{aligned}$$

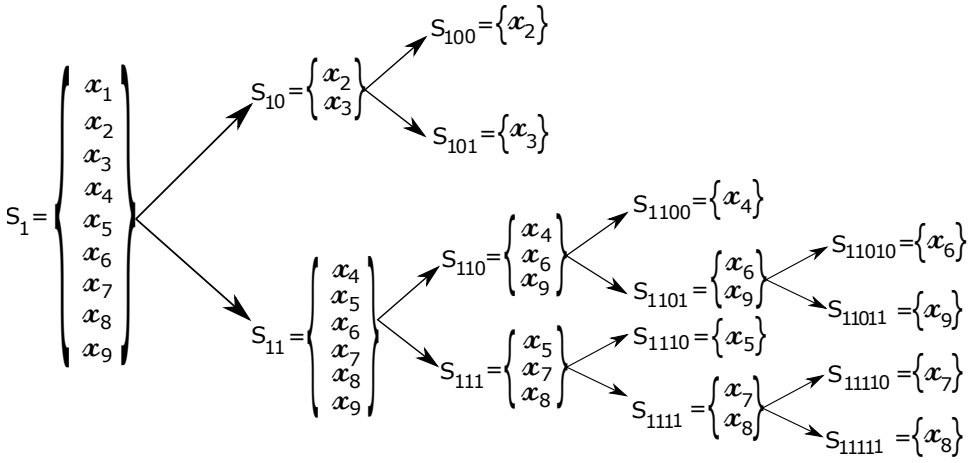
■

Similar Problems:

Example 7.1.2.

Source symbols :	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
Probability :	0.49	0.14	0.14	0.07	0.07	0.04	0.02	0.02	0.01

Solution. Here $p_1 = 0.49$ and $p_2 + \ldots + p_9 = 0.51$. Therefore, we take $S_0 = \{x_1\}$ and



So the code is

$$\begin{aligned} x_1 &\rightarrow 0 \\ x_2 &\rightarrow 100 \\ x_3 &\rightarrow 101 \\ x_4 &\rightarrow 1100 \\ x_5 &\rightarrow 1110 \\ x_6 &\rightarrow 11010 \\ x_7 &\rightarrow 11110 \\ x_8 &\rightarrow 11111 \\ x_9 &\rightarrow 11011 \end{aligned}$$

Therefore,

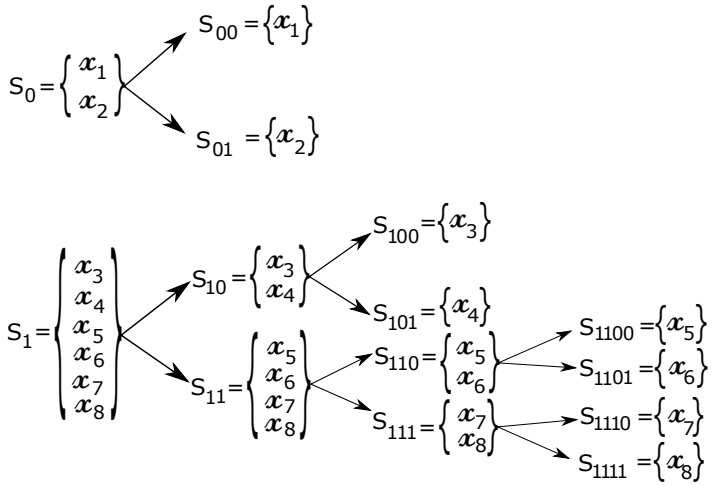
$$\begin{aligned} L(C) &= (1 \times 0.49) + (3 \times 0.14) + (3 \times 0.14) + (4 \times 0.07) + (4 \times 0.07) \\ &\quad +(5 \times 0.04) + (5 \times 0.02) + (5 \times 0.02) + (5 \times 0.01) \\ &= 2.34 \text{ bits/symbol.} \\ H(X) &= 2.3136 \text{ bits} \\ \therefore \eta &= \frac{2.3136}{2.34} = 38.87\% \end{aligned}$$

■

Example 7.1.3.

Source symbols :	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Probability :	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$

Solution. Here $p_1 + p_2 = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$ and $p_3 + p_4 + p_5 + p_6 + p_7 + p_8 = \frac{1}{8} + \frac{1}{8} + \frac{1}{16} + \frac{1}{16} + \frac{1}{16} + \frac{1}{16} = \frac{1}{2}$. Therefore, we take two equiprobable groups as:



So the code is

$$\begin{aligned} x_1 &\rightarrow 00 \\ x_2 &\rightarrow 01 \\ x_3 &\rightarrow 100 \\ x_4 &\rightarrow 101 \\ x_5 &\rightarrow 1100 \\ x_6 &\rightarrow 1101 \\ x_7 &\rightarrow 1110 \\ x_8 &\rightarrow 1111 \end{aligned}$$

Therefore,

$$\begin{aligned} L(C) &= \left(\frac{1}{4}\cdot 2\right)+\left(\frac{1}{4}\cdot 2\right)+\left(\frac{1}{8}\cdot 3\right)+\left(\frac{1}{16}\cdot 4\right)+\left(\frac{1}{16}\cdot 4\right)+\left(\frac{1}{16}\cdot 4\right)+\left(\frac{1}{16}\cdot 4\right) \\ &= 1+\frac{3}{4}+1 \\ &= \frac{11}{4}=2.75 \text{ bits/symbol.} \\ H(X) &= -\sum_{i=1}^{8} p_i \log p_i = \frac{11}{4}=2.75 \text{ bits} \end{aligned}$$

$$\therefore \text{Efficiency of the code} = \eta = \frac{H(X)}{L(C)\log_2 2} = \frac{2.75}{2.75} = 100\%$$

■

7.2 Construction of Haffman binary code

Let X be a random variable with the following distribution

X :	x_1	x_2	$\ldots$	x_n
Probability :	p_1	p_2	$\cdots$	p_n

Step 1: We arrange the source symbols $x_i's$ in decending order of their probabilities. Without loss of generality we may assume that $p_1 \geq p_2 \geq \ldots \geq p_n$. We thus have

$X:$	x_1	x_2	$\ldots$	x_n
Probability :	p_1	p_2	$\cdots$	p_n

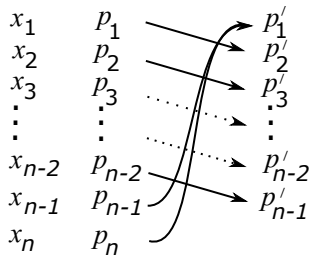
Step 2: We combine the last two symbols to form a new symbol. Then we arrange the source symbols in descending order of their probabilities. Let us suppose that

$$p_{n-1} + p_n \geq p_1.$$

We take

$$x'_1 = x_{n-1} + x_n, \quad x'_2 = x_1, \quad x'_3 = x_2, \quad x'_4 = x_3, \quad \ldots, \quad x'_{n-1} = x_{n-2}$$
$$p'_1 = p_{n-1} + p_n, \quad p'_2 = p_1, \quad p'_3 = p_2, \quad \ldots, \quad p'_{n-1} = p_{n-2}$$

This may be shown as follows:



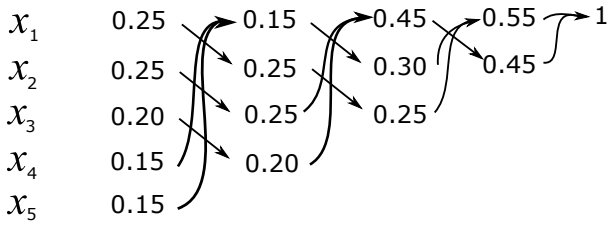
Step 3: Again we combine the last two symbols to form a new symbol and proceed as in Step 2.

Step 4: The process is continued until we reach a stage where we get only one symbol.

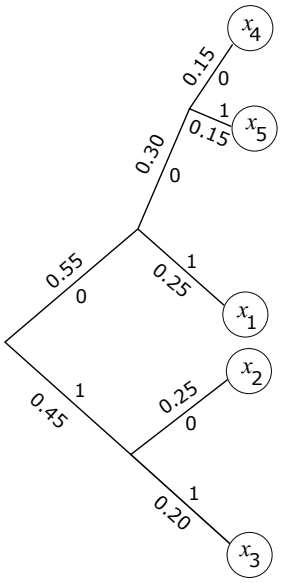
Example 7.2.1. Construct Haffman binary code for the random variable X whose distribution is given by

$X:$	x_1	x_2	x_3	x_4	x_5
Probability :	0.25	0.25	0.2	0.15	0.15

Solution. Consider the following scheme.



We arrange the above scheme as a tree in reverse order from which we can write down the corresponding Haffman binary code.



So the Haffman binary code is

$$\begin{aligned} x_1 &\rightarrow 01 \\ x_2 &\rightarrow 10 \\ x_3 &\rightarrow 11 \\ x_4 &\rightarrow 000 \\ x_5 &\rightarrow 001 \end{aligned}$$

■

7.3 Construction of Haffman D ary code (D>2)

Let the random variable X has the following distribution

$X:$	x_1	x_2	$\ldots$	x_q
Probability :	p_1	p_2	$\cdots$	p_q

Case 1: Let $(q - D)$ is divisible by $(D - 1)$.

Step 1: Arrange the symbols in descending order of their probabilities.

Step 2: We consider last D symbols to a single composite symbol whose probability is equal to the sum of the probabilities of the last D symbols.

Step 3: Repeat Step 1 and Step 2 on the resulting set of symbols until we reach a stage where we get composite symbol only.

Step 4: Following above stage carefully we construct a tree diagram from which codes are assigned for the symbols.

Case 2: If $(q - D)$ is not divisible by $(D - 1)$, then we add new *dummy symbols with zero probability* to make $(q^* - D)$ divisible by $(D - 1)$ where q^* is the number of symbols after addition of dummy symbols.

Now, we proceed as in Case 1. The codes for the dummy symbols are discarded.

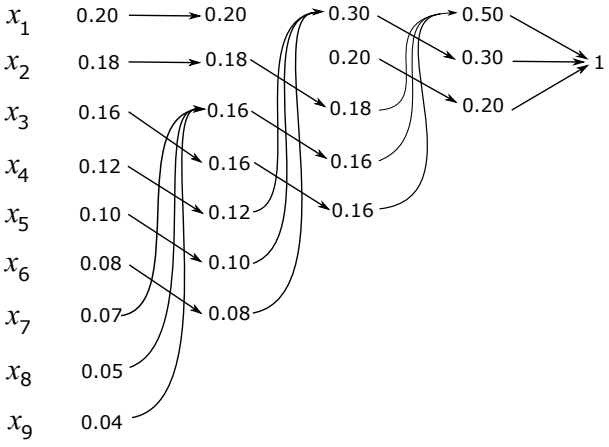
Example 7.3.1.

Source symbols :	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
Probability :	0.20	0.18	0.16	0.12	0.10	0.08	0.07	0.05	0.04

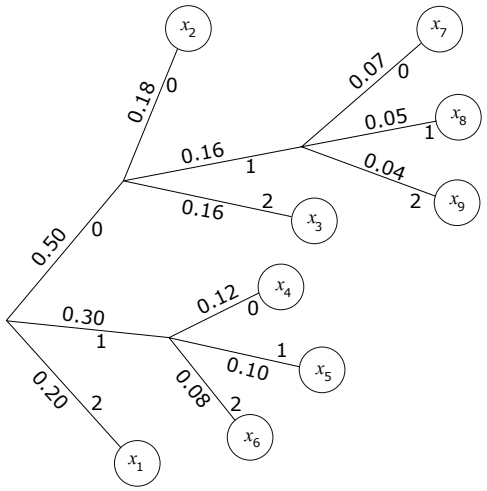
Construct a Haffman ternary code for X. Calculate the expected length and efficiency of the code.

Solution. Here $q = 9$, $\mathfrak{D} = \{0, 1, 2\}$, $D = 3$.

$$\therefore q - D = 9 - 3 = 6 \text{ is divisible by } 2 = 3 - 1 = D - 1$$



We arrange the above scheme as a tree in reverse order from which we can write down the corresponding Haffman binary code.



So the code is

$$\begin{aligned} x_1 &\rightarrow 2 \\ x_2 &\rightarrow 00 \\ x_3 &\rightarrow 02 \\ x_4 &\rightarrow 10 \\ x_5 &\rightarrow 11 \\ x_6 &\rightarrow 12 \\ x_7 &\rightarrow 010 \\ x_8 &\rightarrow 011 \\ x_9 &\rightarrow 012 \end{aligned}$$

Therefore,

$$\begin{aligned} L(C) &= (1 \times 0.20) + (2 \times 0.18) + (2 \times 0.16) + (2 \times 0.12) + (2 \times 0.10) \\ &\quad +(2 \times 0.08) + (3 \times 0.07) + (3 \times 0.05) + (3 \times 0.04) \\ &= 1.96 \text{ bits/symbol.} \\ H(X) &= -\sum_{i=1}^{9} p_i \log p_i = 2.99388 \text{ bits} \end{aligned}$$

$$\therefore \text{Efficiency of the code} = \eta = \frac{H(X)}{L(C)\log_2 3} = 0.9637 = 96.37\%$$

■

Example 7.3.2. Construct Haffman ternary code with the following distribution

Source symbols :	x_1	x_2	x_3	x_4	x_5	x_6
Probability :	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{12}$	$\frac{1}{12}$

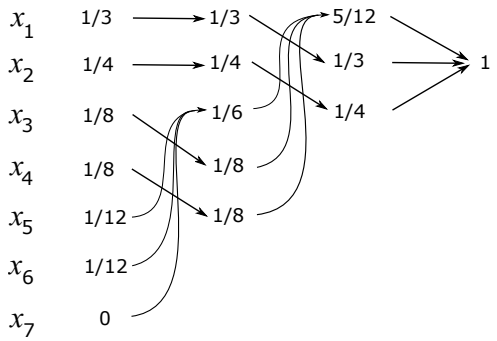
Calculate the expected length and its efficiency.

Solution. Here $q = 6,\ \mathfrak{D} = \{0, 1, 2\},\ D = 3.$

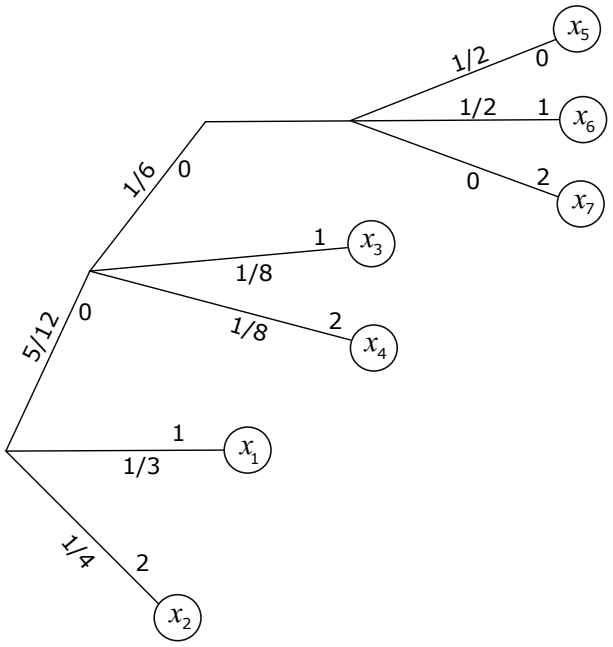
$$\therefore q - D = 6 - 3 = 3 \text{ which is not divisible by } 2. \tag{7.3.1}$$

Therefore, we introduce a dummy source alphabet x_7 with probability zero.

Now, $q^* = 7$, so $q^* - D = 4$ which is divisible by 2



We arrange the above scheme as a tree in reverse order from which we write down the Haffman ternary code.



So the code is

$$\begin{aligned} x_1 &\rightarrow 1 \\ x_2 &\rightarrow 2 \\ x_3 &\rightarrow 01 \\ x_4 &\rightarrow 02 \\ x_5 &\rightarrow 000 \\ x_6 &\rightarrow 001 \\ x_7 &\rightarrow 002 \text{ (discarded)}. \end{aligned}$$

Therefore,

$$\begin{aligned} L(C) &= \left(1 \times \frac{1}{3}\right) + \left(1 \times \frac{1}{4}\right) + \left(2 \times \frac{1}{8}\right) + \left(2 \times \frac{1}{8}\right) + \left(3 \times \frac{1}{12}\right) + \left(3 \times \frac{1}{12}\right) \\ &= \frac{19}{12} \text{ bits/symbol.} \\ H(X) &= -\sum_{i=1}^{6} p_i \log p_i = 1.4990 \text{ bits} \end{aligned}$$

$$\therefore \text{Efficiency of the code} = \eta = \frac{H(X)}{L(C)\log_2 3} = 0.9467 = 94.67\%$$

■

Example 7.3.3. Construct Shannon Fanno ternary code for the following distribution of the random variable X

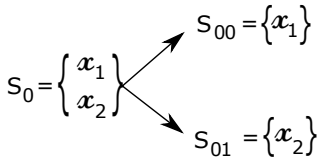
Source symbols :	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
Probability :	0.20	0.18	0.16	0.12	0.10	0.08	0.07	0.05	0.04

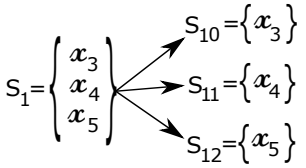
Hence calculate the expected length and efficiency of the code.

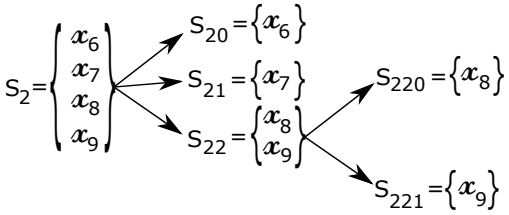
Solution: Here we see that

$$\begin{aligned} p_1 + p_2 &= 0.38 \\ p_3 + p_4 + p_5 &= 0.38 \\ p_6 + p_7 + p_8 + p_9 &= 0.24 \end{aligned}$$

So we take the three equiprobable groups as







Therefore, the Shannon Fanno ternary codes are obtained as

$$\begin{aligned} x_1 &\rightarrow 00 \\ x_2 &\rightarrow 01 \\ x_3 &\rightarrow 10 \\ x_4 &\rightarrow 11 \\ x_5 &\rightarrow 12 \\ x_6 &\rightarrow 20 \\ x_7 &\rightarrow 21 \\ x_8 &\rightarrow 220 \\ x_9 &\rightarrow 221 \end{aligned}$$

Therefore,

$$\begin{aligned} L(C) &= (2 \times 0.20) + (2 \times 0.18) + (2 \times 0.16) + (2 \times 0.12) + (2 \times 0.10) \\ &\quad +(2 \times 0.08) + (2 \times 0.07) + (3 \times 0.05) + (3 \times 0.04) \\ &= 2.09 \text{ bits/symbol.} \\ H(X) &= -\sum_{i=1}^{9} p_i \log p_i = 2.99388 \text{ bits} \end{aligned}$$

$$\therefore \text{Efficiency of the code} = \eta = \frac{H(X)}{L(C)\log_2 3} = 0.9038 = 90.38\%$$

Unit 8

Course Structure

- Error correcting codes
- Construction of linear codes
- Standard form of parity check matrix
- Hamming code, Cyclic code, BCH code

8.1 Error correcting codes

Let F_q be a finite field with q elements and let $n(>1)$ be a given positive integer. We denote by $V_n(F_q)$, the set of all n-tuples $x = (x_1, x_2, \ldots, x_n)$ with $x_i \in F_q,\ i = 1, 2, \ldots, n$. For any $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n) \in V_n(F_q)$ and $\lambda \in F_q$, define

$$\begin{aligned} x + y &= (x_1 + y_1,\ x_2 + y_2,\ \ldots,\ x_n + y_n) \\ \lambda x &= (\lambda x_1, \lambda x_2, \ldots, \lambda x_n) \end{aligned}$$

Then

$$x + y,\ \ \lambda x \in V_n(F_q).$$

It is easy to see that $V_n(F_q)$ is a vector space over the field F_q.

Theorem 8.1.1. For any $x, y \in V_n(F_q)$, if we define $d(x, y) =$ number of i's with $x_i \neq y_i$, then d is a metric on $V_n(F_q)$.

Proof. From definition it is clear that for $x, y \in V_n(F_q)$

$$\begin{aligned} &(i)\ \ d(x, y) \geq 0 \ \text{ and } \ d(x, y) = 0 \ \text{ iff } \ x = y, \\ &(ii)\ \ d(x, y) = d(y, x). \end{aligned}$$

Now, let $x, y, z \in V_n(F_q)$. Then we show that

$$d(x, y) \leq d(x, z) + d(z, y) \tag{8.1.1}$$

If $x = y$, then $d(x, y) = 0$ and so (8.1.1) holds.
If $x = z$, then $d(x, z) = 0$ and $d(z, y) = d(x, y)$.
Hence (8.1.1) holds.
Similarly if $y = z$, then (8.1.1) also holds.
Suppose $x \neq y,\ x \neq z,\ z \neq y$.
Let $E = \{i :\ x_i \neq y_i\}$, $A = \{i : x_i \neq z_i\}$ and $B = \{i : y_i \neq z_i\}$.
$|E|$ denote the number of elements in E. Then $d(x, y) = |E|,\ d(x, z) = |A|,\ d(y, z) = |B|$.
Let $i \in E$. If $x_i \neq z_i$, then $i \in A$ also. Suppose that $x_i = z_i$. Since $x_i \neq y_i$, we have $z_i \neq y_i$. So $i \in B$.

$$\therefore i \in A \cup B.$$

This gives $E \subset A \cup B$. Therefore, $|E| \leq |A| + |B|$.

$$\therefore d(x, y) \leq d(x, z) + d(z, y)$$

Thus d is a metric on $V_n(F_q)$. □

Definition 8.1.2. q-ary code of length n: A non empty subset C of $V_n(F_q)$ is called a q-ary code of length n and members of C are called codeword. If $q = 2$, the corresponding code is called binary code and so on.

Definition 8.1.3. Weight of a codeword: An element x in $V_n(F_q)$ is a codeword. The weight of the codeword x, denoted by $w(x)$ and is defined by

$$w(x) = \text{number of } i's \text{ with } x_i \neq 0.$$
$$\text{e.g., } x = 1\,2\,0\,1\,0\,0\,\ldots\,0. \text{ Then, } w(x) = 3.$$

Definition 8.1.4. Linear code: A linear subspace C of $V_n(F_q)$ is called a linear code of length l over the field F_q and the dimension k of the subspace C is called the dimension of the code C. It is also called an (n, k) linear code over the field F_q.

Definition 8.1.5. Minimum distance of the code: Let C be a code in $V_n(F_q)$. The minimum distance $\delta(C)$ of the code C is defined by

$$\delta(C) = \min\{d(x, y) :\ x, y \in C \text{ and } x \neq y\}$$

Definition 8.1.6. Generator matrix: Let C be an (n, k) linear code over the field F_q with q elements. A $k \times n$ matrix G with entries from the field F_q is said to be the generator matrix of code C if the row space of the matrix G is the same as the subspace C. We also say that the matrix G generates the code C. Since the dimension of C is k, the dimension of the rowspace of G is k which implies that the row vectors of G are linearly independent and so they form a basis of C.

Definition 8.1.7. Parity check matrix: Let C be an (n, k) linear code over the field F_q with q elements. An $(n - k) \times n$ matrix H with entries from the field F_q is called a parity check matrix of code C iff $Hx = 0$ for all $x \in C$.

The matrix H also generates an $(n, n - k)$ linear code over F_q which is denoted by $C^{\perp}$ and is called the dual space of C.

$$\therefore \dim(C) + \dim(C^{\perp}) = n \quad \text{and} \quad \text{rank}(H) = n - k.$$

8.2 Construction of linear codes

• **By using generator matrix:** Let G be a $k \times n$ $(k < n)$ generator matrix with entries from F_q with q elements and rank$(G) = k$.

Let C denote the row space of the matrix G. Then C is and (n, k) linear code denoted by $\alpha_1\alpha_2 \ldots \alpha_k$, the row vectors of G.

Let $a = (a_1 \;\; a_2 \;\; \ldots \;\; a_k) \in V_k(F_q)$. Then

$$u = aG = a_1\alpha_1 + a_2\alpha_2 + \ldots + a_k\alpha_k \in C$$

Thus every u in C is of the form $u = aG$ where $a \in V_k(F_q)$.

Example 8.2.1. Find the codewords determined by the binary generator matrix

$$G = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

Solution. G is a binary generation matrix with 5 columns. Also, it is clear that rank(G)=3. The linear code C generated by G is given by

$$C = \{x : x = aG \text{ and } a \in V_3(F_2)\}$$

The vector $a = (a_1 \;\; a_2 \;\; a_3)$ may be considered in $2^3 = 8$ ways , namely, $(0\;\;0\;\;0)$, $(0\;\;0\;\;1)$, $(0\;\;1\;\;0)$, $(1\;\;0\;\;0)$, $(0\;\;1\;\;1)$, $(1\;\;0\;\;1)$, $(1\;\;1\;\;0)$, $(1\;\;1\;\;1)$.

$$\therefore \begin{pmatrix} 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

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

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

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

$$\begin{pmatrix}0 & 1 & 1\end{pmatrix}\begin{pmatrix}1 & 0 & 0 & 1 & 1\\ 0 & 1 & 0 & 0 & 1\\ 0 & 0 & 1 & 1 & 1\end{pmatrix} = \begin{pmatrix}0 & 1 & 1 & 1 & 0\end{pmatrix} \quad [\because 1+1=0]$$

$$\begin{pmatrix}1 & 0 & 1\end{pmatrix}\begin{pmatrix}1 & 0 & 0 & 1 & 1\\ 0 & 1 & 0 & 0 & 1\\ 0 & 0 & 1 & 1 & 1\end{pmatrix} = \begin{pmatrix}1 & 0 & 1 & 0 & 0\end{pmatrix} \quad [\because 1+1=0]$$

$$\begin{pmatrix}1 & 1 & 0\end{pmatrix}\begin{pmatrix}1 & 0 & 0 & 1 & 1\\ 0 & 1 & 0 & 0 & 1\\ 0 & 0 & 1 & 1 & 1\end{pmatrix} = \begin{pmatrix}1 & 1 & 0 & 1 & 0\end{pmatrix} \quad [\because 1+1=0]$$

$$\begin{pmatrix}1 & 1 & 1\end{pmatrix}\begin{pmatrix}1 & 0 & 0 & 1 & 1\\ 0 & 1 & 0 & 0 & 1\\ 0 & 0 & 1 & 1 & 1\end{pmatrix} = \begin{pmatrix}1 & 1 & 1 & 0 & 1\end{pmatrix} \quad [\because 1+1=0;\ 1+1+1=0+1=1]$$

■

• **By using Parity check matrix:** Let H be an $r \times n (r < n)$ parity check matrix with entries from F_q with q elements and rank$(H) = r$. Let

$$C = \{x : x \in V_n(F_q) \text{ and } Hx = 0\}$$

Take $x, y \in C$ and any $\alpha \in F_q$, then we have

$$\begin{aligned} &Hx = 0, \quad Hy = 0.\\ &H(x+y) = H(x) + H(y) = 0\\ \text{and } &H(\alpha x) = \alpha H(x) = 0. \end{aligned}$$

Therefore, C is a linear subspace of $V_n(F_q)$ and so a linear code over the field F_q.

Clearly H is a parity check matrix for the code C. The dimension of code C is $n - r$.

Example 8.2.2. Find a codeword determined by the binary parity check matrix

$$H = \begin{pmatrix}1 & 0 & 1 & 0\\ 0 & 1 & 1 & 1\end{pmatrix}$$

Solution. Here H is a binary parity check matrix with 4 columns and rank$(H) = 2$. Therefore, the linear

code C determined by the parity check matrix H consists of binary codewords $(x_1\ x_2\ x_3\ x_4)$ satisfies

$$\begin{pmatrix}1&0&1&0\\0&1&1&1\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\end{pmatrix}=0$$

$$\Rightarrow x_1+x_3=0 \quad \text{and} \quad x_2+x_3+x_4=0$$

$$\Rightarrow x_1=x_3 \quad \text{and} \quad x_2=x_3+x_4 \ [\because 2\cdot 1=0;\ 1+1=0;\ -1=1]$$

If the values of x_3 and x_4 are assigned then x_1 and x_2 are determined. There are four ways of choosing x_3 and x_4 i.e., 00, 01, 10, 11, leading to the codewords 0000, 0101, 1110, 1011. ■

The following is a similar problem.

Example 8.2.3. Find the codewords determined by the P.C.M

$$H=\begin{pmatrix}1&0&0&0&1\\0&1&0&1&1\\0&0&1&1&1\end{pmatrix}$$

8.3 Standard form of parity check matrix:

The standard $r\times n$ parity check matrix H is given by

$$H=\begin{bmatrix}1&0&\dots&0&b_{11}&b_{12}&\dots&b_{1\overline{n-r}}\\0&1&\dots&0&b_{21}&b_{22}&\dots&b_{2\overline{n-r}}\\0&0&\dots&0&b_{31}&b_{32}&\dots&b_{3\overline{n-r}}\\\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots\\0&0&\dots&1&b_{r1}&b_{r2}&\dots&b_{r\overline{n-r}}\end{bmatrix}_{r\times n}$$

with entries from the field F_q with q elements.

$$\begin{aligned}x_1 &= b_{11}x_{r+1}+b_{12}x_{r+2}+\dots+b_{1\overline{n-r}}x_n\\x_2 &= b_{21}x_{r+1}+b_{22}x_{r+2}+\dots+b_{2\overline{n-r}}x_n\\ &\dots \quad\quad \dots \quad\quad \dots \quad\quad \dots\\x_r &= b_{r1}x_{r+1}+b_{r2}x_{r+2}+\dots+b_{r\overline{n-r}}x_n\end{aligned}$$

These equations determine $x_1,x_2,\dots,x_r$ when the values of $x_{r+1},\ x_{r+2},\dots,x_n$ are assigned, since there are q^{n-r} ways of choosing the values to obtain the linear code C of dimension $n-r$.

8.4 Hamming Code:

Let r be a given positive integer. We determine a binary matrix H with r rows and with maximum number of columns such that no column of H consist entirely of $0's$ and no two columns of H are same. Then linear code C determined by the parity check matrix H, we correct one error. This code C is called a Hamming code. Since each column of H has r entries and each entry is either 0 or 1, then the maximum number of different column is $n=2^r-1$. (The column consisting of entirely $0's$ being excluded.)

Exercise 8.4.1. Determine the Hamming code by the following P.C.M

$$H=\begin{pmatrix}1&0&0&1&1&0&1\\0&1&0&1&0&1&1\\0&0&1&0&1&1&1\end{pmatrix}$$

8.5 Cyclic Code

Here we shall denote the word "a" of length n by $a_0a_1a_2\ldots a_{n-1}$. The word $\hat{a} = a_{n-1}a_0a_1a_2\ldots a_{n-2}$ is called the 1st cyclic shift of the word a. A code C in $V_n(F_q)$ is said to be cyclic if it is linear and $a \in C \Rightarrow \hat{a} \in C$.

Let C be a cyclic code in $V_n(F_q)$ and $a \in C$. Then the words are obtained from a by n number of cyclic shifts. Any number of cyclic shifts such as

$$a_ia_{i+1}\cdots a_{n-1}a_0a_1\cdots a_{i-1}$$

belong to C.

Cyclic codes are useful for two reasons; from the practical point of view, it is possible to implement by simple devices known as shift resister. On the other hand, cyclic code can be constructed and investigated by means of algebraic theory of rings and polynomials.

Construction of a cyclic code

Let C be a cyclic code in $V_n(F_q)$ generated by $g(x)$. Then $g(x)$ is a divisor of $x^n - 1$. So we have

$$x^n - 1 = h(x)g(x) \tag{8.5.1}$$

Let

$$\begin{aligned} h(x) &= h_0 + h_1x + h_2x^2 + \ldots + h_kx^k \\ g(x) &= g_0 + g_1x + g_2x^2 + \ldots g_{n-k-1}x^{n-k-1} + g_{n-k}x^{n-k} \end{aligned}$$

where $g_{n-k} = 1$.

It is easy to see from (8.5.1) that $h_k = 1$ and $h_0g_0 = -1$, which gives that $h_0 \neq 0, g_0 \neq 0$. The polynomial $g(x)$ corresponds to the codeword

$$g = g_0\ g_1\ g_2\ \cdots\ g_{n-k}\ 0\ 0\ \cdots\ 0 \quad \text{in} \quad V_n(F_q)$$

The polynomial $x^ig(x)$ $(1 \leq i \leq k-1)$ corresponds to the codeword

$$g_{(i)} = 0\ 0\ \cdots\ g_0\ g_1\cdots\ g_{n-k}\ 0\ 0\ \cdots 0$$

There are i zeros at the beginning and $k-1-i$ zeros at the end. We denote by $\overline{h}$, the codeword whose 1st $k+1$ bits are $h_kh_{k-1}\cdots h_1h_0$ followed by $n-k-1$ zeros.

$$\therefore \overline{h} = h_k\ h_{k-1}\cdots\ h_1\ h_0\ 0\ 0\ \cdots 0$$

Let H denote the $(n-k)\times n$ matrix whose rows are $\overline{h},\ \overline{h}_{(1)},\ \overline{h}_{(2)}, \ldots, \overline{h}_{(n-k+1)}$, where $\overline{h}_{(i)}$ is the i-th cyclic shift of the codeword $\overline{h}$. Hence

$$H = \begin{bmatrix} h_k & h_{k-1} & \cdots & h_1 & h_0 & 0 & 0 & \cdots & 0 \\ 0 & h_k & h_{k-1} & \cdots & h_1 & h_0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & h_k & h_{k-1} & \cdots & h_1 & h_0 \end{bmatrix}$$

Example 8.5.1. Determine the binary parity check matrix for the cyclic code $C =< g(x) >$ of length 7 where $g(x) = 1 + x^2 + x^3$ and obtain the code C.

Solution. The factorization of $x^7 - 1$ into irreducible polynomials, i.e.,

$$\begin{aligned} x^7 - 1 &= (1+x)(1+x+x^3)(1+x^2+x^3) \\ &= h(x)g(x) \quad [\because \text{In a binary code, -1} = 1] \end{aligned} \tag{8.5.2}$$

$$\begin{aligned} \therefore h(x) &= (1+x)(1+x+x^3) \\ &= (1+x^2+x^3+x^4) \quad [\because 1+1=0] \end{aligned}$$

$$\therefore h_0 = 1, \quad h_1 = 0, \quad h_2 = 1, \quad h_3 = 1, \quad h_4 = 1.$$

$$H = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 \end{pmatrix}$$

No column of H consist entirely $0's$ and no two columns are exactly same. So the code determined by H is a Hamming code of length 7. ■

8.6 BCH Codes

The most powerful multiple error correcting codes for in random independent errors which have been discovered by Bose-Chowdhuri-Hoequenghem.

This code is known as BCH code. The code was discovered independently around 1960 by Bose and Chowdhuri and by Hoequenghem. Foe moderate wordlengths these codes are very good.

Unit 9

Course Structure

- Markovian decision Process
- Powers of Stochastic Matrices
- Regular matrices

9.1 Introduction

A Markov Process consists of a set of objects and a set of states such that

(i) at any given time, each object must be in a state (distinct objects need not be in distinct states).

(ii) the probability that an object moves from one state to another state which may be the same as the first state, in one time period depends only on those two states.

The integral numbers of time periods past the moment when the process is started represent the stages of the process which may be finite or infinite.

If the number of states is finite or countably infinite, the Markov process is called a **Markov Chain**. A finite Markov chain is one having a finite number of states. We denote the probability of moving from state i to state j in one time period by p_{ij}. For an N state Markov chain, where N is a fixed positive integer, the $N \times N$ matrix $P = [p_{ij}]$ is the **stochastic** or **transition matrix** associated with the process. Necessarily, the elements of each row of P sum to unity.

Theorem 9.1.1. Every stochastic matrix has 1 as an eigen value (possible multiple and none of the eigen values exceed 1 in absolute value).

Because of the way P is defined, it proves convenient in this chapter to indicate N-dimensional vectors as row vectors.

According to the theorem, there exists a vector $X \neq 0$ such that $XP = X$. This left eigen value is called a fixed point of P.

9.2 Powers of Stochastic Matrices

We denote the nth power of a matrix P by

$$P^n \equiv [p_{ij}^{(n)}],$$

where $p_{ij}^{(n)}$ represents the probability that an object moves from state i to state j in n-time periods. P^n is obviously a stochastic matrix.

We write $X^{(0)} = [x_1^{(0)}, x_2^{(0)}, \ldots, x_N^{(0)}]$ which represents the proportion of objects in each state of the beginning of the process whereas

$$X^{(n)} = [x_1^{(n)}, x_2^{(n)}, \ldots, x_N^{(n)}],$$

where, $x_i^{(N)}$ represents the proportion of objects in state i at the end of nth time period, $1 \leq i \leq N$.

$X^{(n)}$ is related to $X^{(0)}$ by the relation $X^{(n)} = X^{(0)}P^n$.

Example 9.2.1. Grapes in Kashmir are classified as either superior, average or poor. Following a superior harvest, the probabilities of having a superior, average and poor harvest in the next year are 0, 0.8 and 0.2. Following an average harvest, the probabilities of a superior, average and poor harvest are 0.2, 0.6 and 0.1. Following a poor harvest, the probabilities of a superior, average and poor harvest are 0.1, 0.8 and 0.1. Determine the probabilities of a superior harvest for each of the next five years if the most recent harvest was average.

Solution. The transition matrix is given by

$$\begin{array}{c} \\ S \\ A \\ P \end{array} \begin{array}{c} \begin{array}{ccc} \text{superior}(S) & \text{average}(A) & \text{poor}(P) \end{array} \\ \begin{bmatrix} 0 & 0.8 & 0.2 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.8 & 0.1 \end{bmatrix} \end{array}$$

Since the most recent harvest rate was average, so,

$$X^{(0)} = \begin{array}{c} \begin{array}{ccc} \text{S} & \text{A} & \text{P} \end{array} \\ \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \end{array}$$

initial probability distribution. Thus,

$$X^{(5)} = X^{(0)}P^5.$$

Now,

$$\begin{aligned} P^2 &= \begin{bmatrix} 0 & 0.8 & 0.2 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.8 & 0.1 \end{bmatrix} \begin{bmatrix} 0 & 0.8 & 0.2 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.8 & 0.1 \end{bmatrix} \\ &= \begin{bmatrix} 0+0.16+0.02 & 0+0.48+0.16 & 0+0.16+0.02 \\ 0+0.12+0.02 & 0.16+0.36+0.16 & 0.04+0.12+0.02 \\ 0+0.16+0.01 & 0.08+0.48+0.08 & 0.02+0.16+0.01 \end{bmatrix} \\ &= \begin{bmatrix} 0.18 & 0.64 & 0.18 \\ 0.14 & 0.68 & 0.18 \\ 0.17 & 0.64 & 0.19 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} P^4 &= \begin{bmatrix} 0.18 & 0.64 & 0.18 \\ 0.14 & 0.68 & 0.18 \\ 0.17 & 0.64 & 0.19 \end{bmatrix} \begin{bmatrix} 0.18 & 0.64 & 0.18 \\ 0.14 & 0.68 & 0.18 \\ 0.17 & 0.64 & 0.19 \end{bmatrix} \\ &= \begin{bmatrix} 0.1526 & 0.6656 & 0.1818 \\ 0.1510 & 0.6672 & 0.1818 \\ 0.1525 & 0.6656 & 0.1819 \end{bmatrix}. \end{aligned}$$

$$\begin{aligned} P^5 &= \begin{bmatrix} 0.1526 & 0.6656 & 0.1818 \\ 0.1510 & 0.6672 & 0.1818 \\ 0.1525 & 0.6656 & 0.1819 \end{bmatrix} \begin{bmatrix} 0.18 & 0.64 & 0.18 \\ 0.14 & 0.68 & 0.18 \\ 0.17 & 0.64 & 0.19 \end{bmatrix} \\ &= \begin{bmatrix} 0.151558 & 0.666624 & 0.181818 \\ 0.151494 & 0.666688 & 0.181818 \\ 0.151557 & 0.666624 & 0.181819 \end{bmatrix}. \end{aligned}$$

Thus,

$$\begin{aligned} X^{(5)} &= \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} P^5 \\ &= \begin{bmatrix} 0.151494 & 0.666688 & 0.181818 \end{bmatrix}. \end{aligned}$$

Hence the probability of a superior harvest for each of the next five years is 0.151494. ■

Definition 9.2.2. (Regular Matrix:) A stochastic matrix is regular if one of its powers contains only positive entries.

Theorem 9.2.3. If a stochastic matrix is regular, then 1 is an eigen value of multiplicity one, and all other eigen values λ_i satisfy $|\lambda_i| \leq 1$.

Example 9.2.4. Is the stochastic matrix

$$P = \begin{bmatrix} 0 & 1 \\ 0.4 & 0.6 \end{bmatrix}$$

regular?

Solution.

$$P^2 = \begin{bmatrix} 0 & 1 \\ 0.4 & 0.6 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0.4 & 0.6 \end{bmatrix} = \begin{bmatrix} 0.40 & 0.60 \\ 0.24 & 0.76 \end{bmatrix}.$$

Since each entry of P^2 is positive, hence P is regular. ■

Unit 10

Course Structure

- Ergodic Matrices

10.1 Ergodic Matrix

Definition 10.1.1. (Ergodic Matrix:) A stochastic matrix P is ergodic if $\lim_{n\to\infty} P^n$ exists, that is, each $P_{ij}^{(n)}$ has a limit as $n \to \infty$. We denote $L = \lim_{n\to\infty} P^n$. Obviously, P is a stochastic matrix. $X^{(\infty)}$ is defined by the equation $X(\infty) = X^{(0)}L$.

The components of $X^{(\infty)}$ are limiting state distributions and represent the approximate proportions of objects in the various states of a Markov chain after a large number of time periods.

Theorem 10.1.2. A stochastic matrix is ergodic if and only if the only eigen value λ of magnitude 1 is 1 itself and if $\lambda = 1$ has multiplicity k, then there exists k linearly independent (left) eigen vectors associated with this eigen value.

Theorem 10.1.3. A regular matrix is ergodic but the converse is not true in general.

If P is regular with limit matrix L, then the rows of L are identical with one another, each being the unique left eigen vector of P associated with the eigen value $\lambda = 1$ and having the sum of its components equal to unity.

Let us denote this eigen vector by E_1. Now, if P is regular, then regardless of the initial distribution $X^{(0)}$, we can write $X^{(\infty)} = E_1 (= X^{(0)}L)$.

Example 10.1.4. Is the stochastic matrix

$$P = \begin{bmatrix} 0 & 1 \\ 0.4 & 0.6 \end{bmatrix}$$

ergodic? Calculate $L = \lim_{n\to\infty} P^n$, if it exists.

Solution. Since each entry of

$$P^2 = \begin{bmatrix} 0.40 & 0.60 \\ 0.24 & 0.76 \end{bmatrix}$$

is positive, P is regular and therefore, ergodic; hence $L = \lim_{n\to\infty} P^n$ exists. Now,

$$\begin{aligned} \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 0.40 & 0.60 \\ 0.24 & 0.76 \end{bmatrix} &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \\ \Rightarrow\ x_1 - 0.4x_2 &= 0 \qquad (10.1.1) \\ \text{and}\ \ x_1 + x_2 &= 1. \qquad (10.1.2) \end{aligned}$$

Solving equation (10.1.1) and (10.1.2), we get,

$$x_1 = \frac{2}{7} \quad \text{and} \quad x_2 = \frac{5}{7}.$$

Thus,

$$E_1 = \begin{bmatrix} \frac{2}{7} & \frac{5}{7} \end{bmatrix} \quad \text{and} \quad \lim_{n\to\infty} P^n = L = \begin{bmatrix} \frac{2}{7} & \frac{5}{7} \\ \frac{2}{7} & \frac{5}{7} \end{bmatrix}.$$

■

Theorem 10.1.5. If every eigen value of a matrix P yields linearly independent (left) eigen vectors in number equal to its multiplicity, then there exists a non-singular matrix M, whose rows are left eigen vectors of P, such that $D \equiv MPM^{-1}$ is a diagonal matrix. The diagonal elements of D are the eigen values of P, repeated according to multiplicity.

We have,

$$\begin{aligned} L &= \lim_{n\to\infty} P^n \\ &= (M^{-1}M) \lim_{n\to\infty} P^n (M^{-1}M) \\ &= M^{-1} \left(\lim_{n\to\infty} MP^nM^{-1} \right) M \\ &= M^{-1} \left(\lim_{n\to\infty} (MPM^{-1})^n \right) M \\ &= M^{-1} \left(\lim_{n\to\infty} D^n \right) M \\ &= M^{-1} \begin{bmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & & \ddots & & & & \\ & & & 1 & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{bmatrix}_{N\times N} M. \end{aligned}$$

The diagonal matrix on the right has k $1's$ and $(N-k)$ $0's$ on the main diagonal.

Example 10.1.6. Is the stochastic matrix

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.4 & 0 & 0.6 & 0 \\ 0.2 & 0 & 0.1 & 0.7 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

regular? Is it ergodic? Calculate $L = \lim_{n\to\infty} P^n$, if it exists.

Solution. The characteristic equation of P is

$$\begin{aligned}
\begin{vmatrix} 1-\lambda & 0 & 0 & 0 \\ 0.4 & -\lambda & 0.6 & 0 \\ 0.2 & 0 & 0.1-\lambda & 0.7 \\ 0 & 0 & 0 & 1-\lambda \end{vmatrix} &= 0 \\
\Rightarrow (1-\lambda)(-\lambda)(0.1-\lambda)(1-\lambda) &= 0 \\
\Rightarrow \lambda &= 1,\ 1,\ 0.1,\ 0.
\end{aligned}$$

Thus, $\lambda_1 = 1$ (multiplicity 2), $\lambda_2 = 0.1$, $\lambda_3 = 0$ are the eigen values of P. Hence P is not regular.

The left eigen vectors for the double eigen value $\lambda_1 = 1$ are $[1, 0, 0, 0]$ and $[0, 0, 0, 1]$, which are linearly independent. Hence P is ergodic. Thus, $L = \lim_{n\to\infty} P^n$ exists.

We now fine the eigen vectors corresponding to $\lambda_2 = 0.1$ and $\lambda_3 = 0$.

$$\begin{aligned}
\begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.4 & 0 & 0.6 & 0 \\ 0.2 & 0 & 0.1 & 0.7 \\ 0 & 0 & 0 & 1 \end{bmatrix} &= 0.1 \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} \\
\Rightarrow (1-0.1)x_1 + 0.4x_2 + 0.2x_3 &= 0 \\
-0.2x_2 &= 0 \\
0.6x_2 + (0.1-0.1)x_3 &= 0 \\
0.7x_3 + (1-0.1)x_4 &= 0 \\
\Rightarrow 0.9x_1 + 0.4x_2 + 0.2x_3 &= 0 \\
-0.1x_2 &= 0 \\
0.6x_2 &= 0 \\
0.7x_3 + 0.9x_4 &= 0.
\end{aligned}$$

Solving these equations, we get,

$$x_1 = -2, \quad x_2 = 0, \quad x_3 = 9, \quad x_4 = -7.$$

Thus, the eigen vector corresponding to λ_2 is $[-2, 0, 9, -7]$. Again,

$$\begin{aligned}
\begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.4 & 0 & 0.6 & 0 \\ 0.2 & 0 & 0.1 & 0.7 \\ 0 & 0 & 0 & 1 \end{bmatrix} &= 0 \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} \\
\Rightarrow x_1 + 0.6x_2 + 0.2x_3 &= 0 \\
0.6x_2 + 0.1x_3 &= 0 \\
0.7x_3 + x_4 &= 0.
\end{aligned}$$

Solving these equations, we get

$$x_1 = 4, \quad x_2 = 5, \quad x_3 = -30, \quad x_4 = 21.$$

Thus, the eigen vector corresponding to λ_3 is $[4, 5, -30, 21]$.

To make P diagonalizable, we consider

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ -2 & 0 & 9 & -7 \\ 4 & 5 & -30 & 21 \end{bmatrix} \text{ and } D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

We now find M^{-1}.

$$\begin{aligned}
[M : I] &= \left[\begin{array}{cccc:cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ -2 & 0 & 9 & -7 & 0 & 0 & 1 & 0 \\ 4 & 5 & -30 & 21 & 0 & 0 & 0 & 1 \end{array}\right] \\
&\xrightarrow[R_4 \to R_2]{R_2 \to R_4} \left[\begin{array}{cccc:cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 4 & 5 & -30 & 21 & 0 & 0 & 0 & 1 \\ -2 & 0 & 9 & -7 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{array}\right] \\
&\xrightarrow[R_3 \to R_3 + 2R_1]{R_2 \to R_2 - 4R_1} \left[\begin{array}{cccc:cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 5 & -30 & 21 & -4 & 0 & 0 & 1 \\ 0 & 0 & 9 & -7 & 2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{array}\right] \\
&\xrightarrow[R_3 \to \frac{1}{9} R_3]{R_2 \to \frac{1}{5} R_2} \left[\begin{array}{cccc:cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & -6 & \frac{21}{5} & -\frac{4}{5} & 0 & 0 & \frac{1}{5} \\ 0 & 0 & 1 & -\frac{7}{9} & \frac{2}{9} & 0 & \frac{1}{9} & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{array}\right] \\
&\xrightarrow{R_2 \to R_2 + 6R_3} \left[\begin{array}{cccc:cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -\frac{7}{15} & \frac{8}{15} & 0 & \frac{2}{3} & \frac{1}{5} \\ 0 & 0 & 1 & -\frac{7}{9} & \frac{2}{9} & 0 & \frac{1}{9} & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{array}\right] \\
&\xrightarrow[R_3 \to R_3 + \frac{7}{9} R_4]{R_2 \to R_2 + \frac{7}{15} R_4} \left[\begin{array}{cccc:cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & \frac{8}{15} & \frac{7}{15} & \frac{2}{3} & \frac{1}{5} \\ 0 & 0 & 1 & 0 & \frac{2}{9} & \frac{7}{9} & \frac{1}{9} & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{array}\right].
\end{aligned}$$

Thus

$$M^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{8}{15} & \frac{7}{15} & \frac{2}{3} & \frac{1}{5} \\ \frac{2}{9} & \frac{7}{9} & \frac{1}{9} & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Thus,

$$\begin{aligned} L &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{8}{15} & \frac{7}{15} & \frac{2}{3} & \frac{1}{5} \\ \frac{2}{9} & \frac{7}{9} & \frac{1}{9} & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ -2 & 0 & 9 & -7 \\ 4 & 5 & -30 & 21 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{8}{15} & \frac{7}{15} & 0 & 0 \\ \frac{2}{9} & \frac{7}{9} & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ -2 & 0 & 9 & -7 \\ 4 & 5 & -30 & 21 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{8}{15} & 0 & 0 & \frac{7}{15} \\ \frac{2}{9} & 0 & 0 & \frac{7}{9} \\ 0 & 0 & 0 & 1 \end{bmatrix}. \end{aligned}$$

■

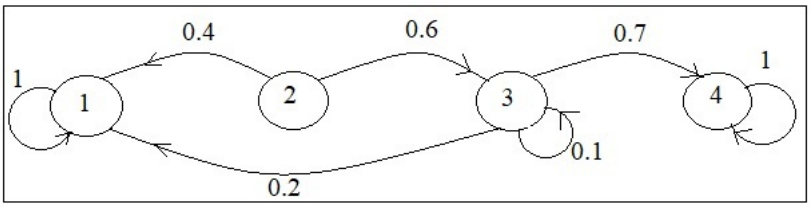
Example 10.1.7. Construct the state-transition diagram for the Markov chain

$$P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.4 & 0 & 0.6 & 0 \\ 0.2 & 0 & 0.1 & 0.7 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

Solution. [A state-transition diagram is an oriented network in which the nodes represent states and the arcs represent possible transitions.]

Labelling the states by $1, 2, 3, 4$, we have the following state-transition diagram.

The number on each arc is the probability of the transition. ■



Example 10.1.8. Prove that if P is regular, then all the rows of $L = \lim_{n\to\infty} P^n$ are identical.

Solution. Given, $L = \lim_{n\to\infty} P^n$. Also, we have, $L = \lim_{n\to\infty} P^{n-1}$. Consequently,

$$L = \lim_{n\to\infty} P^n = \lim_{n\to\infty} (P^{n-1})P = (\lim_{n\to\infty} P^{n-1})P = LP$$

which implies that every row of L is a left eigen vector of P corresponding to the eigen value $\lambda = 1$.

Now, P being regular, all such eigen vectors are scalar multiples of a single vector.

On the other hand, L being stochastic, each row of it sums to unity. Thus it follows that all the rows are identical. ■

Example 10.1.9. Prove that if λ is an eigen value of a stochastic matrix P, then $|\lambda| \leq 1$.

Solution. Let $E = [e_1 \;\; e_2 \;\; \ldots \;\; e_N]^T$ be a right eigen vector corresponding to λ. Then $PE = \lambda E$, and considering the jth component of both sides of this equality, we conclude that

$$\sum_{k=1}^{N} p_{jk} e_k = \lambda e_j. \tag{10.1.3}$$

Let e_i be that component of E having the greatest magnitude, that is,

$$|e_i| = \max\{|e_1|, |e_2|, \ldots, |e_N|\}. \tag{10.1.4}$$

By definition, $E \neq 0$, so that $|e_i| > 0$. Thus, it follows from (10.1.3), with $j = i$ and from (10.1.4) that,

$$|\lambda||e_i| = |\lambda e_i| = \left| \sum_{k=1}^{N} p_{ik} e_k \right| \leq \sum_{k=1}^{N} p_{ik} |e_k| \leq |e_i| \sum_{k=1}^{N} p_{ik} = |e_i|,$$

which implies that $|\lambda| \leq 1$. ■

Example 10.1.10. Formulate the following process as a Markov chain:

The manufacturer of Hi-Glo toothpaste currently controls 60% of the market in a particular city. Data from the previous year show that 88% of Hi-Glo's customers remained loyal to Hi-Glo, while 12% of Hi-Glo's customers switched to rival brands. In addition, 85% of the competition's customers remained loyal to the competition, while the other 15% switched to Hi-Glo. Assuming that these trends continue, determine Hi-Glo's share of the market

(a) in 5 years and (b) over the long run.

Solution. We take state 1 to be consumption of Hi-Glo toothpaste and state 2 to be consumption of a rival brand. Then p_{11} is the probability that a Hi-Glo customer remains loyal to Hi-Glo, that is, 0.88; p_{12} is the probability that a Hi-Glo customer switches to another brand, that is, 0.12; p_{21} is the probability that the customer of another brand switches to Hi-Glo, that is, 0.15; p_{22} is the probability that customer of another brand remains loyal to the competition, that is, 0.85.

The stochastic matrix (Markov chain) defined by these transition probabilities is

$$P = \begin{matrix} \\ 1 \\ 2 \end{matrix} \begin{matrix} \begin{matrix} 1 & \quad 2 \end{matrix} \\ \begin{bmatrix} 0.88 & 0.12 \\ 0.15 & 0.85 \end{bmatrix} \end{matrix}$$

The initial probability distribution vector is $X^{(0)} = [0.60 \;\; 0.40]$, where, the components $x_1^{(0)} = 0.60$ and $x_2^{(0)} = 0.40$ represent the proportions of people initially in states 1 and 2, respectively.

(a) Thus,

$$\begin{aligned} X^{(5)} &= X^{(0)} P^5 \\ &= \begin{bmatrix} 0.60 & 0.40 \end{bmatrix} \begin{bmatrix} 0.6477 & 0.3523 \\ 0.4404 & 0.5596 \end{bmatrix} \\ &= \begin{bmatrix} 0.5648 & 0.4352 \end{bmatrix}. \end{aligned}$$

After 5 years, Hi-Glo's share of the market will have declined to 56.48%. Now,

$$P = \begin{bmatrix} 0.88 & 0.12 \\ 0.15 & 0.85 \end{bmatrix}$$

is regular, since each entry of the first power of P is positive, that is, P is positive. Hence P is ergodic. So, $\lim_{n\to\infty} P^n = L$(say) exists. Now, the left eigen vector corresponding to $\lambda = 1$ is given by

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 0.88 & 0.12 \\ 0.15 & 0.85 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}$$
$$\Rightarrow 0.12x_1 - 0.15x_2 = 0 \quad \text{and} \quad x_1 + x_2 = 1.$$

Solving, we get,

$$x_1 = \frac{5}{9} \quad \text{and} \quad x_2 = \frac{4}{9}$$

and thus

$$E_1 = \begin{bmatrix} x_1 & x_2 \end{bmatrix} = \begin{bmatrix} \frac{5}{9} & \frac{4}{9} \end{bmatrix}.$$

Hence,

$$L = \lim_{n\to\infty} P^n = \begin{bmatrix} \frac{5}{9} & \frac{4}{9} \\ \frac{5}{9} & \frac{4}{9} \end{bmatrix}.$$

(b)

$$\begin{aligned} X^{(\infty)} &= X^{(0)}L \\ &= \begin{bmatrix} 0.60 & 0.40 \end{bmatrix} \begin{bmatrix} \frac{5}{9} & \frac{4}{9} \\ \frac{5}{9} & \frac{4}{9} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{3} + \frac{2}{9} & \frac{12}{45} + \frac{16}{45} \end{bmatrix} = \begin{bmatrix} \frac{5}{9} & \frac{4}{9} \end{bmatrix} = E_1. \end{aligned}$$

Therefore, over the long run, Hi-Glo's share of the market will stabilize at $\frac{5}{9}$, that is, approximately 55.56%.

■

Example 10.1.11. Solve the previous problem, if Hi-Glo currently controls 90% of the market

(a)

$$\begin{aligned} X^{(5)} &= X^{(0)}P^5 \\ &= \begin{bmatrix} 0.90 & 0.10 \end{bmatrix} \begin{bmatrix} 0.6477 & 0.3523 \\ 0.4404 & 0.5596 \end{bmatrix} \\ &= \begin{bmatrix} 0.6270 & 0.3730 \end{bmatrix}. \end{aligned}$$

Therefore, after 5 years, Hi-Glo controls approximately 68% of the market.

(b) Since P is regular,

$$X^{(\infty)} = E_1 = \begin{bmatrix} \frac{5}{9} & \frac{4}{9} \end{bmatrix}.$$

Example 10.1.12. The geriatric ward of a hospital lists its patients as bedridden or ambulatory. Historical data indicate that over a 1-week period, 30% of all ambulatory patients are discharged, 40% remain ambulatory, and 30% are remanded to complete bed rest. During the same period, 50% of all the bedridden patients become ambulatory, 20% remain bedridden, and 30% die. Currently the hospital has 100 patients in its geriatric ward, with 30 bedridden and 70 ambulatory. Determine the status of the patients

(a) after 2 weeks, and

(b) over the long run

(The status of a discharged patient does not change if the patient die).

Solution. We take state 1 to be discharged, sate 2 to be ambulatory, state 3 to be bedridden or bed rest and state 4 to be died patients. Consider 1 time period to be 1 week.

The transition probabilities given by the following transition matrix:

$$P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1\text{(Discharged)} & 2\text{(Ambulatory)} & 3\text{(Bedridden)} & 4\text{(Died)} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.3 & 0.4 & 0.3 & 0 \\ 0 & 0.5 & 0.2 & 0.3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

Since, currently the hospital has 100 patients in its geriatric ward, with 30 bedridden and 70 ambulatory, so the initial probability distribution vector is

$$X^{(0)} = \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 0.7 & 0.3 & 0 \end{bmatrix} \end{array}$$

Now,

$$\begin{aligned} P^2 &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.3 & 0.4 & 0.3 & 0 \\ 0 & 0.5 & 0.2 & 0.3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.3 & 0.4 & 0.3 & 0 \\ 0 & 0.5 & 0.2 & 0.3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.42 & 0.31 & 0.18 & 0.09 \\ 0.15 & 0.30 & 0.19 & 0.36 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \end{aligned}$$

(a)

$$\begin{aligned} X^{(2)} &= X^{(0)}P^2 \\ &= \begin{bmatrix} 0 & 0.7 & 0.3 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.42 & 0.31 & 0.18 & 0.09 \\ 0.15 & 0.30 & 0.19 & 0.36 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 0.339 & 0.307 & 0.183 & 0.171 \end{bmatrix}. \end{aligned}$$

After 2 weeks, there are approximately 34% discharged, 30% ambulatory, 18% bedridden and 17% dead patients.

Now, the characteristic equation of P is

$$\begin{aligned} |P - \lambda I| &= 0 \\ \Rightarrow \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0.3 & 0.4 & 0.3 & 0 \\ 0 & 0.5 & 0.2 & 0.3 \\ 0 & 0 & 0 & 1 \end{vmatrix} &= 0 \\ \Rightarrow (1-\lambda)^2(\lambda^2 - 0.6\lambda - 0.07) &= 0 \\ \Rightarrow \lambda &= 1,\ 1,\ 0.7,\ -0.1. \end{aligned}$$

Since $\lambda_1 = 1$ (multiplicity 2), $\lambda_2 = 0.7$, $\lambda_3 = -0.1$ are the eigen values of P, so P is not regular.

The left eigen vectors for the double eigen value 1 are $[1 \;\; 0 \;\; 0 \;\; 0]$ and $[0 \;\; 0 \;\; 0 \;\; 1]$ which are linearly independent. Hence P is ergodic. Therefore,

$$L = \lim_{n\to\infty} P^n.$$

Now,

$$\begin{aligned}
\begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.3 & 0.4 & 0.3 & 0 \\ 0 & 0.5 & 0.2 & 0.3 \\ 0 & 0 & 0 & 1 \end{bmatrix} &= 0.7 \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} \\
\Rightarrow (1-0.7)x_1 + 0.3x_2 &= 0 \\
(0.4-0.7)x_2 + 0.5x_3 &= 0 \\
0.3x_2 + (0.2-0.7)x_3 &= 0 \\
0.3x_3 + (1-0.7)x_4 &= 0 \\
\Rightarrow 0.3x_1 + 0.3x_2 &= 0 \\
0.3x_2 - 0.5x_3 &= 0 \\
0.3x_3 + 0.3x_4 &= 0.
\end{aligned}$$

Solving the above equations, we get

$$x_1 = -x_2 = -\frac{5}{3}x_3 = \frac{5}{3}x_4.$$

Let $x_4 = 3$. Then we get

$$x_1 = 5, \quad x_2 = -5, \quad x_3 = -3.$$

Thus,

$$\begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} = \begin{bmatrix} 5 & -5 & -3 & 3 \end{bmatrix}$$

is the eigen vector corresponding to $\lambda_2 = 0.7$.

Now,

$$\begin{aligned}
\begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.3 & 0.4 & 0.3 & 0 \\ 0 & 0.5 & 0.2 & 0.3 \\ 0 & 0 & 0 & 1 \end{bmatrix} &= -0.1 \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} \\
\Rightarrow (1+0.1)x_1 + 0.3x_2 &= 0 \\
(0.4+0.1)x_2 + 0.5x_3 &= 0 \\
0.3x_2 + (0.2+0.1)x_3 &= 0 \\
0.3x_3 + (1+0.1)x_4 &= 0 \\
\Rightarrow 1.1x_1 + 0.3x_2 &= 0 \\
x_2 + x_3 &= 0 \\
0.3x_3 + 1.1x_4 &= 0.
\end{aligned}$$

Solving the equations, we get,

$$x_1 = -\frac{3}{11}x_2 = \frac{3}{11}x_3 = -x_4.$$

Taking $x_2 = 11$, we get

$$x_1 = -3, \quad x_2 = 11, \quad x_3 = 3, \quad x_4 = 3.$$

Thus, $[-3 \ \ 11 \ \ 3 \ \ 3]$ is the eigen vector corresponding to $\lambda_3 = -0.1$.

To make P diagonalizable, we consider

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 5 & -5 & -3 & 3 \\ -3 & 11 & 3 & 3 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0.4 & 0 \\ 0 & 0 & 0 & -0.1 \end{bmatrix}.$$

To find M^{-1}:

$$[M : I] = \left[\begin{array}{cccc:cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 5 & -5 & -3 & 3 & 0 & 0 & 1 & 0 \\ -3 & 11 & 3 & 3 & 0 & 0 & 0 & 1 \end{array}\right]$$

$$\xrightarrow[R_3 \to R_3 - 5R_1]{R_2 \leftrightarrow R_4} \left[\begin{array}{cccc:cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ -3 & 11 & 3 & 3 & 0 & 0 & 0 & 1 \\ 0 & -5 & -3 & 3 & -5 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{array}\right]$$

$$\xrightarrow{R_2 \to R_2 + 3R_1} \left[\begin{array}{cccc:cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 11 & 3 & 3 & 3 & 0 & 0 & 1 \\ 0 & -5 & -3 & 3 & -5 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{array}\right]$$

$$\xrightarrow[R_3 \to R_3 - 3R_4]{R_2 \to R_2 - 3R_4} \left[\begin{array}{cccc:cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 11 & 3 & 0 & 3 & -3 & 0 & 1 \\ 0 & -5 & -3 & 0 & -5 & -3 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{array}\right]$$

$$\xrightarrow{R_2 \to \frac{1}{11} R_2} \left[\begin{array}{cccc:cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & \frac{3}{11} & 0 & \frac{3}{11} & -\frac{3}{11} & 0 & \frac{1}{11} \\ 0 & -5 & -3 & 0 & -5 & -3 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{array}\right]$$

$$\xrightarrow{R_3 \to R_3 + 5R_2} \left[\begin{array}{cccc:cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & \frac{3}{11} & 0 & \frac{3}{11} & -\frac{3}{11} & 0 & \frac{1}{11} \\ 0 & 0 & -\frac{18}{11} & 0 & -\frac{40}{11} & -\frac{48}{11} & 1 & \frac{5}{11} \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{array}\right]$$

$$\xrightarrow[R_3 \to -\frac{11}{18} R_3]{R_2 \to R_2 + \frac{1}{6} R_3} \left[\begin{array}{cccc:cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -\frac{1}{3} & -1 & \frac{1}{6} & \frac{1}{66} \\ 0 & 0 & 1 & 0 & \frac{20}{9} & \frac{8}{3} & -\frac{11}{18} & -\frac{5}{18} \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{array}\right].$$

Thus,

$$M^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{3} & -1 & \frac{1}{6} & \frac{1}{66} \\ \frac{20}{9} & \frac{8}{3} & -\frac{11}{18} & -\frac{5}{18} \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Thus,

$$\begin{aligned}\lim_{n\to\infty} P^n = L &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{3} & -1 & \frac{1}{6} & \frac{1}{66} \\ \frac{20}{9} & \frac{8}{3} & -\frac{11}{18} & -\frac{5}{18} \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 5 & -5 & -3 & 3 \\ -3 & 11 & 3 & 3 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{3} & -1 & 0 & 0 \\ \frac{20}{9} & \frac{8}{3} & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 5 & -5 & -3 & 3 \\ -3 & 11 & 3 & 3 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{3} & 0 & 0 & -1 \\ \frac{20}{9} & 0 & 0 & \frac{8}{3} \\ 0 & 0 & 0 & 1 \end{bmatrix}.\end{aligned}$$

(b) Thus, the status of the patients over the long run is

$$\begin{aligned}X^{(\infty)} &= X^{(0)}L \\ &= \begin{bmatrix} 0 & 0.7 & 0.3 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{3} & 0 & 0 & -1 \\ \frac{20}{9} & 0 & 0 & \frac{8}{3} \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{13}{30} & 0 & 0 & \frac{1}{10} \end{bmatrix} = \begin{bmatrix} 0.43 & 0 & 0 & 0.1 \end{bmatrix}.\end{aligned}$$

Therefore, over the long run, there are 43% discharged patients and 10% patients die. No ambulatory or bedridden patients remain in the geriatric ward.

■

Example 10.1.13. The training programme for production supervisors at a particular company consists of two phases. Phase 1, which involves 3 weeks of classroom work, is followed by Phase 2, which is a 3 week apprenticeship program under the direction of working supervisors. From past experience, the company expects only 60% of those beginning classroom training to be graduated into the apprenticeship phase, with the remaining 40% dropped completely from the training program. Of those who make it to the apprenticeship phase, 70% are graduated as supervisors, 10% are asked to repeat the second phase, and 20% are dropped completely from the program. How many supervisors can the company expect from its current training programme if it has 45 people in the classroom phase and 21 people in the apprenticeship phase?

Solution. We consider one time period to be 3 weeks and define states 1 through 4 as the conditions of being dropped, a classroom trainee, an apprentice, and a supervisor, respectively. If we assume that discharged individuals never re-enter the training programme and that supervisors remain supervisors, then the transition probabilities are given by the Markov chain

$$P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.4 & 0 & 0.6 & 0 \\ 0.2 & 0 & 0.1 & 0.7 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

. Since there are $45 + 21 = 66$ people in the training programme currently, so the initial probability vector is given by

$$X^{(0)} = \left[0,\ \frac{45}{66},\ \frac{21}{66},\ 0\right].$$

We have from example 10.1.6,

$$\lim_{n\to\infty} P^n = L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{8}{15} & 0 & 0 & \frac{7}{15} \\ \frac{2}{9} & 0 & 0 & \frac{7}{9} \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

$$\begin{aligned} X^{(\infty)} &= X^{(0)}L \\ &= \begin{bmatrix} 0 & \frac{45}{66} & \frac{21}{66} & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{8}{15} & 0 & 0 & \frac{7}{15} \\ \frac{2}{9} & 0 & 0 & \frac{7}{9} \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 0.4343 & 0 & 0 & 0.5657 \end{bmatrix}. \end{aligned}$$

Eventually, 43.43% of those currently in training (or about 29 people) will be dropped from the programme and 56.67% (or about 37 people) will become supervisors. ■

Example 10.1.14. Solve the previous problem if all 66 people are currently in the classroom phase of training programme.

Solution. Here, $X^{(0)} = [0\ \ 1\ \ 0\ \ 0]$. Thus,

$$\begin{aligned} X^{(\infty)} &= X^{(0)}L \\ &= \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{8}{15} & 0 & 0 & \frac{7}{15} \\ \frac{2}{9} & 0 & 0 & \frac{7}{9} \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{8}{15} & 0 & 0 & \frac{7}{15} \end{bmatrix}. \end{aligned}$$

Thus, $\frac{8}{15} \times 66 \simeq 35$ people will ultimately drop from the program and the remaining $66 - 35 = 31$ people eventually become supervisors. ■

Unit 11

Course Structure

- Geometric programming
- General form of GP (Unconstrained GP)(Primal Problem)

11.1 Geometric Programming

We shall focus our attention on a rather interesting technique called *Geometric Programming* for solving a special type of non-linear programming problem. This technique is initially derived from inequalities rather than the calculus and its extension. This technique was given the name geometric programming because the geometric arithmetic mean inequality was the basis of its development. The advantage here is that it is usually much simpler to work with the dual problem than the primal problem. Geometric programming derives its name from the fact that it is based on the certain geometric concept such as orthogonality and arithmetic geometric inequality. It was developed in early 1960's by Duffin, Peterson and Zener for solving the class of optimization problem that involve special type of functions called posynomial (positive+ polynomial).

A real expression of the form

$$C_j \prod_{i=1}^{n} (x_i)^{a_{ij}}$$

where c_j, a_{ij} are real and $X = (x_1, x_2, \ldots, x_m)^T > 0$ is called monomial in X.

Example: $5.7x_1^3 x_2 - 4x_3^{2.5}$ is a monomial.

Posynomial and Signomial: A generalised polynomial that consist of a finite number of monomials such as

$$f(x) = \sum_{j=1}^{n} C_j \prod_{i=1}^{m} (x_i)^{a_{ij}}$$

is said to be posynomial if all the coefficients C_j are positive; is called the signomial if the coefficients C_j are negative.

The G.P approach instead of solving a non-linear programming problem first finds the optimal value of the objective function by solving its dual problem and then determines an optimal solution to the given NLPP from the optimal solution of the dual.

11.1.1 General form of G.P (Unconstrained G.P) (Primal Problem)

$$\begin{aligned}&\min f(x) = \sum_{j=1}^{n} c_j u_j(x)\\ &\text{such that } \; x_i \geq 0 \quad \text{with } \; c_j > 0\\ &\text{and } \; u_j(x) = \prod_{i=1}^{n} (x_j)^{a_{ij}},\end{aligned}$$

where a_{ij} may be any real number.

11.1.2 Necessary conditions for optimality

The necessary conditions for optimality can be obtained by taking partial derivatives with respect to each x_r and equating the result with 0. Thus

$$\frac{\partial f(x)}{\partial x_r} = \sum_{j=1}^{n} c_j \frac{\partial u_j(x)}{\partial x_r} = 0$$

But,

$$\frac{\partial}{\partial x_r} u_j(x) = \frac{a_{rj}}{x_r} u_j(x).$$

Putting this result in the previous equation, we get,

$$\frac{\partial f(x)}{\partial x_r} = \frac{1}{x_r} \sum_{j=1}^{n} a_{rj} c_j u_j(x) = 0.$$

Let, $f^*(x)$ be the minimum value of $f(x)$. Since, each x_r and c_j is positive, therefore $f^*(x)$ will also be positive. Defining $\dfrac{\partial f(x)}{\partial x_r}$ by $f^*(x)$ we get,

$$\sum_{j=1}^{n} \frac{a_{rj} c_j u_j(x)}{f^*(x)} = 0.$$

Now, we take a simple transformation of variable as

$$y_j = \frac{c_j u_j(x)}{f^*(x)}, \quad j = 1, 2, \ldots, n.$$

Using this transformation, the necessary conditions for local minimum becomes,

$$\sum_{j=1}^{n} a_{rj} y_j = 0; \quad r = 1, 2, \ldots, m. \tag{11.1.1}$$

Thus, due to the definition of y_j, we obtain

$$\sum_{j=1}^{n} y_j = \frac{1}{f^*(x)} \sum_{j=1}^{n} c_j u_j(x) = 1. \tag{11.1.2}$$

At the optimal solution, conditions (11.1.1) and (11.1.2) are the necessary conditions for optimality of non-linear function and also known as orthogonality and normality conditions respectively. This condition give a unique value of y_j for $m+1=n$ and all equations are independent but for $n > (m+1)$, the value of y_j no longer remains independent.

[Degree of G.P difficulty (D.D) of G.P is equal to number of terms in G.P -(1 + number of variables in G.P]

$$\therefore D.D = n - (m+1), \quad (> 0 \quad \text{infinite solution}).$$

Conditions (11.1.1) and (11.1.2) can be expressed as

$$AY = b,$$

where

$$A = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Thus, we require to form the normality and orthogonality condition $AY = B$. This means that the original NLP problem is reduced to one of finding the set of values of Y that satisfy this linear non-homogeneous equation. Hence, to determine the unique value of y_j for the purpose of minimizing effect.

(i) Rank (A, b) > Rank (A), there will be no solution, where (A, b) denote the augmented matrix.

(ii) Rank (A, b) = Rank (A), then a unique solution.

(iii) Rank $(A) < n$, i.e $n > m+1$, that is infinite number of solutions exist.

To find the minimum value of $\mathbf{f(x)}$

At the optimal solution we know that

$$f^*(x) = \frac{c_j u_j(x)}{y_j} = \frac{1}{y_j} c_j \prod_{i=1}^{n} (x_i)^{a_{ij}}$$

Raising both side to power of y_j and taking the product we get,

$$\sum_{j=1}^{n} \left\{ f^*(x) \right\}^{y_j} = \prod_{j=1}^{n} \left\{ \frac{1}{y_j} c_j \prod_{i=1}^{n} (x_i)^{a_{ij}} \right\}^{y_j},$$

Now, since $\sum_{j=1}^{n} y_j = 1$, therefore

$$\prod_{j=1}^{n} \{ f^*(x) \}^{y_j} = \left[f^*(x) \right]^{\sum_{j=1}^{n} y_j} = f^*(x)$$

In R.H.S of the above equation we have

$$\begin{aligned}\prod_{j=1}^{n}\left[\left(\frac{c_j}{y_j}\right)\prod_{i=1}^{m}(x_i)^{a_{ij}}\right]^{y_j} &= \prod_{j=1}^{n}\left(\frac{c_j}{y_j}\right)^{y_j}\prod_{j=1}^{n}\left\{\prod_{i=1}^{m}(x_i)^{a_{ij}}\right\}^{y_j}\\ &= \prod_{j=1}^{n}\left(\frac{c_j}{y_j}\right)^{y_j}\prod_{i=1}^{m}(x_i)^{\sum_{j=1}^{n}a_{ij}y_j}\\ &= \prod_{j=1}^{n}\left(\frac{c_j}{y_j}\right)^{y_j}\prod_{i=1}^{m}(x_i)^{0} \text{ [By Eq. (11.1.1)]}\end{aligned}$$

Thus,

$$\min f(x) = f^*(x) = \prod_{j=1}^{n}\left(\frac{c_j}{y_j}\right)^{y_j} \quad \text{and}$$

$$\therefore f(x) \geq \prod_{j=1}^{n}\left(\frac{c_j}{y_j}\right)^{y_j}.$$

where y_j must satisfy the orthogonality and normality conditions. For the given value of f^* and unique value of y_j, the solution to a set of equations can be obtained from

$$c_j\prod_{i=1}^{m}(x_i)^{a_{ij}} = y_j f^*(x).$$

Dual Problem:

$$\begin{aligned}\max g(y) &= \prod_{j=1}^{n}\left(\frac{c_j}{y_j}\right)^{y_j}\\ \text{subject to} \quad &\sum_{j=1}^{n}a_{ij}y_j = 0\\ \text{and} \quad &\sum_{j=1}^{n}y_j = 1\\ &y_j \geq 0.\end{aligned}$$

Theorem 11.1.1. If x is a feasible solution vector of the unconstraint of a primal geometric programming and y is a feasible solution vector for DP (Dual problem), then

$$f(x) \geq g(y). \qquad \text{(Primal Dual inequality)}$$

Proof. The expression for $f(x)$ can be written as

$$f(x) = \sum_{j=1}^{n}\frac{C_j\prod_{i=1}^{m}(x_i)^{a_{ij}}}{y_j}.$$

Here, weights are $y_1, y_2, \ldots, y_n$ and the positive terms are

$$\frac{C_1\prod_{i=1}^{m}(x_i)^{a_{i1}}}{y_1}, \quad \frac{C_2\prod_{i=1}^{m}(x_i)^{a_{i2}}}{y_2}, \ldots, \quad \frac{C_n\prod_{i=1}^{m}(x_i)^{a_{in}}}{y_n}.$$

Now, applying weighted Arithmetic Mean- Geometric mean inequality,

$$\left(\frac{\dfrac{C_1\prod_{i=1}^{m}(x_i)^{a_{i1}}}{y_1}\cdot y_1+\dfrac{C_2\prod_{i=1}^{m}(x_i)^{a_{i2}}}{y_2}\cdot y_2+\cdots+\dfrac{C_n\prod_{i=1}^{m}(x_i)^{a_{in}}}{y_n}\cdot y_n}{y_1+y_2+\cdots+y_n}\right)^{y_1+y_2+\cdots+y_n}$$

$$\geq\left(\frac{C_1\prod_{i=1}^{m}(x_i)^{a_{i1}}}{y_1}\right)^{y_1}\cdot\left(\frac{C_2\prod_{i=1}^{m}(x_i)^{a_{i2}}}{y_2}\right)^{y_2}\cdots\left(\frac{C_n\prod_{i=1}^{m}(x_i)^{a_{in}}}{y_n}\right)^{y_n}$$

$$\text{or,}\quad f(x)\geq\prod_{j=1}^{n}\left(\frac{C_j\prod_{i=1}^{m}(x_i)^{a_{ij}}}{y_j}\right)^{y_j}\qquad[\text{since}\quad y_1+y_2+\cdots+y_n=1\ \text{ for normality condition}]$$

$$\text{or,}\quad f(x)\geq\prod_{j=1}^{n}\left(\frac{C_j}{y_j}\right)^{y_j}\prod_{i=1}^{m}(x_i)^{\sum_{j=1}^{n}a_{ij}y_j}\tag{11.1.3}$$

$$\text{or,}\quad f(x)\geq\prod_{j=1}^{n}\left(\frac{C_j}{y_j}\right)^{y_j}\qquad\left[\sum_{i=1}^{n}a_{ij}y_j=0,\ \ \text{orthogonality condition}\right]$$

$$\text{or,}\quad f(x)\geq g(y).$$

For constraint, after above

$$g_i(x)=\sum_{r=1}^{P(i)}y_{ir}\left(\frac{C_{ir}\prod_{i=1}^{n}(x_i)^{a_{irj}}}{y_{ir}}\right)$$

Applying weighted arithmetic mean geometric mean inequality, we have

$$\left(\frac{g_i(x)}{\sum_{r=1}^{P(i)}y_{ir}}\right)^{\sum_{r=1}^{P(i)}y_{ir}}\geq\prod_{i=1}^{m}\prod_{r=1}^{P(i)}\left(\frac{C_{ir}\prod_{i=1}^{n}(x_i)^{a_{irj}}}{y_{ir}}\right)^{y_{ir}}$$

$$(g_i(x))^{\sum_{r=1}^{P(i)}y_{ir}}\geq\prod_{i=1}^{m}\prod_{r=1}^{P(i)}\left(\frac{C_{ir}}{y_{ir}}\right)^{y_{ir}}\prod_{i=1}^{n}(x_i)^{\sum_{r=1}^{P(i)}a_{irj}y_{ir}}\left(\sum_{r=1}^{P(i)}y_{ir}\right)^{y_{ir}}.$$

Since $g_i(x)\leq 1$ (constraint), so,

$$1\geq(g_i(x))^{\sum_{r=1}^{P(i)}y_{ir}}.$$

Hence,

$$1 \geq \prod_{i=1}^{m}\prod_{r=1}^{P(i)}\left(\frac{C_{ir}}{y_{ir}}\right)^{y_{ir}} \prod_{i=1}^{n}(x_i)^{\sum_{r=1}^{P(i)} a_{irj}y_{ir}} \left(\sum_{r=1}^{P(i)} y_{ir}\right)^{y_{ir}}. \tag{11.1.4}$$

Multiplying (11.1.3) and (11.1.4), we have

$$f(x) \geq \prod_{j=1}^{n}\left(\frac{C_j}{y_j}\right)^{y_j} \prod_{1=1}^{m}\left[\prod_{r=1}^{P(i)}\left(\frac{C_{ir}}{y_{ir}}\right)^{y_{ir}} \left(\sum_{r=1}^{P(i)} y_{ir}\right)^{y_{ir}}\right] (x_i)^{\sum_{i=1}^{n} a_{ij}y_j + \sum_{i=1}^{m}\sum_{r=1}^{P(i)} a_{irj}y_{ir}}.$$

Using orthogonality condition,

$$\sum_{i=1}^{n} a_{ij}y_j + \sum_{i=1}^{m}\sum_{r=1}^{P(i)} a_{irj}y_{ir} = 0.$$

Thus, we have,

$$f(x) \geq \prod_{j=1}^{n}\left(\frac{C_j}{y_j}\right)^{y_j} \prod_{1=1}^{m}\left[\prod_{r=1}^{P(i)}\left(\frac{C_{ir}}{y_{ir}}\right)^{y_{ir}} \left(\sum_{r=1}^{P(i)} y_{ir}\right)^{y_{ir}}\right]$$

or, $\quad f(x) \geq g(y).$

□

Example 11.1.2. Solve the following NLPP by geometric programming technique.

$$\min z = 7x_1x_2^{-1} + 3x_2x_3^{-2} + 5x_1^{-3}x_2x_3 + x_1x_2x_3$$
$$x_1, x_2, x_3 \geq 0$$

Solution.

$$A = \begin{bmatrix} 1 & 0 & -3 & 1 \\ -1 & 1 & 1 & 1 \\ 0 & -2 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

and we get $AY = b$ with

$$y_1 = \frac{1}{2},\ y_2 = \frac{1}{6},\ y_3 = \frac{5}{24},\ y_4 = \frac{3}{24},\ f^*(x) = \frac{761}{50}$$
$$x_1^* = 1.315,\ x_2^* = 1.21,\ x_3^* = 1.2$$

Now $AY = b$ gives

$$\begin{bmatrix} 1 & 0 & -3 & 1 \\ -1 & 1 & 1 & 1 \\ 0 & -2 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

which leads to the following system of equations

$$y_1 - 3y_3 + y_4 = 0 \tag{11.1.5}$$
$$-y_1 + y_2 + y_3 + y_4 = 0 \tag{11.1.6}$$
$$-2y_2 + y_3 + y_4 = 0 \tag{11.1.7}$$
$$y_1 + y_2 + y_3 + y_4 = 1 \tag{11.1.8}$$

Now, (11.1.6)-(11.1.8) gives

$$-y_1 + y_2 + y_3 + y_4 - y_1 - y_2 - y_3 - y_4 = -1$$
$$\Rightarrow -2y_1 = -1 \Rightarrow y_1 = \frac{1}{2}$$

Now, (11.1.6)-(11.1.7) gives

$$-y_1 + y_2 + y_3 + y_4 + 2y_2 - y_3 - y_4 = 0$$
$$\Rightarrow -y_1 + 3y_2 = 0 \Rightarrow 3y_2 = y_1$$
$$\Rightarrow 3y_2 = \frac{1}{2} \Rightarrow y_2 = \frac{1}{6}.$$

Now, (11.1.5)-(11.1.7) gives

$$y_1 - 3y_3 + y_4 + 2y_2 - y_3 - y_4 = 0$$
$$\Rightarrow y_1 + 2y_2 - 4y_3 = 0 \Rightarrow 4y_3 = y_1 + 2y_2$$
$$\Rightarrow 4y_3 = \frac{1}{2} + \frac{1}{3} \Rightarrow 4y_3 = \frac{5}{6} \Rightarrow y_3 = \frac{5}{24}.$$

Now,

$$\begin{aligned} y_4 &= 1 - (y_1 + y_2 + y_3) \\ &= 1 - \left(\frac{1}{2} + \frac{1}{6} + \frac{5}{24}\right) \\ &= 1 - \frac{12 + 4 + 5}{24} \\ &= 1 - \frac{21}{24} \\ &= \frac{3}{24} \end{aligned}$$

$$\therefore y_1 = \frac{1}{2}, \quad y_2 = \frac{1}{6}, \quad y_3 = \frac{5}{24}, \quad y_4 = \frac{3}{24}$$

$$\begin{aligned} f^*(x) &= \left(\frac{7}{1/2}\right)^{1/2} \times \left(\frac{3}{1/6}\right)^{1/6} \times \left(\frac{5}{5/24}\right)^{5/24} \times \left(\frac{1}{3/24}\right)^{3/24} \\ &= (14)^{1/2} \times (18)^{1/6} \times (24)^{5/24} \times (8)^{3/24} \\ &= 3.74 \times 1.62 \times 1.94 \times 1.297 \\ &= 15.245 \\ &= \frac{761}{50} \end{aligned}$$

Now

$$c_j \prod_{i=1}^{m} (x_i)^{a_{ij}} = y_j f^*(x)$$

$$\therefore 7x_1 x_2^{-1} = \frac{1}{2} \times \frac{761}{50}$$

$$\Rightarrow x_1 x_2^{-1} = \frac{761}{700} \tag{11.1.9}$$

$$\text{and} \quad 3x_2 x_3^{-2} = \frac{1}{6} \times \frac{761}{50}$$

$$\Rightarrow x_2 x_3^{-2} = \frac{761}{900}$$

$$5x_1^{-3} x_2 x_3 = \frac{5}{24} \times \frac{761}{50}$$

$$\Rightarrow x_1^{-3} x_2 x_3 = \frac{761}{1200} \tag{11.1.10}$$

$$\text{and} \quad x_1 x_2 x_3 = \frac{3}{24} \times \frac{761}{50}$$

$$\Rightarrow x_1 x_2 x_3 = \frac{761}{400} \tag{11.1.11}$$

Now (11.1.10) and (11.1.11) gives

$$\frac{x_1^{-3} x_2 x_3}{x_1 x_2 x_3} = \frac{761/1200}{761/400}$$

$$\Rightarrow x_1^{-4} = \frac{1}{3}$$

$$x_1 = \left(\frac{1}{3}\right)^{-1/4} = 3^{1/4} = 1.316.$$

$$\therefore x_1^* = 1.316$$

Now from, (11.1.9) we get

$$x_1 x_2^{-1} = \frac{761}{700}$$

$$\Rightarrow x_2^{-1} = \frac{761}{700} \times \frac{1}{x_1}$$

$$\Rightarrow x_2 = \frac{700}{761} \times x_1$$

$$\Rightarrow x_2 = \frac{700}{761} \times 1.3616$$

$$\Rightarrow x_2 = 1.21$$

$$\therefore x_2^* = 1.21$$

Now, from (11.1.9) we get

$$x_2 x_3^{-2} = \frac{761}{900}$$

$$x_3^2 = \frac{900}{761} x_2$$

$$x_3 = \sqrt{\frac{900}{761}} \times \sqrt{1.21} = 1.2$$

$$\therefore x_3^* = 1.2$$

■

Example 11.1.3. Solve the following NLPP by the geometric programming.

$$\min f(x) = 5x_1x_2^{-1} + 2x_1^{-1}x_2 + 5x_1 + x_2^{-1}; \quad x_1, x_2 \geq 0$$

Solution. The given function may be written as

$$f(x) = 5x_1x_2^{-1} + 2x_1^{-1}x_2 + 5x_1^1x_2^0 + x_1^0x_2^{-1}.$$
$$(c_1, c_2, c_3, c_4) = (5, 2, 5, 1)$$

The orthogonality and normality conditions are given by

$$\begin{bmatrix} 1 & -1 & 1 & 0 \\ -1 & 1 & 0 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Since $n > m + 1$, this equations do not give y_j directly. Solving for y_1, y_2 and y_3 in terms of y_4 we get,

$$\begin{bmatrix} 1 & -1 & 1 \\ -1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 0 \\ y_4 \\ 1 - y_4 \end{bmatrix}$$

$$\text{or } \; y_1 = (1 - 3y_4)/2 = 0.8(1 - 3y_4); \quad y_2 = 0.5(1 - y_4); \quad y_3 = y_4.$$

The corresponding dual problem may be written as

$$\max f(y) = \left[\frac{5}{0.5(1-3y_4)}\right]^{0.5(1-3y_4)} \left[\frac{2}{0.5(1-y_4)}\right]^{0.5(1-3y_4)} \left[\frac{5}{y_4}\right]^{y_4} \left[\frac{1}{y_4}\right]^{y_4}$$

Since, maximization of $f(y)$ is equivalent to $\log f(y)$, taking $\log$ both sides we have

$$\begin{aligned} \log f(y) = {} & 0.5(1 - 3y_4)\{\log 10 - \log(1 - 3y_4)\} + 0.5(1 - y_4)\{\log 4 - \log(1 - y_4)\} \\ & + y_5(\log 5 - \log y_4) + y_4\{\log 1 - \log y_4\} \end{aligned} \tag{11.1.12}$$

The value of y_4 maximizing $\log f(y)$ must be unique, because the primal problem has a unique minimum. Differentiating (11.1.12) with respect to y_4 and equating to zero, we have

$$\begin{aligned} \frac{\partial}{\partial y_4} f(y) \quad = \quad & -\frac{3}{2}\log 10 - \left\{-\frac{3}{2} + \left(-\frac{3}{2}\right)\log(1 - 3y_4)\right\} \\ & -\frac{1}{2}\log 4 - \left\{-\frac{1}{2} + \left(-\frac{1}{2}\right)\log(1 - y_4)\right\} \\ & + \log 5 - \{1 + \log y_4\} + \log 1 - \{1 + \log y_4\} = 0 \end{aligned}$$

Then after simplification, we have

$$-\log\left\{\frac{2 \times 10^{3/2}}{5}\right\} + \log\left\{\frac{(1 - 3y_4)^{3/2}(1 - y_4)^{1/2}}{y_4^2}\right\} = 0.$$
$$\Rightarrow \frac{\sqrt{(1 - 3y_4)^3(1 - y_4)}}{y_4^2} = 12.6$$

After solving we have $y_4^* = 0.16$. Hence

$$y_1^* = 0.26, \quad y_2^* = 0.42, \quad y_3^* = 0.16$$

$$\begin{aligned} \text{The value of } f^*(y) &= f^*(x) \\ &= \left(\frac{5}{0.26}\right)^{0.26} \left(\frac{2}{0.42}\right)^{0.42} \left(\frac{5}{0.16}\right)^{0.16} \left(\frac{1}{0.16}\right)^{0.160} \\ &= 9.661 \end{aligned}$$

$$u_1 = y_1^* f^*(x), \quad u_2 = y_2 f^*(x), \quad u_3 = y_3 f^*(x), \quad u_4 = y_4 f^*(x)$$

$$\begin{aligned} 5x_1 &= 0.16 \times 9.661 \\ \Rightarrow x_1^* &= \frac{0.16 \times 9.661}{5} = 0.309 \end{aligned}$$

and

$$\begin{aligned} x_2^{-1} &= 0.42 \times 9.661 \\ \Rightarrow x_2^* &= \frac{1}{0.16 \times 9.661} = 0.647 \end{aligned}$$

■

Unit 12

Course Structure

- Constraint Geometric Programming Problem

12.1 Constraint Geometric Programming Problem

$$\min z = f(x)$$
$$\text{such that } g_i(x) = \sum_{r=1}^{P(i)} c_{ij} u_{ir}(x) = 1, \quad i = 1, 2, \ldots, M.$$

where $P(i)$ denotes the number of terms in the i-th constraint and $u_{ir}(x) = \prod_{j=1}^{n} (x_i)^{a_{irj}}$.

Forming Lagrange function to obtain normality and orthogonality condition,

$$F(x, \lambda) = f(x) + \sum_{i=1}^{M} \lambda_i [g_i(x) - 1]$$

and require the conditions,

(i) $\dfrac{\partial F}{\partial x_t} = \dfrac{\partial f(x)}{\partial x_t} + \sum_{i=1}^{M} \lambda_i \dfrac{\partial g_i(x)}{\partial x_t} = 0.$

(ii) $\dfrac{\partial F}{\partial \lambda_i} = g_i(x) - 1 = 0;\ i = 1, 2, \ldots, M.$

So, long as right hand side in the second constraint $g_i(x) = 1$, it can be obtained in this form by simple transformation. However, $g_i(x) = 0$ is not admissible because solution space required $x > 0$. Considering once again condition (i), we have

$$\frac{\partial F}{\partial x_t} = \sum_{j=1}^{n} \frac{c_j a_{tj} c_j(x)}{x_t} + \sum_{i=1}^{M} \lambda_i \left[\sum_{r=1}^{P(i)} \frac{c_{ir} a_{irt} u_{ir}(x)}{x_t} \right].$$

Introducing variables y_j for objective and y_{ir} for constraints as follows:

$$y_j = \frac{c_j u_j(x)}{f^*(x)} \quad \text{and} \quad y_{ir} = \frac{\lambda_i c_{ir} u_{ir}(x)}{f^*(x)}$$

By substituting the values of y_j and y_{ir} in $\dfrac{\partial F}{\partial x_t} = 0$, we obtain the orthogonality conditions and normality condition as

$$\sum_{j=1}^{n} a_{tj} y_j + \sum_{i=1}^{M} \sum_{r=1}^{P(i)} a_{irt} y_{ir} = 0; \;\; t = 1, 2, \ldots, n. \quad \text{(Orthogonality Conditions)}$$

$$\sum_{j=1}^{n} y_j = 1 \quad \text{(Normality Condition)}$$

We have seen in earlier discussion that y_j were all positive, because $y_j = \dfrac{c_j u_j(x)}{f^*(x)} > 0$. However, in the equality constraint case, y_j are again positive. But, y_{ir} may be negative because λ_i need not be non-negative. To formulate a dual function it is desirable to all $y_{ir} > 0$. But if one of the y_{ir} is negative, then its sign can be reversed by writing the term in the Lagrange function as $\lambda_q\{1 - g_q(x)\}$. Once again normality and orthogonality conditions can be derived by solving a system of linear equations

$$\sum_{j=1}^{n} a_{tj} y_j$$

When these equations have a unique solution, the optimal of the original problem can be obtained from the definition of y_j and y_{ir} in terms of $f^*(x)$ and x. In case, these equations have an infinite number of solution, we tend to maximize the dual function given by

$$\max f(y) = \prod_{j=1}^{n} \left(\frac{c_j}{y_j}\right)^{y_j} \prod_{i=1}^{M} \left[\prod_{r=1}^{P(i)} \left(\frac{c_{rj}}{y_{ij}}\right)^{y_{rj}}\right] \prod_{i=1}^{M} (v_i)$$

where $v_i = \sum_{r=1}^{P(i)} y_{ir}$ such that the orthogonality and normality constraints.

In the above functions the constraints are linear and therefore it is easy to obtain the optimal solution. Moreover, we may also work with log of the dual function which is linear in the variable $\delta_i = \log y_j$ and $\delta_{ir} = \log y_{ir}$.

Example 12.1.1. Solve the following NLPP by G.P.

$$\min f(x) = 2x_1 x_2^{-3} + 4x_1^{-1} x_2^{-2} + \frac{32}{3} x_1 x_2$$
$$\text{such that} \quad x_1^{-1} x_2^2 = 0$$
$$x_1, x_2 \geq 0.$$

Solution. Given problem derive as

$$\min f(x) = 2x_1 x_2^{-3} + 4x_1^{-1} x_2^{-2} + \frac{32}{3} x_1 x_2$$
$$\text{such that } \; 0.1 x_1^{-1} x_2^2 = 1$$
$$x_1, x_2 \geq 0.$$

Dual problem:

$$\max f(y) = \left(\frac{2}{y_1}\right)^{y_1}\left(\frac{4}{y_2}\right)^{y_2}\left(\frac{32}{3y_3}\right)^{y_3}\left(\frac{0.1}{y_4}\right)^{y_4}(y_4)^{y_4}$$

such that

$$y_1 + y_2 + y_3 = 1$$
$$y_1 - y_2 + y_3 - y_4 = 0$$
$$-3y_1 - 2y_2 + y_3 + 2y_4 = 0$$

Expressing each of the variable in the objective function in terms of y_1, we get

$$\max f(y_1) = \left(\frac{2}{y_1}\right)^{y_1}\left(\frac{4}{1-\frac{4}{3}y_1}\right)^{1-\frac{4}{3}y_1}\left(\frac{32}{y_1}\right)^{\frac{1}{3}y_1}(0.1)^{\frac{8}{3}y_1-1}$$

where

$$y_2 = 1 - \frac{4}{3}y_1$$
$$y_3 = \frac{y_1}{3}$$
$$y_4 = \frac{8}{3}y_1 - 1$$

Taking $\log$ both sides of $f(y_1)$ and differentiating with respect to y_1, we have,

$$\begin{aligned} F(y_1) &= \log f(y_1) \\ &= y_1 \log\left(\frac{2}{y_1}\right) + \left\{1 - \left(\frac{4}{3}\right)y_1\right\}\log 4 - \log\left(1 - \frac{4}{3}y_1\right) \\ &\quad \frac{y_1}{3}\{\log 32 - \log y_1\} + \left(\frac{8}{3}y_1 - 1\right)\log(0.1) \end{aligned}$$

Now,

$$\frac{dF}{dy_1} = \log\left(\frac{2}{y_1}\right) + 2 - \left(\frac{16}{3}\right)y_1 + \log\left(\frac{32}{y_1}\right) + \frac{8}{3}\log(0.1) = 0$$
$$\Rightarrow y_1 = 0.662$$

The values of the other variables are

$$y_1 = 0.662, \quad y_2 = 0.217, \quad y_3 = 0.221, \quad y_4 = 0.766$$

Using the relation $y_j = \dfrac{c_j u_j}{f^*(x)}$ we obtain

$$y_1 = \frac{c_1 u_1}{f^*(x)} = \frac{2x_1 x_2^{-1}}{f^*(x)}$$
$$y_2 = \frac{c_2 u_2}{f^*(x)} = \frac{4x_1^{-1}x_2^{-1}}{f^*(x)}$$
$$y_3 = \frac{c_3 u_3}{f^*(x)} = \frac{32x_1 x_2}{3f^*(x)}$$
$$y_4 = \frac{c_4 u_4}{f^*(x)} = \frac{x_1^{-1}x_2^2}{f^*(x)}$$

■

Exercise 12.1.2. Solve the following NLPP by G.P.

1.

$$\min f(x) = 5x_1x_2^{-1}x_3^2 + x_1^{-2}x_2^{-1} + 10x_2^2 + 2x_1^{-1}x_2x_3^{-2}$$
$$x_1, x_2, x_3 \geq 0$$

Answer: $x_1 = 1.26,\ x_2 = 0.41,\ x_3 = 0.59$ and $\min f(x) = 10.28$

2.

$$\min f(x) = 2x_1 + 4x_2 + \frac{10}{x_1x_2}$$
$$x_1, x_2 \geq 0$$

Answer: $x_1 = 14.1,\ x_2 = 23$ and $\min f(x) = 112.9$

3.

$$\min z = \frac{3x_1}{x_2} + \frac{x_2^2}{x_1} + x_1^2x_2$$
such that
$$\frac{1}{4}x_1^2x_2^{-1} + \frac{1}{9}x_2x_1 = 1$$
$$2\left(\frac{1}{x_1^2}\right) + 4\left(\frac{x_2}{x_1^2}\right) = 2$$
$$x_1, x_2 \geq 0.$$

Unit 13

Course Structure

- Inventory Control/Problem/Model
- The Economic Order Quantity (EOQ) model without shortage

13.1 Inventory Control/Problem/Model

13.1.1 Production Management

In our daily lives, we observe that a small retailer knows roughly the demand of his customers in a month or a week or a day, and accordingly places orders on the wholesaler to meet the demand of his customer. But this is not the case with a manager of a big departmental store or a big retailer because the stocking in such cased depends upon various factors namely demand, time of ordering, lag between orders and actual receives etc. So, the real problem is to have a compromise between over stocking and under stocking. The study of such type of problems is known as material management or production management or inventory control. In broad sense, inventory may be defined as the stock of goods, commodities or other economic resources that are stored or reserved in order to ensure smooth and efficient running of business affairs. The inventory may be kept in any of the following forms:

(i) Raw-material inventory

(ii) Working process inventory

(iii) Finished good inventory

(iv) Inventory also includes furniture, machinery etc.

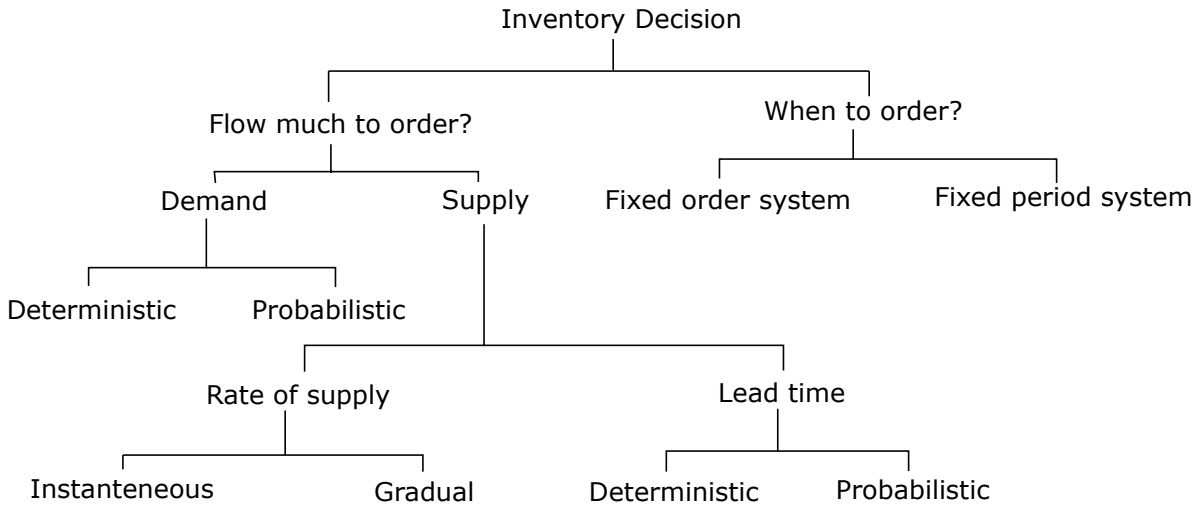
The term inventory may be classified in two main categories, viz.

(1) Direct Inventory

(2) Indirect Inventory

Indirect inventory includes those items which are necessarily required for manufacturing but do not become the component of finished products like oil, grease, lubricants, petrol, office materials, etc.

13.1.2 Inventory Decisions



Lead time: Time between placing an order an actual received.

13.1.3 Inventory related cost:

(1) **Holding Cost (**C_1 **or** C_n**)**: The cost associated with carrying or holding the goods in stock is known as holding cost or carrying cost, which is usually denoted by C_1 per unit of goods per unit time.

(2) **Shortage or stockout cost (**C_2 **or** C_s**)**: The penalty cost which is incurred as a result of running out of stock or shortage is known as shortage or stockout cost. It is usually denoted by C_2 per unit of goods for a specified period. This cost arises due to shortage of goods, sales may be lost, goodwill may be lost and so on.

(3) **Set up or ordering cost (**C_3 **or** C_0**)**: This includes the fixed cost associated with obtaining goods during placing of an order or purchasing or manufacturing or setting up a machinery before starting production. It is usually denoted by C_3 or C_0 per production run (cycle).

13.1.4 Why inventory is maintained?

Mathematically the problem of maintaining the inventory arises due to the fact that if a person decides to have a large stock, his holding cost C_1 increases but his shortage cost C_2 and set up cost C_3 decrease. On the other hand if he has small stock, his holding cost C_1 decreases but shortage cost C_2 and set up cost C_3 increase. Similarly, if he decides to order very frequently, the ordering cost increases when the other cost may decrease. So, it becomes necessary to have a compromise between over stocking and under stocking by making optimum decision by controlling value of some variables.

13.1.5 Variables in Inventory Problems

(i) Controlled variable: $q,\ t$

(ii) Uncontrolled variable: $C_1,\ C_2,\ C_3$, Demand (R), Lead time.

13.1.6 Some Notations

C_1 = Holding cost per quantity per unit time.
C_2 = Shortage cost per quantity per unit time.
C_3 = Set up cost per order.
R = Demand rate.
K = Production rate.
t = Scheduling time period which is variable.
t_p = Prescribed time period.
D = Total demand or annual demand.
q = Quantity already present in the beginning.
L = Lead time.

13.2 The Economic Order Quantity (EOQ) model without shortage

13.2.1 Model I(a): Economic lot size model with uniform demand

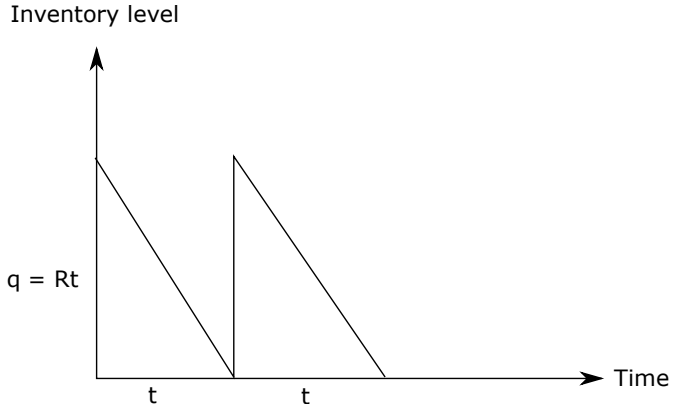
Assumptions:

(i) Demand is uniform at a rate R quantity units per unit time.

(ii) Lead time is zero.

(iii) Production rate is infinite, i.e., instantaneous.

(iv) Shortages are not allowed.



Let each production cycle be made at fixed interval t and therefore the quantity q already present in the beginning should be

$$q = Rt, \tag{13.2.1}$$

where R is a demand rate. Since, the stock in small time dt is $Rt\ dt$, therefore, the stock in total time t will be

$$\int_0^t R\, t\, dt = \frac{1}{2}Rt^2 = \frac{1}{2}qt.$$

Thus,

$$\text{The cost of holding inventory per production run} = C_1\frac{1}{2}qt = C_1\frac{1}{2}Rt^2 \tag{13.2.2}$$

$$\text{The set up cost} = C_3 \text{ per production run for interval } t.$$

$$\text{Total cost} = \frac{1}{2}C_1Rt^2 + C_3 \tag{13.2.3}$$

Therefore, total average cost is given by

$$C(t) = \frac{\frac{1}{2}C_1Rt^2 + C_3}{t} = \frac{1}{2}C_1Rt + \frac{C_3}{t} \text{ (Cost Equation)} \tag{13.2.4}$$

The condition of minimum or maximum of $C(t)$,

$$\begin{aligned} &\frac{d}{dt}\Big[C(t)\Big] = 0 \\ \Rightarrow\ &\frac{1}{2}C_1R - \frac{C_3}{t^2} = 0 \\ \Rightarrow\ &t^* = \sqrt{\frac{2C_3}{C_1R}} \end{aligned} \tag{13.2.5}$$

Also, $\frac{d^2}{dt^2}C(t) = \frac{2C_3}{t^3}$, which is obviously positive for the value of t^*. Hence, $C(t)$ is minimum for optimum time interval t^* and optimum quantity to be produced or ordered at each interval t^* is given by

$$q^* = Rt^* = R\sqrt{\frac{2C_3}{C_1R}} = \sqrt{\frac{2C_3R}{C_1}} \tag{13.2.6}$$

which is called optimal lost size formula and the corresponding minimum cost

$$\begin{aligned} C^*_{\min} &= \frac{1}{2}RC_1\sqrt{\frac{2C_3}{C_1R}} + C_3\sqrt{\frac{C_1R}{2C_3}} \\ &= \sqrt{\frac{C_1C_3R}{2}} + \sqrt{\frac{C_1C_3R}{2}} \\ &= \sqrt{2C_1C_3R} \quad \text{per unit time.} \end{aligned}$$

Note 13.2.1. The cost equation (13.2.4) can also be written as

$$C(q) = \frac{1}{2}C_1q + C_3\frac{R}{q} \quad \text{where} \quad q = Rt.$$

13.2.2 Model I(b): Economic lot size with different rates of demand in different cycles

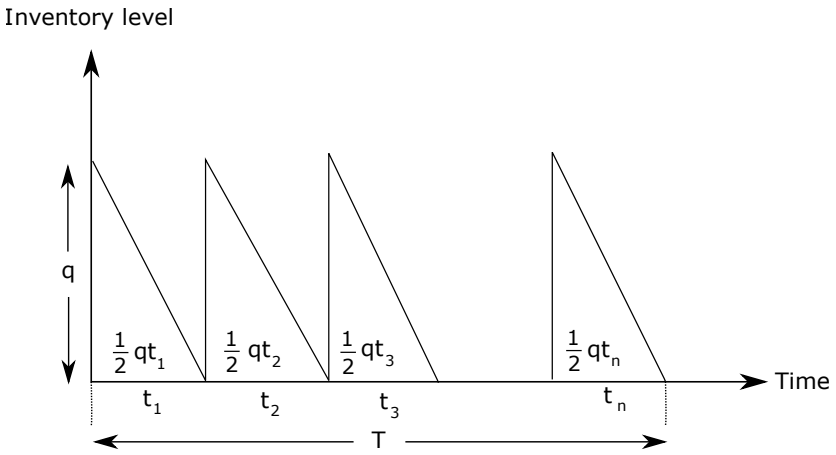
In model I(a), the total demand D is prescribed over the total period T instead of demand rate being constant for each production cycle, that is rate of demand being different in different production cycles.

Let q be the fixed quantity produced in each production cycle. Since, D is the total demand prescribed over the time period T, the number of production cycle will be $n = D/q$. Also, let the total time period $T = t_1 + t_2 + t_3 + \cdots + t_n$. Obviously, the carrying cost for the period T will be

$$\left(\frac{1}{2}qt_1\right)C_1 + \left(\frac{1}{2}qt_2\right)C_1 + \cdots + \left(\frac{1}{2}qt_n\right)C_1 = \frac{1}{2}C_1q(t_1 + t_2 + \cdots + t_n) = \frac{1}{2}C_1qT$$

Set up cost will be equal to $\frac{D}{q}C_3$. Thus, we obtain the cost equation for period T.

$$C(q) = \frac{1}{2}C_1qT + \frac{D}{q}C_3$$



For minimum cost

$$\begin{aligned} &\frac{dC(q)}{dq} = 0 \\ \Rightarrow\ &\frac{1}{2}C_1T - \frac{C_3}{q^2}D = 0 \\ \Rightarrow\ &q^* = \sqrt{\frac{2C_3\left(\frac{D}{T}\right)}{C_1}} \end{aligned}$$

Also, $\frac{d^2C}{dq^2} = \frac{2C_3D}{q^3} > 0$, which minimizes the total cost $C(q)$ and the corresponding minimum value will be

$$\begin{aligned} C_{\min} &= \frac{1}{2}C_1T\sqrt{\frac{2C_3\left(\frac{D}{T}\right)}{C_1}} + C_3D\sqrt{\frac{C_1}{2C_3\left(\frac{D}{T}\right)}} \\ &= \sqrt{\frac{C_1C_3TD}{2}} + \sqrt{\frac{C_1C_3TD}{2}} \\ &= \sqrt{2C_1C_3DT} \end{aligned}$$

Hence, the minimum total average cost will be

$$\begin{aligned} C_{\min} &= \frac{\sqrt{2C_1C_3DT}}{T} \\ &= \sqrt{\frac{2C_1C_3D}{T}} \end{aligned}$$

Note 13.2.2. Here we observed that the fixed demand rate R in model I(a) is replaced by the average demand rate D/T.

Example 13.2.3. You have to supply your customer 100 units of a certain product every Monday. You obtained the product from a local supplier at Rs. 60 per unit. The cost of ordering and transportation from the supplier is Rs. 150 per order. The cost of carrying inventory is estimated at 15% per year of the cost of the product carried.

(i) Describe graphically the inventory system.

(ii) Find the lot size which will minimize the cost of the system.

(iii) How frequently should order be placed?

(iv) Determine the number of orders.

(v) Determine the optimum cost.

Solution. Here

$$\begin{aligned} R &= 100 \text{ units/week.} \\ C_3 &= 150 \text{ per order.} \\ C_1 &= \text{Rs. } \frac{15 \times 60}{100 \times 52} \text{ per unit per week} \\ &= \text{Rs. } \frac{9}{52} \end{aligned}$$

(i)

$$C(t) = 60R + \frac{1}{2}C_1 Rt + \frac{C_3}{t}.$$

(ii)

$$\begin{aligned} q^* &= \sqrt{\frac{2C_3R}{C_1}} \\ &= \sqrt{\frac{2 \times 150 \times 100 \times 52}{9}}. \\ &= 416 \text{ units} \end{aligned}$$

(iii)

$$t^* = \frac{q^*}{R} = \frac{416}{100} = 4.16 \text{ weeks}$$

(iv)

$$\eta = \frac{R}{q^*} = \frac{100}{416} \text{ orders per week}$$

(v)

$$\begin{aligned} C_{\min} &= 60R + \sqrt{2C_1C_3R} \\ &= (60 \times 100) + \sqrt{2 \times \frac{9}{52} \times 150 \times 100} \\ &= 6000 + 72 \\ &= \text{Rs. } 6072 \end{aligned}$$

■

Example 13.2.4. An aircraft company uses rebate at an approximate customer rate of 2500 kg per year. Each unit costs Rs. 30 per kg and the company personal estimate that it cost Rs. 130 to place an order and that the carrying cost of inventory is 10% per year. How frequently should orders be placed? Also determine the optimum size of each order.

Solution. Here

$$\begin{aligned} R &= 2500 \text{ kg per year.} \\ C_3 &= \text{Rs. } 130 \\ C_1 &= \text{Cost of each unit } \times \text{ inventory carrying cost} \\ &= \text{Rs. } 30 \times \frac{1}{30} \\ &= \text{Rs. } 3 \text{ per unit per year} \end{aligned}$$

$$\begin{aligned} q^* &= \sqrt{\frac{2C_3R}{C_1}} \\ &= \sqrt{\frac{2 \times 130 \times 2500}{3}} \\ &= 466 \text{ units} \end{aligned}$$

$$\therefore t^* = \frac{q^*}{R} = \frac{466}{2500} = 0.18 \text{ year} = 0.18 \times 12 \text{ months} = 2.16 \text{ months}$$

■

13.2.3 Model I(c): Economic lot size with finite rate of Replenishment (finite production) [EPQ model]

Some Notations:

$$\begin{aligned} C_1 &= \text{Holding cost per unit item per unit time.} \\ R &= \text{Demand rate.} \\ K &= \text{Production rate is finite, uniform and greater than R.} \\ t &= \text{interval between production cycle.} \\ q &= Rt \end{aligned}$$

In this model, each production cycle time t consists of two parts: t_1 and t_2, where

(i) t_1 is the period during which the stock is growing up at a rate of $(K - R)$ items per unit time.

(ii) t_2 is the period during which there is supply but there is only a constant demand at the rate of R.

It is evident from the graphical situation (see fig. 13.1) that

$$t_1 = \frac{Q}{K - R} \qquad \text{and} \qquad t_2 = \frac{Q}{R}$$

$$\begin{aligned} t &= t_1 + t_2 \\ &= \frac{Q}{K - R} + \frac{Q}{R} \\ &= \frac{QK}{R(K - R)} \end{aligned}$$

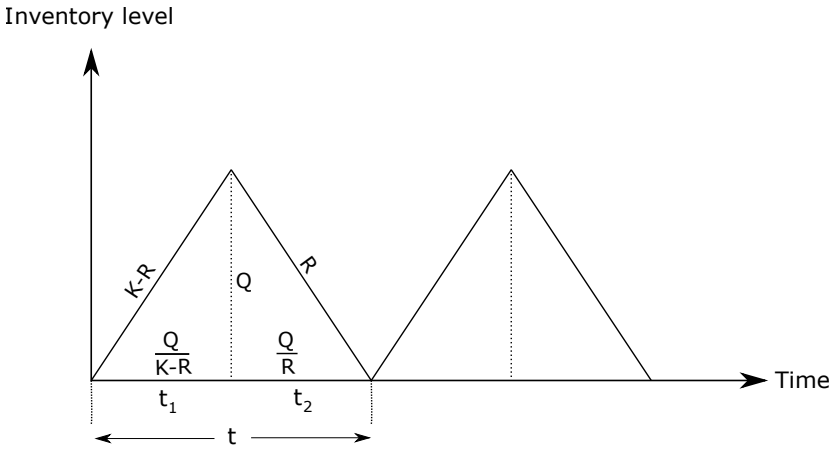


Figure 13.1

which gives

$$\begin{aligned} Q &= \frac{K-R}{K}Rt \\ &= \frac{K-R}{K}q \quad [\because q = Rt] \end{aligned}$$

Now, Holding cost for the time period t is $\frac{1}{2}C_1Qt$ and the set up cost for period t is C_3.

$\therefore$ The total average cost is

$$\begin{aligned} C(t) &= \frac{\frac{1}{2}C_1Qt + C_3}{t} \\ C(q) &= \frac{1}{2}C_1\left(\frac{K-R}{K}\right)q + C_3\frac{R}{q} \quad [\because q = Rt] \end{aligned} \tag{13.2.7}$$

For optimum value of q, we have

$$\begin{aligned} &\frac{dC}{dq} = 0 \\ &\Rightarrow \frac{1}{2}\left(1 - \frac{R}{K}\right)C_1 - \frac{C_3R}{q^2} = 0 \\ &\Rightarrow q = \sqrt{\frac{2C_3RK}{C_1(K-R)}} = \sqrt{\frac{2C_3R}{C_1\left(1-\frac{R}{K}\right)}} \end{aligned}$$

Now, $\quad \frac{d^2C}{dq^2} = \frac{2C_3R}{q^3} > 0$

$\therefore q^* = \sqrt{\frac{2C_3R}{C_1\left(1-\frac{R}{K}\right)}}$ (optimal lot size)

and $\quad t^* = \frac{q^*}{R} = \sqrt{\frac{2C_3}{C_1R\left(1-\frac{R}{K}\right)}}$

and the corresponding minimum total average cost

$$C_{\min} = \sqrt{2C_1\left(1 - \frac{R}{K}\right)C_3R}$$

Note 13.2.5. 1. If $K = R$, $C_{\min} = 0$, which implies that there will be no carrying cost and set up cost.

2. If $K \to \infty$, i.e., production rate is infinite, then this model becomes exactly same as Model I(a).

Example 13.2.6. A contractor has to supply 10,000 bearings per day to an auto-mobile manufacturer. He finds that when he starts a production run, he can produce 25,000 bearings per day. The cost of holding a bearing in stock for one year is 20 paisa and set up cost of a production run is Rs. 180. How frequently (time) should production run be made?

Solution.

$$\begin{aligned} R &= 10000 \quad \text{bearings per day} \\ K &= 25000 \quad \text{bearing per day} \\ C_1 &= \text{Rs. } \frac{0.20}{365} \text{ per bearing per day} \\ &= \text{Rs. } 0.0005 \text{ per bearing per day.} \\ C_3 &= \text{Rs. } 180 \text{ per run.} \end{aligned}$$

$$\therefore t^* = \sqrt{\frac{2 \times 180}{0.0005 \times 10000} \times \frac{3}{5}} = 0.3 \text{ day}$$

■

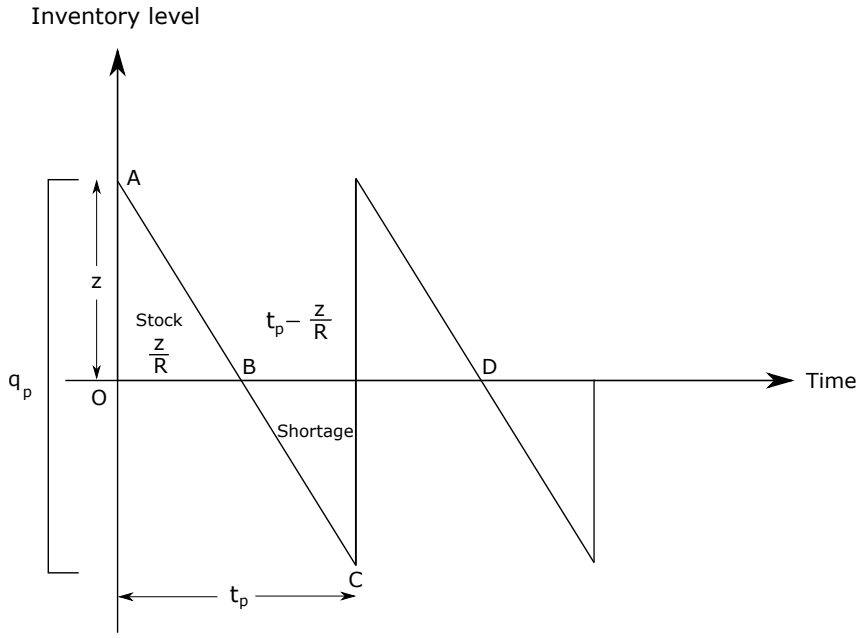
Unit 14

Course Structure

- Model II(a): EOQ model with constant rate of demand scheduling time constant.
- Model II(b): EOQ model with constant rate of demand scheduling time variable.
- Model II(c): EPQ model with shortages.

14.1 Model II(a) : EOQ model with constant rate of demand scheduling time constant

Model II is the extension of Model I allowing shortages.



Some Notations:

C_1 = Holding cost
C_2 = Shortage cost
R = Demand rate
t_p = Scheduling time period is constant
q_p = Fixed lot size (Rt_p)
z = Order level to which the inventory raised in the beginning of each scheduling period.

Here z is the variable. Production rate is infinite. Lead Time is zero.

In this model, we can easily observe that the inventory carrying cost C_1 and also the shortage cost C_2 will be involved only when $0 \leq z \leq q_p$.

$$\begin{aligned}\text{Holding cost per unit time} &= C_1(\Delta OAB)/t_p \\ &= \frac{C_1}{t_p}\left(\frac{1}{2}\cdot z\cdot\frac{z}{R}\right) \\ &= \frac{1}{2}\frac{z^2C_1}{Rt_p} \quad (\because q_p = Rt_p)\end{aligned}$$

$$\begin{aligned}\text{Shortage cost per unit time} &= C_2(\Delta BDC)/t_p \\ &= \frac{C_2}{t_p}\left(\frac{1}{2}\cdot BD\cdot DC\right) \\ &= \frac{C_2}{t_p}\left[\frac{1}{2}(t_p-\frac{z}{R})(q_p-z)\right] \\ &= \frac{1}{2}\frac{C_2}{q_p}(q_p-z)^2\end{aligned}$$

$$\text{Total average cost is } C(z) = \frac{1}{2}\frac{z^2C_1}{q_p} + \frac{1}{2}\frac{C_2}{q_p}(q_p-z)^2 + \frac{C_3}{t_p}$$

Note 14.1.1. Since, the set up cost C_3 and period t_p are constant, the average set up cost $\frac{C_3}{t_p}$ also being constant, will be considered in the cost equation.

Now

$$\frac{dC}{dz} = \frac{1}{2}\cdot\frac{C_1}{q_p}\cdot 2z + \frac{1}{2}\frac{C_2}{q_p}2(q_p-z)(-1) = 0$$
$$\Rightarrow z = \frac{C_2}{C_1+C_2}q_p = \frac{C_2}{C_1+C_2}Rt_p.$$

$$\frac{d^2C}{dz^2} = \frac{C_1}{q_p} + \frac{C_2}{q_p} = \frac{C_1+C_2}{q_p} > 0.$$
$$\therefore z^* = \frac{C_2}{C_1+C_2}Rt_p$$
$$C_{\min} = \frac{C_1C_2}{2(C_1+C_2)}Rt_p.$$

14.2 Model II(b) : EOQ model with constant rate of demand scheduling time variable

Assumptions:

(i) R is the demand rate.

(ii) Production is instantaneous.

(iii) $q = Rt$.

(iv) t is the scheduling time period which is variable.

(v) z is the order level.

(vi) Lead time is zero.

Formulate the model. Show that the optimal order quantity per run which minimizes the total cost is

$$q = \sqrt{\frac{2RC_3(C_1 + C_2)}{C_1C_2}}$$

Since, all the assumptions in this model are same as in Model II(a), except with the difference that the scheduling time period t is not constant here, so, it now becomes important to consider the average set up cost $\frac{C_3}{t}$ in the cost equation.

Thus the cost equation becomes

$$C(z,t) = \frac{C_1z^2}{2Rt} + \frac{1}{2}\frac{C_2}{Rt}(Rt - z)^2 + \frac{C_3}{t}.$$

For the optimization, $\frac{\partial C}{\partial z} = 0$ and $\frac{\partial C}{\partial t} = 0$ which gives

$$\frac{1}{t}\left(\frac{2C_1z}{2R} - \frac{2C_3}{2R}(Rt - z)\right) = 0$$
$$\therefore z = \frac{C_2Rt}{C_1 + C_2}$$

Now

$$-\frac{1}{t^2}\left(\frac{C_1z^2}{2R} + \frac{C_2}{2R}(Rt - z)^2 + C_3\right) + \frac{1}{t}\left(0 + \frac{C_2}{2R}2(Rt - z) + 0\right) = 0$$
$$\Rightarrow \quad -\frac{1}{t^2}\left(\frac{C_1z^2}{2R} + \frac{C_2}{2R}(Rt - z)^2 + C_3\right) + \frac{C_2}{t}(Rt - z) = 0$$

Multiplying this equation by $2Rt^2$ and simplifying we get,

$$-(C_1 + C_2)z^2 + C_2R^2t^2 = 2RC_3$$

Substituting the value of z in the given equation, we have

$$
\begin{aligned}
&-\frac{R^2t^2C_2^2}{C_1+C_2}+C_2R^2t^2=2RC_3\\
&\Rightarrow C_2R^2t^2\left(1-\frac{C_2}{C_1+C_2}\right)=2RC_3\\
&\Rightarrow C_2R^2t^2\left(\frac{C_1}{C_1+C_2}\right)=2RC_3\\
&\Rightarrow t=\sqrt{\frac{2C_3(C_1+C_2)}{RC_1C_2}}
\end{aligned}
$$

For minimum cost, we may further verify that

$$
\begin{aligned}
&\frac{\partial^2C}{\partial t^2}\cdot\frac{\partial^2C}{\partial z^2}-\left(\frac{\partial^2C}{\partial t\partial z}\right)^2>0\\
&\text{and}\quad \frac{\partial^2C}{\partial t^2}>0 \qquad \frac{\partial^2C}{\partial z^2}>0
\end{aligned}
$$

Hence

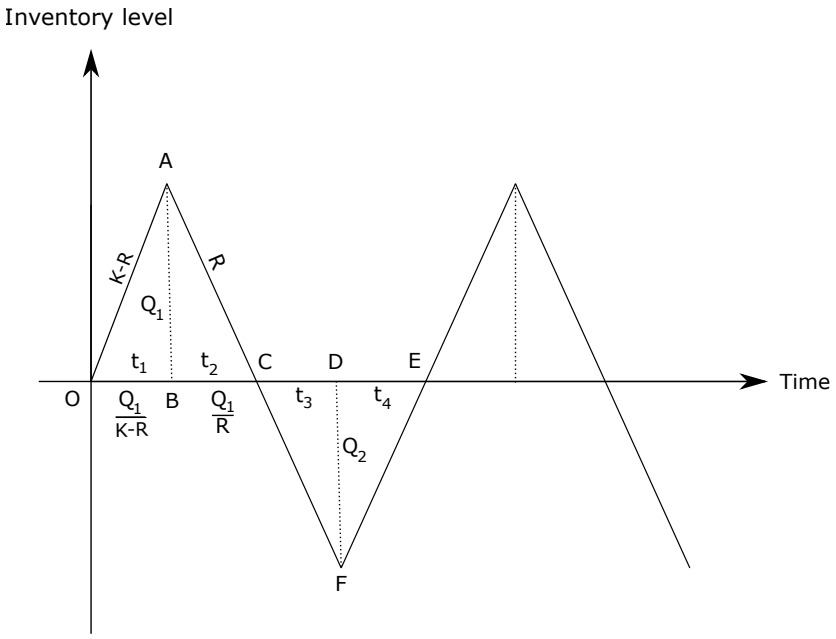
$$
\begin{aligned}
t^* &= \sqrt{\frac{2C_3(C_1+C_2}{RC_1C_2}}\\
q^* &= Rt^*=R\sqrt{\frac{2C_3(C_1+C_2)}{RC_1C_2}}\\
&= \sqrt{\frac{2RC_3(C_1+C_2)}{C_1C_2}} \quad \text{(EOW/lot size)}\\
C_{\min} &= \frac{C_1}{2Rt^*}\left(\frac{C_2Rt^*}{C_1+C_2}\right)^2+\frac{1}{2}\frac{C_2}{Rt^*}\left(Rt^*-\frac{C_2Rt^*}{C_1+C_2}\right)^2+\frac{C_3}{t^*}\\
&= \frac{C_1C_2^2R^2}{2(C_1+C_2)^2Rt^*}t^{*2}+\frac{1}{2}\frac{C_2}{Rt^*}\left(\frac{C_1Rt^*}{C_1+C_2}\right)^2+\frac{C_3}{t^*}\\
&= \frac{C_1C_262}{2(C_1+C_2)^2}(Rt^*)+\frac{C_2C_1^2}{2(C_1+C_2)^2}(Rt^*)+\frac{C_3}{t^*}\\
&= \frac{1}{2}\frac{C_1C_2}{(C_1+C_2)^2}(C_1+C_2)(Rt^*)+\frac{C_3}{t^*}\\
&= \frac{1}{2}\frac{C_1C_2}{(C_1+C_2)}\sqrt{\frac{2RC_3(C_1+C_2)}{C_1C_2}}+C_3\sqrt{\frac{RC_1C_2}{2C_3(C_1+C_2)}}\\
&= \sqrt{\frac{C_1C_2RC_3}{2(C_1+C_3}}+\sqrt{\frac{RC_1C_2C_3}{2(C_1+C_2)}}\\
&= 2\sqrt{\frac{RC_1C_2C_3}{2(C_1+C_2)}}\\
&= \sqrt{2C_1C_3R\times\frac{C_2}{C_1+C_2}}\\
&= \sqrt{2C_1C_3R}\sqrt{\frac{C_2}{C_1+C_2}}
\end{aligned}
$$

Further, it is interesting to note that the minimum cost is less than that already given by Model I(a) $\sqrt{2C_1C_3R}$. (Draw figure as like Model II(a) replaced by t).

14.3 Model II(c) : EPQ model with shortages

The production lot size model with shortage.

Assumptions:



(i) R is the demand rate.

(ii) Lead time is zero.

(iii) Production rate (K) is finite, $K > R$.

(iv) Inventory carrying cost $C_1 = IP$ (For EPQ), $P =$ Finite production cost.

(v) Shortages are allowed and backlogged.

(vi) Shortage cost is Rs. C_2 per quantity unit per unit time.

(vii) Set up cost is Rs. C_3 per order or per set up.

$$\text{Holding Cost} = C_1 \times \Delta\text{OAC} = C_1 \times \frac{1}{2}Q_1(t_1 + t_2).$$

$$\text{Shortage Cost} = C_2 \left(\frac{1}{2}Q_2(t_3 + t_4)\right)$$

and Set up cost C_3.

Thus, the total average cost,

$$\begin{aligned} C &= \frac{\frac{1}{2}C_1Q_1(t_1+t_2)+\frac{1}{2}C_2Q_2(t_3+t_4)+C_3}{t_1+t_2+t_3+t_4} \\ t_1 &= \frac{Q_1}{K-R}, \quad t_2=\frac{Q_1}{R} \\ &= \frac{Rt_2}{K-R}, \quad Q_1=Rt_2. \end{aligned} \tag{14.3.1}$$

Again,

$$\begin{aligned} Q_2 &= Rt_3, & t_4 &= \frac{Q_2}{K-R}. \\ Q_2 &= (K-R)t_4 & &= \frac{Rt_3}{K-R}. \end{aligned}$$

Finally,

$$\begin{aligned} q=Rt &= R(t_1+t_2+t_3+t_4) \\ &= R\left(\frac{Rt_2}{K-R}+t_2+t_3+\frac{Rt_3}{K-R}\right) \\ &= \frac{(t_2+t_3)KR}{K-R} \\ C &= \frac{\frac{1}{2}\left\{C_1(Rt_2)\left(\frac{Rt_2}{K-R}+t_2\right)+C_3Rt_3\left(t_3+\frac{Rt_3}{K-R}\right)\right\}+C_3}{\frac{Rt_2}{K-R}+t_2+t_3+\frac{Rt_3}{K-R}} \\ &= \frac{\frac{1}{2}\left\{\frac{C_1t_2^2RK}{K-R}+\frac{C_2t_3^2RK}{K-R}\right\}+C_3}{(t_2+t_3)\left(1+\frac{R}{K-R}\right)} \\ &= \frac{\frac{1}{2}(C_1t_2^2+C_2t_3^2)\left(\frac{RK}{K-R}\right)+C_3}{(t_2+t_3)\left(\frac{K}{K-R}\right)} \\ &= \frac{\frac{1}{2}(C_1t_2^2+C_2t_3^2)RK+C_3(K-R)}{K(t_2+t_3)}. \end{aligned}$$

This is a function of t_2 and t_3 $C(t_2,t_3)$

$$\frac{\partial C}{\partial t_2}=0, \quad \frac{\partial C}{\partial t_3}=0,$$

$$t_2^*=\sqrt{\frac{2C_3C_2(1-R/K)}{(R(C_1+C_2)C_1}}, \qquad q^*=\sqrt{\frac{2RC_3(C_1+C_2)}{(C_1C_2}\left(\frac{1}{1-R/K}\right)}$$

$$t_3^*=\sqrt{\frac{2C_3C_1(1-R/K)}{(R(C_1+C_2)C_2}}, \qquad C_{\min}=\sqrt{\frac{2RC_1C_2C_3(1-R/K)}{C_1+C_2}}$$

$$C=\frac{\frac{1}{2}(C_1t_2^2+C_2t_3^2)RK+C_3(K-R)}{K(t_2+t_3)}.$$

Now,

$$\begin{aligned}
& \frac{\partial C}{\partial t_2} = 0. \\
\Rightarrow \quad & \frac{K(t_2+t_3)\left[\frac{1}{2}C_1 \times 2t_2\right]RK - \left[\frac{1}{2}(C_1t_2^2 + C_2t_3^2)RK + C_3(K-R)\right]K}{K^2(t_2+t_3)^2} = 0 \\
\Rightarrow \quad & K(t_2+t_3)\cdot C_1t_2RK - \left[\frac{1}{2}(C_1t_2^2 + C_2t_3^2)RK + C_3(K-R)\right]K = 0 \\
\Rightarrow \quad & C_1t_2^2RK^2 + C_1t_2t_3RK^2 - \frac{1}{2}C_1t_2^2RK^2 - \frac{1}{2}C_2t_3^2RK^2 - C_3K(K-R) = 0 \\
\Rightarrow \quad & \frac{1}{2}C_1t_2^2RK^2 + C_1t_2t_3RK^2 - \frac{1}{2}C_2t_3^2RK^2 - C_3K(K-R) = 0 \\
\Rightarrow \quad & \frac{1}{2}C_1t_2^2RK^2 + C_1t_2t_3RK^2 - \frac{1}{2}C_2t_3^2RK^2 = C_3K(K-R) \\
\Rightarrow \quad & \frac{1}{2}RK^2(C_1t_2^2 + 2C_1t_2t_3 - C_2t_3^2) = C_3K(K-R) \\
\Rightarrow \quad & C_1t_2^2 + 2C_1t_2t_3 - C_2t_3^2 = \frac{2C_3(1-R/K)}{R} \\
\Rightarrow \quad & C_1t_2^2 + 2C_1t_2t_3 + C_1t_3^2 - C_1t_3^2 - C_2t_3^2 = \frac{2C_3(1-R/K)}{R} \\
\Rightarrow \quad & C_1(t_2+t_3)^2 - t_3^2(C_1+C_2) = \frac{2C_3(1-R/K)}{R} \\
\Rightarrow \quad & C_1(t_2+t_3)^2 = \frac{2C_3(1-R/K)}{R} + t_3^2(C_1+C_2) \\
\Rightarrow \quad & (t_2+t_3)^2 = \frac{2C_3(1-R/K)}{RC_1} + \frac{t_3^2(C_1+C_2)}{C_1} \\
\Rightarrow \quad & t_2+t_3 = \sqrt{\frac{2C_3(1-R/K)}{RC_1} + \frac{t_3^2(C_1+C_2)}{C_1}}.
\end{aligned}$$

Also,

$$\begin{aligned}
& K(t_2+t_3)\left[\frac{1}{2}\times C_2 \times 2C_3\right]RK - \left[\frac{1}{2}(C_1t_2^2 + C_2t_3^2)RK + C_3(K-R)\right]K = 0 \\
\Rightarrow \quad & C_2t_2t_3RK^2 + C_2t_3^2RK^2 - \frac{1}{2}C_1t_2^2RK^2 - \frac{1}{2}C_2t_3^2RK^2 - C_3(K-R)K = 0 \\
\Rightarrow \quad & \frac{1}{2}C_2t_3^2RK^2 + C_2t_2t_3RK^2 - \frac{1}{2}C_1t_2^2RK^2 = C_3(K-R)K \\
\Rightarrow \quad & C_2t_3^2 + 2C_2t_2t_3 + C_2t_2^2 - (C_1+C_2)t_2^2 = \frac{2C_3(1-R/K)}{R} \\
\Rightarrow \quad & C_2(t_2+t_3)^2 - (C_1+C_2)t_2^2 = \frac{2C_3(1-R/K)}{R}.
\end{aligned}$$

Now,

$$
\begin{aligned}
& C_1(t_2+t_3)^2-(C_1+C_2)t_3^2=\frac{2C_3(1-R/K)}{R}\\
& C_2(t_2+t_3)^2-(C_1+C_2)t_2^2=\frac{2C_3(1-R/K)}{R}\\
\Rightarrow\ & C_1(t_2+t_3)^2-(C_1+C_2)t_3^2=C_2(t_2+t_3)^2-(C_1+C_2)t_2^2\\
\Rightarrow\ & C_1(t_2+t_3)^2-C_2(t_2+t_3)^2=(C_1+C_2)t_3^2-(C_1+C_2)t_2^2\\
\Rightarrow\ & (t_2+t_3)^2(C_1-C_2)=(C_1+C_2)(t_3^2-t_2^2)\\
\Rightarrow\ & (t_2+t_3)^2(C_1-C_2)=(C_1+C_2)(t_3-t_2)(t_3+t_2)\\
\Rightarrow\ & (t_2+t_3)(C_1-C_2)=(C_1+C_2)(t_3-t_2)\\
\Rightarrow\ & C_1t_2-C_2t_2+C_1t_3-C_2t_3=C_1t_3+C_2t_3-C_1t_2-C_2t_2\\
\Rightarrow\ & 2C_1t_2=2C_2t_3\\
\Rightarrow\ & 2C_1t_2=2C_2t_3\\
\Rightarrow\ & t_2=\frac{C_2}{C_1}t_3
\end{aligned}
$$

Thus,

$$
\begin{aligned}
& C_2(t_2+t_3)^2-(C_1+C_2)t_2^2=\frac{2C_3(1-R/K)}{R}\\
\Rightarrow\ & C_2\left(\frac{C_2}{C_1}t_3+t_3\right)^2-(C_1+C_2)\left(\frac{C_2}{C_1}t_3\right)^2=\frac{2C_3(1-R/K)}{R}\\
\Rightarrow\ & C_2\left(\frac{C_2}{C_1}+1\right)^2t_3^2-(C_1+C_2)\frac{C_2^2}{C_1^2}t_3^2=\frac{2C_3(1-R/K)}{R}\\
\Rightarrow\ & \frac{C_2(C_2+C_1)^2}{C_1^2}t_3^2-\frac{(C_1+C_2)C_2^2}{C_1^2}t_3^2=\frac{2C_3(1-R/K)}{R}\\
\Rightarrow\ & \frac{t_3^2}{C_1^2}(C_1+C_2)\left[C_2(C_1+C_2)-C_2^2\right]=\frac{2C_3(1-R/K)}{R}\\
\Rightarrow\ & \frac{t_3^2(C_1+C_2)}{C_1^2}C_1C_2=\frac{2C_3(1-R/K)}{R}\\
\Rightarrow\ & \frac{t_3^2(C_1+C_2)}{C_1}C_2=\frac{2C_3(1-R/K)}{R}\\
\Rightarrow\ & t_3^2=\frac{2C_1C_3(1-R/K)}{R(C_1+C_2)C_2}\\
\Rightarrow\ & t_3^*=\sqrt{\frac{2C_1C_3(1-R/K)}{R(C_1+C_2)C_2}}.
\end{aligned}
$$

Now,

$$\begin{aligned} t_2^* &= \frac{C_2}{C_1} t_3^* \\ &= \frac{C_2}{C_1}\sqrt{\frac{2C_1C_3(1-R/K)}{R(C_1+C_2)C_2}} \\ &= \sqrt{\frac{2C_1C_3(1-R/K)C_2^2}{C_1^2R(C_1+C_2)C_2}} \\ &= \sqrt{\frac{2C_2C_3(1-R/K)}{R(C_1+C_2)C_1}}. \end{aligned}$$

Now,

$$\begin{aligned} q^* &= \frac{KR}{K-R}\left[\sqrt{\frac{2C_2C_3(1-R/K)}{R(C_1+C_2)C_1}} + \sqrt{\frac{2C_1C_3(1-R/K)}{R(C_1+C_2)C_2}}\right] \\ &= \frac{R}{(1-R/K)}\left[\sqrt{\frac{2C_3(1-R/K)}{R(C_1+C_2)}}\left\{\sqrt{\frac{C_2}{C_1}} + \sqrt{\frac{C_1}{C_2}}\right\}\right] \\ &= \sqrt{\frac{2C_3}{R(C_1+C_2)(1-R/K)}} \times \frac{(C_1+C_2)R}{\sqrt{C_1C_2}} \\ &= \sqrt{\frac{2C_3(C_1+C_2)^2R^2}{R(C_1+C_2)C_1C_2(1-R/K)}} \\ &= \sqrt{\frac{2RC_3(C_1+C_2)}{C_1C_2(1-R/K)}} = \sqrt{\frac{2RC_3(C_1+C_2)}{C_1C_2}\left(\frac{1}{1-R/K}\right)}. \end{aligned}$$

So,

$$\begin{aligned} C_{\min} &= \frac{\frac{1}{2}(C_1t_2^{*2} + C_2t_3^{*2})RK + C_3(K-R)}{K(t_2^* + t_3^*)} \\ &= \frac{\frac{1}{2}\left[\frac{2C_1C_2C_3(1-R/K)}{R(C_1+C_2)C_1} + \frac{2C_1C_2C_3(1-R/K)}{R(C_1+C_2)C_2} + C_3(K-R)\right]}{K\left(\sqrt{\frac{2C_3(1-R/K)}{R(C_1+C_2)}} \cdot \frac{(C_1+C_2)}{\sqrt{C_1C_2}}\right)} \\ &= \frac{\left[\frac{C_2C_3(1-R/K)}{R(C_1+C_2)} + \frac{C_1C_3(1-R/K)}{R(C_1+C_2)} + \frac{C_3(K-R)}{2}\right]}{K\sqrt{\frac{2C_3(1-K/R)(C_1+C_2)}{RC_1C_2}}} \\ &= \frac{2C_2C_3(1-R/K) + 2C_1C_3(1-R/K) + C_3K(1-R/K)R(C_1+C_2)}{K\sqrt{\frac{2C_3(1-R/K)(C_1+C_2)}{RC_1C_2}}}. \end{aligned}$$

Example 14.3.1. The demand of an item is uniform at a rate of 25 units per month. The fixed cost is Rs. 15 each time a production run is made (Setup cost). The production cost if Rs. 1 per item and inventory carrying cost is Rs. 0.30 per item per month. If the shortage cost is Rs. 1.50 per item per month, determine how often to make a production run and of what size it should be?

Solution. We have,

$$\begin{aligned} R &= 25 \text{ units per month} \\ C_3 &= \text{Rs. } 15 \text{ per run} \\ I &= \text{Rs. } 0.30 \text{ per item per month. (Inventory carrying cost)} \\ C_2 &= \text{Rs. } 1.50 \text{ per item per month} \\ P &= \text{Rs. } 1 \text{ per item.} \end{aligned}$$

Thus,

$$C_1 = \text{Rs. } 0.30 \text{ per item per month.}$$

Here, the demand of an item is uniform. So,

$$q^* = \sqrt{\frac{2RC_3(C_1 + C_2)}{C_1C_2}} = \sqrt{\frac{2 \times 25 \times 15 \times (0.30 + 1.50)}{0.30 \times 1.50)}} \approx 54 \text{ units.}$$

and

$$t^* = \sqrt{\frac{2C_3(C_1 + C_2)}{RC_1C_2}} = \sqrt{\frac{2 \times 15 \times (0.30 + 1.50)}{25 \times 0.30 \times 1.50}} = 2.19 \text{ months.}$$

■

Unit 15

Course Structure

- Model III: Multi-item inventory model

15.1 Model III: Multi-item inventory model

So far, we have considered each item separately but if there exists a relationship among the items under some limitations, then it is not possible to consider them separately. After constructing the cost equation in such models, we use the method of Lagrange's multiplier to minimize the cost. We consider the problem with the following assumptions

1. n is the number of items to be considered and no lead time.
2. R_{1i} is the uniform demand rate for the ith item $(i = 1, 2, \ldots, n)$.
3. C_{1i} is the holding cost of the ith item
4. Shortages are not allowed
5. C_{3i} is the setup cost for the ith item
6. q_i is the total quantity to be produced of the ith item.

Now, proceeding exactly as in the model I(a), we get,

$$C_i(t) = \frac{1}{2}C_{1i}R_i t + \frac{C_{3i}}{t},$$
$$\text{or,} \quad C_i(q_i) = \frac{1}{2}C_{1i}q_i + \frac{C_{3i}R_i}{q_i} \tag{15.1.1}$$

Then total cost

$$C = \sum_{i=1}^{n}\left\{\frac{1}{2}C_{1i}q_i + \frac{C_{3i}R_i}{q_i}\right\} \tag{15.1.2}$$

To determine the optimum value of q_i, we have

$$\begin{aligned} & \frac{\partial C}{\partial q_i} = 0 \\ \Rightarrow \quad & \frac{1}{2}C_{1i} - \frac{C_{3i}R_i}{q_i^2} = 0 \\ \Rightarrow \quad & q_i = \sqrt{\frac{2c_{3i}R_i}{C_{1i}}}. \end{aligned}$$

Thus,

$$\frac{\partial^2 C}{\partial q_i^2} > 0, \ \ \forall\, q_i.$$

The total cost is minimum. Hence, the optimum cost of

$$q_i^* = \sqrt{\frac{2C_{3i}R_i}{C_{1i}}}, \ \ (i = 1, 2, \ldots, n) \tag{15.1.3}$$

We now proceed to consider the effect of limitations, which are,

1. limitation on investment
2. limitation on stock unit
3. limitation on warehouse floor space

15.1.1 Model III(a): Limitation on Investment

In this case, there is an upper limit M (in Rs.) on the amount to be invested on inventory. Let C_{4i} be the unit price of the ith item. Then

$$\sum_{i=1}^{n} C_{4i}q_i \leq M \tag{15.1.4}$$

Now, our problem is to minimize the total cost C given by equation (15.1.2) subject to the constraint (15.1.4). In this situation, two cases may arise.

Case I: When $\sum_{i=1}^{n} C_{4i}q_i \leq M$ and $q_i^* = \sqrt{\frac{2C_{3i}R_i}{C_{1i}}}$.

In this case, there is no difficulty and hence q_i^* is the optimal solution.

Case II: When $\sum_{i=1}^{n} C_{4i}q_i > M$ and $q_i^* = \sqrt{\frac{2C_{3i}R_i}{C_{1i}}}$.

In this case, q_i^* are not required optimal solutions. Thus, we shall use the Lagrange's multiplier technique.

$$L = \sum_{i=1}^{n} \left(\frac{1}{2}C_{1i}q_i + \frac{C_{3i}R_i}{q_i} \right) + \lambda \left(\sum_{i=1}^{n} C_{4i}q_i - M \right).$$

Here, λ is the Lagrangian multiplier.

The necessary condition

$$\begin{aligned}
&\frac{\partial L}{\partial q_i} = 0 \ (i = 1, 2, \ldots, n) \\
&\frac{\partial L}{\partial \lambda} = 0 \\
\Rightarrow \quad &\frac{1}{2}C_{1i} - \frac{C_{3i}R_i}{q_i^2} + \lambda C_{4i} = 0, \quad \text{and} \quad C_{4i}q_i - M = 0 \\
\Rightarrow \quad &q_i^* = \sqrt{\frac{2C_{3i}R_i}{C_{1i} + 2\lambda C_{4i}}} \quad \text{and} \quad C_{4i}q_i^* = M.
\end{aligned}$$

q_i^* depends on λ. λ can be found by *trial and error method*. By trying positive successive values of λ, the values of λ^* should result in simultaneous value of q_i^* satisfying the given constraint by equality sense.

Example 15.1.1. Consider a shop producing three items, the items are produced in lots. The demand rate for each item is constant and can be assumed to be deterministic. No back order (shortages) are allowed. The following data are given below.

Item	1	2	3
H.C	20	20	20
S.C	50	40	60
Cost per unit item	6	7	5
Yearly demand rate	10,000	12,000	7,500

Determine approximately the EOQ when the total value of average inventory levels of three items if Rs. 1,000.

Solution.

$$\begin{aligned}
q_1^* &= \sqrt{\frac{2C_{31}R}{C_{11}}} = \sqrt{\frac{2 \times 50 \times 10,000}{20}} = 100\sqrt{5} \approx 223 \\
q_2^* &= 40\sqrt{30} \approx 216 \\
q_3^* &= 150\sqrt{2} \approx 210.
\end{aligned}$$

Since the average optimal inventory at any time is $q_i^*/2$, the investment over the average inventory is obtained by replacing q_i by $q_i^*/2$, that is,

$$\sum_{i=1}^{n} C_{4i}\left(\frac{1}{2}q_i^*\right) = \text{Rs. } \left(6 \times \frac{223}{2} + 7 \times \frac{216}{2} + 5 \times \frac{210}{2}\right) = \text{Rs. } 1950.$$

We observe that the amount of Rs. 1950 is greater than the upper limit of Rs. 1000. Thus, we try to find the suitable value of λ by trial and error method for computing q_i^*.

If we put $\lambda = 4$, we get

$$\begin{aligned}
q_1^* &= \sqrt{\frac{2 \times 50 \times 10,000}{20 + 2 \times 4 \times 6}} = 121 \\
q_2^* &= 112 \\
q_3^* &= 123.
\end{aligned}$$

$$\text{Cost of average inventory} = 6 \times \frac{121}{2} + 7 \times \frac{112}{2} + 5 \times \frac{123}{2} = \text{Rs. } 1112.50.$$

Again, if we put $\lambda = 5$, then

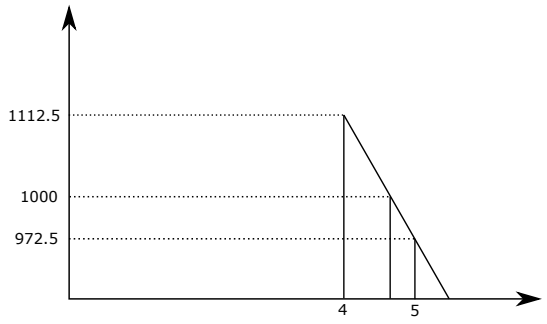
$$q_1^* = 111$$
$$q_2^* = 102$$
$$q_3^* = 113.$$

and

$$\text{Corresponding cost} = \text{Rs. } 972.50$$

which is less than Rs. 1000.

From this, we conclude that, the most suitable value of λ lies between 4 and 5.



To find the most suitable value of λ, we draw a graph between cost and the value of λ as shown in the figure. This graph indicates that $\lambda = 4.7$ is the most suitable value corresponding to which the cost of inventory is Rs. 999.5, which is sufficiently close to Rs. 1000. Hence, for $\lambda = 4.7$, we obtain

$$q_1^* = 114$$
$$q_2^* = 105$$
$$q_3^* = 116.$$

■

15.1.2 Model III(b): Limitation on inventory

In this case, the upper limit of average number of all units in stock is N (say). Hence we have, since the average number of units at any time is $q_i/2$.

$$\text{Min } C = \sum_{i=1}^{n} \left(\frac{1}{2} C_{1i} q_i + \frac{C_{3i} R_i}{q_i} \right)$$
$$\text{subject to } \frac{1}{2} \sum_{i=1}^{n} q_i \leq N.$$

Here also, two cases arise.

Case I: $\frac{1}{2} \sum_{i=1}^{n} q_i \leq N$ and $q_i^* = \sqrt{\frac{2C_{3i}R_i}{C_{1i}}}$, there is no difficulty and the optimum values of q_i^*.

Case II: $\frac{1}{2}\sum_{i=1}^{n} q_i > N$, then q_i^* are not the required values. So, we use Lagrange's multiplier technique. Here, Lagrangian function

$$L = \sum_{i=1}^{n}\left(\frac{1}{2}C_{1i}q_i + \frac{C_{3i}R_i}{q_i}\right) + \lambda\left(\frac{1}{2}\sum_{i=1}^{n} q_i - N\right)$$

where $\lambda > 0$ is a Lagrangian multiplier.

For the minimum value of L, the necessary conditions are

$$\frac{\partial L}{\partial q_i} = \frac{1}{2}C_{1i} - \frac{C_{3i}R_i}{q_i^2} + \frac{\lambda}{2} = 0$$
$$\frac{\partial L}{\partial \lambda} = \frac{1}{2}\sum_{i=1}^{n} q_i - N = 0, \;\; i = 1, 2, \ldots, n.$$

Solving, we get

$$q_i^* = \sqrt{\frac{2C_{3i}R_i}{C_{1i} + \lambda}}$$
$$\frac{1}{2}\sum_{i=1}^{n} q_i = N.$$

To obtain the value of q_i^*, we obtain the value of λ by successive trial and error method and satisfying the given constraint in equality sign.

Example 15.1.2. A company producing three items have a limited storage space of 750 items of all types in average. Determine the optimal production quantity for each item separately when the following information is given

Product	1	2	3
H.S(Rs.)	0.05	0.02	0.04
S.C(Rs.)	50	40	60
D.R(per unit)	100	120	75

Solution. We have

$$q_1^* = 447$$
$$q_2^* = 693$$
$$q_3^* = 464.$$

$$\text{The total average inventory is} = \frac{1}{2}(447 + 693 + 464) = 802 \text{ units},$$

which is greater than 750 units per year. Thus, we have to find the value of the parameter λ by trial and error method.

From these, we observe that the average inventory level is less than the available amount of items. So we try for some other values of λ,

$$\lambda = 0.004,\ 0.003,\ 0.002,\ \text{etc.}$$

For $\lambda = 0.002$,

$$
\begin{aligned}
&q_1^* = 428\\
&q_2^* = 628\\
&q_3^* = 444\\
&\text{Average inventory level} = \frac{1}{2}(428 + 628 + 444) = 750,
\end{aligned}
$$

which is equivalent to the given amount of average inventory. Hence, the optimal solutions are

$$
\begin{aligned}
q_1^* &= 428\\
q_2^* &= 628\\
q_3^* &= 444.
\end{aligned}
$$

For $\lambda = 0.004$,

$$
\begin{aligned}
q_1^* &= \sqrt{\frac{2 \times 50 \times 100}{0.05 + 0.004}} = 430\\
q_2^* &= \sqrt{\frac{2 \times 40 \times 120}{0.02 + 0.004}} = 632\\
q_1^* &= \sqrt{\frac{2 \times 60 \times 75}{0.04 + 0.004}} = 452\\
&\text{Average inventory level} = \frac{1}{2}(430 + 632 + 452) = 757.
\end{aligned}
$$

For $\lambda = 0.003$,

$$
\begin{aligned}
q_1^* &= \sqrt{\frac{2 \times 50 \times 100}{0.05 + 0.003}} = 434\\
q_2^* &= \sqrt{\frac{2 \times 40 \times 120}{0.02 + 0.003}} = 646\\
q_1^* &= \sqrt{\frac{2 \times 60 \times 75}{0.04 + 0.003}} = 457\\
&\text{Average inventory level} = \frac{1}{2}(434 + 646 + 457) = 768.5.
\end{aligned}
$$

■

15.1.3 Model III(c): Limitation on floor space

$A =$ The maximum storage area available for the n items.
$a_i =$ Storage area required per unit of the ith item.

Thus, the total storage requirement constraint becomes

$$\sum_{i=1}^{n} a_i q_i \leq A, \;\; q_i \geq 0.$$

Hence, our problem becomes,

$$\text{Min } C = \sum_{i=1}^{n} \left(\frac{1}{2} C_{1i} q_i + \frac{C_{3i} R_i}{q_i} \right),$$

$$\text{Subject to } \sum_{i=1}^{n} a_i q_i \leq A.$$

$$q_i \geq 0.$$

Case I: If $\sum_{i=1}^{n} a_i q_i \leq A$, then $q_i^* = \sqrt{\frac{2C_{3i}R_i}{C_{1i}}}$. Here we have no difficulty. Hence q_i^* is the optimal solution.

Case II: If $\sum_{i=1}^{n} a_i q_i > A$, then the optimal value q_i^* are not the required value. So we use the Lagrange's multiplier technique. The Lagrangain function is

$$L = \sum_{i=1}^{n} \left(\frac{1}{2} C_{1i} q_i + \frac{C_{3i} R_i}{q_i} \right) + \lambda \left(\sum_{i=1}^{n} a_i q_i - A \right)$$

where $\lambda > 0$ is a Lagrangian multiplier.

The necessary conditions for minimum value of L are

$$\frac{\partial L}{\partial q_i} = 0, \quad \frac{\partial L}{\partial \lambda} = 0.$$

Then, solving we have

$$q_i^* = \sqrt{\frac{2C_{3i}R_i}{C_{1i} + 2\lambda a_i}}, \ \ i = 1, 2, \ldots, n, \quad \text{and} \quad \sum_{i=1}^{n} a_i q_i^* = A.$$

The second equation implies that q_i^* must satisfy the storage constraint in equality sense. The determination of λ by usual trial and error method automatically gives the optimal value of q_i^*.

Unit 16

Course Structure

- Model IV: Deterministic inventory model with price breaks of quantity discount
- Probabilistic Inventory Model

16.1 Model IV: Deterministic inventory model with price breaks of quantity discount

Notations:

$P =$ Cost per item of producing.
$I =$ Unit price per unit item.
$C_3 =$ Setup cost.
$R =$ demand rate.
$t =$ Interval between placing orders.
$q =$ Quantity order.

Assumptions:

1. Demand rate R is constant.
2. Demand is both fixed and known.
3. No shortages are to be permitted.
4. The variable cost associated with the purchasing process.

Determine:

1. How often should be purchased (t^*)?
2. How many units should be purchased at any time (q^*)?

We have,

$$q = Rt \tag{16.1.1}$$

The number of inventories will be given by $\frac{1}{2}qt$.

$$\frac{1}{2}qt = \frac{1}{2}q\frac{q}{R} = \frac{q^2}{R} \tag{16.1.2}$$

The number of lot of inventories will be given by

$$\frac{1}{2}\frac{qt}{q} = \frac{1}{2}\frac{\frac{q^2}{R}}{q} = \frac{1}{2}\frac{q}{R} \tag{16.1.3}$$

$$\begin{aligned} C_3 &= \text{Setup Cost.} \\ qP &= \text{the purchasing cost of q units.} \\ C_3\left(\frac{1}{2}\frac{q}{R}\right)I &= \text{Cost associated with setup of inventory for period t.} \\ qP\left(\frac{1}{2}\frac{q}{R}\right)I &= \text{Cost associated with purchase of inventory for period t.} \end{aligned}$$

Therefore, total cost for period t is given by,

$$C_3 + qP + C_3\frac{1}{2}\frac{q}{R}I + qP \cdot \frac{1}{2}\frac{q}{R}I$$

Hence, average cost per unit time,

$$\begin{aligned} C(q) &= \frac{1}{t}\left(C_3 + qP + C_3\frac{1}{2}\frac{q}{R}I + qP \cdot \frac{1}{2}\frac{q}{R}I\right) \\ C(q) &= \frac{C_3R}{q} + pR + \frac{C_3I}{2} + \frac{qPI}{2} \qquad \left(\text{since } \mathrm{t} = \frac{q}{R}\right) \end{aligned}$$

But the term $\frac{1}{2}C_3I$ being constant throughout the model, it maybe neglected for the purpose of minimization. Therefore,

$$C(q) = \frac{C_3R}{q} + PR + \frac{qPI}{2} \tag{16.1.4}$$

For minimum value of $C(q)$, $\frac{d}{dq}C(q) = 0$.

$$\begin{aligned} &\frac{d}{dq}C(q) = 0 \\ \Rightarrow \quad &\frac{-C_3R}{q^2} + \frac{1}{2}PI = 0 \\ \Rightarrow \quad &q^* = \sqrt{\frac{2C_3R}{IP}} \end{aligned} \tag{16.1.5}$$

Therefore,

$$C(q^*) = \sqrt{2C_3RPI} + PR \tag{16.1.6}$$

Purchase Cost (P) per item	Range of quantity
P_1	$1 \leq q_1 \leq b$
P_2	$q_2 \geq b$

Table 16.1

16.1.1 Model IV(a): Purchase inventory model with one price break

Consider the table 16.1

where b is the quantity at and beyond which the quantity discount applies. Obviously, $P_2 < P_1$. For any purchase quantity q_1 in the range $1 \leq q_1 < b$,

$$C(q_1) = \frac{C_3 R}{q_1} + P_1 R + \frac{P_1 q_1 I}{2} \tag{16.1.7}$$

Similarly, for q_2,

$$C(q_2) = \frac{C_3 R}{q_2} + P_2 R + \frac{P_2 q_2 I}{2} \tag{16.1.8}$$

Rule I Compute q_2^*, using (16.1.5). If $q_2 \geq b$, then the optimum lot size will be q_2^*.

Rule II If, $q_2 < b$, then the quantity discount no longer applies to the purchase quantity q_2^*. Compute q_1^*, then compare $C(q_1^*)$ and $C(b)$ given by,

$$\begin{aligned} C(q_1^*) &= \frac{C_3 R}{q_1^*} + P_1 R + \frac{P_1 q_1^* I}{2} \\ C(b) &= \frac{C_3 R}{b} + \frac{P_2 I b}{2} + P_2 R \end{aligned}$$

It shows that,

$$\frac{C_3 R}{b} + P_2 R < \frac{C_3 R}{q_1^*} + P_1 R \qquad [\text{since } q_1^* < b \text{ and } P_2 < P_1]$$

However, $\frac{P_2 I b}{2}$ may or may not be less than $\frac{P_1 I q_1^*}{2}$. Hence, we must compare the total cost. So, $q^* = b$.

Example 16.1.1. Find the optimum order quantity for a product for which the price breaks are as follows:

Quantity discount	Unit Cost (Rs.)
$0 \leq q_1 < 500$	10.00
$500 \leq q_2$	9.25

The monthly demand for a product is 200 units, the cost of storage is 2% of unit cost and cost of ordering is Rs. 350.

Solution.

$R = 200$ units per month

$I = Rs.\ 0.02$

$C_3 = Rs.\ 350$

$P_1 = Rs.\ 10.00$

$P_2 = Rs.\ 9.25$

$$q_2^* = \sqrt{\frac{2C_3 R}{P_2 I}} = \sqrt{\frac{2 \times 350 \times 200}{9.25 \times 0.02}} = 870 \text{ units} > b = 500.$$

Since $q_2^* = 870$ lies within the range $q_2 \geq 500$, hence the optimum purchase quantity will be $q_2^* = 870$ units. ■

Example 16.1.2. Same as the previous example with $C_3 = Rs.\ 100$. Thus,

$$q_2^* = \sqrt{\frac{2C_3R}{P_2I}} = \sqrt{\frac{2 \times 100 \times 200}{9.25 \times 0.02}} = 447 \text{ units} < b = 500.$$

Then compare C(447) with C(500), that is, the optimum cost of procuring the least quantity which will entitle or price break, that is,

$$C(q^*) = C(447) = Rs.\ 2090.42$$
$$C(500) = Rs.\ 1937.25.$$

Since C(500)<C(447), the optimum purchase quantity will be $q^* = b = 500$.

16.1.2 Model IV(b): Purchase inventory model with two price breaks

Purchase Cost P per item	Range of quantity
P_1	$1 \leq q_1 < b_1$
P_2	$b_1 \leq q_2 2 < b_2$
P_3	$b_2 \leq q_3$

Table 16.2

Consider the table 16.2, where b_1 and b_2 are the quantities which determine the price breaks. The working rule is as follows:

Step 1: Compute q_3^* and compare with b_2.

(i) If $q_3^* \geq b_2$, then the optimum purchase quantity is q_3^*.

(ii) If $q_3^* < b_2$, then go to step 2.

Step 2: Compute q_2^*, since $q_3^* < b_2$ and q_2^* is also less than b_2 because $q_1^* < q_2^* < \ldots < q_n^*$ in general. Thus, there are only two possibilities when $q_2^* < b_2$, that is, either $q_2^* \geq b_1$ or $q_2^* < b_1$.

(i) When $q_2^* < b_2$ but $\geq b_1$, then proceed as in case of one price- break only, that is, compare the cost $C(q_2^*)$ and $C(b_2)$ to obtain the optimum purchase quantity.

The quantity with least cost will naturally be optimum.

(ii) If $q_2^* < b_2$ and b_1, then go to step 3.

Step 3: If $q_2^* < b_2$ (and b_1 both). Then compute q_1^* which will satisfy the inequality $q_1^* < b_1$. In this case, compare the cost $C(q_1^*)$ with $C(b_1)$ and $C(b_2)$ both to determine the optimum purchase quantity.

Example 16.1.3. Find the optimum order quantity for a product for which the price breaks are in table 16.3. The monthly demand for a product is 200 units. The cost of storage is 2% of the unit cost. Cost of ordering is Rs. 350.

Quantity	:	$0 \leq q_1 < 500$	$500 \leq q_2 < 750$	$750 \leq q_3$
Unit Price(Rs.)	:	10.00	9.25	8.75

Table 16.3

Solution.

$$
\begin{aligned}
&R = 200 \text{ units per month} \\
&I = Rs.\ 0.02 \\
&C_3 = Rs.\ 350 \\
&P_1 = Rs.\ 10.00 \\
&P_2 = Rs.\ 9.25 \\
&P_3 = Rs.\ 8.75 \\
&b_1 = 500 \\
&b_2 = 750 \\
&q_3^* = \sqrt{\frac{2 \times 350 \times 200}{8.75 \times 0.02}} = 894 > 750.
\end{aligned}
$$

Thus, the optimum purchase quantity will be $q^* = 894$.

If we choose $C_3 = Rs.\ 100$, and all are the same, then

$$q_3^* = \sqrt{\frac{2 \times 100 \times 200}{8.75 \times 0.02}} = 478 < 750.$$

Step 2:

$$q_2^* = \sqrt{\frac{2 \times 100 \times 200}{9.25 \times 0.02}} = 465 < 500.$$

Again, we compute

$$q_1^* = \sqrt{\frac{2 \times 100 \times 200}{10 \times 0.02}} = 447 < 500.$$

Then, we compare C(447) with C(500) and C(750). Now,

$$C(447) = Rs.\ 2090.42, \quad C(500) = Rs.\ 1937.25, \quad C(750) = Rs.\ 1843.29$$

Thus, $C(750) < C(500) < C(q_1^*)$. This shows that, the optimum purchase quantity is $q^* = 750$ units. ■

16.2 Probabilistic Inventory Model

16.2.1 Instantaneous demand, no set up cost

Discrete Case

Find the optimum order level z which minimizes the total expected cost under the following assumptions

(i) t is the constant interval between orders. (daily, monthly, weekly, etc.)

(ii) z is the stock at the beginning of each period t

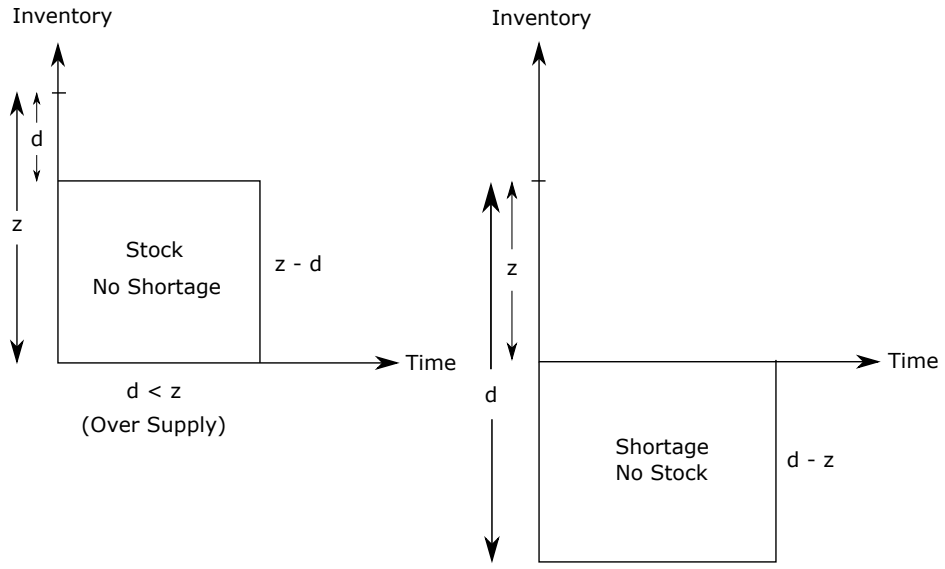
(iii) d is the estimated (random) demand at a discontinuous rate with probability P(d)

(iv) C_1 is holding cost

(v) C_2 is shortage cost

(vi) lead time zero

(vii) demand is instantaneous.



In the model with instantaneous demand, it is assumed that the total demand is fulfilled at the beginning of the period. Thus, depending on the demanded amount the inventory position may either be positive (surplus or stock) or negative (shortage).

Case I: $d \leq z$

$$\begin{aligned} \text{Holding cost} &= (z-d)\,C_1, \qquad \text{for } d \leq z \\ &= C_1 \times 0, \qquad \text{for } d > z \text{ (no stock)} \end{aligned}$$

Case II: $d > z$

$$\begin{aligned} \text{Shortage cost} &= C_2 \times 0 \qquad \text{for } d \leq z \text{ (no shortage)} \\ &= (d-z)\,C_2 \qquad \text{for } d > z \end{aligned}$$

To get the expected cost, we have to multiply the cost by given probability $P(d)$. Further to get the total expected cost we must sum over all the expected cost. So, the total expected cost per unit time is,

$$\begin{aligned} C(z) &= \sum_{d=0}^{z}(z-d)\,C_1P(d) + \sum_{d=z+1}^{\infty} C_1 \cdot 0 \cdot P(d) + \sum_{d=0}^{z} C_2 \cdot 0 \cdot P(d) + \sum_{d=z+1}^{\infty} C_2 \cdot (d-z)\,P(d) \\ &= \sum_{d=0}^{z}(z-d)\,C_1P(d) + \sum_{d=z+1}^{\infty} C_2 \cdot (d-z)\,P(d) \qquad (16.2.1) \end{aligned}$$

For the minimum of $C(z)$, the following must be satisfied:

$$\Delta C\,(z-1) < 0 < \Delta C\,(z) \text{ (finite difference Calculus)} \tag{16.2.2}$$

But, we can difference (16.2.1) under the summation sign for $d = z+1$, the following condition satisfied

$$C_1\{(z+1)-d\}P(d) = C_2(d-(z+1))P(d).$$

Now,

$$\begin{aligned}
\Delta C(z) &= C_1\sum_{d=0}^{z}[((z+1)-d)-(z-d)]P(d) + C_2\sum_{d=z+1}^{\infty}[(d-(z+1))-(d-z)]P(d) \\
&= C_1\sum_{d=0}^{z}P(d) - C_2\sum_{d=z+1}^{\infty}P(d) \\
&= C_1\sum_{d=0}^{z}P(d) - C_2\left[\sum_{d=0}^{\infty}P(d) - \sum_{d=0}^{z}P(d)\right] \\
&= (C_1+C_2)\sum_{d=0}^{z}P(d) - C_2. \quad \left[\text{since } \sum_{d=0}^{\infty}P(d) = 1\right]
\end{aligned}$$

$$\begin{aligned}
&\Delta C(z) > 0 \\
\Rightarrow\ &(C_1+C_2)\sum_{d=0}^{z}P(d) - C_2 > 0 \\
\Rightarrow\ &\sum_{d=0}^{z}P(d) > \frac{C_2}{C_1+C_2}
\end{aligned} \tag{16.2.3}$$

Similarly,

$$\begin{aligned}
&\Delta C(z-1) < 0 \\
&\sum_{d=0}^{z-1}P(d) < \frac{C_2}{C_1+C_2}.
\end{aligned}$$

Combining, we get

$$\sum_{d=0}^{z-1}P(d) < \frac{C_2}{C_1+C_2} < \sum_{d=0}^{z}P(d). \tag{16.2.4}$$

Example 16.2.1. **(Newspaper boy problem)** A newspaper boy buys papers for Rs. 2.60 each and sells them for Rs. 3.60 each. He can not return unsold newspapers. Daily demand has the following probability distribution (Table 16.4).

No. of customers	:	23	24	25	26	27	28	29	30	31	32
Probability	:	0.01	0.03	0.06	0.10	0.20	0.25	0.15	0.10	0.05	0.05

Table 16.4

If each day, demand is independent of the previous days, how many papers should be ordered each day?

Solution. Let z=The number of newspapers ordered per day and d=demand that is, the number that could be sold per day if $z \geq d$, $P(d)$=The probability that the demand will be equal to on a randomly selected day,

$$\begin{aligned} C_1 &= \text{Cost per newspaper} \\ C_2 &= \text{Selling price per newspaper.} \end{aligned}$$

If the demand d exceeds z, his profit would become equal to $(C_2 - C_1)z$, and no newspaper will be let unsold. On the other hand, if d does not exceed z, his profit becomes equal to $(C_2 - C_1)d - (z - d)C_1$, where $(C_2 - C_1)d$ is for the sold papers and $(z - d)C_1$ for the unsold papers. Then the expected net profit per day becomes equal to

$$P(z) = \sum_{d=0}^{z} (C_2 d - C_1 z)P(d) + \sum_{d=z+1}^{\infty} (C_2 - C_1)zP(d).$$

where, $(C_2 d - C_1 z)P(d)$ is for $d \leq z$ and $(C_2 - C_1)zP(d)$ for $d > z$.

Using finite difference calculus, we know that the condition for maximum value of $P(z)$ is

$$\Delta P(z-1) > 0 > \Delta P(z).$$

$$\begin{aligned} \Delta P(z) &= \sum_{d=0}^{z} \left[\{C_2 d - C_1(z+1)\} - (C_2 d - C_1 z)\right] P(d) + \sum_{d=z+1}^{\infty} (C_2 - C_1)\{(z+1) - z\}P(d) \\ &= -C_1 \sum_{d=0}^{z} P(d) + (C_2 - C_1) \sum_{d=z+1}^{\infty} P(d) \\ &= -C_1 \sum_{d=0}^{z} P(d) + (C_2 - C_1) \left\{ \sum_{d=0}^{\infty} P(d) - \sum_{d=0}^{z} P(d) \right\} \\ &= -C_1 \sum_{d=0}^{z} P(d) + (C_2 - C_1) \left\{ 1 - \sum_{d=0}^{z} P(d) \right\} \\ &= -C_2 \sum_{d=0}^{z} P(d) + (C_2 - C_1). \end{aligned}$$

For, maximum of $P(z)$,

$$\begin{aligned} & \Delta P(z) < 0 \\ \text{or,} \quad & -C_2 \sum_{d=0}^{z} P(d) + (C_2 - C_1) < 0 \\ \text{or,} \quad & \sum_{d=0}^{z} P(d) > \frac{C_2 - C_1}{C_2}. \end{aligned} \tag{16.2.5}$$

Similarly, we can find,

$$\sum_{d=0}^{z-1} P(d) < \frac{C_2 - C_1}{C_2}.$$

Combining, we get,

$$\sum_{d=0}^{z} P(d) > \frac{C_2 - C_1}{C_2} > \sum_{d=0}^{z-1} P(d).$$

In this problem, $C_1 = Rs.\ 2.60,\ C_2 = Rs.\ 3.60$. The lower limit for demand d is 23 and upper limit is 32. Therefore, substituting these values in (16.2.5), we get,

$$\sum_{d=0}^{z} P(d) > \frac{3.60 - 2.60}{3.60} = 0.28.$$

Now, we can easily verify that this inequality holds for $z = 27$, that is,

$$\begin{aligned} \sum_{d=23}^{27} P(d) &= P(23) + P(24) + P(25) + P(26) + P(27) \\ &= 0.01 + 0.03 + 0.06 + 0.10 + 0.20 = 0.40 > 0.28. \end{aligned}$$

Similarly,

$$\sum_{d=23}^{26} P(d) = 0.20 < 0.28.$$

■

Continuous Case

This model is same as the previous model except that the stock levels are now assumed to be continuous quantities. So, instead of probability $P(d)$, we shall have $f(x)dx$ and in place of summation, we take integration, where $f(x)$ is the pdf (probability density function). The cost equation for this model becomes

$$C(z) = C_1 \int_0^z (z - x) f(x) dx + C_2 \int_z^\infty (x - z) f(x) dx. \tag{16.2.6}$$

The optimal value of z is obtained by equating z to zero the first derivative of $c(z)$, that is, $\dfrac{dC}{dz} = 0$.

Differentiating (16.2.6), we get,

$$\begin{aligned} \frac{dC}{dz} &= C_1 \int_0^z (1 - 0) f(x) dx + C_1 \left[(z - x) f(x) \frac{dx}{dz} \right]_0^z + C_2 \int_z^\infty (0 - 1) f(x) dx + C_2 \left[(x - z) f(x) \frac{dx}{dz} \right]_z^\infty \\ &= C_1 \int_0^z f(x) dx - C_2 \int_0^\infty f(x) dx \\ &= C_1 \int_0^z f(x) dx - C_2 \left[1 - \int_0^z f(x) dx \right] \\ &= (C_1 + C_2) \int_0^z f(x) dx - C_2. \end{aligned}$$

Thus,

$$\begin{aligned} & \frac{dC}{dz} = 0 \\ \Rightarrow\ & (C_1 + C_2) \int_0^z f(x) dx - C_2 \\ \Rightarrow\ & \int_0^z f(x) dx = \frac{C_2}{C_1 + C_2} \\ & \frac{d^2C}{dz^2} = (C_1 + C_2) \left[f(x) \frac{dx}{dz} \right]_0^z = (C_1 + C_2) f(x) > 0. \end{aligned}$$

Hence, we can get optimum value of z satisfying the sufficient condition for which the total expected cost C is minimum.

Example 16.2.2. A baking company sells cake by the kg weight, it makes a profit of Rs 5.00 per kg on each kg sold on the day it is baked. It disposes off all cakes not sold on the day it is baked at a loss of Rs. 1.20 per kg. If demand is known to be rectangular between 2000 and 3000 kgs, determine the optimal daily amount baked.

Solution.

$$\begin{aligned} C_1 &= \text{profit per kg cake} \\ C_2 &= \text{loss per kg cake for unsold cake} \\ x &= \text{Demand which is continuous with pdf } f(x), \end{aligned}$$

where,

$$\int_{x_1}^{x_2} f(x)dx = \text{the probability of an order within } x_1 \text{ to } x_2.$$

and z=stock level.

Then two cases arise.

Case I: If $x \leq z$, then clearly the demand x is satisfied and unsold $(z - x)$ quantities are returned with a loss of C_2 per kg, so, profit is $C_1 x$ and loss is $C_2(z - x)$. Hence the net profit becomes, $C_1 x - C_2(z - x) = (C_1 + C_2)x - C_2 z$.

Case II: If $x > z$, then the net profit becomes $C_1 z$. Thus, the total expected profit is given by

$$P(z) = \int_{x_1}^{z} \left[(C_1 + C_2)x - C_2 z\right] f(x)dx + \int_{z}^{x_2} C_1 z f(x)dx = P_1(z) + P_2(z) \text{ (say)}.$$

Now, for the maximum value of $P(z)$, we must have,

$$\frac{dP(z)}{dz} = \frac{d}{dz}P_1(z) + \frac{d}{dz}P_2(z) = 0$$

Now,

$$\begin{aligned} P_1(z) &= \int_{x_1}^{z} \left[(C_1 + C_2)x - C_2 z\right] f(x)dx \\ \frac{d}{dz}P_1(z) &= \int_{x_1}^{z} (0 - C_2) f(x)dx + \left[\{(C_1 + C_2)x - C_2 z\} f(x)\frac{dx}{dz}\right]_{x_1}^{z} \\ &= -C_2 \int_{x_1}^{z} f(x)dx + \{(C_1 + C_2)x - C_2 z\} f(x) \\ &= -C_2 \int_{x_1}^{z} f(x)dx + C_1 z f(z). \end{aligned}$$

Similarly,

$$\begin{aligned} \frac{d}{dz}P_2(z) &= \int_{z}^{x_2} C_1 f(x)dx + \left[C_1 z f(z)\frac{dx}{dz}\right]_{z}^{x_2} \\ &= C_1 \int_{z}^{x_2} f(x)dx - C_1 z f(z). \end{aligned}$$

Hence, we have,

$$\frac{dP(z)}{dz} = \left[-C_2 \int_{x_1}^{z} f(x)dx + C_1 z f(z)\right] + \left[C_1 \int_{z}^{x_2} C_1 f(x)dx - C_1 z f(z)\right] = 0$$

$$\Rightarrow \quad -C_2 \int_{x_1}^{z} f(x)dx + C_1 \int_{z}^{x_2} f(x)dx = 0$$

$$\Rightarrow \quad -C_2 \int_{x_1}^{z} f(x)dx + C_1 \left\{\int_{x_1}^{x_2} f(x)dx - \int_{x_1}^{z} f(x)dx\right\} = 0$$

$$\Rightarrow \quad -(C_1 + C_2) \int_{x_1}^{z} f(x)dx + C_1 = 0$$

$$\Rightarrow \quad \int_{x_1}^{z} f(x)dx = \frac{C_1}{C_1 + C_2} \tag{16.2.7}$$

Also,

$$\frac{d^2P(z)}{dz^2} = -(C_1 + C_2)f(z) < 0$$

satisfies the sufficient condition of maximum of $P(z)$.

In this problem,

$$C_1 = \text{Rs. } 5.00, \;\; C_2 = \text{Rs. } 1.20, \;\; x_1 = 2000, \;\; x_2 = 3000.$$

$$f(x) = \frac{1}{x_2 - x_1} = \frac{1}{1000}.$$

Substituting these values in equation (16.2.7), we have

$$\int_{2000}^{z} \frac{1}{1000} dx = \frac{5}{5 + 1.20} = 0.807$$

$$\Rightarrow \quad \frac{1}{1000}(z - 2000) = 0.807$$

$$\Rightarrow \quad z = 2807 \text{ kg}.$$

■

References

1. An Introduction to Information Theory - F. M. Reza.
2. Operations Research : An Introduction - P. K. Gupta and D.S. Hira.
3. Graph Theory with Applications to Engineering and Computer Science - N. Deo.
4. Operations Research - K. Swarup, P. K. Gupta and Man Mohan.
5. Coding and Information Theory - Steven Roman.
6. Coding Theory, A First Course - San Ling r choaping Xing.
7. Introduction to Coding Theory - J. H. Van Lint
8. The Theory of Error Correcting Codes - Mac William and Sloane.
9. Information and Coding Theory - Grenth A. Jones and J. Marry Jones.
10. Information Theory, Coding and Cryptography - Ranjan Bose.

POST GRADUATE DEGREE PROGRAMME (CBCS) IN

MATHEMATICS

SEMESTER IV

SELF LEARNING MATERIAL

PAPER : MATO 4.3
(Applied Stream)

Mathematical Biology

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

Course Preparation Team

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

May, 2020

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, 2017 had been our endeavour. We are happy to have achieved our goal.

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

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

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

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

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

Director

Directorate of Open and Distance Learning

University of Kalyani

Optional Paper

MATO 4.3
Marks : 100 (SEE : 80; IA : 20)

Mathematical Biology
(Applied Stream)

Syllabus

• Unit 1 •
Epidemic models: Simple epidemic; SIS epidemic model; SIS epidemic model with specific rate of infection, SIS epidemic model with constant number of carriers.

• Unit 2 •
General epidemic model; Approximate solution, Recurring epidemic model.

• Unit 3 •
Stochastic epidemic models without removal, Basic system of equations and its solution.

• Unit 4 •
Stochastic epidemic models: with multiple infections; Removal; Carriers; Infectives, immigration and emigration.

• Unit 5 •
Basic model for inheritance of genetic characterises, Hadly-Wienberg law.

• Unit 6 •
Correlation between genetic composition of siblings, Bayes theorem and its applications in genetics

• Unit 7 •
Extension of basic model for inheritance of genetic characteristics, Models for genetic improvement

• Unit 8 •
Genetic Improvement through elimination of Recessives, Selection and Mutation, Alternative Discussion of selection.

• Unit 9 •
Some basic concepts of fluid dynamics, Hegen-Poiseuille Flow, Reynolds number Flow, Non-Newtonian Fluids

• Unit 10 •

Basic concepts about blood, Cardiovascular system, Special Characteristics of Blood flow, Structure and mechanical properties of blood vessels

• Unit 11 •

Non-Newtonian Flow in Circular Tubes, Power-Law, Herschel-Bulkley and Casson fluid flow in circular tubes.

• Unit 12 •

Fahraeus-Lindqvist Effect, Pulsatile Flow in Circular Rigid Tube, Blood Flow through Artery with Mild Stenosis.

• Unit 13 •

Peristaltic Motion in a Channel and Tube, Long-wavelength approximation.

• Unit 14 •

Two-dimensional Flow in Renal Tubule, Function of Renal Tubule, Basic Equations and Boundary Conditions, Solution under approximations.

• Unit 15 •

Diffusion and Diffusion-Reaction Models, Ficks Law of Diffusion, Solutions of One and Two-dimensional Diffusion Equation.

• Unit 16 •

Diffusivity of Population Models, Diffusion on stability of Single Species, Two Species and Prey-Predator Model

Contents

Unit 1

Course Structure

- Terminologies related to epidemic
- Simple epidemic
- SIS Epidemic Model
- SIS Epidemic Model with Specific Rate of Infection as a Function of time
- SIS Model with Constant Number of Carriers
- Simple Epidemic Model with Carriers

1.1 Introduction

The study of mathematical theory of epidemic can be look upon as a continuation of our previous study in the sense that here also our concern is about the population sizes when effected by epidemics. In fact, we will draw our attention in modelling of problems of epidemics in mathematical terms. Sometimes such study is also called the study of mathematical epidemiology.

In order to pose a problem of epidemic, let us think of a small group of individuals who can carry a communicable infection to a large group of individuals, who can therefore be consider to be capable of the conducting the disease. Our immediate problem is to investigate how the disease is develop. In order to have a mathematical model of such situation we need some assumption regarding the characteristic of the disease as well as the mixing of the population. For this we need to consider the basic definition.

• **Susceptible Individuals:** An individual who is capable of conducting the disease directly or indirectly from another infected individual and is thereby become an infectious.

• **Infective Individuals:** An individual who is capable of transmitting the disease to others.

• **Removed Individuals:** An individual who had the disease and has recover or is death and is permanently immune or is latent (existing but not developed) until recovering an permanent immunity occurs.

• **Latent Period:** This it the period during which a disease is developed within a newly infected individual in purely internal way.

• **Infections Period:** This is the period during which the infected is liable to communicate infectious material to susceptible.

• **Incubation Period:** This is the interval between the exposer to disease and the appearance of symptoms.

• **Genial Period:** This is the time interval between the appearance of symptoms in one case and the appearance of symptoms in another case infected from the first.

Remark 1.1.1. The disease under consideration confers permanent immunity upon any individuals who has completely recovers from it and has a negligible short incubation period.

An individuals who conduct the disease becomes infective immediately.

The population is obviously divided into three classes, viz. susceptible class, infective class and removed class.

We next consider some simple cases depending on the nature of the epidemic. We consider three types of epidemic, viz. *simple*, *general* and *recurring* epidemic.

1.2 Simple Epidemic Model

This is the simplest type of epidemic in which a disease may spread among a group of susceptible but there is removal by death or by recovering or by isolation. In reality this may be taken as the reasonable approximation to the early stages of the upper respiratory infection. Over a long time may elapse before an infective is removed.

Let there n susceptible (S) and let us introduce an simple infective (I) into this group at time $t = 0$, so that we have a group of $(n + 1)$ individuals. Let $S(t)$ and $I(t)$ be the respective number of susceptible and infective at time t so that

$$S(t) + I(t) = n + 1 \tag{1.2.1}$$

We now assume that the disease spread in such a way that the average number of new cases of the disease in an interval Δt is proportional to both the number of susceptible and the infective.

Let $\gamma > 0$ be the constant rate between the members at time Δt so that

$$\Delta S = -\gamma SI\Delta t \tag{1.2.2}$$

Procedding to the limit $\Delta t \to 0$, we have

$$\frac{dS}{dt} = -\gamma SI \tag{1.2.3}$$

which can also be written as

$$\begin{aligned} \frac{dS}{d\tau} &= -SI \quad \text{where } \tau = \gamma t \\ \Rightarrow \frac{dS}{d\tau} &= -S[n + 1 - S] \quad [\text{using Eq. (1.2.1)}] \end{aligned} \tag{1.2.4}$$

The solution of Eq. (1.2.4) subject to the initial conditions $S = n$ at $\tau = 0$ is given by

$$S = \frac{n(n+1)}{n + e^{(n+1)\tau}} \tag{1.2.5}$$

Therefore the rate which new cases occur is given by

$$-\frac{dS}{d\tau} = S[n + 1 - S] = \frac{n(n+1)^2 e^{(n+1)\tau}}{\{n + e^{(n+1)\tau}\}^2} \tag{1.2.6}$$

The rate $\dfrac{dS}{d\tau} < 0$ because it represents change in S, and S, the number of susceptible is decreasing as the epidemic develop.

Fig. 1.1 is known as epidemic curve and has a maximum at $\tau = \dfrac{\ln(n)}{n+1}$. We therefore, conclude that the rate appearance of the new cases increases rapidly to begin with rises to a maximum and there after falls to zero.

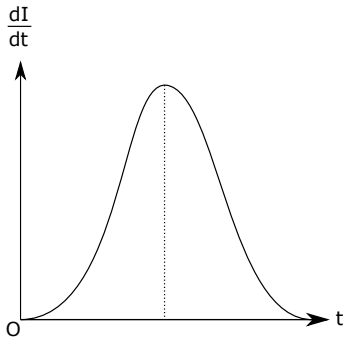


Figure 1.1: Epidemic Curve

Remark 1.2.1. The above analysis does not less tell us the rate at which the infection is spreading. To do this we take the basic equation

$$S + I = N \tag{1.2.7}$$

$$\frac{dS}{dt} = -\gamma SI \tag{1.2.8}$$

where N is size of total population. Therefore we have,

$$\begin{aligned} &\frac{d}{dt}(N - I) = -\gamma(N - I)I \\ \Rightarrow &\frac{dI}{dt} = \gamma(N - I)I \quad [\because N \text{ is time independent}] \end{aligned} \tag{1.2.9}$$

On integration of the Eq. (1.2.9), with the condition $I(0) = 1$, we have

$$I(t) = \frac{N}{(N-1)e^{-\gamma Nt} + 1} \tag{1.2.10}$$

Since γ is positive, $I(t)$ goes to N as $t \to \infty$ one can conclude that every individual in the population will eventually contact the disease. Thus one can calculate $S(t)$ using Eq. (1.2.7) in the form

$$S(t) = \frac{N(N-1)e^{-\gamma Nt}}{1 + (N-1)e^{-\gamma Nt}} \tag{1.2.11}$$

The rate at which the infection takes place is given by

$$\frac{dI}{dt} = \frac{N^2(N-1)\gamma e^{-\gamma Nt}}{\left[1+(N-1)e^{-\gamma Nt}\right]^2} \tag{1.2.12}$$

The curve representing $\frac{dI}{dt}$ vs t is known as the epidemic curve. Now to investigate the maximum value, the rate at which the infections takes place let us compute

$$\frac{d^2I}{dt^2} = \left[\frac{(N-1)N^3\gamma^2 e^{-\gamma Nt}}{1+(N-1)e^{-\gamma Nt}}\right]\left[(N-1)e^{-\gamma Nt}-1\right] \tag{1.2.13}$$

Now the factor inside the first bracket is positive for all values of t. Thus the sign of $\dfrac{d^2I}{dt^2}$ only depends on the other factor namely $\left[(N-1)e^{-\gamma Nt}-1\right]$. At $t=0$, this factor is positive and becomes negative when $t\to\infty$, so there exist an extreme value when $(N-1)e^{-\gamma Nt}-1=0$. So the rate has a maximum at

$$t = \frac{\ln(N-1)}{\gamma N} = t_{max} \tag{1.2.14}$$

Then $\left(\dfrac{dI}{dt}\right)_{max} = \dfrac{N^2\gamma}{4}$ and $I_{max} = \frac{N}{2}$.

Note:

1. One can note from the expression of t_{max} that if γ is small t_{max} tends to the large, i.e., smaller the γ longer it take to reach the peak value. Also the epidemic will be complete in a much shorter time for a dense population than for a sparse one.

2. A serious limitation of this epidemic model is that everyone in the population will contact the disease as many susceptible still remain in the population.

1.3 SIS Epidemic Model

In this model, a susceptible person can become infected at a rate proportional to SI and an infected person can recover and become susceptible again at a rate γI so that we get the model

$$\frac{dS}{dt} = -\beta SI + \gamma I, \tag{1.3.1}$$

$$\frac{dI}{dt} = \beta SI - \gamma I \tag{1.3.2}$$

which gives

$$S(t)+I(t) = N = S(0)+I(0) = S_0+I_0 \quad (I_0 \neq 0). \tag{1.3.3}$$

From (1.3.1)-(1.3.3),

$$\frac{dI}{dt} = (\beta N-\gamma)I - \beta I^2 = kI - \beta I^2. \tag{1.3.4}$$

Integrating (1.3.4), we obtain

$$I(t) = \begin{cases} \dfrac{\exp(kt)}{\beta[\exp(kt)-1]/k + I_0^{-1}} & (k\neq 0) \\ \dfrac{1}{\beta t + I_0^{-1}} & (k=0) \end{cases} \tag{1.3.5}$$

As $t\to\infty$,

$$I(t) \to \begin{cases} N-\rho & \text{if} \quad N>\rho=\gamma/\beta \\ 0 & \text{if} \quad N\leq\rho=\gamma/\beta \end{cases} \tag{1.3.6}$$

1.3.1 SIS Model with Specific Rate of Infection as a Function of t

In this case, (1.3.4) becomes

$$\frac{dI}{dt} = \Big[\beta(t)N - \gamma\Big]I - \beta(t)I^2 \tag{1.3.7}$$

or

$$\frac{dJ}{dt} + \Big[\beta(t)N - \gamma\Big]J = \beta(t), \tag{1.3.8}$$

where

$$J(t) = \big[I(t)\big]^{-1}. \tag{1.3.9}$$

Integrating (1.3.8), we get

$$J(t)\left[\exp\left\{\int_0^t \big[\beta(t)N - \gamma\big]dt\right\}\right] = \int_0^t \beta(t)\left[\exp\left\{\int_0^t [\beta(t)N - \gamma]dt\right\}\right]dt + J_0. \tag{1.3.10}$$

Simplifying this equation and using (1.3.9)

$$I(t) = \frac{\exp\left[N\int_0^t \beta(u)du - \gamma t\right]}{\int_0^t \beta(v)\exp\left[N\int_0^v \beta(u)du - \gamma v\right]dv + I_0^{-1}}. \tag{1.3.11}$$

1.3.2 SIS Model with Constant Number of Carriers

In this model, infection is spread both by infectives and a constant number C of carriers so that (1.3.1) and (1.3.2) becomes

$$\frac{dI}{dt} = \beta(I + C)S - \gamma I = \beta CN + \beta(N - C - \rho)I - \beta I^2. \tag{1.3.12}$$

Integrating, we get

$$I(t) = \frac{\alpha_1(I_0 - \alpha_2)e^{\beta\alpha_1 t} + \alpha_2(\alpha_1 - I_0)e^{\beta\alpha_2 t}}{(I_0 - \alpha_2)e^{\beta\alpha_1 t} + (\alpha_1 - I_0)e^{\beta\alpha_2 t}} \tag{1.3.13}$$

where

$$\alpha_1, \alpha_2 = \frac{1}{2}\Big[(N - C - \rho) \pm \{(N - C - \rho)^2 + 4CN\}\Big]^{1/2}\Big]; \tag{1.3.14}$$

α_1, α_2 correspond to the positive and negative roots respectively of the equations $I^2 - (N - C - \rho)I - NC = 0$. Now, as $t \to \infty$,

$$I(t) \to \alpha_1 \tag{1.3.15}$$

so that $I(t)$ is asymptotic to a positive constant for all values of N and ρ. Thus, with a constant number of carriers, $I(t)$ does not tend to zero.

1.4 Simple Epidemic Model with Carriers

In this model, only carriers spread the disease and their number decreases exponentially with time as they are identified and eliminated. Here, if $S(t),\ I(t)$ and $C(t)$ respectively represent the number of susceptibles,

infectives, and carriers at time t, we have

$$\begin{aligned} \frac{dS}{dt} &= -\beta C(t)S(t) + \gamma I(t), \\ \frac{dI}{dt} &= \beta C(t)S(t) - \gamma I(t), \\ \frac{dC}{dt} &= -\alpha C \end{aligned} \tag{1.4.1}$$

so that

$$\begin{aligned} S(t) + I(t) = S_0 + I_0 = N, \qquad C(t) = C_0 \exp[-\alpha t], \\ \frac{dI}{dt} = \beta C_0 N \exp(-\alpha t) - [\beta C_0 \exp(-\alpha t) + \gamma] I \end{aligned} \tag{1.4.2}$$

whose solution is

$$I(t) = \frac{\beta C_0 N \int_0^t \exp[-\alpha v - \beta C_0 \exp(-\alpha v)/\alpha + \gamma v] dv + I_0 \exp(-\beta C_0/\alpha)}{\exp[-(\beta C_0/\alpha)\exp(-\alpha t) + \gamma t]} \tag{1.4.3}$$

It can be now shown that

$$\lim_{t \to \infty} I(t) = 0. \tag{1.4.4}$$

Unit 2

Course Structure

- General Epidemic Model
- Approximate Solution
- Recurring Epidemic

2.1 General Epidemic Model

The study of general epidemic involves with infection as well as removal. Let us assume that $S(t),\ I(t)$ and $R(t)$ be the respective population sizes of susceptible, infected and removal individual at time t.

Let us make few assumption about the nature of $S,\ I$ and R as follows:

- The population is treated as closed (constant) and continious which can be represented by S moving over to I moving over to R (we ignore both birth and immigration).
- The rate of change of susceptible population is proportional to the number of contacts between the members of the class S and I, in which we take in term, the number of contacts to be proportional to the product of the numbers of S and I. This assumption takes care of uniform mixing of the population.
- Individuals are recovered at a rate proportional to the number I.

Let $r > 0$ be the infective rate and $\gamma > 0$ be the removed rate and if S_0, I_0 be the initial number of members of S and I respectively, then the governing equations are given by

$$\frac{dS}{dt} = -rSI \tag{2.1.1}$$

$$\frac{dI}{dt} = rSI - \gamma I \tag{2.1.2}$$

$$\frac{dR}{dt} = \gamma I \tag{2.1.3}$$

We are to study these equation with the following conditions given by $S = S_0,\ I = I_0$ and $R = 0$ initially at $t = 0$. In addition to these we have

$$S(t) + I(t) + R(t) = \text{constant} \quad i.e.,\ \frac{d}{dt}(S + I + R) = 0 \tag{2.1.4}$$

From Eq. (2.1.2), we have

$$\frac{dI}{dt} = r\left(S - \frac{\gamma}{r}\right) I \tag{2.1.5}$$

If $S_0 < \frac{\gamma}{r}$ then $\frac{dI}{dt} < 0$ and since $S(t) < S_0$, one can conclude that $\frac{dI}{dt} < 0$ for all t. Therefore, it is such a case in which the infection dies out, i.e., non epidemic takes place. This is known as a "*Threshold Phenomena*". We therefore conclude that there exist a critical value for which the initial susceptible has to exceed for their to be an epidemic, in other words the relative removal rate $\frac{\gamma}{r}$ must be sufficiently small so as to allow the epidemic to spread.

The Eqs. (2.1.1)–(2.1.4) also enable us to study another behaviour relative to spread of the disease. Since $S(t)$ is non-increasing and positive $\lim_{t\to\infty} S(t) \to S(\infty)$ exists and since $\frac{dR}{dt} \geq 0$ and $R(t) \leq N$ then $R(\infty)$ exists. Again we have $I(t) = N - R(t) - S(t)$ is follows the $\lim_{t\to\infty} I(t) \to 0$.

Now we consider some other values in dividing Eq. (2.1.1) by Eq. (2.1.3) when we have

$$\frac{dS}{dR} = -\frac{r}{\gamma} S \tag{2.1.6}$$

On integration we have

$$S = S_0 \exp\left\{-\frac{r}{\gamma} R\right\} \tag{2.1.7}$$

Now since $R \leq N$ which implies $-R \geq -N$, so that

$$S = S_0 \exp\left\{-\frac{r}{\gamma} R\right\} \geq S_0 \exp\left\{-\frac{r}{\gamma} N\right\} > 0 = \alpha\ \ (say). \tag{2.1.8}$$

Therefore, $\lim_{t\to\infty} S(t)$ is always positive, one can interpreted this by saying that there will always be susceptible remaining in the population. Thus we conclude that some individual will escape the disease all together and in particular the spread of disease does not stop for the lack of susceptible population. Let us consider a function

$$f(z) = S_0 \exp\left\{-\frac{1}{\rho}(N - z)\right\} - z \quad \text{in which} \quad \rho = \frac{\gamma}{r} \tag{2.1.9}$$

Now, $f(0) > 0$ and $f(N) = S_0 - N < 0$. Therefore, there must be a positive root for $f(z) = 0$. Let z_0 be the root, then we have

$$f'(z) = \frac{1}{\rho} S_0 \exp\left\{-\frac{1}{\rho}(N - z)\right\} - 1 \tag{2.1.10}$$

$$\text{and}\ \ f''(z) = \frac{1}{\rho^2} S_0 \exp\left\{-\frac{1}{\rho}(N - z)\right\} \tag{2.1.11}$$

Now since $f''(z) > 0$ and $f(N) < 0$, there is only one such root $z_0 < N$. Now we have seen that

$$S = S_0 \exp\left(-\frac{R}{\rho}\right) \quad i.e.,\ S_\infty = S_0 \exp\left\{-\frac{1}{\rho}(N - S_\infty)\right\} \tag{2.1.12}$$

Hence, we can say that S_∞ is the root of the equation $f(z) = 0$. Now we can sum up all the results in the form of a theorem as follows:

Theorem 2.1.1. If $S_0 < \rho$ then $I(t)$ decreases monotonically to zero. If $S_0 > \rho$ then the number of infective increases as time t increases and then tends monotonically to zero. Further $\lim_{t\to\infty} S(t)$ exits and S_∞ is a root of the transcendental equation.

Remark 2.1.1. The equation $\dfrac{dS}{dR} = -\dfrac{r}{\gamma}S$ can be solved under certain approximations when R is known.

2.2 Approximate Solution

We have the Eq. (2.1.6) as

$$\frac{dS}{dR} = -\frac{r}{\gamma}S \tag{2.2.1}$$

where S is given by the Eq. (2.1.7) as $S = S_0 \exp\left(-\dfrac{r}{\gamma}R\right)$. Now substituting Eq. (2.1.7) in the Eq. (2.1.3), we have

$$\frac{dR}{dt} = \gamma\left[N - S_0 \exp\left(-\frac{r}{\gamma}R\right) - R\right] \tag{2.2.2}$$

The Eq. (2.2.2) can be solved by standard method by taking some approximate value after expanding upto some power of R. But we are interested in looking for values of R when $t \to \infty$. We note that as $t \to \infty$, $\dfrac{dR}{dt} \to 0$. Further as $t \to \infty$, taking $S_0 \approx N$, we have

$$\begin{aligned} 0 &= \gamma\left[N - N\ \exp\left(-\frac{r}{\gamma}R\right) - R\right] \\ \text{or, } 0 &= \gamma\left[N - N\ \exp\left(-\frac{R}{\rho}R\right) - R\right] \quad (\because \rho = \frac{\gamma}{r}) \end{aligned}$$

Now we expand the exponential term in the right hand side in powers of $\dfrac{R}{\rho}$, which becoming smaller and smaller as $t \to \infty$ and can be approximate upto second power of $\dfrac{R}{\rho}$. Therefore, we have

$$\begin{aligned} 0 &\approx \gamma\left[N - N\left(1 - \frac{R}{\rho} + \frac{R^2}{2\rho^2}\right) - R\right] \\ \Rightarrow R &\approx N\left(\frac{R}{\rho} - \frac{R^2}{2\rho^2}\right) \\ \Rightarrow \frac{1}{N} &\approx \frac{2\rho - R}{2\rho^2} \\ \Rightarrow \frac{2\rho^2}{N} &\approx 2\rho - R \\ \Rightarrow R &\approx 2\rho\left(1 - \frac{\rho}{N}\right) \end{aligned}$$

This is approximate as $t \to \infty$ and hence one should get the ultimate size of the epidemic. If $\rho > N$, there is no true epidemic and hence the appearance of epidemic will be there only when $\rho < N$, i.e., when the effective removal rate is less than the initial number of susceptible and in this case all persons do not get infected. A stage may be reached when all the infected person are immediately removed. So in order of epidemic may

occur, we have $N = \rho + \gamma$, where $\gamma > 0$ is small. Thus we have

$$\begin{aligned} R(\infty) &\approx 2\rho\left(1 - \frac{\rho}{\rho+\gamma}\right) \\ &\approx 2\rho\left[1 - \left(1 + \frac{\gamma}{\rho}\right)^{-1}\right] \\ &\approx 2\rho\left[1 - 1 + \frac{\gamma}{\rho}\right] \\ &\approx 2\gamma \end{aligned}$$

This shows that the initial density of susceptible namely $S_0(= N = \rho + \gamma)$ is reduced to $S_\infty(= \rho - \gamma)$ which means that the final number of susceptible falls at a point as far below the threshold value ρ as originally it was above it. This is known as "*Kermack & McKendric Threshold Theorem*" .

Remark 2.2.1.

- The above theorem corresponds to the general observation of the epidemic tends to built up more rapidly for the density of susceptible is high on account of over crowding and the removal rate is relatively low because of the factors that ignorance and inadequate isolation.
- The Eq. (2.2.2) can also be integrated when the approximation is taken upto second powers to R.

Integration leading to approximate solution. We have

$$\frac{dR}{dt} = \gamma\left[N - R - S_0 \exp\left(-\frac{R}{\rho}\right)\right] \tag{2.2.3}$$

Substituting $\exp\left(-\frac{R}{\rho}\right) = 1 - \frac{R}{\rho} + \frac{R^2}{2\rho^2}$ into the above equation, one get

$$\begin{aligned} \frac{dR}{dt} &= \gamma\left[N - R - S_0\left(1 - \frac{R}{\rho} + \frac{R^2}{2\rho^2}\right)\right] \\ \Rightarrow \frac{dR}{dt} &= \gamma\left[N - S_0 + R\left(\frac{S_0}{\rho} - 1\right) - \frac{S_0}{2}\frac{R^2}{\rho^2}\right] \\ \Rightarrow \frac{dR}{dt} &= a + bR - cR^2 \end{aligned}$$

where $a = \gamma(N - S_0),\ b = \gamma\left(\frac{S_0}{\rho} - 1\right)$ and $c = \frac{\gamma S_0}{2\rho^2}$. On integration, we obtain

$$\begin{aligned} &\frac{2}{q}\tanh^{-1}\left(\frac{2cR - b}{q}\right) = t + c_1,\ \ c_1 \text{ being a constant and } q = \sqrt{b^2 + 4ac} \\ \Rightarrow\ &R(t) = \frac{1}{2c}\left[b + q\tanh\left(\frac{qt}{2} + c_2\right)\right],\ \ c_2 \text{ is a different constant} \\ \Rightarrow\ &R(t) = \frac{1}{2c}\left[b + q\left\{\frac{1 - e^{-qt+c_3}}{1 + e^{-qt+c_3}}\right\}\right] \end{aligned}$$

Since $q > b$ and since $\tanh x$ increases monotonically from -1 to $+1$ when x increases from $-\infty$ to $+\infty$, it follows that the constant c_2 and c_3 exists and have real values. These constants can also be chosen in such a way that $R(0) = 0$. Behaviour of $R(t)$ for large values of time or in other words asymptotic behaviour of $R(t)$ can be found as

$$\lim_{t\to\infty} R(t) = \frac{1}{2c}(b + q) \tag{2.2.4}$$

or in terms of the old parameters of the mode we have,

$$R_\infty = \frac{1}{S_0}\Big[\rho(S_0-\rho)+\rho\{(S_0-\rho)^2+2S_0I_0\}^{1/2}\Big] \tag{2.2.5}$$

Let us now see if some additional assumption regarding the relative size of the parameter gives some result of the threshold theorem mention early.

In particular, it is customary to assume that an epidemic is generated through the introduction of a small number of infected individuals to a population of susceptible. Mathematically, $S_0 > \rho$ and $I_0 > 0$. We now use the quantity

$$\lim_{I_0\to 0} R(\infty) = \frac{2(S_0-\rho)\rho}{S_0} \tag{2.2.6}$$

to represents the asymptotic size of an epidemic resulted from the introduction of a small number of infective into a group of susceptible.

Finally, let us assume that $S-0$ is closed to the threshold value ρ, then the epidemic develop only of $S_0 > \rho$, i.e., $S_0 = \rho+\gamma$, where $\gamma > 0$ is small. Therefore,

$$\lim_{I_0\to\infty} R_\infty = \frac{2\gamma R}{\rho+\gamma} = 2\gamma\left(1+\frac{\gamma}{\rho}\right)^{-1} \approx 2\gamma \tag{2.2.7}$$

Therefore the asymptotic size of the epidemic is approximately equal to 2γ. Hence, we can state as follows:

The total size of the epidemic resulting from an introduction of trace infection into a population of susceptible whose size S_0 is closed to the threshold value ρ is approximately equal to $2(S_0-\rho)$.

Remark 2.2.2.

- It may be remarked that this result is also taken as a part of threshold theorem of epidemiology.
- We can rewrite the expression of $R(t)$ also in the form

$$R(t) = \frac{\rho^2}{S_0}\left[\frac{S_0}{\rho} - 1 + \alpha\tanh\left(\frac{\alpha\gamma t}{2}-\phi\right)\right] \tag{2.2.8}$$

where $\alpha = \left[\left(\frac{S_0}{\rho}-1\right)^2 + \frac{2S_0}{\rho^2}(N-S_0)\right]^{1/2}$ and $\phi = \tanh^{-1}\frac{1}{\alpha}\left(\frac{S_0}{\rho}-1\right)$.

Differentiating, we get

$$\frac{dR}{dt} = \frac{\gamma\rho^2\alpha^2}{2S_0}\,\mathrm{sech}^2\left(\frac{1}{2}\alpha\gamma t-\phi\right) \tag{2.2.9}$$

This equation defines a *symmetrical bell-shaped* curve in $t-\frac{dR}{dt}$ plane (see Fig. 2.1). It may be noted that "*Kermack & McKendric*" compared the values of $\frac{dR}{dt}$ from this equation and found complete agreement with the data from an actual plague which occur during 1905-06 in Bombay. The typical variations of $S(t)$, $I(t)$ and $R(t)$ can be represented graphically in Fig. 2.2.

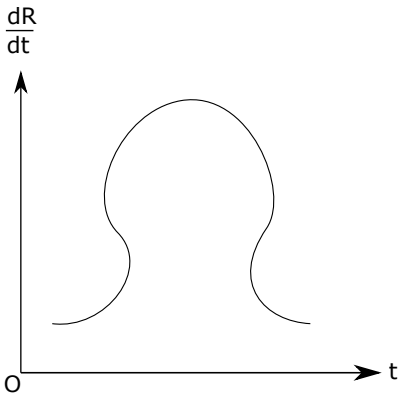


Figure 2.1: Symmetrical bell-shaped curve

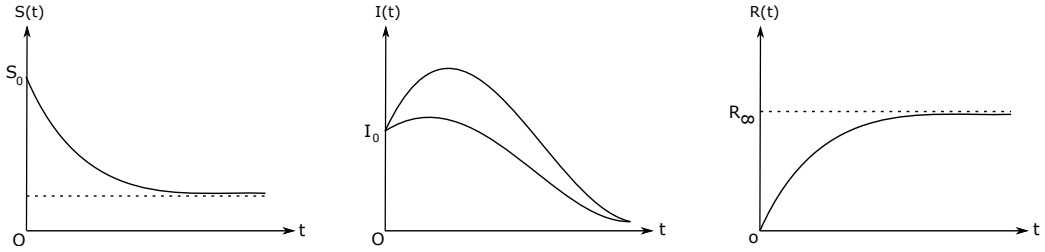


Figure 2.2: Variation of $S(t),\ I(t)$ and $R(t)$ with time t.

2.3 Recurring epidemic

There are many disease that tend to recur in various population with a certain amount of regularity often assuming the character of an epidemic. For example, measles. We assume that the stock of susceptible is replenished at a constant rate μ in time Δt, so that we can take the group of susceptible to be increase by the amount $\mu\Delta t$, which losing $rSI\Delta t$ due to new infections. We can take the total population size to remain constant. By assuming the influx of the new susceptible balanced by an appropriate death rate affecting ony the removed individuals. We can then have the governing equation as

$$\frac{dS}{dt} = -rSI + \mu \tag{2.3.1}$$

$$\frac{dI}{dt} = rSI - \gamma I \tag{2.3.2}$$

The steady state conditions are given by

$$\frac{dS}{dt} = 0 = \frac{dI}{dt}$$

Therefore the steady states are given by

$$S = \frac{\gamma}{r} = S_0 \quad \text{and} \quad I = \frac{\mu}{\gamma} = I_0$$

Let us now study about the equilibrium position through the use of

$$S = S_0(1+u) \quad \text{and} \quad I = I_0(1+v)$$

where u and v are small quantities. Substituting the above quantities in Eqs. (2.3.1) and (2.3.2), we have

$$\frac{1}{rI_0}\frac{du}{dt} = -(u+v+uv) \Rightarrow \sigma\frac{du}{dt} = -(u+v+uv) \text{ where } \sigma = \frac{\gamma}{r\mu} \tag{2.3.3}$$

$$\frac{dv}{dt} = \gamma u(1+v) \Rightarrow \tau\frac{dv}{dt} = u(1+v) \text{ where } \tau = \frac{1}{\gamma} \tag{2.3.4}$$

Since u and v are small, so their products may be neglected so that the Eqs. (2.3.3) and (2.3.4) reduced to

$$\sigma\frac{du}{dt} = -(u+v) \tag{2.3.5}$$

$$\tau\frac{dv}{dt} = u \tag{2.3.6}$$

From these equations, we get

$$\tau\frac{d^2v}{dt^2} = \frac{du}{dt} = -\frac{1}{\sigma}(u+v) = -\frac{1}{\sigma}\left(\tau\frac{dv}{dt} + v\right)$$

$$\Rightarrow \frac{d^2v}{dt^2} + \frac{1}{\sigma}\frac{dv}{dt} + \frac{1}{\tau\sigma}v = 0 \tag{2.3.7}$$

The general solution of the Eq. (2.3.7) is given by

$$v(t) = Ae^{-t/2\sigma}\cos\xi t + Be^{-t/2\sigma}\sin\xi t, \tag{2.3.8}$$

where $\xi = \sqrt{\frac{1}{\sigma\tau} - \frac{1}{4\sigma^2}}$. Using the initial conditions $v = v_0$ and $\frac{dv}{dt} = 0$ at $t = 0$, we have$v_0 = A$ and $B = \frac{v_0}{2\sigma\xi}$. Therefore,

$$\begin{aligned} v(t) &= v_0e^{-t/2\sigma}\cos\xi t + \frac{v_0}{2\sigma\xi}e^{-t/2\sigma}\sin\xi t \\ &= v_0e^{-t/2\sigma}\left[\cos(\xi t) + \frac{1}{2\sigma\xi}\sin\xi t\right] \end{aligned} \tag{2.3.9}$$

Using Eq. (2.3.9), from Eq. (2.3.6) we get

$$\begin{aligned} u(t) &= \tau\frac{dv}{dt} \\ &= \tau v_0\left[-\frac{1}{2\sigma}\left(\cos\xi t + \frac{1}{2\sigma\xi}\sin\xi t\right) + \left(-\xi\sin\xi t + \frac{1}{2\sigma}\cos\xi t\right)\right]e^{-t/2\sigma} \\ &= \tau v_0e^{-t/2\sigma}\left[\frac{1}{4\sigma^2\xi}\sin\xi t - \xi\sin\xi t\right] \\ &= \tau v_0e^{-t/2\sigma}\left(\frac{1}{4\sigma^2\xi} - \xi\right)\sin\xi t \end{aligned}$$

This clearly represents damp harmonic motion to discuss the small departure from equilibrium.

Unit 3

Course Structure

- Stochastic Epidemic Model without Removal
- Basic System of Equations
- Solution of the System of Equation

3.1 Stochastic Epidemic Model Without Removal

3.2 Basic System of Equation

Let us suppose that $p_n(t)$ be the probability that there are n susceptible individuals at time t in the system. Let $f_j(n)\Delta t + o(\Delta t)$ be the probability that the number changes from n to $n+j$ in the time interval $(t,\ t+\Delta t)$. Here, j is any positive or negative integers, and $o(\Delta t)$ denotes an infinitesimal which is such that

$$\frac{o(\Delta t)}{\Delta t} \to 0 \quad \text{as} \quad \Delta t \to 0. \tag{3.2.1}$$

The probability that there is no change in the time interval $(t,\ t+\Delta t)$ is then given by

$$1 - \sum_j f_j(n)\Delta t + o(\Delta t) \tag{3.2.2}$$

Using the theorem of total and compound probabilities, we get

$$p_n(t+\Delta t) = p_n(t)\left[1 - \sum_j f_j(n)\Delta t\right] + \sum_j p_{n-j}(t) f_j(n-j)\Delta t + o(\Delta t) \tag{3.2.3}$$

so that

$$\frac{p_n(t+\Delta t) - p_n(t)}{\Delta t} = -p_n(t)\sum_j f_j(n) + \sum_j p_{n-j}(t) f_j(n-j) + \frac{o(\Delta t)}{\Delta t}. \tag{3.2.4}$$

Proceeding to the limit as $\Delta t \to 0$, we obtain

$$\frac{dp_n}{dt} = -p_n(t)\sum_j f_j(n) + \sum_j p_{n-j}(t) f_j(n-j). \tag{3.2.5}$$

Multiplying (3.2.5) by x^n, summing for all n, and using the definition of the probability generating function, namely,

$$\phi(x,t) = \sum_{n=0}^{\infty} p_n(t) x^n, \tag{3.2.6}$$

we get

$$\frac{\partial \phi}{\partial t} = -\sum_j \sum_n f_j(n) p_n(t) x^n + \sum_j \sum_n p_{n-j}(t) f_j(n-j) x^{n-j} \tag{3.2.7}$$

which gives the partial differential equation

$$\frac{\partial \phi}{\partial t} = \sum_j (x^{-j} - 1) f_j \left(x \frac{\partial}{\partial x} \right) \phi(x,t). \tag{3.2.8}$$

Now we make use of the relations

$$\begin{aligned}
\left(x\frac{\partial}{\partial x} \right) \phi &= \sum_n n p_n(t) x^n \\
\left(x\frac{\partial}{\partial x} \right)^2 \phi &= \sum_n n_n^p(t) x^n \\
&\cdots \quad \cdots \quad \cdots \\
\left(x\frac{\partial}{\partial x} \right)^m \phi &= \sum_n n^m p_n(t) x^n, \quad m = 1,2,3,\ldots
\end{aligned} \tag{3.2.9}$$

to get

$$\psi\left(x\frac{\partial}{\partial x} \right) \phi = \sum_n \psi(n) p_n(t) x^n, \tag{3.2.10}$$

where $\psi(x)$ is any polynomial functions of x. In order to find all the probabilities, we either sove the finite system of differential-difference equations (3.2.5) or solve the partial differential equation (3.2.8) subject to the initial conditions

$$\phi(x,0) = \sum_n p_n(0) x^n = x^{n_0}, \tag{3.2.11}$$

where n_0 is the number of susceptible in the system at $t = 0$.

3.2.1 Solution of the System of Equation

Initially, at $t = 0$, let there be n susceptibles and one infective in the system. Also, let the probability that there are r susceptible person at time t be $p_r(t)$. We assume that the probability of one more person becoming infected in time Δt is

$$\beta(n+1-r)\Delta t + o(\Delta t) \tag{3.2.12}$$

so that

$$\begin{aligned}
f_j(r) &= \beta r(n+1-r) \qquad (j=1) \\
&= 0 \qquad\qquad\qquad\quad (j \neq 1)
\end{aligned} \tag{3.2.13}$$

Substituting (3.2.13) in (3.2.8), we get

$$\begin{aligned}
\frac{\partial \phi}{\partial t} &= \beta(x^{-1}-1)\left[x\frac{\partial}{\partial x}\left(n+1-x\frac{\partial}{\partial x}\right)\phi\right] \\
\Rightarrow \frac{\partial \phi}{\partial t} &= \beta(x^{-1}-1)\left[x\frac{\partial}{\partial x}(n+1)\phi - x\frac{\partial \phi}{\partial x}\right] \\
\Rightarrow \frac{\partial \phi}{\partial t} &= \beta(1-x)\left[(n+1)\frac{\partial \phi}{\partial x} - \frac{\partial \phi}{\partial x} - x\frac{\partial^2 \phi}{\partial x^2}\right] \\
\Rightarrow \frac{\partial \phi}{\partial t} &= \beta(1-x)\left(n\frac{\partial \phi}{\partial x} - x\frac{\partial^2 \phi}{\partial x^2}\right) \qquad (3.2.14)
\end{aligned}$$

Since there are n susceptibles at time $t = 0$,

$$\phi(x,t) = \sum_r p_r(0)x^r = p_n(0)x^n = x^n. \qquad (3.2.15)$$

Substituting $\phi(x,t) = \sum_{r=0}^{n} p_r(t)x^r$ in (3.2.14) and equating the coefficients of the various powers of x, we get

$$\frac{dp_r}{dt} = \beta(r+1)(n-r)p_{r+1} - \beta r(n-r+1)p_r \quad (r = 0,1,2,\ldots,n-1), \qquad (3.2.16)$$

$$\frac{dp_n}{dt} = -\beta n p_n \qquad (3.2.17)$$

with initial conditions

$$p_n(0) = 1, \quad p_r(0) = 0 \quad (r = 0,1,2,\ldots,n-1). \qquad (3.2.18)$$

We can now follow either of the two procedures:

• We can solve the partial differential equation (3.2.14) subject to initial condition (3.2.15), or

• We can solve the system of $n+1$ differential differential equations, namely, (3.2.16) and (3.2.17), subject to initial conditions (3.2.18). We adopt the second procedure here.

Solving (3.2.17) subject to (3.2.18), we get

$$p_n(t) = e^{-\beta n t} \qquad (3.2.19)$$

Equation (3.2.16) then gives

$$\frac{dp_{n-1}}{dt} + 2\beta(n-1)p_{n-1} = n\beta e^{-n\beta t}. \qquad (3.2.20)$$

Integrating (3.2.20) subject to (3.2.18), we obtain

$$p_{n-1}(t) = e^{-2\beta(n-1)t}\int_0^t n\beta e^{-n\beta t}e^{2(n-1)\beta t}dt = \frac{n}{n-2}\left[e^{-n\beta t} - e^{-(2n-2)\beta t}\right]. \qquad (3.2.21)$$

We can proceed in this way systematically step by step to find $p_{n-2}(t),\ p_{n-3}(t),\ldots,p_0(t)$.

Alternatively, we can use the Laplace transform method to solve (3.2.16) and (3.2.17) subject to (3.2.18). Let

$$q_r(s) = \int_0^\infty e^{-st} p_r(t)\, dt. \tag{3.2.22}$$

Multiplying both sides of (3.2.16) and (3.2.17) by e^{-st} and integrating over the range 0 to ∞, we get

$$\int_0^\infty e^{-st}\frac{dp_r}{dt}dt = \beta\int_0^\infty e^{-st}(r+1)(n-r)p_{r+1}dt - \beta r(n-r+1)\int_0^\infty e^{-st}p_r\, dt$$

$$\int_0^\infty e^{-st}\frac{dp_n}{dt}dt = -\beta n\int_0^\infty e^{-st}p_n\, dt.$$

Using the conditions given by (3.2.18), we obtain

$$sq_r(s) = \beta(r+1)(n-1)q_{r+1}(s) - \beta r(n-r+1)q_r(s), \quad r = 0,1,2,\ldots,n-1, \tag{3.2.23}$$

$$sq_n(s) = 1 - \beta n q_n(s) \tag{3.2.24}$$

From (3.2.23),

$$\begin{aligned} q_r(s) &= \frac{\beta(r+1)(n-r)}{[s+r(n-r+1)\beta]}q_{r+1}(s) \\ &= \frac{\beta^2(r+1)(n-r)(r+2)(n-r-1)}{[s+r(n-r+1)\beta][s+(r+1)(n-r)\beta]}q_{r+2} \\ &\vdots \\ &= \frac{\beta^{n-r}[(r+1)(r+2)\cdots(r+n-r)][(n-r)!]}{\prod_{j=1}^{n-r+1}\Big[s+j(n-j+1)\beta\Big]} \\ &= \frac{\beta^{n-r}[\,n!\,][\,(n-r)!\,]}{r!}\prod_{j=1}^{n-r+1}\frac{1}{\Big[s+j(n-j+1)\beta\Big]} \quad (r=0,1,2,\ldots,n-1) \end{aligned} \tag{3.2.25}$$

$$q_n(s) = \frac{1}{s+n\beta}. \tag{3.2.26}$$

By inverting the Laplace transforms, we can find $p_r(t)$. This can be easily done by splitting the product on the right-hand side of Eq. (3.2.25) into partial fractions.

- If $r > n/2$, there are no repeated factors, and this is relatively easier.
- If $r \le n/2$, repeated factors occur, and care has to be exercised.

The mean of the distribution is found by using

$$m(t) = \sum_{r=1}^{n} r p_r(t). \tag{3.2.27}$$

Unit 4

Course Structure

- Stochastic Epidemic Model with Multiple Infections
- Stochastic Epidemic Model with Removal
- Stochastic Epidemic Model with Removal, Immigration and Emigration
- Stochastic Epidemic Model with Carriers
- Stochastic Epidemic Model with Infectives and Carriers

4.1 Other Stochastic Epidemic Models

4.1.1 Epidemics with Multiple Infections

Particular case for the epidemics with multiple infections. When epidemics with multiple infection occur, there can be j infections in the time interval $(t + \Delta t)$ with the probability $\beta_j s(n + 1 - s)\Delta t + o(\Delta t)$ where $j = 1, 2, \ldots, m$ and s be the number of susceptibles at time t and n be the initial number of susceptibles.

Here r_j be the contact rates for j infections. In such a case, the basic partial differential equation will have the form

$$\begin{aligned}
\frac{\partial \phi}{\partial t} &= \sum_{j=1}^{m}(x^{-j} - 1)\beta_j \; x\frac{\partial}{\partial x}\left[\left(n + 1 - x\frac{\partial}{\partial x}\right)\phi\right] \\
\Rightarrow \frac{\partial \phi}{\partial t} &= \sum_{j=1}^{m}\frac{1}{x^{j-1}}(1 - x^j)\beta_j\left[(n+1)\frac{\partial \phi}{\partial x} - \frac{\partial \phi}{\partial x} - x\frac{\partial^2 \phi}{\partial x^2}\right] \\
\Rightarrow \frac{\partial \phi}{\partial t} &= \sum_{j=1}^{m}\frac{1 - x^j}{x^{j-1}}\beta_j\left[n\frac{\partial \phi}{\partial x} - x\frac{\partial^2 \phi}{\partial x^2}\right] \\
\Rightarrow \frac{\partial \phi}{\partial t} &= \left(n\frac{\partial \phi}{\partial x} - x\frac{\partial^2 \phi}{\partial x^2}\right)\left[\beta_1(1 - x) + \beta_2\left(\frac{1}{x} - x\right) + \beta_3\left(\frac{1}{x^2} - x\right) + \cdots + \beta_m\left(\frac{1}{x^{m+1}} - x\right)\right]
\end{aligned}$$

This results is equivalent to (3.2.14) if one takes for $m = 1$ and $\beta_m = \beta$. We can also write the system of differential difference equations from first principle and solve those one-by-one directly or by using Laplace transformation technique.

4.1.2 Stochastic Epidemic Model with Removal

Let $p_{m,n}(t)$ be the probability that there are m susceptibles and n infectives in the population at time t. If N is the total size of the population, then the number of persons in the removed category is $N - m - n$.

Let the probability of susceptible being infected in the time interval $(t, t + \Delta t)$ be $\beta mn\Delta t + o(\Delta t)$, and let the corresponding probability of one infected being removed in the same time interval be $\gamma n\Delta t + o(\Delta t)$. The probability of not having any change in this time interval is

$$1 - \beta mn\Delta t - \gamma n\Delta t + o(\Delta t). \tag{4.1.1}$$

Now there can be m susceptibles and n infected persons at time $t + \Delta t$ if there are

(i) $m + 1$ susceptibles and $n - 1$ infectives at time t and if one person has become infected in time Δt, or

(ii) m susceptibles and $n + 1$ infectives at time t and if one infected person has been removed in time Δt, or

(iii) m susceptibles and n infectives at time t and if there is no change in time Δt.

We assume, as usual, that the probability of more than one change in time Δt is $o(\Delta t)$. Then, using the theorem of total and compound probability, we get

$$\begin{aligned} p_{m,n}(t + \Delta t) &= p_{m+1,n-1}(t)\beta(m + 1)(n - 1)\Delta t + p_{m,n+1}(t)\gamma(n + 1)\Delta t \\ &\quad + p_{m,n}(t)(1 - \beta mn\Delta t - \gamma n\Delta t) + o(\Delta t) \\ \Rightarrow \frac{p_{m,n}(t + \Delta t) - p_{m,n}(t)}{\Delta t} &= \beta(m + 1)(n - 1)p_{m+1,n-1}(t) - \beta mnp_{m,n}(t) \\ &\quad + \gamma(n + 1)p_{m,n+1}(t) - \gamma np_{m,n}(t) + \frac{o(\Delta t)}{\Delta t}. \end{aligned}$$

Proceeding to the limit as $\Delta t \to 0$, we get

$$\begin{aligned} \frac{d}{dt}\Big[p_{m,n}(t)\Big] &= \beta(m + 1)(n - 1)p_{m+1,n-1}(t) - \beta mnp_{m,n}(t) \\ &\quad + \gamma(n + 1)p_{m,n+1}(t) - \gamma np_{m,n}(t). \end{aligned} \tag{4.1.2}$$

Initially, let there be s susceptibles and a infectives. Then we define the probability generating function by

$$\phi(x, y, t) = \sum_{m=0}^{s} \sum_{n=0}^{s+a-m} p_{m,n}(t)x^m y^n. \tag{4.1.3}$$

Multiplying (4.1.2) by $x^m y^n$ and summing over n from 0 to $s+a-m$ and m from 0 to s, we get

$$\begin{aligned}\frac{\partial}{\partial t}\sum_{m=0}^{s}\sum_{n=0}^{s+a-m} p_{m,n}(t)x^n y^n &= \beta y^2 \sum_{m=0}^{s}\sum_{n=0}^{s+a-m} p_{m+1,n-1}(t)(m+1)(n-1)x^m y^{n-2} \\ &\quad -\beta xy \sum_{m=0}^{s}\sum_{n=0}^{s+a-m} p_{m,n}(t)mnx^{m-1}y^{n-1} \\ &\quad +\gamma \sum_{m=0}^{s}\sum_{n=0}^{s+a-m} p_{m,n-1}(t)(n+1)x^m y^n \\ &\quad -\gamma y \sum_{m=0}^{s}\sum_{n=0}^{s+a-m} p_{m,n}(t)nx^m y^{n-1}. \end{aligned} \tag{4.1.4}$$

From (4.1.3) and (4.1.4), we get

$$\frac{\partial \phi}{\partial t} = \beta(y^2 - xy)\frac{\partial^2 \phi}{\partial x \partial y} + \gamma(1-y)\frac{\partial \phi}{\partial y}. \tag{4.1.5}$$

Now the equation (4.1.5) can be solved subject to the initial condition

$$\phi(x,y,0) = x^s y^a \quad \text{since} \quad p_{m,n}(0) = \begin{cases} 1\,; & m=s, n=a \\ 0\,; & \text{otherwise.} \end{cases} \tag{4.1.6}$$

An Alternative Derivation of the Partial Differential Equation

The stochastic model with removal can be represented as follows:

Event	Transition	Transition rate	Probability
a susceptible becomes infected	$(m,n) \to (m-1, n+1)$	βmn	$\beta mn\Delta t + o(\Delta t)$
an infective becomes removed	$(m,n) \to (m, n-1)$	γn	$\gamma n\Delta t + o(\Delta t)$
there is no change	$(m,n) \to (m,n)$	$-(\beta mn + \gamma n)$	$1 - (\beta mn + \gamma n)\Delta t + o(\Delta t)$

From the tabular representation, we get

$$f_{-1,1}(m,n) = \beta mn, \qquad f_{0,-1}(m,n) = \gamma n \tag{4.1.7}$$

so that the partial differential equation representing the model becomes

$$\begin{aligned}\frac{\partial \phi}{\partial t} &= (x^{-1}y^1 - 1)\beta xy\frac{\partial^2 \phi}{\partial x \partial y} + (x^0 y^{-1} - 1)\gamma y \frac{\partial \phi}{\partial y} \\ \Rightarrow \frac{\partial \phi}{\partial t} &= \beta(y^2 - xy)\frac{\partial^2 \phi}{\partial x \partial y} + \gamma(1-y)\frac{\partial \phi}{\partial y} \end{aligned} \tag{4.1.8}$$

It is worthwhile to note here that Eq. (4.1.5) and Eq. (4.1.8) are identical.

Stochastic Epidemic Model with Removal, Immigration, and Emigration

The model can be represented as follows:

Event	Transition	Transition rate	Probability
a susceptible is removed	$(m,n) \to (m-1, n+1)$	βmn	$\beta mn\Delta t + o(\Delta t)$
an infective is removed	$(m,n) \to (m, n-1)$	γn	$\gamma n\Delta t + o(\Delta t)$
a new susceptible joins	$(m,n) \to (m+1, n)$	μ	$\mu\Delta t + o(\Delta t)$
an infective joins	$(m,n) \to (m, n+1)$	ν	$\nu\Delta t + o(\Delta t)$
a susceptible leaves	$(m,n) \to (m-1, n)$	δm	$\delta m\Delta t + o(\Delta t)$

This model gives

$$f_{-1,-1}(m,n) = \beta mn, \quad f_{0,-1}(m,n) = \gamma n, \quad f_{1,0}(m,n) = \mu, \quad f_{0,1}(m,n) = \gamma, \quad f_{-1,0}(m,n) = \delta n$$

so that the partial differential equation representing the model becomes

$$\begin{aligned} \frac{\partial \phi}{\partial t} &= (x^{-1}y^1 - 1)\beta xy\frac{\partial^2 \phi}{\partial x \partial y} + (x^0 y^{-1} - 1)\gamma y\frac{\partial \phi}{\partial y} + (xy^0 - 1)\mu\phi \\ &\quad + (x^0 y^1 - 1)\nu\phi + (x^{-1}y^0 - 1)\delta x\frac{\partial \phi}{\partial x} \\ \Rightarrow \frac{\partial \phi}{\partial t} &= \beta(y^2 - xy)\frac{\partial^2 \phi}{\partial x \partial y} + \gamma(1-y)\frac{\partial \phi}{\partial y} + \mu(x-1)\phi + \nu(y-1)\phi + \delta(1-x)\frac{\partial \phi}{\partial x}. \end{aligned} \quad (4.1.9)$$

In the absence of immigration and emigration, (4.1.9) gives

$$\frac{\partial \phi}{\partial t} = \beta(y^2 - xy)\frac{\partial^2 \phi}{\partial x \partial y} + \gamma(1-y)\frac{\partial \phi}{\partial y}. \quad (4.1.10)$$

It is worthwhile to note here that Eq. (4.1.8) and Eq. (4.1.10) are identical.

Stochastic Epidemic Model with Carriers

Here we consider a disease spread only by carriers so that our interest is in the two classes of individuals, namely, susceptibles and carriers. Carriers are eliminated by external action. Thus we get the following model:

Event	Transition	Transition rate	Probability
a susceptible becomes infective	$(m,n) \to (m-1, n)$	βmn	$\beta mn\Delta t + o(\Delta t)$
a carrier is removed	$(m,n) \to (m, n-1)$	γn	$\gamma n\Delta t + o(\Delta t)$

From the tabular representation, we get

$$f_{-1,0}(m,n) = \beta mn, \qquad f_{0,-1}(m,n) = \gamma n \quad (4.1.11)$$

so that the partial differential equation representing the model becomes

$$\begin{aligned} \frac{\partial \phi}{\partial t} &= (x^{-1}y^0 - 1)\beta xy\frac{\partial^2 \phi}{\partial x \partial y} + (x^0 y^{-1} - 1)\gamma y\frac{\partial \phi}{\partial y} \\ \Rightarrow \frac{\partial \phi}{\partial t} &= \beta y(1-x)\frac{\partial^2 \phi}{\partial x \partial y} + \gamma(1-y)\frac{\partial \phi}{\partial y} \end{aligned} \quad (4.1.12)$$

If we allow immigration and emigration of susceptibles and carriers, we get

$$\frac{\partial \phi}{\partial t} = \beta y(1-x)\frac{\partial^2 \phi}{\partial x \partial y} + \gamma(1-y)\frac{\partial \phi}{\partial y} + \mu(x-1) + \nu(y-1)\phi + \delta(1-x)\frac{\partial \phi}{\partial x}. \quad (4.1.13)$$

Stochastic Epidemic Model with Infectives and Carriers

Let $m,\ n,\ p$ denote the number of susceptibles, infectives and carriers respectively. A susceptible can become infective by contact with either an infected or a carrier. The result is represented in the following model:

Event	Transition	Transition rate	Probability
a susceptible becomes an infective	$(m,n,p) \to (m-1,n+1,p)$	βmn	$\beta mn\Delta t + o(\Delta t)$
a susceptible becomes an infective due to contact with a carrier	$(m,n,p) \to (m-1,n+1,p)$	γmp	$\gamma mp\Delta t + o(\Delta t)$
a carrier is removed	$(m,n,p) \to (m,n,p-1)$	δp	$\delta p\Delta t + o(\Delta t)$

From the tabular representation, we get

$$f_{-1,1,0}(m,n,p) = \beta mn + \gamma mp, \qquad f_{0,0,-1}(m,n,p) = \gamma p \tag{4.1.14}$$

so that the partial differential equation representing the model becomes

$$\begin{aligned} \frac{\partial \phi}{\partial t} &= (x^{-1}y^1z^0 - 1)\left(\beta xy\frac{\partial^2\phi}{\partial x\partial y} + xz\frac{\partial^2\phi}{\partial x\partial z}\right) + (x^0y^0z^{-1} - 1)\delta z\frac{\partial \phi}{\partial z} \\ \Rightarrow \frac{\partial \phi}{\partial t} &= (y-x)\left(\beta y\frac{\partial^2\phi}{\partial x\partial y} + \gamma z\frac{\partial^2\phi}{\partial x\partial z}\right) + \delta(1-z)\frac{\partial \phi}{\partial z}. \end{aligned} \tag{4.1.15}$$

Unit 5

Course Structure

- Basic model for inheritance of genetic characteristic
- Hardly Wienberg law

5.1 Introduction

Population genetics deals with genetic differences within and between populations, and is a part of evolutionary biology. Studies in this branch of biology examine such phenomena as adaptation, speciation, and population structure. Population genetics was a vital ingredient in the emergence of the modern evolutionary synthesis. Traditionally a highly mathematical discipline, modern population genetics encompasses theoretical, lab, and field work. Population genetic models are used both for statistical inference from DNA sequence data and for proof/disproof of concept.

5.2 Basic Model for Inheritance

Genetic Matrices

Each characteristic (e.g., height, colour of the eye and type of blood) of an individual is determined by two *genes*, either of these being received from each of the parents. Each gene may be in two forms, the *dominant* G or the *recessive* g. Thus an individual may belong to one of the following three *genotypes*:

(i) (G, G) which is called *dominant* and denoted by D.

(ii) (G, g) or (g, G) which is termed *hybrid* and denoted by H.

(iii) (g, g) which is called *recessive* and denoted by R.

When two individuals mate, the offspring gets from each parent either of the two forms of genes with the same probability $1/2$. Thus, if (G, G) is crossed with (G, g), there are four possibilities:

(i) The offspring gets the first G from the first parent and G from the second parent. The probability of this is $1/2 \times 1/2 = 1/4$, and the offspring is D.

(ii) The offspring gets the first G from the first parent and g from the second parent. The probability of this is $1/2 \times 1/2 = 1/4$, and the offspring is H.

(iii) The offspring gets the second G from the first parent and G from the second parent. The probability of this is $1/2 \times 1/2 = 1/4$, and the offspring is D.

(iv) The offspring gets the second G from the first parent and g from the second parent. The probability of this is $1/2 \times 1/2 = 1/4$, and the offspring is H.

Thus the probabilities of the offspring being D, H, R are $1/2,\ 1/2,\ 0$ respectively. Arguing in the same way, we get the results given in below which gives the probabilities of the offspring being $D,\ H,\ R$ when these are crossed with $D,\ H,\ R$, in that order.

Results of Crossing by Two Genotypes

	D			H			R		
	D	H	R	D	H	R	D	H	R
D	1	0	0	1/2	1/2	0	0	1	0
H	1/2	1/2	0	1/4	1/2	1/4	0	1/2	1/2
R	0	1	0	0	1/2	1/2	0	0	1

The three fundamental *genetic matrices* which refer to mating with $D,\ H,\ R$, respectively, are obtained as follows:

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/4 & 1/2 & 1/4 \\ 0 & 1/2 & 1/2 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1/2 & 1/2 \\ 0 & 0 & 1 \end{bmatrix}. \tag{5.2.1}$$

Each of these matrices is a *stochastic matrix* since all of its elements are non-negative and the row sums are unity (this is so because in each case the probability that the offspring is a D or H or R is unity).

Let us consider a population in which the probabilities of a person being D, H, R are p, q, r, respectively, so that $p + q + r = 1$. We shall call $P = (p, q, r)$ the *probability vector* of the population.

If each individual in this population is mated with a dominant, then the first matrix gives:

the probability of the *dominant* (D) offspring as

$$1 \cdot p + \frac{1}{2} \cdot q + 0 \cdot r = p + \frac{1}{2}q; \tag{5.2.2}$$

the probability of the *hybrid* (H) offspring as

$$0 \cdot p + \frac{1}{2} \cdot q + 1 \cdot r = \frac{1}{2}q + r; \tag{5.2.3}$$

the probability of the *recessive* (R) offspring as

$$0 \cdot p + 0 \cdot q + 0 \cdot r = 0. \tag{5.2.4}$$

Thus the probability vector for the first generation, on population being mated with pure dominants, is obtained by taking the product of the row matrix P with the first matrix A, i.e., it is given by PA. Similarly, the probability vector for the first generation when population with the probability vector P is mated with pure hybrids (pure recessives) is given by PB (PC).

If the original population is mated with dominants, hybrids, dominants, recessives, hybrids, in that order, the probability vector for the fifth generation is given by $PABACB$.

When mated with dominants, the first generation has the same probability vector as the original probability vector if $PA = P$ i.e., if $\left(p + \frac{1}{2}q, \frac{1}{2}q + r, 0\right) = (p, q, r)$. In other words, if $p = 1, q = 0, r = 0$, then

$$PA = P \Rightarrow P = (1, 0, 0). \tag{5.2.5}$$

Similarly,

$$PB = P \Rightarrow P = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right), \tag{5.2.6}$$

$$PC = P \Rightarrow P = (0, 0, 1)\,. \tag{5.2.7}$$

Now, if the population with the probability vector P is crossed with pure dominant n times, the probability vector of the n-th generation is given by PA^n. To find this, A^n has to be determined, and this is easily done by first diagonalising the matrix A. Thus, since the eigenvalues of the matrix A are easily found to be $\left(1, \frac{1}{2}, 0\right)$ and the corresponding eigenvectors are $(1, 1, 0),\ (0, 1, 2)$ and $(0, 0, 1)$, we can write

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & -2 & 1 \end{bmatrix} = S \wedge S^{-1} \tag{5.2.8}$$

so that

$$\begin{aligned} A^n &= (S \wedge S^{-1})(S \wedge S^{-1}) \cdots (S \wedge S^{-1}) = S \wedge^n S^{-1} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/2^n & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & -2 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 1 - \frac{1}{2^n} & \frac{1}{2^n} & 0 \\ 1 - \frac{1}{2^{n-1}} & \frac{1}{2^{n-1}} & 0 \end{bmatrix}, \end{aligned} \tag{5.2.9}$$

Therefore,

$$PA^n = \left(1 - \frac{q}{2^n} - \frac{r}{2^{n-1}}, \frac{q}{2^n} + \frac{r}{2^{n-1}}, 0\right) \tag{5.2.10}$$

As n tends to infinity, PA^n approaches to the vector $(1, 0, 0)$.

Thus, if any population is mated at random with only dominants successively, we find that (i) the recessives never appear, (ii) the proportion of hybrids tends to zero, and (iii) the proportion of dominants tends to unity.

Even if the original population consists of only recessives, we shall have a proportion $15/16$ of dominants in the fifth generation and a proportion of $511/512$ of dominants in the tenth generation. Thus, if dominants

are desired, we can transform even a breed of recessives into a breed of dominants in a number of generations by repeated mating with dominants.

Exercise 5.2.1. 1. Prove that $PAB \neq PBA,\ PBC \neq PCB,$ and $PCA \neq PAC$. What conclusions can you draw?

2. Suppose an individual of unknown genotype is crossed with a recessive and the offspring is again crossed with a recessive, and so on. Show that, after a long period of such breeding, it is almost certain that the offspring will be a recessive genotype.

3. Find A^n, B^n, C^n, PA^n, PB^n, PC^n, and consider the limiting cases when n approaches infinity. Do the limiting vector depend on P? Interpret the results obtained.

4. Find $P(AB)^n$, and interpret the limit of this as n tends to infinity.

5.3 Hardy-Weinberg Law

Consider random mating or *panmixia* in a population with probability vector P. The probability vectors for mating with $D,\ H,\ R$ are given by $PA,\ PB,\ PC$, but the relative proportions of $D,\ H,\ R$ in the population are $p,\ q,\ r$ so that the probability vector for the first generation, say F_1, is given by

$$\begin{aligned} p\,PA + q\,PB + r\,PC &= \left[\left(p+\frac{1}{2}q\right)^2,\ 2\left(p+\frac{1}{2}q\right)\left(r+\frac{1}{2}q\right),\left(r+\frac{1}{2}q\right)^2\right] \\ &= (p',\ q',\ r') = P' \text{ (say).} \end{aligned} \tag{5.3.1}$$

The three components of the probability vector for the second generation, say F_2, are then given by

$$\begin{aligned} \left(p'+\frac{1}{2}q'\right)^2 &= \left[\left(p+\frac{1}{2}q\right)^2 + \frac{1}{2}\cdot 2\left(p+\frac{1}{2}q\right)\left(r+\frac{1}{2}q\right)\right]^2 \\ &= \left(p+\frac{1}{2}q\right)^2\left(p+\frac{1}{2}q+r+\frac{1}{2}q\right)^2 \\ &= \left(p+\frac{1}{2}q\right)^2 \\ &= p', \end{aligned} \tag{5.3.2}$$

$$
\begin{aligned}
2\left(p'+\frac{1}{2}q'\right)\left(r'+\frac{1}{2}q'\right) &= 2\left[\left(p+\frac{1}{2}q\right)^2+\left(p+\frac{1}{2}q\right)\left(r+\frac{1}{2}q\right)\right] \\
&\quad \times\left[\left(r+\frac{1}{2}q\right)^2+\left(p+\frac{1}{2}q\right)\left(r+\frac{1}{2}q\right)\right] \\
&= 2\left[\left(p+\frac{1}{2}q\right)\left(p+\frac{1}{2}q+r+\frac{1}{2}q\right)\left(r+\frac{1}{2}q\right)\left(r+\frac{1}{2}q+p+\frac{1}{2}q\right)\right] \\
&= 2\left(p+\frac{1}{2}q\right)\left(r+\frac{1}{2}q\right) \\
&= q', \qquad (5.3.3) \\
\left(r'+\frac{1}{2}q'\right)^2 &= \left[\left(r+\frac{1}{2}q\right)^2+\left(p+\frac{1}{2}q\right)\left(r+\frac{1}{2}q\right)\right]^2 \\
&= \left(r+\frac{1}{2}q\right)^2\left(r+\frac{1}{2}q+p+\frac{1}{2}q\right)^2 \\
&= \left(r+\frac{1}{2}q\right)^2 \\
&= r', \qquad (5.3.4)
\end{aligned}
$$

Thus the probability vector for F_2 is the same as that for F_1. This shows that, due to random mating, the probability vectors for the first generation and all succeeding generations are same. This is known as the *Hardy-Weinberg law*, called after the mathematician G. H. Hardy and the geneticist W. Weinberg.

In any population in which random mating takes place, we have $P = P'$ so that

$$
p=\left(p+\frac{1}{2}q\right)^2, \quad q=2\left(p+\frac{1}{2}q\right)\left(r+\frac{1}{2}q\right), \quad r=\left(r+\frac{1}{2}q\right)^2. \qquad (5.3.5)
$$

Simplifying (5.3.5), we get

$$
p=(1-\sqrt{r})^2, \quad q=2\sqrt{r}(1-\sqrt{r}), \quad r=r. \qquad (5.3.6)
$$

The ratios $p : q : r$ in a genetically stable population are known as *Hardy-Weinberg ratios.* There is only one parameter r that depends on the particular gene under consideration.

Note: We may note that the Hardy-Weinberg law holds for a gene if the mating is random with respect to that gene. Thus, in human populations, the law is likely to hold for genes for blood groups, since, in general, people do not worry about blood groups when marrying, but the law may not hold for the gene determining heights since tall people, in general, tend to marry tall people.

Note: If we can identify the three genotypes for a particular gene in a population and if their relative proportions verify (5.3.6), then it confirms that mating is likely to be random for that gene. If (5.3.6) can not be verified, it may be due to non-random mating or differential mortality of dominant and recessive genes.

Note: In general, however, it is not easy to distinguish between the three genotypes. If G is dominant to g, then individuals having (G, G) (G, g) have the same appearance and belong to the same *phenotype*. Thus, while there are three genotypes, only two distinct phenotypes exist, namely, $\{(G, G),\ (G, g)\ \}$ and $\{(g, g\}$, with respect to a gene.

The individuals with $(G,G),\ (g,g)$ are said to be *homozygous* and the individual with (G,g) are said to be *heterozygous*.

The Hardy-Weinberg law can also be stated in terms of gene frequencies. If $P,\ Q$ are the relative gene frequencies in a population and (p,q,r) are the relative frequencies of $(G,G),\ (G,g)$ and (g,g), then it is easily seen that

$$P = p + \frac{1}{2}q, \qquad Q = \frac{1}{2}q + r. \tag{5.3.7}$$

Knowing $p,\ q,\ r$, we can find P and Q uniquely, but knowing P and Q, we cannot find p,q,r uniquely. However, for random mating, (5.3.6) and (5.3.7) give

$$P = 1 - \sqrt{r}, \qquad Q = \sqrt{r}. \tag{5.3.8}$$

Now, in any generation, if the relative frequencies of genes G and g are P and Q provided $P+Q=1$, in both males and females, then, in random mating, the probability of an offspring getting G from both parents is P^2, the probability of its getting G from one parent and g from the other parent is $2PQ$, and the probability of its getting g from both parent is Q^2, so that the proportions in F_1 are

$$(G,G)\ :\ P^2, \quad (G,g)\ :\ 2PQ, \quad (g,g)\ :\ Q^2. \tag{5.3.9}$$

From (5.3.2) - (5.3.4) and (5.3.7), wet the Hardy-Weinberg ratio. The relative gene frequencies in F_1 are

$$G\ :\ 2P^2 + 2PQ = 2P(P+Q) = 2P, \qquad g\ :\ 2PQ + 2Q^2 = 2Q(P+Q) = 2Q \tag{5.3.10}$$

so that the proportions of genes are the same as in the original population and F_2 has the ratios given by (5.3.9). This again confirms the Hardy-Weinberg law.

Unit 6

Course Structure

- Correlation between Genetic Composition of Siblings
- Bayes Theorem and Its Applications in Genetics

6.1 Correlation between Genetic Composition of Siblings

As an example of the use of our basic model, we find the correlation between genetic composition of offsprings from the same parents, assuming random mating and a genetically stable population. In the original population, let the proportion of dominants, hybrids and recessives be given by (5.3.6), i.e.,

$$p = (1-\sqrt{r})^2, \quad q = 2\sqrt{r}(1-\sqrt{r}), \quad r = r. \tag{6.1.1}$$

Let $Y_1,\ Y_2$ be the offspring from parents $X_1,\ X_2$. Then using the theorems of total and compound probabilities, we get

$$\begin{aligned}
P(Y_1 = D, Y_2 = D) &= P(Y_1 = D, Y_2 = D; X_1 = D, X_2 = D) \\
&\quad + P(Y_1 = D, Y_2 = D; X_1 = D, X_2 = H) \\
&\quad + P(Y_1 = D, Y_2 = D; X_1 = H, X_2 = D) \\
&\quad + P(Y_1 = D, Y_2 = D; X_1 = H, X_2 = H) \\
&= P(X_1 = D, X_2 = D)P(Y_1 = D, Y_2 = D/X_1 = D, X_2 = D) \\
&\quad + P(X_1 = D, X_2 = H)P(Y_1 = D, Y_2 = D/X_1 = D, X_2 = H) \\
&\quad + P(X_1 = H, X_2 = D)P(Y_1 = D, Y_2 = D/X_1 = H, X_2 = D) \\
&\quad + P(X_1 = H, X_2 = H)P(Y_1 = D, Y_2 = D/X_1 = H, X_2 = H) \\
&= p^2 \cdot 1 + pq \cdot \frac{1}{2} \cdot \frac{1}{2} + pq \cdot \frac{1}{2} \cdot \frac{1}{2} + q^2 \cdot \frac{1}{4} \cdot \frac{1}{4} \\
&= p^2 + \frac{1}{2}pq + \frac{1}{16}q^2 \\
&= (1-\sqrt{r})^4 + \sqrt{r}(1-\sqrt{r})^3 + \frac{1}{4}r(1-\sqrt{r})^2 \\
&= \frac{1}{4}(1-\sqrt{r})^2(2-\sqrt{r})^2.
\end{aligned} \tag{6.1.2}$$

Similarly,

$$P(Y_1 = D, Y_2 = H) = P(Y_1 = H, Y_2 = D) = \frac{1}{2}\sqrt{r}(1-\sqrt{r})^2(2-\sqrt{r}), \quad (6.1.3)$$

$$P(Y_1 = D, Y_2 = R) = P(Y_1 = R, Y_2 = D) = \frac{1}{4}r(1-\sqrt{r})^2, \quad (6.1.4)$$

$$P(Y_1 = H, Y_2 = R) = P(Y_1 = R, Y_2 = H) = \frac{1}{2}r(1-\sqrt{r})(1+\sqrt{r}), \quad (6.1.5)$$

$$P(Y_1 = H, Y_2 = H) = \sqrt{r}(1-\sqrt{r})(1+\sqrt{r}-r), \quad (6.1.6)$$

$$P(Y_1 = R, Y_2 = R) = \frac{1}{4}r(1+\sqrt{r})^2, \quad (6.1.7)$$

If we assign arbitrary values 1, 0, - 1 to D, H and R, receptively, we get the bivariate probability distribution as follows:

Y_1	Y_2	Probability
1	1	$\frac{1}{4}(1-\sqrt{r})^2(2-\sqrt{r})^2$
- 1	- 1	$\frac{1}{4}r(1+\sqrt{r})^2$
0	0	$\sqrt{r}(1-\sqrt{r})(1+\sqrt{r}-r)$
1	0	$\frac{1}{2}\sqrt{r}(1-\sqrt{r})^2(2-\sqrt{r})$
0	1	$\frac{1}{2}\sqrt{r}(1-\sqrt{r})^2(2-\sqrt{r})$
1	- 1	$\frac{1}{4}r(1-\sqrt{r})^2$
- 1	1	$\frac{1}{4}r(1-\sqrt{r})^2$
0	- 1	$\frac{1}{2}r(1-\sqrt{r})(1+\sqrt{r})$
- 1	0	$\frac{1}{2}r(1-\sqrt{r})(1+\sqrt{r})$

The marginal distributions of Y_1 and Y_2 are the same and are given by

Y_1 or Y_2	1	0	- 1
Probability	$(1-\sqrt{r})^2$	$2\sqrt{r}(1-\sqrt{r})$	r

From these we easily deduce

$$\overline{Y}_1 = \overline{Y}_2 = 1 - 2\sqrt{r}, \quad (6.1.8)$$

$$\sigma^2_{Y_1} = \sigma^2_{Y_2} = 2\sqrt{r}(1-\sqrt{r}), \quad (6.1.9)$$

$$\text{cov}(Y_1, Y_2) = \sqrt{r}(1-\sqrt{r}), \quad (6.1.10)$$

$$\rho_{Y_1,Y_2} = \frac{1}{2}. \quad (6.1.11)$$

Thus, the correlation coefficient comes out to be independent of the value of r, i.e., it is the same for all genes.

6.2 Bayes Theorem and Its Applications in Genetics

At any time, we have some hypotheses to explain genetic events and probabilities associated with these hypotheses to measure our degrees of confidence in the hypotheses. These probabilities are called *a priori*

probabilities. The occurrence of an event changes our degrees of confidence in the sense that the probabilities of some hypotheses may increase and of other may decrease. The new probabilities are called *a posteriori probabilities.* Bayes theorem connects a posteriori and a priori probabilities.

Let $H_1,\ H_2, \ldots, H_n$ be n mutually exclusive hypothesis, and let their a priori probabilities be $P(H_1),\ P(H_2),\ \ldots,\ P(H_n)$. Now let an event A happen, and let the probabilities of happening of this event on the basis of various hypotheses be given by $P(A/H_1),\ P(A/H_2),\ \ldots,\ P(A/H_n)$. Our object is to find the posteriori probabilities $P(H_1/A),\ P(H_2/A), \ldots, P(H_n/A)$ in terms of the known probabilities $P(H_i),\ P(A/H_i),\ i = 1, 2, \ldots, n$.

From the theorem of compound probability,

$$P(AH_i) = P(H_i)P(A/H_i) = P(A)P(H_i/A)$$

so that

$$P(H_i/A) = \frac{P(H_i)P(A/H_i)}{P(A)}, \quad i = 1, 2, \ldots, n. \tag{6.2.1}$$

Since $H_1, H_2, \ldots, H_n$ are mutually exclusive and exhaustive hypotheses under consideration, we have, by the theorem of total probability,

$$P(A) = P(AH_1) + P(AH_2) + \ldots + P(AH_n) = \sum_{j=1}^{n} P(AH_j) = \sum_{j=1}^{n} P(H_j)P(A/H_j). \tag{6.2.2}$$

From (6.2.1) and (6.2.2), we get

$$P(H_i/A) = \frac{P(H_i)P(A/H_i)}{\sum\limits_{j=1}^{n} P(H_j)P(A/H_j)} \tag{6.2.3}$$

which is the required *Bayes theorem*.

Illustration - I

As an illustration of Bayes theorem in genetics, we investigate the probability that two blue-eyed boy twins are monovular (i.e., from the same egg). Here we have two possible hypothesis:

(i) H_1 : Both are from the same egg, i.e., both are monovular;

(ii) H_2 : Both are from the different eggs, i.e., both are binovular.

To find $P(H_1)$ and $P(H_2)$, we remember that observation that 32 percent of all twin pairs are of unlike sex; of the remaining 68 percent, half are expected to be monovular and the other half are expected to be binovular so that

$$P(H_1) = \frac{0.36}{0.68} = \frac{9}{17}, \qquad P(H_2) = \frac{0.32}{0.68} = \frac{8}{17}. \tag{6.2.4}$$

To find $P(A/H_1)$ and $P(A/H_2)$, we assume that mating is random and is genetically stable so that the proportions of $D,\ H,\ R$ are given by

$$p = (1 - \sqrt{r})^2, \quad q = 2\sqrt{r}(1 - \sqrt{r}), \quad r = r \tag{6.2.5}$$

Also, blue eyes are known to arise due to a recessive gene so that:

$$\begin{aligned}
P(A/H_1) &= \text{probability that both boys are recessive when they are from the same egg} \\
&= r, \qquad (6.2.6)\\
P(A/H_2) &= \text{probability that both boys are recessive when they are from the different egg} \\
&= P(\text{parents are } Bb,\ Bb, \text{ and both children are } bb) \\
&\quad + P(\text{parents are } Bb,\ bb, \text{ and both children are } bb) \\
&\quad + P(\text{parents are } bb,\ Bb, \text{ and both children are } bb) \\
&\quad + P(\text{parents are } bb,\ bb, \text{ and both children are } bb) \\
&= q^2 \cdot \frac{1}{4} \cdot \frac{1}{4} + qr \cdot \frac{1}{2} \cdot \frac{1}{2} + qr \cdot \frac{1}{2} \cdot \frac{1}{2} + r^2 \cdot 1 \cdot 1 \\
&= \frac{r(1-\sqrt{r})^2}{4} + \frac{r\sqrt{r}(1-\sqrt{r})}{1} + r^2 \\
&= \frac{1}{4} r(1 + r - 2\sqrt{r} + 4\sqrt{r} - 4r + 4r) \\
&= \frac{1}{4} r(1+\sqrt{r})^2. \qquad (6.2.7)
\end{aligned}$$

Using (6.2.3), (6.2.4), (6.2.6) and (6.2.7), we can find $P(H_1/A)$ and $P(H_2/A)$.

Illustration - II

To illustrate another application of Bayes theorem, we consider the following problem. For mating of two dominant-looking individuals, a dominant-looking individual is obtained. What is the probability that both the parents are real dominants?

There are four possible hypotheses about parents:

$$\begin{aligned}
H_1 &: \text{ parents are } GG, GG, \\
H_2 &: \text{ parents are } GG, Gg, \\
H_3 &: \text{ parents are } Gg, GG, \\
H_4 &: \text{ parents are } Gg, Gg.
\end{aligned}$$

The event A is that the offspring is GG or Gg, so that

$$P(A/H_1) = 1, \quad P(A/H_2) = 1, \quad P(A/H_3) = 1, \quad P(A/H_4) = \frac{3}{4}.$$

Also, in the absence of any knowledge, we make use of *Bayes hypotheses* which postulates that all the four hypotheses have the same a priori probability so that

$$P(H_1) = \frac{1}{4}, \quad P(H_2) = \frac{1}{4}, \quad P(H_3) = \frac{1}{4}, \quad P(H_4) = \frac{1}{4}.$$

Therefore, on making use of (6.2.3), we get

$$P(H_1/A) = \frac{\frac{1}{4} \cdot 1}{\frac{1}{4} \cdot 1 + \frac{1}{4} \cdot 1 + \frac{1}{4} \cdot 1 + \frac{1}{4} \cdot \frac{3}{4}} = \frac{4}{15}.$$

Exercise 6.2.1. 1. A flock of certain species of fowls consists of 117, 191 and 16 with blue, black, and white plumages. Assuming that black and white plumages are the phenotypes corresponding to the homozygous genotypes (b, b) and (w, w) and the blue plumage corresponds to the heterozygous genotype (w, b), find the genotype and gene frequencies.

2. In a certain human population, dominants, hybrids and recessives are 16 per cent, 48 per cent and 36 per cent, respectively. Given that a man is recessive and has a brother, show that the probability of the brother being recessive is 0.66. What are the probabilities of the brother being a dominant or a hybrid?

3. Assuming Mendel's law of independent assortment which postulates that, when there are tow or more gene pair segregating at the same time, they do independently, prove that the double inter-cross $AaBb \times AaBb$ results in four phenotypes, namely, $AB,\ Ab,\ aB$, and ab, in the ratios $9 : 3 : 3 : 1$.

4. From the mating of two hybrids Gg and Gg, a dominant-looking offspring Gx is obtained. This individual is mated with another hybrid, and as a result, n individuals are obtained, all of whom look dominant. What is the a posteriori probability that $x = G$?

5. From the mating of two dominant-looking individuals, n offspring are produced, of which r are recessives. What is the probability that both the parents are hybrid?

Unit 7

Course Structure

- Extension of basic model for inheritance of genetic characteristics
- Models for genetic improvement: Selection and Mutation

7.1 Further Discussion of Basic Model for Inheritance of Genetic Characteristics

7.1.1 Phenotype Ratios

For one gene, there are 4 possible genetic constitutions, namely, $\{(G,G),\ (G,g),\ (g,G),\ (g,g)\}$; there are 3 genotypes, viz, dominant $D\ :\ (G,G)$, hybrid $H\ :\ (G,g),\ (g,G)$, and recessive $R\ :\ (g,g)$; and there 2 phenotypes, viz, dominant-looking: $\{(G,G),\ (G,g),\ (g,G)\}$ and recessive-looking $\{(g,g)\}$. The ratio of the two phenotypes is $3:1$.

For two genes, there are 16 possibilities:

$G_1\ G_1\ G_2\ G_2$	$G_1\ G_1\ G_2\ g_2$	$G_1\ G_1\ g_2\ G_2$	$G_1\ G_1\ g_2\ g_2$
$G_1\ g_1\ G_2\ G_2$	$G_1\ g_1\ G_2\ g_2$	$g_1\ G_1\ g_2\ G_2$	$G_1\ g_1\ g_2\ g_2$
$G_1\ g_1\ G_2\ G_2$	$G_1\ g_1\ G_2\ g_2$	$g_1\ G_1\ g_2\ G_2$	$G_1\ g_1\ g_2\ g_2$
$G_1\ g_1\ G_2\ G_2$	$G_1\ g_1\ G_2\ g_2$	$g_1\ G_1\ g_2\ G_2$	$G_1\ g_1\ g_2\ g_2$

There are 9 genotypes:

$$D_1\ D_2,\ \ D_1\ H_2,\ \ D_1 R_2, \qquad H_1\ D_2,\ \ H_1\ H_2,\ \ H_1 R_2, \qquad R_1\ D_2,\ \ R_1\ H_2,\ \ R_1 R_2$$

with frequencies 1, 2, 1, 2, 4, 2, 1, 2, 1, respectively. There are 4 phenotypes:

$$D_1\ D_2, \qquad D_1\ R_2, \qquad R_1\ D_2, \qquad R_1\ R_2$$

with frequencies $9:3:3:1$.

Let us now generalize the case of n genes. For each gene, there are 4 possibilities and so the total number of possibilities for n gene is 4^n. For each gene, there are 3 genotypes and so the total number of genotypes is 3^n. For each gene, there are 2 phenotypes and so the total number of phenotypes for n genes is 2^n. With respect to each gene, there are 3 dominant phenotypes for each recessive phenotype.

Let us find how many phenotypes are dominant with respect to r genes. We can choose r genes in $\binom{n}{r}$ ways and, corresponding to each of these, there are 3 dominant phenotypes and 1 recessive phenotype so that the frequency of genotypes, which are dominant with respect to r genes and recessives with respect to $n-r$ genes, is $\binom{n}{r}3^r 1^{n-r}$, and the total of all these frequencies is

$$\sum_{r=0}^{n}\binom{n}{r}3^r = (3+1)^n = 4^n. \tag{7.1.1}$$

Thus of the 4^n possibilities with n genes, we have $\binom{n}{r}$ groups of 3^{n-r}, each dominant with respect to $n-r$ genes for $r = 0,\ 1,\ 2,\ \ldots,\ n$.

Thus we have one group of 3^n, $\binom{n}{1}$ groups of 3^{n-1} each, $\ldots$ $\binom{n}{r}$ groups of 3^{n-r} each, $\ldots$ and one group of 1 so that the phenotype ratios are:

$$\underbrace{3^n}_{\binom{n}{0}},\quad \underbrace{3^{n-1},3^{n-1},\ldots,3^{n-1}}_{\binom{n}{1}},\quad \cdots\quad \underbrace{3,3,\ldots,3}_{\binom{n}{n-1}},\quad \underbrace{3^0}_{\binom{n}{n}}$$

The frequencies of phenotypes are given by coefficients in the expansion of $(3x+y)^n$. Similarly, the frequencies of genotypes are given by coefficients of $(x+2y+z)^n$, and the frequency of a genotype dominant with respect to r genes, hybrid with respect to s genes, and recessive with respect to $n-r-s$ genes is

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

7.2 Multiple Alleles and Application to Blood Groups

So far we have considered the case of two alleles G and g only, but there may be a number of *alleles* corresponding to a given locus. The most important and elementary example is the gene determining blood groups, which has three alleles A, B, O giving rise to 9 possibilities (A,A), (A,B), (A,O), (B,A), (B,B), (B,O), (O,A), (O,B) and (O,O). There are, however, only 6 genotypes since (A,B), (B,A); (A,O), (O,A); and (B,O), (O,B) give the same genotypes. Since A and B dominate over O, there are only four phenotype groups, namely, $\{(A,A),\ (A,O),\ (O,A)\}$, $\{(B,B),\ (B,O),\ (O,B)\}$, $\{(A,B),\ (B,A)\}$, and $\{O,O\}$. These are denoted by A, B, AB and O, respectively.

Table 2 : Possible Blood Groups of Father in terms of Blood Groups of Mother and Child

Mother \ Child	A	B	AB	O
A	A, B, AB, O	B, AB	B, AB	A, B, O
B	A, AB	A, B, AB, O	A, AB	AB, O
AB	A, B, AB, O	A, B, AB, O	A, B, AB	ϕ
O	A, AB	B, AB	ϕ	A, B, O

We now get the results given in Table 1 for genotypes and blood groups of offspring. From Table 1, we can deduce the table for the possible blood groups for the father when we know the blood group of mother and child (Table 2). Table 2 is used in certain disputed legal cases to decide whether a certain child born of a certain mother can be the child of a given male.

Again, if the proportions of alleles A, B, O in the population are p, q, r, then the proportions of persons with blood groups A, B, AB, and O in the population are

$$p^2 + 2pr,\ q^2 + 2qr,\ 2qr,\ r^2 \quad [p^2 + 2pq + q^2 + 2qr + 2pq + r^2 = (p + q + r)^2 = 1]$$

If we know the division of the population according to blood groups, we can calculate $p,\ q,\ r$.

Table 1: Genotypes and Blood Groups of Offspring

Mother	Father	A		B		AB	O
		(*A*, *A*)	(*A*, *O*)	(*B*, *B*)	(*B*, *O*)	(*A*, *B*)	(*O*, *O*)
A	(*A*, *A*)	(*A*, *A*)	(*A*, *A*), (*A*, *O*)	(*A*, *B*)	(*A*, *B*), (*A*, *O*)	(*A*, *A*), (*A*, *B*)	(*A*, *O*)
		A	A	AB	AB, A	A, AB	A
	(*A*, *O*)	(*A*,*A*), (*A*,*O*)	(*A*,*A*), (*A*,*O*), (*O*,*O*)	(*A*,*B*), (*B*,*O*)	(*A*,*B*), (*A*,*O*), (*B*,*O*)	(*A*,*A*), (*A*,*O*), (*A*,*B*), (*B*,*O*)	(*A*,*O*), (*O*,*O*)
		A	A, O	AB, B	AB, A, B, O	A, AB, B	A, O
B	(*B*, *B*)	(*A*, *B*)	(*A*, *B*), (*B*, *O*)	(*B*, *B*)	(*B*, *B*), (*B*, *O*)	(*A*, *B*), (*B*, *B*)	(*B*, *O*)
		AB	AB, B	B	B	AB, B	B
	(*B*, *O*)	(*A*,*B*), (*A*,*O*)	(*A*,*B*), (*A*,*O*), (*B*,*O*), (*O*,*O*)	(*B*,*B*), (*B*,*O*)	(*B*,*B*), (*B*,*O*), (*O*,*O*)	(*A*,*B*), (*B*,*B*), (*A*,*O*), (*B*,*O*)	(*B*,*O*), (*O*,*O*)
		AB, A	AB, A, B, O	B	B, O	AB, B, A	B, O
AB	(*A*, *B*)	(*A*,*A*), (*A*,*B*)	(*A*,*A*), (*B*,*O*), (*A*,*O*), (*A*,*B*)	(*A*,*B*), (*B*,*B*)	(*A*,*B*), (*A*,*O*), (*B*,*O*)	(*A*,*A*), (*A*,*B*), (*B*,*B*)	(*A*,*O*), (*B*,*O*)
		A, AB	A, B, AO	AB, B	AB, A, B	A, AB, B	A, B
O	(*O*, *O*)	(*A*, *O*)	(*A*, *O*), (*O*, *O*)	(*B*, *O*)	(*B*, *O*), (*O*, *O*)	(*A*, *O*), (*B*, *O*)	(*O*, *O*)
		A	A, O	B	B, O	A, B	O

7.3 Models for Genetic Improvement: Selection and Mutation

7.3.1 Genetic Improvement through Cross Breeding

We have already discussed the results of crossing a breed successively with a dominant breed, a recessive breed, or a hybrid breed. We now consider the case when genes carrying undesirable characteristics are to be eliminated from a race and are to be replaced by genes with desirable characteristics. Thus let

$$g_1g_1, \qquad g_2g_2, \qquad g_3g_3, \qquad \cdots \qquad g_ng_n \tag{7.3.1}$$

denote n pairs of genes which we want to replace by the n pairs of genes

$$G_1G_1, \qquad G_2G_2, \qquad G_3G_3, \qquad \cdots \qquad G_nG_n. \tag{7.3.2}$$

We shall call the individual having gene pairs (7.3.2) as belonging to the G-race. We are not implying here that G's are dominant and g's are recessive. On crossing the given generation F_0 with the G-race, we get the first generation F_1, namely,

$$G_1g_1, \qquad G_2g_2, \qquad G_3g_3, \qquad \cdots \qquad G_ng_n, \tag{7.3.3}$$

so that one g in each pair is replaced by the corresponding G. Our object is to replace the other g also by successive crossing with the G-race. Successive crosses give us the generations

$$F_2, \qquad F_3, \qquad \cdots \qquad F_{m+1}, \qquad \ldots. \tag{7.3.4}$$

In every generation, there is a probability 1/2 that g_i has been replaced G_i, and there is a probability 1/2 that g_i has not been replaced by G_i, and so the probability that $(m+1)$-th generation still has g_i is $(1/2)^m$. Also, the probability that r of the n replacements of genes have not taken place is given by the binomial distribution, as

$$\binom{n}{r}\left(\frac{1}{2^m}\right)^r\left(1-\frac{1}{2^m}\right)^{n-r}. \tag{7.3.5}$$

If $r=0$, all the genes g_i have replaced by G_i, and the probability of this is

$$\binom{n}{0}\left(\frac{1}{2^m}\right)^0\left(1-\frac{1}{2^m}\right)^{n-0}=\left(1-\frac{1}{2^m}\right)^n. \tag{7.3.6}$$

As m approaches infinity, the probability approaches unity, regardless of the value of n. Thus, ultimately all genes g_i will be replaced by genes G_i for $i=1,2,\ldots,n$.

Unit 8

Course Structure

- Genetic Improvement through Elimination Recessives
- Selection and Mutation
- An Alternative Discussion of Selection

8.1 Genetic Improvement through Elimination Recessives

Another method of improving genetic composition in plants and animals is repeated elimination of recessives in each generation (e.g., by destroying recessive plants or by not allowing recessive animals to breed) and allowing random mating within the remaining members of the population.

In the n-th generation, if the proportions of dominants, hybrids, and recessives are p_n, q_n and r_n, then, in the $(n+1)$-th generation, these proportions are

$$p_{n+1} = \left(p_n + \frac{1}{2}q_n\right)^2, \quad q_{n+1} = 2\left(p_n + \frac{1}{2}q_n\right)\left(r_n + \frac{1}{2}q_n\right), \quad r_{n+1} = \left(r_n + \frac{1}{2}q_n\right)^2. \tag{8.1.1}$$

In the n-th generation, if recessives are eliminated, then the new proportions in the $(n+1)$-th generation are given by

$$p_{n+1} = \left(p'_n + \frac{1}{2}q'_n\right)^2, \quad q_{n+1} = 2\left(p'_n + \frac{1}{2}q'_n\right)\left(\frac{1}{2}q'_n\right), \quad r_{n+1} = \left(\frac{1}{2}q'_n\right)^2. \tag{8.1.2}$$

where p'_n, q'_n are the new proportions in the n-th generation after elimination of the recessives so that

$$\frac{p'_n}{q'_n} = \frac{p_n}{q_n}, \quad p'_n + q'_n = 1. \tag{8.1.3}$$

From (8.1.2) and (8.1.3),

$$p_{n+1} = \left(1 - \frac{1}{2}q'_n\right)^2, \quad q_{n+1} = q'_n\left(1 - \frac{1}{2}q'_n\right), \quad r_{n+1} = \left(\frac{1}{2}q'_n\right)^2. \tag{8.1.4}$$

After eliminating the recessives from the population, we get

$$\frac{p'_{n+1}}{1-\frac{1}{2}q'_n} = \frac{q'_{n+1}}{q'_n} = \frac{1}{1+\frac{1}{2}q'_n}$$
$$\Rightarrow q'_{n+1} = \frac{q'_n}{1+\frac{1}{2}q'_n} \tag{8.1.5}$$

This is a difference equation for solving for q'_n. Substituting

$$u_n = \frac{1}{q_n} \tag{8.1.6}$$

in (8.1.5), we get

$$u_{n+1} = u_n + \frac{1}{2} \tag{8.1.7}$$

whose solution is

$$u_n = A + \frac{1}{2}n \tag{8.1.8}$$

so that

$$q'_n = \frac{1}{A+\frac{1}{2}n}. \tag{8.1.9}$$

To determine A, we make use of $p = (1-\sqrt{r})^2, \quad q = 2\sqrt{r}(1-\sqrt{r}), \quad r = r$, to get

$$q'_1 = \frac{1}{p+q} = \frac{2\sqrt{r}}{1+\sqrt{r}} \tag{8.1.10}$$

so that

$$A = \frac{1}{2\sqrt{r}}. \tag{8.1.11}$$

Also,

$$q'_n = \frac{2\sqrt{r}}{1+n\sqrt{r}}, \tag{8.1.12}$$
$$r'_{n+1} = \left(\frac{1}{2}q'_n\right)^2 = \frac{r}{(1+n\sqrt{r})^2}. \tag{8.1.13}$$

This gives the proportion of recessives in the $(n+1)$-th generation. Given the proportion of recessives in the original stable population, we can find, by using (8.1.13), the number of generation in which we can reduce the proportion of recessives below any given limit by elimination of recessives at all stages. We can also find that $p_n \to 1,\ q_n \to 0,\ r_n \to 0$ as $n \to \infty$.

Instead of eliminating all the recessives, we may keep a fraction k of the recessives. The basic equations in

this case are

$$p'_n = \frac{1-kr_n}{1-r_n}p_n, \quad q'_n = \frac{1-kr_n}{1-r_n}q_n, \quad r'_n = kr_n, \tag{8.1.14}$$

$$p'_n + q'_n + r'_n = 1, \tag{8.1.15}$$

$$p_{n+1}(p'_n + \frac{1}{2}q'_n)^2 = \left(\frac{1-kr_n}{1-r_n}\right)^2 \left(p_n + \frac{1}{2}q_n\right)^2, \tag{8.1.16}$$

$$q_{n+1} = 2\left(p'_n + \frac{1}{2}q'_n\right)\left(r'_n + \frac{1}{2}q'_n\right) = 2\left(\frac{1-kr_n}{1-r_n}\right)\left(p_n + \frac{1}{2}q_n\right)\left(kr_n + \frac{1}{2}\frac{1-kr_n}{1-r_n}q_n\right), \tag{8.1.17}$$

$$r_{n+1} = \left(r'_n + \frac{1}{2}q'_n\right)^2 = \left(kr_n + \frac{1}{2}\frac{1-kr_n}{1-r_n}q_n\right)^2, \tag{8.1.18}$$

$$p'_{n+1} = \frac{1-kr_{n+1}}{1-r_{n+1}}p_{n+1} = \frac{1-kr_{n+1}}{1-r_{n+1}}\left(p'_n + \frac{1}{2}q'_n\right)^2, \tag{8.1.19}$$

$$q'_{n+1} = \frac{1-kr_{n+1}}{1-r_{n+1}}q_{n+1} = 2\frac{1-kr_{n+1}}{1-r_{n+1}}\left(p'_n + \frac{1}{2}q'_n\right)\left(r'_n + \frac{1}{2}q'_n\right) \tag{8.1.20}$$

$$r'_{n+1} = kr_{n+1} = k\left(r'_n + \frac{1}{2}q'_n\right). \tag{8.1.21}$$

From (8.1.16)-(8.1.18), we get two simultaneous nonlinear difference equations of the first order for determining p_n and q_n and from (8.1.19)-(8.1.21), we obtain two nonlinear simultaneous difference equations to determine p'_n and q'_n. However, the equations are complicated, and closed-form solutions cannot be easily determined.

8.2 Selection and Mutation

Let the proportions of genes G and g in the n-th generation be P_n and Q_n so that in the $(n+1)$-th generation, the proportions of $GG,\ Gg$, and gg are $P_n^2,\ 2P_nQ_n$, and Q_n^2. Suppose the probabilities of survival of these are $S(1-K),\ S$ and $S(1-k)$, respectively, where $|k|$ and $|K|$ are less than unity. Then the relative proportions in the $(n+1)$-th generation are

$$S(1-K)P_n^2, \quad 2SP_nQ_n, \quad S(1-k)Q_n^2 \tag{8.2.1}$$

so that the relative proportions of G and g in this generation are

$$2S(1-K)P_n^2 + 2SP_nQ_n, \quad 2SP_nQ_n + 2S(1-k)Q_n^2 \tag{8.2.2}$$

and hence

$$\frac{P_{n+1}}{Q_{n+1}} = \frac{(1-K)(P_n^2/Q_n^2) + (P_n/Q_n)}{(P_n/Q_n) + (1-k)} \tag{8.2.3}$$

$$\Rightarrow u_{n+1} = \frac{(1-K)u_n^2 + u_n}{u_n + (1-k)}, \quad \text{where} \quad u_n = \frac{P_n}{Q_n}. \tag{8.2.4}$$

This is a nonlinear difference equation of the first order. Knowing u_1, we can find, step by step, u_n. The equilibrium solution is obtained by putting $u_n = u_{n+1} = u$ which gives

$$u = \frac{(1-K)u^2 + u}{u + 1 - k}$$

$$\Rightarrow u(uK - k) = 0 \tag{8.2.5}$$

so that $u = 0$, or $u = k/K$, $1/u = 0$ i.e. either dominants or recessives survive. However, a non-trivial equilibrium solution is

$$u = k/K. \tag{8.2.6}$$

Since this equilibrium solution has to be positive, both k and K have to be either positive or negative, i.e., either the hetrozygotes have to be the fittest or they have to be the least fit. If $K = k$, the equilibrium of G and g are the same.

To discuss the stability of the equilibrium solution of (8.2.5), we note that (8.2.4) give

$$u_{n+1} - u_n = \frac{Ku_n}{u_n + 1 - k}\left(\frac{k}{K} - u_n\right) \tag{8.2.7}$$

or

$$\frac{u_{n+1} - u_n}{k/K - u_n} = \frac{Ku_n}{u_n + 1 - k} \tag{8.2.8}$$

and

$$\frac{u_{n+1} - k/K}{u_n - k/K} = \frac{u_n(1-K) + (1-k)}{u_n + 1 - k} \tag{8.2.9}$$

From (8.2.9), we deduce the following results:

(i) If $0 < k < K < 1$ when $u_n > k/K$, we find that $u_{n+1} < u_n$ and $u_{n+1} > k/K$, i.e., u_{n+1} is nearer to k/K than u_n, and the sequence $\{u_n\}$ monotonically decreases to k/K. On the other hand, if $u_n < k/K$, then $u_{n+1} > u_n$ and $u_{n+1} < k/K$ so that the sequence $\{u_n\}$ monotonically increases to k/K. In the first case, we get a monotonically decreasing sequence bounded below; in the second case, we get a monotonically increasing sequence bounded above. In either case, we find that, if $0 < k < K < 1$, then the equilibrium solution is stable.

(ii) If k and K are both negative and $k/K < 1$, then (8.2.8) and (8.2.9) show that $u_{n+1} - u_n$, $u_n - k/K$, and $u_{n+1} - k/K$ have the same sign so that, if $u_n > k/K$, then $u_{n+1} > k/K$ and $u_{n+1} > k/K$ and $u_{n+1} > u_n$, and hence u_{n+1} is farther from k/K than u_n. Similarly, if $u_n < k/K$, then $u_{n+1} < k/K$ and $u_{n+1} < u_n$ so that here too u_{n+1} is farther from k/K than u_n. Thus, when k and K are both negative, the equilibrium solution is unstable.

Thus, when the hybrids are the fittest, we get a stable equilibrium; when these are the least fit, we obtain an unstable equilibrium. Now (8.2.8) can be written as

$$\frac{u_{n+1} - u_n}{(n+1) - 1} = \frac{Ku_n}{u_n + 1 - k}\left(\frac{k}{K} - u_n\right). \tag{8.2.10}$$

When the change in one generation is not substantially different from the changes in the preceding or succeeding generations (e.g., when K is very small or when there are small oscillations about the equilibrium position), we can replace (8.2.10) by the differential equation

$$\frac{du}{dn} = \frac{Ku}{u + 1 - k}\left(\frac{k}{K} - u\right) \tag{8.2.11}$$

which gives, on integration,

$$\left(\frac{1}{k} - 1\right)\ln u + \left(\frac{1}{K} + \frac{1}{k} - 1\right)\ln\left|\frac{k}{K} - u\right| = n + \text{constant}. \tag{8.2.12}$$

From this we can discuss the variation of u from generation to generation.

Similarly, we can discuss the *balance between selection and mutation.* Let the probabilities of survival of D, H, R be $S(1-K)$, $S(1-K)$, and S, respectively, and let μ be the probability of a mutation from g to G in one generation. Then we get

$$\begin{aligned}\frac{P_{n+1}}{Q_{n+1}} &= \frac{2S(1-K)P_n + 2S(1-K)P_nQ_n + \mu[2S(1-K)P_nQ_n + 2SQ_n^2]}{[2S(1-K)P_nQ_n + 2SQ_n^2](1-\mu]} \\ &= \frac{\left(\frac{P_n}{Q_n}+1\right)(1-K)\frac{P_n}{Q_n} + \mu\left(\frac{P_n}{Q_n}+1-K\frac{P_n}{Q_n}\right)}{\left(\frac{P_n}{Q_n}-1-K\frac{P_n}{Q_n}\right)(1-\mu)}\end{aligned} \tag{8.2.13}$$

so that

$$u_{n+1} = \frac{(u_n+1)(1-K)u_n + \mu(u_n+1-Ku_n}{(u_n+1-Ku_n)(1-\mu)}. \tag{8.2.14}$$

If we assume that u_n is very small (which is justified since mutation rates are small, i.e., of the order of 10^{-5} or less), then genes with lower fitness level can be maintained only at very low frequency by mutation. Equation (8.2.14) can now be written as

$$u_{n+1} = u_n(1-K) + \mu, \tag{8.2.15}$$

In equilibrium, this gives

$$u = \mu/K. \tag{8.2.16}$$

8.3 An Alternative Discussion of Selection

In §8.2, we have discussed the problem of selection in terms of the ratio $\dfrac{P_n}{Q_n}$. We can also discuss the same problem in terms of P_n alone. If $\sigma_1, \sigma_2, \sigma_3$ denote the proportions of $D,\ H,\ R$ which survive from birth to reproduction, we get

$$P_{n+1} = \frac{\sigma_1 P_n^2 + \sigma_2 P_nQ_n}{\sigma_1 P_n^2 + 2\sigma_2 P_nQ_n + \sigma_3 Q_n^2} = f(P_n). \tag{8.3.1}$$

If $n \to \infty$, then $P_n \to P,\ P_{n+1} \to P$, and $Q_n \to 1-P$, so that

$$\sigma_1 P^3 + 2\sigma_2 P^2(1-P) + \sigma_3(1-P)^2P - \sigma_1 P^2 - \sigma_2 P(1-P) = 0 \tag{8.3.2}$$

which gives the three equilibrium solutions

$$P = 0, \qquad P = 1, \qquad P_e = \frac{\sigma_2 - \sigma_3}{2\sigma_2 - \sigma_1 - \sigma_3}. \tag{8.3.3}$$

The third solution exists if $0 < P_e < 1$, and is of special significance, because, in this case, al the three forms, namely, $D,\ H,\ R$ survive. We can now write Eq.(8.3.1) as

$$\left(P_{n+1} - \frac{\sigma_2 - \sigma_3}{2\sigma_3 - \sigma_1 - \sigma_3}\right) = \frac{\sigma_1 P_n + \sigma_3 Q_n}{\sigma_1 P_n^2 + 2\sigma_2 P_nQ_n + \sigma_3 Q_n^2}\left[P_n - \frac{\sigma_2 - \sigma_3}{2\sigma_2 - \sigma_1 - \sigma_3}\right]. \tag{8.3.4}$$

This also shows that, if $P_n \to P_e$, then P_{n+1} also approaches P_e. Now

$$\begin{aligned}\frac{\sigma_1 P_n + \sigma_3 Q_n}{\sigma_1 P_n^2 + 2\sigma_2 P_nQ_n + \sigma_3 Q_n^2} &= \frac{\sigma_1 P_n + \sigma_3 Q_n}{(\sigma_1 P_n + \sigma_3 Q_n)(P_n + Q_n) + (2\sigma_2 - \sigma_1 - \sigma_3)P_nQ_n} \\ &= \frac{1}{1 + \frac{2\sigma_2 - \sigma_1 - \sigma_3}{\sigma_1 P_n + \sigma_3 Q_n}P_nQ_n}\end{aligned} \tag{8.3.5}$$

From (8.3.4) and (8.3.5), we deduce the following results:

(i) If $\sigma_2 > \sigma_1, \sigma_3$, then the first factor on the right-hand side of (8.3.4) is less than unity and

$$|P_{n+1} - P_e| < |P_n - P_e| \tag{8.3.6}$$

so that $P_n \to P_e$ as $n \to \infty$, regardless of the initial value P_0. Therefore, the equilibrium point is stable (see Fig. 1).

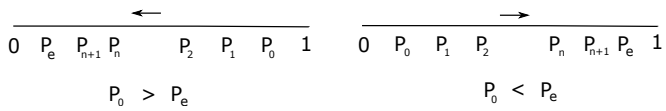


Fig. 1: $\sigma_2 > \sigma_1, \sigma_3$: Stable equilibrium

(ii) If $\sigma_2 < \sigma_1, \sigma_3$, then the first factor on the right-hand side of (8.3.4) is greater than unity and

$$|P_{n+1} - P_e| > |P_n - P_e|. \tag{8.3.7}$$

Hence the equilibrium is unstable. If $P_0 < P_e$, then $P_n \to 0$, and if $P_0 > P_e$, then $P_n \to 1$ (see Fig. 2).

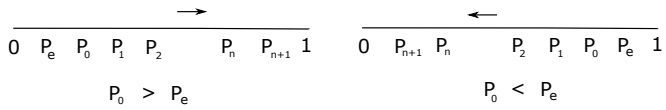


Fig. 2: $\sigma_2 < \sigma_1, \sigma_3$: Unstable equilibrium

These results are the same as those of §8.2 and show that the equilibrium is stable if the heterozygotes have the greatest chance of survival and is unstable if the heterozygotes have the least chance of survival.

For the stability of the equilibrium, it is necessary that

$$\begin{aligned} &2\sigma_2 - \sigma_1 - \sigma_3 > 0, \qquad \frac{\sigma_2 - \sigma_3}{2\sigma_2 - \sigma_1 - \sigma_3} > 0, \qquad \sigma_2 - \sigma_3 > 0, \\ &\frac{\sigma_2 - \sigma_3}{2\sigma_2 - \sigma_1 - \sigma_3} < 1, \qquad \sigma_2 - \sigma_3 < 2\sigma_2 - \sigma_1 - \sigma_3, \qquad \sigma_2 - \sigma_1 > 0. \end{aligned} \tag{8.3.8}$$

Thus, for P_e to represent and equilibrium solution, it is both necessary and sufficient that $\sigma_2 > \sigma_1, \sigma_3$.

The convergence of a sequence $\{P_n\}$ to a limit P_e is said to be *geometric* at the rate c for $0 < |c| < 1,\ 0 < a < |c|$ if

$$\lim_{n\to\infty} \frac{|P_n - P_e|}{c^n} < \infty, \qquad \lim_{n\to\infty} \frac{|P_n - P_e|}{a^n} = \infty. \tag{8.3.9}$$

The convergence is said to be *algebraic* if

$$\lim_{n\to\infty} n^k |P_n - P_e| = \text{a positive constant}. \tag{8.3.10}$$

Using (8.3.9) and (8.3.10), we find that, when $\sigma_2 > \sigma_1, \sigma_3$, the convergence is geometric at the rate

$$1/\left[1 + \frac{2\sigma_2 - \sigma_1 - \sigma_3}{\sigma_1 P_e + \sigma_3 Q_e} P_e Q_e\right] = \frac{\sigma_2(\sigma_1 + \sigma_3) - 2\sigma_1\sigma_3}{\sigma_2^2 - \sigma_1\sigma_3}. \tag{8.3.11}$$

Unit 9

Course Structure

- Some basic concepts of fluid dynamics
- Hegen-Poiseuille Flow
- Inlet Length Flow
- Reynolds Number Flow
- Non-Newtonian Fluids

9.1 Introduction

In large and medium sized arteries, those more typically affected by vascular diseases, blood can be modelled by means of the Navier-Stokes (NS) equation for incompressible homogeneous Newtonian fluids. Non-Newtonian rheological models are necessary for describing some specific flow processes, such as clotting or sickle cell diseases, or more generally flow in capillaries. Let us recall some preliminary concepts of fluid dynamics.

9.2 Some Basic Concepts of Fluid Dynamics

9.2.1 Navier-Stokes Equations for the Flow of a Viscous Incompressible Fluid

Let $u(x, y, z, t)$, $v(x, y, z, t)$, $w(x, y, z, t)$, and $p(x, y, z, t)$ denote respectively the three velocity components are pressure at the point (x, y, z) at time t in a fluid with constant density ρ and viscosity coefficient μ. Then the *equation of continuity*, which expresses the fact that the amount of fluid entering a unit volume per unit time is the same as the amount of the fluid leaving it per unit time, is given by

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0. \tag{9.2.1}$$

The *equations of motion* are obtained from Newton's second law of motion which states that the product of mass and acceleration of any fluid element is equal to the resultant of all the external body forces acting on the element and to the surface forces acting on the fluid volume due to the action of the remaining fluid on

the element. The equations of motion, knows as *Navier-Stokes* equations, for the of a Newtonian viscous incompressible fluid are

$$\rho\left(\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}+w\frac{\partial u}{\partial z}\right) = X-\frac{\partial p}{\partial x}+\mu\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}+\frac{\partial^2 u}{\partial z^2}\right) \tag{9.2.2}$$

$$\rho\left(\frac{\partial v}{\partial t}+u\frac{\partial v}{\partial x}+v\frac{\partial v}{\partial y}+w\frac{\partial v}{\partial z}\right) = Y-\frac{\partial p}{\partial y}+\mu\left(\frac{\partial^2 v}{\partial x^2}+\frac{\partial^2 v}{\partial y^2}+\frac{\partial^2 v}{\partial z^2}\right) \tag{9.2.3}$$

$$\rho\left(\frac{\partial w}{\partial t}+u\frac{\partial w}{\partial x}+v\frac{\partial w}{\partial y}+w\frac{\partial w}{\partial z}\right) = Z-\frac{\partial p}{\partial z}+\mu\left(\frac{\partial^2 w}{\partial x^2}+\frac{\partial^2 w}{\partial y^2}+\frac{\partial^2 w}{\partial z^2}\right) \tag{9.2.4}$$

If the external body forces $X,\ Y,\ Z$ form a conservative system, there exists a potential function Ω such that

$$\begin{aligned} &X=-\frac{\partial\Omega}{\partial x},\quad Y=-\frac{\partial\Omega}{\partial y},\quad Z=-\frac{\partial\Omega}{\partial z}\\ &X-\frac{\partial p}{\partial x}=-\frac{\partial}{\partial x}(\Omega+p),\quad Y-\frac{\partial p}{\partial y}=-\frac{\partial}{\partial y}(\Omega+p),\quad Z-\frac{\partial p}{\partial z}=-\frac{\partial}{\partial z}(\Omega+p)\end{aligned} \tag{9.2.5}$$

so that p is effectively replaced by $p+\Omega$.

If $X,\ Y,\ Z$ are known or are absent, (9.2.1)-(9.2.4) give a system of four coupled nonlinear partial differential equations for the four unknown functions $u,\ v,\ w$, and p. These equations have to be solved subject to certain *initial conditions* giving the motion of the fluid at time $t=0$ and certain prescribed *boundary conditions* on the surfaces with which the fluid may be in contact or conditions which may hold at very large distances from the surfaces. Usually, the boundary conditions are provided by the *no-slip condition* according to which both tangential and normal components of the fluid velocity vanish at all points of the surfaces of the stationary bodies with which the the fluid may be in contact. However, if a body is moving, then the tangential and normal components of the fluid velocity at any point of contact are the same as those of the moving body at that point.

We can simplify the basic equations (9.2.2)-(9.2.4) when

(i) there are *no external body forces,* i.e., when $X=0,\ Y=0,\ Z=0$, or when the external forces form a conservative system

(ii) the motion is *steady*, i.e., when there is no variation with respect to time so that $u,\ v,\ w$, and p are functions of $x,\ y,\ z$ only and $\dfrac{\partial u}{\partial t},\dfrac{\partial v}{\partial t},\dfrac{\partial w}{\partial t}$ and $\frac{\partial p}{\partial t}$ are all zero.

(iii) the motion is *two dimensional*, i.e., when it is the same in all places parallel to $z=0$ plane and, in particular, when $w=0$ and when there is no variation with respect to z. In this case, the three equations we get for the three unknowns, namely, $u(x,y,t),\ v(x,y,t)$ and $p(x,y,t)$, are

$$\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}=0, \tag{9.2.6}$$

$$\rho\left(\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}\right)=-\frac{\partial p}{\partial x}+\mu\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\right), \tag{9.2.7}$$

$$\rho\left(\frac{\partial v}{\partial t}+u\frac{\partial v}{\partial x}+v\frac{\partial v}{\partial y}\right)=-\frac{\partial p}{\partial y}+\mu\left(\frac{\partial^2 v}{\partial x^2}+\frac{\partial^2 v}{\partial y^2}\right). \tag{9.2.8}$$

Equation (9.2.6) can be satisfied by introducing the *stream function* $\psi(x,y)$ which is such that

$$u=\frac{\partial\psi}{\partial y},\qquad v=-\frac{\partial\psi}{\partial x}. \tag{9.2.9}$$

Substituting in (9.2.7) and (9.2.8) and eliminating p between them, we get

$$\frac{\partial}{\partial t}\nabla^2\psi + \frac{\partial\psi}{\partial y}\frac{\partial}{\partial x}\nabla^2\psi - \frac{\partial\psi}{\partial x}\nabla^2\psi + \nu\nabla^4\psi, \tag{9.2.10}$$

where $\nu = \mu/\rho$ is the kinematic viscosity and ∇^2 is the Laplacian operator defined by

$$\nabla^2 \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}, \qquad \nabla^4 \equiv \nabla^2(\nabla^2). \tag{9.2.11}$$

The *vorticity* of this two dimensional flow is defined by

$$\omega = \frac{1}{2}\left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}\right) = -\frac{1}{2}\nabla^2\psi. \tag{9.2.12}$$

From (9.2.10) and (9.2.12), we get

$$\frac{\partial\omega}{\partial t} + \frac{\partial\psi}{\partial y}\frac{\partial\omega}{\partial x} - \frac{\partial\psi}{\partial x}\frac{\partial\omega}{\partial y} = \frac{\partial\omega}{\partial t} + \frac{\partial(\omega,\psi)}{\partial(x,y)} = \nu\nabla^4\psi. \tag{9.2.13}$$

(iv) The basic equations (9.2.2)-(9.2.4) can also be simplified when the motion is axially symmetric, i.e., when it is symmetrical about an axis. Here we use the cylindrical polar coordinate (r,θ,z), where the axis of symmetry is taken as the axis of z. There are, in general, three components of velocity, namely, v_r along the radius vector perpendicular to the axis, v_θ perpendicular to the axis and the radius vector, and v_z parallel to the axis of z. For the axi-symmetric case, we take $v_\theta = 0$, and we also take v_r, v_z and p to be independent of θ. In this case, the equation of continuity and the equations of motion are given by

$$\frac{1}{r}\frac{\partial}{\partial r}(rv_r) + \frac{\partial}{\partial z}v_z = 0, \tag{9.2.14}$$

$$\rho\left(\frac{\partial v_r}{\partial t} + v_r\frac{\partial v_r}{\partial r} + v_z\frac{\partial v_r}{\partial z}\right) = -\frac{\partial p}{\partial r} + \mu\left(\frac{\partial^2 v_r}{\partial r^2} + \frac{\partial^2 v_r}{\partial z^2} + \frac{1}{r}\frac{\partial v_r}{\partial r} - \frac{v_r}{r^2}\right), \tag{9.2.15}$$

$$\rho\left(\frac{\partial v_z}{\partial t} + v_r\frac{\partial v_z}{\partial r} + v_z\frac{\partial v_z}{\partial z}\right) = -\frac{\partial p}{\partial z} + \mu\left(\frac{\partial^2 v_z}{\partial r^2} + \frac{\partial^2 v_r}{\partial z^2} + \frac{1}{r}\frac{\partial v_r}{\partial r}\right). \tag{9.2.16}$$

We can satisfy (9.2.14) by introducing the stream function ψ defined by

$$\frac{1}{r}\frac{\partial\psi}{\partial r} = v_z, \qquad \frac{1}{r}\frac{\partial\psi}{\partial z} = -v_r \tag{9.2.17}$$

Substituting (9.2.17) in (9.2.15) and (9.2.16) and eliminating p, we get the fourth-order partial differential equation for ψ, as

$$\frac{\partial}{\partial t}(D^2\psi) - \frac{1}{r}\frac{\partial(\psi, D^2\psi)}{\partial(r,z)} - \frac{2}{r^2}\frac{\partial\psi}{\partial z}D^2\psi = \nu D^4\psi, \tag{9.2.18}$$

where

$$D^2 \equiv \frac{\partial^2}{\partial r^2} - \frac{1}{r}\frac{\partial}{\partial r} + \frac{\partial^2}{\partial z^2}, \qquad D^2\psi = D^2(D^2\psi). \tag{9.2.19}$$

After solving for ψ, we can obtain pressure p and vorticity ω by using the equation

$$\frac{\partial^2 p}{\partial r^2} + \frac{\partial^2 p}{\partial z^2} + \frac{1}{r}\frac{\partial p}{\partial r} = \frac{2}{r}\left[\frac{\partial^2\psi}{\partial z^2}\left(\frac{\partial^2\psi}{\partial r^2} - \frac{1}{r}\frac{\partial\psi}{\partial r}\right) - \left(\frac{\partial\psi}{\partial z}\right)^2 + \frac{\partial^2\psi}{\partial z\partial r}\left(\frac{1}{r}\frac{\partial\psi}{\partial r} - \frac{\partial^2\psi}{\partial z^2}\right)\right] \tag{9.2.20}$$

$$\omega = -D^2\psi = -\frac{1}{r^2}\left(\frac{\partial^2\psi}{\partial z^2} - \frac{1}{r}\frac{\partial\psi}{\partial r} + \frac{\partial^2\psi}{\partial z^2}\right). \tag{9.2.21}$$

9.3 Hagen-Poiseuille Flow

The equation of fluid flow we have obtained are rather complicated and, in general, have to be integrated either by using approximations or numerically with the help of computers. There are, however, a few exact solutions. One of these was investigated by physician Poiseuille because of his interest in the flow of blood in arteries (see Fig. 9.1).

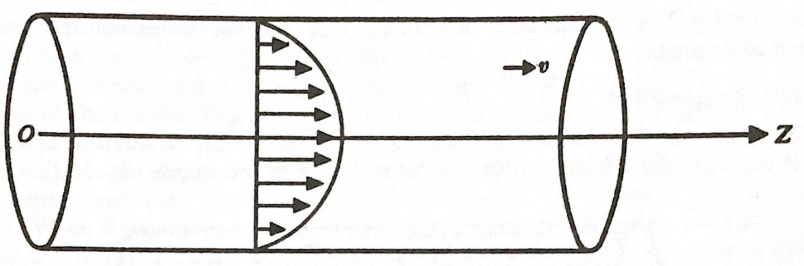


Figure 9.1: Velocity profile for Poiseuille flow

We consider steady flow when there is only one velocity component parallel to the axis so that $v_r = 0, v_\theta = 0$, and $v_z = v$. Then the equation of continuity gives

$$v_z = v(r). \tag{9.3.1}$$

The equations of motion, (9.2.15) and (9.2.16), now give

$$\frac{\partial p}{\partial r} = 0, \qquad \frac{d^2v}{dr^2} + \frac{1}{r}\frac{dv}{dr} = \frac{1}{\mu}\frac{\partial p}{\partial z}. \tag{9.3.2}$$

From (9.3.2), $-\dfrac{\partial p}{\partial z}$ must be a constant. Let us denote this constant pressure gradient by P. Then (9.3.2) gives

$$\frac{1}{r}\frac{d}{dr}\left(r\frac{dv}{dr}\right) = -\frac{P}{\mu}. \tag{9.3.3}$$

Integrating (9.3.3) twice, we get

$$r\frac{dv}{dr} = -\frac{1}{2\mu}Pr^2 + A, \qquad v(r) = -\frac{Pr^2}{4\mu} + A\ln r + B, \tag{9.3.4}$$

but velocity on the axis (i.e., at $r = 0$) must be finite, giving $A = 0$, and it should vanish on $r = a$ because of the no-slip condition so that

$$B = \frac{Pa^2}{4\mu}, \qquad v = \frac{P}{4\mu}(a^2 - r^2). \tag{9.3.5}$$

The velocity is zero on the surface and is maximum on the axis. In fact, the velocity profile is parabolic and in the three-dimensional space, it may be regarded as a paraboloid of revolution.

The total flux across any section, i.e., the total volume of the fluid crossing any section per unit time, is given by

$$Q = \int_0^a 2\pi r\, v\, dr = \frac{\pi a^4}{8\mu}P. \tag{9.3.6}$$

The result that the flux is proportional to the pressure gradient and to the fourth power of the radius of the tube was discovered experimentally by Hagen and rediscovered independently by Poiseuille. The importance of this result is that it can be confirmed experimentally and can be used to determine μ.

9.4 Inlet Length Flow

When a fluid enters a tube from a large reservoir where the velocity is uniform and parallel to the axis of the tube, the velocity profile is a flat surface at the entry (see Fig. 9.2). Immediately after entry, the velocity near the surface is affected by the friction of the surface, but the velocity profile near the axis still remains flat. As the fluid moves further in the tube, the flat portion decreases, and at the section corresponding to A, the paraboloidal velocity profile for the fully develped flow is reached. The flow in the region OA is called the *entry region* (or *inlet*) *flow* and the flow beyond A (in region III) is called the *fully developed flow*. The length OA is called the entry length. the flow in the entry length portion itself consists two parts. The flow in region I near the surface is called the *boundary layer flow*; the flow in region II is called the *core flow*or the *plug flow*. In fact, the flow approaches the parabolic velocity profile asymptotically, and we may define the entry length as the length in which 99 per cent of the final velocity profile is attained.

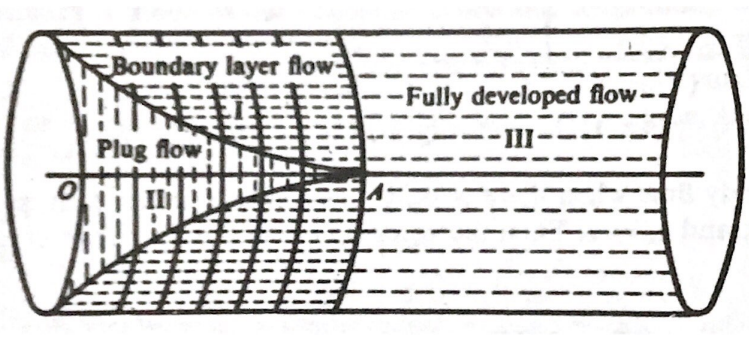


Figure 9.2: Inlet length velocity profiles

9.5 Reynolds Number of Flows

In (9.2.12)-(9.2.14), the terms on the left-hand sides represent the *inertial forces* (mass $\times$ acceleration) while the tree terms on the right-hand side of each equation represent respectively the *body forces*, *pressure forces* and *viscous forces*. If U is a typical velocity and L is a typical length, the inertial forces are of the order $\dfrac{\rho U}{L}$ and the viscous forces are of the order $\dfrac{\mu U}{L}$. The ratio of these forces is of the order

$$Re = \frac{\rho U^2 L^2}{\mu L U} = \frac{\rho U L}{\mu} = \frac{UL}{\mu} \tag{9.5.1}$$

where $\mu = \mu/\rho$ is called the *kinematic viscosity* of the fluid. Now the dimensions of μ and ρUL are given by

$$\mu = \frac{\text{stress}}{\text{strain rate}} = \frac{\text{force per unit area}}{velocity/length} = \frac{MLT^{-2}L^{-2}}{T^{-1}} = ML^{-1}T^{-1}, \tag{9.5.2}$$

$$\rho UL = ML^{-3}LT^{-1}L = ML^{-1}T^{-1}. \tag{9.5.3}$$

Thus, Re is a dimensionless number. It is called *Reynold's number*, after Osborn Reynold who in 1890 showed that the fully developed Poiseuille flow in a circular tube changes from stream line or *laminar flow* to *turbulent flow* when this number, based on the diameter of the tube, exceed a critical value of about 2000.

When Reynold number is small, viscous forces dominate over inertial forces. If we neglect the inertial forces, which we can justifiably do when $Re << 1$, (9.2.13) and (9.2.18) give

$$\nabla^4 \psi = 0. \tag{9.5.4}$$

Low Reynold number flows are also characteristic of

(i) *lubrication theory,* which we shall find useful in our study of lubrication of human joints;

(ii) *microcirculation* or flows of blood in blood vessel of diameter less than $100\ \mu m$;

(iii) *air flows* in alveolar passages of diameter less than a few hundred micron; and

(iv) *swimming of microoranisms* with Re of the order 10^{-3}

9.6 Non-Newtonian Fluids

For the simple motions we shall consider, there is only one non-zero component τ of the stress tensor and only one non-zero component e of the rate of strain. In general, each of these tensor has six distinct components. The functional relations between the components of the two tensors depend on the fluid under consideration and determine the *constitutive equations* for the fluid. For Newtonian viscous fluids,

$$\tau = \mu e, \tag{9.6.1}$$

where μ is the constant coefficient of viscosity. We have fluids for which μ itself may be a function of strain rate, i.e., for which stress becomes a non-linear or non-homogeneous function of strain rate (see Fig. 9.3). Such fluids are called *non-Newtonian fluids*. One important call of non-Newtonian fluids is that of *power-law fluids* with constitutive equations

$$\tau = \mu e^n = \mu e^{n-1} e. \tag{9.6.2}$$

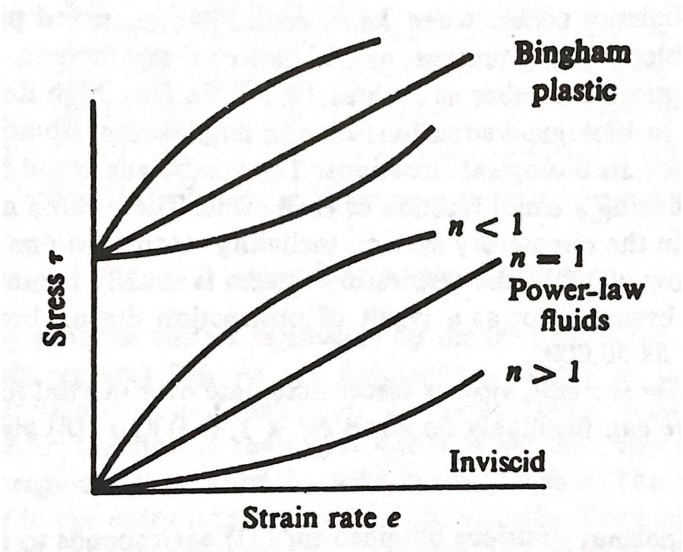


Figure 9.3: Inlet length velocity profiles

If $n < 1$, we get a *pseudo-plastic power-law fluid* in which the effective viscosity coefficient μe^{n-1} decreases with intreaing strain rate.

If $n > 1$, we a *dilatant power-law* fluid in which the effective viscosity coefficient increases with increasing strain rate.

If $n = 1$, Eq.(9.6.2) gives the Newtonian viscous fluid as a special case.

Another important non-Newtonian fluid, namely, the Bingham plastic, has the constitutive equation

$$\begin{aligned} \tau &= \mu e + \tau_0 \qquad (\tau \geq \tau_0), \\ e &= 0 \qquad (\tau \leq \tau_0). \end{aligned} \tag{9.6.3}$$

It shows an yield stress τ_0 and, if $\tau < \tau_0$, no flow takes place. Some other laws which have been proposed for special non-Newtonian fluids are:

- Herschel-Bulkley fluid

$$\begin{aligned} \tau &= \mu e^n + \tau_0 \qquad (\tau \geq \tau_0), \\ e &= 0 \qquad (\tau \leq \tau_0). \end{aligned} \tag{9.6.4}$$

- Casson fluid

$$\begin{aligned} \tau^{\frac{1}{2}} &= \mu^{\frac{1}{2}} e^{\frac{1}{2}} + \tau_0^{\frac{1}{2}} \qquad (\tau \geq \tau_0), \\ e &= 0 \qquad (\tau \leq \tau_0). \end{aligned} \tag{9.6.5}$$

- Prandtl fluid

$$\tau = A \, \sin^{-1}\left(\frac{e}{c}\right) \tag{9.6.6}$$

- Prandtl-Eyring fluid

$$\tau = Ae + B \, \sin^{-1}\left(\frac{e}{c}\right) \tag{9.6.7}$$

Exercise 9.6.1. 1. Verify (9.2.10), (9.2.12), (9.2.18), (9.2.20) and (9.2.21).

2. Discuss the steady motion of a Newtonian viscous incompressible fluid between two parallel plates when

   (i) the plates are at rest and there is an external pressure gradient;

   (ii) one plate is moving in relation to the other and there is no external constant pressure gradient;

   (iii) one plate is moving in relation to the other and there is also an external constant pressure gradient.

3. For steady motion between coaxial circular cylinders, show that

$$v = V\frac{\ln(r/b)}{\ln(a/b)} - \frac{\rho}{4\pi}\left[r^2 - \frac{b^2\ln(r/a) - a^2\ln(r/b)}{\ln(b/a)}\right],$$

   where the inner cylinder moves with velocity V and the outer cylinder is at rest. Show also that

$$Q = \pi V\left[\frac{\frac{1}{2}(b^2 - a^2)}{\ln(b/a)} - a^2\right] + \frac{\pi\rho}{8\mu}\left[b^4 - a^4 - \frac{(b^2 - a^2)}{\ln(b/a)}\right]. \tag{9.6.8}$$

Unit 10

Course Structure

- Basic Concepts about Blood
- Cardiovascular System and Blood Flow
- Special Characteristics of Blood Flow
- Structure, Function and Mechanical properties of Blood Vessels

10.1 Basic Concepts about Blood, Cardiovascular System and Blood Flow

10.1.1 Constitution of Blood

Blood consists of a *suspension of cells* in an aqueous solution called *plasma* which is composed of about 90 per cent water and 7 per cent protein. There are about 5×10^9 cells in a millilitre (1 cc) of healthy human blood, of which about 95 per cent are *red cells* or *erythrocytes* whose main function is to transport oxygen from the lungs to all the cells of the body and the removal of carbon-dioxide formed by metabolic processes in the body to the lungs. About 45 per cent of the blood volume in an average man is occupied by red c ells. This fraction is known as the *hematocrit*. Of the remaining, *white cells* or *leucocytes* constitute about one-sixth or 1 per cent of the total, and these play a role in the resistance of the body to infection; *platelets* form 5 per cent of the total, and they perform a function related to blood clotting.

10.1.2 Viscosity of Blood

Blood is neither homogeneous nor Newtonian. Plasma in isolation may be considered Newtonian with a viscosity of about 1.2 times that of water. For whole blood, we can measure effective viscosity, and this found to depend on shear rate. The constitutive equations proposed for whole blood are as follows:

(i) $\tau = \mu e^n$ (power law equation). This is found to hold good for strain rates between 5 and 200 sec^{-1}, with n having a value between 0.68 and 0.80.

(ii) $\tau = \mu e^n + \tau_0$ $(\tau \geq \tau_0)$ (Herschel-Bulkley equation).

(iii) $\tau^{1/2} = \mu^{1/2}e^{1/2} + \tau_0^{1/2}$ (Casson equation). This holds for strain rated between 0 and 100000 sec^{-1}.

The yield stress arises because, at low shear stress, red cells form aggregates in the form of rouleaux which are stacks of red cells in the shape of a roll of coins (see Fig. 10.1). At some finite stress, which is usually small (of the order of 0.005 dyne/cm^2), the aggregate is disrupted and blood begins to flow.

For hematocrits exceeding 5.8 per cent, it has been found that the yield stress is given by

$$\tau_0^{1/2} = A(H - H_m)/100, \tag{10.1.1}$$

where $A = (0.008 \pm 0.002 \text{ dyne/cm}^2)^{1/3}$, H is the normal hematocrit, and H_m is the hematocrit below which there is no yield stress. Taking H as 45 per cent and H_m as 5 per cent, the yield stress of normal human blood should be between 0.01 and 0.06 dyne/cm^2.

Not only τ_0, but also τ and effective viscosity, depend significantly on the hematocrit. The effective viscosity

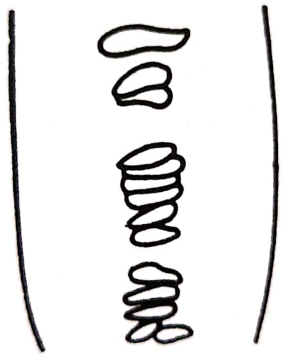


Figure 10.1: Rouleaux formation.

is also apparently found to depend on capillary radius when measurements are made in capillaries of diameters less than 300 μm. This apparent dependence of viscosity on capillary radius is known as *Fahraeus-Lindqvist effect.* We shall explain this effect which is based on the hypothesis of a two layer flow (a plasma layer and a core layer) with different viscosities.

10.1.3 Cardiovascular System

The cardovascular system consists of the following:

(i) The *heart* (which acts as a pump, whose elastic muscular walls contract rhythmically, making possible the pulsatile flow of blood through the vascular system)

(ii) The *distributory system* (comprising arteries and arterioles for sending blood to the various organs of the body)

(iii) The *diffusing system* (made up of fine capillaries which are in contact with the cells of the body)

(iv) The *collecting system of veins* (which collects blood depleted of oxygen and full of products of metabolic processes of the system).

The organs which supplement the function of the cardiovascular system are (i) the lungs which provide a region of inter-phase transfer of O_2 to the blood and removal of CO_2 from it, and (ii) the kidney, liver, and spleen, which help in maintaining the chemical quality of blood under normal conditions and under conditions of extreme stress.

Deoxygenated blood enters the *right atrium* (RA) from where it goes to the *right ventricle* (RV), as shown in Fig. 10.2. When the heart contracts, the *tricuspid valve* between the RA and RV closes and blood is pushed out to the lung through the *pulmonary artery* (PA) which branches to the right and left lungs where CO_2 is removed and blood is oxygenated. The blood returns from the lungs through the *pulmonary vein* (PV) to *left atrium* (LA) and then it goes to the *left ventricle* (LV) and from there, due to contraction of the heart, it enters the aorta from which it travels to other arteries and the rest of the vascular system.

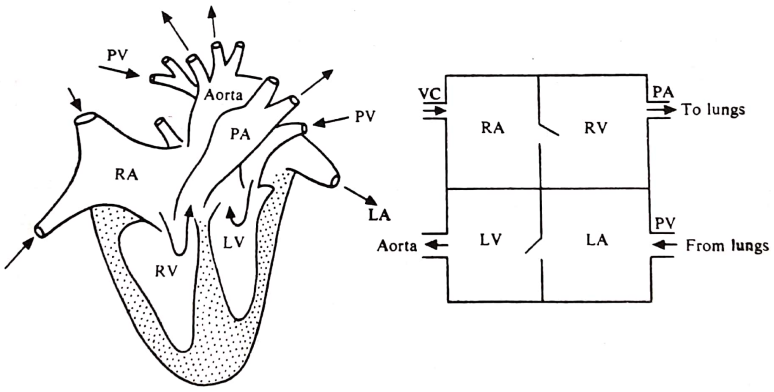


Figure 10.2: The heart.

10.1.4 Special Characteristics of Blood Flow

Blood flow problems are more complicated than the problems of fluid flows in engineering situations for the following reasons:

(i) Unusually high Reynolds number of flows. The flows remain laminar at Reynolds numbers as high as 5000 - 10000. This causes the entry length (which is proportional to the Reynolds number) to be so large that in most cases the fully developed flow is never reached since tube branching start before this stage is attained.

(ii) Unusual curvature of blood vessels. In some cases, this leads to secondary flows and these become more marked at high Reynolds numbers in some of the tubes.

(iii) Unusually large number of branches. Bifurcation takes place 20 - 30 times, leading to millions of blood vessels.

(iv) Unusual distensible properties of containing vessels. These properties arise from the fact that the vessel walls are formed of different substances such as elastin, collagen, and smooth mussels, with entirely different properties.

(v) Unusual fluid properties of blood. These properties are due to the fact that blood is a suspension of millions of cells of different shapes in plasma and these cells can deform when passing through vessels of diameter smaller than their own.

(vi) Unusual pulsatility of flows. This arises from the rhythmic action of the heart.

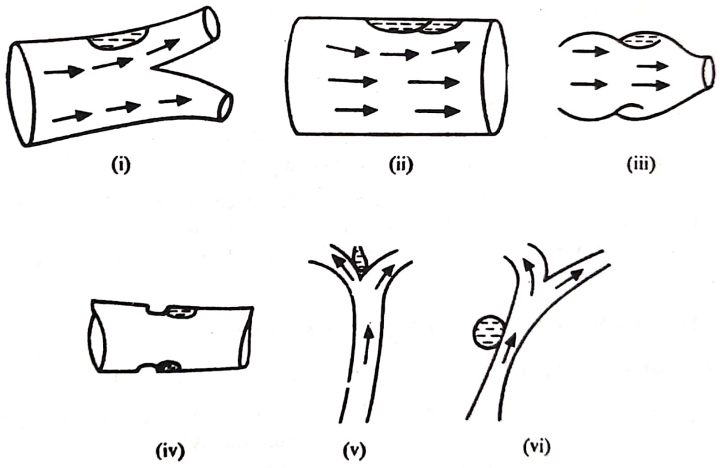


Figure 10.3: Examples of separation of flows in blood vessels.

There is also an unusual separation of flows, leading to increased resistance to flow and undesirable effects, e.g., hardening of arteries. The separation occurs due to various reasons, some of which are as follows (see Fig. **??**):

(i) Bifurcation of blood vessel

(ii) Atheroma of blood vessels or fatty degeneration of the inner walls of the blood vessel

(iii) Stenosis of heart valve or narrowing of the heart valve when the valve is fully open

(iii) Stenosis of blood vessels or narrowing of blood vessels

(iv) Secular aneurysm or a sac-like permanent abnormal blood-filled dilatation of blood vessel, resulting from a disease of the vessel wall

(v) Aortic aneurysm of abnormal blood-filled disatation of the aortic vessel.

10.1.5 Structure and function of Blood Vessels

Blood vessels are well-arranged sophisticated network of branching tubes or pipes conveying blood to the all parts of the body. There are several types of blood vessels, namely aorta, arteries, arterioles, veins, venues, capillaries etc. The arteries ar those blood vessels which carring away from the heart. The blood vessels is composed of three layers.

(i) The innermost layer called *Tunika-Intiama*, consist of thin layer of endothelial cells,

(ii) The middle layer called *Tunika-Median* consists of plain muscles and a network of elastic fibres, and

(iii) The outer most layer, called *Tunika-Adventesia*, is made up of fibrous tissues and elastic tissue.

Veins are the blood vessels which carrying blood to the heart. The venous cross-sectional area at any point is larger than of arteries and the velocity of blood is considerably lower when the arteries break up into minute vessels, they are turned to capillaries.

10.1.6 Principal of Blood Vessels

The principal constitutes of blood vessels are collagen, smooth muscles and elastin.

Collagen: It is the most important structure element of animal. There is a high amount of collagen present in bone materials. Collagen is relatively inextensible fibrous protein. The fibres can be identified by light or electron microscope.

Elastin: Unlike collagen elastin is an extensible fibrous protein present in large amount in skin, blood vessels, lung etc. The elastic behaviour of this structure is solely due to the presence of elastin,. The fact that elastin never appears without collagen, leads us to think that there must be resembles in structure of both.

Smooth muscles: Muscles consist of many fibres held together by connective tissues. Their structure and function varying widely in different organ and animal. One of the basic structure they are divided into smooth and straight muscles.

10.1.7 Mechanical Properties of Blood Vessels

In view of the diverse elastic properties of the components of the arterial wall, a number of theoretical and experimental investigation in the relevant field have established that vascular wall are non-homogeneous, anisotropy, incompressible and visco-elastic.

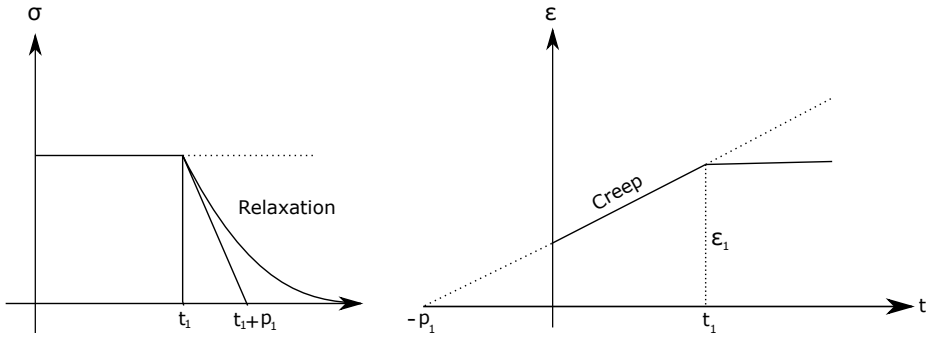
Inhomogeneity: Usually the wall of blood vessels are inhomogeneous. But experimental investigations showed that the outermost layer, adventesia has a very lose network and merges externally with the surrounding tissues. The inner most layer intima, is very tin and can be easily neglected. The remaining layer, the media, is considered homogeneous containing a matrix of smooth muscles elastic and collagen.

Compressibility: A material is said to be compressible if it changes its volume when it subjected to stress. It is said to be incompressible if the change of the volume is ignorable. The experimental studied showed that there is 20-40% change in volume and hence, for practical purpose the compressibility of vascular tissue can be considerably very small.

Anisotropy: Healthy arteries are highly deformable comfit structures and show a non-linear stress strain response with a typical stiffening effect at high pressure. This stiffening effect, common to all biological tissues is based on the recruitment of embedded wavy collagen fibrils which leads to the characteristics of anisotropic behaviour of artery.

Visco-elasticity: For a perfectly elastic body, there must be a single valued relationship between the applied strain and resulting stress. But when artery is subject to a cyclically varying strain the stress response exhibits a hysteresis loop called it cycle. The rate of decreases is very rapid in the beginning, but a steady state is observed after a numbers of cycles.

Moreover, two main characteristic of visco-elastic martial as for example creep and stress relaxation wave also observed in vascular tissue.



In the first stage, ϵ increases under the constant stress. This phenomena is called creep. In the second stage, the stress decrease under constant strain, i.e., the material relaxes. This phenomena is called stress relaxation.

Unit 11

Course Structure

- Steady non-Newtonian fluid flow in circular tubes
- Flow in Power-Law fluid in circular tubes
- Flow in Herschel-Bulkley fluid in circular tubes
- Flow in Casson fluid in circular tubes

11.1 Steady Non-Newtonian Fluid Flow in Circular Tubes

11.1.1 Basic Equations for Fluid Flow

We consider the laminar flow of a non-Newtonian in a circular tube under a constant pressure gradient. Let the control volume be bounded by two coaxial cylinders of raddi r and $r + dr$ and let it be of unit length, as shown in Fig. 11.1.

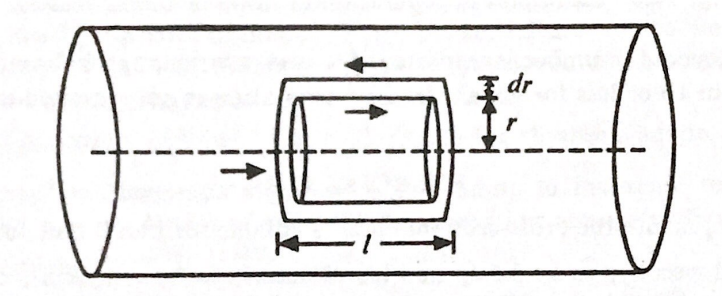


Figure 11.1: Forces on control volume

Due to the pressure gradient, there is a forward force $P \times 2\pi[(r + r\ dr) - r] = P \times 2\pi r\ dr$ on it. Let the stress be $\tau(r)$ at a distance r from the axis. Then the force on the inner cylindrical surface is $2\pi r\tau$, and the

force on the outer cylindrical surface is

$$\begin{aligned} 2\pi(r+dr)\tau(r+dr) &= 2\pi(r+dr)\left[\tau(r)+\frac{d\tau(r)}{dr}dr\right] \\ &= 2\pi\left[r\tau(r)+\tau(r)\,dr+r\frac{d\tau(r)}{dr}dr\right] \\ &= 2\pi\left[[r\tau(r)]+\frac{d}{dr}[r\tau(r)]\,dr\right] \end{aligned} \tag{11.1.1}$$

Balancing the force in the axial direction in the control volume, we get

$$\begin{aligned} 2\pi\frac{d}{dr}[r\tau(r)] &= 2\pi rP \\ \Rightarrow \frac{d}{dr}[r\tau(r)] &= rP. \end{aligned} \tag{11.1.2}$$

Integrating (11.1.2), we obtain

$$\begin{aligned} r\tau(r) &= \frac{1}{2}r^2P+A \\ \Rightarrow \tau(r) &= \frac{1}{2}P\left[r+\frac{D}{r}\right] \quad \text{where} \quad D=2A. \end{aligned} \tag{11.1.3}$$

Since the stress $\tau(r)$ is finite on the axis (i.e. at $r=0$), we have

$$A=0, \qquad D=0, \qquad \tau=\frac{r}{2}P. \tag{11.1.4}$$

The velocity v is parallel to the axis which is also a function of r only and is expected to decrease from a maximum on the axis to zero on the surface so that the only non-zero component of strain rate is

$$e(r)=-\frac{dv}{dr}. \tag{11.1.5}$$

For a non-Newtonian fluid,

$$\tau=f(e) \tag{11.1.6}$$

so that from Eq.(11.1.3), we have

$$\frac{r}{2}P=f\left(-\frac{dv}{dr}\right). \tag{11.1.7}$$

Integrating (11.1.7) subject to the condition that $v=0$ when $r=R$, we get v as a function of r. Then we can obtain the flux Q by using

$$Q=\int_0^R 2\pi rv\,dr. \tag{11.1.8}$$

Integrating the right-hand side of (11.1.8) by parts, we get

$$Q=2\pi\left[\left(\frac{1}{2}r^2v\right)_0^R-\int_0^R\frac{1}{2}r^2\frac{dv}{dr}dr\right]. \tag{11.1.9}$$

Since $v=0$ at $r=R$, we have

$$Q=\pi\int_0^R r^2e(r)\,dr. \tag{11.1.10}$$

11.2 Flow of Power-Law Fluid in Circular Tube

Here $\tau = \mu e^n$, Eq.(11.1.7) gives

$$\frac{dv}{dr} = -\left(\frac{1}{2}\frac{P}{\mu}r\right)^{1/n} \tag{11.2.1}$$

Integrating (11.2.1), we obtain

$$v = \left(\frac{P}{2\mu}\right)^{1/n} \frac{n}{n+1}\left[R^{\frac{1}{n}+1} - r^{\frac{1}{n}+1}\right] \tag{11.2.2}$$

Also,

$$Q = \int_0^R 2\pi r\, v\, dr = \left(\frac{1}{2}\frac{P}{\mu}\right)^{1/n} \frac{n\pi}{3n+1} R^{\frac{1}{n}+3}. \tag{11.2.3}$$

11.3 Flow of Herschel-Bulkley Fluid in Circular Tube

In this case, we have $e = 0$ when $\tau \le \tau_0$, and there is a core region which flows as a plug (see Fig. 11.2). Let the radius of the plug region be r_p. At the surface of the plug, the stress is τ_0 so that, considering the forces on the plug, we get

$$\begin{aligned} &P \times \pi r_p^2 = \tau_0 \times 2\pi r_p \\ &\Rightarrow r_p = 2\tau_0/P \end{aligned} \tag{11.3.1}$$

In the non-core region, $\tau \ge \tau_0$, and

$$\tau = \mu e^n + \tau_0 \tag{11.3.2}$$

so that (11.1.7) gives

$$\left(\frac{\frac{r}{2}P - \tau_0}{\mu}\right)^{1/n} = e = -\frac{dv}{dr} \tag{11.3.3}$$

or, on using (11.3.1), we get

$$\frac{dv}{dr} = -\left(\frac{1}{2}\frac{P}{\mu}\right)^{1/n} (r - r_p)^{1/n}. \tag{11.3.4}$$

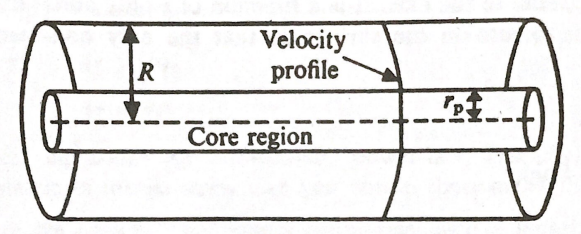


Figure 11.2: Plug flow

Integrating (11.3.4), we obtain

$$v = \frac{n}{n+1}\left(\frac{P}{2\mu}\right)^{1/n} [(R - r_0)^{\frac{1}{n}+1} - (r - r_0)^{\frac{1}{n}+1}]. \tag{11.3.5}$$

If $r = r_p$, then $v = v_p$ (the velocity of the plug flow) so that

$$v_p = \frac{n}{n+1}\left(\frac{P}{2\mu}\right)^{1/n}(R - r_p)^{\frac{1}{n}+1}. \tag{11.3.6}$$

Eq. (11.3.1) determines the radius of the plug, and then using this value of the plug, (11.3.6) determines the velocity of the plug and (11.3.5) determines the velocity in the non-core region. Also,

$$\begin{aligned} Q &= \pi r_p^2 v_p + \int_{r_p}^{R} 2\pi r v \, dr \\ &= \pi r_p^2 \frac{n}{n+1}\left(\frac{P}{2\nu}\right)^{1/n}(R - r_p)^{\frac{1}{n}+1} + \frac{n}{n+1}\left(\frac{P}{2\mu}\right)^{1/n} 2\pi \left[\frac{1}{2}(R - r_p)^{\frac{1}{n}+2}(R + r_p)\right. \\ &\quad \left. - \frac{(R - r_p)^{\frac{1}{n}+3}}{\frac{1}{n}+3} - r_p \frac{(R - r_p)^{\frac{1}{n}+2}}{\frac{1}{n}+3}\right] \end{aligned} \tag{11.3.7}$$

$$\begin{aligned} &= \pi \frac{n}{n+1}\left(\frac{P}{2\mu}\right)^{1/n} R^{\frac{1}{n}+3}\left[c_p^2(1 - c_p)^{\frac{1}{n}+1} + (1 + c_p)(1 - c_p)^{\frac{1}{n}+2}\right. \\ &\quad \left. - \frac{2}{\frac{1}{n}+3}(1 - c_p)^{\frac{1}{n}+3} - \frac{2c_p}{\frac{1}{n}+2}(1 - c_p)^{\frac{1}{n}+2}\right] \end{aligned} \tag{11.3.8}$$

$$= \pi \frac{n}{3n+1}\left(\frac{P}{2\mu}\right)^{1/n} R^{\frac{1}{n}+3} f(c_p) \quad \text{(say)}, \tag{11.3.9}$$

where $c_p = \dfrac{r_p}{R} = \dfrac{2\tau_0}{PR}$, $\quad f(0) = 1$.

If Q_0 denotes the flux when there is no plug flow (i.e., when $\tau_0, c_p = 0$), we get

$$\frac{Q}{Q_0} = f(c_p) = f\left(\frac{r_p}{R}\right) = f\left(\frac{2\tau_0}{PR}\right). \tag{11.3.10}$$

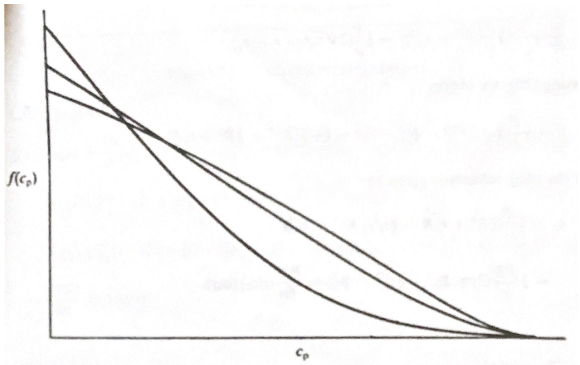


Figure 11.3: Variation of flux with τ_0

This gives the relative change in Q with τ_0. Fig. 11.3 illustrates the variation of $f(c_p)$ with c_p for various values of n. The figure shows that:

(i) As τ_0 increases (μ and n remaining the same), the flux decreases rapidly and approaches zero as c_p approaches unity.

(ii) If $n < 1$, the curve is always concave upwards; when $n = 1$, the curve is always a straight line in the beginning and becomes concave upwards; and when $n > 1$, the curve is convex in the beginning and becomes concave near $c_p = 1$, and, therefore, it has a pint of inflexion.

(iii) If τ_0 and μ are constant, the decline in Q is more when $n < 1$ and less when $n > 1$. If we put $n = 1$ in (11.3.8) and (11.3.9), we get the results for the special case of a Bingham plastic.

(iv) If we put $\tau_0 = 0, r_p = 0$ in (11.3.8), we get results for the special case of a power-law fluid. Further, if we put $n = 1$, we obtain results for Poiseuille flow.

11.4 Flow of Casson Fluid in Circular Tube

Here

$$\tau^{\frac{1}{2}} = \mu^{\frac{1}{2}} e^{\frac{1}{2}} + \tau_0^{\frac{1}{2}} \quad (\tau \geq \tau_0) \tag{11.4.1}$$

so that for the non-core region, (11.3.4) gives

$$-\frac{dv}{dr} = e = \frac{1}{\mu^{1/2}} \left[\left(\frac{1}{2} rP \right)^{1/2} - \left(\frac{1}{2} r_p P \right)^{1/2} \right] \tag{11.4.2}$$

or

$$\frac{dv}{dr} = -\frac{1}{2} \frac{P}{\mu} \left[r^{1/2} - r_p^{1/2} \right]^2 = \frac{1}{2} \frac{P}{\mu} \left[2\sqrt{r_p r} - r - r_p \right]. \tag{11.4.3}$$

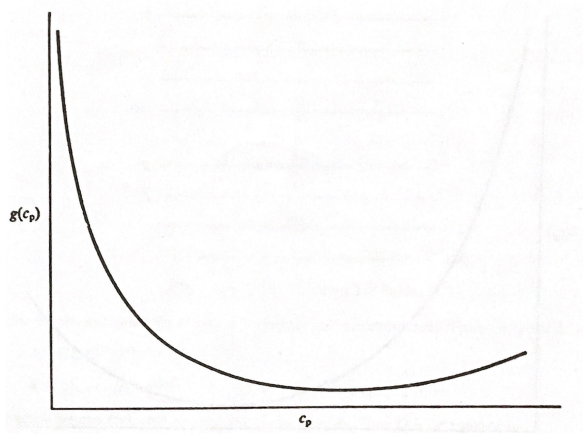


Figure 11.4: Variation of $g(c_p)$ with c_p

Integrating (11.4.3), we obtain

$$v = \frac{1}{2} \frac{P}{\mu} \left[\frac{4}{3} \sqrt{r_p} r^{3/2} - \frac{1}{2} r^2 - r_p r - \frac{4}{3} \sqrt{r_p} R^{3/2} + \frac{1}{2} R^2 + r_p R \right] \tag{11.4.4}$$

so that the plug velocity is given by

$$\begin{aligned} v_p &= \frac{1}{2}\frac{P}{\mu}\left[\frac{1}{2}R^2 + r_pR - \frac{4}{3}\sqrt{r_p}R^{3/2} - \frac{1}{6}r_p^2\right] \\ &= \frac{1}{4}\frac{PR^2}{\mu}\left[1 + 2c_p - \frac{8}{3}c_p^{1/2} - \frac{1}{6}c_p^2\right] \\ &= \frac{PR^2}{4\mu}g(c_p) \quad \text{(say)}. \end{aligned} \tag{11.4.5}$$

Thus

$$\frac{v_p}{(v_p)_0} = g(c_p). \tag{11.4.6}$$

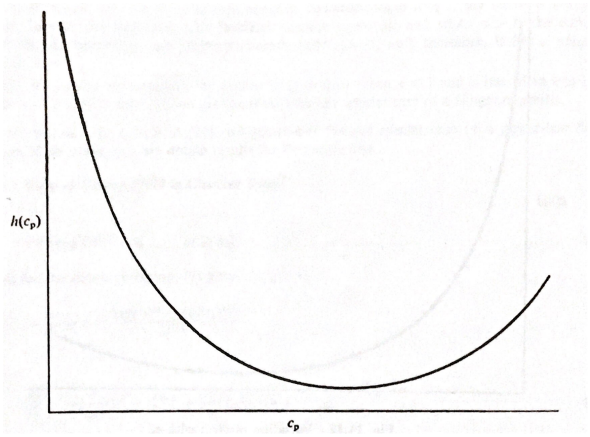


Figure 11.5: Variation of $h(c_p)$ with c_p

Figure 11.4 shows the variation of $g(c_p)$ with c_p. This shows that, as τ_0 increases (μ remaining the same), the plug velocity or the maximum velocity of flow decreases rapidly till c_p reaches 0.6 when the velocity is reduced to about 6 per cent of the value and then it rises sightly. For blood, small changes in τ_0 lead to significant changes in maximum velocity.

The flux Q is given by

$$\begin{aligned} Q &= \pi v_p r_p^2 + \frac{P\pi}{\mu}\left[\frac{8}{21}\sqrt{r_p}(R^{7/2} - r_p^{7/2}) - \frac{1}{8}(R^4 - r_p^4) - \frac{1}{3}r_p(R^3 - r_p^3)\right. \\ &\quad \left. -\frac{2}{3}\sqrt{r_p}R^{3/2}(R^2 - r_p^2) + \frac{1}{4}R^2(R^2 - r_p^2) + \frac{1}{2}r_pR(R^2 - r_p^2)\right] \\ &= \frac{\pi}{4}\frac{PR^4}{\mu}c_p^2 g(c_p) + \frac{\pi PR^4}{\mu}\left[\frac{8}{21}\sqrt{c_p}(1 - c_p^{7/2}) - \frac{1}{8}(1 - c_p^4) - \frac{1}{3}c_p(1 - c_p^3)\right. \\ &\quad \left. -\frac{2}{3}\sqrt{c_p}(1 - c_p^2) + \frac{1}{4}(1 - c_p^2) + \frac{1}{2}c_p(1 - c_p^2)\right] \\ &= \frac{\pi PR^4}{8\mu}h(c_p), \quad \text{(say)} \end{aligned} \tag{11.4.7}$$

so that

$$\frac{Q}{Q_0} = h(c_p). \tag{11.4.8}$$

Figure 11.5 gives the graph of $h(c_p)$ against c_p. Its shows that, as τ_0 increases (μ remaining the same), the flux decreases rapidly till $c_p = 0.6$ and till it has fallen to about 5 per cent of Q_0 and then it rises again. For blood, small changes in τ_0 can make significant changes in Q. The Casson fluid flows in the tube takes place only if $r_p < R$, i.e., if

$$2\tau_0 < PR. \tag{11.4.9}$$

Unit 12

Course Structure

- Fahraeus-Lindqvist Effect
- Pulsatile Flow in Circular Rigid Tube
- Blood Flow through Artery with Mild Stenosis

12.1 Newtonian Fluid Models

12.1.1 Fahraeus-Lindqvist Effect

Fahraeus-Lindqvist effect is an effect where the viscosity of a fluid, in particular blood, changes with the diameter of the tube, it travels through. More precisely, there is a decrease of viscosity as tubes diameter decreases (only if the vessel diameter is between 10 to 300 micrometers).

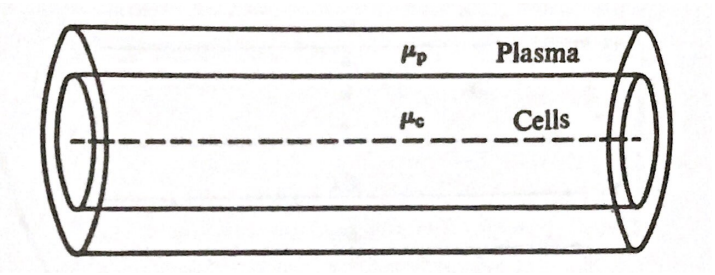


Figure 12.1: Two layer flow

In arteries, blood flows in two layers, a plasma layer near the walls consisting of only the plasma and almost no cells and a core layer consisting of red cells in plasma (see Fig. 12.1). If μ_p and μ_c are the viscosities of the two fluids, which are assumed Newtonian, we get

$$v_p = \frac{P}{4\mu_p}(R^2 - r^2), \qquad R - \delta \leq r \leq R, \tag{12.1.1}$$

$$v_c = \frac{P}{4\mu_c}(R^2 - r^2) + \frac{P}{\mu_c}\left[R^2 - (R-\delta)^2\right]\left(\frac{\mu_c}{\mu_p} - 1\right), \quad 0 \le r \le R - \delta. \tag{12.1.2}$$

Thus the velocity in the plasma layer is the same as it would be when the whole tube is filled with plasma, but the velocity in the core layer is more than it would be when the whole tube is filled with the core fluid. This is what is expected.

Now,

$$\begin{aligned} Q &= \int_0^{R-\delta} 2\pi r v_c \, dr + \int_{R-\delta}^{R} 2\pi r v_p \, dr \\ &= \frac{\pi P R^4}{8\mu_p}\left[1 - \left(1 - \frac{\delta}{R}\right)^4 \left(1 - \frac{\mu_p}{\mu_c}\right)\right]. \end{aligned} \tag{12.1.3}$$

If the whole tube were filled with a single Newtonian fluid with viscosity coefficient μ, we would have

$$Q = \frac{\pi P R^4}{8\mu}. \tag{12.1.4}$$

The two fluxes would be the same if

$$\mu = \mu_p \left[1 - \left(1 - \frac{\delta}{R}\right)^4 \left(1 - \frac{\mu_p}{\mu_c}\right)\right]^{-1}, \tag{12.1.5}$$

where μ is the *effective viscosity* of the two fluids taken together. From (12.1.5), it can be seen that the effective viscosity depends on R. In practice, $\dfrac{\delta}{R} << 1$, and hence (12.1.5) gives

$$\mu = \mu_p \left[1 - \frac{4\delta}{R}\left(\frac{\mu_c}{\mu_p} - 1\right)\right]. \tag{12.1.6}$$

We find that, as R decreases, μ decreases. This explains the Fahraeus-Lindqvist effect. Here it has been assumed that δ is independent of R.

Pulsatile Flow in Circular Rigid Tube

We consider axially-symmetric flow in a rigid circular tube of radius R for which

$$v_r = 0, \quad v_\theta = 0, \quad v_z = v(r, z, t), \quad p = p(r, z, t) \tag{12.1.7}$$

so that the equation of continuity and the equation of motion are given by

$$\frac{\partial v}{\partial z} = 0, \qquad \frac{\partial p}{\partial r} = 0, \tag{12.1.8}$$

$$\frac{\partial v}{\partial t} + v\frac{\partial v}{\partial z} = -\frac{1}{\rho}\frac{\partial p}{\partial z} + \nu\left(\frac{\partial^2 v}{\partial r^2} + \frac{1}{r}\frac{\partial v}{\partial r} + \frac{\partial^2 v}{\partial z^2}\right). \tag{12.1.9}$$

By using (12.1.8) and (12.1.9) becomes

$$\frac{\partial v}{\partial t} = -\frac{1}{\rho}\frac{\partial p}{\partial z} + \frac{\nu}{r}\frac{\partial}{\partial r}\left(r\frac{\partial v}{\partial r}\right). \tag{12.1.10}$$

From (12.1.7) and (12.1.8), v is a function of r and t only and p is a function of z and t only. From (12.1.10), $\partial p/\partial z$ is a function of t only. Thus for a pulsatile sinusoidal flow, we take

$$\frac{\partial p}{\partial z} = -Pe^{i\omega t}, \quad (i = \sqrt{-1}), \tag{12.1.11}$$

$$v(r,t) = V(r)e^{i\omega t}. \tag{12.1.12}$$

This means that the real part gives the velocity for pressure gradient $P\cos(\omega t)$ and the imaginary part gives the velocity for the pressure gradient $P\sin\omega t$.

From (12.1.10)-(12.1.12),

$$i\omega V\rho = P + \mu\left(\frac{d^2V}{dr^2} + \frac{1}{r}\frac{dV}{dr}\right) \tag{12.1.13}$$

$$\Rightarrow \frac{d^2V}{dr^2} + \frac{1}{r}\frac{dV}{dr} - \frac{i\omega}{\mu}\rho V = -\frac{P}{\mu}. \tag{12.1.14}$$

Now the general solution of the equation

$$\frac{d^2y}{dx^2} + \frac{1}{x}\frac{dy}{dx} - k^2y = 0 \tag{12.1.15}$$

is

$$y = AJ_0(ikx) + BY_0(ikx), \tag{12.1.16}$$

where both $J_0(x)$ and $Y_0(x)$ are Bessel functions of zero order and are of the first and second kind, respectively.

Thus the solution of (12.1.14) is

$$V = AJ_0\left[\left(i^{\frac{3}{2}}\sqrt{\frac{\omega\rho}{\mu}}\right)r\right] + BY_0\left[\left(i^{\frac{3}{2}}\sqrt{\frac{\omega\rho}{\mu}}\right)r\right] + \frac{P}{\omega\rho i}. \tag{12.1.17}$$

Since v and V have to be finite on the axis (i.e., at $r = 0$) and $Y_0(0)$ is not finite, B has to be zero. Also, because of the no-slip condition $v(r) = 0$ when $r = R$, we have

$$AJ_0\left[\left(i^{\frac{3}{2}}\sqrt{\frac{\omega\rho}{\mu}}\right)R\right] + \frac{P}{\omega\rho i} = 0, \quad B = 0. \tag{12.1.18}$$

Let

$$\alpha^2 = \frac{\omega\rho}{\mu}R^2 = \frac{\omega R^2}{\nu} \tag{12.1.19}$$

so that

$$A = \frac{P}{\omega\rho}i\frac{1}{J_0(i^{3/2}\alpha}, \tag{12.1.20}$$

$$V(r) = -\frac{P}{\omega\rho}i\left[1 - \frac{J_0(i^{3/2}\alpha s)}{J_0(i^{3/2}\alpha}\right], \tag{12.1.21}$$

where

$$s = \frac{r}{R}. \tag{12.1.22}$$

Finally, we get

$$v(r,t) = -\frac{PR^2}{\mu\alpha^2} i \left[1 - \frac{J_0(i^{3/2}\alpha s)}{J_0(i^{3/2}\alpha)}\right] e^{i\omega t}. \tag{12.1.23}$$

The volumetric flow rate Q is given by

$$\begin{aligned} Q &= \int_0^R v 2\pi r \, dr \\ &= 2\phi R^2 \int_0^1 vs \, ds \\ &= -\frac{2\pi PR^4}{\mu\alpha^2} i e^{i\omega t} \left[\int_0^1 s \, ds - \frac{1}{J_0(i^{3/2}\alpha)} \int_0^1 J_0(i^{3/2}\alpha s) \, s \, ds\right] \\ &= -\frac{\pi PR^4}{\mu\alpha^2} i e^{i\omega t} \left[1 - \frac{2}{J_0(i^{3/2}\alpha)} \int_0^{i^{3/2}\alpha} \left(\frac{xJ_0(x)}{i^3\alpha^2}\right) dx\right]. \end{aligned} \tag{12.1.24}$$

But

$$\int xJ_0(x) \, dx = xJ_1(x) \tag{12.1.25}$$

so that

$$\begin{aligned} Q &= -\frac{\pi PR^4}{\mu\alpha^2} i e^{i\omega t} \left[1 - \frac{2i}{J_0(i^{3/2}\alpha)} \frac{i^{3/2}\alpha J_1(i^{3/2}\alpha)}{\alpha^2}\right] \\ &= -\frac{\pi R^4}{\mu\alpha^2} iP \left[1 - \frac{2J_1(i^{3/2}\alpha)}{i^{3/2}\alpha J_0(i^{3/2}\alpha)}\right] e^{i\omega t} \\ &= \frac{\pi R^4 P}{\mu\alpha^2 i} X(\alpha) e^{i\omega t} \quad \text{(say)}. \end{aligned} \tag{12.1.26}$$

Now the series expansion for $J_0(x)$ and $J_1(x)$ are given by

$$J_0(x) = 1 - \frac{1}{2}x^2 + \ldots, \tag{12.1.27}$$

$$J_1(x) = \frac{x}{2} - \frac{(x/2)^3}{1^2 \cdot 2} + \frac{(x/2)^5}{1^2 \cdot 2^2 \cdot 3} - \ldots \tag{12.1.28}$$

For small values of α,

$$X(\alpha) = 1 - \frac{2\left[\frac{i^{3/2}\alpha}{2} - \left(\frac{i^{3/2}\alpha}{2/2}\right)^3 + \ldots\right]}{i^{3/2}\left[1 - \left(\frac{i^{3/2}\alpha}{2}\right)^2 + \ldots\right]} = 1 - \frac{1 - \frac{i^3\alpha^2}{8} + \ldots}{1 - \frac{i^3\alpha^2}{4} + \ldots} = \frac{i\alpha^2}{8} + O(\alpha^4) \tag{12.1.29}$$

From (12.1.24) and (12.1.29), we have

$$Q = \left[\frac{\pi R^4 P}{8} + O(\alpha^2)\right] e^{i\omega t}. \tag{12.1.30}$$

From (12.1.19) as $\alpha \to 0,\ \omega \to 0$ and then from (12.1.30), $Q \to Q_0$, where

$$Q_0 = \frac{\pi R^4 P}{8\mu} e^{i\omega t}, \quad |Q_0| = \frac{\pi R^4 P}{8\mu}, \tag{12.1.31}$$

and $|Q_0|$ is the volumetric flow rate for a constant pressure gradient and is the same as for Poiseuille law for steady flow. If

$$X(\alpha) = X_1(\alpha) + iX_2(\alpha), \tag{12.1.32}$$

(12.1.24) gives

$$Q = \frac{\pi R^4}{\mu\alpha^2}\left[\Big\{X_2(\alpha)\cos\omega t + X_1(\alpha)\sin\omega t\Big\} - i\Big\{X_1(\alpha)\cos(\omega t) - X_2(\alpha)\sin\omega t\Big\}\right] \tag{12.1.33}$$

The real part gives the flux when the pressure gradient is $P\cos\omega t$ and the imaginary part gives the flux when it is $P\sin\omega t$.

12.2 Blood Flow through Artery with Mild Stenosis

12.2.1 Effect of Stenosis

The term *stenosis* denotes the narrowing of the artery due to the development of arteriosclerotic plaques or other types of abnormal tissue development. As the growth projects into the lumen (cavity) of the artery, blood flow is obstructed. The obstruction may damage the internal cells of the wall and may lead to further growth of the stenosis. Thus there is a coupling between the growth of a stenosis and the flow of blood in the artery since each affects the other.

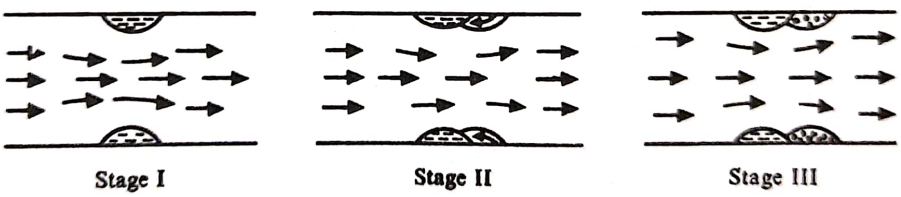


Figure 12.2: Three stages of stenosis growth.

The stenosis growth usually passes through three stages, as shown in Fig. 12.2. In stage I, there is no separation of flow and there is no back flow. In stage II, the flow is laminar, but separation occurs and there is back flow. In stage III, turbulence develops in a certain region of the down stream. We shall discuss here only Stage I, called *mild stenosis.*

The development of stenosis in artery can have serious consequences and can disrupt the normal functioning of the circulatory system. In particular, it may lead to

(i) increased resistance to flow, with possible severe reduction in blood flow;

(ii) increased danger of complete occlusion (obstruction);

(iii) abnormal cellular growth in the vicinity of the stenosis, which increases the intensity of the stenosis; and

(iv) tissue damage leading to post-stenosis dilatation.

12.2.2 Analysis of Mild Stenosis

We shall consider the steady flow of a Newtonian fluid past an axially-symmetric stenosis whose surface is given by

$$\frac{R}{R_0} = 1 - \frac{\delta}{2R_0}\left(1 + \cos\frac{\pi z}{z_0}\right), \tag{12.2.1}$$

where the notations are clear from Fig. XX. We shall assume further that

$$\frac{\delta}{R_0} << 1, \qquad \frac{R_0}{z_0} \approx 0(T), \qquad Re\frac{\delta}{z_0} << 1, \tag{12.2.2}$$

where Re is the Reynolds number of fluid flow. By carrying out an order of magnitude analysis on these basic

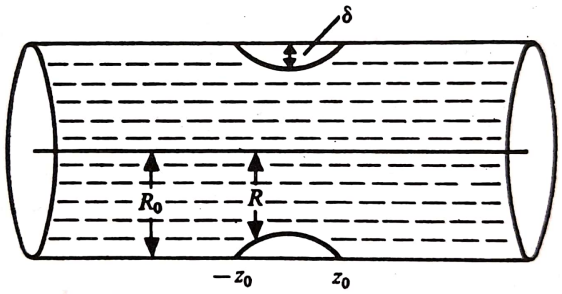


Figure 12.3: Mild stenosis.

equations of motion in cylindrical polar coordinates, it can be shown that the radial velocity can be neglected in relation to axial velocity v which is determined by

$$0 = -\frac{\partial p}{\partial z} + \mu\left(\frac{\partial^2 v}{\partial r^2} + \frac{1}{r}\frac{\partial v}{\partial r}\right), \tag{12.2.3}$$

$$0 = -\frac{\partial p}{\partial r}, \tag{12.2.4}$$

or

$$-P(z) = \frac{\mu}{r}\frac{\partial}{\partial r}\left(r\frac{\partial v}{\partial r}\right). \tag{12.2.5}$$

The no-slip condition on the stenosis surface gives

$$\begin{aligned} &v = 0 \quad \text{at} \quad r = R(z), \quad -z_0 \le z \le z_0, \\ &v = 0 \quad \text{at} \quad r = R_0, \qquad\quad |z| \ge z_0. \end{aligned} \tag{12.2.6}$$

Thus for a mild stenosis, the main difference from the usual Poiseuille flow is that the pressure gradient and axial velocity are functions of z also. However, for a stenosis in stage II or stage III, the radial velocity can be significant, and turbulence may have to be considered. Obviously then, the analysis is more complicated.

Integrating (12.2.5), we get

$$r\frac{\partial v}{\partial r} = -P(z)\frac{r^2}{2\mu} + A(z), \tag{12.2.7}$$

but $A(z) = 0$ since $\dfrac{\partial v}{\partial r} = 0$ on the axis. Integrating again and using (12.2.6), we get

$$v = -\frac{P(z)}{4\mu}\Big[r^2 - R^2(z)\Big]. \tag{12.2.8}$$

If Q is the flux through the tube, then

$$Q = \int_0^{R(z)} v 2\pi r \, dr = \frac{\pi P(z)}{8\mu} R^4(z). \tag{12.2.9}$$

Since Q is constant for all section of the tube, the pressure gradient varies inversely as the fourth power of the surface distance of the stenosis from the axis of the artery so that it (the pressure gradient) is minimum at the middle of the stenosis and is maximum at the ends.

Unit 13

Course Structure

- Peristaltic Flows in Tubes and Channel
- Peristaltic Motion in a Channel
- Long-wavelength Approximation
- Further discussion on Long-wavelength Approximation

13.1 Peristaltic Flows in Tubes and Channel

13.1.1 Peristaltic Flows in Biomechanics

Peristaltic flow is the motion generated in the fluid contained in a distensible tube when a progressive wave of area contraction and expansion travels along the wall of the tube. The elasticity of the tube wall does not directly enter into our calculations, but it affects the flow through the progressive wave travelling along its length. This wave determines the boundary conditions sice the no-slip condition has to be used now on a moving undulating wall surface.

Peristaltic motion is involved in

(i) expansion and contractions (or vasomotion) of small blood vessels,

(ii) celia transport through the ducts efferents of the male reproductive organ,

(iii) transport of spermatozoa in cervical canal,

(iv) transport of chyme in small intestines,

(v) functioning of ureter, and

(i) transport of bile.

The wide occurrence of peristaltic motion should not be surprising since it results physiologically from neuro-muscular properties of any tubular smooth muscle.

We now consider peristaltic motion in channels or tubes. The fluid involved may be non-Newtonian (e.g., power-law, viscoelastic, or micropolar fluid) or Newtonian, and the flow may take place in two layers (a core layer and a peripheral layer). The equations of motion in their complete generality do nt admit of simple solutions and we have to look for reasonable approximations. For this we first transform these equations in terms of dimensionless variables.

Peristaltic Motion in a Channel : Characteristic Dimensionless Parameters

We consider the flow of a homogeneous Newtonian fluid through a channel of width $2a$. Travelling sinusoidal waves are supposed on the elastic walls of the channel. Taking the x-axis along the centre line of the channel and the y-axis normal to it, the equations of the walls are given by

$$Y = \eta(X,T) = \pm a\left[1 + \epsilon\cos\left\{\frac{2\pi}{\lambda}(x - ct)\right\}\right] \tag{13.1.1}$$

where ϵ is the amplitude ratio, λ the wavelength, and c the phase velocity of the waves. Now using Eq. 9.10 of Unit 9, the stream function $\psi(X,Y)$ for the two-dimensional motion satisfies the equation

$$\nu\nabla^4\psi = \nabla^2\Psi_T + \Psi_Y\nabla^2\Psi_X - \Psi_X\nabla^2\Psi_Y, \tag{13.1.2}$$

where the velocity components are given by

$$U = \Psi_Y, \qquad V = -\Psi_X. \tag{13.1.3}$$

Assuming that the walls have only transverse displacements at all times, we get the boundary conditions as

$$U = 0, \quad V = \pm\frac{2\pi ac\epsilon}{\lambda}\sin\left\{\frac{2\pi}{\lambda}(X - cT)\right\} \quad \text{at} \quad Y = \pm\eta(X,T). \tag{13.1.4}$$

We now introduce the dimensionless variables and parameters

$$x = \frac{X}{\lambda}, \quad y = \frac{Y}{a}, \quad t = \frac{cT}{\lambda}, \quad \psi = \frac{\Psi}{ac}, \quad \delta = \frac{a}{\lambda}, \quad Re = \frac{ac}{\nu} \tag{13.1.5}$$

so that (13.1.2) becomes

$$\frac{1}{\delta\ Re}\left[\delta^2\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right]^2\psi = \left[\delta^2\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right]\psi_t + \psi_y\left[\delta^2\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right]\psi_x - \psi_x\left[\delta^2\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right]\psi_y. \tag{13.1.6}$$

The boundary conditions becomes

$$\psi_y = 0, \quad \psi_x = 2\pi\epsilon\sin(x - \epsilon). \tag{13.1.7}$$

Thus the basic partial differential equations and the boundary consdition together involve three dimensionless parameters:

(i) The Reynolds number, Re determined by the phase velocity, half the mean distance between the plates, and the kinematic viscosity. (This number is small if the distance between the walls is small or the phase velocity is small or the kinematic viscosity is large.)

(ii) The wave number δ which is small if the wavelength is large as compared to the distance between the walls.

(iii) The amplitude ratio ϵ which is small if the amplitude of the wave is small as compared to the distance between the walls.

In obtaining the equations for the stream funtion, the pressure gradient was eliminated. Hence there may arise a fourth dimensionless parameter, depending on the pressure gradient. Non-Newtonian fluids give rise to additional dimensionless parameters, depending on the parameters occurring in the constitutive equations of the fluids.

It is not possible to solve (13.1.2) for arbitrary values of δ, ϵ, Re and, therefore, this equation is solved under, among others, the following alternative sets of assumptions:

(i) $\epsilon << 1$, and Stoke's assumption of slow motion so that inertial terms can be neglected.

(ii) $\epsilon << 1,\ \delta << 1$.

(iii) $\delta << 1,\ Re << 1$, but ϵ is arbitrary

(iv) $\epsilon << 1,\ Re << 1$, but δ is arbitrary.

The initial flow may be taken as the Hagen-Poiseuille flow.

Long-wavelength Approximation to Peristaltic Flow in a Tube

Let the equation of the tube surface be given by

$$h(Z,t) = a\left[1 + \epsilon \sin\left\{\frac{2\pi}{\lambda}(Z - ct)\right\}\right], \tag{13.1.8}$$

where a is the undisturbed radius of the tube and ϵ the amplitude ratio, $a(1+\epsilon)$ and $a(1-\epsilon)$ are the maximum and minimum disturbed radii, and λ is the wave velocity and c the phase velocity (see Fig. 13.1). Under the

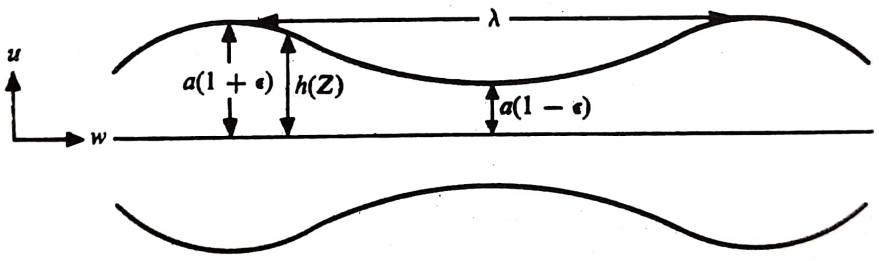


Figure 13.1: Tube geometry.

assumptions $\frac{a}{\lambda} << 1$ and $\frac{ac}{\nu} << 1$, we conduct an order of magnitude study of the various terms in the equation of continuity and equations of motion in cylindrical polar coordinates to find

$$\frac{\partial p}{\partial R} << \frac{\partial p}{\partial Z} \tag{13.1.9}$$

so that p is only weakly dependent on R and we can take

$$p = p(Z, t). \tag{13.1.10}$$

Now it is convenient to use the moving coordinate system (r, z) travelling with the wave so that

$$r = R, \qquad z = Z - ct. \tag{13.1.11}$$

In this system, p is a function of z only. The equations of continuity and motion reduce respectively to

$$\frac{\partial}{\partial r}(ru) + \frac{\partial}{\partial z}(rw) = 0, \tag{13.1.12}$$

$$\frac{dp}{dz} = \mu\left(\frac{\partial^2 w}{\partial r^2} + \frac{1}{r}\frac{\partial w}{\partial r}\right) = \frac{\mu}{r}\frac{\partial}{\partial r}\left(r\frac{\partial w}{\partial r}\right), \tag{13.1.13}$$

where u and w are the velocity components for the motion of the fluid in relation to the moving coordinate system.

The boundary conditions for solving (13.1.12) and (13.1.13) are

$$u = \frac{\partial h}{\partial t}, \quad w = -c \quad \text{at} \quad r = h. \tag{13.1.14}$$

Integrating (13.1.13) at the constant z, we obatin

$$w = -c - \frac{1}{4\mu}\frac{dp}{dz}(h^2 - r^2). \tag{13.1.15}$$

To an observer moving with velocity c in the axial direction, the pressure and flow appear stationary. Hence the flow rate q measured in the moving coordinate system is a constant, independent of position and time. Now

$$q = 2\pi \int_0^h rw \, dr. \tag{13.1.16}$$

Using (13.1.15) we have

$$q = \pi h^2 c - \frac{\pi h^4}{8\mu}\frac{dp}{dz} \tag{13.1.17}$$

or

$$\frac{dp}{dz} = -\frac{8\mu q}{\pi h^4} - \frac{8\mu c}{h^2}. \tag{13.1.18}$$

Substituting in (13.1.15), we get

$$w = -c + 2(h^2 - r^2)\left[\frac{q}{\pi h^4} + \frac{c}{h^2}\right]. \tag{13.1.19}$$

To find the transverse velocity component u, we integrate the continuity equation (13.1.12) at the constant z. Remembering that $u = 0$ at $r = 0$, we obtain

$$ru = -\int_0^r r\frac{\partial w}{\partial z} dr. \tag{13.1.20}$$

Using (13.1.19) and remembering that $u(0,z)=0$, we get

$$u=-\frac{dh}{dz}\left(\frac{cr^3}{h^3}-\frac{2qr}{\pi h^3}+\frac{2qr^3}{\pi h^5}\right). \tag{13.1.21}$$

We now revert to the stationary coordinate system with the coordinates $R,\ Z$, the velocity components $U,\ W$, and the flow rate Q so that

$$W=w+c, \qquad U=u, \tag{13.1.22}$$

$$Q=2\pi\int_0^h WR\,dR \quad \text{or} \quad Q=q+\pi ch^2. \tag{13.1.23}$$

Let $\overline{Q}$ denote the time average of Q over a complete time period T for h so that

$$T = \frac{\lambda}{c} \tag{13.1.24}$$

$$\overline{Q} = \frac{1}{T}\int_0^T Q\,dt = q+\pi ca^2\left(1+\frac{1}{2}\epsilon^2\right). \tag{13.1.25}$$

Further Discussion on Long-wavelength Approximation

From (13.1.8) and (13.1.11),

$$h(z) = a\left[1+\epsilon\sin\left\{\frac{2\pi}{\lambda}(Z-ct)\right\}\right]=a\left[1+\epsilon\sin\left(\frac{2\pi}{\lambda}z\right)\right] \tag{13.1.26}$$

$$\frac{dh}{dz} = \frac{2\pi a\epsilon}{\lambda}\cos\left(\frac{2\pi}{\lambda}z\right)=\frac{2\pi a\epsilon}{\lambda}\cos\left\{\frac{2\pi}{\lambda}(Z-ct)\right\}. \tag{13.1.27}$$

From (13.1.11), (13.1.19), (13.1.21) and (13.1.22) we have

$$U = -\frac{2\pi a\epsilon}{\lambda}\cos\left\{\frac{2\pi}{\lambda}(Z-ct)\right\}\left[\frac{cR^3}{h^3}-\frac{2qR}{\pi h^3}+\frac{2qR^3}{\pi h^5}\right] \tag{13.1.28}$$

$$W = 2\left[\frac{q}{\pi h^4}+\frac{c}{h^2}\right](h^2-R^2). \tag{13.1.29}$$

Here h is determined as a function of Z and t from (13.1.26), and q is known from (13.1.25) after $\overline{Q}$ is determined experimentally.

To determine the pressure drop across a length equal to the wavelength λ, we integrate (13.1.18) to get

$$\begin{aligned}(\Delta p)_k &= -\frac{8\mu q}{\pi a^4}\int_0^\lambda \frac{dz}{\left[1+\epsilon\sin\left(\frac{2\pi}{\lambda}z\right)\right]^4}-\frac{8\mu c}{\pi a^2}\int_0^\lambda\frac{dz}{\left[1+\epsilon\sin\left(\frac{2\pi}{\lambda}z\right)\right]^2}\\ &= -\frac{4\mu\lambda}{\pi^2a^4}\int_0^{2\pi}\left[\frac{q}{[1+\epsilon\sin\tau]^4}+\frac{\pi ca^2}{[1+\epsilon\sin\tau]^2}\right]dr\\ &= -\frac{4\mu\lambda}{\pi a^4}\left[q\frac{2+3\epsilon^2}{(1-\epsilon^2)^{7/2}}+\frac{2\pi ca^2}{(1-\epsilon^2)^{3/2}}\right].\end{aligned} \tag{13.1.30}$$

The pressure drop across one wavelength would be zero if

$$q = -2\pi c \frac{a^2(1-\epsilon^2)^2}{2+3\epsilon^2}, \tag{13.1.31}$$

and then from (13.1.25),

$$\overline{Q} = \frac{\pi a^2 c(16\epsilon^2 - \epsilon^4)}{2(2+3\epsilon^2)}. \tag{13.1.32}$$

Substituting (13.1.31) in (13.1.28) and (13.1.29), we get

$$U = -\frac{2\pi ac\epsilon R}{\lambda h^3}\cos\left\{\frac{2\pi}{\lambda}(Z-ct)\right\}\left[R^2 + \frac{4ca^2(1-\epsilon^2)^2}{2+3\epsilon^2}\left(1-\frac{R^2}{h^2}\right)\right], \tag{13.1.33}$$

$$W = 2c\left[1 - \frac{2a^2(1-\epsilon^2)^2}{h^2(2+3\epsilon^2)}\right]\left(1-\frac{R^2}{h^2}\right). \tag{13.1.34}$$

For every fixed z, we can draw the velocity profiles U/c and W/C in the special case $(\Delta p)_\lambda = 0$. If $(\Delta p)_\lambda \neq 0$, then velocity profiles will depend also on q.

Unit 14

Course Structure

- Two Dimensional Flow in Renal Tubule
- Function of Rennal Tubule
- Basic Equations and Boundary Conditions
- Solution When Radial Velocity at Wall Decreases Linearly with z

14.1 Two Dimensional Flow in Renal Tubule

14.1.1 Function of Renal Tubule

The functional unit of the kidney is called the *nephron* or *renal tubule*, and each kidney has about 1 million of these tubules. One major part of a nephron is the glomerular tuft through which blood coming from the renal artery and afferent arterioles is filtered. THe glomerular filtrate is essentially identical to plasma, and no chemical separation occurs upto this point. If the kidneys deliver this filtrate for excretion, the body loses many valuable materials, including water, at a rate faster than the one at which they can supplied by synthesis or feeding. The rest of the nephron therefore recovers these valuable materials and returns them to the blood. Thus about 80 per cent of the filtrate is reabsorbed in the proximal tubule, and of the remaining, about 95 per cent is further reabsorbed by the end of the collecting ducts.

This reabsorption or seepage creates a radial component of the velocity in the cylindrical tubule, which must be considered along with the axial component of the velocity (see Fig. 14.1). Due to loss of fluid from the walls, both the radial and axial velocities decrease with z. Mathematically, we have to solve the problem of flow of viscous fluid in circular cylinder when there are axial and radial components of velocity and the radial velocity at all points on the surface of the cylinder is prescribed and is a decreasing function $\phi(z)$ of z.

14.1.2 Basic Equations and Boundary Conditions

At the outset, we may note that the equation of motion can be simplified since the inertial term in relation to the viscous terms can be neglectedd. THe average tubular radius is about 10^{-3} cm, the average velocity

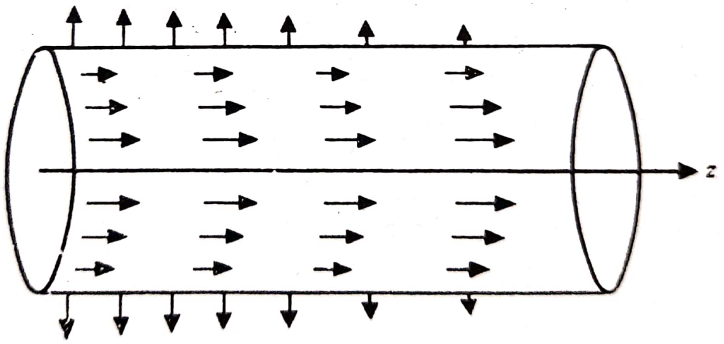


Figure 14.1: Two-dimensional flow in renal tubule.

is abut 10^{-1} cm/sec, since this is very much less than one, we neglect the inertial terms to get the following equations of continuity and motion

$$\frac{1}{r}\frac{\partial}{\partial r}(rv_r) + \frac{\partial v_z}{\partial z} = 0, \tag{14.1.1}$$

$$\frac{1}{\mu}\frac{\partial p}{\partial r} = \frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial}{\partial r}(rv_r)\right) + \frac{\partial^2 v_r}{\partial z^2}, \tag{14.1.2}$$

$$\frac{1}{\mu}\frac{\partial p}{\partial z} = \frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial}{\partial r}(rv_z)\right) + \frac{\partial^2 v_z}{\partial z^2}, \tag{14.1.3}$$

The boundary conditions are

$$\frac{\partial v_z}{\partial r} = 0, \quad v_r = 0, \quad v_z = \text{finite} \quad \text{at} \quad r = 0, \tag{14.1.4}$$

$$v_z = 0, \quad v_r = \phi(z) \quad \text{at} \quad r = R, \tag{14.1.5}$$

$$p = p_0 \quad \text{at} \quad z = 0, \tag{14.1.6}$$

$$p = p_L \quad \text{at} \quad z = L. \tag{14.1.7}$$

Eliminating p between (14.1.2) and (14.1.3), we get

$$\frac{\partial^2}{\partial r \partial z}\left[\frac{1}{r}\frac{\partial}{\partial z}(rv_z)\right] + \frac{\partial^3 v_r}{\partial z^3} = \frac{\partial}{\partial r}\left[\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{v_z}{\partial r}\right)\right] + \frac{\partial^3 v_z}{\partial z^2 \partial r}. \tag{14.1.8}$$

Taking the partial derivative of this equation with respect to z and substituting from (14.1.1), we get

$$\left[\frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial}{\partial r}\right)\right)\right) + 2\frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial}{\partial r}\left(\frac{\partial^2}{\partial z^2}\right)\right) + \frac{1}{r}\frac{\partial^4}{\partial z^4}\right](rv_r) = 0. \tag{14.1.9}$$

Alternatively, we can satisfy (14.1.1) by taking

$$v_r = \frac{1}{r}\frac{\partial \psi}{\partial z}, \quad v_z = -\frac{1}{r}\frac{\partial \psi}{\partial r}. \tag{14.1.10}$$

Substituting (14.1.9) in (14.1.7), we get

$$D^2(D^2\psi) = 0, \tag{14.1.11}$$

where the operator D^2 is defined by

$$D^2 \equiv \frac{\partial^2}{\partial r^2} - \frac{1}{r}\frac{\partial}{\partial r} + \frac{\partial^2}{\partial z^2}. \tag{14.1.12}$$

If

$$v_r = f(r)g(z), \tag{14.1.13}$$

then the form of (14.1.8) suggests that an analytical solution may be possible if

$$g(z) = A_0 + A_1 z \quad \text{or} \quad g(z) = A_2 e^{-\gamma z}. \tag{14.1.14}$$

From (14.1.5) since $v_r = \phi(z)$ when $r = R$, we get

$$f(R)g(z) = \phi(z). \tag{14.1.15}$$

This suggests that we may get an analytical solution when the radial component of velocity on the surface of the cylinder is given by

$$\phi(z) = a_0 + a_1 z \quad \text{or} \quad \phi(z) = ce^{\gamma z}. \tag{14.1.16}$$

We shall give the solutions for a special cases in §14.1.3.

14.1.3 Solution When Radial Velocity at Wall Decreases Linearly with z

For (14.1.11), we try the solution

$$\psi(r,z) = F(r)\left(a_0 z + \frac{1}{2}a_1 z^2\right) + G(r) \tag{14.1.17}$$

so that using (14.1.10), we get

$$\begin{aligned}
v_r &= \frac{1}{r}F(r)(a_0 + a_1 z), && (14.1.18)\\
v_z &= -\frac{1}{r}F'(r)\left(a_0 z + \frac{1}{2}a_1 z^2\right) - \frac{1}{r}G'(r), && (14.1.19)\\
D^2\psi &= \left(\frac{d^2}{dr^2} - \frac{1}{r}\frac{d}{dr}\right)F(r)\left(a_0 z + \frac{1}{2}a_1 z^2\right) + \left(\frac{d^2}{dr^2} - \frac{1}{r}\frac{d}{dr}\right)G(r) + a_1 F(r), && (14.1.20)\\
D^2(D^2\psi) &= \left(\frac{d^2}{dr^2} - \frac{1}{r}\frac{d}{dr}\right)^2 F(r)\left(a_0 z + \frac{1}{2}a_1 z^2\right) + \left(\frac{d^2}{dr^2} - \frac{1}{r}\frac{d}{dr}^2\right)G(r)\\
&\quad +2a_1\left(\frac{d^2}{dr^2} - \frac{1}{r}\frac{d}{dr}\right)F(r) = 0. && (14.1.21)
\end{aligned}$$

From (14.1.1) and (14.1.21), we get

$$\left(\frac{d^2}{dr^2} - \frac{1}{r}\frac{d}{dr}\right)^2 F(r) = 0, \tag{14.1.22}$$

$$\left(\frac{d^2}{dr^2} - \frac{1}{r}\frac{d}{dr}\right)^2 G(r) + 2a_1\left(\frac{d^2}{dr^2} - \frac{1}{r}\frac{d}{dr}\right)F(r) = 0. \tag{14.1.23}$$

Equation (14.1.22) gives

$$\left(\frac{d^2}{dr^2} - \frac{1}{r}\frac{d}{dr}\right)H(r) = 0, \quad \left(\frac{d^2}{dr^2} - \frac{1}{r}\frac{d}{dr}\right)F(r) = H(r). \tag{14.1.24}$$

Solving (14.1.24), we get

$$\begin{aligned}
H(r) &= Ar^2 + B, && (14.1.25)\\
r^2\frac{d^2F}{dr^2} - r\frac{dF}{dr} &= Ar^4 - Br^2. && (14.1.26)
\end{aligned}$$

Integrating (14.1.26), we obtain

$$F(r) = C + Dr^2 + \frac{Ar^4}{8} + \frac{Br^2}{2}\ln r. \tag{14.1.27}$$

From (14.1.23) and (14.1.27), we have

$$\left(\frac{d^2}{dr^2} - \frac{1}{r}\frac{d}{dr}\right)\left[\left(\frac{d^2}{dr^2} - \frac{1}{r}\frac{d}{dr}\right)G(r) + 2a_1F(r)\right] = 0. \tag{14.1.28}$$

Using (14.1.24) and (14.1.25), we get

$$\left(\frac{d^2}{dr^2} - \frac{1}{r}\frac{d}{dr}\right)G(r) + 2a_1F(r) = Mr^2 + N. \tag{14.1.29}$$

Now from (14.1.24), (14.1.25), (14.1.18) and (14.1.19), we have

$$\frac{d}{dr}\left[\frac{1}{r}F'(r)\right] = 0 \quad \text{at} \quad r = 0, \tag{14.1.30}$$

$$\frac{d}{dr}\left[\frac{1}{r}G'(r)\right] = 0 \quad \text{at} \quad r = 0, \tag{14.1.31}$$

$$\frac{1}{r}F(r) = 0 \quad \text{at} \quad r = 0, \tag{14.1.32}$$

$$\frac{1}{r}F'(r) \quad \text{and} \quad \frac{1}{r}G'(r) \quad \text{are finite at} \quad r = 0, \tag{14.1.33}$$

$$F'(R) = 0, \quad G'(R) = 0, \quad F(R) = R. \tag{14.1.34}$$

From (14.1.27), (14.1.32) and (14.1.33), we obtain

$$C = 0, \qquad B = 0. \tag{14.1.35}$$

From (14.1.27), (14.1.34) and (14.1.35)

$$2DR + \frac{1}{2}AR^3 = 0, \quad DR^2 + \frac{AR^4}{8} = R \tag{14.1.36}$$

so that

$$F(r) = \frac{2r^2}{R} - \frac{r^4}{R^3} = R\left[2\left(\frac{r}{R}\right)^2 - \left(\frac{r}{R}\right)^4\right]. \tag{14.1.37}$$

Substituting (14.1.37) in (14.1.29), we get

$$\frac{d^2G}{dr^2} - \frac{1}{r}\frac{dG}{dr} = Mr^2 + N - 4a_1\frac{r^2}{R} + 2a_1\frac{r^4}{R^3}. \tag{14.1.38}$$

Integrating (14.1.38), we obtain

$$G(r) = M_1r^2 + N_1 + \frac{Mr^4}{8} + \frac{Nr^2\ln r}{2} - \frac{a_1}{2}\frac{r^4}{R} + \frac{a_1}{12}\frac{r^6}{R^3}. \tag{14.1.39}$$

From (14.1.33) and (14.1.39), we have

$$N = 0. \tag{14.1.40}$$

From (14.1.34) and (14.1.39), we have

$$2M_1R + \frac{1}{2}MR^3 - \frac{3a_1}{2}R^2 = 0. \tag{14.1.41}$$

Equation 14.1.41 can determine only one of the two unknown constants M and M_1. To determine both of them, we need one more relation. This relation can be found in terms of Q_0 which is the total flux at $z = 0$. Using (14.1.18) and (14.1.19), we get

$$\begin{aligned} Q(z) &= \int_0^R 2\pi r v_z(r,z)\, dr \\ &= 2\pi \int_0^R \left[\left(\frac{4r^3}{R^3} - \frac{4r}{R} \right) \left(a_0 z + \frac{1}{2} a_1 z^2 \right) - 2M_1 r - \frac{M r^3}{2} - \frac{2a_1}{R} r^3 + \frac{a_1 r^5}{2R^3} \right] dr, \end{aligned} \tag{14.1.42}$$

$$\therefore \frac{Q_0}{2\pi R^2} = \frac{MR^2}{8} - \frac{a_1}{3} R, \tag{14.1.43}$$

$$\Rightarrow M = \frac{8}{R^2} \left(\frac{Q_0}{2\pi R^2} + \frac{a_1 R}{3} \right), \tag{14.1.44}$$

$$\Rightarrow M_1 = -\frac{Q_0}{\pi R^2} + \frac{a_1 R}{12}. \tag{14.1.45}$$

From (14.1.39), (14.1.40), 14.1.44 and 14.1.45,

$$G(r) = \left(\frac{a_1 R}{12} - \frac{Q_0}{\pi R^2} \right) r^2 + N_1 + \frac{1}{R^2} \left(\frac{Q_0}{2\pi R^2} + \frac{a_1 R}{3} \right) r^4 - \frac{a_1}{2} \frac{r^4}{R} + \frac{a_1}{12} \frac{r^6}{R^3}. \tag{14.1.46}$$

The constant N_1 need not to be determined since $\psi(r,z)$ can always contain an arbitrary constant without affecting the velocity components.

From (14.1.18) and (14.1.19), (14.1.27), (14.1.46), we have

$$v_r(r,z) = \left[2\frac{r}{R} - \left(\frac{r}{R} \right)^3 \right] (a_0 + a_1 z), \tag{14.1.47}$$

$$\begin{aligned} v_z(r,z) &= -4 \left(\frac{r}{R} - \frac{r^3}{R^3} \right) \left(a_0 z + \frac{1}{2} a_1 z^2 \right) - 2 \left(\frac{a_1 R}{12} - \frac{Q_0}{\pi R^2} \right. \\ &\quad \left. - \frac{4}{R^2} \left(\frac{Q_0}{2\pi R^2} + \frac{a_1 R}{3} \right) r^2 + 2a_1 \frac{r^2}{R} - \frac{a_1}{2} \frac{r^4}{R^3} \right) \\ &= \left(1 - \frac{r^2}{R^2} \right) \left[\frac{2Q_0}{\pi R^2} - \frac{2}{R} (2a_0 z + a_1 z^2) - \frac{a_1 R}{2} \left(\frac{1}{3} - \frac{r^2}{R^2} \right) \right]. \end{aligned} \tag{14.1.48}$$

Differentiating (14.1.42), we obtain

$$\frac{dQ}{dz} = 8\pi (a_0 + a_1 z) \int_0^R \left(\frac{r^3}{R^3} - \frac{r}{R} \right) dr = -2\pi R (a_0 + a_1 z) \tag{14.1.49}$$

so that the decrease of flux is equal to the amount of the fluid coming out of the cylinder per unit length per unit time. Integrating (14.1.49), we get

$$Q(z) = Q_0 - \pi R (2a_0 z + a_1 z^2). \tag{14.1.50}$$

From (14.1.48) and (14.1.50), we have

$$v_z = \left(1 - \frac{r^2}{R^2} \right) \left[\frac{2Q(z)}{\pi R^2} - \frac{a_1 R}{2} \left(\frac{1}{3} - \frac{r^2}{R^2} \right) \right]. \tag{14.1.51}$$

For Hagen-Poiseuille flow in a circular tube, we have

$$v_z = \left(1 - \frac{r^2}{R^2}\right)\frac{2Q}{\pi R^2}. \tag{14.1.52}$$

Comparing (14.1.51) and (14.1.52), we find that there are two changes:

(i) Q is replaced by the variable $Q(z)$, and

(ii) there is further distortion due to the varying nature of the radial flow.

Using (14.1.2), (14.1.3), (14.1.47), (14.1.48) and (14.1.50), we get

$$\frac{\partial p}{\partial r} = -\frac{8\mu r}{R^2}(a_0 + a_1 z), \tag{14.1.53}$$

$$\frac{\partial p}{\partial z} = -\frac{4a_1\mu}{R}\left[\frac{r^2}{R^2} + \frac{2Q(z)}{a_1\pi R^3} + \frac{1}{2}\right]. \tag{14.1.54}$$

Integrating (14.1.53), we obtain

$$p(r,z) = -\frac{4\mu r^2}{R^3}(a_0 + a_1 z) + K(z). \tag{14.1.55}$$

Differentiating (14.1.55) partially with respect to z and then substituting $\partial p/\partial z$ in (14.1.54), we get

$$K'(z) = -\frac{4a_1\mu}{R}\left[\frac{1}{3} + \frac{2Q(z)}{a_1\pi R^2}\right] \tag{14.1.56}$$

so that

$$K(z) = -\frac{4a_1\mu}{R}\left[\frac{1}{3}z + \frac{2z\overline{Q}(z)}{a_1\pi R^3}\right] + K_0, \tag{14.1.57}$$

where

$$\overline{Q}(z) = \int_0^z Q(z)\, dz. \tag{14.1.58}$$

Substituting from (14.1.57) in (14.1.56), we get

$$p(r,z) - p(0,0) = -\frac{4\mu}{R}(a_0 + a_1 z)\frac{r^2}{R^2} - \mu\left(\frac{4a_1}{3R} + \frac{8\overline{Q}}{\pi R^4}\right)z. \tag{14.1.59}$$

The average pressure $\overline{p}(z)$ at any section is given by

$$\overline{p}(z) = \frac{\int_0^R p(r,z)2\pi r\, dr}{\int_0^R 2\pi r\, dr} = -\mu\left[\frac{2a_0}{R} + \left(\frac{8\overline{Q}(z)}{\pi R^4} + \frac{10a_1}{3R}\right)z\right] \tag{14.1.60}$$

Thus the pressure drop over the tube length L is

$$\Delta\overline{p} = \overline{p}(0) - \overline{p}(L) = \mu\left[\frac{8\overline{Q}(L)}{\pi R^4} + \frac{10a_1}{3R}\right]L. \tag{14.1.61}$$

Unit 15

Course Structure

- Diffusion and Diffusion-Reaction Models
- Fick's Laws of Diffusion
- Solution of the One-dimensional Diffusion Equation
- Solution of the Two-dimensional Diffusion Equation

15.1 The Diffusion Equation

15.1.1 Fick's Laws of Diffusion

Let $c(x, y, z, t)$ be the concentration of a solute or the amount of the solute per unit volume at the point (x, y, z) at time t. Due to the concentration gradient grad c, there is a flow of solute given by the current density vector **j**, which, according to *Fick's first law of diffusion,* is given by

$$\mathbf{j} = D \text{ grad } c = -D\nabla c \tag{15.1.1}$$

or

$$j_x = -D\frac{\partial c}{\partial x}, \quad j_y = -D\frac{\partial c}{\partial y}, \quad j_z = -D\frac{\partial c}{\partial z}. \tag{15.1.2}$$

Here the quantities j_x, j_y, j_z give respectively the amounts of the solute crossing the planes perpendicular to $x,\ y,\ z$ axes per unit area per unit time so that the dimensions of D are

$$\frac{ML^{-2}T^{-1}}{ML^{-3}L^{-1}} = L^2T^{-1}. \tag{15.1.3}$$

The negative signs in (15.1.1) and (15.1.2) indicate that the flow takes place in the direction of decresing concentration. D can vary with $x,\ y,\ z$ but we shall take it to be constant. Its values for some common biological solutes in water lie between 0.05×10^{-6} and $10 \times 10^{-6}\ cm^2/sec$.

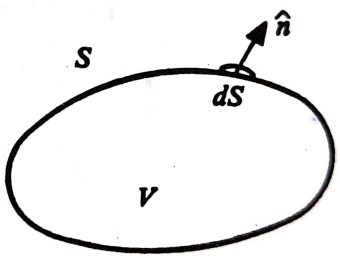


Figure 15.1: Control volume.

Now, consider a volume V with surface S (see Fig. 15.1). The rate of change of the amount of the solute is given by

$$\frac{\partial}{\partial t}\int\limits_V c(x,y,z,t)\,dx\,dy\,dz. \tag{15.1.4}$$

The amount of the solute which comes out of the surface S per unit time is given by

$$\int\limits_S \mathbf{j}\cdot\hat{\mathbf{n}}\,dS, \tag{15.1.5}$$

where $\hat{\mathbf{n}}$ is the unit normal vector to the surface. If there is no source or sink inside the volume, then on using (15.1.1), (15.1.4) and (15.1.5), and Gauss' divergence theorem, we get

$$\begin{aligned}\frac{\partial}{\partial t}\int\limits_V c(x,y,z,t)\,dx\,dy\,dz &= -\int\limits_S \mathbf{j}\cdot\hat{\mathbf{n}}\,dS\\ &= \int\limits_S (D\ \text{grad}\ c)\cdot\hat{\mathbf{n}}\,dS\\ &= \int\limits_V \text{div}\ (D\ \text{grad}\ c)\,dx\,dy\,dz\end{aligned} \tag{15.1.6}$$

so that

$$\int\limits_V \left[\frac{\partial c}{\partial t} - \text{div}\ (D\ \text{grad}\ c)\right]\,dx\,dy\,dz = 0. \tag{15.1.7}$$

Since (15.1.7) holds for all volumes, we get *Fick's second law of diffusion* as

$$\frac{\partial c}{\partial t} = \text{div}\ (D\ \text{grad}\ c). \tag{15.1.8}$$

Since D is assumed to be constant, we get the diffusion equation

$$\frac{\partial c}{\partial t} = D\ \text{div}\ (\text{grad}\ c) = D\nabla^2 c = D\left(\frac{\partial^2 c}{\partial x^2}+\frac{\partial^2 c}{\partial y^2}+\frac{\partial^2 c}{\partial z^2}\right). \tag{15.1.9}$$

The equation governing the temperature θ of a heat-conducting homogeneous solid is given by

$$\frac{\partial \theta}{\partial t} = k\left(\frac{\partial^2 \theta}{\partial x^2}+\frac{\partial^2 \theta}{\partial y^2}+\frac{\partial^2 \theta}{\partial z^2}\right). \tag{15.1.10}$$

where k is called the *thermal diffusivity* of the solid. The diffusion equation is therefore also known as the *heat-conduction equation.*

15.1.2 Some Solution of the One-dimensional Diffusion Equation

Solution I

If there is diffusion only in the direction of the x-axis, (15.1.9) gives

$$\frac{\partial c}{\partial t} = D\frac{\partial^2 c}{\partial x^2}. \tag{15.1.11}$$

By differentiating and substituting in (15.1.11), it can be easily verified that

$$c = c(x,t) = \frac{m}{(4\pi Dt)^{1/2}} \exp\left[-\frac{x^2}{4Dt}\right] \tag{15.1.12}$$

is a solution of (15.1.11). Also,

$$\begin{aligned} \int_{-\infty}^{\infty} c(x,t)\, dx &= \frac{m}{(4\pi Dt)^{1/2}} \int_{-\infty}^{\infty} \exp\left[-\frac{x^2}{4Dt}\right] dx \\ &= \frac{m}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}y^2\right] dy = m \end{aligned} \tag{15.1.13}$$

so that m denotes the total amount of the diffusing solute. It is easily seen that $\dfrac{c}{m}$ is the density function for the normal probability distribution with mean zero and variance $2Dt$. The graphs of $\frac{c}{m}$ against x for $Dt = 4, 1, 1/4, 1/9$ and $1/6$ are given in Fig. 15.2.

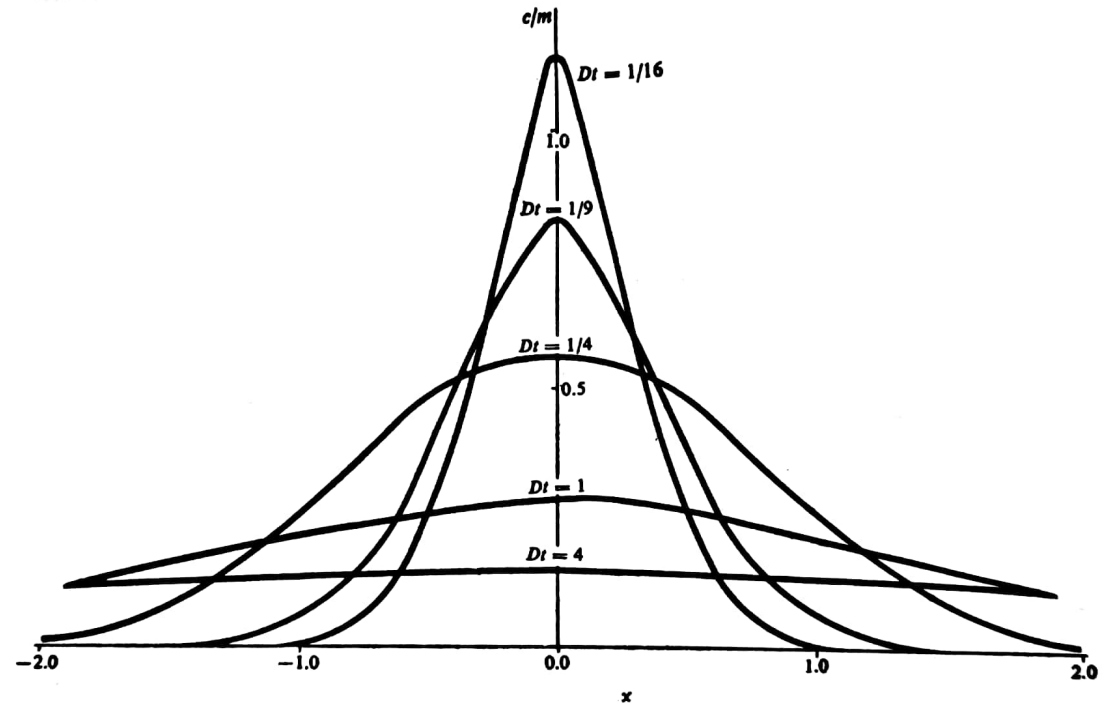


Figure 15.2: Graph of c/m against x.

The area under each of these curves is unity. As $t \to 0$, the variance tends to zero and we get Dirac delta-function $\delta(x)$ which vanishes everywhere except at $x = 0$ and is such that

$$\int_{-\infty}^{\infty} \delta(x)\, dx = 1, \quad \int_{-\infty}^{\infty} f(x)\delta(x)\, dx = f(0). \tag{15.1.14}$$

It thus appears that

$$\delta(x) = \lim_{t\to 0} \frac{1}{(4\pi Dt)^{1/2}} \exp\left[-\frac{x^2}{4Dt}\right]. \tag{15.1.15}$$

Thus (15.1.12) gives the concentration due to a solute mass m placed at $x = 0$ at time $t = 0$. If a unit mass of solute is placed at $x = \xi$, the concentration $c(x, t)$ is given by

$$c(x,t) = \frac{1}{(4\pi Dt)^{1/2}} \exp\left[-\frac{(x-\xi)^2}{Dt}\right]. \tag{15.1.16}$$

If the solute has an initial density distribution $A(\xi)\, d\xi$, then the concentration of the solute at time t is given by

$$c(x,t) = \frac{1}{(4\pi Dt)^{1/2}} \int_0^{\infty} A(\xi) \exp\left[-\frac{(x-\xi)^2}{Dt}\right] d\xi. \tag{15.1.17}$$

Solution II

For obtaining the second solution of (15.1.11), if $c(x, t)$ satisfies (15.1.11), then $\dfrac{\partial c}{\partial x}$ also satisfies it. Conversely, if (15.1.12) is a solution of (15.1.11), then

$$\frac{m}{(4\pi Dt)^{1/2}} \int_{-\infty}^{x} \exp\left[-\frac{x^2}{4Dt}\right] dx = \frac{m}{\sqrt{\pi}} \int_0^{\eta} \exp[-\eta^2]\, d\eta, \tag{15.1.18}$$

where

$$\eta = \frac{x}{(4Dt)^{1/2}}, \tag{15.1.19}$$

is also a solution of (15.1.11). If we define *error function* $\text{erf}(z)$ and *error function complement* $\text{erfc}(z)$ as

$$\begin{aligned} \text{erf}(z) &= \frac{2}{\sqrt{\pi}} \int_0^{z} \exp[-\eta^2]\, d\eta, \\ \text{ercf}(z) &= 1 - \text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_z^{\infty} \exp[-\eta^2]\, d\eta, \end{aligned} \tag{15.1.20}$$

then we find that $\text{erfc}\left[\dfrac{x}{(4Dt)^{1/2}}\right]$ is a solution of the one-dimensional diffusion equation. We may note that

$$\text{erf}(-z) = erf(z), \quad \text{erf}(0) = 0, \quad \text{erf}(\infty) = 1. \tag{15.1.21}$$

Since both $\text{erf}(z)$ and $\text{erfc}(z)$ are tabulated functions, we have a convenient solution of the one dimensional diffusion equation.

Solution III

For solving the boundary value problem for which there is no flux at $x = 0$ and $x = a$, i.e., for solving (15.1.11) subject to the boundary conditions

$$\frac{\partial c}{\partial x} = 0 \quad \text{at} \quad x = 0,\ x = a, \tag{15.1.22}$$

we use the method of *separation of variables* and try the solution of the form

$$c(x,t) = X(x)T(t) \tag{15.1.23}$$

for (15.1.11) to get

$$\frac{1}{T}\frac{dT}{dt} = \frac{D}{X}\frac{d^2X}{dx^2} = -k^2 \ \text{(say)} \tag{15.1.24}$$

so that

$$c(x,t) = \sum_k \exp[-k^2 t]\left[A_k \cos\left(\frac{kx}{\sqrt{D}}\right) + B_k \sin\left(\frac{kx}{\sqrt{D}}\right)\right]. \tag{15.1.25}$$

Equation (15.1.22) then give

$$B_k = 0, \quad \frac{k}{\sqrt{D}} = \frac{n\pi}{a} \tag{15.1.26}$$

so that

$$c(x,t) = \sum_{n=0}^{\infty} C_n \exp\left[-\frac{n^2\pi^2 Dt}{a^2}\right] \cos\left[\frac{n\pi x}{a}\right]. \tag{15.1.27}$$

To determine the constants C_n, we make use of the knowledge of the initial distribution of concentration $c(x,0) = f(x)$, so that

$$f(x) = \sum_{n=0}^{\infty} C_n \cos\left(\frac{n\pi x}{a}\right). \tag{15.1.28}$$

Expanding $f(x)$ in a half-range cosine series, we get

$$C_0 = \frac{1}{a}\int_0^a f(x)\,dx, \tag{15.1.29}$$

$$C_n = \frac{2}{a}\int_0^a f(x)\,\cos\left(\frac{n\pi x}{a}\right)\,dx \quad (n = 1,2,3,\ldots). \tag{15.1.30}$$

As $t \to \infty$, we get, from (15.1.27) and (15.1.29),

$$\lim_{t\to\infty} c(x,t) = C_0 = \frac{1}{a}\int_0^a f(x)\,dx \tag{15.1.31}$$

which is only the average value of the initial concentration. This shows that, as $t \to \infty$, the concentration tends to become uniform and equal to the average value of the initial concentration. In fact, from (15.1.27),

$$\int_0^a c(x,t)\,dx = C_0 a = \int_0^a f(x)\,dx \tag{15.1.32}$$

so that the total amount of the solute at any time t is equal to the initial total amount. This result is expected since, according to boundary conditions (15.1.22), no solute enters or leaves the boundaries.

15.1.3 Some Solutions of the Two-dimensional Diffusion Equation

Solution I

By using the method of separation of variables, it is easily seen that the solution of the equation

$$\frac{\partial c}{\partial t} = D\left(\frac{\partial^2 c}{\partial x^2} + \frac{\partial^2 c}{\partial y^2}\right) \tag{15.1.33}$$

is

$$c(x,y,t) = \sum_{\lambda}\sum_{\mu} C_{\lambda\mu}\cos(\lambda x + \epsilon_k)\cos(\mu x + \epsilon_m u)\exp[-(\lambda^2+\mu^2)Dt]. \tag{15.1.34}$$

If the boundary conditions are

$$\begin{aligned} \frac{\partial c}{\partial x} &= 0 \quad \text{when} \quad x = 0, a, \\ \frac{\partial c}{\partial y} &= 0 \quad \text{when} \quad y = 0, b, \end{aligned} \tag{15.1.35}$$

we get

$$\epsilon_\lambda = 0, \quad \epsilon_\mu = 0, \quad \lambda = \frac{m\pi}{a}, \quad \mu = \frac{n\pi}{b} \tag{15.1.36}$$

so that

$$c(x,y,t) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} C_{mn}\cos\left(\frac{m\pi x}{a}\right)\cos\left(\frac{n\pi x}{b}\right)\exp\left[-\left(\frac{m^2\pi^2}{a^2}+\frac{n^2\pi^2}{b^2}\right)Dt\right]. \tag{15.1.37}$$

If the initial concentration is $f(x,y)$, then we get

$$C_{00} = \frac{1}{ab}\int_0^b\int_0^a f(x,y)\,dy\,dx, \tag{15.1.38}$$

$$C_{m0} = \frac{2}{ab}\int_0^b\int_0^a f(x,y)\,\cos\left(\frac{m\pi x}{a}\right)\,dy\,dx, \tag{15.1.39}$$

$$C_{0n} = \frac{2}{ab}\int_0^b\int_0^a f(x,y)\,\cos\left(\frac{m\pi y}{b}\right)\,dy\,dx, \tag{15.1.40}$$

$$C_{mn} = \frac{4}{ab}\int_0^b\int_0^a f(x,y)\,\cos\left(\frac{m\pi x}{a}\right)\,\cos\left(\frac{n\pi y}{b}\right)\,dy\,dx, \tag{15.1.41}$$

so that, as expected

$$\lim_{t\to\infty} c(x,y,t) = C_{00} = \frac{1}{ab}\int_0^b\int_0^a f(x,y)\,dy\,dx, \tag{15.1.42}$$

$$\int_0^b\int_0^a c(x,y,t)\,dy\,dx = abC_{00} = \int_0^b\int_0^a f(x,y)\,dy\,dx. \tag{15.1.43}$$

Solution II

For the axially-symmetric case, the diffusion equation in cylindrical polar coordinates is

$$\frac{\partial c}{\partial t} = D\left(\frac{\partial^2 c}{\partial r^2} + \frac{1}{r}\frac{\partial c}{\partial r} + \frac{\partial^2 c}{\partial z^2}\right). \tag{15.1.44}$$

By using the method of separation of variables, we get a solution of (15.1.44), namely,

$$c(x,z,t) = \sum_{\mu}\sum_{k} A_{\lambda\mu} J_0(\sqrt{\lambda^2+\mu^2}r)\exp(-\lambda^2 Dt \pm \mu z). \tag{15.1.45}$$

A solution independent of z is

$$c(r,t) = \sum_{\lambda} A_k J_0(\lambda r)\exp(-\lambda^2 Dt). \tag{15.1.46}$$

If the flux $\dfrac{\partial c}{\partial r} = 0$ across the cylindrical boundary $r = a$, then

$$J_1(\lambda a) = 0 \quad \text{or} \quad \lambda = \frac{\xi}{a}, \tag{15.1.47}$$

where ξ is a zero of the first order Bessel function. Hence

$$c(r,t) = \sum_{n=1}^{\infty} B_n J_0\left(\xi_n \frac{r}{a}\right)\exp\left[-\left(\frac{\xi_n^2}{a^2}\right)Dt\right], \tag{15.1.48}$$

where ξ_n is the n-th zero of $J_1(x)$. The constants B_n are to be determined from

$$c(r,0) = f(r) = \sum_{n=1}^{\infty} B_n J_0\left(\xi_n \frac{r}{a}\right). \tag{15.1.49}$$

If the boundary condition is $c = 0$ at $r = a$, then (15.1.46) gives

$$J_0(\lambda a) = 0 \tag{15.1.50}$$

so that

$$c(r,t) = \sum_{n=1}^{\infty} D_n \exp\left[-\frac{\eta_n^2 Dt}{a} J_0\left(\frac{\eta_n r}{a}\right)\right], \tag{15.1.51}$$

where

$$D_n = \frac{2}{a^2 J_1^2(\eta_n)} \int_0^a r f(r) J_0\left(\frac{\eta_n r}{a}\right)\, dr \tag{15.1.52}$$

and η_n is the n-th zero of zero order Bessel function.

Unit 16

Course Structure

- Ecological Application of Diffusion Models
- Diffusion on the Stability of Single Species Model
- Diffusion on the Stability of Two Species Model
- Diffusion on the Stability of Prey-Predator Models

16.1 Application of Diffusion and Diffusion-Reaction Models in Population Biology

In the absence of diffusion, if an ecological model for n species is

$$\frac{dc_i}{dt} = Q_i(c_1, c_2, \ldots, c_n), \tag{16.1.1}$$

then a model with diffusion is represented by

$$\frac{\partial c_i}{\partial t} = D_i \nabla^2 c_i + Q_i(c_1, c_2, \ldots, c_n), \quad i = 1, 2, \ldots, n, \tag{16.1.2}$$

where D_i is the coefficient of diffusion of the i-th substance and Q_i is the rate of its generation per unit time. Equation (16.1.2) is called *diffusion-reaction equation.* IN particular, if $N_1(x,t)$, $N_2(x,t)$ denote the densities of the two species at the point x at time t in a medium in which both species are diffusing in the direction of the x-axis only, then a *competition model with diffusion is*

$$\begin{aligned} \frac{\partial N_1}{\partial t} &= N_1(a_1 - b_{11}N_1 - b_{12}N_2) + D_1\frac{\partial^2 N_1}{\partial x^2}, \\ \frac{\partial N_2}{\partial t} &= N_2(a_2 - b_{21}N_1 - b_{22}N_2) + D_2\frac{\partial^2 N_2}{\partial x^2}, \end{aligned} \tag{16.1.3}$$

and a *prey-predator model with diffusion* is

$$\begin{aligned} \frac{\partial N_1}{\partial t} &= N_1(a_1 - b_{11}N_1 - b_{12}N_2) + D_1\frac{\partial^2 N_1}{\partial x^2}, \\ \frac{\partial N_2}{\partial t} &= N_2(-a_2 - b_{21}N_1 - b_{22}N_2) + D_2\frac{\partial^2 N_2}{\partial x^2}, \end{aligned} \tag{16.1.4}$$

Now we will discuss the stabilities of the equilibrium states of these models.

16.2 Absence of Diffusive Instability for Single Species

In the absence of diffusion, let a population grow according to the law

$$\frac{dN}{dt} = f(N). \tag{16.2.1}$$

Let the population be confined to the volume $0 \le x \le a,\ 0 \le y \le b,\ 0 \le z \le c$, and let there be diffusion. Let there be no flux across the faces of the rectangular parallelepiped so that (16.2.1) becomes

$$\frac{\partial N}{\partial t} = f(N) + D\left(\frac{\partial^2 N}{\partial x^2} + \frac{\partial^2 N}{\partial y^2} + \frac{\partial^2 N}{\partial z^2}\right). \tag{16.2.2}$$

The boundary conditions are

$$\begin{aligned} \frac{\partial N}{\partial x} &= 0 \quad \text{at} \quad x = 0, a, \\ \frac{\partial N}{\partial y} &= 0 \quad \text{at} \quad y = 0, b, \\ \frac{\partial N}{\partial z} &= 0 \quad \text{at} \quad z = 0, c. \end{aligned} \tag{16.2.3}$$

If $\overline{N}$ gives an equilibrium value for (16.2.1), it also gives an equilibrium value for (16.2.2). Let

$$N(x, y, z, t) = \overline{N} + u(x, y, z, t), \tag{16.2.4}$$

where u is sufficiently small so its squares and higher powers can be neglected. Then (16.2.2) gives

$$\frac{\partial u}{\partial t} = u\frac{\partial f}{\partial \overline{N}} + D\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}\right). \tag{16.2.5}$$

where $\dfrac{\partial f}{\partial \overline{N}}$ denote the value of $\dfrac{\partial f}{\partial N}$ at the equilibrium point $\overline{N}$. Now the boundary condition (16.2.3) becomes

$$\begin{aligned} \frac{\partial u}{\partial x} &= 0 \quad \text{at} \quad x = 0, a, \\ \frac{\partial u}{\partial y} &= 0 \quad \text{at} \quad y = 0, b, \\ \frac{\partial u}{\partial z} &= 0 \quad \text{at} \quad z = 0, c. \end{aligned} \tag{16.2.6}$$

For (16.2.5), we try the solution

$$u(x, y, z, t) = e^{\lambda t} \sum_p \sum_n \sum_m A_{mnp} \cos\left(\frac{m\pi x}{a}\right) \cos\left(\frac{n\pi y}{b}\right) \cos\left(\frac{p\pi z}{c}\right) \tag{16.2.7}$$

which automatically satisfies boundary conditions (16.2.6). Substituting (16.2.7) in (16.2.5), we get

$$\lambda - \frac{\partial f}{\partial \overline{N}} + D\left(\frac{m^2\pi^2}{\partial a^2} + \frac{n^2\pi^2}{b^2} + \frac{p^2\pi^2}{\partial c^2}\right) = 0 \tag{16.2.8}$$

or

$$\lambda = \frac{\partial f}{\partial \overline{N}} - D\sigma^2, \qquad \text{where} \quad \sigma^2 = \left(\frac{m^2}{\partial a^2} + \frac{n^2}{b^2} + \frac{p^2}{\partial c^2}\right)\pi^2. \tag{16.2.9}$$

If, in the absence of diffusion, the equilibrium position is unstable, then $\dfrac{\partial f}{\partial \overline{N}}$ is negative, and so λ is also negative. Therefore, a position of equilibrium, which is stable in the absence of diffusion remains stable when there is diffusion in a finite domain with no flux across its surfaces. Thus there is no possibility of diffusion-induced instability when there is only one single species.

16.3 Possibility of Diffusive Instability for Two Species

If $N_1(x, y, z, t)$ and $N_2(x, y, z, t)$ are the populations of the two species, then the basic diffusion reaction equations are

$$\frac{\partial N_1}{\partial t} = f_1(N_1, N_2) + D_1\left(\frac{\partial^2 N_1}{\partial x^2} + \frac{\partial^2 N_1}{\partial y^2} + \frac{\partial^2 N_1}{\partial z^2}\right), \tag{16.3.1}$$

$$\frac{\partial N_2}{\partial t} = f_2(N_1, N_2) + D_2\left(\frac{\partial^2 N_2}{\partial x^2} + \frac{\partial^2 N_2}{\partial y^2} + \frac{\partial^2 N_2}{\partial z^2}\right). \tag{16.3.2}$$

The equilibrium position for these equations is given by

$$f_1(\overline{N_1}, \overline{N_2}) = 0, \qquad f_2(\overline{N_1}, \overline{N_2}) = 0. \tag{16.3.3}$$

If

$$\begin{aligned} N_1(x, y, z, t) &= \overline{N_1} + u_1(x, y, z, t), \\ N_2(x, y, z, t) &= \overline{N_2} + u_2(x, y, z, t), \end{aligned} \tag{16.3.4}$$

then, after substituting (16.3.1) and (16.3.2) and linearlizing, we get

$$\frac{\partial u_1}{\partial t} = u_1\frac{\partial f_1}{\partial \overline{N_1}} + u_2\frac{\partial f_1}{\partial \overline{N_2}} + D_1\left(\frac{\partial^2 u_1}{\partial x^2} + \frac{\partial^2 u_1}{\partial y^2} + \frac{\partial^2 u_1}{\partial z^2}\right), \tag{16.3.5}$$

$$\frac{\partial u_2}{\partial t} = u_1\frac{\partial f_2}{\partial \overline{N_1}} + u_2\frac{\partial f_2}{\partial \overline{N_2}} + D_2\left(\frac{\partial^2 u_2}{\partial x^2} + \frac{\partial^2 u_2}{\partial y^2} + \frac{\partial^2 u_2}{\partial z^2}\right) \tag{16.3.6}$$

where $\dfrac{\partial f_i}{\partial \overline{N_i}}$, $i = 1, 2$, denotes the value of $\dfrac{\partial f_i}{\partial N_i}$ at the equilibrium point $\overline{N_1}, \overline{N_2}$. When there is no flux, the boundary conditions are

$$\begin{aligned} \frac{\partial u_i}{\partial x} &= 0 \quad \text{at} \quad x = 0, a, \\ \frac{\partial u_i}{\partial y} &= 0 \quad \text{at} \quad y = 0, b, \\ \frac{\partial u_i}{\partial z} &= 0 \quad \text{at} \quad z = 0, c. \end{aligned} \tag{16.3.7}$$

where $i = 1, 2$. Trying the solution

$$\begin{aligned} u_1 &= e^{\lambda t}\sum_p\sum_n\sum_m a_{mnp}\cos\frac{m\pi x}{a}\cos\frac{n\pi y}{b}\cos\frac{p\pi z}{c}, \\ u_2 &= e^{\lambda t}\sum_p\sum_n\sum_m b_{mnp}\cos\frac{m\pi x}{a}\cos\frac{n\pi y}{b}\cos\frac{p\pi z}{c}, \end{aligned} \tag{16.3.8}$$

we get

$$\begin{vmatrix} \lambda - \dfrac{\partial f_1}{\partial \overline{N_1}} + D_1\sigma^2 & -\dfrac{\partial f_1}{\partial \overline{N_2}} \\ -\dfrac{\partial f_2}{\partial \overline{N_1}} & \lambda - \dfrac{\partial f_2}{\partial \overline{N_2}} + D_2\sigma^2 \end{vmatrix} = 0 \tag{16.3.9}$$

where $\sigma^2 = \left(\dfrac{m^2}{a^2} + \dfrac{n^2}{b^2} + \dfrac{p^2}{c^2}\right)\pi^2$.

or

$$\begin{aligned} \lambda^2 &+ \lambda\left[(D_1+D_2)\sigma^2 - \frac{\partial f_1}{\partial \overline{N_1}} - \frac{\partial f_2}{\partial \overline{N_2}}\right] + \left(\frac{\partial f_1}{\partial \overline{N_1}}\frac{\partial f_2}{\partial \overline{N_2}} - \frac{\partial f_1}{\partial \overline{N_2}}\frac{\partial f_2}{\partial \overline{N_1}}\right) \\ &- \sigma^2\left(D_1\frac{\partial f_2}{\partial \overline{N_2}} + D_2\frac{\partial f_1}{\partial \overline{N_1}}\right) + D_1D_2\sigma^4 = 0. \end{aligned} \tag{16.3.10}$$

In the absence of diffusion, the equation corresponding to (16.3.10) is

$$\lambda^2 - \lambda\left(\frac{\partial f_1}{\partial \overline{N_1}} + \frac{\partial f_2}{\partial \overline{N_2}}\right) + \left(\frac{\partial f_1}{\partial \overline{N_1}}\frac{\partial f_2}{\partial \overline{N_2}} - \frac{\partial f_1}{\partial \overline{N_2}}\frac{\partial f_2}{\partial \overline{N_1}}\right) = 0. \tag{16.3.11}$$

We assume that the equilibrium position $(\overline{N_1}, \overline{N_2})$ is stable in the absence of diffusion so that

$$\left(\frac{\partial f_1}{\partial \overline{N_1}} + \frac{\partial f_2}{\partial \overline{N_2}}\right) < 0, \quad \left(\frac{\partial f_1}{\partial \overline{N_1}}\frac{\partial f_2}{\partial \overline{N_2}} - \frac{\partial f_1}{\partial \overline{N_2}}\frac{\partial f_2}{\partial \overline{N_1}}\right) > 0. \tag{16.3.12}$$

Inequalities (16.3.12) show that the coefficient of λ in (16.3.10) is positive and the constant term in (16.3.10) is also positive if

$$D_1\frac{\partial f_2}{\partial \overline{N_2}} + D_2\frac{\partial f_1}{\partial \overline{N_1}} < 0. \tag{16.3.13}$$

Thus, if (16.3.13) is satisfied, the equilibrium position which is stable in the absence of diffusion remains stable when there is diffusion. In particular, in view of the first inequality in (16.3.12), if the diffusion coefficients are equal, diffusion fails to induce instability. Thus for diffusion-induced instability to occur, it is necessary that D_1 and D_2 should be unequal; but his condition is obviously not sufficient. Even when inequality (16.3.13) is reversed, the constant term in (16.3.10) *may be* (but need not to be) negative, and the equilibrium position may be unstable when there is diffusion. A sufficient condition for diffusion-induced instability is

$$\left(\frac{\partial f_1}{\partial \overline{N_1}}\frac{\partial f_2}{\partial \overline{N_2}} - \frac{\partial f_1}{\partial \overline{N_2}}\frac{\partial f_2}{\partial \overline{N_1}}\right) + D_1D_2\sigma^4 - \sigma^2\left(D_1\frac{\partial f_2}{\partial \overline{N_2}} + D_2\frac{\partial f_1}{\partial \overline{N_1}}\right) < 0 \tag{16.3.14}$$

for some integral values of m, n, p. We may note that the stable equilibrium remain stable in spite of diffusion (16.3.13) is satisfied or if $D_1 = D_2$ or if

$$\frac{\partial f_1}{\partial \overline{N_1}} < 0, \qquad \frac{\partial f_2}{\partial \overline{N_2}} < 0. \tag{16.3.15}$$

16.4 Influence of Diffusion on the Stability of Prey-Predator Models

We now consider the influence of diffusion on the stability of three prey-predator models:

(i) The simplest prey-predator model is

$$\frac{dN_1}{dt} = N_1(a_1 - \alpha_1 N_2), \qquad \frac{dN_2}{dt} = N_2(-a_2 + \alpha_2 N_1) \tag{16.4.1}$$

so that

$$\begin{aligned} &\overline{N_1} = \frac{a_2}{\alpha_2}, \qquad \overline{N_2} = \frac{a_1}{\alpha_1}, \\ &\frac{\partial f_1}{\partial \overline{N_1}} = 0, \quad \frac{\partial f_2}{\partial \overline{N_2}} = 0, \\ &\frac{\partial f_1}{\partial \overline{N_1}} = -\alpha_1 \frac{a_2}{\alpha_2}, \quad \frac{\partial f_2}{\partial \overline{N_1}} = \alpha_2 \frac{a_1}{\alpha_1}, \end{aligned} \tag{16.4.2}$$

Then (16.3.10) and (16.3.11) becomes

$$\lambda^2 + \lambda[(D_1 + D_2)\sigma^2] + a_1 a_2 + \sigma^4 D_1 D_2 = 0, \tag{16.4.3}$$

$$\lambda^2 + a_1 a_2 = 0, \tag{16.4.4}$$

so that the equilibrium is neutral without diffusion and is neutral or stable with diffusion. Thus diffusion may 'increase' stability; at least it does not 'decrease stability'.

(ii) For the more general prey-predator model given by (16.3.1) and (16.3.2), we have

$$\frac{\partial f_1}{\partial \overline{N_1}} \geq 0, \quad \frac{\partial f_2}{\partial \overline{N_2}} \leq 0, \quad \frac{\partial f_1}{\partial \overline{N_2}} < 0, \quad \frac{\partial f_2}{\partial \overline{N_1}} > 0. \tag{16.4.5}$$

Thus, if the equilibrium is stable without diffusion and is unstable with diffusion, we get

$$\left|\frac{\partial f_2}{\partial \overline{N_2}}\right| \geq \left|\frac{\partial f_1}{\partial \overline{N_1}}\right|, \quad D_2 \left|\frac{\partial f_1}{\partial \overline{N_1}}\right| > D_1 \left|\frac{\partial f_2}{\partial \overline{N_2}}\right| \tag{16.4.6}$$

which give $D_1 < D_2$. Thus for diffusion-induced instability, it is necessary that the coefficient of diffusion for prey should be less than the diffusion coefficient for predator. Again, this condition is not sufficient.

(iii) Consider the model which, in the absence of diffusion, is given by

$$\begin{aligned} \frac{dN_1}{dt} &= N_1[f(N_1) - N_2], \\ \frac{dN_2}{dt} &= N_2[N_1 - g(N_2)]. \end{aligned} \tag{16.4.7}$$

Then we have

$$\begin{aligned} &f(\overline{N_1}) = \overline{N_2}, && g(\overline{N_2}) = \overline{N_2}, \\ &\frac{\partial f_1}{\partial N_1} = \overline{N_1} f'(\overline{N_1}), && \frac{\partial f_1}{\partial \overline{N_2}} = -\overline{N_1}, \\ &\frac{\partial f_2}{\partial \overline{N_1}} = \overline{N_2}, && \frac{\partial f_2}{\partial \overline{N_2}} = -\overline{N_2} g'(\overline{N_2}). \end{aligned} \tag{16.4.8}$$

Now, (16.3.10) and (16.3.11) gives

$$\begin{aligned} &\lambda^2 + \lambda[(D_1 + D_2)\sigma^2 - \overline{N_1} f'(\overline{N_1}) + \overline{N_2} g'(\overline{N_2})] + [-\overline{N_1 N_2} f'(\overline{N_1}) g'(\overline{N_2}) + \overline{N_1 N_2}] \\ &\quad + D_1 D_2 \sigma^4 - \sigma^2[D_2 \overline{N_1} f'(\overline{N_1}) - D_1 \overline{N_2} g'(\overline{N_2})] = 0, \end{aligned} \tag{16.4.9}$$

$$\Rightarrow \lambda^2 + \lambda[-\overline{N_1} f'(\overline{N_1}) + \overline{N_2} g'(\overline{N_2})] + \overline{N_1 N_2}[1 - f'(\overline{N_1}) g'(\overline{N_2})] = 0. \tag{16.4.10}$$

When there is no diffusion, the equilibrium is stable if

$$\overline{N_2}g'(\overline{N_2}) - \overline{N_1}f'(\overline{N_1}) > 0, \qquad 1 - f'(\overline{N_1})g'(\overline{N_2}) > 0. \tag{16.4.11}$$

When there is diffusion, the equilibrium can be unstable if

$$D_2\overline{N_1}f'(\overline{N_1}) > D_1\overline{N_2}g'(\overline{N_2}). \tag{16.4.12}$$

By the same reasoning as before, $D_1 < D_2$. If $f'(\overline{N_1}) > 0$, $g'(\overline{N_2}) > 0$, we can find the values of $m,\ n,\ p$ so that the equilibrium with diffusion is unstable. However, if $f'(\overline{N_1}) < 0$, then $g'(\overline{N_2}) > 0$. This is not possible, and the equilibrium continues to be stable.

References

1. K. E. Watt : Ecology and Resource Management-A Quantitative Approach.
2. R. M. May : Stability and Complexity in Model Ecosystem.
3. Y. M. Svirezhev and D. O. Logofet : Stability of Biological Communities.
4. A. Segel : Modelling Dynamic Phenomena in Molecular Biology.
5. J. D. Murray : Mathematical Biology. Springer and Verlag.
6. N. T. J. Bailey : The Mathematical Approach to Biology and Medicine.
7. L. Perko (1991): Differential Equations and Dynamical Systems, Springer Verlag.
8. F. Verhulust (1996): Nonlinear Differential Equations and Dynamical Systems, Springer Verlag.
9. H. I. Freedman - Deterministic Mathematical Models in Population Ecology.
10. Mark Kot (2001): Elements of Mathematical Ecology, Cambridge Univ. Press

POST GRADUATE DEGREE PROGRAMME (CBCS) IN

MATHEMATICS

SEMESTER IV

SELF LEARNING MATERIAL

PAPER : MATO 4.3
(Pure Stream)

Advanced Complex Analysis I

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

Course Preparation Team

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

May, 2020

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, 2017 had been our endeavour. We are happy to have achieved our goal.

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

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

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

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

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

Director

Directorate of Open and Distance Learning

University of Kalyani

Optional Paper

MATO 4.3
Marks : 100 (SEE : 80; IA : 20)

Advanced Complex Analysis I
(Pure Stream)

Syllabus

- **Unit 1:** The functions- $M(r)$ and $A(r)$. Hadamard theorem on the growth of $\log M(r)$
- **Unit 2:** Schwarz inequality, Borel-Caratheodory inequality, Open mapping theorem.
- **Unit 3:** Dirichlet series, abscissa of convergence and abscissa of absolute convergence, their representations in terms of the coefficients of the Dirichlet series.
- **Unit 4:** The Riemann Zeta function, the product development and the zeros of the zeta functions.
- **Unit 5:** Entire functions, growth of an entire function, order and type and their representations in terms of the Taylor coefficients.
- **Unit 6:** Distribution of zeros of entire functions, the exponent of convergence of zeros.
- **Unit 7:** Weierstrass factorization theorem.
- **Unit 8:** Canonical product, Borel's first theorem. Borel's second theorem (statement only), Hadamard's factorization theorem, Schottky's theorem (no proof), Picard's first theorem.
- **Unit 9:** Multiple-valued functions, Riemann surface for the functions $\sqrt{z}$, $\log z$.
- **Unit 10:** Analytic continuation, uniqueness, continuation by the method of power series
- **Unit 11:** Continuation by the method of natural boundary. Existence of singularity on the circle of convergence.
- **Unit 12:** Functions element, germ and complete analytic functions. Monodromy theorem.
- **Unit 13:** Conformal transformations, Riemanns theorems for circle, Schwarz principle of symmetry
- **Unit 14:** Schwarz-Christoffel formula (statement only) with applications.
- **Unit 15:** Univalent functions, general theorems
- **Unit 16:** Sequence of univalent functions, sufficient conditions for univalence.

Contents

Unit 1

Course Structure

- The functions- $M(r)$ and $A(r)$.
- Hadamard theorem on the growth of $\log M(r)$

1.1 Introduction

We have read about the Maximum and Minimum modulus theorems for a non-constant analytic functions on bounded set G. We mainly introduce two new terms, viz., $M(r)$ and $A(r)$, and derive their properties. The main motive is to study the growth of an analytic function f. f, being a complex function, is not comparable with the real functions when we come to measure their orders of growth. So, to be able to measure the order of growth of such functions, we need to define real function and find the desired results with respect to them.

Objectives

After reading this unit, you will be able to

- define $M(r)$ and $A(r)$ of an analytic function
- deduce the properties of them with the help of the maximum modulus theorem
- learn preliminary definitions of convex functions
- deduce the Hadamard's three-circles theorem

1.2 The functions $M(r)$ and $A(r)$

Let us first state the maximum modulus theorem and then we will define the new terms $M(r)$ and $A(r)$ and derive its properties as a consequence of the maximum modulus theorem.

Theorem 1.2.1. If a function f is analytic in a bounded region G, and continuous on $\overline{G}$ and $M = \max\{|f(z)| : z \in \partial G\}$, where ∂G is the boundary of G, then $|f(z)| < M$ in G, unless f is a constant function.

Example 1.2.1. Consider the function $f(z) = z^2$ defined on the closed disc $D = \{z : |z - 1 - i| \leq 1\}$. Let us show that the maximum value of $|f(z)|$ is attained at $z = (1 + 1/\sqrt{2})(1 + i)$. To do this, set

$$z = 1 + i + \mathrm{e}^{i\theta} = (1 + \cos\theta) + i(1 + \sin\theta),\ \theta \in [0, 2\pi).$$

Then $|f(z)| = 3 + 2(\cos\theta + \sin\theta)$. It follows that the maximum value of $|f(z)|$ is attained at $\theta = \pi/4$ and the maximum value is $3 + 2\sqrt{2}$. The maximum value is attained at $z = 1 + i + \mathrm{e}^{i\frac{\pi}{4}}$.

Corollary 1.2.1. Suppose that f is analytic in a bounded region G and continuous on $\overline{G}$. Then, each of $\mathrm{Re} f(z)$, $-\mathrm{Re} f(z)$, $\mathrm{Im} f(z)$ and $-\mathrm{Im} f(z)$ attains its maximum at some point on the boundary ∂G of G.

Proof. Let $u(x, y) = \mathrm{Re}\ f(z)$ and $g(z) = \mathrm{e}^{f(z)}$. By the Maximum Modulus theorem, $|g(z)| = \mathrm{e}^{u(x,y)}$ cannot assume the maximum value in G. Since e^u is maximized when u is maximized, obtain that $u(x, y)$ cannot attain its maximum value in G. Similarly, the other cases can be proved. □

The minimum modulus theorem comes as a direct corollary of the above theorem which is

Theorem 1.2.2. Let f be a non-constant analytic function in a bounded region G and continuous on $\overline{G}$. If $f(z) \neq 0$ inside ∂G, then $|f(z)|$ must attain its minimum value on ∂G.

Example 1.2.2. Suppose that f and g are analytic on the closed unit disc $|z| \leq 1$ such that

1. $|f(z)| \leq M$ for all $|z| \leq 1$;
2. $f(z) = z^n g(z)$ for all $|z| \leq 1/3$ and for some $n \in \mathbb{N}$.

We wish to use the Maximum modulus theorem to find the maximum value of $|f(z)|$ on $|z| \leq 1/3$. To do this, we proceed as follows. On $|z| = 1$, we have

$$m \geq |f(z)| = |z^n g(z)| = |g(z)|$$

and so, $|g(z)| \leq M$ for $|z| \leq 1$. Now, for $|z| \leq 1/3$, we have,

$$|f(z)| = |z^n g(z)| = |z^n||g(z)| = 3^{-n}|g(z)| \leq 3^{-n}M.$$

Thus, $|f(z)| \leq 3^{-n}M$ for all $|z| \leq 1/3$.

The hypothesis that G is bounded can't be dropped however as we see in the following example.

Example 1.2.3. Define $f(z) = \mathrm{e}^{-iz}$ on $G = \{z : \mathrm{Im}\ z > 0\}$. Then $|f(z)| = 1$ on the boundary $\partial G = \{z : \mathrm{Im}\ z = 0\}$, that is, the real axis. But, for $z = x + iy \in G$, we have,

$$|f(x + iy)| = \mathrm{e}^y \to \infty \ \text{ as } \ y \to +\infty;$$

that is, f itself is not bounded. And the Maximum modulus theorem fails.

We will now define the terms $M(r)$ and $A(r)$ related to an analytic function f as follows

Definition 1.2.1. Let f be a non-constant analytic function defined in $|z| \leq R$. Then, for $0 \leq r < R$, we define

1. $M(r) = \max\{|f(z)| : |z| = r\}$; and
2. $A(r) = \max\{\mathrm{Re}\ f(z) : |z| = r\}$.

Theorem 1.2.3. Let f be a non-constant analytic function defined in $|z| \leq R$. Then, $0 \leq r < R$

1. $M(r)$ is a strictly increasing function of r;

2. $A(r)$ is a strictly increasing function of r.

Proof. 1. Let $0 \leq r_1 < r_2 < R$. Since f is analytic in $|z| \leq r_2$, the maximum value of $|f(z)|$ for $|z| \leq r_2$ is attained on $|z| = r_2$. Let z_2 be a point on $|z| = r_2$ such that $|f(z_2)| = M(r_2)$. Similarly, the maximum value of $|f(z)|$ for $|z| \leq r_1$ is attained on $|z| = r_1$. Let z_1 be a point on $|z| = r_1$ such that $|f(z_1)| = M(r_1)$. Since z_1 is an interior point of the closed region $|z| \leq r_2$, hence by maximum modulus theorem, $|f(z_1)| < M(r_2)$, that is, $M(r_1) < M(r_1)$. This completes the proof of the first part of the theorem.

2. Let $0 \leq r_1 < r_2 < R$. Since f is analytic in $|z| \leq r_2$, the maximum value of Re $f(z)$ for $|z| \leq r_2$ is attained on $|z| = r_2$. Let z_2 be a point on $|z| = r_2$ such that Re $f(z_2) = A(r_2)$. Similarly, the maximum value of Re $f(z)$ for $|z| \leq r_1$ is attained on $|z| = r_1$. Let z_1 be a point on $|z| = r_1$ such that Re $f(z_1) = A(r_1)$. Since z_1 is an interior point of the closed region $|z| \leq r_2$, hence by corollary 1.2.1, Re $f(z_1) < A(r_2)$, that is, $A(r_1) < A(r_2)$.

□

1.3 Hadamard's theorem on the growth of $\log M(r)$

Definition 1.3.1. Let $[a, b]$ be an interval in $\mathbb{R}$. A function $f : [a, b] \to \mathbb{R}$ is said to be **convex** if for any two points $x_1, x_2 \in [a, b]$,

$$f(tx_2 + (1-t)x_1) \leq f(x_2) + (1-t)f(x_1) \quad \text{for} \quad 0 \leq t \leq 1.$$

Geometrically, $f(x)$ is said to be convex downwards, or simply convex in $[a, b]$ if the curve $y = f(x)$ between any two points x_1 and x_2 in $[a, b]$ always lies below the chord joining the points $(x_1, f(x_1))$ and $(x_2, f(x_2))$.

1.3.1 Analytical condition for convexity

Consider the figure 1.1.

Let $y = f(x)$ be the curve and $(x_1, f(x_1))$, $(x_2, f(x_2))$ be two points as shown in fig. 1.1. Let $x = tx_1 + (1-t)x_2$ be any point between x_1 and x_2 and $0 \leq t \leq 1$. Then, $x_1 \leq x \leq x_2$ and the equation of the chord joining $(x_1, f(x_1))$ and $(x_2, f(x_2))$ is

$$\begin{aligned} y - f(x_1) &= \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1) \\ \text{or,} \qquad y &= f(x_1) + \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1). \end{aligned}$$

Let the coordinates of any point on the chord joining $(x_1, f(x_1))$ and $(x_2, f(x_2))$ be (x, y). According to the definition of convexity, $f(x)$ will be convex if and only if $f(x) \leq y$, that is,

$$f(x) \leq f(x_1) + \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1)$$

that is, if and only if,

$$f(x) \leq \frac{x_2 - x}{x_2 - x_1} f(x_1) + \frac{x - x_1}{x_2 - x_1} f(x_2) \tag{1.3.1}$$

holds for all $x_1 \leq x \leq x_2$.

Let us check through some quick results on convex functions that we will need in the sequel.

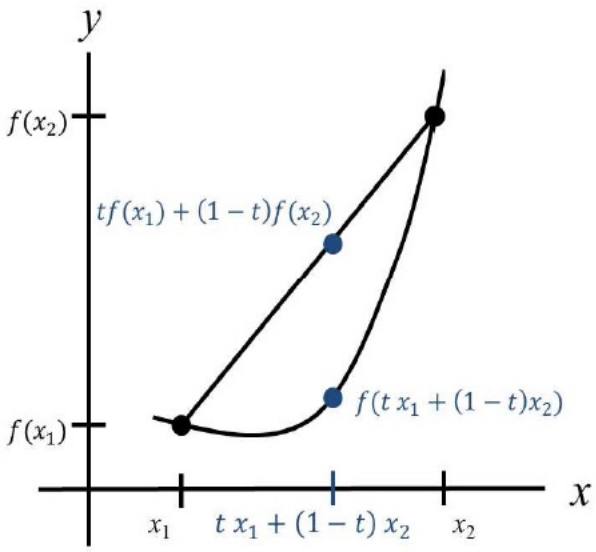


Figure 1.1

Result 1.3.1. 1. A differentiable function f on $[a, b]$ is convex if and only if f' is increasing.

2. A sufficient condition for f to be convex is that, $f''(x) \geq 0$.

3. If $f : (a, b) \to \mathbb{R}$ is convex, then it is continuous there.

We now state the Hadamard's three circles theorem.

Theorem 1.3.1. (Hadamard's three-circles theorem) Let f be analytic on the closed annulus $0 < r_1 \leq |z| \leq r_3$ (see fig. 1.2). If $r_1 < r_2 < r_3$, then

$$M(r_2)^{\log\left(\frac{r_3}{r_1}\right)} \leq M(r_1)^{\log\left(\frac{r_3}{r_2}\right)} M(r_3)^{\log\left(\frac{r_2}{r_1}\right)}.$$

Proof. Let $\phi(z) = z^\lambda f(z)$, where λ is a real constant to be chosen later. If λ is not an integer, $\phi(z)$ is multi-valued in $r_1 \leq |z| \leq r_3$. So we cut the annulus along the negative part of the real axis obtaining a simply connected region G in which the principal branch of ϕ is analytic.

The maximum modulus of this branch of ϕ in G, that is, the cut-annulus is obtained on the boundary of G. Since λ is real, all the branches of ϕ have the same modulus. By considering another branch of ϕ which is analytic in another cut-annulus obtained by using a different cut, it is clear that the principal branch of ϕ must attain its maximum modulus on at least one of the bounding circles of the annulus. Thus, $|\phi(z)| \leq \max\{r_1^\lambda M(r_1), r_3^\lambda M(r_3)\}$. Hence, on $|z| = r_2$, we have,

$$r_2^\lambda M(r_2) \leq \max\{r_1^\lambda M(r_1), r_3^\lambda M(r_3)\}. \tag{1.3.2}$$

We now choose λ such that

$$r_1^\lambda M(r_1) = r_3^\lambda M(r_3);$$

that is,

$$\lambda = -\frac{\log\left(\frac{M(r_3)}{M(r_1)}\right)}{\log\left(\frac{r_3}{r_1}\right)}.$$

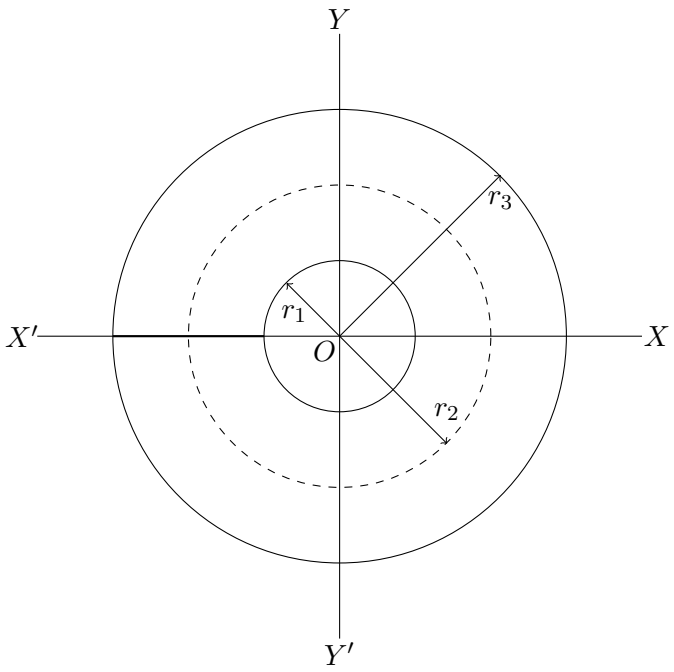


Figure 1.2: Hadamard's Three-Circles Theorem

With this λ, we get from (1.3.2), and the fact that $a^{\log b} = b^{\log a}$ holds for all positive real numbers a and b, we get

$$\begin{aligned}
& M(r_2) \leq \left(\frac{r_1}{r_2}\right)^{\lambda} M(r_1) \\
\Rightarrow \quad & M(r_2)^{\log\left(\frac{r_3}{r_1}\right)} \leq \left(\frac{r_1}{r_2}\right)^{\log\left(\frac{M(r_1)}{M(r_3)}\right)} \cdot M(r_1)^{\log\left(\frac{r_3}{r_1}\right)} \\
\Rightarrow \quad & M(r_2)^{\log\left(\frac{r_3}{r_1}\right)} \leq \left(\frac{M(r_1)}{M(r_3)}\right)^{\log\left(\frac{r_1}{r_2}\right)} \cdot M(r_1)^{\log\left(\frac{r_3}{r_1}\right)} \\
\Rightarrow \quad & M(r_2)^{\log\left(\frac{r_3}{r_1}\right)} \leq M(r_1)^{\log\left(\frac{r_3}{r_2}\right)} \cdot M(r_3)^{\log\left(\frac{r_2}{r_1}\right)}.
\end{aligned}$$

□

Note 1.3.1. The equality in the above theorem can be achieved when $\phi(z)$ is constant, that is, $f(z)$ is of the form $f(z) = kz^{\lambda}$, for some real λ, k being a constant.

Theorem 1.3.2. Let $f(z) = \sum_{n=0}^{\infty} a_n z^n$ be analytic in $|z| \leq r$. Then

$$|a_n| r^n \leq \max\{4A(r), 0\} - 2\text{Re } f(0), \ \forall n > 0.$$

Proof. Let $z = r\,\mathrm{e}^{i\theta}$, $f(z) = \sum_{n=0}^{\infty} a_n z^n = u(r,\theta) + iv(r,\theta)$ and $a_n = \alpha_n + i\beta_n$. Thus,

$$\begin{aligned} u(r,\theta) + iv(r,\theta) &= \sum_{n=0}^{\infty} (\alpha_n + i\beta_n) r^n\,\mathrm{e}^{in\theta} \\ &= \sum_{n=0}^{\infty} (\alpha_n + i\beta_n) r^n (\cos n\theta + i \sin n\theta) \\ &= \sum_{n=0}^{\infty} r^n \{(\alpha_n \cos n\theta - \beta_n \sin n\theta) + i(\alpha_n \sin n\theta + \beta_n \cos n\theta)\}. \end{aligned}$$

Equating real parts, we get,

$$u(r,\theta) = \sum_{n=0}^{\infty} r^n (\alpha_n \cos n\theta - \beta_n \sin n\theta). \tag{1.3.3}$$

The series (1.3.3) converges uniformly with respect to θ. Hence we may multiply it by $\sin n\theta$ or $\cos n\theta$ and integrate it term by term. Now,

$$u(r,\theta) = \alpha_0 + (\alpha_1 \cos\theta - \beta_1 \sin\theta) r + \cdots + (\alpha_n \cos n\theta - \beta_n \sin n\theta) r^n + \cdots$$

Hence,

$$\int_0^{2\pi} u(r,\theta)\,d\theta = \alpha_0 2\pi \Rightarrow \alpha_0 = \frac{1}{2\pi} \int_0^{2\pi} u(r,\theta)\,d\theta.$$

Also,

$$u(r,\theta)\cos n\theta = \alpha_0 \cos n\theta + (\alpha_1 \cos\theta \cos n\theta - \beta_1 \sin\theta \cos n\theta) r + \cdots + (\alpha_n \cos^2 n\theta - \beta_n \sin n\theta \cos n\theta) r^n + \cdots$$

So,

$$\int_0^{2\pi} u(r,\theta) \cos n\theta\,d\theta = \alpha_n r^n \int_0^{2\pi} \cos^2 n\theta\,d\theta = \pi \alpha_n r^n.$$

Hence, for $n > 0$,

$$\alpha_n r^n = \frac{1}{\pi} \int_0^{2\pi} u(r,\theta) \cos n\theta\,d\theta.$$

Similarly, multiplying (1.3.3) by $\sin n\theta$ and integrating term by term from 0 to 2π, we get,

$$-\beta_n r^n = \frac{1}{\pi} \int_0^{2\pi} u(r,\theta) \sin n\theta\,d\theta, \quad \text{for} \quad n > 0.$$

Hence,

$$\begin{aligned} a_n r^n = (\alpha_n + i\beta_n) r^n &= \frac{1}{\pi} \int_0^{2\pi} u(r,\theta) \cos n\theta\,d\theta - \frac{1}{\pi} \int_0^{2\pi} u(r,\theta) \sin n\theta\,d\theta \\ &= \frac{1}{\pi} \int_0^{2\pi} u(r,\theta)\,\mathrm{e}^{-in\theta}\,d\theta, \ \ n > 0. \end{aligned}$$

Thus,

$$|a_n| r^n \le \frac{1}{\pi} \int_0^{2\pi} |u(r,\theta)|\,d\theta, \ \ n > 0.$$

Hence,

$$|a_n| r^n + 2\alpha_0 = \frac{1}{\pi} \int_0^{2\pi} \{|u(r,\theta)| + u(r,\theta)\}\,d\theta. \tag{1.3.4}$$

Now, $|u(r,\theta)| + u(r,\theta) = 0$ when $u(r,\theta) < 0$. Hence if $A(r) < 0$, the right hand side of (1.3.4) is 0. Again, if $A(r) \geq 0$, the right hand side of (1.3.4) does not exceed

$$\frac{1}{\pi}\int_0^{2\pi} 2A(r)d\theta = 4A(r).$$

Thus,

$$\begin{aligned} & |a_n|r^n + 2\alpha_0 \leq \max\{4A(r), 0\} \\ \Rightarrow \quad & |a_n|r^n + 2\text{Re } f(0) \leq \max\{4A(r), 0\} \\ \Rightarrow \quad & |a_n|r^n \leq \max\{4A(r), 0\} - 2\text{Re } f(0). \end{aligned}$$

□

1.4 Few Probable Questions

1. State the Maximum modulus theorem. Show that $M(r)$ is an increasing function of r.
2. Show that for any non-constant analytic function f defined on a bounded region G and continuous on $\overline{G}$, Re $f(z)$ attains its maximum at some point on the boundary ∂G of G. Show that $A(r)$ is an increasing function of r.
3. State and prove the Hadamard's three-circles theorem.

Unit 2

Course Structure

- Schwarz Lemma, Open mapping theorem.
- Borel-Caratheodory inequality

2.1 Introduction

Let us denote $\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\}$. In this unit, we start with a simple but one of the classical theorems in complex analysis, named Schwarz' Lemma, which states that if f is analytic and satisfies $|f(z)| < 1$ in $\mathbb{D}$ and $f(0) = 0$, then $|f(z)| \leq |z|$ for each $z \in \mathbb{D}$ with equality sign if and only if f has the form $f(z) = \mathrm{e}^{i\alpha} z$, for some $\alpha \in \mathbb{R}$. Furthermore, $|f'(0)| \leq 1$ with the equality if and only if f has the form as stated previously. This result has important role in the proof of Riemann mapping theorem.

This unit also deals with deducing the Open mapping theorem and the Borel-Caratheodory theorem.

Objectives

After reading this unit, you will be able to

- deduce Schwarz lemma and its various variants; also applying it in various problems
- deduce the Borel Caratheodory inequality from Schwarz lemma and discuss some of its consequences
- derive the Open mapping theorem

2.2 Schwarz Lemma

Let us now begin with a sharp version of the classical Schwarz lemma.

Theorem 2.2.1. (Schwarz Lemma) Let $f : \mathbb{D} \to \overline{\mathbb{D}}$ be analytic having a zero of order n at the origin. Then

1. $|f(z)| \leq |z|^n$ for all $z \in \mathbb{D}$,
2. $|f^{(n)}(0)| \leq n!$

and the equality holds either in 1 for some point $0 \neq z_0 \in \mathbb{D}$ or in 2 occurs if and only if $f(z) = \epsilon z^n$ with $|\epsilon| = 1$.

Proof. Let $f : \mathbb{D} \to \overline{\mathbb{D}}$ be analytic on $\mathbb{D}$ and has nth order zero at the origin. Then we have,

$$f(0) = 0 = f'(0) = \cdots = f^{(n-1)}(0) \quad \text{and} \quad f^{(n)}(0) \neq 0.$$

So we can write

$$f(z) = \sum_{k=n}^{\infty} a_k z^k = z^n g(z), \quad \text{for} \quad z \in \mathbb{D},$$

where

$$a_k = \frac{f^{(k)}(0)}{k!} \quad \text{and} \quad g(z) = \sum_{k=n}^{\infty} a_k z^{k-n}.$$

The function $g(z) = f(z)/z^n$ has a removable singularity at the origin so that if

$$\begin{aligned} g(z) &= z^{-n} f(z) \quad \text{for } z \in \mathbb{D} \setminus \{0\} \\ &= a_n \qquad \text{for } z = 0. \end{aligned}$$

then g is analytic in $\mathbb{D} \setminus \{0\}$ and continuous in $\mathbb{D}$. Since we will have by Cauchy's theorem for a disc, $\int_C g(z)dz = 0$ for all closed contours C inside $\mathbb{D}$ so by Morera's theorem, g is analytic on $\mathbb{D}$.

We claim that $|g(z)| \leq 1$ for all $z \in \mathbb{D}$. Now, for $0 < r < 1$, g is analytic on the bounded domain $\mathbb{D}_r = \{z : |z| < r\}$ and g is continuous on the closure of $\mathbb{D}_r$. Thus, the maximum modulus theorem is applicable here. As $|f(z)| \leq 1$ for all $z \in \mathbb{D}$, it follows that for $|z| = r$,

$$|g(z)| = \frac{|f(z)|}{|z|^n} \leq \frac{1}{r^n}.$$

By Maximum modulus theorem, $|g(z)| \leq r^{-n}$, for all z with $|z| \leq r$. Since r is arbitrary, by letting $r \to 1$, we find that $|g(z)| \leq 1$, that is,

$$|g(z)| \leq 1 \quad \text{for all} \quad z \in \mathbb{D} \tag{2.2.1}$$

and this implies that

$$|f(z)| \leq |z|^n \quad \text{for all} \quad z \in \mathbb{D}.$$

Equality in 1 holds for some point z_0 in $\mathbb{D} \setminus \{0\}$ implies that $|g(z_0)| = 1$. It follows that g achieves its maximum modulus at an interior point z_0. Consequently, by the Maximum modulus theorem, g must reduce to a constant, say ϵ. Then, $f(z) = \epsilon z^n$, where $|\epsilon| = 1$.

Also, note that $|g(z)| \leq 1$ throughout the disc $\mathbb{D}$. Since $|a_n| = |g(0)|$, we get by equation (2.2.1), $|g(0)| \leq 1$ so, we get $|f^{(n)}(0)|/n! \leq 1$ and hence, 2 follows.

Again, if $|f^{(n)}(0)| = n!$ then $|g(0)| = 1$ showing that g achieves its maximum modulus 1 at some interior point $'0'$. So, g is a constant function of absolute value 1 and as before, it means that $f(z) = \epsilon z^n$ with $|\epsilon| = 1$. □

Remark 2.2.1. Note that the case $n = 1$ of the previous theorem is the original Schwarz lemma state in the beginning of this unit.

For example, if f is an analytic function over $\mathbb{D}$ with $|f(z)| \leq 1$ and $f(0) = 0$, then what kind of function is f if $f(1/2) = 1/2$? It must be none other than the identity function since the equality in the preceding theorem holds with $n = 1$ and $z = 1/2 \in \mathbb{D}$.

Corollary 2.2.1. If f is analytic and satisfies $|f(z)| \leq M$ in $B(a; R)$ and $f(a) = 0$, then

1. $|f(z)| \leq M|z-a|/R$ for every $z \in B(a;R)$,
2. $|f'(a)| \leq M/R$

with the equality sign if and only if f has the form $f(z) = M\epsilon(z-a)/R$, for some constant ϵ with $|\epsilon| \leq 1$.

Proof. Use Schwarz lemma with $g(z) = f(Rz+a)/M$, $|z| < 1$. □

Another generalisation of the above theorem is as follows

Corollary 2.2.2. If f is analytic and satisfies $|f(z)| \leq M$ in $B(a;R)$ and a is a zero of f of order n. Then

1. $|f(z)| \leq M|z-a|^n/R^n$ for every $z \in B(a;R)$,
2. $|f^{(n)}(a)| \leq M/R^n$

with the equality sign if and only if f has the form $f(z) = M\epsilon(z-a)^n/R^n$, for some constant ϵ with $|\epsilon| \leq 1$.

Proof. Prove the corollary independently without using the Schwarz lemma. □

Does the Schwarz lemma hold for real-valued functions? Consider the function

$$u(x) = \frac{2x}{x^2+1}.$$

Then u is infinitely differentiable on $\mathbb{R}$. In particular, $u'(x)$ is continuous on $[-1,1]$, $u(0) = 0$ and $|u(x)| \leq 1$. But $|u(x)| > |x|$ for $0 < |x| < 1$.

Example 2.2.1. Let $\omega = \mathrm{e}^{2\pi i/n}$ be the nth root of unity, where $n \in \mathbb{N}$ is fixed. Suppose that $f : \mathbb{D} \to \mathbb{D}$ is analytic such that $f(0) = 0$. We wish to apply Schwarz lemma to show that

$$|f(z) + f(\omega z) + f(\omega^2 z) + \cdots + f(\omega^{n-1} z)| \leq n|z|^n \tag{2.2.2}$$

and the equality for some point $0 \neq z_0 \in \mathbb{D}$ occurs if and only if $f(z) = \epsilon z^n$ with $|\epsilon| = 1$. To do this, we define $F : \mathbb{D} \to \mathbb{D}$ by

$$F(z) = \frac{1}{n}\sum_{k=0}^{n-1} f(\omega^k z).$$

Clearly F is analytic on $\mathbb{D}$, $F(0) = 0$ and, for $1 \leq m \leq n-1$,

$$F^{(m)}(z) = \frac{1}{n}\sum_{k=0}^{n-1} (\omega^k)^m f^{(m)}(\omega^k z)$$

so that (as $\omega^n = 1$)

$$F^{(m)}(0) = \frac{1}{n}\sum_{k=0}^{n-1} (\omega^k)^m f^{(m)}(0) = \frac{f^{(m)}(0)}{n}\left(\frac{1-(\omega^m)^n}{1-\omega^m}\right) = 0.$$

By Schwarz lemma, it follows that $|F(z)| \leq |z|^n$ for all $z \in \mathbb{D}$ which is the same as (2.2.2). The equality in this inequality for some point $z_0 \neq 0$ occurs if and only if $F(z) = \epsilon z^n$ with $|\epsilon| = 1$, or equivalently,

$$\sum_{k=0}^{n-1} [f(\omega^k z) - \epsilon z^n] = 0. \tag{2.2.3}$$

We claim that the above equation implies that $f(z) = \epsilon z^n$. If we let $f(z) = \sum_{m=1}^{\infty} a_m z^m$, then (2.2.3) becomes

$$\sum_{m=1}^{\infty} a_m \left(\sum_{k=0}^{n-1} \omega^{km} \right) z^m = n\epsilon z^n.$$

In view of the identity

$$\begin{aligned} \sum_{k=0}^{n-1} \omega^{km} &= n \text{ if } m \text{ is a multiple of } n \\ &= 0 \text{ otherwise} \end{aligned}$$

the last equation implies that $a_n = \epsilon$ and $a_{2n} = a_{3n} = \cdots = 0$. On the other hand, as $|f(z)| \leq 1$ on $\mathbb{D}$ and $|a_n| = 1$, we have

$$\lim_{r \to 1^-} \frac{1}{2\pi} \int_0^{2\pi} |f(r\,\mathrm{e}^{i\theta})|^2 d\theta = \sum_{m=1}^{\infty} |a_m|^2 \leq 1$$

which shows that all the Taylor's coefficients of f (except a_n) must vanish and so, $f(z) = \mathrm{e}^{i\theta}\, z^n$.

2.2.1 Borel-Caratheodory theorem

Borel-Caratheodory theorem is an important theorem that establishes the relationship between $M(r)$ and $A(r)$. We deduce this with the help of Schwarz theorem.

Theorem 2.2.2. (Borel Caratheodory theorem) Let f be analytic on $D : |z| \leq R$ and $M(r)$ and $A(r)$ are as we defined in the previous unit. Then for $0 < r < R$,

$$M(r) \leq \frac{2r}{R-r} A(R) + \frac{R+r}{R-r} |f(0)|.$$

Proof. We consider the following cases.

Case I. When $f(z) =$ constant $= a + ib$, where a and b are real constants. Then $M(r) = \sqrt{a^2+b^2}$, and $|f(0)| = \sqrt{a^2+b^2}$, and $A(R) = a$. Now,

$$\begin{aligned} \frac{2r}{R-r} A(R) + \frac{R+r}{R-r} |f(0)| - M(r) &= \frac{2r}{R-r} a + \left(\frac{R+r}{R-r} - 1 \right) \sqrt{a^2+b^2} \\ &= \frac{2r}{R-r} (a + \sqrt{a^2+b^2}) \geq 0. \end{aligned}$$

Hence,

$$M(r) \leq \frac{2r}{R-r} A(R) + \frac{R+r}{R-r} |f(0)|.$$

Case II. When $f(z) \neq$ constant and $f(0) = 0$. Then $A(R) > A(0) = 0$, since $A(r)$ is an increasing function of r. Let,

$$g(z) = \frac{f(z)}{2A(R) - f(z)}. \tag{2.2.4}$$

$2A(R) - f(z) \neq 0$ for all $z \in D$, since the real part of $2A(R) - f(z)$ does not vanish in D and $g(0) = 0$. Let $f(z) = u + iv$. Then

$$|g(z)|^2 = \frac{u^2+v^2}{(2A(R)-u)^2+v^2} \leq 1, \;\; z \in D,$$

since $2A(R) - u \geq u$. Hence by Schwarz lemma,

$$|g(z)| \leq \frac{r}{R} \quad \text{for} \quad |z| = r < R.$$

From (2.2.4), we get,

$$\begin{aligned} f(z) &= 2A(R)g(z) - f(z)g(z) \\ \text{or} \quad , \quad f(z)(1 + g(z)) &= 2A(R)g(z). \end{aligned}$$

Hence,

$$|f(z)| = \left|\frac{2A(R)g(z)}{1+g(z)}\right| \leq \frac{2A(R)|g(z)|}{1-|g(z)|} \leq \frac{2A(R)\frac{r}{R}}{1-\frac{r}{R}} = \frac{2r}{R-r}A(R),$$

for $|z| = r < R$. Thus, $M(r) \leq \frac{2r}{R-r}A(R)$ and

$$M(r) \leq \frac{2r}{R-r}A(R) + \frac{R+r}{R-r}|f(0)|.$$

Case III. When $f(z) \neq$ constant and $f(0) \neq 0$. Let $h(z) = f(z) - f(0)$. Then $h(0) = 0$. So, by case II, we have

$$\max\{|h(z)| : |z| = r\} \leq \frac{2r}{R-r} \max\{\text{Re } h(z) : |z| = R\}. \tag{2.2.5}$$

Now,

$$\begin{aligned} \max\{|h(z)| : |z| = r\} &= \max\{|f(z) - f(0)| : |z| = r\} \\ &\geq \max\{|f(z)| : |z| = r\} - |f(0)| \\ &= M(r) - |f(0)|, \end{aligned}$$

and

$$\begin{aligned} \max\{\text{Re } h(z) : |z| = R\} &= \max\{\text{Re } (f(z) - f(0)) : |z| = R\} \\ &= \max\{(\text{Re } f(z) - \text{Re } f(0)) : |z| = R\} \\ &\leq \max\{\text{Re } f(z) : |z| = R\} + |f(0)| \\ &= A(R) + |f(0)|. \end{aligned}$$

Hence from (2.2.5), we get,

$$\begin{aligned} M(r) - |f(0)| &\leq \frac{2r}{R-r}\left(A(R) + |f(0)|\right) \\ \text{or,} \quad M(r) &\leq \frac{2r}{R-r}A(R) + \frac{R+r}{R-r}|f(0)|. \end{aligned}$$

□

Corollary 2.2.3. If $A(R) \geq 0$, then

$$M(r) \leq \frac{2r}{R-r}A(R) + \frac{R+r}{R-r}|f(0)| \leq \frac{R+r}{R-r}\left(A(R) + |f(0)|\right)$$

$\left[\text{since } \frac{2r}{R-r} < \frac{r+r}{R-r} \text{ as } R + r > 2r\right]$.

Corollary 2.2.4. If $A(R) \geq 0$, then

$$\max\{|f^{(n)}(z)| : |z| = r\} \leq \frac{2^{n+2} \cdot n! R}{(R-r)^{n+1}}(A(R) + |f(0)|).$$

Proof. By Cauchy's Integral formula for derivatives, we have,

$$f^{(n)}(z) = \frac{n!}{2\pi i} \int_\gamma \frac{f(t)dt}{(t-z)^{n+1}} \tag{2.2.6}$$

where, $\gamma : |t - z| = \delta = (R - r)/2$.

On γ,

$$|t| = |t - z + z| \leq |t - z| + |z| = \frac{1}{2}(R - r) + r = \frac{1}{2}(R + r) < R,$$

which ensures that γ lies within $|z| = R$. By Borel Caratheodory theorem, we have,

$$\begin{aligned} \max |f(t)| &\leq \frac{R + \frac{1}{2}(R+r)}{R - \frac{1}{2}(R+r)}(A(R) + |f(0)|) \\ &= \frac{3R + r}{R - r}(A(R) + |f(0)|) \\ &< \frac{4R}{R-r}(A(R) + |f(0)|). \end{aligned}$$

Hence from (2.2.6),

$$\begin{aligned} |f^{(n)}(z)| &= \frac{n!}{2\pi} \left| \int_\gamma \frac{f(t)dt}{(t-z)^{n+1}} \right| \\ &\leq \frac{n!}{2\pi\delta^{n+1}} \frac{4R}{R-r}(A(R) + |f(0)|) \cdot 2\pi\delta \\ &= \frac{n!}{\delta^n} \cdot \frac{4R}{R-r}(A(R) + |f(0)|) \\ &= \frac{2^{n+2} \cdot n! R}{(R-r)^{n+1}}(A(R) + |f(0)|). \end{aligned}$$

Hence,

$$\max\{|f^{(n)}(z)| : |z| = r\} \leq \frac{2^{n+2} \cdot n! R}{(R-r)^{n+1}}(A(R) + |f(0)|).$$

□

Exercise 2.2.1. If f is an analytic function defined on $\mathbb{D}$ such that $|f(z)| < 1$ for all z in $\mathbb{D}$ and f fixes two distinct points of $\mathbb{D}$, then show that f is the identity function.

2.3 Open Mapping Theorem

In this section, we are interested in functions that send open sets to open sets. A function f defined on an open set U is said to be an **open mapping** if for every open subset V of U, the image $f(V)$ is open. Consider the mappings $f, g, h : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = x^2, \quad g(x) = \sin x, \quad h(x) = \frac{\mathrm{e}^x + \mathrm{e}^{-x}}{2},$$

respectively. Clearly,

$$f(\mathbb{R}) = [0, \infty), \quad g(\mathbb{R}) = (0, 1], \quad h(\mathbb{R}) = [1, \infty)$$

showing that each of f, g and h are not open mappings. Each of the above functions are infinitely differentiable non-constant real valued functions defined on the real line. Thus, the above examples show that the following theorem does not hold for real line $\mathbb{R}$. Let us now state the open mapping theorem.

Theorem 2.3.1. Let G be a region and suppose that f is a non-constant analytic function on G. Then for any open set U in G, $f(U)$ is open.

Proof. Let $U \subset G$ be open. To show that $f(U)$ is open, we show that for each $a \in U$, $\exists\ \delta > 0$ such that the open ball $B(f(a); \delta) \subset f(U)$. Let $\phi(z) = f(z) - f(a)$. Then a is a zero of ϕ. Since the zeros of a non-constant analytic functions are isolated points, so there exists an open ball $B(a; r)$ with $\overline{B}(a; r) \subset U$ such that $\phi(z) \neq 0$ in $0 < |z - a| < r$. In particular, $\phi(\alpha) \neq 0$ for $\alpha \in \partial B(a; \rho)$ where $\rho < r$.

Let

$$2\delta = \min\{|\phi(\alpha)| :\ \alpha \in \partial B(a; \rho)\}.$$

Then $\delta > 0$. Now, for any $w \in B(f(a); \delta)$, we have,

$$\begin{aligned} |f(\alpha) - w| &\geq |f(\alpha) - f(a)| - |f(a) - w| \\ &= |\phi(\alpha)| - |f(a) - w| \\ &> 2\delta - \delta \\ &= \delta \\ &> |f(a) - w|, \end{aligned}$$

for all $\alpha \in \partial B(a; \rho)$. This implies that

$$\min\{|f(\alpha) - w| :\ \alpha \in \partial B(a; \rho)\} > |f(a) - w|. \tag{2.3.1}$$

Let $F(z) = f(z) - w$. Then F has a zero in $B(a; \rho)$. For, if $F(z) \neq 0$ in $B(a; \rho)$, there exists a nbd $N(a)$ of a containing $\overline{B}(a; \rho)$ lying in G such that $F(z) \neq 0$ in $N(a)$. Then $1/F(z)$ will be analytic in $N(a)$ and

$$\left|\frac{1}{F(a)}\right| < \max\left\{\left|\frac{1}{F(\alpha)}\right| :\ \alpha \in \partial B(a; \rho)\right\} = \frac{1}{\min\{|F(\alpha)| :\ \alpha \in \partial B(a; \rho)\}}$$

that is,

$$\min\{|f(\alpha) - w| :\ \alpha \in \partial B(a; \rho)\} < |f(a) - w|,$$

which contradicts (2.3.1). Hence $\exists\ z_0 \in B(a; \rho)$ such that $f(z_0) = w$. Since w is an arbitrary point of $B(f(a); \delta)$, it follows that $B(f(a); \delta) \subset f(U)$, and hence the theorem. □

2.4 Few Probable Questions

1. State and prove Schwarz lemma.
2. State and prove Borel-Caratheodory theorem.
3. State and prove the Open Mapping theorem.

Unit 3

Course Structure

- Dirichlet series, abscissa of convergence and abscissa of absolute convergence,
- Their representations in terms of the coefficients of the Dirichlet series.

3.1 Introduction

In mathematics, a Dirichlet series is any series of the form

$$f(s) = \sum_{n=1}^{\infty} \frac{a_n}{n^s}$$

where s is a complex number, and a_n is a complex sequence. It is a special case of general Dirichlet series.

Dirichlet' s series were, as their name implies, first introduced into analysis by Dirichlet, primarily with a view to applications in the theory of numbers. A number of important theorems concerning them were proved by Dedekind, and incorporated by him in his later editions of Dirichlet's Vorlesungen uber Zahlentheorle. Dirichlet and Dedekind, however, considered only real values of the variable s. The first theorems involving complex values of s are due to Jensen, who determined the nature of the region of convergence of the general series; and the first attempt to construct a systematic theory of the function $f(s)$ was made by Cahent in a memoir which, although much of the analysis which it contains is open to serious criticism, has served and possibly just for that reason as the starting point of most of the later researches in the subject. We will however, not go into a very vigorous treatment of the subject. We will mainly concern ourselves with the preliminaries of Dirichlet series and gain some idea about their convergence.

Objectives

After reading this unit, you will be able to

- define general and ordinary Dirichlet's series and its examples
- deduce various conditions for convergence of Dirichlet's series
- define certain terms related to the convergence of Dirichlet's series and deduce certain properties

3.2 Dirichlet Series

We formally define the Dirichlet series as follows.

Definition 3.2.1. The series of the form

$$f(s) = \sum_{n=1}^{\infty} a_n \, \mathrm{e}^{-\lambda_n s} \tag{3.2.1}$$

where, $\{\lambda_n\}$ is an increasing sequence of real numbers whose limit is infinity, and $s = \sigma + it$ is a complex variable, whose real and imaginary parts are σ and t respectively. Such a series is called a Dirichlet's series of type λ_n. If $\lambda_n = n$, then (3.2.1) is a power series in e^{-n}. If $\lambda_n = \log n$, then (3.2.1) becomes

$$f(s) = \sum_{n=1}^{\infty} a_n n^{-s} \tag{3.2.2}$$

is called an ordinary Dirichlet's series. In this unit, we will mainly deal with the ordinary Dirichlet's series.

It is clear that all but a finite number of the numbers λ_n must be positive. It is often convenient to suppose that they are all positive, or at any rate $\lambda_1 \geq 0$. Sometimes, an additional assumption is needed, such as the Bohr condition, namely $\lambda_{n+1} - \lambda_n \geq c/n$ for some $c > 0$.

We will look into certain examples of Dirichlet's series now.

Example 3.2.1. A very important example is the Riemann zeta function which is

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}.$$

For $t = 0$, that is, $s = \sigma \in \mathbb{R}$, it is proved from elementary calculus, that $\zeta(\sigma)$ diverges for $\sigma = 1$ and is absolutely convergent for $\sigma > 1$. This is called the "p-test", where $p = \sigma$. We will learn more about this in our next unit.

Example 3.2.2. Another familiar example is the alternating zeta series

$$\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n^s},$$

again when $s = \sigma \in \mathbb{R}$ known as the Euler-Dedekind function. It is proved in elementary calculus that this series converges for $\sigma > 0$, where the convergence is conditional for $0 < \sigma \leq 1$ and absolute for $1 < \sigma$.

In this section we shall prove that very similar results hold, with appropriate hypotheses on the coefficients a_n, for $s \in \mathbb{C}$, that is, dropping the condition $t = 0$.

3.2.1 Convergence of Dirichlet's series

Recall that for a power series $\sum_{n=1}^{\infty} a_n z^n$, there exists a value $R \in [0, \infty]$, called the radius of convergence, such that

1. if $|z| < R$, then the power series converges;
2. if $|z| < R$, then the power series diverges;

3. for any $r < R$, the series converges uniformly and absolutely in $\{|z| \leq R\}$ and the sum is bounded on this set;

4. on the circle $\{|z| = R\}$, the behavior is more delicate.

As we will see, a Dirichlet's series has an **abscissa of convergence** $\sigma_0(f)$ such that the series converges for all $s \in \mathbb{C}$ with Re$s > \sigma_0(f)$ and diverges for all $s \in \mathbb{C}$ with Re$s < \sigma_0(f)$. For instance, the abscissa of convergence for the Riemann zeta function $\zeta(s)$ is 1.

Before we go into further details, we prove the following lemmas.

Lemma 3.2.1. Let $\alpha, \beta, \sigma \in \mathbb{R}$, $0 < \sigma$, $0 < \alpha < \beta$. Then

$$|\, \mathrm{e}^{-\alpha s} - \mathrm{e}^{-\beta s} \,| \leq \frac{|s|}{\sigma}(\mathrm{e}^{-\alpha\sigma} - \mathrm{e}^{-\beta\sigma}).$$

Proof. We have,

$$\mathrm{e}^{-\alpha s} - \mathrm{e}^{-\beta s} = s \int_\alpha^\beta \mathrm{e}^{-us}\, du,$$

hence,

$$|\, \mathrm{e}^{-\alpha s} - \mathrm{e}^{-\beta s} \,| \leq |s| \int_\alpha^\beta |\, \mathrm{e}^{-us} \,| du = |s| \int_\alpha^\beta \mathrm{e}^{-u\sigma}\, du = \frac{|s|}{\sigma}(\mathrm{e}^{-\alpha\sigma} - \mathrm{e}^{-\beta\sigma}).$$

□

Note 3.2.1. Setting $\alpha = \log(m),\ \beta = \log(n)$ in the above lemma, $0 < m < n,\ \sigma > 0$, then

$$|m^{-s} - n^{-s}| \leq \frac{|s|}{\sigma}(m^{-\sigma} - n^{-\sigma}).$$

Lemma 3.2.2. (Abel's Summation by parts formula) Let $A_n = \sum_{k=1}^n a_k$, then

$$\sum_{k=1}^n a_k b_k = A_n b_{n+1} - \sum_{k=1}^n A_k (b_{k+1} - b_k).$$

Proof. Since $a_k = A_k - A_{k-1}$, we have

$$\begin{aligned} \sum_{k=1}^n a_k b_k &= \sum_{k=1}^n [A_k - A_{k-1}] b_k \\ &= \sum_{k=1}^n A_k b_k - \sum_{k=1}^n A_k b_{k+1} + A_n b_{n+1}. \end{aligned}$$

Hence the result. □

Corollary 3.2.1. The sum $\sum_{k=1}^\infty a_k b_k$ converges if both $\sum_{k=1}^\infty A_k(b_{k+1} - b_k)$ and $\{A_n b_{n+1}\}$ are convergent.

We remark that Abels summation formula can be thought of as a discrete version of the familiar integration by parts formula from calculus, this should be clear by writing them side by side as

$$\sum_{k=1}^n a_k b_k = A_n b_{n+1} - \sum_{k=1}^n A_k (b_{k+1} - b_k), \quad \int u dv = vu - \int v du.$$

Theorem 3.2.1. Consider $\sum_{n=1}^\infty a_n n^{-s}$, $a_n \in \mathbb{C}$. Let $A_n = a_1 + a_2 + \cdots$. If $\{|A_n|\}$ is bounded, then the series converges for $\sigma > 0$.

Proof. We have, $|A_n| \leq C$, for some $C > 0$ and for all n. We shall use corollary 3.2.1, with $a_n = A_n$ and $b_n = n^{-s}$. Then

$$|A_n b_{n+1}| = |A_n| \cdot |b_{n+1}| \leq C \cdot (n+1)^{-\sigma} \to 0, \text{ as } n \to \infty.$$

Hence the second condition of Corollary (3.2.1) is satisfied, that is, $\{A_n b_{n+1}\}$ converges (in this case, to 0).

For the first condition, we apply the Cauchy convergence criterion to $\sum_{k=1}^{\infty} A_k((k+1)^{-s} - k^{-s})$. Given $\epsilon > 0$ and using note 3.2.1, if $\{S_n\}$ is the partial sum associated with the series, then we have,

$$\begin{aligned}
|S_n - S_m| &= \left| \sum_{k=m+1}^{n} A_k((k+1)^{-s} - k^{-s}) \right| \\
&\leq C \sum_{k=m+1}^{n} |(k+1)^{-s} - k^{-s})| \\
&\leq \frac{C|s|}{\sigma} \sum_{k=m+1}^{n} \left(\frac{1}{k^{\sigma}} - \frac{1}{(k+1)^{\sigma}} \right) \\
&= \frac{C|s|}{\sigma} \left(\frac{1}{(m+1)^{\sigma}} - \frac{1}{(n+1)^{\sigma}} \right) \\
&\leq \frac{C|s|}{\sigma(m+1)^{\sigma}} < \epsilon
\end{aligned}$$

for sufficiently large m. Hence, the result follows. □

We now state an elementary theorem for the convergence of Dirichlet's series.

Theorem 3.2.2. If the series is convergent for $s = \sigma + it$, then it is convergent for any value of s whose real part is greater than σ.

This theorem is included in the more general and less elementary theorem which follows. The above theorem can be obtained as a corollary of the theorem that follows.

Theorem 3.2.3. If the series $\sum_{n=1}^{\infty} a_n n^{-s}$ converges at some $s_0 \in \mathbb{C}$, then, for every $\delta > 0$, it converges uniformly in the sector

$$\left\{ s : \ -\frac{\pi}{2} + \delta < \arg(s - s_0) < \frac{\pi}{2} - \delta \right\}.$$

Proof. Without any loss of generality, we may assume that $s_0 = 0$, that is, the series $\sum_{n=1}^{\infty} a_n$ converges. Let $r_n = \sum_{k=n+1}^{\infty} a_k$, and fix $\epsilon > 0$. Then there exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$, $|r_n| < \epsilon$. Using summation by parts, for s in the sector and $M, N > n_0$, we get,

$$\begin{aligned}
\sum_{n=M}^{N} a_n n^{-s} &= \sum_{n=M}^{N} (r_{n-1} - r_n) n^{-s} \\
&= \sum_{n=M}^{N-1} r_n \left[\frac{1}{(n+1)^s} - \frac{1}{n^s} \right] + \frac{r_{M-1}}{M^s} - \frac{r_N}{N^s} \qquad (3.2.3)
\end{aligned}$$

The absolute values of the last two terms are bounded by ϵ, numerators are bounded by ϵ, while the denominators have absolute value at least 1. To estimate the summation part of (3.2.3), note that

$$\frac{1}{(n+1)^s} - \frac{1}{n^s} = \int_n^{n+1} \frac{-s}{x^{s+1}} dx,$$

so that

$$\left|\frac{1}{(n+1)^s}-\frac{1}{n^s}\right| \le |s|\int_n^{n+1}\frac{dx}{|x^{s+1}|} = \frac{|s|}{\sigma}\left[\frac{1}{n^\sigma}-\frac{1}{(n+1)^\sigma}\right]. \tag{3.2.4}$$

Thus the absolute value of the summation part of (3.2.3) satisfies for $M, N > n_0$,

$$\begin{aligned}\left|\sum_{n=M}^{N-1} r_n\left[\frac{1}{(n+1)^s}-\frac{1}{n^s}\right]\right| &\le \sum_{n=M}^{N-1}|r_n|\frac{|s|}{\sigma}\left[\frac{1}{n^\sigma}-\frac{1}{(n+1)^\sigma}\right]\\ &\le \epsilon\frac{|s|}{\sigma}\sum_{n=M}^{N-1}\left[\frac{1}{n^\sigma}-\frac{1}{(n+1)^\sigma}\right]\\ &\le \epsilon\frac{|s|}{\sigma}\left[\frac{1}{M^\sigma}-\frac{1}{N^\sigma}\right]\\ &\le c(\delta)\epsilon,\end{aligned}$$

since

$$\frac{|s|}{\sigma} = \left|\frac{1}{\cos(\arg s)}\right| \le \frac{1}{\cos\left(\frac{\pi}{2}-\delta\right)} =: c(\delta).$$

This proves that the series is uniformly Cauchy, and hence uniformly convergent. □

There are now three possibilities as regards the convergence of the series. It may converge for all, or no, or some values of s. In the last case it follows from theorem 3.2.2, by a classical argument, that we can find a number σ_0 such that the series is convergent for $\sigma > \sigma_0$ and divergent or oscillatory for $\sigma < \sigma_0$.

Theorem 3.2.4. The series may be convergent for all values of s, or for none, or for some only. In the last case there is a number σ_0 such that the series is convergent for $\sigma > \sigma_0$ and divergent or oscillatory for $\sigma < \sigma_0$.

Proof. If the series converges at some $s_0 \in \mathbb{C}$, the theorem follows from the inclusion

$$\{s:\ \text{Re } s = \sigma > \sigma_0\} \subset \bigcup_{\delta>0}\{s:\ |s-s_0| < \pi/2-\delta\}.$$

□

In other words the region of convergence is a half-plane. We call σ_0 as the **abscissa of convergence**, and the line $\sigma = \sigma_0$ as the **line of convergence**. It is convenient to write $\sigma_0 = -\infty$ or $\sigma_0 = \infty$ when the series is convergent for all or no values of s. On the line of convergence the question of the convergence of the series remains open, and requires considerations of a much more delicate character.

We formally define the abscissa of convergence as follows:

Definition 3.2.2. The abscissa of convergence of the Dirichlet series $\sum_{n=1}^{\infty} a_n n^{-s}$ is the extended real number $\sigma_0 \in [-\infty, \infty]$ with the following properties

1. If Re $s > \sigma_0$, then the series converges;
2. If Re $s < \sigma_0$, then the series diverges; If Re $s = \sigma_0$, nothing can be said about the convergence of the series.

Note 3.2.2. To determine the abscissa of convergence, it is enough to look at convergence of the series for $s \in \mathbb{R}$.

Example 3.2.3. 1. The series $\sum_{n=1}^{\infty} a^n n^{-s}$, where $|a| < 1$, is convergent for all s. If $|a| > 1$, then the series converges for no value of s. And for $a = 1$, it is not convergent at any point of the line of convergence, diverging to $+\infty$ for $s = 1$ and oscillating finitely for other values of s.

2. The series $\sum_{n=2}^{\infty} (\log n)^{-2} n^{-s}$ has the same line of convergence as the last series, but is convergent (indeed absolutely convergent) at all points of the line.

3. The series $\sum_{n=2}^{\infty} a_n n^{-s}$ where $a_n = (-1)^n + (\log n)^{-2}$, has the same line of convergence, and is convergent (though not absolutely) at all points of it.

We also have an abscissa of absolute convergence of a Dirichlet's series $\sum_{n=1}^{\infty} a_n n^{-s}$.

Definition 3.2.3. Given a Dirichlet's series $\sum_{n=1}^{\infty} a_n n^{-s}$, the abscissa of absolute convergence is defined as

$$\begin{aligned} \sigma_a &= \inf \left\{ \rho : \sum_{n=1}^{\infty} a_n n^{-s} \text{ converges absolutely for some s with } Re\ s = \rho \right\} \\ &= \inf \left\{ \rho : \sum_{n=1}^{\infty} a_n n^{-s} \text{ converges absolutely for all s with } Re\ s \geq \rho \right\}. \end{aligned}$$

The following theorem gives the relationship between σ_0 and σ_a for a Dirichlet's series.

Theorem 3.2.5. For any Dirichlet series, we have

$$\sigma_0 \leq \sigma_a \leq \sigma_0 + 1.$$

Proof. The first inequality is obvious. For the second, assume, that $\sigma_0 = 0$. We need to show that for $\sigma > 1$, $\sum_{n=1}^{\infty} |a_n n^{-s}|$ converges. Take $\epsilon > 0$ such that $\sigma - \epsilon > 1$. Then,

$$\sum_{n=1}^{\infty} |a_n n^{-s}| = \sum_{n=1}^{\infty} \frac{|a_n|}{n^{\sigma}} = \sum_{n=1}^{\infty} \frac{|a_n|}{n^{\epsilon}} \cdot \frac{1}{n^{\sigma-\epsilon}} \leq C \sum_{n=1}^{\infty} \frac{1}{n^{\sigma-\epsilon}} < \infty,$$

where, $C := \sup_n |a_n/n^{\epsilon}|$ is finite, since $\sigma_0 = 0$. □

Remark 3.2.1. If $a_n > 0$ for all $n \in \mathbb{N}$, then $\sigma_0 = \sigma_a$. This follows immediately by considering $s \in \mathbb{R}$.

Recall that for the radius of convergence of a power series, we have the following formula

$$1/R = \limsup_{n\to\infty} |a_n|^{\frac{1}{n}}.$$

The following is an analogous formula for the abscissa of convergence of a Dirichlet's series.

Theorem 3.2.6. Let $\sum_{n=1}^{\infty} a_n n^{-s}$ be a Dirichlet's series, and let σ_0 be its abscissa of convergence. Let $s_n = a_1 + a_2 + \cdots + a_n$ and $r_n = a_{n+1} + a_{n+1} + \cdots$

1. If $\sum a_n$ diverges, then
$$0 \leq \sigma_0 = \limsup_{n\to\infty} \frac{\log |s_n|}{\log n}.$$

2. If $\sum a_n$ converges, then
$$0 \geq \sigma_0 = \limsup_{n\to\infty} \frac{\log |r_n|}{\log n}.$$

Proof. 1. We assume that the series $\sum_{n=1}^{\infty} a_n n^{-s}$ diverges and define

$$\alpha = \limsup_{n\to\infty} \frac{\log|s_n|}{\log n}.$$

We will first show that $\alpha \leq \sigma_0$. Assume that $\sum_{n=1}^{\infty} a_n n^{-\sigma}$ converges. Thus, $\sigma > 0$ and we need to show that $\sigma \geq \alpha$. Let $b_n = a_n n^{-s}$ and $B_n = \sum_{k=1}^{n} b_k$ (so that $B_0 = 0$). By assumption, the sequence $\{B_n\}$ is bounded, say by M, and we can use the summation by parts as follows:

$$\begin{aligned} s_N &= \sum_{n=1}^{N} a_n \\ &= \sum_{n=1}^{N} b_n n^{\sigma} \\ &= \sum_{n=1}^{N-1} B_n[n^{\sigma} - (n+1)^{\sigma}] + B_N N^{\sigma} \end{aligned}$$

so that

$$|s_N| \leq M \sum_{n=1}^{N-1} [(n+1)^{\sigma} - n^{\sigma}] + MN^{\sigma} \leq 2MN^{\sigma}.$$

Applying the natural logarithm to both sides yields

$$\log|s_N| \leq \sigma \log N + \log 2M,$$

so,

$$\frac{\log|s_n|}{\log N} \leq \sigma + \frac{\log 2M}{\log N},$$

and this tends to σ as $N \to \infty$, giving the desired upper bound for α.

We need to show the other inequality $\sigma_0 \leq \alpha$. Suppose $\sigma > \alpha$. We need to show that $\sum_{n=1}^{\infty} a_n n^{-\sigma}$ converges. Chosen an $\epsilon > 0$ such that $\alpha + \epsilon < \sigma$. By definition, there exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$,

$$\frac{\log|s_n|}{\log n} \leq \alpha + \epsilon.$$

This implies that

$$\log|s_n| \leq (\alpha + \epsilon) \log n = \log\left(n^{\alpha+\epsilon}\right).$$

Using summation by parts, we can compute

$$\begin{aligned} \sum_{n=M+1}^{N} \frac{a_n}{n^{\sigma}} &= \sum_{n=M}^{N} s_n[n^{-\sigma} - (n+1)^{\sigma}] + S_N(N+1)^{\sigma} - s_M M^{-\sigma} \\ &\leq \sum_{n=M}^{N} n^{\alpha+\epsilon}[\sigma n^{-\sigma-1}] + N^{\alpha+\epsilon}N^{-\sigma} + M^{\alpha+\epsilon}M^{-\sigma} \\ &\lesssim (M-1)^{\alpha+\epsilon-\sigma}, \end{aligned}$$

and the last quantity tends to zero as M tends to ∞.

We estimated $\sum_{n=M}^{N} n^{\alpha+\epsilon-\sigma-1}$ by the integral

$$\int_{M-1}^{N-1} x^{\alpha+\epsilon-\sigma-1}dx \lesssim (M-1)^{\alpha+\epsilon-\sigma},$$

and the symbol $\lesssim$ means less than or equal to a constant times the right hand-side (where the constant depends on $\alpha+\epsilon-\sigma$, but, critically, not on M).

2. Similar to the first part.

□

From the formulae above we can simply deduce formulae for the abscissa of absolute convergence, although these can be derived easily on their own.

Corollary 3.2.2. For a Dirichlet's series $\sum_{n=1}^{\infty} a_n n^{-s}$, we have

1. if $\sum |a_n|$ diverges, then

$$\sigma_a = \limsup_{n\to\infty} \frac{\log(|a_1|+|a_2|+\cdots+|a_n|)}{\log n} \geq 0,$$

2. if $\sum |a_n|$ converges, then

$$\sigma_a = \limsup_{n\to\infty} \frac{\log(|a_{n+1}|+|a_{n+2}|+\cdots)}{\log n} \leq 0.$$

Example 3.2.4. The series

$$\sum_{n=1}^{\infty} \frac{(-1)^n}{p_n^s}$$

(where, p_n are primes) has $\sigma_0 = 0$ and $\sigma_a = 1$.

The series of coefficients diverges and so we use the first of the pair of formulae for each abscissae

$$\sigma_0 = \limsup_{n\to\infty} \frac{\log 1}{\log n} = 0,$$

and, using the prime number theorem,

$$\sigma_a = \limsup_{n\to\infty} \frac{\log(\pi(n))}{\log n} = \limsup_{n\to\infty} \frac{\log n - \log(\log n)}{\log n} = 1,$$

where, $\pi(x)$ denotes the number of primes less than or equal to x.

Theorem 3.2.7. Suppose that the series $\sum_{n=1}^{\infty} a_n n^{-s}$ converges absolutely to some $f(s)$ in some half-plane $\mathbb{H}_c = \{s : \text{ Re } s > c\}$ and $f(s) \equiv 0$ in the half-plane $\mathbb{H}_c$. Then $a_n = 0$ for all $n \in \mathbb{N}$.

Proof. We may assume that $c < 0$, so, in particular, $\sum |a_n| < \infty$. Suppose that all a_n's are not zero, and let n_0 be the be the smallest natural number such that $a_{n_0} \neq 0$.

We claim that $\lim_{\sigma\to\infty} f(\sigma) n_0^{\sigma} = a_{n_0}$. To prove the claim that

$$\begin{aligned} 0 &\leq n_0^{\sigma} \left| \sum_{n>n_0} a_n n^{-\sigma} \right| \\ &\leq \sum_{n>n_0} |a_n| \left(\frac{n_0}{n}\right)^{\sigma} \\ &\leq \left(\frac{n_0}{n_0+1}\right)^{\sigma} \sum_{n>n_0} |a_n|, \end{aligned}$$

and the last term tends to 0 as $\sigma \to \infty$, since $\sum |a_n|$ converges. As

$$f(\sigma)n_0^\sigma = a_{n_0} + n_0^\sigma \sum_{n>n_0} a_n n^{-\sigma},$$

the claim is proved.

The proof is also finished, because the limit in the claim is obviously 0, a contradiction. □

3.3 Few Probable Questions

1. Define the abscissa of convergence of a Dirichlet's series $\sum_{n=1}^\infty a_n n^{-s}$. Show that if the series diverges, then
$$0 \leq \sigma_0 = \limsup_{n\to\infty} \frac{\log |s_n|}{\log n}.$$

2. Define the abscissa of absolute convergence of a Dirichlet's series $\sum_{n=1}^\infty a_n n^{-s}$. Show that $\sigma_0 \leq \sigma_a \leq \sigma_0 + 1$.

3. Show that if the series $\sum_{n=1}^\infty a_n n^{-s}$ converges for some $s_0 \in \mathbb{C}$, then, for every $\delta > 0$, it converges uniformly in the sector
$$\left\{ s : \ -\frac{\pi}{2} + \delta < \arg(s - s_0) < \frac{\pi}{2} - \delta \right\}.$$

Unit 4

Course Structure

- The Riemann Zeta function
- The product development and the zeros of the zeta functions.

4.1 Introduction

As we have introduced in the previous unit, the Riemann zeta function $\zeta(s)$ is a function of the complex variable s, defined as

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s},$$

which plays a pivotal role in the analytic number theory and has applications in physics, probability theory, and applied statistics.

As a function of a real variable, Leonhard Euler first introduced and studied it in the first half of the eighteenth century without using complex analysis, which was not available at the time. Bernhard Riemann's 1859 article "On the Number of Primes Less Than a Given Magnitude" extended the Euler definition to a complex variable, proved its meromorphic continuation and functional equation, and established a relation between its zeros and the distribution of prime numbers.

The values of the Riemann zeta function at even positive integers were computed by Euler. The first of them, $\zeta(2)$, provides a solution to the Basel problem. In 1979 Roger Apry proved the irrationality of $\zeta(3)$. The values at negative integer points, also found by Euler, are rational numbers and play an important role in the theory of modular forms. Many generalizations of the Riemann zeta function, such as Dirichlet series, Dirichlet L-functions and L-functions, are known. We will however not indulge into such rigorous treatments of the zeta function. We will only restrict ourselves to some preliminary ideas, starting with the definition, convergence, etc.

Objectives

After reading this unit, you will be able to

- define the Riemann zeta function and know about its origins in a preliminary level

- deduce the product development of the zeta function
- deduce the functional equation and its other forms

4.2 Riemann Zeta Function

The Riemann zeta function, as we have already seen, is the function

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s},$$

where, $s = \sigma + it$ is a complex number. First, we will discuss the convergence of the function. See that

$$\begin{aligned}
|\zeta(s)| &= \left|\sum_{n=1}^{\infty} \frac{1}{n^s}\right| \\
&\leq \sum_{n=1}^{\infty} \frac{1}{|n^s|} \\
&= \sum_{n=1}^{\infty} \frac{1}{|n^{\sigma+it}|} \\
&= \sum_{n=1}^{\infty} \frac{1}{|n^{\sigma}| \cdot |n^{it}|} \\
&= \sum_{n=1}^{\infty} \frac{1}{n^{\sigma}|\, \mathrm{e}^{it\log(n)}\,|} \\
&= \sum_{n=1}^{\infty} \frac{1}{n^{\sigma}}.
\end{aligned}$$

which converges for all $\sigma > 1$. Hence, $\sigma_0 = 1$ is the abscissa of convergence of the series $\sum n^{-s}$. Also, the series is not convergent on the line of convergence. Also, since for $\sigma > 1$,

$$\left|\frac{1}{n^s}\right| \leq \frac{1}{n^{\sigma}},$$

and since the series $\sum_{n=1}^{\infty} n^{-\sigma}$ converges in the said region, so by Weierstrass M-test, the series $\sum n^{-s}$ converges uniformly and absolutely in the hlf-plane $\sigma > 1$, and thus, defines an analytic function in the plane $\mathbb{H}_1 = \{s \in \mathbb{C} : \text{ Re } s > 1\}$.

4.3 The Product Development

The number-theoretic properties of $\zeta(s)$ are inherent in the following connection between the zeta function and the ascending sequence of primes $p_1, p_2, \ldots, p_n, \ldots$

Theorem 4.3.1. For $\sigma > 1$, we have

$$\frac{1}{\zeta(s)} = (1 - p_1^{-s})(1 - p_2^{-s}) \cdots (1 - p_n^{-s}) \cdots = \Pi_{n=1}^{\infty}(1 - p_n^{-s}),$$

where, $p_1, p_2, \ldots, p_n, \ldots$, are prime numbers and the term on the right hand side of the above equation is the infinite product of the numbers $(1 - p_n^{-s})$.

Proof. Under the assumption that $\sigma > 1$, we see that

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}. \tag{4.3.1}$$

Multiplying equation (4.3.1) by 2^{-s}, we get

$$\zeta(s) \cdot \frac{1}{2^s} = \sum_{n=1}^{\infty} \frac{1}{n^s} \cdot \frac{1}{2^s} = \sum_{n=1}^{\infty} \frac{1}{(2n)^s}. \tag{4.3.2}$$

Subtracting equation (4.3.2) from (4.3.1), we get,

$$\zeta(s)\left(1 - \frac{1}{2^s}\right) = \sum_{n=1}^{\infty} \frac{1}{n^s} - \sum_{n=1}^{\infty} \frac{1}{(2n)^s} = \sum_{n=1;\ n\neq 2k}^{\infty} \frac{1}{n^s}. \tag{4.3.3}$$

The last term in the above equation is the sum of all terms n^{-s}, excluding the terms n which are multiples of 2.

Again, multiplying the equation (4.3.3) by 3^{-s}, we get

$$\zeta(s)\left(1 - \frac{1}{2^s}\right) \cdot \frac{1}{3^s} = \sum_{n=1;\ n\neq 2k}^{\infty} \frac{1}{n^s} \cdot \frac{1}{3^s} = \sum_{n=1;\ n\neq 2k}^{\infty} \frac{1}{(3n)^s}. \tag{4.3.4}$$

Subtracting equation (4.3.4) from (4.3.3), we get,

$$\zeta(s)\left(1 - \frac{1}{2^s}\right)\left(1 - \frac{1}{3^s}\right) = \sum_{n=1;\ n\neq 2k;\ n\neq 3k}^{\infty} \frac{1}{n^s}.$$

The last term in the above equation is the sum of all terms n^{-s}, excluding the terms n which are multiples of 2 and 3.

Continuing in this way, we get,

$$\zeta(s)\left(1 - \frac{1}{2^s}\right)\left(1 - \frac{1}{3^s}\right)\cdots\left(1 - \frac{1}{p_n^s}\right) = \sum_{\cdots\ n\neq p_n k}^{\infty} \frac{1}{n^s},$$

where, the term on the right hand side of the above equation is the sum of all those terms n^{-s}, which are not the multiples of the primes $2, 3, 5, \ldots, p_n$ arranged in ascending order. Thus, taking limit as $n \to \infty$, we get

$$\lim_{n\to\infty} \zeta(s)\left(1 - \frac{1}{2^s}\right)\left(1 - \frac{1}{3^s}\right)\cdots\left(1 - \frac{1}{p_n^s}\right) = \sum_{n\neq p_n k} \frac{1}{n^s},$$

where, the sum is taken over all such n^{-s}, such that n is not a multiple of any prime p_n and such number can be none other than 1. So, the above equation becomes,

$$\zeta(s)\left(1 - \frac{1}{2^s}\right)\left(1 - \frac{1}{3^s}\right)\cdots\left(1 - \frac{1}{p_n^s}\right)\cdots = 1,$$

which finally gives,

$$\frac{1}{\zeta(s)} = (1 - p_1^{-s})(1 - p_2^{-s})\cdots(1 - p_n^{-s})\cdots,$$

where, $p_1, p_2, p_3, \ldots, p_n, \ldots$ is the complete list of prime numbers arranged in ascending order. □

The above representation of the zeta function is called the Euler product representation of the zeta function. Also, notice that the product development explained above, includes the introduction of an infinite product. Infinite products, like the infinite series, are convergent when the sequence of partial products converge as we will see in subsequent units. The infinite product in this case is convergent uniformly in the region $\sigma > 1$.

We have taken for granted that there are infinitely many primes. Actually, the reasoning can be used to prove this fact. For if p_n were the largest prime, then we would have got

$$\zeta(s)\left(1-\frac{1}{2^s}\right)\left(1-\frac{1}{3^s}\right)\cdots\left(1-\frac{1}{p_n^s}\right)=1$$

and it would follow that $\zeta(\sigma)$ has a finite limit when $\sigma \to 1$. This contradicts the divergence of the series $\sum_{n=1}^{\infty} n^{-1}$.

4.4 Functional Equations

4.4.1 Relationship with the Gamma Function

We are familiar with the gamma function which is written as

$$\Gamma(s)=\int_0^\infty t^{s-1}\,\mathrm{e}^{-t}\,dt,$$

and the integral converges for all values of Re $s > 0$. We make the substitution

$$t = nu \Rightarrow dt = ndu,$$

where n is positive integer. So the above integral changes to

$$\Gamma(s)\int_0^\infty (nu)^{s-1}\,\mathrm{e}^{-nu}\,ndu=\int_0^\infty n^s u^{s-1}\,\mathrm{e}^{-nu}\,du,$$

which gives

$$\Gamma(s)\frac{1}{n^s}\int_0^\infty u^{s-1}\,\mathrm{e}^{-nu}\,du.$$

Summing over n from 1 to ∞, we get,

$$\Gamma(s)\sum_{n=1}^{\infty}\frac{1}{n^s}=\sum_{n=1}^{\infty}\int_0^\infty u^{s-1}\,\mathrm{e}^{-nu}\,du$$

Since the integral on the right hand side is absolutely converging, so the sum and integral can be exchanged. Thus, the above equation changes to

$$\begin{aligned}
\Gamma(s)\zeta(s) &= \int_0^\infty u^{s-1}\sum_{n=1}^{\infty}\mathrm{e}^{-nu}\,du \\
&= \int_0^\infty u^{s-1}\left(\frac{1}{1-\mathrm{e}^{-u}}-1\right)du \\
&= \int_0^\infty u^{s-1}\frac{\mathrm{e}^{-u}}{1-\mathrm{e}^{-u}}du \\
&= \int_0^\infty \frac{u^{s-1}}{\mathrm{e}^{u}-1}du.
\end{aligned}$$

for Re $s > 1$. Thus, the celebrated relationship between the gamma and zeta function is given by

$$\Gamma(s)\zeta(s) = \int_0^\infty \frac{u^{s-1}}{\mathrm{e}^u - 1} du$$

for Re $s > 1$.

4.4.2 Theta Function

We need to learn certain basics of the Jacobi theta function that we will need in the sequel. The theta function is given by

$$\vartheta(x) = \sum_{n\in\mathbb{Z}} \mathrm{e}^{-\pi n^2 x}.$$

Let f be any complex function that is analytic in the strip $\{z \in \mathbb{C} : |\mathrm{Im}\, z| < a\}$, and $|f(x+iy)| \leq A/(1+x^2)$ for some constant $A > 0$ and all $x \in \mathbb{R}$ such that $|y| < a$ for $a > 0$. $\mathrm{e}^{-\pi n^2 z}$ satisfies the properties stated thus.

By Poisson summation formula, we have for such f as described above,

$$\sum_{n\in\mathbb{Z}} f(n) = \sum_{n\in\mathbb{Z}} \hat{f}(n),$$

where, $\hat{f}$ is the Fourier transform of f. The Fourier transform of the function $\mathrm{e}^{\pi x^2}$ is the function itself, that is,

$$\int_{-\infty}^{\infty} \mathrm{e}^{-\pi x^2}\, \mathrm{e}^{-2\pi i x\xi}\, dx = \mathrm{e}^{-\pi\xi^2}.$$

For fixed values of $t > 0$ and $a \in \mathbb{R}$, the change of variables $x \mapsto t^{1/2}(x + a)$ in the above integral show that the Fourier transform of the function

$$f(x) = \mathrm{e}^{-\pi t(x+a)^2},$$

for fixed values of $t > 0$ and $a \in \mathbb{R}$, we get

$$\hat{f}(\xi) = t^{-1/2}\, \mathrm{e}^{-\pi\xi^2/t}\, \mathrm{e}^{2\pi i a\xi}.$$

Applying the Poisson summation formula to the above pair we get,

$$\sum_{n\in\mathbb{Z}} \mathrm{e}^{-\pi t(n+a)^2} = \sum_{n\in\mathbb{Z}} t^{-1/2}\, \mathrm{e}^{-\pi n^2/t}\, \mathrm{e}^{2\pi i n a}.$$

This identity has noteworthy consequences. For instance, the special case $a = 0$ is the transformation law for the theta function we defined above. Thus, we get,

$$\vartheta(t) = t^{-1/2}\vartheta(1/t), \tag{4.4.1}$$

for $t > 0$.

4.4.3 Functional equations

Now, we will derive the Riemann functional equation. The equation is

$$\pi^{-\frac{s}{2}}\Gamma\left(\frac{s}{2}\right)\zeta(s) = \pi^{-\frac{1-s}{2}}\Gamma\left(\frac{1-s}{2}\right)\zeta(1-s).$$

We have

$$\Gamma(s) = \int_0^\infty t^{s-1}\,\mathrm{e}^{-t}\,dt.$$

Thus,

$$\Gamma\left(\frac{s}{2}\right) = \int_0^\infty t^{\frac{s}{2}-1}\,\mathrm{e}^{-t}\,dt. \tag{4.4.2}$$

We use the substitution

$$t = \pi n^2 x \Rightarrow dt = \pi n^2 dx$$

Thus, equation (4.4.2) becomes

$$\Gamma\left(\frac{s}{2}\right) = \int_0^\infty (\pi n^2 x)^{\frac{s}{2}-1}\,\mathrm{e}^{-\pi n^2 x}\,\pi n^2 dx = \int_0^\infty \pi^{\frac{s}{2}} n^s x^{\frac{s}{2}-1}\,\mathrm{e}^{-\pi n^2 x}\,dx. \tag{4.4.3}$$

Multiplying (4.4.3) by $\pi^{-s/2}\cdot n^{-s}$, we get

$$\pi^{-\frac{s}{2}}\Gamma\left(\frac{s}{2}\right)\cdot\frac{1}{n^s} = \int_0^\infty x^{\frac{s}{2}-1}\,\mathrm{e}^{-\pi n^2 x}\,dx.$$

Summing the above equation over, from 1 to ∞, we get,

$$\sum_{n=1}^\infty \pi^{-\frac{s}{2}}\Gamma\left(\frac{s}{2}\right)\cdot\frac{1}{n^s} = \sum_{n=1}^\infty \int_0^\infty x^{\frac{s}{2}-1}\,\mathrm{e}^{-\pi n^2 x}\,dx$$

which gives,

$$\pi^{-\frac{s}{2}}\Gamma\left(\frac{s}{2}\right)\zeta(s) = \int_0^\infty x^{\frac{s}{2}-1}\sum_{n=1}^\infty \mathrm{e}^{-\pi n^2 x}\,dx. \tag{4.4.4}$$

We have,

$$\vartheta(x) = \sum_{n\in\mathbb{Z}} \mathrm{e}^{-\pi n^2 x} = 1 + 2\sum_{n=1}^\infty \mathrm{e}^{-\pi n^2 x} = 1 + 2\psi(x).$$

Replacing this in the right hand side of equation (4.4.4) we get,

$$\begin{aligned}\int_0^\infty x^{\frac{s}{2}-1}\sum_{n=1}^\infty \mathrm{e}^{-\pi n^2 x}\,dx &= \int_0^\infty x^{\frac{s}{2}-1}\psi(x)dx \\ &= \int_0^1 x^{\frac{s}{2}-1}\psi(x)dx + \int_1^\infty x^{\frac{s}{2}-1}\psi(x)dx. \end{aligned}\tag{4.4.5}$$

We have, by (4.4.1),

$$2\psi(x) + 1 = \frac{1}{\sqrt{x}}\left(2\psi\left(\frac{1}{x}\right) + 1\right)$$

which gives,

$$\psi(x) = \frac{1}{\sqrt{x}}\psi\left(\frac{1}{x}\right) + \frac{1}{2\sqrt{x}} - \frac{1}{2}.$$

Thus,

$$\begin{aligned}\int_0^1 x^{\frac{s}{2}-1}\psi(x)dx &= \int_0^1 x^{\frac{s}{2}-1}\left(\frac{1}{\sqrt{x}}\psi\left(\frac{1}{x}\right) + \frac{1}{2\sqrt{x}} - \frac{1}{2}\right)dx \\ &= \int_0^1 \left[x^{\frac{s}{2}-\frac{3}{2}}\psi\left(\frac{1}{x}\right) + \frac{1}{2}\left(x^{\frac{s}{2}-\frac{3}{2}} - x^{\frac{s}{2}-1}\right)\right]dx \\ &= \int_0^1 x^{\frac{s}{2}-\frac{3}{2}}\psi\left(\frac{1}{x}\right)dx + \frac{1}{s(s-1)}. \end{aligned}\tag{4.4.6}$$

Using the substitution in the integral on the right hand of the above equation

$$x = \frac{1}{u} \Rightarrow dx = -\frac{1}{u^2}du$$

and the limits also change accordingly and the equation (4.4.6) becomes, with a change in the dummy variable u to x,

$$\int_0^1 x^{\frac{s}{2}-1}\psi(x)dx = \int_1^\infty x^{-\frac{s}{2}-\frac{1}{2}}\psi(x)dx + \frac{1}{s(s-1)}.$$

Finally, (4.4.5) becomes

$$\begin{aligned}\int_0^\infty x^{\frac{s}{2}-1}\psi(x)dx &= \int_1^\infty x^{\frac{s}{2}-1}\psi(x)dx + \int_1^\infty x^{-\frac{s}{2}-\frac{1}{2}}\psi(x)dx + \frac{1}{s(s-1)} \\ &= \int_1^\infty \left(x^{\frac{s}{2}-1} + x^{-\frac{s}{2}-\frac{1}{2}}\right)\psi(x)dx + \frac{1}{s(s-1)}.\end{aligned}$$

Thus, (4.4.4) and (4.4.5) together give

$$\begin{aligned}\pi^{-\frac{s}{2}}\Gamma\left(\frac{s}{2}\right)\zeta(s) &= \int_1^\infty \left(x^{\frac{s}{2}-1} + x^{-\frac{s}{2}-\frac{1}{2}}\right)\psi(x)dx + \frac{1}{s(s-1)} \\ &= \int_1^\infty \left(x^{\frac{s}{2}} + x^{\frac{1-s}{2}}\right)\frac{\psi(x)}{x}dx + \frac{1}{s(s-1)}. \qquad (4.4.7)\end{aligned}$$

Putting s by $1-s$ in the above equation, we get,

$$\begin{aligned}\pi^{-\frac{1-s}{2}}\Gamma\left(\frac{1-s}{2}\right)\zeta(1-s) &= \int_1^\infty \left(x^{\frac{1-s}{2}} + x^{\frac{1-(1-s)}{2}}\right)\frac{\psi(x)}{x}dx + \frac{1}{(1-s)(1-s-1)} \\ &= \int_1^\infty \left(x^{\frac{1-s}{2}} + x^{\frac{s)}{2}}\right)\frac{\psi(x)}{x}dx + \frac{1}{(1-s)s}. \qquad (4.4.8)\end{aligned}$$

Notice that the right hand sides of the equations (4.4.7) and (4.4.8) are the same. So, we get

$$\pi^{-\frac{s}{2}}\Gamma\left(\frac{s}{2}\right)\zeta(s) = \pi^{-\frac{1-s}{2}}\Gamma\left(\frac{1-s}{2}\right)\zeta(1-s).$$

Which is our required functional equation.

4.5 Few Probable Questions

1. With proper justification, find the abscissa of convergence of the zeta function.
2. Establish the Euler product representation of the zeta function.
3. Establish the relation between zeta function and the gamma function.
4. Deduce the Riemann functional equation.
5. Write the Riemann functional equation. Hence deduce that

$$\zeta(s) = 2^s\pi^{s-1}\sin\frac{\pi s}{2}\Gamma(1-s)\zeta(1-s).$$

Unit 5

Course Structure

- Entire functions, growth of an entire function
- Order and type and their representations in terms of the Taylor coefficients.

5.1 Introduction

In complex analysis, an entire function, also called an integral function, is a complex-valued function that is holomorphic at all finite points over the whole complex plane. Typical examples of entire functions are polynomials and the exponential function, and any finite sums, products and compositions of these, such as the trigonometric functions sine and cosine and their hyperbolic counterparts sinh and cosh, as well as derivatives and integrals of entire functions such as the error function. If an entire function $f(z)$ has a root at w, then $f(z)/(z-w)$, taking the limit value at w, is an entire function. On the other hand, neither the natural logarithm nor the square root is an entire function, nor can they be continued analytically to an entire function. Also, a transcendental entire function is an entire function that is not a polynomial. For example, the exponential function, sine and cosine functions are most common transcendental entire functions. This unit is dedicated to the study of the growth of entire functions.

Objectives

After reading this unit, you will be able to

- get more idea about the various entire functions
- learn the behaviour of the maximum modulus function $M(r)$ for entire functions
- define the order and type of entire functions
- define the order and type of entire functions with the help of the Taylor coefficients

5.2 Entire Functions

An entire function is a single valued function having Taylor series expansion

$$f(z) = a_0 + a_1 z + \cdots + a_n z^n + \cdots$$

which converges for all finite z. Moreover, according to the Cauchy Hadamard formula,

$$\lim_{n\to\infty} \sqrt[n]{|a_n|} = 0.$$

There are three possible ways in which an entire function $f(z)$ can behave at infinity:

1. $f(z)$ can have a regular point at infinity, then, according to Liouville's theorem, f is a constant function;
2. f can have a pole of order $k \geq 1$ at infinity, and then f reduces to a polynomial;
3. f can have an essential singularity at infinity, and then f is said to be an entire transcendental function.

Note that, the behaviour of $f(z)$ at infinity is determined by the action of $f(1/z)$ at 0. We will be mainly concerned with the transcendental entire functions from now on. If f is such a function, then we will clearly have, since $M(r)$ is a strictly increasing function of r,

$$\lim_{r\to\infty} M(r) = \infty.$$

We have the following theorem for transcendental entire functions f.

Theorem 5.2.1. If $f(z)$ is a transcendental entire function, with maximum modulus function $M(r)$, then

$$\liminf_{r\to\infty} \frac{\log M(r)}{\log r} = \infty.$$

Proof. Let $f(z) = \sum_{k=0}^{\infty} a_k z^k$ be a transcendental entire function with maximum modulus function $M(r)$. If possible, let

$$\liminf_{r\to\infty} \frac{\log M(r)}{\log r} = \mu < \infty.$$

Then for $\epsilon > 0$, we can find an increasing sequence $\{r_n\}$, tending to ∞, such that

$$\frac{\log M(r_n)}{\log r_n} < \mu + \epsilon$$

for every r_n, that is,

$$\log M(r_n) < (\mu + \epsilon) \log r_n \Rightarrow M(r_n) < r_n^{\mu+\epsilon}$$

for every r_n. Hence, by Cauchy's inequality,

$$|a_k| \leq \frac{M(r_n)}{r_n^k} < r_n^{\mu+\epsilon-k}$$

for $k = 0, 1, 2, \ldots$. Since r_n can be chosen arbitrarily large, it follows that $a_k = 0$ for all $k > \mu + \epsilon > \mu$. Hence, f is a polynomial of degree not greater than $[\mu]$, that is, the largest integer less than or equal to μ. This contradicts our assumption that f is transcendental. Hence the result. □

We have another result analogous to the above for non-transcendental entire functions as follows.

Theorem 5.2.2. For an entire function f, if there exists a positive integer k such that

$$\lim_{r\to\infty} \frac{M(r)}{r^k} < \infty,$$

then f is a polynomial of degree k atmost.

Proof. Let $f(z) = \sum_{n=0}^{\infty} a_n z^n$ be an entire function with maximum modulus function $M(r)$. Let

$$\lim_{r\to\infty} \frac{M(r)}{r^k} = \mu < \infty.$$

Then,

$$M(r) < (\mu + \epsilon) r^k,$$

for any positive ϵ and all $r \geq r_0$, for some r_0. By Cauchy's inequality,

$$|a_n| \leq \frac{M(r)}{r^n} < (\mu + \epsilon) r^{k-n}$$

for all $r \geq r_0$. Since we can choose r sufficiently large, $|a_n| \to 0$ as $r \to \infty$ for $n > k$. Hence, $a_n = 0$ for all $n > k$ and thus f is a polynomial of degree at most k. □

Note that the above theorem would also be true if

$$\liminf_{r\to\infty} \frac{M(r)}{r^k} < \infty.$$

5.2.1 Order of an entire function

The preceding discussions assert that a transcendental entire function $f(z)$ grows faster than any fixed positive power of r. This suggests that using the exponential function (that is, the simplest "rapidly growing function") to measure the growth of $f(z)$. We now formally define the order of an entire function.

Definition 5.2.1. An entire function f is said to be of **finite** order if there exists a positive number k such that the inequality

$$\log M(r) < r^k,$$

or

$$M(r) < \mathrm{e}^{r^k}$$

holds for sufficiently large r. Then

$$\rho = \inf \{ k : M(r) < \mathrm{e}^{r^k} \text{ holds for sufficiently large } r \}$$

is called the **order** of f. If $\rho = \infty$, that is, for any number k, there exists arbitrarily large values of r such that $\log M(r) > r^k$, then f is said to be of infinite order.

For example, e^z if of finite order (in fact, of order 1), while $\mathrm{e}^{\mathrm{e}^z}$ is of infinite order. From the definition, it is clear that the order of an entire function is always non-negative.

Theorem 5.2.3. The order ρ of an entire function f is given by the formula

$$\rho = \limsup_{r\to\infty} \frac{\log\log M(r)}{\log r}.$$

Proof. Let ρ be the order of f. Then, from the definition of ρ, we have, for any $\epsilon > 0$, there exists a number $r_0(\epsilon) > 0$ such that

$$\log M(r) < r^{\rho+\epsilon}$$

holds for all $r > r_0$. On the other hand, there exists an increasing sequence $\{r_n\}$ tending to infinity, such that

$$\log M(r) > r^{\rho-\epsilon}$$

for $r = r_n$. In other words,

$$\frac{\log \log M(r)}{\log r} < \rho + \epsilon, \quad \forall r > r_0 \tag{5.2.1}$$

and

$$\frac{\log \log M(r)}{\log r} > \rho - \epsilon, \tag{5.2.2}$$

for a sequence of values of r tending to infinity. Equations (5.2.1) and (5.2.2) precisely means

$$\rho = \limsup_{r\to\infty} \frac{\log \log M(r)}{\log r}.$$

□

5.2.2 Type of an entire function of finite non-zero order

Next, we subdivide the class of entire functions of finite non-zero order ρ.

Definition 5.2.2. Let f be an entire function with finite non-zero order ρ. By the **type** τ of f, we mean the greatest lower bound of positive numbers k such that the inequality

$$\log M(r) < kr^{\rho} \tag{5.2.3}$$

holds for all sufficiently large r, that is,

$$\tau = \inf\{k : \text{ (5.2.3) holds for all } r > r_0(k)\}.$$

However, suppose that $\tau = \infty$, that is, suppose that given any positive number k, there exists arbitrarily large values of r such that

$$\log M(r) > kr^{\rho}.$$

Then f is said to be of **infinite** (or, maximum) type. And when, $\tau = 0$, then f is said to be of minimum type. From the definition of type, it is clear that τ is always non-negative. When $0 < \tau < \infty$, then f is said to be of normal type.

Example 5.2.1. 1. If $f(z) = \mathrm{e}^z$, then $\rho = 1$ and $\tau = 1$.

2. If $f(z) = \mathrm{e}^{pz^n}$, where $p > 0$ and n is a positive integer, then $\rho = n$ and $\tau = p$.

Theorem 5.2.4. The type τ of an entire function f with finite non-zero order ρ is given by the formula

$$\tau = \limsup_{r\to\infty} \frac{\log M(r)}{r^{\rho}}.$$

Proof. Let τ be the type of an entire function f of order $\rho(\neq 0)$. Then, from the definition of τ, we have, for any $\epsilon > 0$, there exists a number $r_0(\epsilon) > 0$ such that

$$\log M(r) < (\tau + \epsilon) r^\rho, \quad \forall\, r > r_0.$$

On the other hand, there exists an increasing sequence $\{r_n\}$ tending to infinity, such that

$$\log M(r) > (\tau - \epsilon) r^\rho, \quad \forall\, r = r_n.$$

In other words,

$$\frac{\log M(r)}{r^\rho} < \tau + \epsilon, \quad \forall\, r > r_0 \tag{5.2.4}$$

and

$$\frac{\log M(r)}{r^\rho} > \tau - \epsilon, \tag{5.2.5}$$

for a sequence if values of r tending to infinity. Equations (5.2.4) and (5.2.5) together means

$$\tau = \limsup_{r\to\infty} \frac{\log M(r)}{r^\rho}.$$

□

Example 5.2.2. We show that the order of any polynomial is zero. Let $f(z) = a_0 + a_1 z + \cdots + a_n z^n$ be a polynomial. Then

$$\begin{aligned} |f(z)| &= |a_0 + a_1 z + \cdots + a_n z^n| \\ &\leq |a_0| + |a_1||z| + \cdots + |a_n||z|^n. \end{aligned}$$

Thus,

$$M(r) \leq |a_0| + |a_1| r + \cdots + |a_n| r^n \leq r^n(|a_0| + |a_1| + \cdots + |a_n|) = Br^n.$$

(taking $r \geq 1$. This choice is justified since ultimately $r \to \infty$.) where,

$$B_n = |a_0| + |a_1| + \cdots + |a_n|.$$

Hence,

$$\log M(r) \leq \log B + n \log r \leq logr + n \log r$$

taking r sufficiently large. Thus,

$$\log M(r) \leq (n+1) \log r$$

for large r. Hence,

$$\rho = \limsup_{r\to\infty} \frac{\log \log M(r)}{\log r} \leq \limsup_{r\to\infty} \frac{\log(n+1) + \log \log r}{\log r} = 0,$$

that is, $\rho \leq 0$. Also, we know that by definition, $\rho \geq 0$. Hence, $\rho = 0$.

From the above example, it is clear that the order of any constant function is zero. But, it does not mean that any zero order entire function is always a polynomial.

Example 5.2.3. The order of a transcendental entire function may also be zero. For example, if

$$f(z) = \sum_{n=0}^{\infty} \frac{z^n}{n^{n^{1+\delta}}}, \ \ \delta > 0,$$

is a transcendental entire function having order $\rho = 0$ as we will soon see.

5.2.3 Order for sum and multiplications of entire functions

This section is dedicated to study the growth of the sum and multiplication of entire functions with reference to the orders of the parent functions.

Theorem 5.2.5. Let ρ_1 and ρ_2 be the orders of the entire functions f_1 and f_2 respectively. Then

1. order of $f_1 f_2 \leq \max\{\rho_1, \rho_2\}$
2. order of $f_1 \pm f_2 \leq \max\{\rho_1, \rho_2\}$

Proof. 1. Let $\phi(z) = f_1(z) f_2(z)$ and let ρ be the order of ϕ and let $\rho_1 \geq \rho_2$. Also, let

$$M(r, \phi) = \max_{|z|=r} |\phi(z)|, \quad M(r, f_1) = \max_{|z|=r} |f_1(z)|, \quad M(r, f_2) = \max_{|z|=r} |f_2(z)|.$$

Since ρ_1 and ρ_2 are orders of f_1 and f_2 respectively, we have for any given $\epsilon > 0$,

$$\log M(r, f_1) < r^{\rho_1 + \epsilon}$$
$$\log M(r, f_2) < r^{\rho_2 + \epsilon}$$

for all sufficiently large r. Now,

$$|\phi(z)| = |f_1(z)||f_2(z)| \leq M(r, f_1) \cdot M(r, f_2), \quad \forall\, z \text{ in } |z| \leq r.$$

Hence,

$$\begin{aligned} & M(r, \phi) \leq M(r, f_1) \cdot M(r, f_2) \\ \text{or,} \quad & \log M(r, \phi) \leq \log M(r, f_1) + \log M(r, f_2) \\ & \leq r^{\rho_1 + \epsilon} + r^{\rho_2 + \epsilon} \\ & \leq r^{\rho_1 + \epsilon} + r^{\rho_1 + \epsilon} \\ & \leq 2r^{\rho_1 + \epsilon} < r^{\epsilon} \cdot r^{\rho_1 + \epsilon} = r^{\rho_1 + 2\epsilon} \end{aligned}$$

for large r. Thus,

$$M(r, \phi) < r^{\rho_1 + 2\epsilon} \tag{5.2.6}$$

for sufficiently large r. Since $\epsilon > 0$ is arbitrary, so from (5.2.6), it follows that

$$\rho \leq \rho_1 = \max\{\rho_1, \rho_2\}.$$

Similarly, the result follows when $\rho_2 > \rho_1$.

2. Let $\psi(z) = f_1(z) \pm f_2(z)$ be of order ρ and let $\rho_1 \geq \rho_2$. Also, let $M(r, \psi) = \max_{|z|=r} |\psi(z)|$. Since ρ_1 and ρ_2 are orders of f_1 and f_2 respectively, we have for any given $\epsilon > 0$,

$$M(r, f_1) < \mathrm{e}^{r^{\rho_1 + \epsilon}}, \quad \text{and} \quad M(r, f_2) < \mathrm{e}^{r^{\rho_2 + \epsilon}},$$

for sufficiently large r. Now,

$$\begin{aligned} |\psi(z)| \leq |f_1(z)| + |f_2(z)| \Rightarrow M(r, \psi) \quad & \leq \quad M(r, f_1) + M(r, f_2) \\ & < \quad \mathrm{e}^{r^{\rho_1 + \epsilon}} + \mathrm{e}^{r^{\rho_2 + \epsilon}} < 2\,\mathrm{e}^{r^{\rho_1 + \epsilon}} \end{aligned}$$

for sufficiently large r [since exponential function and r^n are both increasing]. Thus,

$$M(r, \psi) < \mathrm{e}^{r^{\rho_1 + 2\epsilon}} \Rightarrow \log M(r) < r^{\rho_1 + 2\epsilon}$$

for all large r. Since $\epsilon > 0$ is arbitrary, it follows that the order of ψ can't exceed ρ_1, that is, $\rho \leq \rho_1 = \max\{\rho_1, \rho_2\}$. Similarly, if $\rho_2 > \rho_1$, the result can be proved.

□

Corollary 5.2.1. If ρ be the order of $f_1 \pm f_2$ and $\rho_1 \neq \rho_2$, then $\rho = \max\{\rho_1, \rho_2\}$.

Proof. Let $\rho_1 > \rho_2$. There exists a sequence $r_n \to \infty$ such that

$$M(r_n, f_1) > \mathrm{e}^{r_n^{\rho_1 - \epsilon}}.$$

Hence,

$$M(r, \psi) \geq \mathrm{e}^{r_n^{\rho_1-\epsilon}} - \mathrm{e}^{r_n^{\rho_2+\epsilon}} = \exp(r_n^{\rho_1-\epsilon})\{1 - \exp(r_n^{\rho_2+\epsilon} - r_n^{\rho_1-\epsilon})\} > \frac{1}{2}\exp(r_n^{\rho_1-\epsilon})$$

if ϵ is chosen so small that $\rho_2 + \epsilon < \rho_1 - \epsilon$ for sufficiently large n. Hence, $\rho \geq \rho_1$. But already we have, $\rho \leq \rho_1$. Thus, $\rho = \rho_1 = \max\{\rho_1, \rho_2\}$. □

Corollary 5.2.2. If ρ be the order of $f_1 f_2$ and $\rho_1 \neq \rho_2$, then $\rho = \max\{\rho_1, \rho_2\}$.

Remark 5.2.1. The result of the above two corollaries are not true if $\rho_1 = \rho_2$. For example, let $f_1(z) = \mathrm{e}^z$ and $f_2(z) = -\mathrm{e}^z$. Then the orders of f_1 and f_2 are both 1. But the order of $f_1 + f_2$ is 0. Similarly, if $f_1(z) = \mathrm{e}^z$ and $f_2(z) = \mathrm{e}^{-z}$ then the orders of f_1 and f_2 are both 1 and the order of $f_1 f_2$ is 0.

Example 5.2.4. Let $P(z)$ be a polynomial of degree n, then the order of $\mathrm{e}^{P(z)}$ is n and the type of $\mathrm{e}^{P(z)}$ is the modulus of the coefficient of the highest degree term in $P(z)$.

Let $P(z) = a_0 + a_1 z + \cdots + a_n z^n$, $a_n \neq 0$ and $f(z) = \exp(a_0 + a_1 z + \cdots + a_n z^n)$. Then,

$$\begin{aligned} M(r, f) = \max_{|z|=r} |f(z)| &= \max_{|z|=r} |\exp(a_0 + a_1 z + \cdots + a_n z^n)| \\ &= \max_{|z|=r} |\mathrm{e}^{a_0} \cdot \mathrm{e}^{a_1 z} \cdots \mathrm{e}^{a_n z^n}| \\ &= \max_{|z|=r} \{|\mathrm{e}^{a_0}| \cdot |\mathrm{e}^{a_1 z}| \cdots |\mathrm{e}^{a_n z^n}|\}. \end{aligned} \tag{5.2.7}$$

Let us find $\max\limits_{|z|=r} |\mathrm{e}^{a_m z^m}|$. Let $a_m = t\,\mathrm{e}^{i\phi}$ and $z = r\,\mathrm{e}^{i\theta}$. Then

$$a_m z^m = t\,\mathrm{e}^{i\phi} \cdot r^m\,\mathrm{e}^{im\theta} = t r^m\,\mathrm{e}^{i(m\theta+\phi)}.$$

Hence,

$$\begin{aligned} \max_{|z|=r} |\mathrm{e}^{a_m z^m}| &= \max_{\theta} |\exp\{t r^m\,\mathrm{e}^{i(m\theta+\phi)}\}| \\ &= \max_{\theta} |\exp(t r^m \{\cos(m\theta + \phi) + i\sin(m\theta + \phi)\})| \\ &= \max_{\theta} |\exp(t r^m \cos(m\theta + \phi))| = \exp(t r^m) = \exp\{|a_m| r^m\}. \end{aligned}$$

Hence, from (5.2.7),

$$M(r, f) = \mathrm{e}^{|a_0|} \cdot \mathrm{e}^{|a_1| r} \cdots \mathrm{e}^{|a_n| r^n} = \mathrm{e}^{|a_0| + |a_1| r + \cdots + |a_n| r^n}.$$

Hence,

$$\log M(r, f) = |a_0| + |a_1| r + \cdots + |a_n| r^n.$$

Hence,

$$\begin{aligned}
\rho = \limsup_{r\to\infty} \frac{\log\log M(r,f)}{\log r} &= \limsup_{r\to\infty} \frac{\log(|a_0| + |a_1|r + \cdots + |a_n|r^n)}{\log r} \\
&= \limsup_{r\to\infty} \frac{\log r^n \left(\frac{|a_0|}{r^n} + \frac{|a_1|}{r^{n-1}} + \cdots + |a_n|\right)}{\log r} \\
&= \limsup_{r\to\infty} \frac{n\log r + \log\left(\frac{|a_0|}{r^n} + \frac{|a_1|}{r^{n-1}} + \cdots + |a_n|\right)}{\log r} \\
&= \limsup_{r\to\infty} \left(n + \frac{\log\left(\frac{|a_0|}{r^n} + \frac{|a_1|}{r^{n-1}} + \cdots + |a_n|\right)}{\log r}\right) = n.
\end{aligned}$$

And,

$$\begin{aligned}
\tau = \limsup_{r\to\infty} \frac{\log M(r)}{r^\rho} &= \limsup_{r\to\infty} \frac{|a_0| + |a_1|r + \cdots |a_n|r^n}{r^n} \\
&= \limsup_{r\to\infty} \left\{\frac{|a_0|}{r^n} + \frac{|a_1|}{r^{n-1}} + \cdots + |a_n|\right\} = |a_n|.
\end{aligned}$$

5.2.4 Order and coefficients in terms of Taylor's Coefficients

Let $f(z) = \sum_{n=0}^{\infty} a_n z^n$ be an entire function with order ρ and type τ. We now state the formulae for order and type of f in terms of the Taylor's coefficients.

Theorem 5.2.6. Let $f(z) = \sum_{n=0}^{\infty} a_n z^n$ be an entire function of finite order ρ. Then

$$\rho = \limsup_{n\to\infty} \frac{\log n}{\log\left(\frac{1}{|a_n|^{1/n}}\right)}.$$

Theorem 5.2.7. Let $f(z) = \sum_{n=0}^{\infty} a_n z^n$ be an entire function of finite order ρ. Then

$$\tau = \frac{1}{\mathrm{e}\,\rho} \limsup_{n\to\infty} n|a_n|^{\rho/n}.$$

Example 5.2.5. Let

$$f(z) = \sum_{n=0}^{\infty} \frac{z^n}{(n!)^\alpha}, \quad \alpha > 0.$$

Let ρ and τ be the order and type respectively for f. Here,

$$a_n = \frac{1}{(n!)^\alpha} \Rightarrow \log\left(\frac{1}{|a_n|}\right) = \alpha \log n!$$

Hence,

$$\begin{aligned}
\rho = \limsup_{r\to\infty} \frac{n\log n}{\alpha \log n!} &= \frac{1}{\alpha}\limsup_{r\to\infty} \frac{(n+1)\log(n+1) - n\log n}{\log(n+1)! - \log n!} \\
&= \frac{1}{\alpha}\limsup_{r\to\infty} \frac{(n+1)\log n(1+1/n) - n\log n}{\log(n+1)} \\
&= \frac{1}{\alpha}\limsup_{r\to\infty} \frac{(n+1)\log n + (n+1)\log(1+1/n) - n\log n}{\log(n+1)} \\
&= \frac{1}{\alpha}\limsup_{r\to\infty} \frac{\log n + \log(1+1/n)^n + \log(1+1/n)}{\log n + \log(1+1/n)} \\
&= \frac{1}{\alpha}\limsup_{r\to\infty} \frac{1 + \frac{\log(1+1/n)^n}{\log n} + \frac{\log(1+1/n)}{\log n}}{1 + \frac{\log(1+1/n)}{\log n}} = \frac{1}{\alpha}.
\end{aligned}$$

Hence $\rho = 1/\alpha$. Thus,

$$\begin{aligned}
\tau = \frac{\alpha}{\mathrm{e}} \limsup_{r\to\infty} n\left(\frac{1}{(n!)^\alpha}\right)^{\frac{1}{n\alpha}} &= \frac{\alpha}{\mathrm{e}} \limsup_{r\to\infty} n\left(\frac{1}{n!}\right)^{\alpha\cdot\frac{1}{n\alpha}} \\
&= \frac{\alpha}{\mathrm{e}} \limsup_{r\to\infty} n\left(\frac{1}{n!}\right)^{\frac{1}{n}} \\
&= \frac{\alpha}{\mathrm{e}} \limsup_{r\to\infty} \left(\frac{n^n}{n!}\right)^{\frac{1}{n}} = \frac{\alpha}{\mathrm{e}}\cdot\alpha = \alpha.
\end{aligned}$$

Hence, $\tau = \alpha$.

Exercise 5.2.1. 1. Find the order and type of the following functions

(a) e^z

(b) $\mathrm{e}^{z^4}\cdot z^4$

(c) $\sin z$

2. Find the order and type of the following functions

$$\sum_{n=0}^{\infty}\left(\frac{z}{n}\right)^n, \quad \sum_{n=0}^{\infty}\left(\frac{\log n}{n}\right)^{n/a} z^n, \ \ a>0.$$

5.3 Few Probable Questions

---

1. Show that for a transcendental entire function f with maximum modulus function $M(r)$,

$$\liminf_{r\to\infty} \frac{\log M(r)}{\log r} = \infty.$$

2. If for an entire function f with maximum modulus $M(r)$, the relation

$$\lim_{r\to\infty} \frac{M(r)}{r^k} < \infty,$$

holds, then show that f is a polynomial of degree at most k.

3. Show that the order ρ of an entire function with maximum modulus function $M(r)$ is given by

$$\rho = \limsup_{r\to\infty} \frac{\log\log M(r)}{\log r}.$$

Hence find the order of $\cos z$.

4. Show that any polynomial is of order zero. Is the converse true? Justify.

5. Define the type of an entire function f having finite non-zero order ρ. Also, show that

$$\tau = \limsup_{r\to\infty} \frac{\log M(r)}{r^\rho}.$$

6. Show that for two entire functions f_1 and f_2 of orders ρ_1 and ρ_2 respectively,

$$\text{order of } f_1 f_2 \leq \max\{\rho_1, \rho_2\}.$$

7. Show that for two entire functions f_1 and f_2 of orders ρ_1 and ρ_2 respectively,

$$\text{order of } f_1 \pm f_2 \leq \max\{\rho_1, \rho_2\}.$$

Unit 6

Course Structure

- Distribution of zeros of entire functions.
- The exponent of convergence of zeros.

6.1 Introduction

The zeroes of entire functions play an important role in determining their growth rates. We will start off with the Jensen's theorems for analytic functions. In complex analysis, Jensen's formula, introduced by Johan Jensen (1899), relates the average magnitude of an analytic function on a circle with the number of its zeros inside the circle. It forms an important statement in the study of entire functions as we will soon come to see.

Objectives

After reading this unit, you will be able to

- study Jensen's theorems and related results
- define convergence exponent of the zeros of an entire function and deduce various related results

6.2 Distribution of zeros of analytic functions

Theorem 6.2.1. (Jensen's theorem) Let f be analytic on $|z| \leq R$, $f(0) \neq 0$ and $f(z) \neq 0$ on $|z| = R$. If $a_1, a_2, \ldots, a_n$ are the zeros of f in $|z| < R$, multiple zeros being repeated, and $|a_i| = r_i$, then

$$\log \frac{R^n}{r_1 r_2 \cdots r_n} = \frac{1}{2\pi} \int_0^{2\pi} \log |f(R\,\mathrm{e}^{i\theta})| d\theta - \log |f(0)|.$$

Proof. Let

$$\begin{aligned} \phi(z) &= f(z) \cdot \frac{R^2 - \overline{a_1} z}{R(z - a_1)} \cdot \frac{R^2 - \overline{a_2} z}{R(z - a_2)} \cdots \frac{R^2 - \overline{a_n} z}{R(z - a_n)} \\ &= f(z) \prod_{k=1}^{n} \frac{R^2 - \overline{a_k} z}{R(z - a_k)}. \end{aligned} \tag{6.2.1}$$

The zeroes of the denominator of ϕ are also the zeros of f of the same order. Hence the zeros of f cancel the poles a_n in the product and so ϕ is analytic on $|z| \leq R$. Also, $\phi(z) \neq 0$ on $|z| \leq R$. Since

$$R^2 - \overline{a_k} z = 0 \Rightarrow z = \frac{R^2}{\overline{a_k}} \Rightarrow |z| = \frac{R^2}{|\overline{a_k}|} = \frac{R^2}{|a_k|} > R$$

since $|a_k| < R$ for all $k = 1, 2, \ldots, n$. Thus, any zero of $\phi(z)$ lies outside the circle $|z| = R$. So, ϕ has neither zeros nor poles in $|z| \leq R$. Thus, the function $\log \phi(z)$ is analytic in $|z| \leq R$. Thus, by Cauchy's Integral theorem, we have

$$\log \phi(0) = \frac{1}{2\pi i} \int_{|z|=R} \frac{1}{z} \log \left(f(z) \prod_{k=1}^{n} \frac{R^2 - \overline{a_k} z}{R(z - a_k)} \right) dz. \tag{6.2.2}$$

On $|z| = R$, we have, $z = R\,\mathrm{e}^{i\theta}$, $\theta \in [0, 2\pi]$, which implies that $dz = R\,\mathrm{e}^{i\theta}\, i d\theta$. Also,

$$|\phi(z)| = |f(z)| \left| \frac{R^2 - \overline{a_1} z}{R(z - a_1)} \right| \cdot \left| \frac{R^2 - \overline{a_2} z}{R(z - a_2)} \right| \cdots \left| \frac{R^2 - \overline{a_n} z}{R(z - a_n)} \right|.$$

On $|z| = R$, we have

$$\left| \frac{R^2 - \overline{a_k} z}{R(z - a_k)} \right| = \left| \frac{z\overline{z} - \overline{a_k} z}{R(z - a_k)} \right| = \frac{|z|}{R} \left| \frac{\overline{z} - \overline{a_k}}{z - a_k} \right| = \left| \frac{\overline{z - a_k}}{z - a_k} \right| = 1$$

and thus,

$$|\phi(z)| = |f(z)|, \quad \text{on } |z| = R.$$

The equation (6.2.2) changes to

$$\begin{aligned} \log \phi(0) &= \frac{1}{2\pi i} \int_0^{2\pi} \frac{1}{R\,\mathrm{e}^{i\theta}} \log \left(f(R\,\mathrm{e}^{i\theta}) \prod_{k=1}^{n} \frac{R^2 - \overline{a_k} z}{R(z - a_k)} \right) R\,\mathrm{e}^{i\theta}\, i d\theta \\ &= \frac{1}{2\pi} \int_0^{2\pi} \left[\log f(R\,\mathrm{e}^{i\theta}) + \sum_{k=1}^{n} \log \left(\frac{R^2 - \overline{a_k} z}{R(z - a_k)} \right) \right] d\theta. \end{aligned} \tag{6.2.3}$$

Taking real parts of equation (6.2.3), we get, by using the conditions deduced in the previous discussions,

$$\log |\phi(0)| = \frac{1}{2\pi} \int_0^{2\pi} \log |f(R\,\mathrm{e}^{i\theta})| d\theta. \tag{6.2.4}$$

Since equation (6.2.1) gives

$$|\phi(0)| = |f(0)| \prod_{k=1}^{n} \frac{R}{|a_k|} = |f(0)| \prod_{k=1}^{n} \frac{R}{r_k} = |f(0)| \frac{R^n}{r_1 r_2 \cdots r_n}$$

Thus,

$$\log |\phi(0)| = \log |f(0)| + \log \frac{R^n}{r_1 r_2 \cdots r_n}$$

and thus, equation (6.2.4) gives

$$\log \frac{R^n}{r_1 r_2 \cdots r_n} = \frac{1}{2\pi} \int_0^{2\pi} \log |f(R\,\mathrm{e}^{i\theta})| d\theta - \log |f(0)|.$$

□

Note 6.2.1. Jensen's theorem can also be written as

$$\log|f(0)| + \sum_{i=1}^{n} \log\frac{R}{|a_k|} = \frac{1}{2\pi}\int_0^{2\pi} \log|f(R\,\mathrm{e}^{i\theta})|d\theta.$$

Theorem 6.2.2. (Jensen's Inequality) Let f be analytic on $|z| \le R$, $f(0) \neq 0$ and $f(z) \neq R$ on $|z| = R$. If $a_1, a_2, \ldots, a_n$ are the zeros of f in $|z| < R$, multiple zeros being repeated, and $|a_i| = r_i$, then

$$\frac{R^n|f(0)|}{r_1 r_2 \cdots r_n} \le M(R).$$

Proof. Let

$$\begin{aligned}\phi(z) &= f(z)\cdot\frac{R^2-\overline{a_1}z}{R(z-a_1)}\cdot\frac{R^2-\overline{a_2}z}{R(z-a_2)}\cdots\frac{R^2-\overline{a_n}z}{R(z-a_n)}\\ &= f(z)\prod_{k=1}^{n}\frac{R^2-\overline{a_k}z}{R(z-a_k)}.\end{aligned}$$

The zeroes of the denominator of ϕ are also the zeros of f of the same order. Hence the zeros of f cancel the poles a_n in the product and so ϕ is analytic on $|z| \le R$. Also,

$$|\phi(z)| = |f(z)|\left|\frac{R^2-\overline{a_1}z}{R(z-a_1)}\right|\cdot\left|\frac{R^2-\overline{a_2}z}{R(z-a_2)}\right|\cdots\left|\frac{R^2-\overline{a_n}z}{R(z-a_n)}\right|.$$

On $|z| = R$, we have

$$\left|\frac{R^2-\overline{a_k}z}{R(z-a_k)}\right| = \left|\frac{z\overline{z}-\overline{a_k}z}{R(z-a_k)}\right| = \frac{|z|}{R}\left|\frac{\overline{z}-\overline{a_k}}{z-a_k}\right| = \left|\frac{\overline{z-a_k}}{z-a_k}\right| = 1$$

and thus,

$$|\phi(z)| = |f(z)|, \quad \text{on } |z| = R.$$

By Maximum modulus theorem, $|\phi(z)| \le M(R)$ for $|z| \le R$. In particular, $|\phi(0)| \le M(R)$, that is,

$$|f(0)|\left|\frac{R}{-a_1}\right|\cdots\left|\frac{R}{-a_n}\right| \le M(R) \Rightarrow \frac{R^n|f(0)|}{r_1 r_2\cdots r_n} \le M(R).$$

□

Definition 6.2.1. Let f be analytic on $|z| \le R$, with zeros at the points $a_1, a_2, \ldots$, arranged in the order of non-decreasing modulus, multiple zeros being repeated. We define the function $n(r)$ as the number of zeros of f in $|z| \le r$, $r \le R$. Evidently, $n(r)$ is a non-negative, non-decreasing function of r which is constant in any interval which does not contain the modulus of a zero of f. Observe that if $f(0) \neq 0$, then $n(r) = 0$ for $r < |a_1|$. Also, $n(r) = n$ for $|a_n| \le r < |a_{n+1}|$.

We will rewrite Jensen's inequality in terms of $n(r)$ as follows.

Theorem 6.2.3. Let f be analytic on $|z| \le R$, $f(0) \neq 0$. Let its zeros, arranged in order of non-decreasing modulus be $a_1, a_2, \ldots$, multiple zeros being repeated according to their multiplicities. If $|a_n| \le r < |a_{n+1}|$, then

$$\int_0^x \frac{n(x)}{x}dx \le \log M(r) - \log|f(0)|.$$

Proof. Let $|a_i| = r_i$, $i = 1, 2, \ldots$, and r be a positive number such that $r_N \leq r < r_{N+1}$, $(r \leq R)$. Let $x_1, x_2, \ldots, x_m$ be the distinct numbers of the set $E = \{r_1, r_2, \ldots, r_N\}$ so that $x_1 = r_1, \ldots, x_m = r_N$. Suppose x_i is repeated p_i times in E. Then $p_1 + \cdots + p_m = N$. Also, $s_i = p_1 + \cdots + p_i$, $i = 1, 2, \ldots, m$. We consider two cases.

Case I: Let $r_N < r$. Then,

$$\begin{aligned}
\int_0^x \frac{n(x)}{x}dx &= \lim_{\epsilon\to 0}\left\{\int_{x_1}^{x_2-\epsilon}\frac{n(x)}{x}dx + \int_{x_2}^{x_3-\epsilon}\frac{n(x)}{x}dx + \cdots + \int_{x_{m-1}}^{x_m-\epsilon}\frac{n(x)}{x}dx\right\} + \int_{x_m}^{r}\frac{n(x)}{x}dx \\
&\qquad\qquad \left[\text{since } \int_0^{x_1-\epsilon}\frac{n(x)}{x}dx = 0 \text{ as } n(x) = 0 \text{ when } 0 \leq x < x_1\right] \\
&= \lim_{\epsilon\to 0}\left\{\int_{x_1}^{x_2-\epsilon}\frac{s_1}{x}dx + \int_{x_2}^{x_3-\epsilon}\frac{s_2}{x}dx + \cdots + \int_{x_{m-1}}^{x_m-\epsilon}\frac{s_{m-1}}{x}dx\right\} + \int_{x_m}^{r}\frac{N}{x}dx \\
&= \lim_{\epsilon\to 0}\left\{[s_1\log x]_{x_1}^{x_2-\epsilon} + [s_2\log x]_{x_2}^{x_3-\epsilon} + \cdots + [s_{m-1}\log x]_{x_{m-1}}^{x_m-\epsilon}\right\} + [N\log x]_{x_N}^{r} \\
&= \lim_{\epsilon\to 0}[s_1\{\log(x_2-\epsilon) - \log x_1\} + s_2\{\log(x_3-\epsilon) - \log x_2\} + \cdots \\
&\qquad +s_{m-1}\{\log(x_m-\epsilon) - \log x_{m-1}\}] + N(\log r - \log r_N) \\
&= s_1(\log x_2 - \log x_1) + s_2(\log x_3 - \log x_2) + \cdots + s_{m-1}(\log x_m - \log x_{m-1}) \\
&\qquad +N(\log r - \log r_N) \\
&= p_1\log x_2 - p_1\log x_1 + (p_1+p_2)\log x_3 - (p_1+p_2)\log x_2 + \cdots \\
&\quad +(p_1+\cdots+p_{m-1})\log x_m - (p_1+\cdots+p_{m-1})\log x_{m-1} \\
&\qquad +N\log r - (p_1+\cdots+p_m)\log x_m \\
&= N\log r - (p_1\log x_1 + p_2\log x_2 + \cdots + p_m\log x_m) \\
&= \log r^N - \log x_1^{p_1}x_2^{p_2}\cdots x_m^{p_m} \\
&= \log\frac{r^N}{x_1^{p_1}x_2^{p_2}\cdots x_m^{p_m}} = \log\frac{r^N}{r_1r_2\cdots r_N}.
\end{aligned}$$

Case II: Let $r_N = r$. Then as before,

$$\begin{aligned}
\int_0^r \frac{n(x)}{x}dx &= \lim_{\epsilon\to 0}\left\{\int_{x_1}^{x_2-\epsilon}\frac{s_1}{x}dx + \int_{x_2}^{x_3-\epsilon}\frac{s_2}{x}dx + \cdots + \int_{x_{m-1}}^{x_m-\epsilon}\frac{s_{m-1}}{x}dx\right\} \\
&= \sum_{i=1}^{m-1} s_i(\log x_{i+1} - \log x_i) + s_m(\log r - \log r_N) \qquad [\text{ since } r = r_N] \\
&= \log\frac{r^N}{r_1\cdots r_N} \qquad (\text{ Proceeding as in Case I})
\end{aligned}$$

Thus, in any case,

$$\int_0^r \frac{n(x)}{x}dx = \log\frac{r^N}{r_1r_2\cdots r_N}$$

But Jensen's inequality gives us

$$\frac{r^N}{r_1r_2\cdots r_N} \leq \frac{M(r)}{|f(0)|}$$

Hence,

$$\int_0^r \frac{n(x)}{x}dx = \log\frac{r^N}{r_1\cdots r^N} \leq \log M(r) - \log|f(0)|$$

□

Note 6.2.2. Jensen's inequality is also true for entire functions where $R \to \infty$.

6.3 Distribution of zeros of entire functions

Theorem 6.3.1. Let f be an entire function with finite order ρ and $f(0) \neq 0$. Then $n(r) = O\left(r^{\rho+\epsilon}\right)$ for any $\epsilon > 0$ and for sufficiently large values of r.

Proof. By Jensen's inequality,

$$\int_0^r \frac{n\left(x\right)}{x} dx \leq \log M\left(r\right) - \log\left|f\left(0\right)\right| \tag{6.3.1}$$

Replacing r by $2r$ in (6.3.1) we get,

$$\int_0^{2r} \frac{n\left(x\right)}{x} dx \leq \log M\left(2r\right) - \log\left|f\left(0\right)\right| \tag{6.3.2}$$

Since ρ is the order of f we have for each $\epsilon > 0$,

$$\log M\left(2r\right) < \left(2r\right)^{(\rho+\epsilon)} = 2^{(\rho+\epsilon)} \cdot r^{(\rho+\epsilon)} = kr^{(\rho+\epsilon)}$$

for all large values of r, where $k = 2^{(\rho+\epsilon)} =$ a constant. Hence from (6.3.2)

$$\int_0^{2r} \frac{n\left(x\right)}{x} dx < Ar^{(\rho+\epsilon)}, \forall \text{ large } r, A = \text{ constant independent of } r.$$

Now, since $n\left(x\right)$ is non-negative and non-decreasing function of x we have,

$$\int_r^{2r} \frac{n\left(x\right)}{x} dx \leq \int_0^{2r} \frac{n\left(x\right)}{x} dx < Ar^{(\rho+\epsilon)}$$

and also

$$\int_r^{2r} \frac{n\left(x\right)}{x} dx \geq \int_r^{2r} \frac{n\left(x\right)}{x} dx = n\left(r\right) \int_r^{2r} \frac{dx}{x} = n\left(r\right) \log 2$$

Hence,

$$n\left(r\right) \log 2 \leq \int_r^{2r} \frac{n\left(x\right)}{x} dx < Ar^{(\rho+\epsilon)};$$

that is,

$$n\left(r\right) < \frac{A}{\log 2} r^{(\rho+\epsilon)}, \text{ for all large } r.$$

Hence, $n\left(r\right) = O\left(r^{(\rho+\epsilon)}\right)$ □

6.3.1 Convergence exponent of zeros of entire functions

Definition 6.3.1. Let f be an entire function with the zeros $z_1, z_2, \ldots$, arranged in order of non-decreasing modulus. We associate with this sequence of zeros a number ρ_1 defined by the equation

$$\rho_1 = \lim_{n\to\infty} \frac{\log n}{\log r_n}$$

$\left[\text{or }, \rho_1 = \lim_{r\to\infty} \frac{\log n\left(r\right)}{\log r}\right]$, where $|z_n| = r_n$. This number ρ_1 is called the convergence exponent or exponent of convergence of the zeros of function f.

Theorem 6.3.2. Let f be an entire function with zeros $z_1, z_2, \ldots$, arranged in order of non-decreasing modulus and $|z_n| = r_n$. If the convergence exponent ρ_1 of the zeros of f be finite, then the series $\sum_{n=1}^{\infty} \frac{1}{{r_n}^{\alpha}}$ converges when $\alpha > \rho_1$ and diverges when $\alpha < \rho_1$. If ρ_1 is infinite, the above series diverges for all positive values of α.

Proof. Let ρ_1 be finite and $\alpha > \rho_1$. Then $\rho_1 < \frac{1}{2}(\rho_1 + \alpha)$. Hence, from the definition of ρ_1 we have, $\frac{\log n}{\log r_n} < \frac{1}{2}(\rho_1 + \alpha)$ for all large n. Hence,

$$\log n < \frac{1}{2}(\rho_1 + \alpha) \log r_n = \log {r_n}^{\frac{1}{2}(\rho_1+\alpha)}$$

that is,

$$n < {r_n}^{\frac{1}{2}(\rho_1+\alpha)}, \text{ or, } n^{\frac{2}{\rho_1+\alpha}} < r_n,$$

$$r_n^{\alpha} > n^{\frac{2\alpha}{\rho_1+\alpha}} = n^{1+\frac{\alpha-\rho_1}{\alpha+\rho_1}} = n^{1+p}, \text{ where } p = \frac{\alpha-\rho_1}{\alpha+\rho_1} > 0$$

Hence, $\frac{1}{{r_n}^{\alpha}} < \frac{1}{n^{1+p}}$ for large n. Hence, $\sum_{n=1}^{\infty} \frac{1}{{r_n}^{\alpha}}$ converges.

Next, let $\alpha < \rho_1$. Then, $\frac{\log n}{\log r_n} > \alpha$ for a sequence of values of n, tending to ∞, that is, $\log n > \alpha \log r_n$ $= \log {r_n}^{\alpha}$. Hence, $n > {r_n}^{\alpha}$, or $\frac{1}{{r_n}^{\alpha}} > \frac{1}{n}$ for a sequence of values of n tending to infinity. Let N be such a value of n for which the above inequality holds, that is, $\frac{1}{{r_N}^{\alpha}} > \frac{1}{N}$ and let m be the least integer greater than $\frac{N}{2}$. Then, since r_n is non-decreasing with n, we have

$$\sum_{N-m}^{N} = \frac{1}{{r_{N-m}}^{\alpha}} + \frac{1}{{r_{N-m+1}}^{\alpha}} + \cdots + \frac{1}{{r_N}^{\alpha}} \geq \frac{1}{{r_N}^{\alpha}} + \cdots + \frac{1}{{r_N}^{\alpha}} = \frac{m+1}{{r_N}^{\alpha}} > \frac{m}{N} > \frac{1}{2}$$

Since these are values of N as large as we please, by Cauchy's principle of convergence, the series $\sum_{n=1}^{\infty} \frac{1}{{r_n}^{\alpha}}$ diverges.

If ρ_1 is infinite, then for any value of α, $\frac{\log n}{\log r_n} > \alpha$ for a sequence of values of n tending to infinity, that is, $\log n > \log {r_n}^{\alpha}$, that is, $n > {r_n}^{\alpha}$ for a sequence of values of n tending to infinity, from which we may similarly conclude that the series $\sum_{n=1}^{\infty} \frac{1}{{r_n}^{\alpha}}$ diverges for any positive α. □

Note 6.3.1. We may also define convergence exponent ρ_1 as the $g.l.b$ of the positive numbers α for which the series $\sum_{n=1}^{\infty} \frac{1}{{r_n}^{\alpha}}$ is convergent. For an entire function with no zeros we define $\rho_1 = 0$ and if the series $\sum_{n=1}^{\infty} \frac{1}{{r_n}^{\alpha}}$ diverges for all positive α, then $\rho_1 = \infty$.

Note 6.3.2. If ρ_1 is finite, the series $\sum_{n=1}^{\infty} \frac{1}{{r_n}^{\rho_1}}$ may be either convergent or divergent. For example, if $r_n = n$ we have, $\rho_1 = \limsup_{n\to\infty} \frac{\log n}{\log r_n} = \limsup_{n\to\infty} \frac{\log n}{\log n} = 1$ and $\sum_{n=1}^{\infty} \frac{1}{{r_n}^{\rho_1}} = \sum_{n=1}^{\infty} \frac{1}{n}$ diverges.

Again, if $r_n = n(\log n)^2$ we have,

$$\rho_1 = \limsup_{n\to\infty} \frac{\log n}{\log r_n} = \limsup_{n\to\infty} \frac{\log n}{\log n + 2\log\log n} = \limsup_{n\to\infty} \frac{1}{1 + 2\frac{\log\log n}{\log n}} = 1$$

and $\sum_{n=1}^{\infty} \frac{1}{{r_n}^{\rho_1}} = \sum_{n=1}^{\infty} \frac{1}{n(\log n)^2}$ converges.

Theorem 6.3.3. If f is an entire function with finite order ρ and $r_1, r_2, \ldots$ are the moduli of the zeros of f, then $\sum_{n=1}^{\infty} \frac{1}{{r_n}^{\alpha}}$ converges if $\alpha > \rho$.

Proof. Let β be a number such that $\rho < \beta < \alpha$. Since $n(r) = O(r^{\rho+\epsilon})$ for any $\epsilon > 0$. We have, $n(r) < Ar^{\beta}$ for all large r, A being a constant.

Putting $r = r_n$, n being large, this inequality gives $n < A{r_n}^{\beta}$, that is, ${r_n}^{\beta} > \frac{n}{A}$, or, $r_n > \frac{n^{\frac{1}{\beta}}}{A^{\frac{1}{\beta}}}$ or, ${r_n}^{\alpha} > \frac{n^{\frac{\alpha}{\beta}}}{A^{\frac{\alpha}{\beta}}} = Bn^{\frac{\alpha}{\beta}}$, B = constant. Hence, $\frac{1}{{r_n}^{\alpha}} < \frac{B_1}{n^{\frac{\alpha}{\beta}}}$ for large n, B_1 = constant. Since $\frac{\alpha}{\beta} > 1$, it follows that $\sum_{n=1}^{\infty} \frac{1}{{r_n}^{\alpha}}$ converges. □

Corollary 6.3.1. Since convergence exponent ρ_1 of the zeros of f is the lower bound of the positive numbers α for which $\sum_{n=1}^{\infty} \frac{1}{{r_n}^{\alpha}}$ is convergent, it follows from the above theorem that $\rho_1 \leq \rho$.

Remark 6.3.1. We can prove that the result $\rho_1 \leq \rho$ without using the last theorem.

Proof. We have

$$\begin{aligned}
\rho_1 &= \limsup_{n\to\infty} \frac{\log n}{\log r_n} \\
&= \limsup_{r\to\infty} \frac{\log n(r)}{\log r} \leq \limsup_{r\to\infty} \frac{\log(Ar^{\rho+\epsilon})}{\log r} \\
&= \limsup_{r\to\infty} \frac{\log A + (\rho+\epsilon)\log r}{\log r}, \quad A = \text{ constant} \\
&= \limsup_{r\to\infty} \left\{ \rho + \epsilon + \frac{\log A}{\log r} \right\} \\
&= \rho + \epsilon, \text{ for any } \epsilon > 0
\end{aligned}$$

Hence, $\rho_1 \leq \rho$. □

Note 6.3.3. Convergence exponent may be 0 or ∞. For example, if $r_n = e^n$, then $\rho_1 = \limsup_{n\to\infty} \frac{\log n}{n} = 0$. Also,if $r_n = \log n$, then

$$\rho_1 = \limsup_{n\to\infty} \frac{\log n}{\log\log n} = \infty$$

We may have, $\rho_1 < \rho$. For example, if $f(z) = e^z$, then $\rho = 1$, $\rho_1 = 0$, since there are no zeros of f. For $\sin z$ or $\cos z$, $\rho = \rho_1 = 1$.

Theorem 6.3.4. Let f be an entire function of finite order. If convergence exponent ρ_1 of the zeros of f is greater than zero,then f has infinite number of zeros.

Proof. If possible, let f has finite number of zeros. Let $r_1, r_2, \ldots, r_N$ be the moduli of the zeros of f arranged in non-decreasing order. The series $\sum_{n=1}^{N} \frac{1}{{r_n}^{\alpha}}$, being a series of finite number of terms, converges for every positive value of α. It follows $\rho_1 = 0$ which contradicts our assumption. Hence f contains infinite number of zeros. □

Note 6.3.4. For an entire function with finite number of zeros, $\rho_1 = 0$.

Example 6.3.1. Find the convergence exponent of the zeros of $\cos z$.

Solution. The zeros of $\cos z$ are $\frac{\pi}{2}, \frac{-\pi}{2}, \frac{3\pi}{2}, \frac{-3\pi}{2}, \ldots$. Now,

$$\begin{aligned}\sum_{n=1}^{\infty} \frac{1}{{r_n}^{\alpha}} &= \left(\frac{2}{\pi}\right)^{\alpha} + \left(\frac{2}{\pi}\right)^{\alpha} + \left(\frac{2}{\pi}\right)^{\alpha} \cdot \frac{1}{3^{\alpha}} + \left(\frac{2}{\pi}\right)^{\alpha} \cdot \frac{1}{3^{\alpha}} + \cdots \\ &= 2\left(\frac{2}{\pi}\right)^{\alpha} \left(1 + \frac{1}{3^{\alpha}} + \frac{1}{5^{\alpha}} + \cdots\right)\end{aligned}$$

The series $\frac{1}{1^{\alpha}} + \frac{1}{3^{\alpha}} + \frac{1}{5^{\alpha}} + \cdots$ converges when $\alpha > 1$ and diverges when $\alpha < 1$. Hence, the lower bound of positive numbers α for which the series $\sum_{n=1}^{\infty} \frac{1}{{r_n}^{\alpha}}$ converges us 1. Hence, $\rho_1 = 1$.

Aliter: The zeros of $\cos z$ are $(2n+1)\frac{\pi}{2}$, $n = 0, \pm 1, \pm 2, \ldots$. Let $z_1 = \frac{\pi}{2},\ z_1' = -\frac{\pi}{2},\ z_2 = \frac{3\pi}{2}$, $z_2' = -\frac{3\pi}{2}, \ldots, z_n = (2n-1)\frac{\pi}{2},\ z_n' = -(2n-1)\frac{\pi}{2}, \ldots$. Hence, $r_1 = |z_1| = |z_1'| = \frac{\pi}{2}$, $r_2 = |z_2| = |z_2'| = \frac{3\pi}{2}, \ldots, r_n = |z_n| = |z_n'| = (2n-1)\frac{\pi}{2}, \ldots$

Hence,

$$\begin{aligned}\rho_1 &= \limsup_{n\to\infty} \frac{\log n}{\log r_n} \\ &= \limsup_{n\to\infty} \frac{\log n}{\log(2n-1)\pi/2} \\ &= \limsup_{n\to\infty} \frac{\log n}{\log(2n-1) + \log \pi/2} \\ &= \limsup_{n\to\infty} \frac{\log n}{\log n(2-1/n) + \log \pi/2} \\ &= \limsup_{n\to\infty} \frac{1}{1 + \frac{\log(2-1/n)}{\log n} + \frac{\log \pi/2}{\log n}} = 1.\end{aligned}$$

■

Theorem 6.3.5. If f is an entire function having no zeros, then f is of the form $f(z) = \mathrm{e}^{g(z)}$, where $g(z)$ is an entire function.

Proof. Since $f(z) \neq 0$ for all $z \in \mathbb{C}$, then the function

$$h(z) = \frac{f'(z)}{f(z)} \tag{6.3.3}$$

is also an entire function. Integrating (6.3.3) along any path joining the two points z_0 and z, we get

$$\int_{z_0}^{z} h(z)dz = \int_{z_0}^{z} \frac{f'(z)}{f(z)}dz = \log f(z) - \log f(z_0),$$

where principal branch of logarithm is taken. Hence,

$$\log f(z) = \log f(z_0) + \int_{z_0}^{z} h(z)dz. \tag{6.3.4}$$

The right hand side of equation (6.3.4) is an entire function, say $g(z)$. Hence,

$$\log f(z) = g(z) \Rightarrow f(z) = \mathrm{e}^{g(z)},$$

where $g(z)$ is an entire function. □

Exercise 6.3.1. Find the convergence exponent of the zeros of $\sin z$.

6.4 Few Probable Questions

1. State and prove Jensen's theorem.
2. State and prove Jensen's inequality.
3. For a function f analytic in $|z| \leq R$, $f(0) \neq 0$, if $a_1, a_2, \ldots$ are its zeros, arranged in the order of non-decreasing modulus, multiple zeros repeated according to their multiplicities, show that, for $|a_n| \leq r < |a_{n+1}|$,
$$\int_0^x \frac{n(x)}{x}dx \leq \log M(r) - \log |f(0)|.$$
4. Show that for an entire function f of finite order ρ and $f(0) \neq 0$, $n(r) = O(r^{\rho+\epsilon})$ for any $\epsilon > 0$ and for sufficiently large values of r.
5. Define convergence exponent ρ_1 of the zeros of an entire function. Show that if r_i be the moduli of the zeros of an entire function f, arranged in order of non-decreasing modulus, then the series $\sum_{n=1}^{\infty} \frac{1}{r_n^{\alpha}}$ converges for $\alpha > \rho_1$ and diverges for $\alpha < \rho_1$.
6. Show that an entire function f having no zeros, is of the form $f(z) = \mathrm{e}^{g(z)}$, where $g(z)$ is an entire function.

Unit 7

Course Structure

- Infinite products and infinite product of functions
- Weierstrass factor theorem.

7.1 Introduction

In this unit, our main objective is to deduce Weierstrass' factorization theorem, which asserts that every entire function can be represented as a (possibly infinite) product involving its zeroes. The theorem may be viewed as an extension of the fundamental theorem of algebra, which asserts that every polynomial may be factored into linear factors, one for each root.

The theorem, which is named for Karl Weierstrass, is closely related to a second result that every sequence tending to infinity has an associated entire function with zeroes at precisely the points of that sequence.

A generalization of the theorem extends it to meromorphic functions and allows one to consider a given meromorphic function as a product of three factors: terms depending on the function's zeros and poles, and an associated non-zero analytic function.

We will start off with the infinite products of complex numbers and study the conditions required for their convergence and thereafter establish the results for factorisation of entire functions in the upcoming units.

Objectives

After reading this unit, you will be able to

- define the conditions of convergence of infinite products and also the infinite product of functions
- define the Weierstrass' primary factors and related results
- learn about the factorisations of entire functions and deduce the Weierstrass' factorisation theorem
- deduce related results for the structures of entire functions

7.2 Infinite Products

An expression of the form

$$\prod_{n=1}^{\infty} u_n = u_1 u_2 \cdots u_n \cdots \tag{7.2.1}$$

where $\{u_n\}$ is a sequence of non-zero complex numbers is called an infinite product. Let $P_n = \prod_{n=1}^{\infty} u_n = u_1 u_2 \cdots u_n$. Then $\{P_n\}$ is called the sequence of partial products of (7.2.1). The infinite product (7.2.1) is said to be convergent if the sequence $\{P_n\}$ converges to a non-zero limit u as $n \to \infty$. If $\lim_{n\to\infty} P_n = u \neq 0$, then u is called the value the value of the infinite product (7.2.1) and we write $\prod_{n=1}^{\infty} u_n = u$. If an infinite product does not converge to a non-zero limit, it is said to be divergent.

However, sometimes we need to modify the definition as follows:

An infinite product $\prod_{n=1}^{\infty} u_n$ is said to be convergent if atmost a finite number of factors u_n are zero and the sequence of partial products formed by the non-vanishing factors tends to a non-zero finite limit.

Analogous to the necessary condition for the convergence of series, we have the following theorem for infinite products.

Theorem 7.2.1. (Necessary Condition of convergence) If an infinite product $\prod_{n=1}^{\infty} u_n$ is convergent, then, $\lim_{n\to\infty} u_n = 1$.

Proof. Let $\prod_{n=1}^{\infty} u_n = u$, then $P_n = u_1 u_2 \cdots u_n \to u \neq 0$ as $n \to \infty$, and $P_{n-1} = u_1 u_2 \cdots u_{n-1} \to u$. Hence

$$\lim_{n\to\infty} u_n = \lim_{n\to\infty} \frac{P_n}{P_{n-1}} = \frac{\lim_{n\to\infty} P_n}{\lim_{n\to\infty} P_{n-1}} = \frac{u}{u} = 1.$$

□

Remark 7.2.1. The condition is however, not sufficient. For example, if we take the product $\prod_{n=1}^{\infty} \frac{n}{n+1}$, then we have,

$$P_n = \frac{1}{2} \cdot \frac{2}{3} \cdots \frac{n-1}{n} \cdot \frac{n}{n+1} = \frac{1}{n+1} \to 0 \quad \text{as } n \to \infty$$

and thus the product is divergent. But, $\lim_{n\to\infty} u_n = \frac{n}{n+1} = 1$. So, the condition in the preceding theorem is not sufficient.

In view of the necessary condition for convergence, we write the general term of the product (7.2.1) in the form $u_n = (1 + a_n),\ (a_n \neq 1)$, so that a necessary condition for convergence of the product $\prod_{n=1}^{\infty}(1 + a_n)$ is $\lim_{n\to\infty} a_n = 0$.

Definition 7.2.1. (Absolute Convergence) An infinite product $\prod_{n=1}^{\infty}(1+a_n)$ is said to be absolutely convergent if the product $\prod_{n=1}^{\infty}(1+|a_n|)$ is convergent.

In case of absolute convergence, the factors of the product can be rearranged arbitrarily without affecting the convergence of the product or changing the value of the product.

Theorem 7.2.2. The infinite product $\prod_{n=1}^{\infty}(1+a_n)$, $(a_n \neq 1)$, converges if and only if the series $\sum_{n=1}^{\infty}\log(1+a_n)$ converges, where each logarithm has its principal value. Also,

$$\prod_{n=1}^{\infty}(1+a_n) = \exp\left(\sum_{n=1}^{\infty}\log(1+a_n)\right).$$

Proof. Let $P_n = \prod_{k=1}^{n}(1+a_k)$, $S_n = \sum_{k=1}^{n}\log(1+a_k)$ and $\lim_{n\to\infty} S_n = S$. Since $\mathrm{e}^{z+w} = \mathrm{e}^z \cdot \mathrm{e}^w$ for all $z, w \in \mathbb{C}$, we have,

$$\mathrm{e}^{S_n} = \mathrm{e}^{\log(1+a_1)} \cdot \mathrm{e}^{\log(1+a_2)} \cdots \mathrm{e}^{\log(1+a_n)} = (1+a_1)(1+a_2)\cdots(1+a_n) = P_n.$$

Since e^z is a continuous function on $\mathbb{C}$, $S_n \to S$ implies $P_n = \mathrm{e}^{S_n} \to \mathrm{e}^S \neq 0$. Hence, if the series $\sum_{n=1}^{\infty}\log(1+a_n)$ converges to S, then the product $\prod_{n=1}^{\infty}(1+a_n)$ converges to e^S.

Conversely, suppose that $P_n \to P(\neq 0)$ as $n \to \infty$. Without any loss of generality, we may assume that $P \notin (-\infty, 0]$. For, if $P \in (-\infty, 0]$, then we may consider in place of $\{1+a_n\}$ a new sequence $\{1+b_n\}$ with $1+b_1 = -(1+a_1)$ and $1+b_n = 1+a_n$ for $n \geq 2$. Then

$$\prod_{n=1}^{\infty}(1+b_n) = -\prod_{n=1}^{\infty}(1+a_n) = -P \notin (-\infty, 0].$$

As $P_n \to P \neq (-\infty, 0]$, we have, $P_n \in \mathbb{C} \setminus (-\infty, 0]$ for large n and since $\log z$ is continuous at P, $\log P_n \to \log P$ as $n \to \infty$. Since $\mathrm{e}^{S_n} = P_n$, we can write

$$S_n = \log P_n + 2\pi k_n i. \tag{7.2.2}$$

Then,

$$\log(1+a_{n+1}) = S_{n+1} - S_n = \log P_{n+1} - \log P_n + 2\pi i(k_{n+1} - k_n), \tag{7.2.3}$$

where k_{n+1} and k_n are integers. Equating the imaginary parts on both sides of (7.2.3), we get

$$\arg(1+a_{n+1}) = \arg P_{n+1} - \arg P_n + 2\pi(k_{n+1} - k_n). \tag{7.2.4}$$

Since the product $\prod_{n=1}^{\infty}(1+a_n)$ converges, we have $(1+a_n) \to 1$ and since $\arg(1+a_n)$ is continuous at 1, we have, $\arg(1+a_n) \to \arg 1 = 0$ as $n \to \infty$. Taking limit as $n \to \infty$ in (7.2.4), it follows that $(k_{n+1} - k_n) \to 0$ as $n \to \infty$. Since k_{n+1} and k_n are integers, there exists an integer N such that $k_{n+1} = k_n = \text{constant} = m$, for $n \geq N$, m being an integer. By (7.2.2), we conclude that $S_n \to \log P_n + 2\pi m i$ as $n \to \infty$ for some integer m. This completes the proof. □

Theorem 7.2.3. If $a_n \geq 0$ for all $n \in \mathbb{N}$, then the product $\prod_{n=1}^{\infty}(1+a_n)$ converges if and only if the series $\sum_{n=1}^{\infty} a_n$ converges.

Proof. Let $P_n = \prod_{k=1}^{n}(1+a_k)$, and $S_n = \sum_{k=1}^{n} a_k$. Since $a_n \geq 0$ for all n, $\{P_n\}$ and $\{S_n\}$ are both increasing sequences. Since $1 + x \leq \mathrm{e}^x$ for $x \geq 0$, we have, $a_1 < 1 + a_1 \leq \mathrm{e}^{a_1}$. Hence,

$$(a_1 + a_2 + \cdots + a_n) < (1+a_1)(1+a_2)\cdots(1+a_n) \leq \mathrm{e}^{a_1+a_2+\cdots+a_n},$$

that is, $S_n < P_n \leq \mathrm{e}^{S_n}$. Thus, by the monotonic bounded principle, the series and the product are both convergent or both divergent according as both are bounded or unbounded. Let S_n be bounded. Then $S_n \leq M$ for some $M > 0$ and for all n. Now, $P_n \leq \mathrm{e}^{S_n}$ and so $P_n \leq \mathrm{e}^M$ for all n, that is P_n is bounded. Thus, both the series and the product are either both bounded or both unbounded. Hence the product $\prod_{n=1}^{\infty}(1+a_n)$ converges if and only if the series $\sum_{n=1}^{\infty} a_n$ converges. □

Corollary 7.2.1. The product $\prod_{n=1}^{\infty}(1+a_n)$ is absolutely convergent if and only if the series $\sum_{n=1}^{\infty} a_n$ is absolutely convergent.

Proof. Since $|a_n| \geq 0$ for all $n \in \mathbb{N}$, so by the previous theorem, the product $\prod_{n=1}^{\infty}(1+|a_n|)$ is convergent if and only if the series $\sum_{n=1}^{\infty} |a_n|$ is convergent. Hence the result. □

Theorem 7.2.4. The three series $\sum_{n=1}^{\infty} |a_n|$, $\sum_{n=1}^{\infty} |\log(1+a_n)|$ and $\sum_{n=1}^{\infty} \log(1+|a_n|)$ either converge or diverge together.

Proof. We know that $\lim_{z\to 0} \dfrac{\log(1+z)}{z} = 1$. Hence for ϵ with $0 < \epsilon < 1$, we get, $\left|\dfrac{\log(1+z)}{z} - 1\right| < \epsilon$ whenever $z \to 0$. The triangle inequality shows that

$$(1-\epsilon) < \left|\frac{\log(1+z)}{z}\right| < 1+\epsilon, \quad \text{whenever } z \to 0$$

that is,

$$(1-\epsilon)|z| < |\log(1+z)| < (1+\epsilon)|z|, \quad \text{whenever } z \to 0. \tag{7.2.5}$$

Similarly, $\lim_{t\to 0} \dfrac{\log(1+t)}{t} = 1$ for $0 < t \leq 1$. Then,

$$(1-\epsilon)t < \log(1+t) < (1+\epsilon)t, \quad \text{whenever } t \to 0. \tag{7.2.6}$$

We now assume that $a_n \to 0$ as $n \to \infty$. Consequently, from (7.2.5) and (7.2.6), we have

$$(1-\epsilon)|a_n| < |\log(1+a_n)| < (1+\epsilon)|a_n|, \quad \text{as } n \to \infty \tag{7.2.7}$$

$$\text{and} \qquad (1-\epsilon)|a_n| < \log(1+|a_n|) < (1+\epsilon)|a_n|, \quad \text{as } n \to \infty. \tag{7.2.8}$$

From (7.2.7) and (7.2.8), by comparison test, for any sequence $\{a_n\}$ convergent to 0, the three series $\sum_{n=1}^{\infty} |a_n|$, $\sum_{n=1}^{\infty} |\log(1+a_n)|$ and $\sum_{n=1}^{\infty} \log(1+|a_n|)$ converge or diverge together. □

7.2.1 Infinite product of functions

We now consider the infinite products whose factors are functions defined on a set.

Given a sequence $\{f_n\}$ of functions defined on some set $E \subseteq \mathbb{C}$, the infinite product $\prod_{n=1}^{\infty}(1+f_n(z))$ is said to be convergent on E if for each $a \in E$,

$$\lim_{n\to\infty} P_n(a) = \lim_{n\to\infty} \prod_{k=1}^{n}(1+f_k(a)) \text{ exists and is non-zero.}$$

And, the infinite product $\prod_{n=1}^{\infty}(1+f_n(z))$ is said to be uniformly convergent to a function $f(z)$ in E if the sequence $\{P_n(z)\}$ of partial products, defined by $P_n(z) = \prod_{k=1}^{n}(1+f_k(z))$ is uniformly convergent to the function $f(z)$ in E, with $f(z) \neq 0$ in E.

Theorem 7.2.5. Let every term of the sequence of functions $\{f_n(z)\}$ be analytic in a region G and suppose the infinite series $\sum_{n=1}^{\infty} \log(1+f_n(z))$ is uniformly convergent on every compact subset of G (in particular, none of the terms $f_n(z)$ can take the value -1 at any point of G). Then the infinite product $\prod_{n=1}^{\infty}(1+f_n(z))$ converges uniformly on every compact subset of G.

Theorem 7.2.6. (M test) Let every term of the sequence of functions $\{f_n(z)\}$ be analytic in a region G and suppose none of the terms $f_n(z)$ takes the value -1 at any point of G. Moreover, suppose that there is a convergent series $\sum_{n=1}^{\infty} M_n$, whose terms are non-negative constants, such that $|f_n(z)| \leq M_n$ for all $z \in G$, and for all $n \geq N$, N being a positive integer. Then the infinite product $\prod_{n=1}^{\infty}(1+f_n(z))$ converges uniformly and absolutely to a non-vanishing analytic function $f(z)$ on every compact subset of G.

Exercise 7.2.1. 1. Prove that every absolutely convergent product is convergent.

2. Prove the following

(a) $\prod_{n=2}^{\infty}\left(1-\frac{1}{n^2}\right) = \frac{1}{2}$

(b) $\prod_{n=1}^{\infty}\left(1+\frac{(-1)^{n-1}}{n}\right) = 1$

7.3 Factorization of Entire functions

In this section, we try to construct an entire function $f(z)$ with the given zeros. Let f be an entire function with inly finite number of zeros $z_1, z_2, \ldots, z_n$ (multiple zeros being repeated according to their multiplicities). Then, the function $\dfrac{f(z)}{(z-z_1)(z-z_2)\cdots(z-z_n)}$ is an entire function with no zeros. Hence, by the last theorem in the preceding unit, $\dfrac{f(z)}{(z-z_1)(z-z_2)\cdots(z-z_n)} = \mathrm{e}^{g(z)}$, where $g(z)$ is an entire function. Then, $f(z) = (z-z_1)(z-z_2)\cdots(z-z_n)\,\mathrm{e}^{g(z)}$, $g(z)$ is an entire function.

If, however, an entire function $f \neq 0$ has an infinite number of zeros, then set of zeros cannot have a limit point in any finite region of $\mathbb{C}$ since such a limit point would be a singularity of f. The only limit point is therefore, the point at infinity. We now consider the construction of an entire function with prescribed infinite number of zeros.

Definition 7.3.1. (Weierstrass' Primary factors): The functions

$$E(z,0) = 1-z, \;\; E(z,p) = (1-z)\exp\left(z + \frac{z^2}{2} + \cdots \frac{z^p}{p}\right), \;\; p = 1, 2, \ldots$$

are called Weierstrass' primary factors. Each primary factor is an entire function with only one zero, a simple zero at $z = 1$.

Lemma 7.3.1. If $|z| \leq \dfrac{1}{2}$, $|\log E(z,p)| \leq 2|z|^{p+1}$.

Proof. First let $|z| < 1$. Then,

$$\begin{aligned}
\log E(z,p) &= \log(1-z) + z + \frac{z^2}{2} + \cdots \frac{z^p}{p} \\
&= \left(-z - \frac{z^2}{2} - \cdots \frac{z^p}{p} - \frac{z^{p+1}}{p+1} - \cdots\right) + \left(z + \frac{z^2}{2} + \cdots \frac{z^p}{p}\right) \\
&= -\frac{z^{p+1}}{p+1} - \frac{z^{p+2}}{p+2} - \cdots
\end{aligned} \tag{7.3.1}$$

Now if $|z| \leq \frac{1}{2}$,

$$\begin{aligned}
|\log E(z,p)| &\leq |z|^{p+1} + |z|^{p+2} + \cdots \\
&= |z|^{p+1}\left(1 + |z| + |z|^2 + \cdots\right) \\
&\leq |z|^{p+1}\left(1 + \frac{1}{2} + \frac{1}{2^2} + \cdots\right) \\
&= 2|z|^{p+1}
\end{aligned}$$

□

Theorem 7.3.1. Weierstrass' Factorization theorem: Let $\{a_n\}$ be an arbitrary sequence of complex numbers whose only limit point is ∞, that is, $a_n \to \infty$ as $n \to \infty$. Then it is possible to construct an entire function $f(z)$ with zeros precisely at these points.

Proof. We may suppose that the origin is not a zero of the entire function $f(z)$ to be constructed so that $a_n \neq 0 \; \forall n$. Because if origin is a zero of $f(z)$ of order n, we need only multiply the constructed function by

z^m. We also arrange the zeros in order of non-decreasing modulus(if several distinct points a_n have the same modulus, we take them in any order) so that $|a_1| \leq |a_2| \leq \cdots$.

Let $|a_n| = r_n$. Since $r_n \to \infty$, we can always find a sequence of positive integers $\{p_n\}$ such that $\sum_{n=1}^{\infty} \left(\frac{r}{r_n}\right)^{p_n}$ converges for all $r > 0$. In fact, if $p_n = n$, for any given value of r, the inequality

$$\left(\frac{r}{r_n}\right)^n < \frac{1}{2^n}$$

holds for all sufficiently large values of n and hence the series is convergent.

Next, we take an arbitrary positive number R and choose the integer N such that $r_N \leq 2R < r_{N+1}$. Then, for $n > N$ and $|z| \leq R$, we have

$$\left|\frac{z}{a_n}\right| \leq \frac{R}{r_n} \leq \frac{R}{r_{N+1}} < \frac{1}{2}.$$

From the previous lemma, we get $\left|\log E\left(\frac{z}{a_n}, p_n - 1\right)\right| \leq 2\left(\frac{R}{r_n}\right)^{p_n}$. By Weierstrass' M-test, the series $\log E\left(\frac{z}{a_n}, p_n - 1\right)$ converges absolutely and uniformly in $|z| \leq R$. This implies that the infinite product $\prod_{n=1}^{\infty} E\left(\frac{z}{a_n}, p_n - 1\right)$ converges absolutely and uniformly in the disk $|z| \leq R$, however large R may be. Hence, the above product represents an entire function, say $G(z)$. Thus $G(z) = \prod_{n=1}^{\infty} E\left(\frac{z}{a_n}, p_n - 1\right)$ with the same value of R. We choose another integer K such that $r_K \leq R < r_{K+1}$. Then each of the functions of the sequence $\prod_{n=1}^{m} E\left(\frac{z}{a_n}, p_n - 1\right)$, $m = K+1, K+2, \ldots$ vanishes at the points $a_1, a_2, \ldots a_k$ and nowhere else in $|z| \leq R$. Hence by **Hurwitz's theorem**(Let each function of the sequence $\{f_n\}$ be analytic in the closed region D and bounded by a closed contour γ. The sequence $\{f_n\}$ converges uniformly to f in D. If $f(z) \neq 0$ on γ, then f and the functions f_n, for all large values of n have the same number of zeros within γ. Also, a zero of f is either a zero of f_n for all sufficiently large values of n, or, else is a limit point of the set of zeros of the functions of the sequence), the only zeros of G in $|z| \leq R$ are $a_1, a_2, \ldots, a_K$. Since R is arbitrary this implies that the only zeros of G are the points of the sequence $\{a_n\}$. Thus, G is our required entire function. Now, if the origin is a zero function of order m of the required entire function $f(z)$, then $f(z)$ is of the form:

$$f(z) = z^m G(z)$$

□

Note 7.3.1. Since there are many possible sequences $\{p_n\}$ in the construction of the function $G(z)$ and ultimately of $f(z)$, the function $f(z)$ is not uniquely determined. Again, for any entire function $g(z)$, $\mathrm{e}^{g(z)}$ is also an entire function without any zeros. Hence, the general form of the required entire function $f(z)$ is of the form:

$$f(z) = z^m \, \mathrm{e}^{g(z)} \, G(z) = z^m \, \mathrm{e}^{g(z)} \prod_{n=1}^{\infty} E\left(\frac{z}{a_n}, p_n - 1\right).$$

Theorem 7.3.2. If f is an entire function and $f(0) \neq 0$, then $f(z) = f(0)G(z)\,\mathrm{e}^{g(z)}$, where $G(z)$ is a product of primary factors and $g(z)$ is an entire function.

Proof. We form $G(z)$ as in the previous theorem from the zeros of f. Let $\phi(z) = \dfrac{f'(z)}{f(z)} - \dfrac{G'(z)}{G(z)}$. Then ϕ is an entire function, since the poles of one term are cancelled by those of the other. Let

$$g(z) = \int_0^z \phi(t)dt.$$

Then $g(z)$ is also an entire function. Now,

$$\begin{aligned} g(z) &= \int_0^z \left(\frac{f'(t)}{f(t)} - \frac{G'(t)}{G(t)}\right) dt \\ &= \int_0^z \frac{d}{dt}\left(\log f(t) - \log G(t)\right) dt \\ &= \left[\log f(t) - \log G(t)\right]_0^z \\ &= \log f(z) - \log f(0) - \log G(z) + \log G(0) \\ &= \log f(z) - \log f(0) - \log G(z) \quad [\text{since} \quad \log G(0) = 1] \\ &= \log \frac{f(z)}{f(0)G(z)}. \end{aligned}$$

Hence,

$$\mathrm{e}^{g(z)} = \frac{f(z)}{f(0)G(z)} \Rightarrow f(z) = f(0)G(z)\,\mathrm{e}^{g(z)}\,.$$

□

Theorem 7.3.3. If the real part of an entire function f satisfies the inequality Re $f < r^{k+\epsilon}$ for any $\epsilon > 0$ and for a sequence of values of r tending to infinity, then f is a polynomial of degree not exceeding k.

Proof. By Taylor's theorem, $f(z) = \sum_{n=0}^{\infty} a_n z^n$, where $a_n = \dfrac{1}{2\pi i}\displaystyle\int_\gamma \frac{f(z)}{z^{n+1}}dz$, $\gamma : |z| = r$, r being a positive number. On γ, $z = r\,\mathrm{e}^{i\theta}$, $0 \le \theta \le 2\pi$. Now, when $n > 0$,

$$\begin{aligned} \int_\gamma \frac{\overline{f(z)}}{z^{n+1}}dz &= \int_\gamma \sum_{m=0}^{\infty} \overline{a}_m \overline{z}^m \frac{dz}{z^{n+1}} \\ &= \sum_{m=0}^{\infty} \int_\gamma \overline{a}_m \overline{z}^m \frac{dz}{z^{n+1}} \\ &= \sum_{m=0}^{\infty} \int_0^{2\pi} \frac{\overline{a}_m r^m\,\mathrm{e}^{-im\theta} \cdot ir\,\mathrm{e}^{i\theta}}{r^{n+1}\,\mathrm{e}^{i(n+1)\theta}} d\theta \\ &= \sum_{m=0}^{\infty} \int_0^{2\pi} \overline{a}_m r^{m-n}\,\mathrm{e}^{-i(m+n)\theta}\, id\theta = 0, \end{aligned}$$

the term by term integration is valid since the series $\displaystyle\sum_{m=0}^{\infty} \overline{a}_m \overline{z}^m$ converges uniformly. Hence, for $n > 0$,

$$\begin{aligned} a_n &= \frac{1}{2\pi i}\int_\gamma \frac{f(z)}{z^{n+1}}dz + \frac{1}{2\pi i}\int_\gamma \frac{\overline{f(z)}}{z^{n+1}}dz \\ &= \frac{1}{2\pi i}\int_\gamma \left(f(z) + \overline{f(z)}\right)\frac{dz}{z^{n+1}} \\ &= \frac{1}{\pi i}\int_\gamma \text{Re } f(z)\frac{dz}{z^{n+1}} = \frac{1}{\pi}\int_0^{2\pi} \frac{\text{Re } f(r\,\mathrm{e}^{i\theta})}{r^n\,\mathrm{e}^{in\theta}} d\theta. \end{aligned}$$

Hence,

$$a_n r^n = \frac{1}{\pi}\int_0^{2\pi} \frac{\text{Re } f(r\,\mathrm{e}^{i\theta})}{\mathrm{e}^{in\theta}} d\theta. \Rightarrow |a_n| r^n \leq \frac{1}{\pi}\int_0^{2\pi} |\text{Re } f(r\,\mathrm{e}^{i\theta})| d\theta, \text{ for } n > 0.$$

Also,

$$a_0 = \frac{1}{2\pi i}\int_\gamma \frac{f(z)}{z} dz = \frac{1}{2\pi}\int_0^{2\pi} f(r\,\mathrm{e}^{i\theta}) d\theta$$

and so,

$$\text{Re } a_0 = \frac{1}{2\pi}\int_0^{2\pi} \text{Re } f(r\,\mathrm{e}^{i\theta}) d\theta.$$

Hence,

$$\begin{aligned} 2\text{Re } a_0 + |a_n| r^n \leq \frac{1}{\pi}\int_0^{2\pi} (|\text{Re } f| + \text{Re } f)\, d\theta &= 2\text{Re } f; \text{ if Re } f > 0 \\ &= 0; \text{ if Re } f \leq 0. \end{aligned}$$

Since Re $f < r^{k+\epsilon}$ for any $\epsilon > 0$ and for a sequence of values $\{r_m\}$ of r tending to infinity, we have,

$$2\text{Re } a_0 + |a_n| r_m^n < 4 r_m^{k+\epsilon} \Rightarrow |a_n| < 4 r_m^{k+\epsilon-n} - 2\text{Re } a_0 r_m^{-n}$$

for any $\epsilon > 0$. Taking limit as $r_m \to \infty$, we have, $a_n = 0$ when $n > k$ and so, f is a polynomial of degree not exceeding k. □

Theorem 7.3.4. The function $\mathrm{e}^{f(z)}$ is an entire function of finite order with no zeros if and only if $f(z)$ is a polynomial.

Proof. We already know that $\mathrm{e}^{f(z)}$ is an entire function with no zeros if and only if f is an entire function. Moreover, if f is a polynomial of degree k, then $\mathrm{e}^{f(z)}$ is of finite order k.

Conversely, we assume that $\mathrm{e}^{f(z)}$ is an entire function with finite order ρ and without any zeros. Then, f is an entire function. Also, $\left|\mathrm{e}^{f(z)}\right| < \mathrm{e}^{r^{\rho+\epsilon}}$, for all $r > r_0$, $r = |z|$ and $\epsilon > 0$ is arbitrary, that is,

$$\mathrm{e}^{\text{Re } f} < \mathrm{e}^{r^{\rho+\epsilon}} \Rightarrow \text{Re } f < r^{\rho+\epsilon}, \ \forall r > r_0 \text{ and } \epsilon > 0.$$

Hence, by the previous theorem, f is a polynomial of degree not exceeding ρ. □

7.4 Few Probable Questions

1. When is an infinite product $\prod_{n=1}^{\infty} u_n$, $u_n \neq 0$ for all n, said to be convergent. Deduce a necessary condition for the convergence of the product.
2. Show that if an infinite product $\prod_{n=1}^{\infty} u_n$, $u_n \neq 0$ for all n, is convergent, then $\lim_{n\to\infty} u_n = 1$. Is the condition sufficient? Justify your answer.
3. Show that the infinite product $\prod_{n=1}^{\infty}(1 + a_n)$, $(a_n \neq -1)$ converges if and only if the series $\sum_{n=1}^{\infty} \log(1 + a_n)$ converges and
$$\prod_{n=1}^{\infty}(1 + a_n) = \exp\left(\sum_{n=1}^{\infty} \log(1 + a_n)\right).$$

4. Show that the infinite product $\prod_{n=1}^{\infty}(1+a_n)$, $(a_n \geq 0)$ converges if and only if the series $\sum_{n=1}^{\infty} a_n$ converges.

5. State and prove Weierstrass' Factorization theorem.

6. If for an entire function f satisfies the inequality Re $f < r^{k+\epsilon}$ for any $\epsilon > 0$ and a sequence of values of r tending to infinity, then show that f is a polynomial of degree not exceeding k.

Unit 8

Course Structure

- Canonical product, Borel's first theorem. Borel's second theorem (statement only),
- Hadamard's factorization theorem,
- Schottky's theorem (no proof), Picard's first theorem.

8.1 Introduction

This unit deals with the factorization of entire functions of finite order with the help of the newly defined canonical product. Hence we have deduced several results related to the relationship between the order of an entire function f and the convergence exponent of its zeros culminating in the Picard's little theorem. Little Picard's theorem, named after Charles E. Picard, is an important theorem that puts light on the range set of a non-constant entire function. We will discuss few examples in this light.

Objectives

After reading this unit, you will be able to

- define the canonical product and genus of entire functions
- deduce the Borel's first theorem
- deduce Hadamard's factorization theorem
- deduce related results and the celebrated Picard's little theorem

8.2 Canonical Product

Let f be an entire function with infinite number of zeros a_n, $n = 1, 2, \ldots$ where $a_n \neq 0$ for all n and $|a_n| = r_n$. If there exists a least non-negative integer p such that the series $\sum_{n=1}^{\infty} \frac{1}{r_n^{p+1}}$ is convergent. We

form the infinite product $G(z) = \prod_{n=1}^{\infty} E\left(\frac{z}{a_n}, p\right)$. By Weierstrass' factorization theorem, $G(z)$ represents an entire function having zeros precisely at the points a_n. We call $G(z)$ as the canonical product formed with the sequence $\{a_n\}$ of zeros of f and the integer p is called its **genus**.

If $z = 0$ is a zero of f of order m, then the canonical product is $z^m G(z)$.

Observe that, if the convergence exponent $\rho_1 \neq$ an integer, then $p = [\rho_1]$ and if $\rho_1 =$ an integer, then

1. $p = \rho_1$, if $\sum_{n=1}^{\infty} \frac{1}{r_n^{\rho_1}}$ is divergent;

2. $p = \rho_1 - 1$, if $\sum_{n=1}^{\infty} \frac{1}{r_n^{\rho_1}}$ is convergent.

In any case, $\rho_1 - 1 \leq p \leq \rho_1 \leq \rho$, where $\rho =$ order of $f(z)$.

Example 8.2.1.

1. Let $a_n = n$. Then $\sum_{n=1}^{\infty} \frac{1}{r_n^2} = \sum_{n=1}^{\infty} \frac{1}{n^2}$ is convergent, while $\sum_{n=1}^{\infty} \frac{1}{r_n} = \sum_{n=1}^{\infty} \frac{1}{n}$ is divergent. Hence, the genus $p = 1$.

2. Let $a_n = \mathrm{e}^n$. Then the genus is $p = 0$.

3. Let $a_1 = \frac{1}{2}\log 2$, $a_n = \log n$, $n \geq 2$. Then there exists no finite p such that the series $\sum_{n=1}^{\infty} \frac{1}{r_n^{p+1}}$ is convergent.

Theorem 8.2.1. (Borel's theorem) The order of a canonical product is equal to the convergence exponent of its zeros.

Proof. Let

$$G(z) = \prod_{n=1}^{\infty} E\left(\frac{z}{a_n}, p\right)$$

be a canonical product with zeros at the points $a_1, a_2, \ldots$ and genus p. Let ρ_1 and ρ be the convergence exponent and order respectively of $G(z)$. Since $\rho_1 \leq \rho$ for any entire function, we only need to prove that $\rho \leq \rho_1$ for $G(z)$. Let $|a_n| = r_n$ and $K(> 1)$ be a constant. Then for $|z| = r$,

$$\log|G(z)| = \sum_{r_n \leq Kr} \log\left|E\left(\frac{z}{a_n}, p\right)\right| + \sum_{r_n > Kr} \log\left|E\left(\frac{z}{a_n}, p\right)\right| = \Sigma_1 + \Sigma_2 \text{ (say)}.$$

We first estimate Σ_2. In Σ_2, $\frac{r}{r_n} < \frac{1}{K} < 1$. Hence,

$$\begin{aligned}
\log E\left(\frac{z}{a_n}, p\right) &= -\frac{1}{p+1}\left(\frac{z}{a_n}\right)^{p+1} - \frac{1}{p+2}\left(\frac{z}{a_n}\right)^{p+2} - \cdots \\
\Rightarrow \left|\log E\left(\frac{z}{a_n}, p\right)\right| &< \frac{1}{p+1}\left\{\left(\frac{r}{r_n}\right)^{p+1} + \left(\frac{r}{r_n}\right)^{p+2} + \cdots\right\} \\
&= \frac{1}{p+1}\left(\frac{\left(\frac{r}{r_n}\right)^{p+1}}{1 - \frac{r}{r_n}}\right) < A\left(\frac{r}{r_n}\right)^{p+1},
\end{aligned}$$

A being a constant. Also, we know that $\log|f| = \text{Re}(\log f) \le |\log f|$ [since $\log f(z) = \log|f(z)| + i\arg f(z)$], for any function f.

Hence,

$$\Sigma_2 = \sum_{r_n > Kr} \log\left|E\left(\frac{z}{a_n}, p\right)\right| = O\left(\sum_{r_n>Kr}\left(\frac{r}{r_n}\right)^{p+1}\right) = O\left(r^{p+1}\sum_{r_n>Kr}\frac{1}{{r_n}^{p+1}}\right) = O\left(r^{p+1}\right)$$

$\left[\text{since } \sum_{r_n>Kr}\frac{1}{{r_n}^{p+1}} \text{ is convergent by the definition of } p \text{ and converges to } B(\text{say})\right]$. If $p+1 = \rho_1$, then $\sum_2 = O\left(r^{\rho_1}\right)$. Otherwise, $p+1 > \rho_1 + \epsilon$, $\epsilon > 0$ being small enough. Then

$$r^{p+1}\sum_{r_n>Kr} r_n^{-p-1} = r^{p+1}\sum_{r_n>Kr}\left(r_n^{\rho_1+\epsilon-p-1}r_n^{-\rho_1-\epsilon}\right) < r^{p+1}(Kr)^{\rho_1-\epsilon-1}\sum_{r_n>Kr} r_n^{-\rho_1-\epsilon} = O\left(r^{\rho_1+\epsilon}\right)$$

$\left[\text{since } \sum_{r_n>Kr} r_n^{-\rho_1-\epsilon} \text{ is convergent}\right]$. Thus, in any case,

$$\Sigma_2 = O\left(r^{\rho_1+\epsilon}\right). \tag{8.2.1}$$

Next, we estimate $\sum_1$. In $\sum_1$, $\frac{r}{r_n} \ge \frac{1}{K}$. Now,

$$\begin{aligned}\log\left|E\left(\frac{z}{a_n}, p\right)\right| &= \log\left|\left(1-\frac{z}{a_n}\right)\exp\left(\frac{z}{a_n} + \ldots\frac{1}{p}\left(\frac{z}{a_n}\right)^p\right)\right| \\ &\le \log\left(1+\frac{r}{r_n}\right) + \frac{r}{r_n} + \cdots\frac{1}{p}\left(\frac{r}{r_n}\right)^p.\end{aligned}$$

Also, $\log\left(1+\frac{r}{r_n}\right) < \frac{r}{r_n}$ [since $1+|x| < \mathrm{e}^{|x|}$, hence $\log(1+|x|) < |x|$]. Hence, $\log\left|E\left(\frac{z}{a_n}, p\right)\right| < A\left(\frac{r}{r_n}\right)^p$ where A depends only on K. Hence $\sum_1 = O\left(\sum_{r_n\le Kr}\left(\frac{r}{r_n}\right)^p\right) = O\left(r^p\sum_{r_n\le Kr}{r_n}^{-p}\right)$

$= O\left(r^p\sum_{r_n\le Kr} r_n^{\rho_1+\epsilon-p}\cdot r_n^{-\rho_1-\epsilon}\right) = O\left(r^p(Kr)^{\rho_1+\epsilon-p}\cdot\sum_{r_n\le Kr} r_n^{-\rho_1-\epsilon}\right) = O\left(r^{\rho_1+\epsilon}\right)$. Using this and equation (8.2.1), we get, $\log|G(z)| = O(r^{\rho_1+\epsilon})$. This implies that $\rho \le \rho_1$. Combining, we have $\rho = \rho_1$. □

Example 8.2.2. We find the canonical product of $\sin z$.

Let $f(z) = \sin z$. Then f is an entire function with infinite number of zeros at $z = n\pi$, n being an integer. First we consider the zeros of f exceeding the simple zero at $z = 0$. Let $a_n = n\pi$, $n = \pm1, \pm2, \ldots$ and $|a_n| = r_n = |n\pi|$. Now,

$$\sum_{n=1}^{\infty}\frac{1}{r_n} = \sum_{n=1}^{\infty}\frac{1}{|n\pi|} = \frac{1}{\pi}\sum_{n=1}^{\infty}\frac{1}{n}$$

is divergent, but

$$\sum_{n=1}^{\infty}\frac{1}{r_n^2} = \frac{1}{\pi^2}\sum_{n=1}^{\infty}\frac{1}{n^2}$$

is convergent. Hence the genus of the required canonical product is 1. Hence the canonical product is given by

$$\begin{aligned} G(z) &= \prod_{n=-\infty}^{-1}\left(1-\frac{z}{a_n}\right)\mathrm{e}^{\frac{z}{a_n}}\cdot\prod_{n=1}^{\infty}\left(1-\frac{z}{a_n}\right)\mathrm{e}^{\frac{z}{a_n}} \\ &= \prod_{n=1}^{\infty}\left(1-\frac{z}{n\pi}\right)\mathrm{e}^{\frac{z}{n\pi}}\left(1+\frac{z}{n\pi}\right)\mathrm{e}^{-\frac{z}{n\pi}} \\ &= \prod_{n=1}^{\infty}\left(1-\frac{z^2}{n^2\pi^2}\right). \end{aligned}$$

Now, since zero is a simple zero of f, the canonical product of f will be

$$G(z) = z\prod_{n=1}^{\infty}\left(1-\frac{z^2}{n^2\pi^2}\right).$$

Exercise 8.2.1. Find the canonical product of $\cos z$.

8.3 Hadamard's Factorization theorem and results

Theorem 8.3.1. (Hadamard's Factorization theorem) If f is an entire function of order ρ with infinite number of zeros and $f(0)\neq 0$, then $f(z)=\mathrm{e}^{Q(z)}\,G(z)$, where $G(z)$ is the canonical product formed with the zeros of f and $Q(z)$ is a polynomial of degree not greater than ρ.

Proof. By Weierstrass' factorization theorem, we have,

$$f(z)=\mathrm{e}^{Q(z)}\,G(z), \tag{8.3.1}$$

where $Q(z)$ is an entire function and $G(z)$ is the canonical product with genus p formed with the zeros $a_1,a_2,\ldots$ of f. Since ρ is finite, we need to show that $Q(z)$ is a polynomial of degree less than or equal to ρ. Let $m=[\rho]$. Then genus $p\leq m$. Taking logarithms on both sides of (8.3.1), we get

$$\begin{aligned} \log f(z) &= Q(z)+\log G(z) \\ &= Q(z)+\sum_{n=1}^{\infty}\log\left(1-\frac{z}{a_n}\right)+\sum_{n=1}^{\infty}\left\{\frac{z}{a_n}+\frac{1}{2}\left(\frac{z}{a_n}\right)^2+\cdots+\frac{1}{p}\left(\frac{z}{a_n}\right)^p\right\} \qquad (8.3.2) \\ &\quad\left[\text{since } G(z)=\prod_{n=1}^{\infty}E\left(\frac{z}{a_n},p\right)=\prod_{n=1}^{\infty}\left(1-\frac{z}{a_n}\right)\exp\left(\frac{z}{a_n}+\cdots+\frac{1}{p}\left(\frac{z}{a_n}\right)^p\right)\right]. \end{aligned}$$

Differentiating (8.3.2) $m+1$ times, we get

$$\begin{aligned} &\frac{d^m}{dz^m}\left(\frac{f'(z)}{f(z)}\right)=Q^{(m+1)}(z)-m!\sum_{n=1}^{\infty}\frac{1}{(a_n-z)}^{m+1} \qquad (8.3.3) \\ &\left[\text{since } p\leq m,\ \frac{d^{m+1}}{dz^{m+1}}\sum_{n=1}^{\infty}\left\{\frac{z}{a_n}+\frac{1}{2}\left(\frac{z}{a_n}\right)^2+\cdots+\frac{1}{p}\left(\frac{z}{a_n}\right)^p\right\}=0\right] \\ &\left[\text{and } \frac{d^{m+1}}{dz^{m+1}}\log\left(1-\frac{z}{a_n}\right)=\frac{d^{m+1}}{dz^{m+1}}\log(a_n-z)=-m!\frac{1}{(a_n-z)}^{m+1}\right]. \end{aligned}$$

Now, $Q(z)$ will be a polynomial of degree n at most if we can show that $Q^{(m+1)}(z) = 0$.

Let

$$g_R(z) = \frac{f(z)}{f(0)} \prod_{|a_n| \le R} \left(1 - \frac{z}{a_n}\right)^{-1}.$$

Then since $f(z)$ is entire and $f(0) \neq 0$ and $\prod_{|a_n| \le R} \left(1 - \frac{z}{a_n}\right)^{-1} g_R(z)$ cancels with the factors in $f(z)$, so $g_R(z)$ is an entire function and $g_R(z) \neq 0$ in $|z| \le R$. For $|z| = 2R$ and $|a_n| \le R$, we have, $\left|1 - \frac{z}{a_n}\right| \ge 1$. Hence,

$$|g_R(z)| \le \frac{|f(z)|}{|f(0)|} < A \exp\left((2R)^{\rho+\epsilon}\right), \quad \text{for } |z| = 2R, \ A = \text{constant}.$$

By maximum modulus theorem,

$$|g_R(z)| < A \exp\left((2R)^{\rho+\epsilon}\right), \quad \text{for } |z| < 2R. \tag{8.3.4}$$

Let $h_R(z) = \log g_R(z)$, the logarithm being determined such that $h_R(0) = 0$. Then $h_R(z)$ is analytic in $|z| \le R$. Now, from (8.3.4), we have

$$\text{Re } h_R(z) = \log |g_R(z)| < \log A + 2^{\rho+\epsilon} R^{\rho+\epsilon}.$$

Hence,

$$\text{Re } h_R(z) < K R^{\rho+\epsilon}, \quad K = \text{constant}. \tag{8.3.5}$$

Hence, from the second corollary of Borel Caratheodory theorem, we have,

$$|h_R^{(m+1)}(z)| \le \frac{2^{m+3}(m+1)! \cdot R}{(R-r)^{m+2}} K R^{\rho+\epsilon}, \quad \text{for } |z| = r < R.$$

Hence, for $|z| = r = \frac{R}{2}$,

$$h_R^{(m+1)}(z) = O\left(R^{\rho+\epsilon-m-1}\right). \tag{8.3.6}$$

But,

$$h_R(z) = \log g_R(z) = \log f(z) - \log f(0) - \sum_{|a_n| \le R} \log\left(1 - \frac{z}{a_n}\right).$$

Hence,

$$h_R^{(m+1)}(z) = \frac{d^m}{dz^m}\left(\frac{f'(z)}{f(z)}\right) + m! \sum_{|a_n| \le R} \frac{1}{(a_n - z)^{m+1}}.$$

Hence, from (8.3.2), we have

$$\begin{aligned} Q^{(m+1)}(z) &= h_R^{(m+1)}(z) + m! \sum_{|a_n| \le R} \frac{1}{(a_n - z)^{m+1}} \\ &= O\left(R^{\rho+\epsilon-m-1}\right) + O\left(\sum_{|a_n| > R} \frac{1}{|a_n|^{m+1}}\right), \quad \text{for } |z| = \frac{R}{2} \text{ [by (8.3.6)]} \end{aligned} \tag{8.3.7}$$

and so also for $|z| < \frac{R}{2}$ by maximum modulus theorem. The first term on the right hand side of (8.3.7) tends to zero as $R \to \infty$ if $\epsilon > 0$ is small enough, since $m + 1 > \rho$. Also, the second term tends to 0

since $\sum_{n=1}^{\infty} \frac{1}{|a_n|^{m+1}}$ is convergent. In fact, $\sum_{|a_n|>R} \frac{1}{|a_n|^{m+1}}$ becomes the remainder term for large R. Hence, $Q^{(m+1)}(z) = 0$, since $Q^{(m+1)}$ is independent of R. Thus, $Q(z)$ is a polynomial of degree not greater than ρ. □

Note 8.3.1. If $z = 0$ is a zero of f of order m, then $f(z) = z^m \, \mathrm{e}^{Q(z)} \, G(z)$.

Theorem 8.3.2. If f is an entire function of order ρ and ρ_1 is the convergence exponent of its zeros, then $\rho_1 = \rho$ if ρ is not an integer.

Proof. Since the zeros of f coincide with the zeros of its canonical product $G(z)$, we can take ρ_1 to be the convergence exponent of the zeros of $G(z)$. By Hadamard's factorization theorem, we have, $f(z) = \mathrm{e}^{Q(z)} \, G(z)$, where $Q(z)$ is a polynomial of degree not exceeding ρ. In any case, $\rho_1 \leq \rho$. Suppose if possible $\rho_1 < \rho$. Also, if degree of $Q(z) = q$, then $\mathrm{e}^{Q(z)}$ is of order $q \leq \rho$. In this case, $q < \rho$, since q is an integer and ρ is not an integer. Thus, f is the product of two entire functions each of order less than ρ. Hence, order of f is less than ρ which contradicts the given hypothesis. Hence, $\rho_1 = \rho$. □

Theorem 8.3.3. Let f be an entire function with order ρ and g be an entire function with order $\leq \rho$. If the zeros of g are all zeros of f, then $H(z) = \dfrac{f(z)}{g(z)}$ is an entire function of order ρ at most.

Proof. Since the zeros of g are all zeros of f, $H(z)$ is an entire function. Let $G_1(z)$ and $G_2(z)$ be the canonical products formed with the zeros of f and g respectively. By Hadamard's factorization theorem, we have

$$f(z) = G_1(z) \, \mathrm{e}^{Q_1(z)} \quad \text{and} \quad g(z) = G_2(z) \, \mathrm{e}^{Q_2(z)},$$

where $Q_1(z)$ and $Q_2(z)$ are polynomials with degrees less than or equal to ρ. Then,

$$H(z) = G(z) \, \mathrm{e}^{Q_1(z) - Q_2(z)}, \quad \text{where } G(z) = \frac{G_1(z)}{G_2(z)}$$

is the canonical product formed with the zeros of $G_1(z)$ that are not zeros of $G_2(z)$. Since the convergence exponent of a sequence is not increased by removing some of the terms, the convergence exponent and hence the order of $G(z)$ does not exceed ρ. Also, $Q_1(z) - Q_2(z)$ is a polynomial of degree $\leq \rho$. Hence, order of $\mathrm{e}^{Q_1(z) - Q_2(z)}$ is $\leq \rho$. Thus, H is the product of two entire functions, each of order $\leq \rho$. Hence, order of $H(z)$ is ρ at most. □

Theorem 8.3.4. (Picard's little theorem) An entire function of finite order takes any complex number except at most one number.

Proof. Let f be an entire function of finite order. If possible, let f do not take two values a and b. Then $f(z) - a \neq 0$ and $f(z) - b \neq 0$ for all $z \in \mathbb{C}$. Thus, there exists an entire function g such that $f(z) - a = \mathrm{e}^{g(z)}$. Since f is of finite order, the function $f(z) - a$ is also of finite order. By Hadamard's factorization theorem, $g(z)$ must be a polynomial. Now,

$$f(z) - b = f(z) - a + (a - b) = \mathrm{e}^{g(z)} + (a - b) \neq 0 \ \ \forall z \in \mathbb{C}.$$

Hence, $\mathrm{e}^{g(z)} \neq b - a$ for all $z \in \mathbb{C}$. This is a contradiction, since $g(z)$ being a polynomial, by fundamental theorem of algebra, $g(\mathbb{C}) = \mathbb{C}$. Hence the theorem. □

Example 8.3.1.
1. The most common example is the non-constant entire function e^z which omits only the value 0.

2. Any non-constant polynomial f takes all the values of the finite complex plane. This is due to the fact that for any complex number a, the function $f(z) - a$ is also a polynomial in $\mathbb{C}$ having zero in $\mathbb{C}$ by the fundamental theorem of algebra.

Theorem 8.3.5. Let f be an entire function of finite order ρ which is not an integer. Then f has infinitely many zeros.

Proof. Let f be an entire function of finite order ρ which is not an integer. If possible, suppose that the zeros of f are $\{a_1, a_2, \ldots, a_n\}$, finite in number, counted according to multiplicities. Then $f(z)$ can be expressed as $f(z) = (z - a_1)(z - a_2)\cdots(z - a_n)\,\mathrm{e}^{g(z)}$, where $g(z)$ is an entire function. By Hadamard's factorization theorem, $g(z)$ is a polynomial whose degree is less than or equal to ρ. Clearly, $f(z)$ and $\mathrm{e}^{g(z)}$ are of same order. But the order of $\mathrm{e}^{g(z)}$ is exactly the degree of $g(z)$, which is an integer. This implies that ρ is an integer. This is a contradiction and hence the result. □

Example 8.3.2. If $\alpha > 1$, we show that the entire function $\prod_{n=1}^{\infty}\left(1 - \frac{z}{n^\alpha}\right)$ is of finite order $\frac{1}{\alpha}$. Firstly, the infinite product $\prod_{n=1}^{\infty}\left(1 - \frac{z}{n^\alpha}\right)$ is of the form $\prod_{n=1}^{\infty} E\left(\frac{z}{n^\alpha}, 0\right)$. Here, $p = 0$ is the least non-negative integer for which,

$$\sum_{n=1}^{\infty} \frac{1}{r_n^{p+1}} = \sum_{n=1}^{\infty} \frac{1}{n^\alpha}$$

is convergent since $\alpha > 1$. Hence, the given infinite product is a canonical product. So, the order of $\prod_{n=1}^{\infty}\left(1 - \frac{z}{n^\alpha}\right)$ is the same as the convergence exponent of its zeros. Here, zeros are $a_n = n^\alpha$. Hence, $r_n = |a_n| = n^\alpha$. Hence the convergence exponent

$$\rho_1 = \limsup_{n\to\infty} \frac{\log n}{\log r_n} = \limsup_{n\to\infty} \frac{\log n}{\alpha \log n} = \frac{1}{\alpha}.$$

Exercise 8.3.1. 1. Prove the following.

(a) $\sin \pi z = \pi z \prod_{n=1}^{\infty}\left(1 - \frac{z^2}{n^2}\right)$

(b) $\cos z = \prod_{n=1}^{\infty}\left(1 - \frac{4z^2}{\pi^2(2n-1)^2}\right)$

2. Prove Fundamental theorem of algebra using Picard's little theorem.

3. With proper justifications, write which of the following functions is constant.

   (a) An entire function that omit two values on the complex plane;

   (b) An entire function that omit values on the negative real axis;

   (c) An entire function whose range set is a dense set;

   (d) An entire function whose range set is not a dense set;

   (e) An entire function whose range set is a straight line;

   (f) An entire function whose range set is $B(0; R)$, $R > 0$ is a real number;

   (g) An entire function for which $M(r_1) = M(r_2)$ for $r_1 < r_2$;

(h) An entire function whose range set is a closed set;

(i) An entire function that takes the value $a \in \mathbb{C}$ at all points in the set E that has a limit point in $\mathbb{C}$.

8.4 Few Probable Questions

1. Define canonical product of the zeros of an entire function. Find the canonical product of the function $\sin z$.
2. State and prove Borel's theorem.
3. State and prove Hadamard's Factorization theorem.
4. Show that for an entire function f of finite order ρ, which is not an integer, and convergence exponent ρ_1, $\rho = \rho_1$.
5. State the Hadamard's factorization theorem. If f and g be entire functions of order ρ and ρ' respectively, such that $\rho' \leq \rho$, and also if the zeros of g are all zeros of f, the show that the function $H(z) = \dfrac{f(z)}{g(z)}$ is an entire function of order at most ρ.
6. State and prove the Picard's little theorem.
7. Show that an entire function f of finite order ρ, which is not an integer, has infinitely many zeros.
8. State Picard's Little theorem. Hence prove the Fundamental theorem of Algebra.

Unit 9

Course Structure

- Multiple-valued functions,
- Riemann surface for the functions $\sqrt{z}$, $\log z$.

9.1 Introduction

Multifunctions are hard to avoid. Many complex functions, like the complex exponential, are not globally one-to-one. We may view such a function as having an inverse, so long as we allow the inverse to be a multifunction. Constructing, at least locally, a well-behaved functional inverse will involve extracting a suitable value from this multifunction at each point of the domain. But in order to treat more complicated examples, such as $\sqrt{z}$, $z^{2/3}$, etc., we begin a deeper analysis of many-valuedness. which will enable us to handle logarithms and powers of rational functions.

Objectives

After reading this unit, you will be able to

- define multifunctions and visualize certain examples related to them
- discuss in detail the argument function and its rle in the many-valuedness of complex functions
- define branches and branch cuts of multifunctions
- discuss the Riemann surfaces of the square root function and logarithm function

9.2 Multiple-Valued Functions

Thus far, we have considered a complex function f to be a rule that assigns to each point z, a single complex number $f(z)$. This familiar conception of a function is unduly restricted. Using examples, we now discuss how we may broaden the definition of a function to allow $f(z)$ to have many different values for a single value of z. In such a case, f is called a many-valued function or a multifunction.

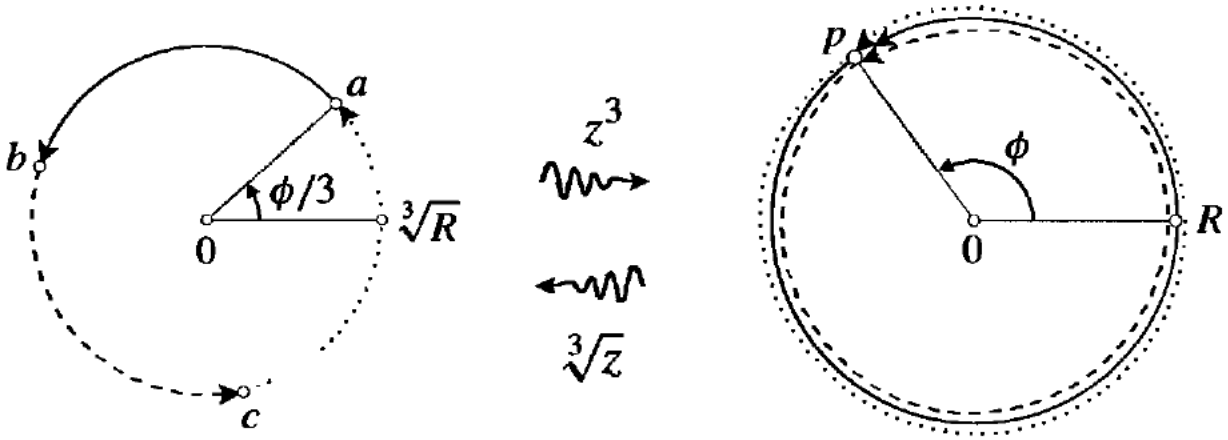


Figure 9.1

For example, if we consider the function $\sqrt[3]{z}$, then it has three different values (if z is non-zero) for a single value of z and hence is a three-valued multifunction.

Let us see how. The function $z \mapsto \sqrt[3]{z}$, that is, if $w = \sqrt[3]{z}$, and a is a solution of this equation, let us find the other two solutions too. If $z = r\,\mathrm{e}^{i\theta}$ orbits round an origin-centred circles, $z^3 = r^3\,\mathrm{e}^{3i\theta}$ orbits three times faster executing a complete revolution each time z executes one third of a revolution. Put differently, reversing the direction of the mapping divides the angular speed by three. This is an essential ingredient, which we will now study in detail.

Writing $z = r\,\mathrm{e}^{i\theta}$, we have $\sqrt[3]{z} = \sqrt[3]{r}\,\mathrm{e}^{i(\theta/3)}$. Here, $\sqrt[3]{r}$ uniquely defined as the real cube root of the length of z; the sole source of the three-fold ambiguity in the formula is the fact that there are infinitely many different choices for the angle θ of a given point z.

Think of z as a moving point that is initially at $z = p$. If we arbitrarily choose θ to be the angle ϕ as shown in fig. 9.1, then $\sqrt[3]{p} = a$. As z gradually moves away from p, θ gradually changes from its initial value ϕ, and $\sqrt[3]{z} = \sqrt[3]{r}\,\mathrm{e}^{i(\theta/3)}$ gradually moves away from its initial position a, but in a completely determined way-its distance from the origin is the cubic root of the distance of z, and its speed of movement is one-third that of z.

9.3 Argument as a function

The angle θ in the expression $z = |z|\,\mathrm{e}^{i\theta}$ is not uniquely determined. Indeed, this is the fundamental cause of many-valuedness in complex function theory. For $z \neq 0$, we define the argument of z to be

$$[\arg z] = \{\theta \in \mathbb{R} :\ z = |z|\,\mathrm{e}^{i\theta}\}.$$

The bracket notation $[\arg z]$ is designed to emphasize that the argument of z is a set of numbers, not a single number. In fact, $[\arg z]$ is an infinite set, consisting of all numbers from $\theta + 2k\pi$ for $k \in \mathbb{Z}$, where θ is any fixed real number such that $\mathrm{e}^{i\theta} = z/|z|$.

The restriction $-\pi < \theta \leq \pi$, or alternatively, $0 \leq \theta < 2\pi$, uniquely determines θ in the equation $0 \neq z = |z|\,\mathrm{e}^{i\theta}$.

Now, consider what happens to a principal value determination of argument $\operatorname{Arg} z = \theta$, where $z = |z|\,\mathrm{e}^{i\theta}$, $0 \leq \theta < 2\pi$, where z performs a complete anticlockwise circuit round the unit circle starting from $z = 1$, with $\theta \in [0, 2\pi)$. Within the chosen range $[0, 2\pi)$, θ has value 0 at the start and increases steadily towards 2π as z moves round the circle until it arrives back at 1, when θ must jump back to 0. Thus, $\operatorname{Arg} z$ has a jump discontinuity. On the other hand, if we insist on choosing θ so that it varies continuously with z, then its final

value has to be 2π, a different choice from $[\arg 1]$ from that we made at the start. We can give a more formal treatment of the issues just discussed.

We show that there is no way to impose a restriction which selects $\theta(z) \in [\arg z]$ for all $z \in \mathbb{C} \setminus \{0\}$, so $\theta : z \mapsto \theta(z)$ is continuous as a function of z. We assume for a contradiction that such a continuous function θ does not exist and consider

$$k(t) = \frac{1}{2\pi}\left(\theta(\mathrm{e}^{it}) + \theta(\mathrm{e}^{-it})\right), \quad t \in \mathbb{R}.$$

Then k is continuous and

$$k(t) = \frac{1}{2\pi}\left((t + 2m_t\pi) + (-t + 2n_t\pi)\right), \quad m_t, n_t \in \mathbb{Z},$$

so k takes only integer values. Also $k(0)$ is even and $k(\pi)$ is odd, so k is non-constant. This contradicts the intermediate value theorem from real analysis.

This result has implications for other multifunctions. For example, it tells us that there cannot be a continuous logarithm in $\mathbb{C} \setminus \{0\}$: if there were one, then its imaginary part - an argument function - would be continuous too.

9.4 Branch Points

Take a multifunction $f(z)$, so that $w(z)$ is a non-empty subset of $\mathbb{C}$ for each z in the domain of definition of f. Assume that the many-valuedness arises because, for one or more points a, the definition of $f(z)$ explicitly or implicitly involves the angle θ, where $z - a = |z - a|\,\mathrm{e}^{i\theta}$. Such points are called branch points. Any branch point is excluded from the domain of definition of $f(z)$. More formally, a is a branch point for $f(z)$, if for all sufficiently small $r > 0$, it is not possible to choose a particular value of $w = f(z)$ such that f is continuous in the open ball $B(a; r)$. The motivation comes from the previous section, no continuous argument function can be drawn from $[\arg(z - a)]$ for z on a circle with centre at a.

We illustrate this with the example of $\sqrt[3]{z}$ (see fig. 9.2). Usually, we draw mappings from left to right, but here we have reversed this convention to to facilitate comparison with 9.1. As z travels along the closed

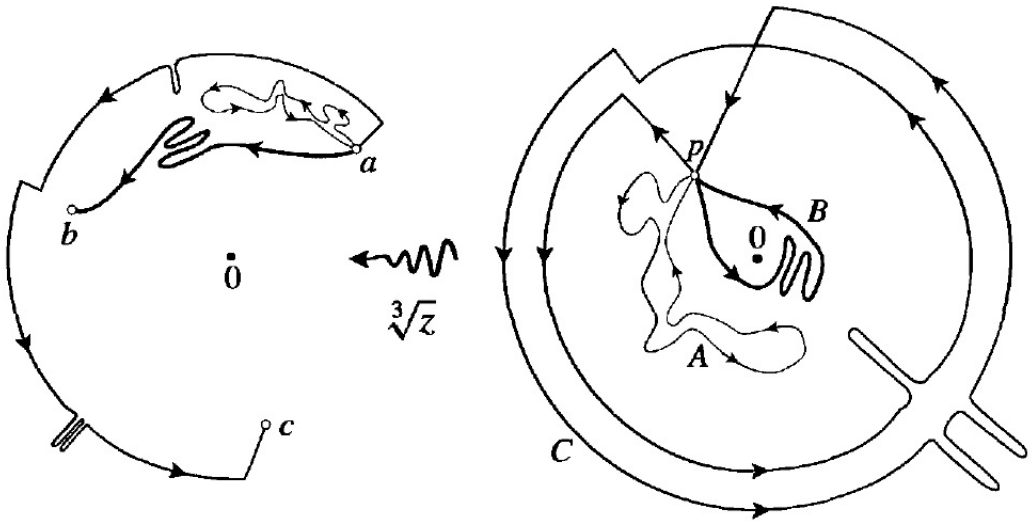


Figure 9.2

loop A (finally returning to p), $\sqrt[3]{z}$ travels along the illustrated closed loop and returns to its original value a.

However, if z instead travels along the closed loop B, which goes round the origin once, then $\sqrt[3]{z}$ does not return to its original value, but instead it ends up at a different cube root of p, namely b. Note that the detailed shape of B is irrelevant, all that matters is that it encircles the origin once. Similarly, if z travels along C, encircling the origin twice, then $\sqrt[3]{z}$ ends up at c, the third and final cube root of p. Clearly, if z ere to travel along the loop (not shown) that encircled the origin three times, then $\sqrt[3]{z}$ would return to the original value a.

The premise for this picture of $z \mapsto \sqrt[3]{z}$ was the arbitrary choice of $\sqrt[3]{p} = a$, rather than b or c. If we instead chose $\sqrt[3]{p} = b$, then the orbits on the left of fig. 9.2 would simply be rotated by $2\pi/3$. Similarly, if we chose $\sqrt[3]{p} = c$, then the orbits would be rotated by $4\pi/3$.

The point $z = 0$ is the branch point of $\sqrt[3]{z}$. More generally, let $f(z)$ be a multifunction and let $a = f(p)$ be one of its values at some point $z = p$. Arbitrarily choosing the initial position of $f(z)$ to be a, we may follow the movement of $f(z)$ as z travels along a closed loop beginning and ending at p. When z returns to p, $f(z)$ will either return to a or it will not. A branch point $z = q$ of f is a point such that $f(z)$ fails to return to a as z travels along any loop that encircles q once.

Returning to the specific example $f(z) = \sqrt[3]{z}$, we have seen that if z executes three revolutions round the branch point at $z = 0$ then $f(z)$ returns to its original value. If $f(z)$ were an ordinary, single-valued function then it would return to its original value after only one revolution. Thus, relative to an ordinary function, $two\ extra$ revolutions are needed to restore the original value of $f(z)$. We summarize this by saying that 0 is a branch point of $\sqrt[3]{z}$ of $order\ two$.

Definition 9.4.1. If q is a branch point of some multifunction $f(z)$, and $f(z)$ first returns to its original value after N revolutions round q, then q is called an $algebraic\ branch\ point$ of order $(N-1)$; an algebraic branch point of order 1 is called a $simple\ branch\ point$. We should stress that it is perfectly possible that $f(z)$ never returns to its original value, no matter how many times z travels round q. In this case q is the $logarithmic\ branch\ point$-the name will be explained in the next section.

By extending the above discussion of $\sqrt[3]{z}$, check for yourself that if n is an integer, then $z^{1/n}$ is an n-valued multifunction whose only (finite) branch point is at $z = 0$, the order of this branch point being $(n-1)$. More generally, the same is true for any fractional power $z^{m/n}$, where m/n is a fraction reduced to lowest terms.

9.4.1 Multibranches

Suppose we are given a multifunction $f(z)$. Our goal is to select a value $w = f(z)$ from various possible values, for each z in as large a domain as possible, so that f is holomorphic. In particular. f has to be continuous. We now introduce multibranches. These provide a stepping stone on the way to our goal.

There is a sense in which we can make continuous selections from multifunctions in a natural way. The key idea is the following. Rather than considering z as our variable we introduce, for each branch point a, new variables (r, θ), where $z = a + r\,\mathrm{e}^{i\theta}$. Let us illustrate this with the example of $\sqrt[3]{z}$.

By arbitrarily picking one of the three values of $\sqrt[3]{p}$ at $z = p$, and then allowing z to move, we see that we obtain a unique value of $\sqrt[3]{Z}$ associated with any particular path from p to Z. However, we are still dealing with multifunction: by going round the branch point 0, we can end up at any one of the three possible values of $\sqrt[3]{Z}$.

On the other hand, the value of $\sqrt[3]{Z}$ does not depend on the detailed shape of the path: *if we continuously deform the path without crossing the branch point, then we obtain the same value of* $\sqrt[3]{Z}$. This shows us how we may obtain a single-valued function. If we restrict z to a simply connected set S that contains p, but does not contain the branch point, then every path in S from p to Z will yield the same value of $\sqrt[3]{Z}$, which we will call $f_1(Z)$. Since the path is irrelevant, f_1 is an ordinary, single-valued function of position in S, which is a branch of the $\sqrt[3]{z}$.

Fig. 9.3 illustrates such a set S, together with its image under the branch f_1 of $\sqrt[3]{z}$. Here, we have reverted to our normal practice of depicting the mapping going from left to right. If we instead choose $\sqrt[3]{p} = b$, then

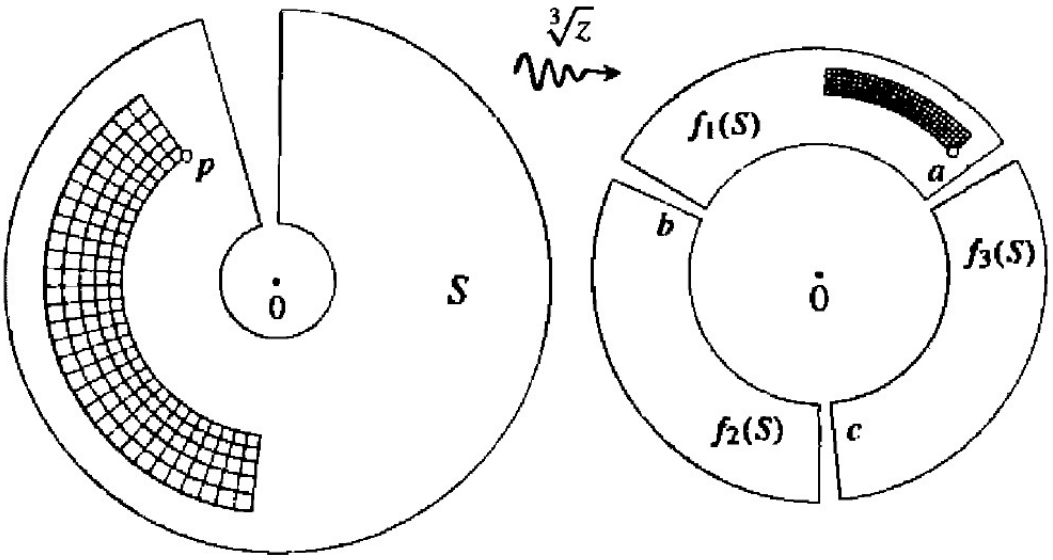


Figure 9.3

we obtain a second branch f_2 of $\sqrt[3]{z}$, while $\sqrt[3]{p} = c$ yields a third and final branch f_3.

We can simplify it as considering three different single-valued functions

$$f_1(r,\theta) = \sqrt[3]{r}\,\mathrm{e}^{i\theta/3}, \quad f_2(r,\theta) = \sqrt[3]{r}\,\mathrm{e}^{i(\theta+2\pi)/3}, \quad f_3(r,\theta) = \sqrt[3]{r}\,\mathrm{e}^{i(\theta+4\pi)/3}$$

such that for $z = r\,\mathrm{e}^{i\theta}$,

$$\begin{aligned} \sqrt[3]{z} &= f_1(r,\theta) \quad \text{if } \theta \in [0, 2\pi) \\ &= f_2(r,\theta) \quad \text{if } \theta \in [2\pi, 4\pi) \\ &= f_3(r,\theta) \quad \text{if } \theta \in [4\pi, 6\pi). \end{aligned}$$

And, as z circles the origin and θ changes from 4π through 6π, $\sqrt[3]{z}$ jumps back to the previous branch $f_1(r,\theta)$.

9.4.2 Branch Cuts

How can we prevent the interchange of f_3 to f_1? f_3 changes to f_1 due to the change of argument θ. So, if we restrict the movement of θ, then we can remain in a particular branch of $\sqrt[3]{z}$, which is single-valued and holomorphic. We draw an arbitrary curve C from the branch point 0 to infinity. We are now restricting the domain of z which includes all points of S excluding those on C. This prevents the closed path in S from encircling the branch point; or simplifying, we can say that we restrict the increase of θ by a value of 2π since changing the value of θ by 2π changes the branch of $\sqrt[3]{z}$.

This C is called a *branch cut*. As we have just seen, this shows up in the fact that the resulting branches are discontinuous on C, despite the fact that the three values of $\sqrt[3]{z}$ move continuously as z move continuously. As z crosses C travelling counterclockwise, then we must switch from one branch to the next in order to maintain continuous motion of $\sqrt[3]{z}$. If z executes three counterclockwise revolutions round the branch point, then the branches permute cyclically as shown in fig. 9.4.

A common choice for C is the negative real axis. If we do not allow z to cross the cut, then we restrict the angle θ to lie in the range $-\pi < \theta \leq \pi$. This is called the *principal value of argument*, written as Arg z, as we have encountered in the beginning of this unit. With this choice of θ, the single-valued function $\sqrt[3]{r}\,\mathrm{e}^{i\theta/3}$ is

$$\begin{Bmatrix} f_1 \\ f_2 \\ f_3 \end{Bmatrix} \to \begin{Bmatrix} f_2 \\ f_3 \\ f_1 \end{Bmatrix} \to \begin{Bmatrix} f_3 \\ f_1 \\ f_2 \end{Bmatrix} \to \begin{Bmatrix} f_1 \\ f_2 \\ f_3 \end{Bmatrix}$$

Figure 9.4

called the *principal branch* of the cube root. Let us denote this as $[\sqrt[3]{z}]$. Note that the principal branch agrees with the real cube root function on the positive real axis, but not the negative real axis and note that the other two branches can be expressed in terms of the principal branch as $\mathrm{e}^{i(2\pi/3)}\,[\sqrt[3]{z}]$ and $\mathrm{e}^{i(4\pi/3)}\,[\sqrt[3]{z}]$.

9.5 Riemann Surfaces

Riemann surfaces were first studied by Bernhard Riemann. Riemann surfaces are the easiest way to geometrically understand the multifunctions. The main interest in Riemann surfaces is that holomorphic functions may be defined between them. We discuss the Riemann surfaces of two major functions: the complex square root function and the logarithm function.

9.5.1 Square Root function

Just as we have discussed the case of $\sqrt[3]{z}$, we can simply show that $w = u+iv = f(z) = \sqrt{z}$ is a multifunction having a branch at 0. We define two branches

$$f_1(z) = \sqrt{r}\,\mathrm{e}^{i\theta/2}, \qquad f_2(z) = \sqrt{r}\,\mathrm{e}^{i(\theta+2\pi)/2} = -\,\mathrm{e}^{i\theta/2} = -f_1(z);$$

so f_1 and f_2 can be thought of as "plus" and "minus" square root functions. The negative real axis is called a branch cut for the functions f_1 and f_2. Each point on the branch cut is a point of discontinuity for both functions f_1 and f_2.

Example 9.5.1. We show that f_1 is discontinuous along the negative real axis. Let $z_0 = r\,\mathrm{e}^{i\pi}$ denote a negative real number. We compute the limit as z approaches z_0 through the upper half plane and the limit as z approaches z_0 through the lower half plane. In polar coordinates, these are given by

$$\lim_{(r,\theta)\to(r_0,\pi)} f_1(r\,\mathrm{e}^{i\theta}) = \lim_{(r,\theta)\to(r_0,\pi)} \sqrt{r_0}\left(\cos\frac{\theta}{2} + i\sin\frac{\theta}{2}\right) = i\sqrt{r_0}, \quad \text{and}$$

$$\lim_{(r,\theta)\to(r_0,-\pi)} f_1(r\,\mathrm{e}^{i\theta}) = \lim_{(r,\theta)\to(r_0,-\pi)} \sqrt{r_0}\left(\cos\frac{\theta}{2} + i\sin\frac{\theta}{2}\right) = -i\sqrt{r_0}.$$

The two limits are distinct, so f_1 is discontinuous at z_0. Since z_0 is arbitrary, so f_1 is discontinuous on the whole negative real axis.

We will now draw the Riemann surface for f. $f(z)$ has two values for any $z \neq 0$. Each functions f_1 and f_2 are single-valued on the domain formed by cutting the z plane along the negative real axis. Let D_1 (see fig. 9.5) and D_2 (see fig. 9.6) be the domains of f_1 and f_2 respectively. The range set for f_1 is H_1 consisting of the right-half plane and the positive v-axis; and the range set for f_2 is H_2 consisting of the left-half plane and the negative v-axis. The sets H_1 and H_2 are "glued together" along the positive v-axis and the negative v-axis to form the w plane with the origin deleted.

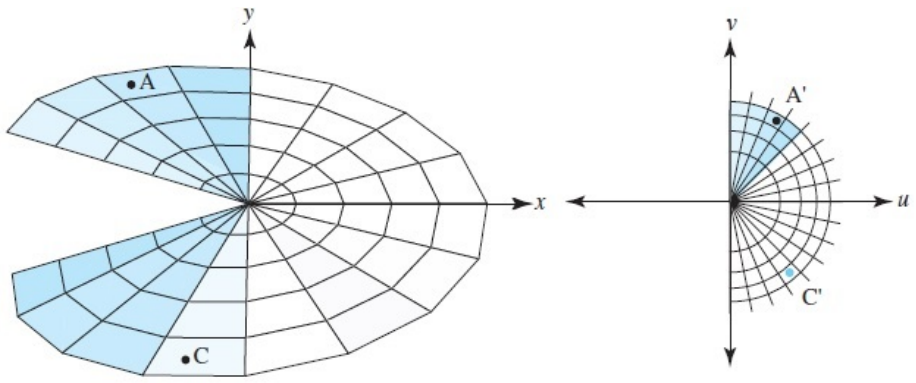


Figure 9.5: A portion of D_1 and its image under $w = \sqrt{z}$

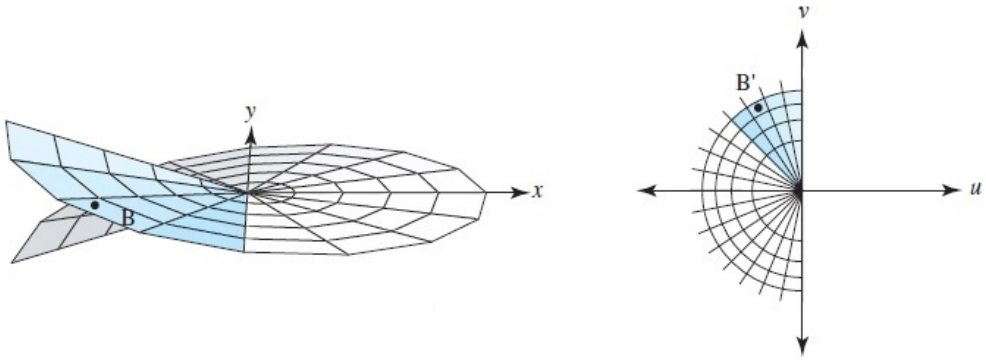


Figure 9.6: A portion of D_2 and its image under $w = \sqrt{z}$

We stack D_1 directly above D_2. The edge of D_1 in the upper half-plane is joined to the edge of D_2 in the lower half-plane, and the edge of D_1 in the lower half-plane is joined to the edge of D_2 in the upper half-plane. When these domains are glued together in this manner, they form R, which is a Riemann surface domain for the mapping $w = f(z) = \sqrt{z}$. The portions of $D_1,\ D_2$ and R that lie in $\{z : \ |z| < 1\}$ are shown in fig. 9.7.

The beauty of this structure is that it makes this "full square root function" continuous for all $z \neq 0$. Normally, the principal square root function would be discontinuous along the negative real axis, as points near -1 but above that axis would get mapped to points close to i, and points near -1 but below the axis would get mapped to points close to $-i$. As fig. 9.7 indicates, however, between the point A and the point B, the domain switches from the edge of D_1 in the upper half-plane to the edge of D_2 in the lower half plane. The corresponding mapped points A' and B' are exactly where they should be. The surface works in such a way that going directly between the edges of D_1 in the upper and lower half planes is impossible (likewise for D_2). Going counterclockwise, the only way to get from the point A to the point C, for example, is to follow the path indicated by the arrows in fig. 9.7.

We now move on to the logarithmic function.

Exercise 9.5.1. Show that f_2 is discontinuous at every point on the negative real axis.

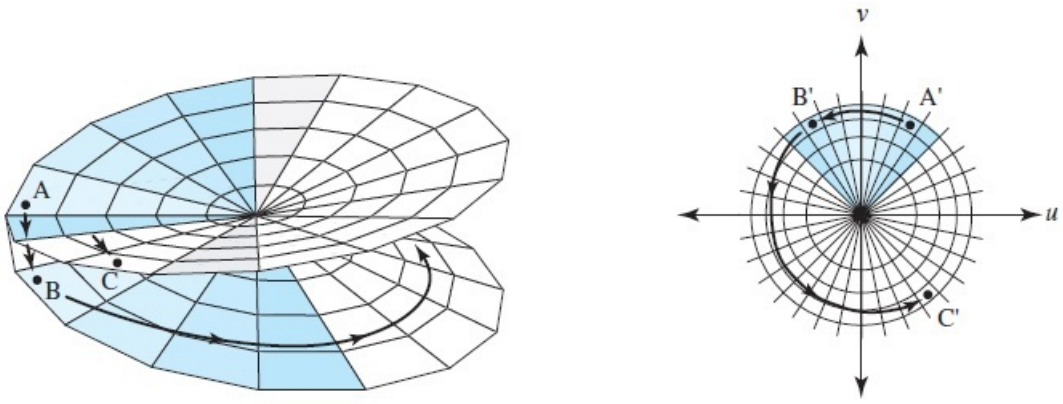


Figure 9.7: A portion of R and its image under $w = \sqrt{z}$

9.5.2 Logarithm Function

The complex logarithm function $\log(z)$ may be introduced as the "inverse" of e^z. More precisely, we define $\log z$ to be any complex number z that satisfies

$$e^{\log z} = z.$$

It follows that

$$\log z = \ln|z| + i \arg(z).$$

Since $\arg(z)$ takes infinitely many values, differing from each other by multiples of 2π, we see that $\log(z)$ is a multifunction taking infinitely many values, differing from each other by multiples of $2\pi i$. For example,

$$\log(2+2i) = \ln 2\sqrt{2} + i\frac{\pi}{4} + 2n\pi i,$$

where n is an arbitrary integer. The reason we get infinitely many values is clear if we see the behaviour of the exponential function e^z. Each time z travels straight upwards by $2\pi i$, e^z executes a complete revolution and returns to its original value. Clearly, $\log(z)$ has a branch point at 0. However, this branch point is quite unlike that of $\sqrt[n]{z}$, for no matter how many times we loop around the origin, $\log(z)$ never returns to its original value, rather it continues moving upwards forever. You can now understand previously introduced term "logarithmic branch point".

Here is another difference between the branch point of $\sqrt[n]{z}$ and $\log(z)$. As z approaches the origin, say along a ray, $|\sqrt[n]{z}|$ tends to zero, but $|\log(z)|$ tends to infinity, and in this sense, origin is a singularity as well as a branch point.

To define single-valuedness of $\log(z)$, we make a branch cut from 0 out to infinity. The most common choice for this cut is the negative real axis. In this cut plane, we may restrict $\arg(z)$ to its principal value Arg z. This yields the principal branch of logarithm, written as Log z, defined as

$$\text{Log } z = \ln|z| + \text{Arg } z.$$

We are in a position to draw the Riemann surface for $\log(z)$. Let define the multibranches of logarithm function as follows

$$F_k(z) = \log|z| + i(\theta + 2k\pi), \quad k \in \mathbb{Z}.$$

Each F_k is a continuous function of z. Furthermore, for any fixed $c \in \mathbb{R}$, and for $0 \neq z = r\,e^{i\theta}$,

$$\log(z) = \{F_k(z) : \quad k \in \mathbb{Z}, \quad \theta \in [c, c+2\pi)\},$$

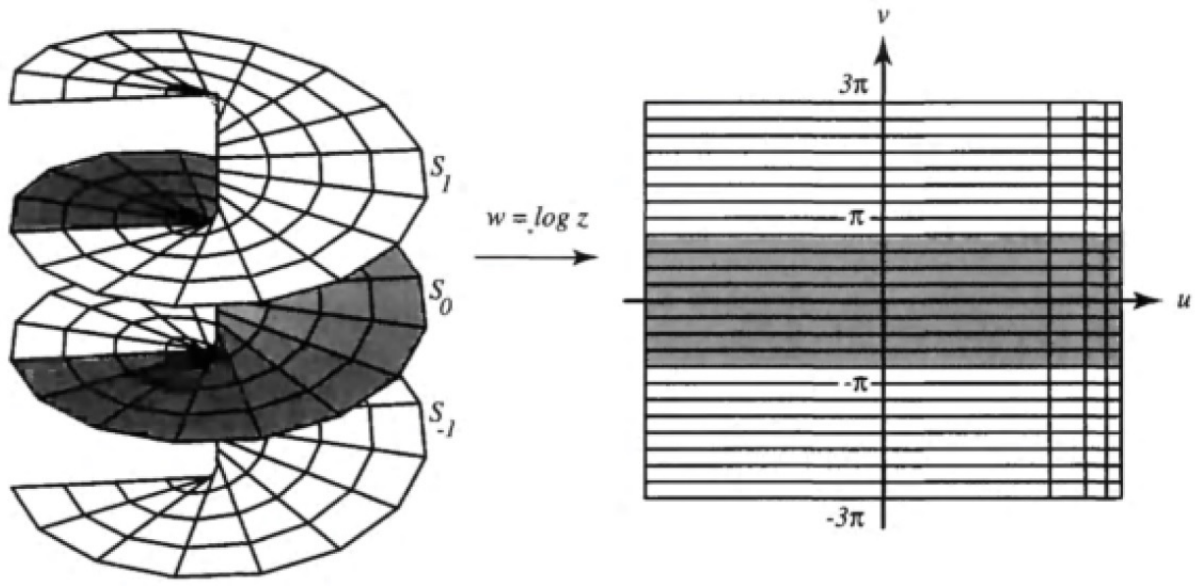


Figure 9.8

with no values repeated, and if similarly, θ is restricted to any interval $(c, c + 2\pi]$.

The domain of definition of each F_k are none other than the whole complex plane with different arguments. Let S_k be the domain of definition of the branch F_k of $\log(z)$. (These copies of the complex plane are each cut along the negative real axis) These cut planes are then stacked directly on top of each other and joined as follows. For each integer k, the edge of S_k in the upper half plane is joined to the edge of S_{k+1} in the lower half plane. The resulting Riemann surface of $\log(z)$ looks like a spiral staircase that extends upwards on $S_1, S_2, \ldots$ and downwards on $S_{-1}, S_{-2}, \ldots$ as shown in fig. 9.8. If we start on S_0 and make a counterclockwise circuit around the origin, we end up on S_1, and the next circuit brings us to S_2, etc, so each time we cross the negative real axis, we end up on a new branch of $\log(z)$.

9.6 Few Probable Questions

1. Which of the following are multifunctions?

   A. $\cos z$ B. $\sqrt{z}$ C. e^z D. None

2. The branch point of the function $\log(z)$ is at

   A. 0 B. 1 C. ∞ D. -1

3. The function $\sqrt[3]{z^2}$ has an algebraic branch point of order ..... at $z = 0$.

   A. $3/2$ B. $2/3$ C. 2 D. 3

4. For the function $\sqrt{z-1}$, the point $z = 1$ is a/an ..... branch point.

A. transcendental B. logarithmic C. algebraic D. not a branch point

5. Define branch of a multifunction. Show that the branches of the square root function are discontinuous at each point of the negative real axis.

Unit 10

Course Structure

- Analytic continuation, uniqueness,
- Continuation by the method of power series

10.1 Introduction

Analytic continuation is an important idea since it provides a method for making the domain of definition of an analytic function as large as possible. Usually, analytic functions are defined by means of some mathematical expressions such as polynomials, infinite series, integrals, etc. The domain of definition of such an analytic function is often restricted by the manner of defining the function. For instance the power series representation of such analytic functions does not provide any direct information as to whether we could have a function analytic in a domain larger than disc of convergence which coincides with the given function. We have previously seen that an analytic function is determined by its behaviour at a sequence of points having limit point. This was precisely the content of the identity theorem which is also referred to as the principle of analytic continuation. For example, as a consequence, there is precisely a unique entire function on $\mathbb{C}$ which agrees with $\sin x$ on the real axis, namely $\sin z$.

Objectives

After reading this unit, you will be able to

- define analytic continuation of an analytic function and consider examples of such process
- show that the analytic continuation of an analytic function is always unique
- define chain and function elements and discuss the condition for analytic continuation from a domain into another
- discuss the power series method and its examples

10.2 Analytic Continuation

As we have seen in the introduction that if G is a region and f is an analytic function on G. Also, let g be an analytic function defined on an open set $D \subseteq G$, such that $f(z) = g(z)$ in D. Then, by uniqueness theorem (identity theorem), we have $f \equiv g$ on G. It is a simple process to extend the domain of g over to G. A natural question is the following: Is it always possible to have such an extension? Clearly that is not the case. For example,

$$f(z) = \frac{1}{z}, \;\; z \in \mathbb{C} \setminus \{0\}$$

does not have an extension to $\mathbb{C}$. Similarly, if we take two domains

$$D_1 = \mathbb{C} \setminus \{z : \text{ Re } z \leq 0, \text{ Im } z = 0\}, \quad \text{and} \quad D_2 = \mathbb{C},$$

then for $f(z) = \text{Log } (z)$, which is the principal branch of logarithm function that is analytic on the domain D_1, but can't be extended on to D_2.

However, if the extension is possible, there are ways to carry out the process of continuation so that the given analytic function becomes analytic on a larger domain. To make this point more precise, let us start by examining the analytic continuation of the function

$$f(z) = \sum_{n \geq 0} z^n. \tag{10.2.1}$$

The series on the right hand side of (10.2.1), as is well known, is convergent for $|z| < 1$ and diverges for $|z| \geq 1$. On the other hand, we know that the series given by the formula (10.2.1) represents an analytic function for $|z| < 1$ and the sum of the series (10.2.1) for $|z| < 1$ is $1/(1 - z)$. However, the function F defined by the formula

$$F(z) = \frac{1}{1 - z}$$

is analytic for $z \in \mathbb{C}_\infty \setminus \{1\} = D$, since

$$F\left(\frac{1}{z}\right) = \frac{1}{1 - z^{-1}} = \frac{z}{z - 1}$$

is analytic at $z = \infty$. Now, $f(z) = F(z)$ for all $z \in \mathbb{D} \cap D$, and we call F an analytic continuation of f from $\mathbb{D}$ into D, that is, the function f, given at first for $|z| < 1$, has been extended to the extended complex plane but for the point 1, at which the function has a simple pole. Thus, it seems that F, which is analytic globally, is represented by a power series only locally.

We now, formally define analytic continuation of a function f as follows.

Definition 10.2.1. Suppose that f and F are two functions such that

1. f is analytic on some domain $D \subset \mathbb{C}$;
2. F is analytic in a domain D_1 such that $D_1 \cap D \neq \emptyset$ and $D \subset D_1$, such that $f(z) = F(z)$ for $z \in D \cap D_1$.

Then we call F an analytic continuation or holomorphic extension of f from domain D into D_1. In other words, f is said to be analytically continuable into D_1.

The definition can also be given as follows.

Definition 10.2.2. A function $f(z)$, together with a domain D in which it is analytic, is said to be a **function element** and is denoted by (f, D). Two function elements (f_1, D_1) and (f_2, D_2) are called **direct analytic continuations** of each other if and only if

$$D_1 \cap D_2 \neq \emptyset \quad \text{and} \quad f_1 = f_2 \;\text{ on }\; D_1 \cap D_2.$$

Remark 10.2.1. Whenever there exists a direct analytic continuation of (f_1, D_1) into a domain D_2, it must be uniquely determined, for any two direct analytic continuations would have to agree on $D_1 \cap D_2$, and by identity theorem, would consequently have to agree throughout D_2. That is, given an analytic function f_1 on D_1, there is at most one way to extend f_1 from D_1 into D_2 so that the extended function is analytic in D_2.

The property of being a direct analytic continuation is not transitive. That is, even if (f_1, D_1) and (f_2, D_2) are direct analytic continuations of each other, and (f_2, D_2) and (f_3, D_3) are direct analytic continuations of each other, we cannot conclude that (f_1, D_1) and (f_3, D_3) are direct analytic continuations of each other. A simple example of this occurs whenever D_1 and D_3 have no points in common. However, there is a relationship between $f_1(z)$ and $f_3(z)$ that is worth explaining.

Definition 10.2.3. Suppose that $\{(f_1, D_1), (f_2, D_2), \ldots, (f_n, D_n)\}$ is a finite set of function elements with the property that (f_k, D_k) and (f_{k+1}, D_{k+1}) are direct analytic continuations of each other for $k = 1, 2, \ldots, n-1$. Then the set of function elements are said to be analytic continuations of one another. Such a set of function elements is then called a **chain**.

Example 10.2.1. Consider the figure 10.1. Let

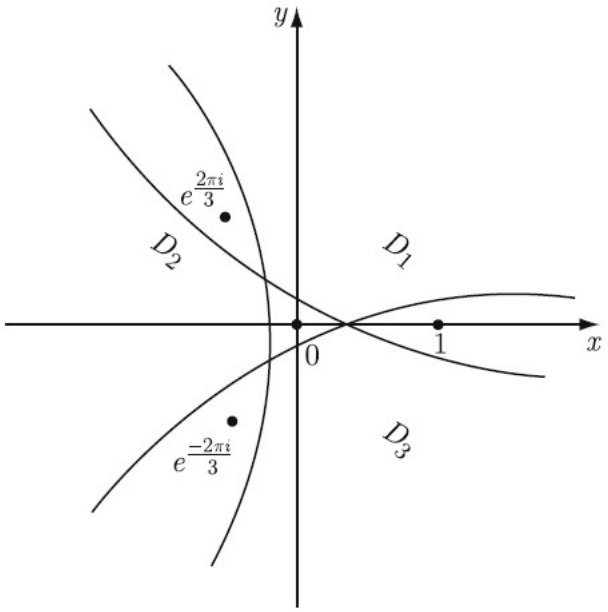


Figure 10.1

$$\begin{aligned} f_1(z) &= \text{Log } z, \ \ z \in D_1 \\ f_2(z) &= \text{Log } z, \ \ z \in D_2 \\ f_3(z) &= \text{Log } z + 2\pi i, \ \ z \in D_3. \end{aligned}$$

Then $\{(f_1, D_1), (f_2, D_2), (f_3, D_3)\}$ is a chain with $n = 3$. Note that $0 = f_1(1) \neq f_3(1) = 2\pi i$.

Note that (f_i, D_i) and (f_j, D_j) are analytic continuations of each other if and only if they can be connected by finitely many direct analytic continuations.

10.3 Analytic Continuation along a curve

Definition 10.3.1. If $\gamma : [0,1] \to \mathbb{C}$ is a curve and if there exists a chain $\{(f_i, D_i)\}_i$ of function elements such that

$$\gamma([0,1]) \subset \cup_{i=1}^{n} D_i, \;\; z_0 = \gamma(0) \in D_1, \;\; z_n = \gamma(1) \in D_n,$$

then we say that the function element (f_n, D_n) is an analytic continuation of (f_1, D_1) along the curve γ. That is a function element (f, D) can be analytically continued along a curve if there is a chain containing (f, D) such that each point on the curve is contained in the domain of some function element of the chain.

As another example, the domains of a chain are also shown in Figure 10.2.

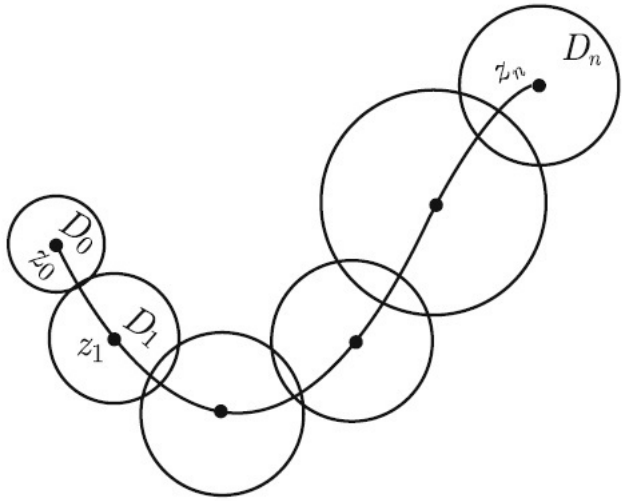


Figure 10.2

The definition can also be given as

Definition 10.3.2. Let $\gamma : [a,b] \to \mathbb{C}$ be a curve such that

$$\gamma(t) = z(t) = x(t) + iy(t), \;\; a \le t \le b.$$

Let us consider a partition $a = t_0 < t_1 < \ldots < t_n = b$ of $[a,b]$. If there is a chain $\{(f_1, D_1), (f_2, D_2), \ldots, (f_n, D_n)\}$ of function elements such that (f_{k+1}, D_{k+1}) is a direct analytic continuation of (f_k, D_k) for $k = 1, 2, \ldots, n-1$ and $z(t) \in D_k$ for $t_{k-1} \le t \le t_k$, $k = 1, 2, \ldots, n$, then (f_n, D_n) is said to be an analytic continuation of (f_1, D_1) along the curve γ.

Thus we shall obtain a well defined analytic function in a nbd. of the end point of the path, which is called the analytic continuation of (f_1, D_1) along the path γ. Here, D_k may be taken as discs containing $z(t_{k-1})$. Further, we say that the sequence $\{D_1, D_2, \ldots, D_n\}$ connected by the curve γ along the partition if the image $z([t_{k-1}, t_k])$ is contained in D_k.

Theorem 10.3.1. (Uniqueness of Analytic Continuation along a Curve) Analytic continuation of a given function element along a given curve is unique. In other words, if (f_n, D_n) and (g_m, D_m) are two analytic continuations of (f_1, D_1) along the curve γ defined by

$$\gamma(t) = z(t) = x(t) + iy(t), \;\; a \le t \le b.$$

Then $f_n = g_m$ on $D_n \cap E_m$.

Proof. Suppose there are two analytic continuations of (f_1, D_1) along the curve γ, namely,

$$(f_1, D_1), (f_2, D_2), \ldots, (f_n, D_n)$$
$$(g_1, E_1), (g_2, E_2), \ldots, (g_m, E_m)$$

where $f_1 = g_1$ and $E_1 = D_1$. Then there exist partitions

$$a = t_0 < t_1 < \ldots < t_n = b$$
$$a = s_0 < s_1 < \ldots < s_m = b$$

such that $z(t) \in D_i$ for $t_{i-1} \leq t \leq t_i$, for $i = 1, 2, \ldots, n$ and $z(t) \in E_j$ for $s_{j-1} \leq t \leq s_j$ for $j = 1, 2, \ldots, m$. We claim that if $1 \leq i \leq n$, $1 \leq j \leq m$ and

$$[t_{i-1}, t_i] \cap [s_{j-1}, s_j] \neq \emptyset$$

then (f_i, D_i) and (g_j, E_j) are direct analytic continuations of each other. This is certainly true when $i = j = 1$, since $f_1 = g_1$ and $E_1 = D_1$. If it is not true for all i and j, then we may pick from all (i, j), for which the statement is false and such that $i + j$ is minimal. Suppose that $t_{i-1} \geq s_{j-1}$, where $i \geq 2$. Since $[t_{i-1}, t_i] \cap [s_{j-1}, s_j] \neq \emptyset$ and $s_{j-1} \leq t_{i-1}$, we must have $t_{i-1} \leq s_j$. Thus, $s_{j-1} \leq t_{i-1} \leq s_j$. It follows that $z(t_{i-1}) \in D_{i-1} \cap E_i \cap E_j$. In particular, this intersection is non-empty. None of (f_i, D_i) is a direct analytic continuation of (f_{i-1}, D_{i-1}). Moreover, (f_{i-1}, D_{i-1}) is a direct analytic continuation of (g_j, E_j) since $i+j$ is minimal, where we observe that $t_{i-1} \in [t_{i-2}, t_{i-1}] \cap [s_{j-1}, s_j]$ so that the hypothesis of the claim is satisfied. Since $D_{i-1} \cap D_i \cap E_j \neq \emptyset$, (f_i, D_i) must be direct analytic continuation of (g_j, E_j) which is a contradiction. Hence our claim holds for all i and j. In particular, it holds for $i = n$ and $j = m$, which proves the theorem. □

Given a chain $\{(f_1, D_1), (f_2, D_2), \ldots, (f_n, D_n)\}$, can a function $f(z)$ be defined such that $f(z)$ is analytic in the domain $D_1 \cup D_2 \cup \cdots \cup D_n$? Certainly this can be done when $n = 2$. The function

$$\begin{aligned} f(z) &= f_1(z) \text{ if } z \in D_1 \\ &= f_2(z) \text{ if } z \in D_2 \end{aligned}$$

is analytic in $D_1 \cup D_2$. If $D_1 \cap D_2 \cap \cdots \cap D_n \neq \emptyset$, we can show by induction that f, defined by $f(z) = f_i(z)$ for $z \in D_i$, $i = 1, 2, \ldots, n$ is analytic.

However, the proof for the general case fails. Consider the four domains illustrated in Figure 10.3.

For a fixed branch of $\log z$, set $f_1(z) = \log z$ in D_1. The function element (f_1, D_1) determines a unique direct analytic continuation (f_2, D_2) which determines (f_3, D_3) which determines (f_4, D_4). We thus have the chain $\{(f_1, D_1), (f_2, D_2), (f_3, D_3), (f_4, D_4)\}$. However, in the domain $D_1 \cap D_4$, it is not true that $f_1(z) = f_4(z)$. We actually have $f_4(z) = f_1(z) + 2\pi i$ for all points in $D_1 \cap D_4$. The difference in the two functions lies in the fact that the argument of the multiple-valued logarithmic function has increased by 2π after making a complete revolution around the origin. Note also that we can continue (f_1, D_1) into the domain D_3 by different chains and come up with different functions. For the chains $\{(f_1, D_1), (f_2, D_2), (f_3, D_3)\}$ and $\{(f_1, D_1), (g_1, D_4), (g_2, D_3)\}$, we have the values of f_3 and g_2 differing by $2\pi i$. We shall continue this discussion which ultimately culminates into Monodromy theorem, in the upcoming units.

10.4 Power Series Method

We have seen that for a given analytic function f_1 on D_1, if there exists an analytic continuation f_2 on D_2, then it is unique. When does a power series represent a function which is analytic beyond the disc of convergence

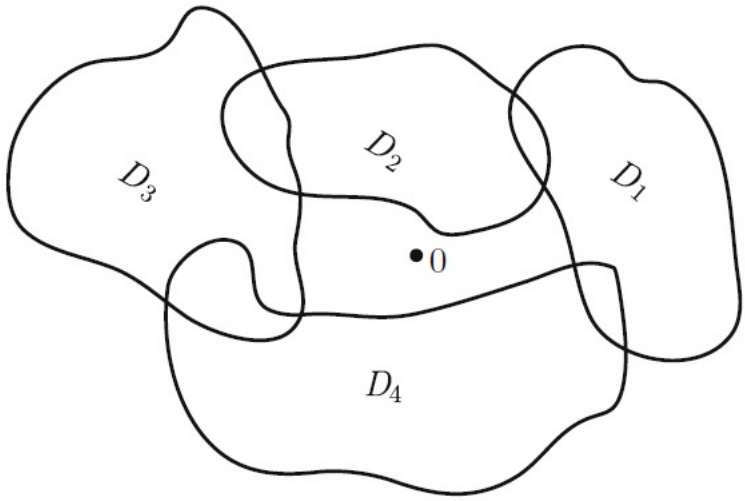


Figure 10.3

of the original series? One way to provide an affirmative answer is by the power series method. Let us start our discussion on this method and see how one can use the power series to go beyond the boundary of the disc of convergence. A fundamental fact about a function f, analytic in a domain D, is that, for each $a \in D$, there exists a sequence $\{a_n\}_{n\geq 0}$ and a number $r_a \in (0, \infty]$ such that

$$f(z) = \sum_{n=0}^{\infty} a_n (z-a)^n \quad \text{for all } \; z \in B(a; r_a).$$

To extend f, we choose a point b other than a in the disc of convergence $B(a; r_a)$. Then $|b-a| < r_a$ and

$$\begin{aligned}
\sum_{n=0}^{\infty} a_n (z-a)^n &= \sum_{n=0}^{\infty} a_n [z-b+b-a]^n \\
&= \sum_{n=0}^{\infty} a_n \left(\sum_{k=0}^{n} \binom{n}{k} (b-a)^{n-k} (z-b)^k \right) \\
&= \sum_{k=0}^{\infty} \left(\sum_{n=k}^{\infty} a_n \binom{n}{k} (b-a)^{n-k} \right) (z-b)^k \\
&= \sum_{k=0}^{\infty} A_k (z-b)^k.
\end{aligned}$$

The interchange of summation is justified since

$$\sum_{n=0}^{\infty} |a_n| \sum_{k=0}^{n} \binom{n}{k} |b-a|^{n-k} |z-b|^k = \sum_{n=0}^{\infty} |a_n| \left(|z-b| + |b-a| \right)^n < \infty$$

whenever $|z-b| + |b-a| < r_a$. Therefore, the series about b converges at least for $|z-b| < r_a - |b-a|$. However, this may happen that the disc of convergence $B(b; r_b)$ for this new series extends outside $B(a; r_a)$, that is, it may be possible that $r_b > r_a - |b-a|$. In this case, the function can be analytically continued to the union of these two discs. This process may be continued.

Example 10.4.1. Let

$$f(z) = \frac{1}{1-z}.$$

Then, for $z \in \mathbb{D}$, with $a = 0$, $r_a = 1$, we have

$$\frac{1}{1-z} = \sum_{n\geq 0} z^n, \quad z \in \mathbb{D}.$$

Take $b = i$. In order to get the expression for $z \in B(i; r_b)$, we write

$$\begin{aligned}
\frac{1}{1-z} &= \frac{1}{1-i-(z-i)} \\
&= \sum_{n=0}^{\infty} \frac{1}{(1-i)^{n+1}}(z-i)^n, \ \ |z-i| < |1-i| = \sqrt{2}, \\
&= \sum_{n=0}^{\infty} A_n(z-i)^n, \quad A_n = \frac{(1+i)^{n+1}}{2^{n+1}}, \ |z-i| < r_b = \sqrt{2}.
\end{aligned}$$

Thus, $\sum_{n=0}^{\infty} A_n(z-i)^n$ is an analytic continuation of $\sum_{n=0}^{\infty} z^n$ in $\mathbb{D}$ to the disc $B(i;\sqrt{2})$. Similarly, one can see that $\sum_{n=0}^{\infty} \frac{1}{2^{n+1}}(z+1)^n$ is an analytic continuation of $\sum_{n=0}^{\infty} z^n$ from $\mathbb{D}$ to the disc $B(-1;2)$.

Example 10.4.2. We show that the function

$$f(z) = \frac{1}{a} + \frac{z}{a^2} + \frac{z^2}{a^3} + \cdots$$

can be continued analytically. This series converges within the circle $C_0 : |z| = |a|$ and has the sum

$$f(z) = \frac{1}{a}\frac{1}{1-\frac{z}{a}} = \frac{1}{a-z}.$$

The only singularity of $f(z)$ on C_0 is at $z = a$. Hence the analytic continuation of $f(z)$ beyond C_0 is possible. For this purpose we take a point $z = b$ not lying on the line segment joining $z = 0$ and $z = a$. We draw a circle C_1 with centre b and radius $|a-b|$ that is, C_1 is $|z-b| = |a-b|$. This new circle C_1 clearly extends beyond C_0 as shown in the figure 10.4

Now, we reconstruct the power series given in powers of $(z-b)$ in the form

$$\sum_{n=0}^{\infty} \frac{(z-b)^n}{(a-b)^{n+1}}, \ \text{ where } \ f^{(n)}(b) = \frac{n!}{(a-b)^{n+1}}. \tag{10.4.1}$$

This power series has the cirlce of convergence C_1 and has cum function $\dfrac{1}{a-z}$. Thus, the power series (10.4.1) and the given power series represent the same function in the region common to C_0 and C_1. Hence (10.4.1) represents an analytic continuation of the given series.

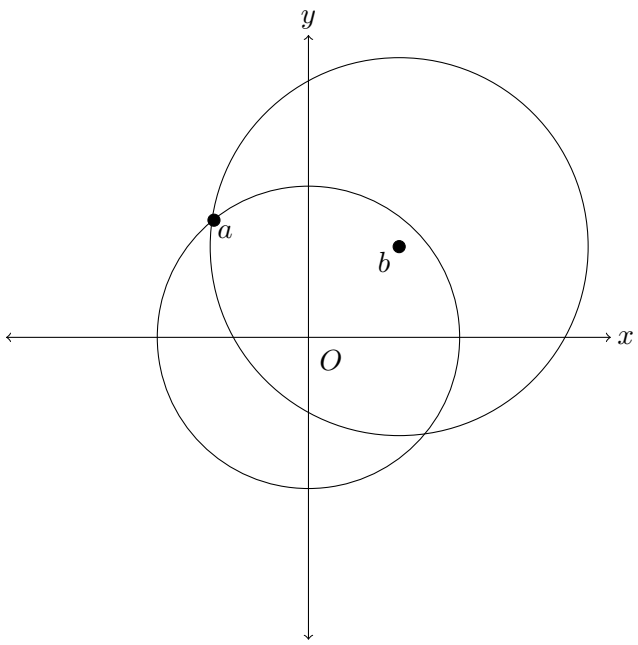


Figure 10.4

10.5 Few Probable Questions

1. Define analytic continuation of an analytic function f from a domain D_1 into another domain D_2. Show that such a continuation is unique. Is the analytic continuation of an analytic function always possible? Justify your answer.

2. Define analytic continuation along a curve. Show that it is unique.

3. Let $\{(f_1, D_1), (f_2, D_2), \ldots, (f_n, D_n)\}$ be a chain. With proper justifications, show that a function defined by $f(z) = f_i(z)$, for $z \in D_i$ is analytic in $D_1 \cup D_2 \cup \cdots \cup D_n$ if $D_1 \cap D_2 \cap \cdots \cap D_n \neq \emptyset$. Is the result true for any general case? Justify.

Unit 11

Course Structure

- Continuation by the method of natural boundary,
- Existence of singularity on the circle of convergence

11.1 Introduction

This unit deals with the continuation by natural boundary. Suppose that a power series has radius of convergence R and defines an analytic function f inside that disc. Consider points on the circle of convergence. A point for which there is a neighbourhood on which f has an analytic extension is regular, otherwise singular. Convergence is limited to within by the presence of at least one singularity on the boundary of . If the singularities on are so densely packed on the circle, that analytic continuation cannot be carried out on a path that crosses, then it is said to form a natural boundary. In particular, the circle is a natural boundary if all its points are singular. More generally, we may apply the definition to any open connected domain on which f is analytic, and classify the points of the boundary of the domain as regular or singular: the domain boundary is then a natural boundary if all points are singular. We will study about this in details.

Objectives

After reading this unit, you will be able to

- define natural boundary of an analytic function f on a domain D
- deduce certain results on the existence of singularities on the circle of convergence

11.2 Continuation by method of natural boundary

We start by the definition of natural boundary.

Definition 11.2.1. (Natural Boundary) Let f be analytic on a domain D. If f cannot be continued analytically across the boundary ∂D, then ∂D is called $natural\ boundary$ of f. A point $z_0 \in \partial D$ is said to be a regular point of $f(z)$ if f can be continued analytically to a region D_1 with $z_0 \in D_1$. Otherwise, $f(z)$ is said to have a $singular\ point$ at z_0.

Example 11.2.1. Consider the power series

$$f(z) = \sum_{k\geq 0} z^{2^k} \tag{11.2.1}$$

A direct consequence of the Root test is that the radius of convergence of the above series is 1 and so, $f(z)$ defined as above is analytic for $|z| < 1$. If $|z| \geq 1$, then $\lim_{n\to\infty} |z^{2n}| \neq 0$ is therefore, the series diverges for $|z| \geq 1$.

Let $\zeta = \mathrm{e}^{2\pi i m/2^n}$, $m = 0, 1, 2, \ldots, 2^n - 1$, $(n \in \mathbb{N})$ be the 2^nth root of unity. If $z = r\,\mathrm{e}^{2\pi i m/2^n} \in \mathbb{D}$, then

$$f(z) = \sum_{k=0}^{n-1} z^{2^k} + \sum_{k=n}^{\infty} z^{2^k}$$

and so for $r \to 1^-$, we have

$$|f(\zeta r)| \geq \sum_{k=n}^{\infty} r^{2^k} - \left|\sum_{k=0}^{n-1} z^{2^k}\right| \geq \sum_{k=n}^{\infty} r^{2^k} - n,$$

and hence, for every 2^nth root of ζ, we have,

$$\lim_{r\to 1^-} |f(\zeta r)| = \infty.$$

Therefore, if D is a domain containing points of $\mathbb{D}$ and of its complement, then D contains the points $\zeta = \mathrm{e}^{2\pi i m/2^n}$ and so any function F in D which coincides with f in $\mathbb{D} \cap D$ cannot be continued analytically through $\zeta^{2n} = 1$ for each $n \in \mathbb{N}$. In other words, any root of the equation

$$z^2 = 1,\ \ z^4 = 1,\ \ \ldots, z^{2n} = 1\ \ (n \in \mathbb{N})$$

is a singular point of f and hence any arc, however small it may be, of $\partial\mathbb{D}$ contains an infinite number of singularities. Thus, f on $\mathbb{D}$ cannot be continued analytically across the boundary $\partial\mathbb{D}$ of $\mathbb{D}$. This observation shows that the unit circle $|z| = 1$ is a natural boundary for the power series defined by (11.2.1).

Example 11.2.2. Similarly, if

$$f(z) = \sum_{k\geq 0} z^{k!} \tag{11.2.2}$$

then f is analytic in $\mathbb{D}$. Upon taking $\zeta = \mathrm{e}^{2\pi i m/n}$, $m = 0, 1, 2, \ldots, n-1$, $z = r\zeta$,(where m/n is the irreducible fraction), and choosing r close to 1 from below along a radius of the unit circle it can be seen that $\lim_{r\to 1^-} |f(\zeta r)| = \infty$. Hence, f is singular at every n-th root of unity for any $n \in \mathbb{N}$. Since, every point on $|z| = 1$ is a singular point, f cannot be continued analytically through the n-th root of unity for any natural number n. In other words, there can be no continuation anywhere across $|z| = 1$ and hence, $|z| = 1$ is a natural boundary for the power series defined by (11.2.2).

11.3 Existence of singularities on the circle of convergence

Theorem 11.3.1. If $f(z) = \sum_{n\geq 0} a_n z^n$ has a radius of convergence $R > 0$, then f must have at least one singularity on $|z| = R$.

Proof. Suppose, on contrary that f has no singularity on $|z| = R$. Then f must be analytic at all points of $|z| = R$. This implies f is analytic on $|z| \leq R$. It follows, from the definition of analyticity at a point, that for each $\zeta \in \partial\mathbb{D}_R$ there exists for some $R_\zeta > 0$ and a function f_ζ which is analytic in $B(\zeta; R_\zeta)$ and

$$f = f_\zeta \quad \text{on B}(\zeta; R_\zeta) \cap \mathbb{D}_R$$

In this way, if ζ_k and $\zeta_l \in \partial\mathbb{D}_R (k \neq l)$ with $G = B(\zeta_k; R_{\zeta_k}) \cap B(\zeta_l; R_{\zeta_l}) \neq \phi$, then we have two functions f_{ζ_k} and f_{ζ_l} which are respectively analytic in $B(\zeta_k; R_{\zeta_k})$ and $B(\zeta_l; R_{\zeta_l})$ such that

$$f = f_{\zeta_k} = f_{\zeta_l} \quad G \cap \mathbb{D}_R$$

Since G is connected and $G \cap \mathbb{D}_R$ is an open subset of G, by the uniqueness theorem, $f_{\zeta_k} = f_{\zeta_l}$ on G. Since

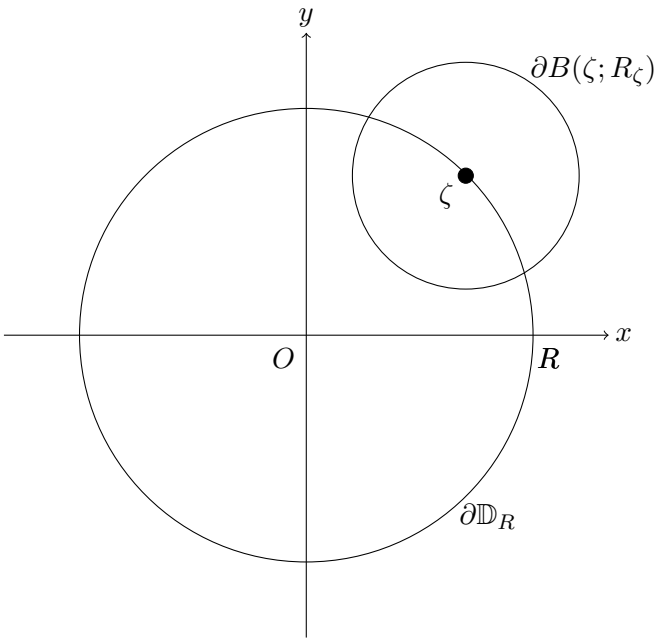


Figure 11.1: Illustration for singularity on circle $|z| = R$.

$|\zeta| = R$ is compact, by the Heine-Borel theorem, we may select a finite number of $B(\zeta_1; R_{\zeta_1}), B(\zeta_2; R_{\zeta_2}), \ldots,$ $B(\zeta_n; R_{\zeta_n})$ from the collection $\{B(\zeta; R_\zeta) : \zeta \in \partial\mathbb{D}_R\}$ such that it covers the circle $\partial\mathbb{D}_R$. Let

$$\Omega = \cup_{k=1}^{n} B(\zeta_k; R_{\zeta_k}) \text{ and } \delta = \text{ dist}(\partial\mathbb{D}_R\, \Omega)$$

Then, as $R_{\zeta_k} > 0$ for each k, we have $\delta > 0$. Moreover,

$$\{z : R - \delta < |z| < R + \delta\} \subset \Omega \quad \text{and} \quad \mathbb{D}_{R+\delta} \subset D = \mathbb{D}_R \cup \Omega$$

Then g is defined by

$$\begin{aligned} g(z) &= f(z) \qquad \text{for } |z| < R \\ &= f_{\zeta_k}(z) \quad \text{for } |z - \zeta_k| < R_{\zeta_k}, \quad k = 1, 2, \ldots, n \end{aligned}$$

as well defined, single-valued and analytic on D and has same power series representation as f for $|z| < R$. Thus there exists an analytic function, say ϕ, in $\mathbb{D}_{R+\delta}$, which coincides with f on $\mathbb{D}_R$. But, then by Taylor's theorem we have the power series representation

$$\phi(z) = \sum_{n \geq 0} b_n z^n \quad \text{for} \quad z \in \mathbb{D}_{R+\delta}$$

Since $f = g$ on $\mathbb{D}_R$, by the uniqueness theorem, we have $a_n = b_n$ for each n. This shows that the radius of convergence of f is $R + \delta$, which is a contradiction. □

Theorem 11.3.2. If $a_n \geq 0$ and $f(z) = \sum_{n\geq 0} a_n z^n$ has radius of convergence 1, then $(f, \mathbb{D})$ has no direct analytic continuation to a function element (F, D) with $1 \in D$.

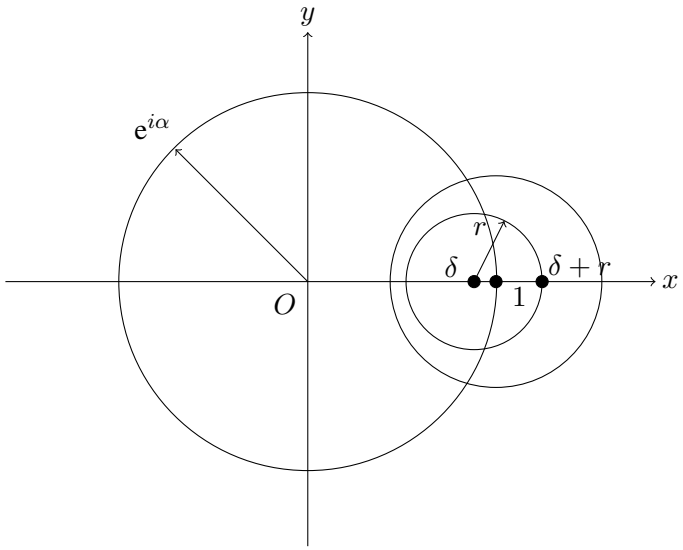


Figure 11.2: Existence of singularity on the circle of convergence

Proof. For each $z = r\,\mathrm{e}^{i\theta} \in \mathbb{D}(0 < r < 1; \theta \in [0, 2\pi])$, we have

$$f^{(k)}(z) = \sum_{n\geq k} n(n-1)\cdots(n-(k-1))a_n z^{n-k} \tag{11.3.1}$$

so that since $a_n \geq 0$

$$|f^{(k)}(r\,\mathrm{e}^{i\theta})| \leq \sum_{n\geq k} n(n-1)\cdots(n-(k-1))a_n z^{n-k} = f^{(k)}(r) \tag{11.3.2}$$

We have to show that 1 is a singular point of f. Suppose, on the contrary that 1 is a regular point of f. Then, f can be analytically continued in a neighbourhood of $z = 1$ and so there is a δ with $0 < \delta < 1$ (see fig. (11.2)) for which the Taylor's series expansion of f about δ, namely the series

$$\sum_{k\geq 0} \frac{f^{(k)}(\delta}{k!}(z-\delta)^k, \tag{11.3.3}$$

would be convergent for $|z - \delta| < r$ with $\delta + r > 1$. Then by (11.3.2), we find that

$$\frac{|f^{(k)}(\delta\,\mathrm{e}^{i\theta})|}{k!} \leq \frac{f^{(k)}(\delta)}{k!}.$$

From this, the root test and the comparison test with (11.3.3), it follows that the radius of convergence of the Taylor series about $\delta\,\mathrm{e}^{i\theta}$ is at least r. This observation implies that the Taylor series

$$\sum_{k\geq 0} \frac{f^{(k)}(\delta\,\mathrm{e}^{i\theta})}{k!}\left(z - \delta\,\mathrm{e}^{i\theta}\right)^k$$

would be convergent in the disc $|z - \delta\, e^{i\theta}| < r$ for each θ, with $\delta + r > 1$. In other words, the Taylor series

$$\sum_{k\geq 0} \frac{f^{(k)}(z_0)}{k!}\,(z - z_0)^k$$

about each z_0 with $|z_0| = \delta$ would have radius of convergence $\geq r > 1 - \delta$. Since this contradicts the previous theorem, and hence 1 must be a singular point of f. This completes the proof. □

Notice that the last series is actually a rearrangement of $\sum_{n\geq 0} a_n z^n$. Indeed, by (11.3.1),

$$\begin{aligned}\sum_{k\geq 0}\left(\sum_{n\geq k}\binom{n}{k} a_n z_0^{n-k}\right)(z - z_0)^k &= \sum_{n\geq 0}\sum_{k=0}^{n}\binom{n}{k} a_n z_0^{n-k}(z - z_0)^k \\ &= \sum_{n\geq 0} a_n (z - z_0 + z_0)^n \\ &= \sum_{n\geq 0} a_n z^n.\end{aligned}$$

Corollary 11.3.1. If $a_n \geq 0$ and $f(z) = \sum_{n\geq 0} a_n z^n$ has the radius of convergence $R > 0$, then $z = R$ is a singularity of $f(z)$.

Finally, we state the following result.

Theorem 11.3.3. If $f(z) = \sum_{k\geq 0} a_k z^{n_k}$ and $\liminf_{k\to\infty} \frac{n_{k+1}}{n_k} > 1$. Then the circle of convergence of the power series is the natural boundary for f.

Example 11.3.1. Consider the function

$$f(z) = \sum_{k=0}^{\infty} \frac{z^{3^k}}{3^k}.$$

Clearly, $n_k = 3^k$. Thus,

$$\liminf_{k\to\infty} \frac{n_{k+1}}{n_k} = \liminf_{k\to\infty} \frac{3^{k+1}}{3^k} = \liminf_{k\to\infty} 3\frac{3^k}{3^k} = 3 > 1.$$

Thus, by the previous theorem, the circle of convergence of the given series is the natural boundary for f. We find the radius of convergence by Cauchy-Hadamard's theorem. Let R be the radius of convergence of the power series. Writing the given power series as $\sum_{n=0}^{\infty} a_n z^n$, we get the terms a_n of the power series as

$$f(z) = 0 \cdot z^0 + 1 \cdot z^1 + 0 \cdot z^2 + \frac{1}{3}z^3 + \cdots + \frac{1}{9}z^9 + \cdots$$

Thus, the set $|a_n|^{\frac{1}{n}}$ is given below

$$\left\{|1|^1,\ |0|^{\frac{1}{2}},\ \left|\frac{1}{3}\right|^{\frac{1}{3}},\ 0,\ 0,\ \ldots, \left|\frac{1}{9}\right|^{\frac{1}{9}}, \ldots\right\}.$$

Thus, Cauchy Hadamard's theorem yields

$$\begin{aligned} R = \frac{1}{\limsup\limits_{n\to\infty} |a_n|^{\frac{1}{n}}} &= \frac{1}{\limsup\limits_{k\to\infty} |a_{3^k}|^{\frac{1}{3^k}}} \\ &= \frac{1}{\limsup\limits_{k\to\infty} \left|\frac{1}{3^k}\right|^{\frac{1}{3^k}}} \\ &= \limsup_{k\to\infty} \left(3^k\right)^{\frac{1}{3^k}} = 1. \end{aligned}$$

Hence, $R = 1$ is the radius of convergence of the given power series and $|z| = 1$ is the natural boundary of f.

11.4 Few Probable Questions

---

1. Show that the function $f(z) = \sum\limits_{n\geq 0} a_n z^n$ having radius of convergence $R > 0$ must have at least one singularity on $|z| = R$.
2. If $a_n \geq 0$, and
3. Find the natural boundary of the function $f(z) = \sum\limits_{n\geq 0} a_n z^n$ has radius of convergence 1, the show that $(f, \mathbb{D})$ has no direct analytic continuation to a function element (F, D), with $1 \in D$.

$$f(z) = \sum_{k=0}^{\infty} \frac{z^{2^k}}{2^{k^2}}.$$

Unit 12

Course Structure

- Monodromy theorem
- germs

12.1 Introduction

This unit is a continuation of the previous unit an deals in the Monodromy theorem. In complex analysis, the Monodromy theorem is an important result about analytic continuation of a complex-analytic function to a larger set. The idea is that one can extend a complex-analytic function (from here on called simply analytic function) along curves starting in the original domain of the function and ending in the larger set. A potential problem of this analytic continuation along a curve strategy is there are usually many curves which end up at the same point in the larger set. The Monodromy theorem gives sufficient conditions for analytic continuation to give the same value at a given point regardless of the curve used to get there, so that the resulting extended analytic function is well-defined and single-valued.

Objectives

After reading this unit, you will be able to

- define homotopy of two curves
- define germ of an analytic function f at a point a
- deduce Monodromy theorem

12.2 Monodromy Theorem

We first give some definitions.

Definition 12.2.1. Let $\gamma_0, \gamma_1 : [0,1] \to G$ be two closed rectifiable curves in a region G then γ_0 is homotopic to γ_1 in G if there is a continuous function

$$F : [0,1] \times [0,1] \to G$$

such that

$$\begin{aligned} F(s,0) &= \gamma_0(s) \\ F(s,1) &= \gamma_1(s) \qquad (0 \le s \le 1) \\ F(0,t) &= F(1,t) \quad (0 \le t \le 1) \end{aligned}$$

Definition 12.2.2. Let $\gamma_0, \gamma_1 : [0,1] \to G$ be two closed rectifiable curves in G such that $\gamma_0(0) = \gamma_1(0) = a$ and $\gamma_0(1) = \gamma_1(1) = b$. Then γ_0 and γ_1 are fixed-end-point homotopic (FEP homotopic) if there is a continuous map $F : [0,1] \times [0,1] \to G$ such that

$$\begin{aligned} &F(s,0) = \gamma_0(s), \;\; F(s,1) = \gamma_1(s) \\ &F(0,t) = a, \;\; F(1,t) = b, \quad \text{for } \; 0 \le s,t \le 1. \end{aligned}$$

We note that the relation of FEP homotopic is an equivalence relation on the curves from one given point to another.

Definition 12.2.3. An open set G is called simply connected if G is connected and every closed curve in G is homotopic to zero.

This is equivalent to the definition of simply connected region which we had learnt previously which states that a set is simply connected if every closed rectifiable curve can be continuously deformed to a single point without passing through any point outside the set. Now, we define the germ of a function f.

Definition 12.2.4. Let (f, G) be a function element. Then the germ of f at a is the collection of all function elements (g, D) such that $a \in D$ and $f(z) = g(z)$ for all z in a neighbourhood of a. The germ of f at a is denoted by $[f]_a$.

Notice that $[f]_a$ is a collection of function elements.

Definition 12.2.5. Let $\gamma : [0,1] \to \mathbb{C}$ be a path and suppose that for each $t \in [0,1]$ there is a function element (f_t, D_t) such that

1. $\gamma(t) \in D_t$;
2. for each $t \in [0,1]$, there is a $\delta > 0$ such that $|s - t| < \delta$ implies that $\gamma(s) \in D_t$ and

$$[f_s]_{\gamma(s)} = [f_t]_{\gamma(s)}.$$

Then (f_1, D_1) is called analytic continuation of (f_0, D_0) along the path γ.

Remark 12.2.1. Since γ is a continuous function and $\gamma(t)$ is in the open set D_t, so there is a $\delta > 0$ such that $\gamma(s) \in D_t$ for $|s - t| < \delta$.

So, part 2 of the previous definition implies

$$f_s(z) = f_t(z) \;\; \text{for all} \;\; z \in D_s \cap D_t,$$

whenever $|s - t| < \delta$.

Theorem 12.2.1. Let $\gamma : [0,1] \to \mathbb{C}$ be a path from a to b and let $\{(f_t, D_t) : 0 \leq 1\}$ and $\{(g_t, B_t) : 0 \leq 1\}$ be analytic continuation along γ such that $[f_0]_a = [g_0]_a$. Then $[f_1]_b = [g_1]_b$.

Proof. Consider the set

$$T = \{t \in [0,1] : [f_t]_{\gamma(t)} = [g_t]_{\gamma(t)}\}$$

Since $[f_0]_a = [g_0]_a$, so $0 \in T$. Thus $T \neq \phi$.

We shall show that T is both open and closed. To show T is open, let t be a fixed point of T such that $t \neq 0$. By definition of analytic continuation, there is $\delta > 0$ such that for $|s - t| < \delta$.

$$\begin{aligned} &\gamma(s) \in D_t \cap B_t \quad \text{and} \\ &[f_s]_{\gamma(s)} = [f_t]_{\gamma(s)} \\ &[g_s]_{\gamma(s)} = [g_t]_{\gamma(s)} \end{aligned}$$

But $t \in T$ implies

$$f_t(z) = g_t(z) \qquad \forall z \in D_t \cap B_t$$

Hence, $[f_t]_{\gamma(s)} = [g_t]_{\gamma(s)}$ for all $\gamma(s) \in D_t \cap B_t$. So, $[f_s]_{\gamma(s)} = [g_s]_{\gamma(s)}$ whenever $|s - t| < \delta$. That is, $s \in T$ whenever $|s - t| < \delta$ or $(t - \delta, t + \delta) \subset T$.

If $t = 0$ then the above argument shows that $[a, a + \delta) \subset T$ for some $\delta > 0$. Hence T is open.

To show that T is closed let t be a limit point of T. Again by definition of analytic continuation there is a $\delta > 0$ such that $|s - t| < \delta$, $\gamma(s) \in D_t \cap B_t$ and

$$\begin{aligned} &[f_s]_{\gamma(s)} = [f_t]_{\gamma(s)} \\ &[g_s]_{\gamma(s)} = [g_t]_{\gamma(s)} \end{aligned} \tag{12.2.1}$$

Since t is a limit point of T, there is a point s in T such that $|s - t| < \delta$. Let $G = D_t \cap B_t \cap D_s \cap B_s$. Then $\gamma(s) \in G$. So, G is non-empty open set. Thus by definition of T, $f_s(z) = g_s(z)$ for all $z \in G$. But, (12.2.1) implies

$$\begin{aligned} &f_s(z) = f_t(z) \quad \text{and} \quad g_s(z) = g_t(z) \quad \text{for all } z \in G \\ &f_t(z) = g_t(z) \quad \forall z \in G. \end{aligned}$$

Since, G has a limit point in $D_t \cap B_t$, this gives $[f_t]_{\gamma(t)} = [g_t]_{\gamma(t)}$. Thus, $t \in T$ and so T is closed.

Now,T is non-empty subset of $[0,1]$ such that T is both open and closed. So, connectedness of $[0,1]$ implies $T = [0,1]$. Thus $1 \in T$ and hence $[f_1]_{\gamma(1)} = [g_1]_{\gamma(1)}$, that is, $[f_1]_b = [g_1]_b$ as $\gamma(1) = b$ □

Definition 12.2.6. If $\gamma : [0,1] \to \mathbb{C}$ is a path from a to b and $\{(f_t, D_t) : 0 \leq t \leq 1\}$ is an analytic continuation along γ then the germ $[f_1]_b$ is the analytic continuation of $[f_0]_a$ along γ.

Remark 12.2.2. Suppose a and b are two complex numbers and let γ and σ be two paths from a to b. Suppose, $\{(f_t, D_t)\}$ and $\{(g_t, D_t)\}$ are analytic continuations along γ and σ respectively such that $[f_0]_a = [g_0]_a$. Now, the question is, does it follow that $[f_1]_b = [g_1]_b$? If γ and σ are the same path then the above result gives an affirmative answer. However, if γ and σ are distinct then the answer can be no.

Lemma 12.2.1. Let $\gamma : [0,1] \to \mathbb{C}$ be a path and let $\{(f_t, D_t) : 0 \leq t \leq 1\}$ be an analytic continuation along γ. For $0 \leq t \leq 1$, let $R(t)$ be the radius of convergence of the power series expansion of f about $z = \gamma(t)$. Then either $R(t) \equiv \infty$ or $R : [0,1] \to (0, \infty)$ is continuous.

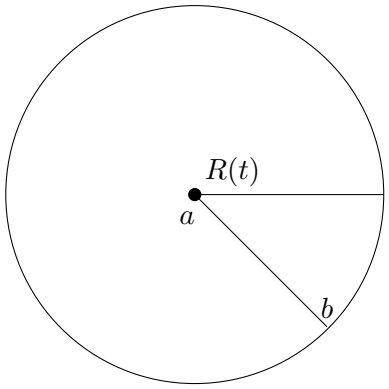


Figure 12.1

Proof. Suppose $R(t) = \infty$ for some value of t. Then, f_t can be extended to an entire function. It follows that $f_s(z) = f_t(z)$ for all $z \in D_s$ so that $R(s) = \infty$ for all $s \in [0,1]$. That is $R(s) \equiv \infty$. Now, suppose that $R(t) < \infty$ for all t. Let t be a fixed number in $[0,1]$ and let $a = \gamma(t)$. Let

$$f_t(z) = \sum_{n=0}^{\infty} a_n (z-a)^n$$

be the power series expansion of f_t about a. Now, let $\delta_1 > 0$ be such that $|s-t| < \delta_1$ implies that $\gamma(s) \in D_t \cap B(a; R(t))$ and $[f_s]_{\gamma(s)} = [f_t]_{\gamma(s)}$. Fix s with $|s-t| < \delta_1$ and let $b = \gamma(s)$. Now, f_t can be extended to an analytic function on $B(a; R(t))$. Since, f_s agrees with f_t on a neighbourhood of f_s can be extended so that it is also analytic on $B(a; R(t)) \cup D_s$. If f_s has power series expansion

$$f_s(z) = \sum_{n=0}^{\infty} b_n (z-b)^n \quad \text{about} \quad z = b$$

Then the radius of convergence $R(s)$ must be at least as big as the distance from b to the circle $|z-a| = R(t)$; that is,

$$\begin{aligned} R(s) &\geq d(b, \{z : |z-a| = R(t)\}) \\ &\geq R(t) - |a-b| \end{aligned}$$

This implies $R(t) - R(s) \leq |a-b|$ that is $R(t) - R(s) \leq |\gamma(t) - \gamma(s)|$. Similarly, we can show $R(s) - R(t) \leq |\gamma(t) - \gamma(s)|$. Hence,

$$|R(s) - R(t)| \leq |\gamma(t) - \gamma(s)| \quad \text{for} \quad |s-t| < \delta_1.$$

Since, $\gamma : [0,1] \to \mathbb{C}$ is continuous so given $\epsilon > 0$, $\exists \delta_2 > 0$ so that $|\gamma(t) - \gamma(s)| < \epsilon$ for $|s-t| < \delta_2$. Let $\delta = \min\{\delta_1, \delta_2\}$. Then $\delta > 0$ and $|R(s) - R(t)| < \epsilon$ for $|s-t| < \delta$. Hence R is continuous at t.

□

Lemma 12.2.2. Let $\gamma : [0,1] \to \mathbb{C}$ be a path from a to b and let $\{(f_t, D_t) : 0 \leq t \leq 1\}$ be an analytic continuation along γ. There is a number $\epsilon > 0$ such that if $\sigma : [0,1] \to \mathbb{C}$ is any path from a to b with $|\gamma(t) - \sigma(t)| < \epsilon$ for all t and if $\{(g_t, B_t) : 0 \leq t \leq 1\}$ is any continuation along γ with $[g_0]_a = [f_0]_a$; the $[g_1]_b = [f_1]_b$.

Proof. For $0 \leq t \leq 1$, let $R(t)$ be the radius of convergence of the power series expansion of f_t about $z = \gamma(t)$. If $R(t) \equiv \infty$ then any value of ϵ will be sufficient. So, suppose $R(t) < \infty$ for all t. Since R is a

continuous function and $R(t) > 0$ for all t, R has a positive minimum value. Let $0 < \epsilon < \frac{1}{2}\min\{R(t) : 0 \leq t \leq 1\}$. Suppose $\sigma : [0,1] \to \mathbb{C}$ is any path from a to b with $|\gamma(t) - \sigma(s)| < \epsilon$ for all t and $\{(g_t, B_t) : 0 \leq t \leq 1\}$ is any continuation along σ with $[g_0]_a = [f_0]_a$. Suppose D_t is a disk of radius $R(t)$ about $\gamma(t)$. Since $|\sigma(t) - \gamma(t)| < \epsilon < R(t)$, $\sigma(t) \in B_t \cap D_t$ for all t.

Define the set $T = \{t \in [0,1] : f_t(z) = g_t(z) \ \forall \ z \in B_t \cap D_t\}$. Then $0 \in T$, since $[g_0]_a = [f_0]_a$. So, $T \neq \phi$. We will show that $1 \in T$. For this, it is sufficient to show that T is both open and closed subset of $[0,1]$.

To show T is open, let t be any fixed point of T. Choose $\delta > 0$

$$\begin{aligned} &|\gamma(s) - \gamma(t)| < \epsilon, \quad [f_s]_{\gamma(s)} = [f_t]_{\gamma(s)} \\ &|\sigma(s) - \sigma(t)| < \epsilon, \quad [g_s]_{\sigma(s)} = [g_t]_{\sigma(s)} \end{aligned} \tag{12.2.2}$$

and $\sigma(s) \in B_t$ whenever $|s - t| < \delta$.

We now show that $B_s \cap B_t \cap D_s \cap D_t \neq \phi$ for $|s-t| < \delta$. For this we will show $\sigma(s) \in B_s \cap B_t \cap D_s \cap D_t$ for $|s - t| < \delta$. If $|s - t| < \delta$, then

$$|\sigma(s) - \gamma(s)| < \epsilon < R(s)$$

so that $\sigma(s) \in D_s$. Also

$$|\sigma(s) - \gamma(t)| = |\sigma(s) - \gamma(s) + \gamma(s) - \gamma(t)| \leq |\sigma(s) - \gamma(s)| + |\gamma(s) - \gamma(t)| < 2\epsilon < R(t)$$

So, $\sigma(s) \in D_t$. Since we already have $\sigma(s) \in B_s \cap B_t$, so (12.2.2) we have $\sigma(s) \in B_s \cap B_t \cap D_s \cap D_t = G$. Since, $t \in T$, it follows that $f_t(z) = g_t(z)$ for all $z \in G$. Also, (12.2.2) implies $f_s(z) = f_t(z)$ and $g_s(z) = g_t(z)$ for all $z \in G$. Thus, $f_s(z) = g_s(z)$ for all $z \in G$. But since G has a limit point in $B_s \cap D_s$, we must have $s \in T$. That is, $(t - \delta, t + \delta) \subset T$. Hence, T is open.

Similarly, we can show T is closed. T is non-empty and closed subset of $[0,1]$. As, $[0,1]$ is connected, we have $[0,1] = T$. Thus $1 \in T$ and the result follows. □

Definition 12.2.7. Let (f, D) be a function element and let G be a region which contains D. Then (f, D) admits unrestricted analytic continuation in G if for any path γ in G with initial point in D there is an analytic continuation of (f, D) along γ.

Theorem 12.2.2. (Monodromy Theorem) Let (f, D) be a function element and let G be a region containing D such that (f, D) admits unrestricted continuation in G. Let $a \in D$, $b \in G$ and let γ_0 and γ_1 be paths in G from a to b; let $\{(f_t, D_t) : \ 0 \leq t \leq 1\}$ and $\{(g_t, D_t) : \ 0 \leq t \leq 1\}$ be analytic continuations of (f, D) along γ_0 and γ_1 respectively. If γ_0 and γ_1 are FEP homotopic in G, then

$$[f_1]_b = [g_1]_b.$$

Proof. Since γ_0 and γ_1 are fixed end point homotopic in G, there is a continuous function $F : [0,1] \times [0,1] \to G$ such that

$$\begin{aligned} &F(t,0) = \gamma_0(t), \quad F(t,1) = \gamma_1(t) \\ &F(0,u) = a, \quad F(1,u) = b. \end{aligned}$$

For all t and u in $[0,1]$. Let u be a fixed point of $[0,1]$. Consider the path γ_u, defined by

$$\gamma_u(t) = F(t,u) \ \text{ for } \ t \in [0,1].$$

Then,

$$\gamma_u(0) = F(0,u) = a, \ \ \gamma_u(1) = F(1,u) = b.$$

Thus, γ_u is a path from a to b. By hypothesis, there is an analytic continuation

$$\{(h_{t,u}, D_{t,u}) : \ 0 \leq t \leq 1\}$$

of (f, D) along γ_u. Now, $\{(h_{t,u}, D_{t,u}) : \ 0 \leq t \leq 1\}$ and $\{(f_t, D_t) : \ 0 \leq t \leq 1\}$ are analytic continuations along γ_0 so by theorem 12.2.1, we have

$$[f_1]_b = [h_{1,0}]_b.$$

Similarly,

$$[g_1]_b = [h_{1,1}]_b.$$

To prove the theorem, it is sufficient to show

$$[h_{1,0}]_b = [h_{1,1}]_b.$$

Consider the set

$$U = \{u \in [0, 1] : \ [h_{1,u}]_b = [h_{1,0}]_b\}.$$

We will show that $1 \in U$. Now, $0 \in U$. So, $U \neq \emptyset$. We claim that U is both open and closed subset of $[0, 1]$.

Let $u \in [0, 1]$ be arbitrary. We assert that there is $\delta > 0$ such that if $|u - v| < \delta$, then

$$[h_{1,u}]_b = [h_{1,v}]_b. \tag{12.2.3}$$

By lemma 12.2.2, there is an $\epsilon > 0$ such that if σ is any path from a to b with $|\gamma_u(t) - \sigma(t)| < \epsilon$ for all t and if $\{(k_t, E_t)\}$ is any continuation of (f, D) along σ, then

$$[h_{1,u}]_b = [k_1]_b. \tag{12.2.4}$$

Now, F is uniformly continuous function so there is $\delta > 0$ such that

$$\begin{aligned} & |F(t, u) - F(t, v)| < \epsilon \ \text{ whenever } \ |u - v| < \delta \\ \Rightarrow \quad & |\gamma_u(t) - \gamma_v(t)| < \epsilon \quad \text{whenever } \ |u - v| < \delta. \end{aligned}$$

So, for $|u - v| < \delta$, γ_v is a path from a to b with

$$\gamma_u(t) - \gamma_v(t)| < \epsilon$$

for all t and $\{(h_{t,v}, D_{t,v})\}$ is a continuation of (f, D) along γ_v, so by (12.2.4),

$$[h_{1,u}]_b = [h_{1,v}]_b.$$

Suppose $u \in U$ such that $[h_{1,u}]_b = [h_{1,0}]_b$. Then as proved above, there is a $\delta > 0$ such that $|u - v| < \delta$ which implies that

$$\begin{aligned} & [h_{1,u}]_b = [h_{1,v}]_b \\ \text{i.e.} \quad & v \in (u - \delta, u + \delta) \Rightarrow [h_{1,v}]_b = [h_{1,0}]_b \\ \text{i.e.} \quad & v \in (u - \delta, u + \delta) \Rightarrow v \in U \\ \text{i.e.} \quad & (u - \delta, u + \delta) \subset U. \end{aligned}$$

Hence U is open.

To show that U is closed, we show that $\overline{U} = U$. Let $u \in U$ and δ be the positive number satisfying (12.2.3). Then there is a $v \in U$ such that $|u - v| < \delta$. So, by (12.2.3), $[h_{1,u}]_b = [h_{1,v}]_b$. Since $v \in U$, so $[h_{1,v}]_b = [h_{1,0}]_b$. Thus, $[h_{1,u}]_b = [h_{1,0}]_b$ so that $u \in U$. Thus, U is closed as $\overline{U} = U$.

Now, U is a non-empty open and closed subset of $[0, 1]$ and since $[0, 1]$ is connected, so, $U = [0, 1]$. So, $1 \in U$ and the result is proved. $\square$

The following corollary is the main consequence of the Monodromy theorem.

Corollary 12.2.1. Let (f, D) be a function element which admits unrestricted continuation in the simply connected region G. Then there is an analytic function $F : G \to \mathbb{C}$ such that $F(z) = f(z)$ for all $z \in D$.

Proof. Let a be a fixed point in D and z is any point in G. If γ is a path in G from a to z and $\{(f_t, D_t) : 0 \leq t \leq 1\}$ is an analytic continuation of (f, D) along γ, then let $F(z, \gamma) = f_1(z)$ since G is simply connected.

$F(z, \gamma) = F(z, \sigma)$ for any two paths γ and σ in G from a to z. Thus, $F(z) = F(z, \gamma)$ is a well defined function from G to $\mathbb{C}$. To show that F is analytic, let $z \in G$. Let γ be a path in G from a to z and $\{(f_t, D_t)\}$ be the analytic continuation of (f, D) along γ. Then $F(\omega) = f_1(\omega)$ for all ω in a neighbourhood of z. Hence F must be analytic. □

12.3 Few Probable Questions

1. Let (f, G) be a function element. Define germ of f at a. If $\gamma : [0, 1] \to \mathbb{C}$ be a path from a to b and $\{(f_t, D_t) : 0 \leq t \leq 1\}$ and $\{(g_t, B_t) : 0 \leq t \leq 1\}$ be analytic continuation along γ such that $[f_0]_a = [g_0]_a$, then show that $[f_1]_b = [g_1]_b$.
2. State and prove Monodromy theorem.
3. If $\gamma : [0, 1] \to \mathbb{C}$ be a path and $\{(f_t, D_t) : 0 \leq t \leq 1\}$ is an analytic continuation along γ. For $0 \leq t \leq 1$, if $R(t)$ is the radius of convergence of the power series expansion of f_t about $z = \gamma(t)$, then show that either $R(t) \equiv \infty$ or $R : [0, 1] \to (0, \infty)$ is continuous.

Unit 13

Course Structure

- Conformal transformations,
- Riemanns theorems for circle.

13.1 Introduction

In mathematics, a conformal map is a function that locally preserves angles, but not necessarily lengths. We shall see that the derivative relates the angle between two curves to the angle between their images. In addition, the derivative will be seen to measure the "distortion" of image curves. They are also worth studying because of their usefulness in solving certain physical problems, for example, problems about two-dimensional fluid flow, the idea being to transform a given problem into an equivalent one which is easier to solve. So we wish to consider the problem of mapping a given region G onto a geometrically simpler region G'. For example the open unit disc or the open upper half-plane.

Objectives

After reading this unit, you will be able to

- define conformal and isogonal maps and see certain examples
- deduce further conditions satisfied by conformal maps
- define conformally equivalent regions and see certain examples of them
- define M*ö*bius transformation and related terms and deduce few results related to symmetry

13.2 Conformal Transformations

Any straight line in the plane that passes through the origin may be parameterized by $\sigma(s) = s\,\mathrm{e}^{i\alpha}$, where s traverses the set of real numbers and α is the angle measured in radians between the positive real axis and the line. More generally, a straight line passing through the point z_0 and making an angle α with the real axis can be expressed as $\sigma(s) = z_0 + s\,\mathrm{e}^{i\alpha}$, s is real.

Suppose now that a function f is analytic on a smooth (parameterized) curve whose derivative is given by $f'(z(t))z'(t)$ (by chain rule). A smooth curve is characterized by having a tangent at each point. So, we interpret $z'(t)$ as a vector in the direction of the tangent vector at the point $z(t)$. Our purpose is to compare the inclination of the tangent to the curve at a point with the inclination of the tangent to the image curve at the image of the point.

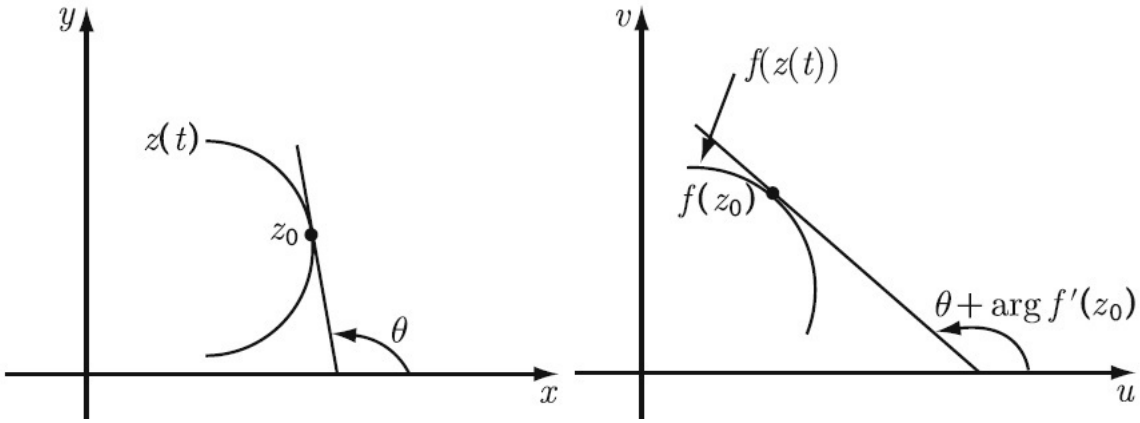
Let $z_0 = z(t_0)$ be a point on the curve $z = z(t)$. Then the vector $z'(t_0)$ is tangent to the curve at the point z_0 and $\arg z'(t_0)$ is the angle this directed tangent makes with the positive x-axis. Suppose that $w = w(t) = f(z(t))$ with $w_0 = f(z_0)$. For any point z on the curve other than z_0, we have the identity

$$w - w_0 = \frac{f(z) - f(z_0)}{z - z_0}(z - z_0).$$

Thus,

$$\arg(w - w_0) = \arg \frac{f(z) - f(z_0)}{z - z_0} + \arg(z - z_0) \qquad (\mod 2\pi), \tag{13.2.1}$$

where it is assumed that $f(z) \neq f(z_0)$ so that (13.2.1) has meaning. Note that $\arg(z - z_0)$ is the angle in the z plane between the x axis and the straight line passing through the points z and z_0, while $\arg(w - w_0)$ is the angle in the w plane between the u axis and the straight line passing through the points w and w_0. Hence, as



z approaches z_0 along the curve $z(t)$, $\arg(z - z_0)$ approaches a value θ, which is the angle that the tangent to the curve $z(t)$ at z_0 makes with the x-axis. Similarly, $\arg(w - w_0)$ approaches a value ϕ, the angle that the tangent to the curve $f(z(t))$ at w_0 makes with the u axis.

Suppose that $f'(z_0) \neq 0$ so that $\arg f'(z_0)$ has a meaning. Then, taking limits in (13.2.1), we find ($\mod 2\pi$) that

$$\phi = \arg f'(z_0) + \theta, \quad \text{or} \quad \arg w'(t_0) = \arg f'(z_0) + \arg z'(t_0). \tag{13.2.2}$$

That is, the difference between the tangent to a curve at a point and the tangent to the image curve at the image of the point depends only on the derivative of the function at the point.

Theorem 13.2.1. Suppose $f(z)$ is analytic at z_0 with $f'(z_0) \neq 0$. Let $C_1 : z_1(t)$ and $C_2 : z_2(t)$ be smooth curves in the z plane that intersect at $z_0 = z_1(t_0) = z_2(t_0)$ with $C_1' : w_1(t)$ and $C_2' : w_2(t)$ the images of C_1 and C_2, respectively. Then the angle between C_1 and C_2, measured from C_1 to C_2, is equal to the angle between C_1' and C_2' measured from C_1' to C_2'.

Proof. Let the tangents to C_1 and C_2 make angles θ_1 and θ_2 respectively with the x-axis. Then the angle between C_1 and C_2 at z_0 is $\theta_2 - \theta_1$ (see fig. 13.1). According to (13.2.2), the angle between C_1' and C_2', which

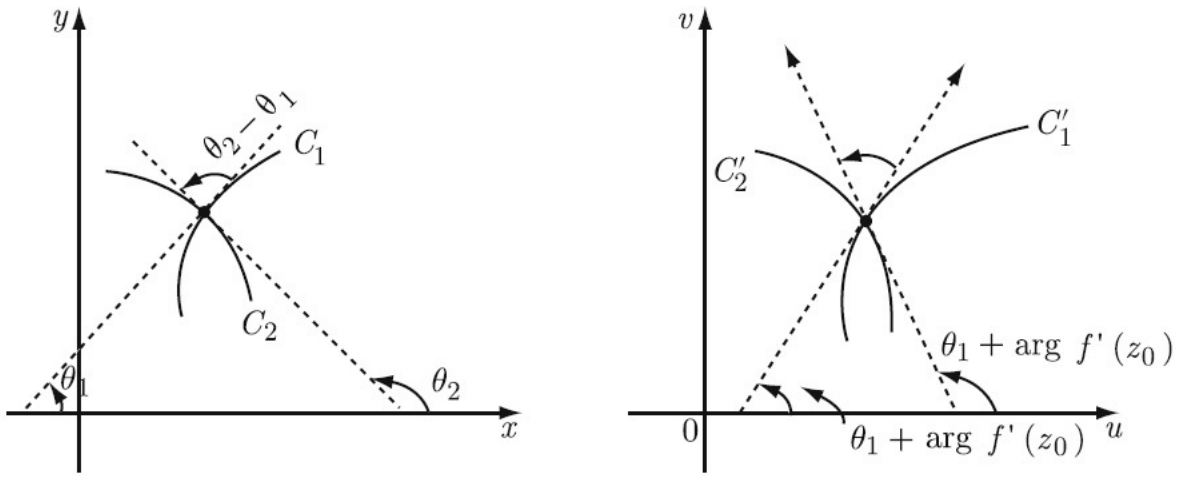


Figure 13.1

is the angle between the tangent vectors $f'(z_0)z_1'(t_0)$ and $f'(z_0)z_2'(t_0)$, of the image curves is

$$\theta_2 + \arg f'(z_0) - (\theta_1 + \arg f'(z_0)) = \theta_2 - \theta_1,$$

and the theorem is proved. □

A function that preserves both angle size and orientation is said to be **conformal**. Theorem 13.2 says that an analytic function is conformal at all points where the derivative is non-zero. For example, the function $f(z) = \mathrm{e}^z$ maps vertical and horizontal lines into circles and orthogonal radial rays, respectively.

A function that preserves angle size but not orientation is said to be **isogonal**. An example of such a function is $f(z) = \overline{z}$. To illustrate, $\overline{z}$ maps maps the positive real axis and the positive imaginary axis onto the positive real axis and the negative real axis respectively (see fig. 13.2). Although the two curves intersect at right angles in each plane, a "counterclockwise" angle is mapped onto a "clockwise" angle.

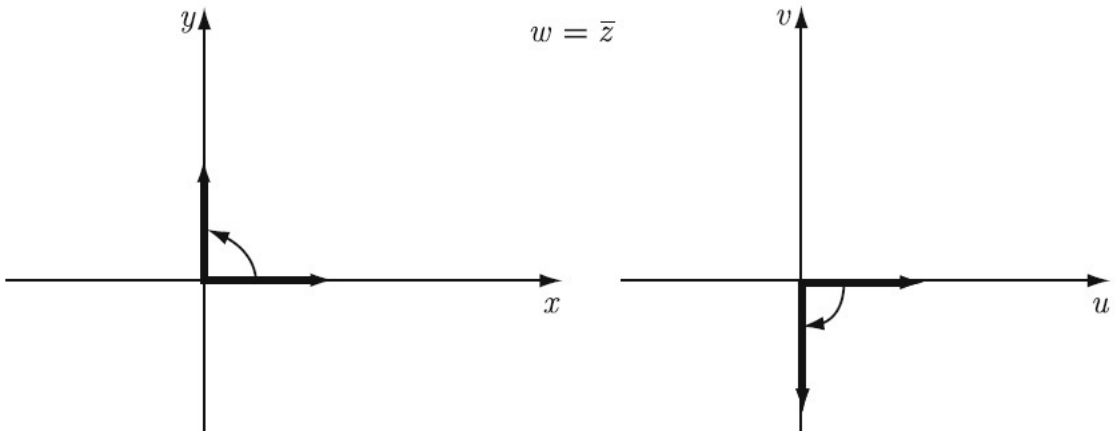


Figure 13.2

The non-zero derivatives of f has certain implications which we shall see now.

Theorem 13.2.2. If $f(z)$ is analytic at z_0 with $f'(z_0) \neq 0$, then $f(z)$ is one-to-one in some neighbourhood of z_0.

Proof. Since $f'(z_0) \neq 0$ and $f'(z)$ is continuous at z_0, there exists a $\delta > 0$ such that

$$|f'(z) - f'(z_0)| < \frac{|f'(z_0)|}{2} \quad \text{for all} \quad |z| < \delta$$

Let z_1 and z_2 be two distinct points in $|z| < \delta$, and γ be a line segment connecting z_1 and z_2. Set $\phi(z) = f(z) - f'(z_0)z$ so that $|\phi'(z)| < |f'(z_0)|/2$ for all $|z| < \delta$. Now we have,

$$|\phi(z_2) - \phi(z_1)| = \left| \int_\gamma \phi'(z) dz \right| < (|f'(z_0)|/2)|z_2 - z_1|$$

or equivalently,

$$|f(z_2) - f(z_1) - f'(z_0)(z_2 - z_1)| < (|f'(z_0)|/2)|z_2 - z_1|.$$

Thus, by the triangle inequality, we obtain

$$|f(z_2) - f(z_1)| > (|f'(z_0)|/2)|z_2 - z_1| > 0.$$

That is, $f(z)$ is one-to-one in $|z| < \delta$. □

The vanishing of a derivative does not preclude the possibility of real function being one-to-one. Although the derivative of $f(x) = x^3$ is zero at the origin, the function is still one-to-one on the real line. That this cannot occur for complex functions is seen by

Theorem 13.2.3. If $f(z)$ is analytic and one-to-one in a domain D, then $f'(z) \neq 0$ in D, so that f is conformal in D.

Proof. If $f'(z) = 0$ at some point z_0 in D, then

$$f(z) - f(z_0) = \frac{f''(z_0)}{2!}(z - z_0)^2 + \cdots$$

has a zero of order $k (k \geq 2)$ at z_0. Since zeros of an analytic function are isolated, there exists an $r > 0$ so small that both $f(z) - f(z_0)$ and $f'(z)$ have no zeros in the punctured disk $0 < |z - z_0| \leq r$. Let $g(z) := f(z) - f(z_0)$, $C = \{z : |z - z_0| = r\}$ and $m = \min_{z \in C} |g(z)|$.

Then, g has a zero of order $k (k \geq 2)$ and $m > 0$. Let $b \in \mathbb{C}$ be such that $0 < |b - f(z_0)| < m$. Then, as $m \leq |g(z)|$ on C,

$$|f(z_0) - b| < |g(z)| \quad \text{on} \quad C$$

It follows from Rouche's theorem that $g(z)$ and $g(z) + (f(z_0) - b) = f(z) - b$ have same number of zeros inside C. Thus, $f(z) - b$ has at least two zeros inside C. Observe that none of these zeros can be at z_0. Since $f'(z) \neq 0$ in the punctured disk $0 < |z - z_0| \leq r$, these zeros must be simple and so, distinct. Thus, $f(z) = b$ at two or more points inside C. This contradicts the fact that f is one-to-one on D. □

We sum up our results for differentiable functions. In the real case,the nonvanishing of a derivative on an interval is a sufficient but not a necessary condition for the function to be one-to-one on the interval; whereas in the complex case, the nonvanishing of a derivative on a domain is a necessary but not a sufficient condition for the function to be one-to-one on the domain.

An analytic function $f : D \to \mathbb{C}$ is called *locally bianalytic* at $z_0 \in D$ if there exists a neighbourhood N of z_0 such that the restriction of f from N onto $f(N)$ is bianalytic. Clearly, a locally bianalytic map on D need not be bianalytic on D, as the example $f(z) = z^n (n > 2)$ on $\mathbb{C} - \{0\}$ illustrates.

Combining 13.2.2 and 13.2.3 leads to the following criterion for local bianalytic maps.

Theorem 13.2.4. Let $f(z)$ be analytic in a domain D and $z_0 \in D$. Then f is bianalytic at z_0 iff $f'(z_0) \neq 0$.

A sufficient condition for an analytic function to be one-to-one in a simply connected domain is that it be one-to-one on its boundary. More formally, we have

Theorem 13.2.5. Let $f(z)$ be analytic in a simply connected domain D and on its boundary, the simple closed contour C. If $f(z)$ is one-to-one on C, then $f(z)$ is one-to-one in D.

Proof. (See fig. 13.3) Choose a point $z_0 \in D$ such that $w_0 = f(z_0) \neq f(z)$ for z on C. According to the argument principle, the number of zeros of $f(z) - f(z_0)$ in D is given by $(1/2\pi i)\int_C \{f(z) - f(z_0)\}dz$. By hypothesis, the image of C must be a simple closed contour, which we shall denote by C'. Thus the net change in the argument of $w - w_0 = f(z) - f(z_0)$ as $w = f(z)$ traverses the contour C' is either $+2\pi$ or -2π, according to whether the contour is traversed counterclockwise or clockwise. Since $f(z)$ assumes the value w_0 at least once in D, we must have

$$\frac{1}{2\pi i}\int_C \{f(z) - f(z_0)\}dz = \frac{1}{2\pi i}\int_C \{w - w_0\} = 1.$$

That is, $f(z)$ assumes the value $f(z_0)$ exactly once in D.

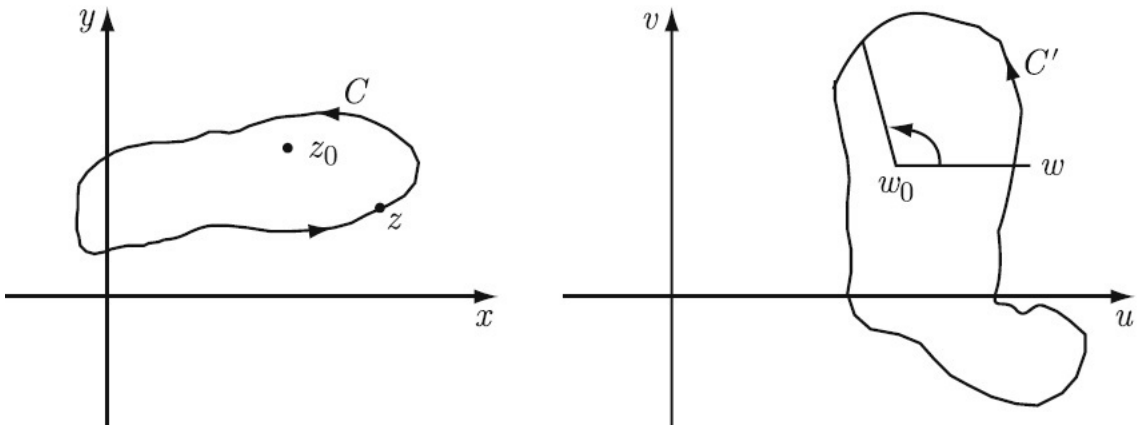


Figure 13.3

This proves the theorem for all points z_0 in D at which $f(z) \neq f(z_0)$ when z is on C. If $f(z) = f(z_0)$ at some point on C, then the expression $\int_C \{f(z) - f(z_0)\}dz$ is not defined. We leave for the reader the completion of the proof in this special case. □

13.3 Conformal Equivalences and Examples

An analytic map $f : G \to G'$ which is bijective is called a bianalytic map as we have already come across. Given such a map f, we say that G and G' are conformally equivalent or simply biholomorphic. An important fact is that the inverse of f is analytic in that case automatically. We have also seen that if an analytic map $f : G \to G'$ is injective, then $f'(z) \neq 0$ for all $z \in G$, that is, f is conformal. We begin our study of conformal mappings by looking at a number of specific examples. The first gives the conformal equivalence between the unit disc and the upper half-plane, which plays an important role in many problems.

Example 13.3.1. Let $\mathbb{H} = \{z \in \mathbb{C} : \text{ Im } z > 0\}$ be the upper half plane. A remarkable fact, which at first seems surprising, is that the unbounded set $\mathbb{H}$ is conformally equivalent to the unit disc. Moreover, an explicit formula giving this equivalence exists. Indeed, let

$$F(z) = \frac{i-z}{i+z} \quad \text{and} \quad G(w) = i\frac{1-w}{1+w}.$$

Then it is a regular exercise to check that map $F : \mathbb{H} \to \mathbb{D}$ is conformal with inverse $G : \mathbb{D} \to \mathbb{H}$. An interesting aspect of these functions is their behaviour on the boundaries of our open sets. Observe that F is analytic everywhere on $\mathbb{C}$ except at $z = -i$, and in particular it is continuous everywhere on the boundary of $\mathbb{H}$, namely, the real line. If we take $z = x$ real, then the distance from x to i is the same as the distance from x to $-i$, therefore $|F(x)| = 1$. Thus, F maps $\mathbb{R}$ onto the boundary of $\mathbb{D}$. We get more information by writing

$$F(z) = \frac{i-x}{i+x} = \frac{1-x^2}{1+x^2} + i\frac{2x}{1+x^2},$$

and parametrizing the real line by $x = \tan t$ with $t \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. Since

$$\sin 2a = \frac{2\tan a}{1+\tan^2 a} \quad \text{and} \quad \cos 2a = \frac{1-\tan^2 a}{1+\tan^2 a},$$

we have, $F(x) = \cos 2t + i \sin 2t = \mathrm{e}^{i2t}$. Hence the image of the real line is the arc consisting of the circle omitting the point -1. Moreover, as x travels from $-\infty$ to ∞, $F(x)$ travels along the arc starting from -1 and first going through that part of the circle that lies in the lower half-plane. The point -1 on the circle corresponds to the "point at infinity" of the upper half-plane.

Example 13.3.2. Mappings of the form

$$z \mapsto \frac{az+b}{cz+d},$$

where a, b, c and d are complex numbers, and where the denominator is assumed not to be a multiple of the numerator, are usually referred to as **fractional linear transformations**.

Example 13.3.3. The map

$$f(z) = \frac{1+z}{1-z}$$

takes the upper half-disc $\{z = x + iy : \ |z| < 1 \ \text{ and } \ y > 0\}$ conformally to the first quadrant $\{w = u + iv : \ u > 0, \ \ v > 0\}$ (see fig. 13.4).

Indeed, if $z = x + iy$, we have so f maps the half-disc in the upper half-plane into the first quadrant. The inverse map, given by

$$g(w) = \frac{w-1}{w+1},$$

is clearly analytic in the first quadrant. Moreover, $|w+1| > |w-1|$ for all w in the first quadrant because the distance from w to -1 is greater than the distance from w to 1; thus g maps into the unit disc. Finally, an easy calculation shows that the imaginary part of $g(w)$ is positive whenever w is in the first quadrant. So g transforms the first quadrant into the desired half-disc and we conclude that f is conformal because g is the inverse of f.

To examine the action of f on the boundary, note that if $z = \mathrm{e}^{i\theta}$ belongs to the upper half-circle, then

$$f(z) = \frac{1+\mathrm{e}^{i\theta}}{1-\mathrm{e}^{i\theta}} = \frac{\mathrm{e}^{-i\theta/2} + \mathrm{e}^{i\theta/2}}{\mathrm{e}^{-i\theta/2} - \mathrm{e}^{i\theta/2}} = \frac{i}{\tan(\theta/2)}.$$

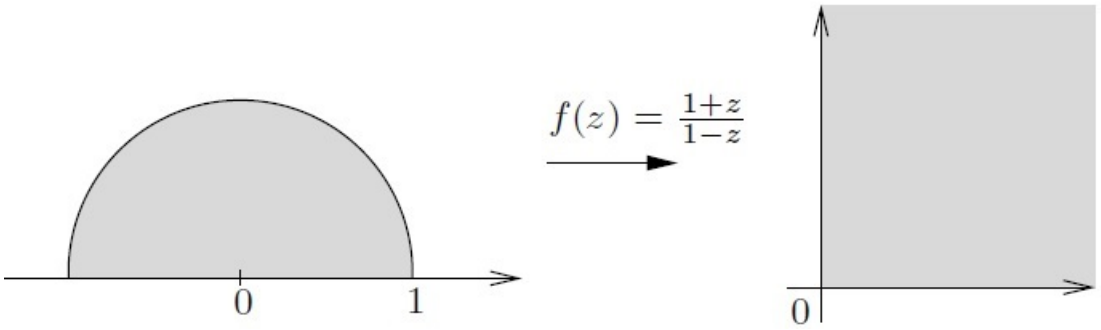


Figure 13.4

As θ travels from 0 to π we see that $f(e^{i\theta})$ travels along the imaginary axis from infinity to 0. Moreover, if $z = x$ is real, then

$$f(z) = \frac{1+x}{1-x}$$

is also real; and one sees from this, that f is actually a bijection from $(-1, 1)$ to the positive real axis, with $f(x)$ increasing from 0 to infinity as x travels from -1 to 1. Note also that $f(0) = 1$.

Exercise 13.3.1.

1. Show that for $h \in \mathbb{C}$, the translation map $f(z) = z + h$ is a conformal map from $\mathbb{C}$ to itself.
2. Show that the map $f(z) = e^{iz}$ takes the half-strip $\left\{z = x + iy : \ -\frac{\pi}{2} < x < \frac{\pi}{2}, \ y > 0\right\}$ conformally to the half-disc $\{w = u + iv : \ |w| < 1, \ u > 0\}$.

13.4 Möbius Transformations

We have already seen that the functions of the form

$$f(z) = \frac{az + b}{cz + d} \tag{13.4.1}$$

is a linear fractional transformation. If $ad - bc \neq 0$, then $f(z)$ is called a Möbius Transformation. If f is a Möbius Transformation, then

$$f^{-1}(z) = \frac{dz - b}{-cz + a}$$

is the inverse map of f. Also, if f and g are two linear fractional transformations, then their composition $f \circ g$ is also so. Hence, the set of all Möbius Transformations form a group under group composition.

Theorem 13.4.1. If f is a Möbius Transformation, then f is the composition of translations, dilations and inversion.

The fixed points of a Möbius Transformation (13.4.1) are the points where $f(z) = z$, that is,

$$cz^2 + (d - a)z - b = 0.$$

Hence a Möbius Transformation has at most two fixed points unless it is the identity transformation.

Now, let f be a Möbius Transformation and let a, b, c be distinct points in $\mathbb{C}_\infty$ such that $f(a) = \alpha,\ f(b) = \beta,\ f(c) = \gamma$. Suppose that g is another Möbius Transformation with the same property. Then $g^{-1} \circ f$ has a, b and c as fixed points and hence it is the identity transformation and thus, $f \equiv g$. Thus a Möbius Transformation is uniquely determined by its action on three points in $\mathbb{C}_\infty$.

Let z_2, z_3 and z_4 be points on $\mathbb{C}_\infty$. Define $f : \mathbb{C}_\infty \to \mathbb{C}_\infty$ by

$$\begin{aligned} f(z) &= \frac{z - z_3}{z - z_4} \cdot \frac{z_2 - z_4}{z_2 - z_3} && \text{if } z_2, z_3, z_4 \in \mathbb{C}_\infty \\ &= \frac{z - z_3}{z - z_4} && \text{if } z_2 = \infty \\ &= \frac{z_2 - z_4}{z - z_4} && \text{if } z_3 = \infty \\ &= \frac{z - z_3}{z_2 - z_3} && \text{if } z_4 = \infty. \end{aligned}$$

In any case, $f(z_2) = 1,\ f(z_3) = 0,\ f(z_4) = \infty$ and f is the only transformation having this property.

Definition 13.4.1. If $z_1 \in \mathbb{C}_\infty$, then the **cross ratio** of z_1, z_2, z_3 and z_4 is the image of z_1 under the unique Mö transformation which takes z_2 to 1, z_3 to 0 and z_4 to ∞. The cross ratio of z_1, z_2, z_3 and z_4 is denoted by (z_1, z_2, z_3, z_4).

For example, $(z_2, z_2, z_3, z_4) = 1$ and $(z, 1, 0, \infty) = z$. Also, if M is a Möbius map and w_2, w_3, w_4 are the points such that $Mw_2 = 1,\ Mw_3 = 0$ and $Mw_4 = \infty$, then $Mz = (z, w_2, w_3, w_4)$.

Theorem 13.4.2. If z_2, z_3 and z_4 are distinct points and T is any Möbius transformation, then

$$(z_1, z_2, z_3, z_4) = (T(z_1), T(z_2), T(z_3), T(z_4))$$

for any point z_1.

Proof. Let $S(z) = (z, z_2, z_3, z_4)$. Then S is a Möbius map. If $M = ST^{-1}$, then $M(T(z_2)) = 1,\ M(T(z_3)) = 0,\ M(T(z_4)) = \infty$. Hence, $ST^{-1}(z) = (z, T(z_2), T(z_3), T(z_4))$ for all $z \in \mathbb{C}_\infty$. In particular, if $z = T(z_1)$, the desired result follows. □

Theorem 13.4.3. If z_2, z_3, z_4 are distinct points in $\mathbb{C}_\infty$ and w_2, w_3, w_4 are also distinct points of $\mathbb{C}_\infty$, then there is one and only one Möbius transformation S such that $S(z_2) = w_2, S(z_3) = w_3, S(z_4) = w_4$.

Proof. Let $T(z) = (z, z_2, z_3, z_4)$, $M(z) = (z, w_2, w_3, w_4)$ and put $S = M^{-1}T$. Clearly, S has the desired property. If R is another Möbius transformation with $Rz_j = w_j$ for $j = 2, 3, 4$ then $R^{-1} \cdot S$ has three fixed points (z_2, z_3 and z_4). Hence, $R^{-1} \cdot S = I$ or $S = R$. □

It is well known that three points in the plane determine a circle. The next result explains when four points lie on a circle.

Theorem 13.4.4. Let z_1, z_2, z_3, z_4 be four distinct points in $\mathbb{C}_\infty$. Then (z_1, z_2, z_3, z_4) is a real number iff all four points lie on a circle.

Proof. Let $S : \mathbb{C}_\infty \to \mathbb{C}_\infty$ be defined by $S(z) = (z, z_2, z_3, z_4)$; then $S^{-1}(\mathbb{R})$=the set of z such that (z, z_2, z_3, z_4) is real. Hence, we will be finished if we can show that the image of $\mathbb{R}_\infty$ under a Möbius map us a circle.

Let

$$S(z) = \frac{az + b}{cz + d} \tag{13.4.2}$$

If $z = w \in \mathbb{R}$ and $w = S^{-1}(x)$ then $x = S(w)$ implies that $S(w) = \overline{S(w)}$. That is,

$$\frac{aw+b}{cw+d} = \frac{\overline{a}\overline{w}+\overline{b}}{\overline{c}\overline{w}+\overline{d}}$$

Cross multiplying this gives

$$(a\overline{c}-\overline{a}c)|w|^2 + (a\overline{d}-\overline{b}c)w + (b\overline{c}-d\overline{a})\overline{w} + (b\overline{d}-\overline{b}d) = 0 \tag{13.4.3}$$

If $a\overline{c}$ is real then $a\overline{c}-\overline{a}c = 0$; putting $\alpha = 2(a\overline{d}-\overline{b}c)$, $\beta = i(b\overline{d}-\overline{b}d)$ and multiplying (13.4.3) by i gives

$$0 = \text{Im}\,(\alpha w) - \beta = \text{Im}\,(\alpha w - \beta) \tag{13.4.4}$$

since β is real. That is, w lies on the line determined by (13.4.4) for fixed α and β. If $a\overline{c}$ is not real then (13.4.3) becomes

$$|w|^2 + \overline{\gamma}w + \gamma\overline{w} - \delta = 0 \tag{13.4.5}$$

for some constants γ in $\mathbb{C}$, δ in $\mathbb{R}$. Hence,

$$|w+\gamma| = \lambda \tag{13.4.6}$$

where

$$\lambda = \sqrt{|\gamma|^2+\delta} = \left|\frac{ad-bc}{\overline{a}c-a\overline{c}}\right| > 0.$$

Since γ and λ are independent of x and since (13.4.6) is the equation of a circle, the proof is done. □

Theorem 13.4.5. A Möbius transformation takes circles into circles.

Proof. Let Γ be any circle in $\mathbb{C}_\infty$ and let S be any Möbius transformation. Let z_2, z_3, z_4 be three distinct points on Γ and put $w_j = S(z_j)$ for $j = 2, 3, 4$. Then w_2, w_3, w_4 determine a circle Γ'. We claim that $S(\Gamma) = \Gamma'$. In fact, for any z in $\mathbb{C}_\infty$,

$$(z, z_2, z_3, z_4) = (S(z), w_2, w_3, w_4) \tag{13.4.7}$$

by theorem 13.4.2. By the preceding theorem, if z is on Γ, then both sides of (13.4.7) are real. But this says that $S(z) \in \Gamma'$. □

Now, let Γ and Γ' be two circles in $\mathbb{C}_\infty$ and let $z_2, z_3, z_4 \in \Gamma$; $w_2, w_3, w_4 \in \Gamma'$. Put $R(z) = (z, z_2, z_3, z_4)$, $S(z) = (z, w_2, w_3, w_4)$. Then $T = S^{-1} \circ R$ maps Γ onto Γ'. In fact, $T(z_j) = w_j$ for $j = 2, 3, 4$ and, as in the above proof, it follows that $T(\Gamma) = \Gamma'$.

Theorem 13.4.6. For any given circles Γ and Γ' in $\mathbb{C}_\infty$, there is a Möbius transformation T such that $T(\Gamma) = \Gamma'$. Furthermore we can specify that T takes any three points in Γ onto any three points on Γ'. If we specify $T(z_j)$ for $j = 2, 3, 4$ (distinct z_j in Γ) then T is unique.

Now that we know that a Möbius map takes circles to circles, the next question is: What happens to the inside and the outside of these circles? To answer this we introduce some new concepts.

Definition 13.4.2. Let Γ be a circle through points z_2, z_3, z_4. The points z, z^* in $\mathbb{C}_\infty$ are said to be symmetric with respect to Γ is

$$(z^*, z_2, z_3, z_4) = \overline{(z, z_2, z_3, z_4)}. \tag{13.4.8}$$

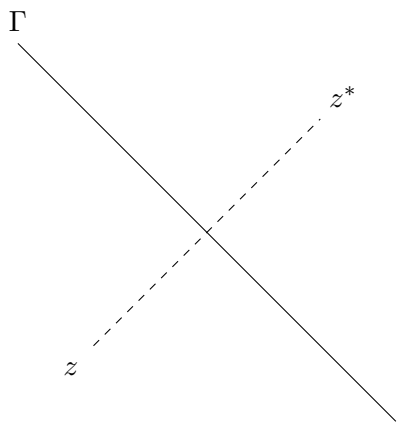


Figure 13.5

As it stands, this definition not only depends on the circle but also on the points z_2, z_3, z_4.

Also by theorem 13.4.4, z is symmetric to itself with respect to Γ if and only if $z \in \Gamma$. Let us investigate what it means for z and z^* to be symmetric. If Γ is a straight line then our linguistic prejudices lead us to believe that z and z^* are symmetric with respect to Γ if the line through z and z^* are the same distance from Γ but on the opposite sides of it (see fig. 13.5.

If Γ is a straight line then, choosing $z_4 = \infty$, (13.4.8) becomes

$$\frac{z^* - z_3}{z_2 - z_3} = \frac{\overline{z} - \overline{z}_3}{\overline{z}_2 - \overline{z}_3}.$$

This gives $|z^* - z_3| = |z - z_3$. Since z_3 was not specified, we have that z and z^* are equidistant from each point on Γ. Also,

$$\mathrm{Im}\frac{z^* - z_3}{z_2 - z_3} = \mathrm{Im}\frac{\overline{z} - \overline{z}_3}{\overline{z}_2 - \overline{z}_3} = -\mathrm{Im}\frac{z - z_3}{z_2 - z_3}.$$

Hence, we have (unless $z \in \Gamma$) that z and z^* in different half planes determined by Γ. It now follows that $[z, z^*]$ is perpendicular to Γ.

Now, suppose that $\Gamma = \{z : \ |z - a| = R\}$ $(0 < R < \infty)$. Let z_2, z_3, z_4 be points in Γ. Using (13.4.8) and theorem 13.4.2 for a number of M*ö*bius transformations gives

$$\begin{aligned}
(z^*, z_2, z_3, z_4) &= \overline{(z, z_2, z_3, z_4)} \\
&= \overline{(z - a, z_2 - a, z_3 - a, z_4 - a)} \\
&= \left(\overline{z} - \overline{a}, \frac{R^2}{z_2 - a}, \frac{R^2}{z_3 - a}, \frac{R^2}{z_4 - a}\right) \\
&= \left(\frac{R^2}{\overline{z} - \overline{a}}, z_2 - a, z_3 - a, z_4 - a\right) \\
&= \left(\frac{R^2}{\overline{z} - \overline{a}} + a, z_2, z_3, z_4\right).
\end{aligned}$$

Hence $z^* = a + R^2(\overline{z} - \overline{a})^{-1}$ or $(z^* - z)(\overline{z} - \overline{a}) = R^2$. From this it follows that

$$\frac{z^* - a}{z - a} = \frac{R^2}{|z - a|^2} > 0,$$

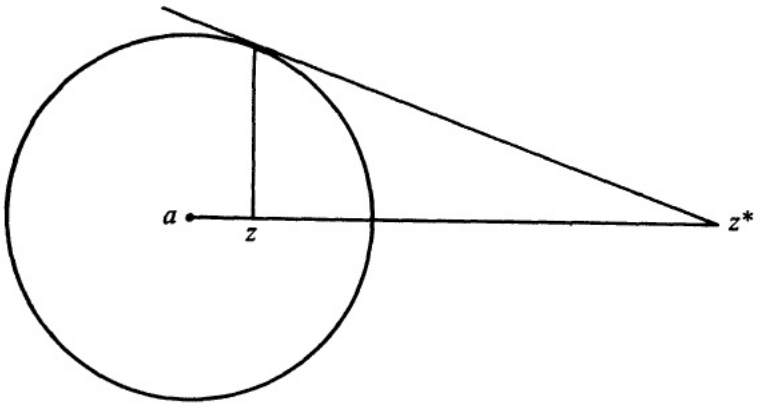


Figure 13.6

so that z^* lies on the ray $\{a+t(z-a) : 0 < t < \infty\}$ from a through z. Using the fact that $|z-a||z^*-a| = R^2$, we obtain z^* from z (if z lies inside Γ) as in the figure 13.6. That is, let L be the ray from a through z. Construct a line P perpendicular to L at z and at the point where P intersects Γ; construct the tangent to Γ. The point of intersection of this tangent with L is the point z^*. Thus, the points a and ∞ are symmetric with respect to Γ.

Theorem 13.4.7. (Symmetry Principle) If a Möbius transformation T takes a circle Γ_1 onto the circle Γ_2, then any pair of points symmetric with respect to Γ_1 are mapped by T onto a pair of points symmetric with respect to Γ_2.

13.5 Few Probable Questions

1. Define conformal maps. Show that a map f, analytic at z_0 with $f'(z_0) \neq 0$ is one-to-one in a neighbourhood of z_0.
2. Show that a one-to-one analytic function in a domain is conformal there.
3. If a function f is analytic in a simply connected domain D and on its boundary C (which is a simple closed contour), then f one-to-one on C implies it is so in D.
4. Define conformally equivalent regions. Show that the upper half disc $\{z : |z| < 1, \text{ Im } z > 0\}$ is conformally equivalent to the first quadrant $\{w = u + iv : u > 0, v > 0\}$.
5. Define cross ratio of z_1, z_2, z_3, z_4. For $z_1, z_2, z_3, z_4 \in \mathbb{C}_\infty$, show that the cross ratio is a real number if and only if all the four points lie on a circle.
6. Show that a Möbius transformation takes circles into circles. When are two points said to be symmetric with respect to a circle Γ?

Unit 14

Course Structure

- Schwarz principle of symmetry
- Schwarz-Christoffel formula (statement only)
- Applications of Schwarz-Christoffel formula.

14.1 Introduction

In mathematics, the Schwarz reflection principle, or the Schwarz principle of symmetry, is a way to extend the domain of definition of a complex analytic function, i.e., it is a form of analytic continuation. It states that if an analytic function is defined on the upper half-plane, and has well-defined (non-singular) real values on the real axis, then it can be extended to the conjugate function on the lower half-plane as we shall see. This unit is also dedicated to a preliminary study of the Schwarz-Christoffel mapping which is mainly a conformal transformation of the upper half-plane onto the interior of a simple polygon. SchwarzChristoffel mappings are used in potential theory and some of its applications, including minimal surfaces and fluid dynamics. They are named after Elwin Bruno Christoffel and Hermann Amandus Schwarz.

Objectives

After reading this unit, you will be able to

- define symmetric open set and deduce the symmetry principle
- deduce the Schwarz principle of symmetry
- have some preliminary idea about the Schwarz Christoffel mappings

14.2 Schwarz Principle of Symmetry

In real analysis, there are various situations where one wishes to extend a function from a given set to a larger one. Several techniques exist that provide extensions for continuous functions, and more generally for functions with varying degrees of smoothness. Of course, the difficulty of the technique increases as we impose

more conditions on the extension. The situation is very different for holomorphic functions. Not only are these functions indefinitely differentiable in their domain of definition, but they also have additional characteristically rigid properties, which make them difficult to mould. For example, there exist holomorphic functions in a disc which are continuous on the closure of the disc, but which cannot be continued (analytically) into any region larger than the disc.

Let Ω be an open subset of $\mathbb{C}$ that is symmetric with respect to the real line, that is

$$z \in \Omega \text{ if and only if } \overline{z} \in \Omega.$$

Let Ω^+ denote a part of Ω that lies in the upper half-plane and Ω^- that part that lies in the lower half-plane (see fig.14.1 for illustration).

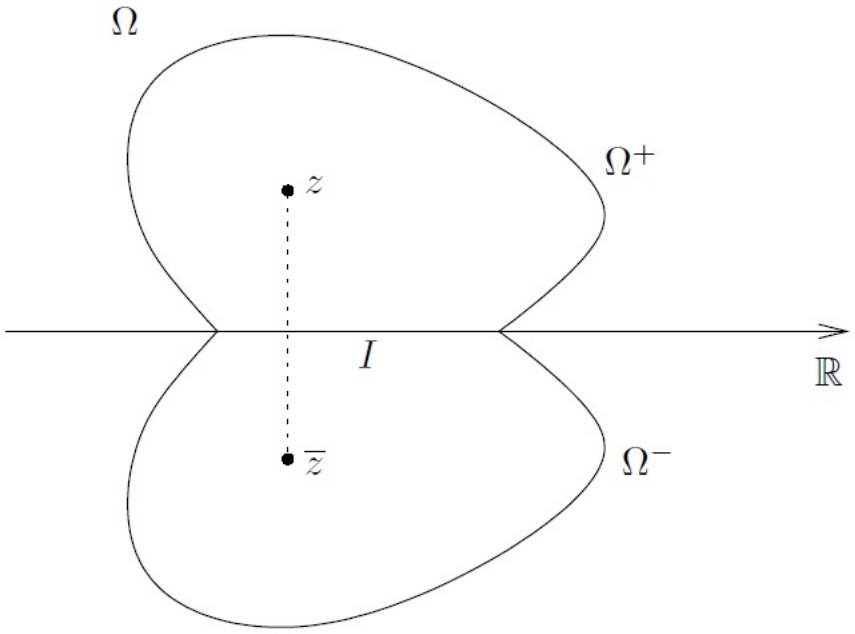


Figure 14.1

Also, let $I = \Omega \cap \mathbb{R}$ so that I denotes the interior of that part of the boundary of Ω^+ and Ω^- that lies on the real axis. Then we have

$$\Omega^+ \cup I \cup \Omega^- = \Omega$$

and the only interesting case of the next theorem occurs, of course, when I is non-empty.

Theorem 14.2.1. (Symmetry principle) If f^+ and f^- are analytic in Ω^+ and Ω^- respectively, that extend continuously to I and

$$f^+(x) = f^-(x), \ \forall\, x \in I,$$

then the function f defined on Ω by

$$\begin{aligned} f(z) &= f^+(z) \quad \text{if } z \in \Omega^+ \\ &= f^+(z) = f^-(z) \quad \text{if } z \in I \\ &= f^-(z) \quad \text{if } z \in \Omega^- \end{aligned}$$

is analytic on all of Ω.

Proof. One notes first that f is continuous throughout Ω. The only difficulty is to prove that f is analytic at points of I. Suppose D is a disc centred at a point on I and entirely contained in Ω. We prove that f is analytic in D by Moreras theorem. Suppose T is a triangle in D. If T does not intersect I, then

$$\int_T f(z)dz = 0$$

since f is analytic in the upper and lower half-discs. Suppose now that one side or vertex of T is contained in I, and the rest of T is in, say, the upper half-disc. If T_ϵ is the triangle obtained from T by slightly raising the edge or vertex which lies on I, we have $\int_{T_\epsilon} f = 0$ since T_ϵ is entirely contained in the upper half-disc an (illustration of the case when an edge lies on I is given in Figure 14.2). When we let $\epsilon \to 0$, and by continuity, we conclude that

$$\int_T f(z)dz = 0$$

If the interior of T intersects I, we can reduce the situation to the previous one by writing T as the union of

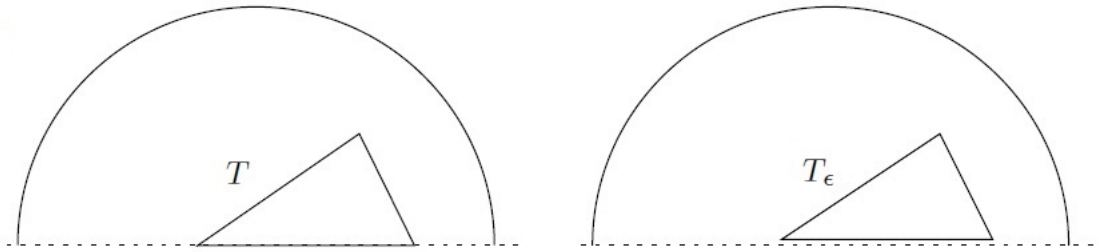


Figure 14.2: Raising a vertex

triangles each of which has an edge or vertex on I as shown in Figure 14.3. By Moreras theorem we conclude that f is analytic in D, as was to be shown. □

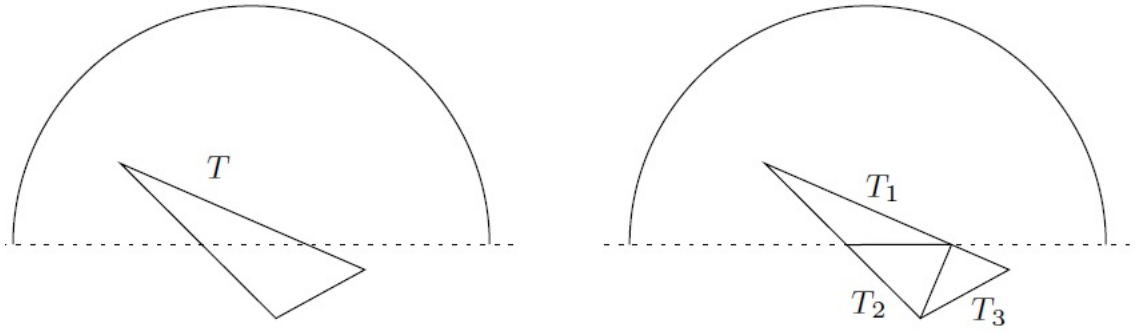


Figure 14.3: Splitting a triangle

We can now state the extension principle, where we use the above notation.

Theorem 14.2.2. (Schwarz reflection principle) Suppose that f is a analytic function in Ω^+ that extends continuously to I and such that f is real-valued on I. Then there exists a function F analytic in all of Ω such that $F = f$ on Ω^+.

Proof. The idea is simply to define $F(z)$ for $z \in \Omega^-$ by

$$F(z) = \overline{f(\overline{z})}.$$

To prove that F is analytic in Ω^- we note that if $z, z_0 \in \Omega^-$, then $\overline{z}, \overline{z}_0 \in \Omega^+$ and hence, the power series expansion of f near $\overline{z}_0$ gives

$$f(\overline{z}) = \sum a_n(\overline{z} - \overline{z}_0)^n.$$

As a consequence we see that

$$F(z) = \sum \overline{a}_n(z - z_0)^n$$

and F is analytic in Ω^-. Since f is real valued on I we have, $\overline{f(x)} = f(x)$, whenever $x \in I$ and hence F extends continuously up to I. The proof is complete once we invoke the symmetry principle. □

14.3 Schwarz Christoffel formula

We represent the unit vector which is tangent to a smooth are C at a point z_0 by the complex number t, and we let the number τ denote the unit vector tangent to the image Γ of C at the corresponding point w_0 under a transformation $w = f(z)$. We assume that f is analytic at z_0 and that $f'(z_0) \neq 0$. We know,

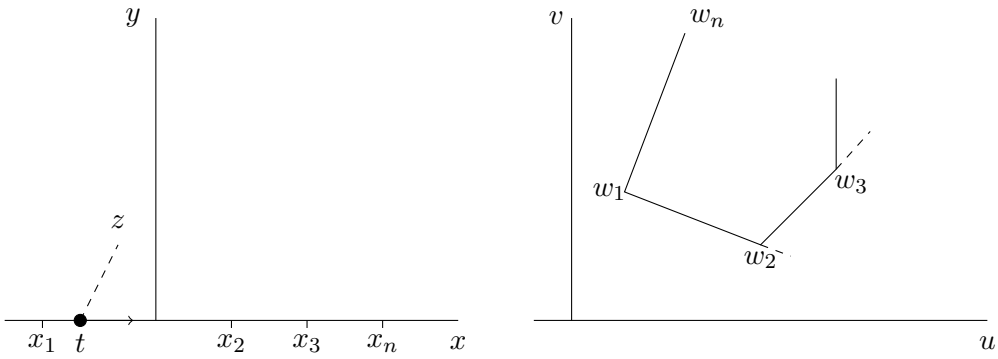
$$\arg \tau = \arg f'(z_0) + \arg t \tag{14.3.1}$$

In particular, if C is a segment of the x-axis with positive sense to the right then $t = 1$ and $\arg t = 0$ at each point $z_0 = x$ on C. In that case, equation (14.3.1) becomes

$$\arg \tau = \arg f'(x) \tag{14.3.2}$$

If $f'(z)$ has a constant argument along that segment, it follows that $\arg \tau$ is constant. Hence, the image Γ of C is also a segment of a straight line.

Let us now construct a transformation $w = f(z)$ that maps the whole x-axis onto a polygon of n sides, where $x_1, x_2, \ldots, x_{n-1}$ and ∞ are the points on that axis whose images are to be the vertices of the polygon and where $x_1 < x_2 < \cdots < x_{n-1}$. The vertices are the n points $w_j = f(x_j)(j = 1, 2, \ldots n-1)$ and $w_n = f(\infty)$. The function f should be such that $\arg f'(z)$ jumps from one constant value to another at the points $z = x_j$ as the point z traces out the x-axis. If the function f is chosen such that



$$f'(z) = A(z - x_1)^{k_1}(z - x_2)^{k_2} \cdots (z - x_{n-1})^{k_{n-1}}. \tag{14.3.3}$$

where A is a complex constant and each k_j is a real constant, then the argument of $f'(z)$ changes in the prescribed manner as z describes the real axis. This is seen writing the argument of the derivative (14.3.3) as

$$\arg f'(z) = \arg A - k_1 \arg(z - x_1) - k_2 \arg(z - x_2) - \cdots - k_{n-1} \arg(z - x_{n-1}) \tag{14.3.4}$$

When $z = x$ and $x < x_1$,

$$\arg(z - x_1) = \arg(z - x_2) = \cdots = \arg(z - x_{n-1}) = \pi$$

When $x_1 < x < x_2$, the argument $\arg(z - x_1)$ is 0 and each of the other arguments is π. According to equation (14.3.4), then $\arg f'(z)$ increases abruptly by the angle $k_1\pi$ as z moves to the right through the point $z = x_1$. It again jumps in value, by the amount $k_2\pi$, as z passes through the point x_2, etc.

In view of (14.3.2), the unit vector τ is constant in direction as z moves from x_{j-1} to x_j; the point w thus moves in that fixed direction along a straight line. The direction of τ changes abruptly by the angle $k_j\pi$ at the image point w_j of x_j. Those angles $k_j\pi$ are the exterior angles of the polygon described by the point w.

The exterior angles can be limited to angles between $-\pi$ to π, in which case $-1 < k_j < 1$. We assume that the sides of the polygon never cross one another and that the polygon is given is given a positive or counterclockwise orientation. The sum of the exterior angles of a $closed$ polygon is, then 2π and the exterior angle at the vertex w_n which is the image of the point $z = \infty$, can be written

$$k_n\pi = 2\pi - (k_1 + k_2 + \cdots + k_{n-1})\pi$$

Thus the numbers k_j must necessarily satisfy the conditions

$$k_1 + k_2 + \cdots + k_{n-1} + k_n = 2, \qquad -1 < k_j < 1 \quad (j = 1, 2, \ldots, n) \tag{14.3.5}$$

Note that $k_n = 0$ if $k_1 + k_2 + \cdots + k_{n-1} = 2$. This means that the direction of τ does not change at the point w_n. So, w_n is not a vertex, and the polygon has $n - 1$ sides.

14.4 Few Probable Questions

1. State and prove the symmetry principle.
2. State and prove the Schwarz principle of symmetry.

Unit 15

Course Structure

- Univalent functions, general theorems

15.1 Introduction

In mathematics, in the branch of complex analysis, an analytic function on an open subset of the complex plane is called univalent if it is injective. The theory of univalent functions is an old subject, born around the turn of the century, yet it remains an active field of research. This unit introduces the class S of univalent functions and some of its subclasses defined by geometric conditions. A number of basic questions are answered by elementary methods. Most of the results concerning the class S are direct consequences of the area theorem, which may be regarded as the cornerstone of the entire subject.

Objectives

After reading this unit, you will be able to

- define univalent functions
- define normal families of analytic functions and related terms
- learn about preliminary types of univalent functions

15.2 Normal Families

A family $\mathscr{F}$ of functions analytic in a domain D is called a **normal family** if every sequence of functions $f_n \in \mathscr{F}$ has a subsequence which converges uniformly on each compact subset of D.

A family $\mathscr{F}$ is compact if whenever $f_n \in \mathscr{F}$ and $f_n(z) \to f(z)$ uniformly on compact subsets of D, it follows that $f \in \mathscr{F}$. The defining property of a normal family is analogous to the Bolzano-Weierstrass property of a bounded set of points in Euclidean space. Compact families are analogous to closed sets.

A family $\mathscr{F}$ of functions analytic in D is said to be **locally bounded** is the functions are **uniformly bounded** on each closed disc $B \subset D$; that is, if $|f(z)| \leq M$ for all $z \in B$ and for every $f \in \mathscr{F}$, where the bound M depends only on B. In view of the Heine Borel theorem, it then follows that the functions are

uniformly bounded on each compact subset of D. If $\mathscr{F}$ is a locally bounded family of analytic functions, then by the Cauchy integral formula, the family of derivatives $\{f' : f \in \mathscr{F}\}$ is also locally bounded.

We have the following theorem concerning locally bounded family of analytic functions.

Theorem 15.2.1. A necessary and sufficient condition for a family of analytic functions to be locally bounded is that, it is normal.

15.3 Univalent Functions

Definition 15.3.1. A single valued function f is said to be univalent (or schlicht) in a domain $D \subset \mathbb{C}$ if it never takes the same value twice; that is, if $f(z_1) \neq f(z_2)$ for all points z_1 and z_2 in D with $z_1 \neq z_2$. The function f is said to be locally univalent at a point $z_0 \in D$ if it is univalent in some neighbourhood of z_0.

For analytic functions f, the condition $f'(z_0) \neq 0$ is equivalent to local univalence at z_0 as we have seen previously. An analytic univalent function is a conformal mapping because of its angle-preserving property.

We shall be concerned primarily with the class S of functions f analytic and univalent in the unit disc $\mathbb{D}$, satisfied by the conditions $f(0) = 0$ and $f'(0) = 1$. Thus each $f \in S$ has a Taylor series expansion of the form

$$f(z) = z + a_2 z^2 + a_3 z^3 + \cdots, \quad |z| < 1.$$

The Riemann mapping theorem states that for any simply connected domain D, which is a proper subset of the complex plane and any point $\zeta \in D$, there is a unique function f which maps D conformally onto the unit disc and has properties $f(\zeta) = 0$ and $f'(\zeta) > 0$. That is, it says that any simply connected domain D, which is a proper subset of $\mathbb{C}$, is conformally equivalent to the unit disc $\mathbb{D}$.

In view of the Riemann mapping theorem, most of the geometric theorems concerning functions of class S are readily translated to statements about univalent functions in arbitrary simply connected domains with more than one boundary point. The leading example of a function of class S is the *Koebe* function

$$k(z) = z(1-z)^{-2} = z + 2z^2 + 3z^3 + \cdots$$

The Koebe function maps the disc $\mathbb{D}$ onto the entire plane minus the part of the negative real axis from $-1/4$ to infinity. This is best seen by writing

$$k(z) = \frac{1}{4}\left(\frac{1+z}{1-z}\right)^2 - \frac{1}{4}$$

and observing that the function

$$w = \frac{1+z}{1-z}$$

maps $\mathbb{D}$ conformally onto the right half-plane $\mathrm{Re}\{w\} > 0$.

Other simple examples of functions in S are

1. $f(z) = z$, the identity mapping;
2. $f(z) = z(1-z)^{-1}$, which maps $\mathbb{D}$ conformally onto the half plane $\mathrm{Re}\{w\} > -1/2$;
3. $f(z) = z(1-z^2)^{-1}$, which maps $\mathbb{D}$ onto the entire plane minus the two half lines $\frac{1}{2} \leq x < \infty$ and $-\infty < x \leq -\frac{1}{2}$;
4. $f(z) = z - \frac{1}{2}z^2 = \frac{1}{2}[1 - (1-z)^2]$, which maps $\mathbb{D}$ onto the interior of a cardioid.

The sum of two functions in S need not be univalent. For example, the sum of $z(1-z)^{-1}$ and $z(1+iz)^{-1}$ has a derivative which vanishes at $\frac{1}{2}(1+i)$ (verify!). However, the class S is preserved under a number of elementary transformations.

1. **Conjugation:** If $f \in S$ and
$$g(z) = \overline{f(\overline{z})} = z + \overline{a}_2 z^2 + \overline{a}_3 z^3 + \cdots,$$
then $g \in S$.

2. **Rotation:** If $f \in S$ and
$$g(z) = \mathrm{e}^{-i\theta} f(\mathrm{e}^{-i\theta} z),$$
then $g \in S$.

3. **Dilation:** If $f \in S$ and
$$g(z) = \frac{1}{r} f(rz), \quad \text{where } 0 < r < 1,$$
then $g \in S$.

4. **Range Transformation:** If $f \in S$ and ψ is a function analytic and univalent on the range of f, with $\psi(0) = 0$ and $\psi'(0) = 1$, then $g = \psi \circ f \in S$.

5. **Omitted-value transformation:** If $f \in S$ and $f(z) \neq \omega$, then
$$g = \frac{\omega f}{\omega - f} \in S.$$

6. **Square-root transformation:** If $f \in S$ and $g(z) = \sqrt{f(z^2)}$, the $g \in S$.

The square root transformation requires a word of explanation. Since $f(z) = 0$ only at the origin, a single valued branch of the square root may be chosen by writing

$$\begin{aligned} g(z) &= \sqrt{f(z^2)} = z\{1 + a_2 z^2 + a_3 z^4 + \cdots\}^{\frac{1}{2}} \\ &= z + c_3 z^3 + c_5 z^5 + \cdots, \quad |z| < 1. \end{aligned}$$

Note that g is an odd analytic function, so that $g(-z) = -g(z)$. If $g(z_1) = g(z_2)$, then $f(z_1^2) = f(z_2^2)$ and $z_1^2 = z_2^2$, which gives $z_1 = \pm z_2$. But, if $z_1 = -z_2$, then $g(z_1) = g(z_2) = -g(z_1)$. Thus $g(z_1) = 0$, and $z_1 = 0$. This shows that $z_1 = z_2$ in either case, proving that g is univalent.

Closely related to S, is the class Σ of functions

$$g(z) = z + b_0 + b_1 z^{-1} + b_2 z^{-2} + \cdots$$

is analytic and univalent in the domain $\mathbb{E} = \{z : |z| > 1\}$, exterior to the domain $\mathbb{D}$, except for a simple pole at infinity with residue 1. Each function $g \in \Sigma$ maps $\mathbb{E}$ onto the complement of a compact connected set E. It is useful to consider the subclass Σ' of functions $g \in \Sigma$ for which $0 \in E$; that is, for which $g(z) \neq 0$ in $\mathbb{E}$. Any function $g \in \Sigma$ will belong to Σ' after suitable adjustment of the constant term b_0. Such an adjustment will only translate the range of g and will not destroy the univalence.

For each $f \in S$, the function

$$g(z) = \left\{ f\left(\frac{1}{z}\right) \right\}^{-1} = z - a_2 + (a_2^2 - a_3) z^{-1} + \cdots$$

belongs to Σ'. This transformation is called an inversion. It actually establishes a one-to-one correspondence between S and Σ'. The class Σ' is preserved under the square-root transformation

$$G(z) = \sqrt{g(z^2)} = z\{1 + b_0 z^{-2} + b_1 z^{-4} + \cdots\}^{\frac{1}{2}}.$$

It is important to observe that this operation cannot be applied to every function $g \in \Sigma$, but is permissible only if $g \in \Sigma'$, because the square root will introduce a branch point wherever $g(z^2) = 0$.

Sometimes, it is convenient to consider the subclass Σ_0 consisting of all $g \in \Sigma$ with $b_0 = 0$. Obviously this can be achieved by suitable translation, but it may not be possible to translate a given function $g \in \Sigma$ simultaneously to both Σ_0 and Σ'.

It is also useful to distinguish the subclass $\tilde{\Sigma}$ of all functions $g \in \Sigma$ whose omitted set E has two dimensional Lebesgue measure zero. The functions $g \in \tilde{\Sigma}$ will be called full mappings.

15.4 Few Probable Questions

1. Define normal family.
2. Define locally bounded family of analytic functions in a domain D.
3. Define univalent function on a domain D. Show that for an analytic function f on D, $f'(z_0) \neq 0$ at $z_0 \in D$ is equivalent to the local univalence of f at z_0.

Unit 16

Course Structure

- Area theorem
- Growth and Distortion theorems

16.1 Introduction

The univalence of a function

$$g(z) = z + b_0 + \sum_{n=1}^{\infty} b_n z^{-1}, \qquad |z| > 1$$

places strong restriction on the size of the Laurent coefficients b_n, $n = 1, 2, \ldots$. This is expressed by the area theorem, which is fundamental to the theory of univalent functions. The reason for the name will be apparent from the proof. Gronwall discovered the theorem in 1914.

Objectives

After reading this unit, you will be able to

- deduce the area theorem and related results
- deduce the growth and distortion theorems

16.2 Area Theorem

Theorem 16.2.1. Area Theorem: If $g \in \sum$, then

$$\sum_{n=1}^{\infty} n|b_n|^2 \leq 1$$

with equality if and only if $g \in \tilde{\sum}$.

Proof. Let E be the set omitted by g. For $r > 1$, let C_r be the image under g of the circle $|z| = r$. Since g is univalent, C_r is a simple closed curve which encloses a domain $E_r \supset E$. By Green's theorem, the area of E_r is

$$\begin{aligned} A_r &= \frac{1}{2i}\int_{C_r} \overline{w}dw = \frac{1}{2i}\int_{|z|=r} \overline{g(z)}g'(z)dz \\ &= \frac{1}{2}\int_0^{2\pi} \{r\,\mathrm{e}^{-i\theta} + \sum_{n=0}^{\infty} \overline{b_n} r^{-n}\,\mathrm{e}^{in\theta}\} \times \{1 - \sum_{v=1}^{\infty} v b_v r^{-v-1}\,\mathrm{e}^{-i(v+1)\theta}\} r\,\mathrm{e}^{i\theta}\,d\theta \\ &= \pi\{r^2 - \sum_{n=1}^{\infty} n|b_n|^2 r^{-2n}\}, \qquad r > 1 \end{aligned} \tag{16.2.1}$$

Let r decrease to 1, we obtain

$$m(E) = \pi\{r^2 - \sum_{n=1}^{\infty} n|b_n|^2\}$$

where $m(E)$ is the outer measure of E. Since, $m(E) \geq 0$, this proves the theorem. □

An immediate corollary is the inequality $|b_n| \leq n^{-1/2}$, $n = 1, 2, \ldots$. This inequality is not sharp if $n \geq 2$, since the function

$$g(z) = z + n^{-1/2} z^{-n}$$

is not univalent. Indeed, its derivative

$$g'(z) = 1 - n^{1/2} z^{-n-1}$$

vanishes at certain points in $\mathbb{E}$ if $n \geq 2$. However, the inequality $|b_1| \leq 1$ is sharp and has important consequences.

Corollary 16.2.1. If $g \in \sum$, then $|b_1| \leq 1$, with equality if and only if g has the form

$$g(z) = z + b_0 + \frac{b_1}{z}, \qquad |b_1| = 1$$

This is a conformal mapping of $\mathbb{E}$ onto the complement of a line segment of length 4.

From this result it is a short step to a theorem of Bieberbach estimating the second coefficient a_2 of a function of class S. This theorem was given in 1916 and was the main basis for the famous $Bieberbach\ conjecture$.

Theorem 16.2.2. (Bieberbach's Theorem). If $f \in S$, then $|a_2| \leq 2$, with equality if and only if f is a rotation of the Koebe function.

Proof. A square-root transformation and an inversion applied to $f \in S$ will produce a function

$$g(z) = \{f(1/z)\}^{-1/2} = z - (a_2/2)z^{-1} + \cdots$$

of class $\sum$. Thus $|a_2| \leq 2$, by the corollary to the area theorem. Equality occurs if and only if g has the form

$$g(z) = z - \mathrm{e}^{i\theta}/z$$

A simple calculation shows that this is equivalent to

$$f(\zeta) = \zeta(1 - \mathrm{e}^{i\theta}\zeta)^{-2} = \mathrm{e}^{-i\theta} k(\mathrm{e}^{i\theta}\zeta),$$

a rotation of the Koebe function. □

As a first application of Bieberbach's theorem, we shall now prove a famous covering theorem due to Koebe. Each function $f \in S$ is an open mapping with $f(0) = 0$, so its range contains some disk centered at the origin. As early as 1907, Koebe discovered that the ranges of all functions in S contain a common disk $|w| < \rho$, where ρ is an absolute constant. The Koebe function shows that $\rho \leq \frac{1}{4}$, and Bieberbach later established Koebe's conjecture that ρ may be taken to be $\frac{1}{4}$.

Theorem 16.2.3. (Koebe One-Quarter Theorem): The range of every function of class S contains the disk $\{w : |w| < \frac{1}{4}\}$.

Proof. If a function $f \in S$ omits the value $\omega \in \mathbb{C}$, then

$$g(z) = \frac{\omega f(z)}{\omega - f(z)} = z + \left(a_2 + \frac{1}{\omega}\right) z^2 + \cdots$$

is analytic and univalent in $\mathbb{D}$. This is the omitted-value transformation, which is the composition of f with a linear fractional mapping. Since, $g \in S$, Bieberbach's theorem gives

$$\left|a_2 + \frac{1}{\omega}\right| \leq 2$$

Combined with the inequality $|a_2| \leq 2$ this shows that $|1/\omega| \leq 4$, or $|\omega| \geq \frac{1}{4}$. Thus every omitted value must lie outside the disk $|w| < \frac{1}{4}$. □

This proof actually shows that the Koebe function and its rotations are the only functions in S which omit a value of modulus $\frac{1}{4}$. Thus the range of every other function in S covers a disk of larger radius.

It should be observed that $univalence$ is the key to Koebe's theorem. For example, the analytic functions

$$f_n(z) = \frac{1}{n}(\mathrm{e}^{nz} - 1), \qquad n = 1, 2, \ldots,$$

have the properties $f_n(0) = 0$ and $f_n'(0) = 1$, yet f_n omits the value $-1/n$, which may be chosen arbitrarily close to the origin.

16.3 Growth and Distortion Theorems

Bieberbach's inequality $|a_2| \leq 2$ has further implications in the geometric theory of conformal mapping. One important consequence is the $Koebe\ distortion\ theorem$, which provides sharp upper and lower bounds for $|f'(z)|$ as f ranges over the class S, The term "distortion" arises from the geometric interpretation $|f'(z)|$ as the infinitesimal magnification factor of arclength under the mapping f, or from that of the Jacobian $|f'(z)|^2$ as the infinitesimal magnification factor of area. The following theorem gives a basic estimate which leads to the distortion theorem and related results.

Theorem 16.3.1. For each $f \in S$,

$$\left|\frac{zf''(z)}{f'(z)} - \frac{2r^2}{1 - r^2}\right| \leq \frac{4r}{1 - r^2}, \qquad |z| = r < 1 \tag{16.3.1}$$

Proof. Given $f \in S$, fix $\zeta \in \mathbb{D}$ and perform a disk automorphism to construct

$$F(z) = \frac{f\left(\frac{z+\zeta}{1+\bar{\zeta}z}\right) - f(\zeta}{(1 - |\zeta|^2)f'(\zeta)} = z + A_2(\zeta)z^2 + \cdots. \tag{16.3.2}$$

Then $F \in S$ and a calculation gives

$$A_2(\zeta) = \frac{1}{2}\{(1-|\zeta|^2)\frac{f''(\zeta)}{f'(\zeta)} - 2\zeta\}.$$

But by Bieberbach's theorem, $|A_2(\zeta)| \leq 2$. Simplifying this inequality and replacing ζ by z, we obtain the inequality (16.3.1). A suitable rotation of the Koebe function shows that the estimate is sharp for each $z \in \mathbb{D}$. □

Theorem 16.3.2. (Distortion Theorem). For each $f \in S$

$$\frac{1-r}{(1+r)^3} \leq |f'(z)| \leq \frac{1+r}{(1-r)^3}, \qquad |z| = r < 1 \tag{16.3.3}$$

For each $z \in \mathbb{D}$, $z \neq 0$, equality occurs if and only if f is a suitable rotation of the Koebe function.

Proof. Since, an inequality $|\alpha| \leq c$ implies $-c \leq Re\{\alpha\} \leq c$, it follows from (16.3.1) that

$$\frac{2r^2-4r}{1-r^2} \leq Re\{\frac{zf''(z)}{f'(z)}\} \leq \frac{2r^2+4r}{1-r^2}$$

Because $f'(z) \neq 0$ and $f'(0) = 1$, we can choose a single-valued branch of $\log f'(z)$ which vanishes at the origin. Now, observe that

$$Re\{\frac{zf''(z)}{f'(z)}\} = r\frac{\partial}{\partial r}Re\{\log f'(z)\}, \qquad z = \mathrm{e}^{i\theta}.$$

Hence,

$$\frac{2r^2-4r}{1-r^2} \leq \frac{\partial}{\partial r}|f'(r\,\mathrm{e}^{r\theta})| \leq \frac{2r^2+4r}{1-r^2} \tag{16.3.4}$$

Holding θ fixed, integrate with respect to r from 0 to R. A calculation then produces the inequality

$$\log\frac{1-R}{(1+R)^3} \leq \log|f'(R\,\mathrm{e}^{i\theta})| \leq \log\frac{1+R}{(1-R)^3},$$

ad the distortion theorem follows by exponentiation.

A suitable rotation of the Koebe function, whose derivative is

$$k'(z) = \frac{1+z}{(1-z)^3}, \tag{16.3.5}$$

shows that both estimates of $|f'(z)|$ are best possible. Furthermore, whenever equality occurs for $z = R\,\mathrm{e}^{i\theta}$ in either the upper or the lower estimate of (16.3.3), the equality must hold in the corresponding part of (16.3.4) for all r, $0 \leq r \leq R$. In particular,

$$Re\{\mathrm{e}^{i\theta}\,\frac{f''(0)}{f'(0)}\} = \pm 4,$$

which implies that $|a_2| = 2$. Hence by Bieberbach's theorem, f must be a rotation of the Koebe function. □

The distortion theorem will now be applied to obtain the sharp upper and lower bounds for $|f(z)|$. This result is as follows.

Theorem 16.3.3. (Growth Theorem). For each $f \in S$,

$$\frac{r}{(1+r)^2} \leq |f(z)| \leq \frac{r}{(1-r)^2}, \qquad |z| = r < 1. \tag{16.3.6}$$

For each $z \in \mathbb{D}$, $z \neq 0$, equality occurs if and only if f is a suitable rotation of the Koebe function.

Proof. Let $f \in S$ and fix $z = \mathrm{e}^{i\theta}$ with $0 < r < 1$. Observe that

$$f(z) = \int_0^r f'(\rho\,\mathrm{e}^{i\theta})\,\mathrm{e}^{i\theta}\,d\rho,$$

since $f(0) = 0$. Thus by the distortion theorem,

$$|f(z)| \le \int_0^r |f'(\rho\,\mathrm{e}^{i\theta})|d\rho \le \int_0^r \frac{1+\rho}{(1-\rho)^3}d\rho = \frac{r}{(1-r)^2}.$$

The lower estimate is more subtle. It holds trivially if $|f(z)| \ge \frac{1}{4}$, since $r(1+r)^{-2} < \frac{1}{4}$ for $0 < r < 1$. If $|f(z)| < \frac{1}{4}$, the Koebe one-quarter theorem implies that the radial segment from 0 to $f(z)$ lies entirely in the range of f. Let C be the preimage of this segment. Then C is a simple arc from 0 to z, and

$$f(z) = \int_C f'(\zeta)d\zeta$$

But $f'(\zeta)d\zeta$ has constant signum along C, by construction, so the distortion theorem gives

$$|f(z)| = \int_C |f'(\zeta)||d\zeta| \ge \int_0^r \frac{1-\rho}{(1+\rho)^3}d\rho = \frac{r}{(1+r)^2}.$$

Equality in either part of (16.3.6) implies equality in the corresponding part of (16.3.3), which implies that f is a rotation of the Koebe function.

All of this information was obtained by passing to the real part in the basic inequality (16.3.1). Taking the imaginary part instead, one finds

$$-\frac{4r}{1-r^2} \le Im\{\frac{zf''(z)}{f'(z)}\} \le \frac{4r}{1-r^2}$$

$$-\frac{4r}{1-r^2} \le \frac{\partial}{\partial r}\arg f'(r\,\mathrm{e}^{i\theta}) \le \frac{4r}{1-r^2}$$

Radial integration now produces the inequality

$$|\arg f'(z)| \le 2\log\frac{1+r}{1-r}, \qquad f \in S \tag{16.3.7}$$

Here it is understood that $\arg f'(z)$ is the branch which vanishes at the origin. The quantity $\arg f'(z)$ can be interpreted geometrically as the local rotation factor under the conformal mapping f. For this reason the inequality (16.3.7) may be called a $rotation\ theorem$. Unfortunately, however, it is not sharp at any point $z \ne 0$ in the disk. The true rotation theorem

$$\begin{aligned} |\arg f'(z)| &\le 4\sin^{-1} r, \ \ r \le 1/\sqrt{2} \\ &\le \pi + \log\frac{r^2}{1-r^2}, \ \ r \ge 1/\sqrt{2}, \end{aligned}$$

lies much deeper. The splitting of the sharp bound at $r = 1/\sqrt{2}$ is one of the most remarkable phenomena in the univalent function theory. □

One further inequality, a combined growth and distortion theorem, is sometimes useful.

16.4 Few Probable Questions

1. State and prove the Area theorem.
2. State and prove Bieberbach's theorem.
3. State and Koebe One-Quarter theorem.
4. State and prove Distortion theorem.
5. State and prove Growth theorem.

References

1. E. T. Copson : An Introduction to the Theory of Functions of a Complex Variable.
2. E. C. Titthmarsh : The Theory of Functions.
3. A. I. Markushevich : Theory of Functions of a Complex Variable (Vol. I, II & III).
4. L. V. Ahlfors : Complex Analysis.
5. J. B. Conway : Functions of One Complex Variable.
6. A. I. Markushevich : The Theory of Analytic Functions, A Brief Course.
7. G. Valiron : Integral Functions.
8. C. Caratheodory : Theory of Functions of a Complex Variable.
9. R. P. Boas : Entire Functions.