

**POST GRADUATE DEGREE PROGRAMME (CBCS)
IN
MATHEMATICS**

**SEMESTER II
SELF LEARNING MATERIAL**

PAPER - MATCB 2.1

**History of Mathematics
Operation Research
Linear Algebra
Dynamical Systems**



**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, 2019

Directorate of Open and Distance Learning, University of Kalyani
Published by the Directorate of Open and Distance Learning,
University of Kalyani, Kalyani-741235, West Bengal

All rights reserved. No part of this work should be reproduced in any form without the permission in writing from the Directorate of Open and Distance Learning, University of Kalyani.

Director's Note

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

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

Self Learning Materials (SLMs) have been published by the Directorate of Open and Distance Learning, University of Kalyani, Kalyani-741235, West Bengal and all the copyright reserved for University of Kalyani. No part of this work should be reproduced in any form without permission in writing from the appropriate authority of the University of Kalyani.

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

Director
Directorate of Open and Distance Learning
University of Kalyani

CONTENTS

Sl. No.	Block	Unit	Page
1	History of Mathematics	1	2 – 9
		2	10 – 14
		3	15 – 17
		4	18 – 22
2	Operations Research	5	25 – 32
		6	33 - 36
		7	37 - 48
		8	49 - 70
3	Linear Algebra	9	72 – 78
		10	79 – 85
		11	86 – 95
		12	96 – 104
4	Dynamical Systems	13	106 – 117
		14	118 – 122
		15	123 - 132
		16	133 – 141

Choice Paper

MATCB 2.1

Block - I

Marks : 20 (SEE : 25; IA : 05)

History of Mathematics

Syllabus

Unit	Topic
1	Babylonian and Egyptian mathematics, Greek mathematics, Pythagoras, Euclid and the elements of geometry, Archimedes, Apollonius.
2	Development of Trigonometry, Development of Algebra, Development of Analytic Geometry,
3	Development of Calculus, Development of Selected Topics of Modern Mathematics
4	Development of Modern geometries, Modern algebra, Methods of real analysis.

Unit I

Course Structure

- Introduction
 - Babylonian and Egyptian mathematics
 - Greek mathematics
 - Pythagoras, Euclid and the elements of geometry
 - Archimedes, Apollonius
-

Introduction

The area of study known as the history of mathematics is primarily an investigation into the origin of discoveries in mathematics and, to a lesser extent, an investigation into the mathematical methods and notation of the past. Before the modern age and the worldwide spread of knowledge, written examples of new mathematical developments have come to light only in a few locales. From 3000 BC the Mesopotamian states of Sumer, Akkad and Assyria, together with Ancient Egypt and Ebla began using arithmetic, algebra and geometry for purposes of taxation, commerce, trade and also in the field of astronomy and to formulate calendars and record time.

Babylonian and Egyptian mathematics

Our first knowledge of mankind's use of mathematics comes from the Egyptians and Babylonians. Both civilizations developed mathematics that was similar in scope but different in particulars. There can be no denying the fact that the totality of their mathematics was profoundly elementary, but their astronomy of later times did achieve a level comparable to the Greeks.

Babylonian Mathematics

The mathematics developed and practised by the people of Mesopotamia from the days of the early Sumerians to the fall of Babylon in 539 BCE. The Babylonian Mathematics can be categorised into two: one of the Old Babylonian Period (1830-1531 BCE) and the other mainly Seleucid from the last three of four centuries BCE. The Babylonians were somewhat more advanced than the Egyptians in mathematics. The main features of the Babylonian Mathematics were:

- Their mathematical notation was positional but sexagesimal. From this we derive the modern day usage of 60 seconds in a minute, 60 minutes in an hour, and 360 degrees in a circle.

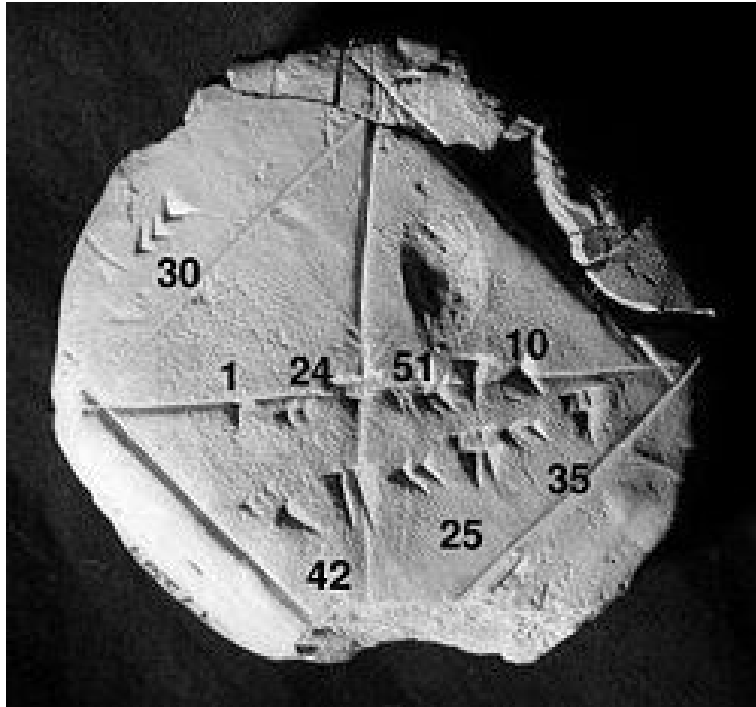


Figure 1: Babylonian clay tablet with annotations

- They used no zero.
- More general fractions, though not all fractions, were admitted.
- They could extract square roots and solve linear systems.
- The Pythagorean triples were frequently used.
- They solved cubic equations with the help of tables.
- They studied circular measurement.
- They studied circular measurement.

Most clay tablets that describe Babylonian mathematics belong to the Old Babylonian, which is why the mathematics of Mesopotamia is commonly known as Babylonian mathematics. Some clay tablets contain mathematical lists and tables, others contain problems and worked solutions.

For enumeration the Babylonians used symbols for 1, 10, 60, 600, 3,600, 36,000, and 216,000, similar to the earlier period. Below are some of the symbols. They did arithmetic in base 60, sexagesimal.

There is no clear reason why the Babylonians selected the sexagesimal system⁶. It was possibly selected in the interest of metrology, this according to Theon of Alexandria, a commentator of the fourth century A.D.: i.e. the values 2,3,5,10,12,15,20, and 30 all divide 60. Remnants still exist

1		11		21		31		41		51	
2		12		22		32		42		52	
3		13		23		33		43		53	
4		14		24		34		44		54	
5		15		25		35		45		55	
6		16		26		36		46		56	
7		17		27		37		47		57	
8		18		28		38		48		58	
9		19		29		39		49		59	
10		20		30		40		50		60	

Figure 2: Babylonian numerals

today with time and angular measurement. However, a number of theories have been posited for the Babylonians choosing the base of 60. For example

1. The number of days, 360, in a year gave rise to the subdivision of the circle into 360 degrees, and that the chord of one sixth of a circle is equal to the radius gave rise to a natural division of the circle into six equal parts. This in turn made 60 a natural unit of counting. (Moritz Cantor, 1880)
2. The Babylonians used a 12 hour clock, with 60 minute hours. That is, two of our minutes is one minute for the Babylonians. (Lehmann-Haupt, 1889) Moreover, the (Mesopotamian) zodiac was divided into twelve equal sectors of 30 degrees each.
3. The number 60 is the product of the number of planets (5 known at the time) by the number of months in the year, 12. (D. J. Boorstin, 6Recall, the very early use of the sexagesimal system in China. There may well be a connection. 7See Georges Ifrah, The Universal History of Numbers, Wiley, New York, 2000. Babylonian Mathematics 8 1986)

Because of the large base, arithmetic was carried out with the aide of a pre-calculated table. For example, two tablets found at Senkerah on the Euphrates in 1854, dating from 2000 BC, give lists of the squares of numbers up to 59 and the cubes of numbers up to 32. The Babylonians used the lists

of squares together with the formulae

$$\begin{aligned}
 ab &= \frac{(a+b)^2 - a^2 - b^2}{2} \\
 ab &= \frac{(a+b)^2 - a^2 - b^2}{2} \\
 ab &= \frac{(a+b)^2 - (a-b)^2}{4} \\
 ab &= \frac{(a+b)^2 - (a-b)^2}{4}
 \end{aligned}$$

to simplify multiplication. The Babylonians did not have an algorithm for long division. Instead, they based their method on the fact that

$$\frac{a}{b} = a \times \frac{1}{b}$$

together with a table of reciprocals. Numbers whose only prime factors are 2, 3 or 5 (known as 5-smooth or regular numbers) have finite reciprocals in sexagesimal notation, and tables with extensive lists of these reciprocals have been found.

Egyptian Mathematics

Our sources of Egyptian mathematics are scarce. Indeed, much of our knowledge of ancient Egyptian mathematics comes not from the hieroglyphics (carved sacred letters or sacred letters) inscribed on the hundreds of temples but from two papyri containing collections of mathematical problems with their solutions.

1. The Rhind Mathematical Papyrus named for A.H. Rhind (1833- 1863) who purchased it at Luxor in 1858. Origin: 1650 BCE but it was written very much earlier. It is 18 feet long and 13 inches wide. It is also called the Ahmes Papyrus after the scribe that last copied it.
2. The Moscow Mathematical Papyrus purchased by V. S. Golenishchev (d. 1947). Origin: 1700 BC. It is 15 ft long and 3 inches wide. Two sections of this chapter offer highlights from these papyri.

The Egyptian counting system was decimal. Though non positional, it could deal with numbers of great scale. Yet, there is no apparent way to construct numbers arbitrarily large. The number system was decimal with special symbols for 1, 10, 100, 1,000, 10,000, 100,000, and 1,000,000. Addition was accomplished by grouping and regrouping. Multiplication and division were essentially based on binary multiples. Fractions were ubiquitous but only unit fractions, with two exceptions, were allowed. All other fractions were required to be written as a sum of unit fractions. Geometry was limited to areas, volumes, and similarity. Curiously, though, volume measures for the fractional portions of the hekat a volume measuring about 4.8 liters, were symbolically expressed differently from others. Simple algebraic equations were solvable, even systems of equations in two dimensions could be solved.



Figure 3: Rhind Mathematical Papyrus

	READING FROM RIGHT TO LEFT					READING FROM LEFT TO RIGHT				
1										
10	∩					∩				
100	  					  				
1000										
10,000										
100,000										
1,000,000										

Figure 4: Mathematical Notations in Hieroglyphics

1	α	alpha	10	ι	iota	100	ρ	rho
2	β	beta	20	κ	kappa	200	σ	sigma
3	γ	gamma	30	λ	lambda	300	τ	tau
4	δ	delta	40	μ	mu	400	υ	upsilon
5	ϵ	epsilon	50	ν	nu	500	ϕ	phi
6	ζ	vau*	60	ξ	xi	600	χ	chi
7	ζ	zeta	70	$\omicron$	omicron	700	ψ	psi
8	η	eta	80	π	pi	800	ω	omega
9	θ	theta	90	$\koppa$ *	koppa*	900	$\sampi$	sampi

*vau, koppa, and sampi are obsolete characters

Figure 5: Greek Numerals

Greek mathematics

Greek mathematics refers to mathematics texts and advances written in Greek, developed from the 7th century BC to the 4th century AD around the shores of the Eastern Mediterranean. Our knowledge of Greek Mathematics is less reliable than our than that of the older Egyptian and Babylonian mathematics, because none of the original manuscripts are extant. There are two sources:

1. Byzantine Greek codices (manuscript books) written 500-1500 years after the Greek works were composed.
2. Arabic translations of Greek works and Latin translations of the Arabic versions.

Historians traditionally place the beginning of Greek mathematics proper to the age of Thales of Miletus (ca. 624–548 BC). The two earliest mathematical theorems, Thales' theorem and Intercept theorem are attributed to Thales. The former, which states that an angle inscribed in a semicircle is a right angle, may have been learned by Thales while in Babylon but tradition attributes to Thales a demonstration of the theorem. It is for this reason that Thales is often hailed as the father of the deductive organization of mathematics and as the first true mathematician. It is also known that within two hundred years of Thales the Greeks had introduced logical structure and the idea of proof into mathematics.

Another important figure in the development of Greek mathematics is Pythagoras of Samos (ca. 580–500 BC) who had established an order called the Pythagoreans, which held knowledge and property in common and hence all of the discoveries by individual Pythagoreans were attributed to the order. And since in antiquity it was customary to give all credit to the master, Pythagoras himself was given credit for the discoveries made by his order.

Thales is supposed to have used geometry to solve problems such as calculating the height of pyramids based on the length of shadows, and the distance of ships from the shore. He is also credited by tradition with having made the first proof of two geometric theorems—the "Theorem of Thales" and the "Intercept theorem" described above. Pythagoras is widely credited with recognizing the mathematical basis of musical harmony and, according to Proclus' commentary on Euclid, he discovered

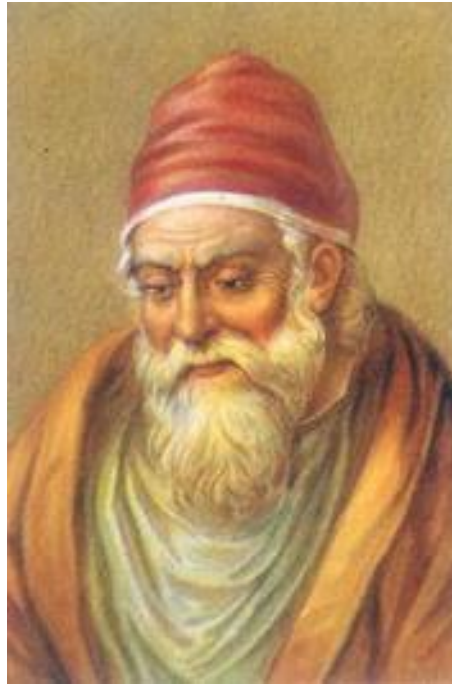


Figure 6: Euclid

the theory of proportionals and constructed regular solids. The Pythagoreans regarded numerology and geometry as fundamental to understanding the nature of the universe and therefore central to their philosophical and religious ideas. They are credited with numerous mathematical advances, such as the discovery of irrational numbers.

Euclid and the elements of geometry

Euclid, (born c. 300 BCE, Alexandria, Egypt), the most prominent mathematician of Greco-Roman antiquity, best known for his treatise on geometry, the **Elements**. Euclid compiled his Elements from a number of works of earlier men. Among these are Hippocrates of Chios (flourished c. 440 BCE), not to be confused with the physician Hippocrates of Cos (c. 460–375 BCE). The latest compiler before Euclid was Theudius, whose textbook was used in the Academy and was probably the one used by Aristotle (384–322 BCE).

The older elements were at once superseded by Euclid's and then forgotten. For his subject matter

Euclid drew upon all his predecessors, but it is clear that the whole design of his work was his own, culminating in the construction of the five regular solids, now known as the Platonic solids. A brief survey of the Elements belies a common belief that it concerns only geometry. This misconception may be caused by reading no further than Books I through IV, which cover elementary plane geometry. Euclid understood that building a logical and rigorous geometry (and mathematics) depends on the foundation—a foundation that Euclid began in Book I with 23 definitions (such as “a point is that which has no part” and “a line is a length without breadth”), five unproved assumptions that Eu-

clid called postulates (now known as axioms), and five further unproved assumptions that he called common notions. (See the table of Euclid's 10 initial assumptions.) Book I then proves elementary theorems about triangles and parallelograms and ends with the Pythagorean theorem. (For Euclid's proof of the theorem, see Sidebar: Euclid's Windmill Proof.)

Archimedes, Apollonius

Archimedes was a Greek mathematician, philosopher and inventor who wrote important works on geometry, arithmetic and mechanics. In mechanics he defined the principle of the lever and is credited with inventing the compound pulley and the hydraulic screw for raising water from a lower to higher level. He is most famous for discovering the law of hydrostatics, sometimes known as 'Archimedes' principle', stating that a body immersed in fluid loses weight equal to the weight of the amount of fluid it displaces. During the Roman conquest of Sicily in 214 BC Archimedes worked for the state, and several of his mechanical devices were employed in the defence of Syracuse. Among the war machines attributed to him are the catapult and - perhaps legendary - a mirror system for focusing the sun's rays on the invaders' boats and igniting them. After Syracuse was captured, Archimedes was killed by a Roman soldier.

Apollonius of Perga (late 3rd – early 2nd centuries BC) was a Greek geometer and astronomer known for his theories on the topic of conic sections. Beginning from the theories of Euclid and Archimedes on the topic, he brought them to the state they were in just before the invention of analytic geometry. His definitions of the terms ellipse, parabola, and hyperbola are the ones in use today.

Apollonius worked on many other topics, including astronomy. Most of the work has not survived except in fragmentary references in other authors. His hypothesis of eccentric orbits to explain the apparently aberrant motion of the planets, commonly believed until the Middle Ages, was superseded during the Renaissance.

Suggested Readings

1. D.M. Burton, The History of Mathematics, Allyn and Bacon, 5th edition.
2. Florian Cajori, A History of Mathematics

Unit II

Course Structure

- Introduction
 - Development of Trigonometry
 - Development of Algebra
 - Development of Analytic Geometry
-

Introduction

The development of mathematics is intimately interwoven with the progress of civilization, influencing the course of history through its application to science and technology. Mathematics has changed with each step of advancement in civilizations. Even the mathematics of the 1800s can seem quite strange now, so greatly has mathematics evolved in the past 100 years and so thoroughly has it been reworked in the post-modern approach.

Despite its arcane appearance from the outside looking in, the present, abstract and highly specialized state of mathematics is the natural evolution of the subject, and there is much ahead that is exciting.

Development of Trigonometry

It began as a branch of geometry and was utilized extensively by early Greek mathematicians to determine unknown distances. The most notable examples are the use by Aristarchus (310-250 B.C.) to determine the distance to the Moon and Sun, and by Eratosthenes (c. 276-195 B.C.) to calculate the Earth's circumference. The general principles of trigonometry were formulated by the Greek astronomer, Hipparchus of Nicaea (162-127 B.C.), who is generally credited as the founder of trigonometry. His ideas were worked out by Ptolemy of Alexandria (A.D. c. 90-168), who used them to develop the influential Ptolemaic theory of astronomy. Much of the information we know about the work of Hipparchus and Ptolemy comes from Ptolemy's compendium, *The Almagest*, written around 150.

Trigonometry was initially considered a field of the science of astronomy. It was later established as a separate branch of mathematics—largely through the work of the mathematicians Johann Bernoulli (1667-1748) and Leonhard Euler (1707-1783). The major trigonometric functions, including sine, cosine, and tangent, were first defined as ratios of sides in a right triangle. Since trigonometric functions are intrinsically related, they can be used to determine the dimensions of any triangle given limited information. In the eighteenth century, the definitions of trigonometric functions were

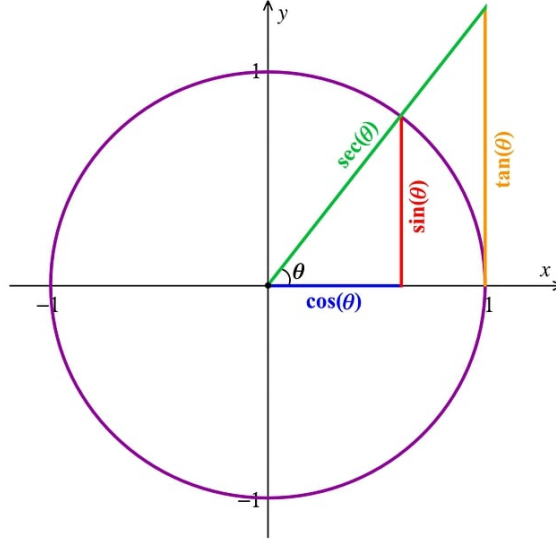


Figure 7: Trigonometric Functions Using Circle

broadened by being defined as points on a unit circle. This allowed the development of graphs of functions related to the angles they represent which were periodic. Today, using the periodic nature of trigonometric functions, mathematicians and scientists have developed mathematical models to predict many natural periodic phenomena. Ancient Greek and Hellenistic mathematicians made

use of the chord. Given a circle and an arc on the circle, the chord is the line that subtends the arc. A chord's perpendicular bisector passes through the center of the circle and bisects the angle. One half of the bisected chord is the sine of one half the bisected angle, that is,

$$\text{chord } \theta = 2 \sin \frac{\theta}{2}, \text{ chord } \theta = 2 \sin \frac{\theta}{2},$$

and consequently the sine function is also known as the half-chord. Due to this relationship, a number of trigonometric identities and theorems that are known today were also known to Hellenistic mathematicians, but in their equivalent chord form.

Some of the early and very significant developments of trigonometry were in India. Influential works from the 4th–5th century, known as the *Siddhantas* (of which there were five, the most important of which is the *Surya Siddhanta*) first defined the sine as the modern relationship between half an angle and half a chord, while also defining the cosine, versine, and inverse sine. Soon afterwards, another Indian mathematician and astronomer, Aryabhatta (476–550 AD), collected and expanded upon the developments of the *Siddhantas* in an important work called the *Aryabhatya*. The *Siddhantas* and the *Aryabhatya* contain the earliest surviving tables of sine values and versine values, to an accuracy of 4 decimal places. They used the words *jya* for sine, *kojya* for cosine, *utkrama-jya* for versine, and *otkram jya* for inverse sine. The words *jya* and *kojya* eventually became sine and cosine respectively after a mistranslation described above. Madhava (c. 1400) made early strides in the analysis of

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots$$

$$\sin x^2 = x^2 - \frac{(x^2)^3}{3!} + \frac{(x^2)^5}{5!} - \frac{(x^2)^7}{7!} + \dots$$

$$\sin x^2 = x^2 - \frac{x^6}{3!} + \frac{x^{10}}{5!} - \frac{x^{14}}{7!} + \dots$$

trigonometric functions and their infinite series expansions. He developed the concepts of the power series and Taylor series, and produced the power series expansions of sine, cosine, tangent, and arctangent. Using the Taylor series approximations of sine and cosine, he produced a sine table to 12 decimal places of accuracy and a cosine table to 9 decimal places of accuracy. His works were expanded by his followers at the Kerala School up to the 16th century.

Development of Algebra

The roots of the word, algebra, can be found in Medieval Latin and the Arabic word, *al-jabr*, which means 'the reduction'. The word was first used as early as 1551. The history of algebra can be divided into three parts:

- The rhetorical, or written stage, where only words were used in equations
- The syncopated, or shortened, stage, where some symbols were used in equations
- The symbolic or modern stage

Early works of algebra of ancient Babylonians and Egyptians lack the abstract notation that algebra has today. The Babylonians had methods of solving quadratic equations, while the Egyptians used the symbol heap for the unknown.

In China, a treatise called Nine Chapters was compiled in the first century CE composed of 246 problems. The text shows methods of solving determinate and indeterminate equations. More sophisticated than the works of the Babylonians and Egyptians, the treatise is mostly what is known today as rhetorical algebra—problems and solutions are expressed in words rather than in algebraic notations.

Algebra to the ancient Greeks is an unknown science except for the Greek mathematician Diophantus of Alexandria (3rd century BCE). His work *Arithmetica* contains the first suggestions of algebraic notations and is probably the earliest treatise on algebra. He used algebraic equations and



Figure 8: Khwarzimi

notations in presenting problems and solutions in *Arithmetica*.

In the 6th century CE in India, Aryabhata's works show knowledge in summing an arithmetic series, and solving quadratic equations and indeterminate linear equations. Shortly after, in the 7th century, Brahmagupta applied algebra to astronomy; his works give the rules for using negative numbers, and solving quadratic equations.

The Islamic scholars made several contributions to algebra as well—most notable of them is al-Khwārizmī. His treatise *A short book on the calculus al-jabr and al-muqabalah*, deals with the solution of quadratic equations. He did not use algebraic notations in his treatise but employed rhetorical algebra. This is one reason why some consider Diophantus as the “Father of Algebra” rather than Al-Khwārizmī.

Development of Analytic Geometry

Analytic geometry, also called coordinate geometry, is a mathematical subject in which algebraic symbolism and methods are used to represent and solve problems in geometry. Apollonius of Perga foreshadowed the development of analytic geometry by more than 1,800 years with his book **Conics**. He defined a conic as the intersection of a cone and a plane (see figure).

Further development of coordinate systems in mathematics emerged only after algebra had matured under Islamic and Indian mathematicians. With the power of algebraic notation, mathemati-

cians were no longer completely dependent upon geometric figures and geometric intuition to solve problems. The more daring began to leave behind the standard geometric way of thinking in which linear (first power) variables corresponded to lengths, squares (second power) to areas, and cubics (third power) to volumes, with higher powers lacking “physical” interpretation. Two Frenchmen, the mathematician-philosopher René Descartes and the lawyer-mathematician Pierre de Fermat, were among the first to take this daring step.

Descartes and Fermat independently founded analytic geometry in the 1630s by adapting Viète’s algebra to the study of geometric loci. They used letters to represent distances that are variable instead of fixed. Descartes used equations to study curves defined geometrically, and he stressed the need to consider general algebraic curves—graphs of polynomial equations in x and y of all degrees. He demonstrated his method on a classical problem: finding all points P such that the product of the distances from P to certain lines equals the product of the distances to other lines.

Fermat emphasized that any relation between x and y coordinates determines a curve. Using this idea, he recast Apollonius’s arguments in algebraic terms and restored lost work. Fermat indicated that any quadratic equation in x and y can be put into the standard form of one of the conic sections.

Their ideas gained general acceptance only through the efforts of other mathematicians in the latter half of the 17th century. In particular, the Dutch mathematician Frans van Schooten translated Descartes’s writings from French to Latin. He added vital explanatory material, as did the French lawyer Florimond de Beaune, and the Dutch mathematician Johan de Witt. In England, the mathematician John Wallis popularized analytic geometry, using equations to define conics and derive their properties. He used negative coordinates freely, although it was Isaac Newton who unequivocally used two (oblique) axes to divide the plane into four quadrants.

Unit III

Course Structure

- Development of Calculus
 - Development of Selected Topics of Modern Mathematics
-

Development of Calculus

The discovery of calculus is often attributed to two men, Isaac Newton and Gottfried Leibniz, who independently developed its foundations. Although they both were instrumental in its creation, they thought of the fundamental concepts in very different ways. While Newton considered variables changing with time, Leibniz thought of the variables x and y as ranging over sequences of infinitely close values. He introduced dx and dy as differences between successive values of these sequences. Leibniz knew that dy/dx gives the tangent but he did not use it as a defining property. On the other hand, Newton used quantities x' and y' , which were finite velocities, to compute the tangent. Of course neither Leibniz nor Newton thought in terms of functions, but both always thought in terms of graphs. For Newton the calculus was geometrical while Leibniz took it towards analysis. It is interesting to note that Leibniz was very conscious of the importance of good notation and put a lot of thought into the symbols he used. Newton, on the other hand, wrote more for himself than anyone else. Consequently, he tended to use whatever notation he thought of on that day. This turned out to be important in later developments. Leibniz's notation was better suited to generalizing calculus to multiple variables and in addition it highlighted the operator aspect of the derivative and integral. As a result, much of the notation that is used in Calculus today is due to Leibniz.

The development of Calculus can roughly be described along a timeline which goes through three periods: Anticipation, Development, and Rigorization. In the Anticipation stage techniques were being used by mathematicians that involved infinite processes to find areas under curves or maximize certain quantities. In the Development stage Newton and Leibniz created the foundations of Calculus and brought all of these techniques together under the umbrella of the derivative and integral. However, their methods were not always logically sound, and it took mathematicians a long time during the Rigorization stage to justify them and put Calculus on a sound mathematical foundation.

In their development of the calculus both Newton and Leibniz used "infinitesimals", quantities that are infinitely small and yet nonzero. Of course, such infinitesimals do not really exist, but Newton and Leibniz found it convenient to use these quantities in their computations and their derivations of results. Although one could not argue with the success of calculus, this concept of infinitesimals bothered mathematicians. Lord Bishop Berkeley made serious criticisms of the calculus referring to infinitesimals as "the ghosts of departed quantities".

Berkeley's criticisms were well founded and important in that they focused the attention of mathematicians on a logical clarification of the calculus. It was to be over 100 years, however, before Calculus was to be made rigorous. Ultimately, Cauchy, Weierstrass, and Riemann reformulated Calculus in terms of limits rather than infinitesimals. Thus the need for these infinitely small (and nonexistent) quantities was removed, and replaced by a notion of quantities being "close" to others. The derivative and the integral were both reformulated in terms of limits. While it may seem like a lot of work to create rigorous justifications of computations that seemed to work fine in the first place, this is an important development. By putting Calculus on a logical footing, mathematicians were better able to understand and extend its results, as well as to come to terms with some of the more subtle aspects of the theory.

When we first study Calculus we often learn its concepts in an order that is somewhat backwards to its development. We wish to take advantage of the hundreds of years of thought that have gone into it. As a result, we often begin by learning about limits. Afterward we define the derivative and integral developed by Newton and Leibniz. But unlike Newton and Leibniz we define them in the modern way – in terms of limits. Afterward we see how the derivative and integral can be used to solve many of the problems that precipitated the development of Calculus.

Development of Modern Mathematics

When we consider the history of modern Mathematics, two questions arise:

1. what limitations shall be placed upon the term mathematics?
2. what force shall be assigned to the word Modern?

Here, we mainly limit ourselves into the domain of pure science. Questions of applications of the various branches will be considered incidentally. Such great contributions as those of Newton in the realm of Mathematical Physics, of Laplace in celestial mechanics, of Lagrange and Cauchy in wave theory, belong rather to the field of applications.

In particular, in the domain of numbers, reference will be made to certain of the contributions to the general theory, to the men who have placed the study of irrational and transcendental numbers upon a scientific foundation, and to those who have developed the modern theory of complex numbers and its elaboration in the field of quaternions and Ausdehnungslehre. In the theory of equations, the names of some of the leading investigators will be mentioned, together with a brief statement of the results which they secured. This phase of higher algebra will be followed by the theory of forms, or quantics. The later development of calculus, leading to differential equations and the theory of functions, will complete the algebraic side, save for a brief reference to the theory of probabilities. In the domain of geometry, some of the contributors to the later development of the analytic and synthetic fields will be mentioned, together with the most noteworthy results of their labours.

The term Modern Mathematics is not well defined. Algebra cannot be modern yet the theory of equations has received some of its most significant additions during the nineteenth century, while the theory of forms is a recent creation. Similarly with elementary geometry; the labours of Lobachevsky and Bolyai during the second quarter of the century threw a new light upon the whole subject, and more recently the study of the triangle has added another chapter to the theory. Thus, the history

of modern mathematics must also be the modern history of ancient branches, while subjects which seem the product of late generations have root in other centuries than the present.

The nineteenth century has been a period of intense study of first principles, of the recognition of necessary limitations of various branches, of a great spread of mathematical knowledge, and of the opening of extensive fields of applied mathematics. Especially influential has been the establishment of scientific schools and journals and university chairs. About the middle of the century, these schools began to exert a still greater influence through the custom of calling to them mathematicians of high repute, thus making Zurich, Karlsruhe, Munich, Dresden, and other cities well known as mathematical centres.

Unit IV

Course Structure

- Development of Modern geometries
 - Development of Modern algebra
 - Development of the methods of real analysis.
-

Development of Modern geometry

Descriptive, Projective, and Modern Synthetic Geometry are so interwoven in their historic development that it is even more difficult to separate them from one another than from the analytic geometry just mentioned. Monge had been in possession of his theory for over thirty years before the publication of his *Géométrie Descriptive* (1800), a delay due to the jealous desire of the military authorities to keep the valuable secret. It is true that certain of its features can be traced back to Desargues, Taylor, Lambert, and Frézier, but it was Monge who worked it out in detail as a science, although Lacroix (1795), inspired by Monge's lectures in the *École Polytechnique*, published the first work on the subject. After Monge's work appeared, Hachette (1812, 1818, 1821) added materially to its symmetry, subsequent French contributors being Leroy (1842), Olivier (from 1845), de la Gournerie (from 1860), Vallée, de Fourcy, Adhémar, and others. In Germany leading contributors have been Ziegler (1843), Anger (1858), and especially Fiedler (3d edn. 1883-88) and Wiener (1884-87). At this period Monge by no means confined himself to the descriptive geometry. So marked were his labors in the analytic geometry that he has been called the father of the modern theory. He also set forth the fundamental theorem of reciprocal polars, though not in modern language, gave some treatment of ruled surfaces, and extended the theory of polars to quadrics.

Monge and his school concerned themselves especially with the relations of form, and particularly with those of surfaces and curves in a space of three dimensions. Inspired by the general activity of the period, but following rather the steps of Desargues and Pascal, Carnot treated chiefly the metrical relations of figures. In particular he investigated these relations as connected with the theory of transversals, a theory whose fundamental property of a four-rayed pencil goes back to Pappos, and which, though revived by Desargues, was set forth for the first time in its general form in Carnot's *Géométrie de Position* (1803), and supplemented in his *Théorie des Transversales* (1806). In these works he introduced negative magnitudes, the general quadrilateral and quadrangle, and numerous other generalizations of value to the elementary geometry of to-day. But although Carnot's work was important and many details are now commonplace, neither the name of the theory nor the method employed have endured. The present Geometry of Position (*Geometrie der Lage*) has little in common with Carnot's *Géométrie de Position*.

Projective Geometry had its origin somewhat later than the period of Monge and Carnot. Newton had discovered that all curves of the third order can be derived by central projection from five fundamental types. But in spite of this fact the theory attracted so little attention for over a century that its origin is generally ascribed to Poncelet. A prisoner in the Russian campaign, confined at Saratoff on the Volga (1812-14), “privé,” as he says, “de toute espèce de livres et de secours, surtout distrait par les malheurs de ma patrie et les miens propres,” he still had the vigor of spirit and the leisure to conceive the great work which he published (1822) eight years later. In this work was first made prominent the power of central projection in demonstration and the power of the principle of continuity in research. His leading idea was the study of projective properties, and as a foundation principle he introduced the anharmonic ratio, a concept, however, which dates back to Pappos and which Desargues (1639) had also used. Möbius, following Poncelet, made much use of the anharmonic ratio in his *Barycentrische Calcül* (1827), but under the name “Doppelschnitt-Verhältniss” (ratio bisectionalis), a term now in common use under Steiner’s abbreviated form “Doppelverhältniss.” The name “anharmonic ratio” or “function” (*rapport anharmonique*, or *fonction anharmonique*) is due to Chasles, and “cross-ratio” was coined by Clifford. The anharmonic point and line properties of conics have been further elaborated by Brianchon, Chasles, Steiner, and von Staudt. To Poncelet is also due the theory of “figures homologiques,” the perspective axis and perspective center (called by Chasles the axis and center of homology), an extension of Carnot’s theory of transversals, and the “cordes idéales” of conics which Plücker applied to curves of all orders. He also discovered what Salmon has called “the circular points at infinity,” thus completing and establishing the first great principle of modern geometry, the principle of continuity. Brianchon (1806), through his application of Desargues’s theory of polars, completed the foundation which Monge had begun for Poncelet’s (1829) theory of reciprocal polars.

Among the most prominent geometers contemporary with Poncelet was Gergonne, who with more propriety might be ranked as an analytic geometer. He first (1813) used the term “polar” in its modern geometric sense, although Servois (1811) had used the expression “pole.” He was also the first (1825-26) to grasp the idea that the parallelism which Maurolycus, Snell, and Viète had noticed is a fundamental principle. This principle he stated and to it he gave the name which it now bears, the Principle of Duality, the most important, after that of continuity, in modern geometry. This principle of geometric reciprocation, the discovery of which was also claimed by Poncelet, has been greatly elaborated and has found its way into modern algebra and elementary geometry, and has recently been extended to mechanics by Genese. Gergonne was the first to use the word “class” in describing a curve, explicitly defining class and degree (order) and showing the duality between the two. He and Chasles were among the first to study scientifically surfaces of higher order.

Steiner (1832) gave the first complete discussion of the projective relations between rows, pencils, etc., and laid the foundation for the subsequent development of pure geometry. He practically closed the theory of conic sections, of the corresponding figures in three-dimensional space and of surfaces of the second order, and hence with him opens the period of special study of curves and surfaces of higher order. His treatment of duality and his application of the theory of projective pencils to the generation of conics are masterpieces. The theory of polars of a point in regard to a curve had been studied by Bobillier and by Grassmann, but Steiner (1848) showed that this theory can serve as the foundation for the study of plane curves independently of the use of coordinates, and introduced those noteworthy curves covariant to a given curve which now bear the names of himself, Hesse, and Cayley. This whole subject has been extended by Grassmann, Chasles, Cremona, and Jonquières. Steiner was the first to make prominent (1832) an example of correspondence of a more

complicated nature than that of Poncelet, Möbius, Magnus, and Chasles. His contributions, and those of Gudermann, to the geometry of the sphere were also noteworthy.

While Möbius, Plücker, and Steiner were at work in Germany, Chasles was closing the geometric era opened in France by Monge. His *Aperçu Historique* (1837) is a classic, and did for France what Salmon's works did for algebra and geometry in England, popularizing the researches of earlier writers and contributing both to the theory and the nomenclature of the subject. To him is due the name "homographic" and the complete exposition of the principle as applied to plane and solid figures, a subject which has received attention in England at the hands of Salmon, Townsend, and H. J. S. Smith.

Von Staudt began his labors after Plücker, Steiner, and Chasles had made their greatest contributions, but in spite of this seeming disadvantage he surpassed them all. Joining the Steiner school, as opposed to that of Plücker, he became the greatest exponent of pure synthetic geometry of modern times. He set forth (1847, 1856-60) a complete, pure geometric system in which metrical geometry finds no place. Projective properties foreign to measurements are established independently of number relations, number being drawn from geometry instead of conversely, and imaginary elements being systematically introduced from the geometric side. A projective geometry based on the group containing all the real projective and dualistic transformations, is developed, imaginary transformations being also introduced. Largely through his influence pure geometry again became a fruitful field. Since his time the distinction between the metrical and projective theories has been to a great extent obliterated,⁵⁹ the metrical properties being considered as projective relations to a fundamental configuration, the circle at infinity common for all spheres. Unfortunately von Staudt wrote in an unattractive style, and to Reye is due much of the popularity which now attends the subject.

Cremona began his publications in 1862. His elementary work on projective geometry (1875) in Leudesdorf's translation is familiar to English readers. His contributions to the theory of geometric transformations are valuable, as also his works on plane curves, surfaces, etc.

In England Mulcahy, but especially Townsend (1863), and Hirst, a pupil of Steiner's, opened the subject of modern geometry. Clifford did much to make known the German theories, besides himself contributing to the study of polars and the general theory of curves.

Development of Modern algebra

Modern algebra, also called abstract algebra, branch of mathematics concerned with the general algebraic structure of various sets (such as real numbers, complex numbers, matrices, and vector spaces), rather than rules and procedures for manipulating their individual elements.

Prior to the nineteenth century, algebra meant the study of the solution of polynomial equations. By the twentieth century algebra came to encompass the study of abstract, axiomatic systems such as groups, rings, and fields. This presentation provides an account of the history of the basic concepts, results, and theories of abstract algebra. The development of abstract algebra was propelled by the need for new tools to address certain classical problems that appeared unsolvable by classical means. A major theme of the approach in this book is to show how abstract algebra has arisen in attempts to solve some of these classical problems, providing context from which the reader may gain a deeper appreciation of the mathematics involved. Key features: Begins with an overview of classical algebra Contains separate chapters on aspects of the development of groups, rings, and fields Examines the evolution of linear algebra as it relates to other elements of abstract algebra Highlights

the lives and works of six notables: Cayley, Dedekind, Galois, Gauss, Hamilton, and especially the pioneering work of Emmy Noether. Offers suggestions to instructors on ways of integrating the history of abstract algebra into their teaching. Each chapter concludes with extensive references to the relevant literature. Mathematics instructors, algebraists, and historians of science will find the work a valuable reference. The book may also serve as a supplemental text for courses in abstract algebra or the history of mathematics.

During the second half of the 19th century, various important mathematical advances led to the study of sets in which any two elements can be added or multiplied together to give a third element of the same set. The elements of the sets concerned could be numbers, functions, or some other objects. As the techniques involved were similar, it seemed reasonable to consider the sets, rather than their elements, to be the objects of primary concern. A definitive treatise, *Modern Algebra*, was written in 1930 by the Dutch mathematician Bartel van der Waerden, and the subject has had a deep effect on almost every branch of mathematics.

Development of Real Analysis

The study of real functions has played a fundamental role in the development of mathematics over the last three centuries. The discovery of calculus by eighteenth century mathematicians, notably Newton and Leibniz, was largely due to increased understanding of the behavior of real functions. The birth of analysis is often traced to the early nineteenth century work of Cauchy, who gave precise definitions of concepts such as continuity and limits for real functions. Convergence problems while approximating real functions by Fourier series gave rise to both the Riemann and Lebesgue integrals. Cantor developed his set theory in an effort to answer uniqueness questions about Fourier series.

During this time, different techniques have been used as the theory behind them became available. For example, after Cauchy, various limiting operations such as pointwise and uniform convergence were studied, giving rise to various approximation techniques. At the turn of this century, measure theoretic techniques were exploited, leading to stochastic convergence ideas in the 1920's. Also, at about the same time topology was developed, and its applications to analysis gave rise to functional analysis.

In recent years, a new research trend has appeared which indicates the emergence of a yet another branch of inquiry that could be called set theoretic real analysis. This area is the study of families of real functions using modern techniques of set theory. These techniques include advanced forcing methods, special axioms of set theory such as Martin's axiom (MA) and proper forcing axiom (PFA), as well as some of their weaker consequences like additivity of measure and category.

Set theoretic real analysis is closely allied with descriptive set theory, but the objects studied in the two areas are different. The objects studied in descriptive set theory are various classes of (mostly nice) sets and their hierarchies, such as Borel sets or analytic sets. Set theoretic real analysis uses the tools of modern set theory to study real functions and is interested mainly in more pathological objects. Thus, the results concerning subsets of the real line (like the series of studies on "small" subsets $\mathbb{R}$, or deep studies of the duality between measure and category) are considered only remotely related to the subject. (However, some of these duality studies spread to real analysis too.)

Set theoretic real analysis already has a long history. Its roots can be traced back to the 1920's, where powerful new techniques based on the Axiom of Choice (AC) and the Continuum Hypothesis (CH) can be seen in many papers from such journals as *Fundamenta Mathematicae* and *Studia*

Mathematica. The most interesting consequences of the Continuum Hypothesis discovered in this period have been collected in 1934 monograph of Sierpinski, The influence of Sierpinski's results (and the monograph) on the set theoretic real analysis can be best seen in the next section.

The new emergence of the field was sparked by the discovery of powerful new techniques in set theory and can be compared to the parallel development of set theoretic topology during the late 1950's and 1960's. In fact, it is a bit surprising that the development of set theoretic analysis is so much behind that of set theoretic topology, since at the beginning of the century the applicability of set theory in analysis was at least as intense as in topology. This, however, can be probably attributed to the simple fact, that in the past half of a century there were many mathematicians that knew well both topology and set theory, and very few that knew well simultaneously analysis and set theory.

Suggested Readings

1. D.M. Burton, The History of Mathematics, Allyn and Bacon, 5th edition.
2. Carl B. Boyer and Uta C. Merzbach, A History of Mathematics 3rd Edition .
3. Florian Cajori , A History of Mathematics (Paperback).
4. J.H. Eves, An Introduction to the History of Mathematics, Saunders, 1990.

Choice Based Paper

MATCB 2.1

Block - II

Marks : 25 (SSE : 20; IA : 05)

Operations Research

Syllabus

• Unit 5 •

Formulation of linear programming models. Graphical solution. Basic solution (BS) and Basic Feasible Solution (BFS), Degenerate and non-degenerate BFS, Convex set, convex hull, convex polyhedron, extreme points, hyper plane.

• Unit 6 •

Standard form of LPP, Simplex method, Charnes' Big – M method.

• Unit 7 •

Transportation and assignment problems.

• Unit 8 •

A brief introduction to PERT and CPM, Components of PERT/CPM Network and precedence relationships, Critical path analysis.

Unit 5

Simple Linear Programming Problems

When both the objective and all the constraints in Expression 1.5 are linear functions, then the optimization problem is called a *linear programming problem*. This has the general form:

$$(2.1) \quad \left\{ \begin{array}{l} \max \quad z(x_1, \dots, x_n) = c_1x_1 + \dots + c_nx_n \\ \text{s.t.} \quad a_{11}x_1 + \dots + a_{1n}x_n \leq b_1 \\ \qquad \qquad \qquad \vdots \\ \qquad \qquad \qquad a_{m1}x_1 + \dots + a_{mn}x_n \leq b_m \\ \qquad \qquad \qquad h_{11}x_1 + \dots + h_{1n}x_n = r_1 \\ \qquad \qquad \qquad \vdots \\ \qquad \qquad \qquad h_{l1}x_1 + \dots + h_{ln}x_n = r_l \end{array} \right.$$

DEFINITION 2.1 (Linear Function). A function $z : \mathbb{R}^n \rightarrow \mathbb{R}$ is *linear* if there are constants $c_1, \dots, c_n \in \mathbb{R}$ so that:

$$(2.2) \quad z(x_1, \dots, x_n) = c_1x_1 + \dots + c_nx_n$$

LEMMA 2.2 (Linear Function). If $z : \mathbb{R}^n \rightarrow \mathbb{R}$ is linear then for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$ and for all scalar constants $\alpha \in \mathbb{R}$ we have:

$$(2.3) \quad z(\mathbf{x}_1 + \mathbf{x}_2) = z(\mathbf{x}_1) + z(\mathbf{x}_2)$$

$$(2.4) \quad z(\alpha\mathbf{x}_1) = \alpha z(\mathbf{x}_1)$$

EXERCISE 9. Prove Lemma 2.2.

For the time being, we will eschew the general form and focus exclusively on linear programming problems with two variables. Using this limited case, we will develop a graphical method for identifying optimal solutions, which we will generalize later to problems with arbitrary numbers of variables.

EXAMPLE 2.3. Consider the problem of a toy company that produces toy planes and toy boats. The toy company can sell its planes for \$10 and its boats for \$8 dollars. It costs \$3 in raw materials to make a plane and \$2 in raw materials to make a boat. A plane requires 3 hours to make and 1 hour to finish while a boat requires 1 hour to make and 2 hours to finish. The toy company knows it will not sell anymore than 35 planes per week. Further, given the number of workers, the company cannot spend anymore than 160 hours per week

finishing toys and 120 hours per week making toys. The company wishes to maximize the profit it makes by choosing how much of each toy to produce.

We can represent the profit maximization problem of the company as a linear programming problem. Let x_1 be the number of planes the company will produce and let x_2 be the number of boats the company will produce. The profit for each plane is $\$10 - \$3 = \$7$ per plane and the profit for each boat is $\$8 - \$2 = \$6$ per boat. Thus the total profit the company will make is:

$$(2.5) \quad z(x_1, x_2) = 7x_1 + 6x_2$$

The company can spend no more than 120 hours per week making toys and since a plane takes 3 hours to make and a boat takes 1 hour to make we have:

$$(2.6) \quad 3x_1 + x_2 \leq 120$$

Likewise, the company can spend no more than 160 hours per week finishing toys and since it takes 1 hour to finish a plane and 2 hour to finish a boat we have:

$$(2.7) \quad x_1 + 2x_2 \leq 160$$

Finally, we know that $x_1 \leq 35$, since the company will make no more than 35 planes per week. Thus the complete linear programming problem is given as:

$$(2.8) \quad \begin{cases} \max & z(x_1, x_2) = 7x_1 + 6x_2 \\ & s.t. \quad 3x_1 + x_2 \leq 120 \\ & \quad \quad x_1 + 2x_2 \leq 160 \\ & \quad \quad x_1 \leq 35 \\ & \quad \quad x_1 \geq 0 \\ & \quad \quad x_2 \geq 0 \end{cases}$$

EXERCISE 10. A chemical manufacturer produces three chemicals: A, B and C. These chemical are produced by two processes: 1 and 2. Running process 1 for 1 hour costs \$4 and yields 3 units of chemical A, 1 unit of chemical B and 1 unit of chemical C. Running process 2 for 1 hour costs \$1 and produces 1 units of chemical A, and 1 unit of chemical B (but none of Chemical C). To meet customer demand, at least 10 units of chemical A, 5 units of chemical B and 3 units of chemical C must be produced daily. Assume that the chemical manufacturer wants to minimize the cost of production. Develop a linear programming problem describing the constraints and objectives of the chemical manufacturer. [Hint: Let x_1 be the amount of time Process 1 is executed and let x_2 be amount of time Process 2 is executed. Use the coefficients above to express the cost of running Process 1 for x_1 time and Process 2 for x_2 time. Do the same to compute the amount of chemicals A, B, and C that are produced.]

1. Modeling Assumptions in Linear Programming

Inspecting Example 2.3 (or the more general Problem 2.1) we can see there are several assumptions that must be satisfied when using a linear programming model. We enumerate these below:

Proportionality Assumption: A problem can be phrased as a linear program only if the contribution to the objective function *and* the left-hand-side of each constraint by each decision variable $(x_1, \dots, x_n)$ is proportional to the value of the decision variable.

Additivity Assumption: A problem can be phrased as a linear programming problem only if the contribution to the objective function *and* the left-hand-side of each constraint by any decision variable x_i ($i = 1, \dots, n$) is completely independent of any other decision variable x_j ($j \neq i$) and additive.

Divisibility Assumption: A problem can be phrased as a linear programming problem only if the quantities represented by each decision variable are infinitely divisible (i.e., fractional answers make sense).

Certainty Assumption: A problem can be phrased as a linear programming problem only if the coefficients in the objective function and constraints are known with certainty.

The first two assumptions simply assert (in English) that both the objective function and functions on the left-hand-side of the (in)equalities in the constraints are linear functions of the variables $x_1, \dots, x_n$.

The third assumption asserts that a valid optimal answer could contain fractional values for decision variables. It's important to understand how this assumption comes into play—even in the toy making example. Many quantities can be divided into non-integer values (ounces, pounds etc.) but many other quantities cannot be divided. For instance, can we really expect that it's reasonable to make $1/2$ a plane in the toy making example? When values must be constrained to true integer values, the linear programming problem is called an *integer programming problem*. These problems are outside the scope of this course, but there is a *vast* literature dealing with them. For many problems, particularly when the values of the decision variables may become large, a fractional optimal answer could be obtained and then rounded to the nearest integer to obtain a reasonable answer. For example, if our toy problem were re-written so that the optimal answer was to make 1045.3 planes, then we could round down to 1045.

The final assumption asserts that the coefficients (e.g., profit per plane or boat) is known with absolute certainty. In traditional linear programming, there is no lack of knowledge about the make up of the objective function, the coefficients in the left-hand-side of the constraints or the bounds on the right-hand-sides of the constraints. There is a literature on *stochastic programming* that relaxes some of these assumptions, but this too is outside the scope of the course.

EXERCISE 11. In a short sentence or two, discuss whether the problem given in Example 2.3 meets all of the assumptions of a scenario that can be modeled by a linear programming problem. Do the same for Exercise 10. [Hint: Can you make $2/3$ of a toy? Can you run a process for $1/3$ of an hour?]

EXERCISE 12. Suppose the costs are not known with certainty but instead a probability distribution for each value of c_i ($i = 1, \dots, n$) is known. Suggest a way of constructing a linear program from the probability distributions. [Hint: Suppose I tell you that I'll give you a uniformly random amount of money between \$1 and \$2. How much money do you expect to receive? Use the same reasoning to answer the question.]

2. Graphically Solving Linear Programs Problems with Two Variables (Bounded Case)

Linear Programs (LP's) with two variables can be solved graphically by plotting the feasible region along with the level curves of the objective function. We will show that we can find a point in the feasible region that maximizes the objective function using the level curves of the objective function. We illustrate the method first using the problem from Example 2.3.

EXAMPLE 2.4 (Continuation of Example 2.3). Let's continue the example of the Toy Maker begin in Example 2.3. To solve the linear programming problem graphically, begin by drawing the feasible region. This is shown in the blue shaded region of Figure 2.1.

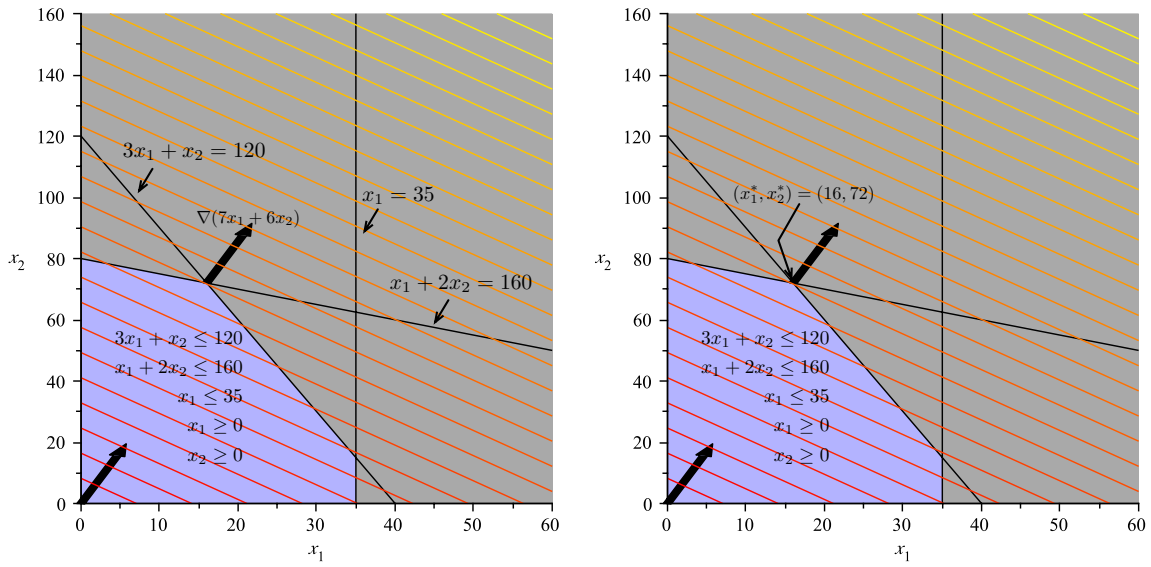


Figure 2.1. Feasible Region and Level Curves of the Objective Function: The shaded region in the plot is the feasible region and represents the intersection of the five inequalities constraining the values of x_1 and x_2 . On the right, we see the optimal solution is the “last” point in the feasible region that intersects a level set as we move in the direction of increasing profit.

After plotting the feasible region, the next step is to plot the level curves of the objective function. In our problem, the level sets will have the form:

$$7x_1 + 6x_2 = c \implies x_2 = \frac{-7}{6}x_1 + \frac{c}{6}$$

This is a set of parallel lines with slope $-7/6$ and intercept $c/6$ where c can be varied as needed. The level curves for various values of c are parallel lines. In Figure 2.1 they are shown in colors ranging from red to yellow depending upon the value of c . Larger values of c are more yellow.

To solve the linear programming problem, follow the level sets along the gradient (shown as the black arrow) until the last level set (line) intersects the feasible region. If you are doing this by hand, you can draw a single line of the form $7x_1 + 6x_2 = c$ and then simply

draw parallel lines in the direction of the gradient $(7, 6)$. At some point, these lines will fail to intersect the feasible region. The last line to intersect the feasible region will do so at a point that maximizes the profit. In this case, the point that maximizes $z(x_1, x_2) = 7x_1 + 6x_2$, subject to the constraints given, is $(x_1^*, x_2^*) = (16, 72)$.

Note the point of optimality $(x_1^*, x_2^*) = (16, 72)$ is at a corner of the feasible region. This corner is formed by the intersection of the two lines: $3x_1 + x_2 = 120$ and $x_1 + 2x_2 = 160$. In this case, the constraints

$$3x_1 + x_2 \leq 120$$

$$x_1 + 2x_2 \leq 160$$

are both *binding*, while the other constraints are non-binding. In general, we will see that when an optimal solution to a linear programming problem exists, it will always be at the intersection of several binding constraints; that is, it will occur at a corner of a higher-dimensional polyhedron.

Convex Sets, Functions and Cones and Polyhedral Theory

In this chapter, we will cover all of the geometric prerequisites for understanding the theory of linear programming. We will use the results in this section to prove theorems about the Simplex Method in other sections.

1. Convex Sets

DEFINITION 4.1 (Convex Set). Let $X \subseteq \mathbb{R}^n$. Then the set X is convex if and only if for all pairs $\mathbf{x}_1, \mathbf{x}_2 \in X$ we have $\lambda \mathbf{x}_1 + (1 - \lambda) \mathbf{x}_2 \in X$ for all $\lambda \in [0, 1]$.

The definition of convexity seems complex, but it is easy to understand. First recall that if $\lambda \in [0, 1]$, then the point $\lambda \mathbf{x}_1 + (1 - \lambda) \mathbf{x}_2$ is on the line segment connecting $\mathbf{x}_1$ and $\mathbf{x}_2$ in $\mathbb{R}^n$. For example, when $\lambda = 1/2$, then the point $\lambda \mathbf{x}_1 + (1 - \lambda) \mathbf{x}_2$ is the midpoint between $\mathbf{x}_1$ and $\mathbf{x}_2$. In fact, for every point $\mathbf{x}$ on the line connecting $\mathbf{x}_1$ and $\mathbf{x}_2$ we can find a value $\lambda \in [0, 1]$ so that $\mathbf{x} = \lambda \mathbf{x}_1 + (1 - \lambda) \mathbf{x}_2$. Then we can see that, convexity asserts that if $\mathbf{x}_1, \mathbf{x}_2 \in X$, then every point on the line connecting $\mathbf{x}_1$ and $\mathbf{x}_2$ is also in the set X .

DEFINITION 4.2 (Positive Combination). Let $\mathbf{x}_1, \dots, \mathbf{x}_m \in \mathbb{R}^n$. If $\lambda_1, \dots, \lambda_m > 0$ and then

$$(4.1) \quad \mathbf{x} = \sum_{i=1}^m \lambda_i \mathbf{x}_i$$

is called a *positive combination* of $\mathbf{x}_1, \dots, \mathbf{x}_m$.

DEFINITION 4.3 (Convex Combination). Let $\mathbf{x}_1, \dots, \mathbf{x}_m \in \mathbb{R}^n$. If $\lambda_1, \dots, \lambda_m \in [0, 1]$ and

$$\sum_{i=1}^m \lambda_i = 1$$

then

$$(4.2) \quad \mathbf{x} = \sum_{i=1}^m \lambda_i \mathbf{x}_i$$

is called a *convex combination* of $\mathbf{x}_1, \dots, \mathbf{x}_m$. If $\lambda_i < 1$ for all $i = 1, \dots, m$, then Equation 4.2 is called a *strict convex combination*.

REMARK 4.4. If you recall the definition of linear combination, we can see that we move from the very general to the very specific as we go from linear combinations to positive combinations to convex combinations. A linear combination of points or vectors allowed us to choose any real values for the coefficients. A positive combination restricts us to positive values, while a convex combination asserts that those values must be non-negative and sum to 1.

EXAMPLE 4.5. Figure 4.1 illustrates a convex and non-convex set. Non-convex sets have

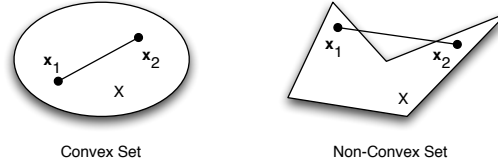


Figure 4.1. Examples of Convex Sets: The set on the left (an ellipse and its interior) is a convex set; every pair of points inside the ellipse can be connected by a line contained entirely in the ellipse. The set on the right is clearly not convex as we've illustrated two points whose connecting line is not contained inside the set.

some resemblance to crescent shapes or have components that look like crescents.

THEOREM 4.6. *The intersection of a finite number of convex sets in $\mathbb{R}^n$ is convex.*

PROOF. Let $C_1, \dots, C_n \subseteq \mathbb{R}^n$ be a finite collection of convex sets. Let

$$(4.3) \quad C = \bigcap_{i=1}^n C_i$$

be the set formed from the intersection of these sets. Choose $\mathbf{x}_1, \mathbf{x}_2 \in C$ and $\lambda \in [0, 1]$. Consider $\mathbf{x} = \lambda \mathbf{x}_1 + (1 - \lambda) \mathbf{x}_2$. We know that $\mathbf{x}_1, \mathbf{x}_2 \in C_1, \dots, C_n$ by definition of C . By convexity, we know that $\mathbf{x} \in C_1, \dots, C_n$ by convexity of each set. Therefore, $\mathbf{x} \in C$. Thus C is a convex set. $\square$

2. Convex and Concave Functions

DEFINITION 4.7 (Convex Function). A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a convex function if it satisfies:

$$(4.4) \quad f(\lambda \mathbf{x}_1 + (1 - \lambda) \mathbf{x}_2) \leq \lambda f(\mathbf{x}_1) + (1 - \lambda) f(\mathbf{x}_2)$$

for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$ and for all $\lambda \in [0, 1]$.

This definition is illustrated in Figure 4.2. When f is a univariate function, this definition can be shown to be equivalent to the definition you learned in Calculus I (Math 140) using first and second derivatives.

DEFINITION 4.8 (Concave Function). A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a concave function if it satisfies:

$$(4.5) \quad f(\lambda \mathbf{x}_1 + (1 - \lambda) \mathbf{x}_2) \geq \lambda f(\mathbf{x}_1) + (1 - \lambda) f(\mathbf{x}_2)$$

for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$ and for all $\lambda \in [0, 1]$.¹

To visualize this definition, simply flip Figure 4.2 upside down. The following theorem is a powerful tool that can be used to show sets are convex. It's proof is outside the scope of the class, but relatively easy.

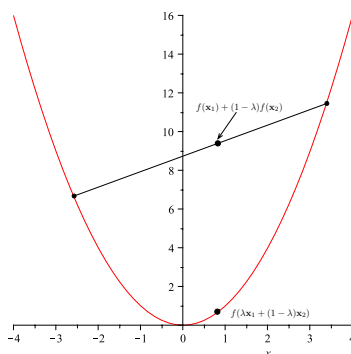


Figure 4.2. A convex function: A convex function satisfies the expression $f(\lambda \mathbf{x}_1 + (1 - \lambda) \mathbf{x}_2) \leq \lambda f(\mathbf{x}_1) + (1 - \lambda) f(\mathbf{x}_2)$ for all $\mathbf{x}_1$ and $\mathbf{x}_2$ and $\lambda \in [0, 1]$.

THEOREM 4.9. *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a convex function. Then the set $C = \{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) \leq c\}$, where $c \in \mathbb{R}$, is a convex set.*

EXERCISE 40. Prove the Theorem 4.9. [Hint: Skip ahead and read the proof of Lemma 4.15. Follow the steps in that proof, but apply them to f .]

3. Polyhedral Sets

Important examples of convex sets are polyhedral sets, the multi-dimensional analogs of polygons in the plane. In order to understand these structures, we must first understand hyperplanes and half-spaces.

DEFINITION 4.10 (Hyperplane). Let $\mathbf{a} \in \mathbb{R}^n$ be a constant vector in n -dimensional space and let $b \in \mathbb{R}$ be a constant scalar. The set of points

$$(4.6) \quad H = \{\mathbf{x} \in \mathbb{R}^n | \mathbf{a}^T \mathbf{x} = b\}$$

is a *hyperplane* in n -dimensional space. *Note the use of column vectors for $\mathbf{a}$ and $\mathbf{x}$ in this definition.*

EXAMPLE 4.11. Consider the hyper-plane $2x_1 + 3x_2 + x_3 = 5$. This is shown in Figure 4.3. This hyperplane is composed of the set of points $(x_1, x_2, x_3) \in \mathbb{R}^3$ satisfying $2x_1 + 3x_2 + x_3 = 5$. This can be plotted implicitly or explicitly by solving for one of the variables, say x_3 . We can write x_3 as a function of the other two variables as:

$$(4.7) \quad x_3 = 5 - 2x_1 - 3x_2$$

DEFINITION 4.12 (Half-Space). Let $\mathbf{a} \in \mathbb{R}^n$ be a constant vector in n -dimensional space and let $b \in \mathbb{R}$ be a constant scalar. The sets of points

$$(4.8) \quad H_l = \{\mathbf{x} \in \mathbb{R}^n | \mathbf{a}^T \mathbf{x} \leq b\}$$

$$(4.9) \quad H_u = \{\mathbf{x} \in \mathbb{R}^n | \mathbf{a}^T \mathbf{x} \geq b\}$$

are the half-spaces defined by the hyperplane $\mathbf{a}^T \mathbf{x} = b$.

Unit 6

Extreme Points

DEFINITION 4.27 (Extreme Point of a Convex Set). Let C be a convex set. A point $\mathbf{x}_0 \in C$ is a *extreme point* of C if there are *no points* $\mathbf{x}_1$ and $\mathbf{x}_2$ ($\mathbf{x}_1 \neq \mathbf{x}_0$ or $\mathbf{x}_2 \neq \mathbf{x}_0$) so that $\mathbf{x} = \lambda\mathbf{x}_1 + (1 - \lambda)\mathbf{x}_2$ for some $\lambda \in (0, 1)$.²

An extreme point is simply a point in a convex set C that cannot be expressed as a strict convex combination of any other pair of points in C . We will see that extreme points must be located in specific locations in convex sets.

DEFINITION 4.28 (Boundary of a set). Let $C \subseteq \mathbb{R}^n$ be (convex) set. A point $\mathbf{x}_0 \in C$ is on the *boundary* of C if for all $\epsilon > 0$,

$$B_\epsilon(\mathbf{x}_0) \cap C \neq \emptyset \text{ and}$$

$$B_\epsilon(\mathbf{x}_0) \cap \mathbb{R}^n \setminus C \neq \emptyset$$

EXAMPLE 4.29. A convex set, its boundary and a boundary point are illustrated in Figure 4.8.

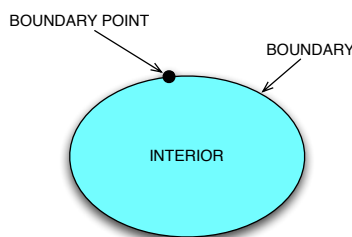
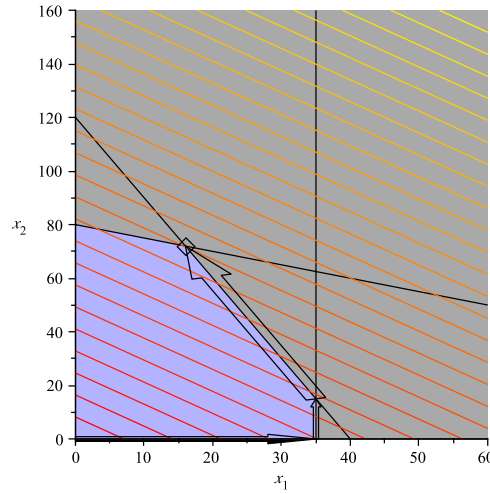


Figure 4.8. Boundary Point: A boundary point of a (convex) set C is a point in the set so that for *every* ball of any radius centered at the point contains some points inside C and some points outside C .

LEMMA 4.30. Suppose C is a convex set. If $\mathbf{x}$ is an extreme point of C , then $\mathbf{x}$ is on the boundary of C .

PROOF. Suppose not, then $\mathbf{x}$ is not on the boundary and thus there is some $\epsilon > 0$ so that $B_\epsilon(\mathbf{x}_0) \subset C$. Since $B_\epsilon(\mathbf{x}_0)$ is a hypersphere, we can choose two points $\mathbf{x}_1$ and $\mathbf{x}_2$ on the boundary of $B_\epsilon(\mathbf{x}_0)$ so that the line segment between these points passes through the center of $B_\epsilon(\mathbf{x}_0)$. But this center point is $\mathbf{x}_0$. Therefore $\mathbf{x}_0$ is the mid-point of $\mathbf{x}_1$ and $\mathbf{x}_2$ and since $\mathbf{x}_1, \mathbf{x}_2 \in C$ and $\lambda\mathbf{x}_1 + (1 - \lambda)\mathbf{x}_2 = \mathbf{x}_0$ with $\lambda = 1/2$ it follows that $\mathbf{x}_0$ cannot be an extreme point, since it is a strict convex combination of $\mathbf{x}_1$ and $\mathbf{x}_2$. This completes the proof. $\square$

Most important in our discussion of linear programming will be the extreme points of polyhedral sets that appear in linear programming problems. The following theorem establishes the relationship between extreme points in a polyhedral set and the intersection of hyperplanes in such a set.



The Simplex Algorithm: The path around the feasible region is shown in the figure. Each exchange of a basic and non-basic variable moves us along an edge of the polygon in a direction that increases the value of the objective function.

Simplex Method–Tableau Form

No one executes the simplex algorithm in algebraic form. Instead, several representations (tableau representations) have been developed to lessen the amount of writing that needs to be done and to collect all pertinent information into a single table.

To see how a *Simplex Tableau* is derived, consider Problem P in standard form:

$$P \begin{cases} \max & \mathbf{c}^T \mathbf{x} \\ s.t. & \mathbf{A} \mathbf{x} = \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0} \end{cases}$$

We can re-write P in an unusual way by introducing a new variable z and separating $\mathbf{A}$ into its basic and non-basic parts to obtain:

$$(5.22) \quad \begin{aligned} \max & \quad z \\ s.t. & \quad z - \mathbf{c}_B^T \mathbf{x}_B - \mathbf{c}_N^T \mathbf{x}_N = 0 \\ & \quad \mathbf{B} \mathbf{x}_B + \mathbf{N} \mathbf{x}_N = \mathbf{b} \\ & \quad \mathbf{x}_B, \mathbf{x}_N \geq 0 \end{aligned}$$

From the second equation, it's clear

$$(5.23) \quad \mathbf{x}_B + \mathbf{B}^{-1} \mathbf{N} \mathbf{x}_N = \mathbf{B}^{-1} \mathbf{b}$$

We can multiply this equation by $\mathbf{c}_B^T$ to obtain:

$$(5.24) \quad \mathbf{c}_B^T \mathbf{x}_B + \mathbf{c}_B^T \mathbf{B}^{-1} \mathbf{N} \mathbf{x}_N = \mathbf{c}_B^T \mathbf{B}^{-1} \mathbf{b}$$

If we add this equation to the equation $z - \mathbf{c}_B^T \mathbf{x}_B - \mathbf{c}_N^T \mathbf{x}_N = 0$ we obtain:

$$(5.25) \quad z + \mathbf{0}^T \mathbf{x}_B + \mathbf{c}_B^T \mathbf{B}^{-1} \mathbf{N} \mathbf{x}_N - \mathbf{c}_N^T \mathbf{x}_N = \mathbf{c}_B^T \mathbf{B}^{-1} \mathbf{b}$$

Here $\mathbf{0}$ is the vector of zeros of appropriate size. This equation can be written as:

$$(5.26) \quad z + \mathbf{0}^T \mathbf{x}_B + (\mathbf{c}_B^T \mathbf{B}^{-1} \mathbf{N} - \mathbf{c}_N^T) \mathbf{x}_N = \mathbf{c}_B^T \mathbf{B}^{-1} \mathbf{b}$$

We can now represent this set of equations as a large matrix (or tableau):

	z	$\mathbf{x}_B$	$\mathbf{x}_N$	RHS	
z	1	$\mathbf{0}$	$\mathbf{c}_B^T \mathbf{B}^{-1} \mathbf{N} - \mathbf{c}_N^T$	$\mathbf{c}_B^T \mathbf{B}^{-1} \mathbf{b}$	Row 0
$\mathbf{x}_B$	$\mathbf{0}$	$\mathbf{1}$	$\mathbf{B}^{-1} \mathbf{N}$	$\mathbf{B}^{-1} \mathbf{b}$	Rows 1 through m

The augmented matrix shown within the table:

$$(5.27) \quad \left[\begin{array}{cc|cc|c} 1 & \mathbf{0} & \mathbf{c}_B^T \mathbf{B}^{-1} \mathbf{N} - \mathbf{c}_N^T & \mathbf{c}_B^T \mathbf{B}^{-1} \mathbf{b} \\ \mathbf{0} & \mathbf{1} & \mathbf{B}^{-1} \mathbf{N} & \mathbf{B}^{-1} \mathbf{b} \end{array} \right]$$

is simply the matrix representation of the simultaneous equations described by Equations 5.23 and 5.26. We can see that the first row consists of a row of the first row of the $(m+1) \times (m+1)$ identity matrix, the reduced costs of the non-basic variables and the current objective function values. The remainder of the rows consist of the rest of the $(m+1) \times (m+1)$ identity matrix, the matrix $\mathbf{B}^{-1} \mathbf{N}$ and $\mathbf{B}^{-1} \mathbf{b}$ the current non-zero part of the basic feasible solution.

This matrix representation (or tableau representation) contains all of the information we need to execute the simplex algorithm. An entering variable is chosen from among the columns containing the reduced costs and matrix $\mathbf{B}^{-1} \mathbf{N}$. Naturally, a column with a negative reduced cost is chosen. We then chose a leaving variable by performing the minimum ratio test on the chosen column and the right-hand-side (RHS) column. We pivot on the element at the entering column and leaving row and this transforms the tableau into a new tableau that represents the new basic feasible solution.

EXAMPLE 5.10. Again, consider the toy maker problem. We will execute the simplex algorithm using the tableau method. Our problem in standard form is given as:

$$\left\{ \begin{array}{l} \max \quad z(x_1, x_2) = 7x_1 + 6x_2 \\ \text{s.t.} \quad 3x_1 + x_2 + s_1 = 120 \\ \quad \quad x_1 + 2x_2 + s_2 = 160 \\ \quad \quad x_1 + s_3 = 35 \\ \quad \quad x_1, x_2, s_1, s_2, s_3 \geq 0 \end{array} \right.$$

We can assume our initial basic feasible solution has s_1 , s_2 and s_3 as basic variables and x_1 and x_2 as non-basic variables. Thus our initial tableau is simply:

$$(5.28) \quad \begin{array}{c|cccccc|c} & z & x_1 & x_2 & s_1 & s_2 & s_3 & \text{RHS} \\ \hline z & 1 & -7 & -6 & 0 & 0 & 0 & 0 \\ s_1 & 0 & 3 & 1 & 1 & 0 & 0 & 120 \\ s_2 & 0 & 1 & 2 & 0 & 1 & 0 & 160 \\ s_3 & 0 & 1 & 0 & 0 & 0 & 1 & 35 \end{array}$$

Note that the columns have been swapped so that the identity matrix is divided and $\mathbf{B}^{-1} \mathbf{N}$ is located in columns 2 and 3. This is because of our choice of basic variables. The reduced cost vector is in Row 0.

Using this information, we can see that either x_1 or x_2 can enter. We can compute the minimum ratio test (MRT) next to the RHS column. If we chose x_2 as the entering variable, then the MRT tells us s_2 will leave. We put a box around the element on which we will pivot:

$$(5.29) \quad \begin{array}{c} z \\ s_1 \\ s_2 \\ s_3 \end{array} \left[\begin{array}{c|cccccc|c} z & x_1 & x_2 & s_1 & s_2 & s_3 & \text{RHS} \\ \hline 1 & -7 & -6 & 0 & 0 & 0 & 0 \\ 0 & 3 & 1 & 1 & 0 & 0 & 120 \\ 0 & 1 & \boxed{2} & 0 & 1 & 0 & 160 \\ 0 & 1 & 0 & 0 & 0 & 1 & 35 \end{array} \right] \quad \begin{array}{c} \text{MRT } (x_2) \\ 120 \\ 80 \\ - \end{array}$$

If we pivot on this element, then we transform the column corresponding to x_2 into the identity column:

$$(5.30) \quad \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

This process will correctly compute the new reduced costs and $\mathbf{B}^{-1}$ matrix as well as the new cost information. The new tableau becomes:

$$(5.31) \quad \begin{array}{c} z \\ s_1 \\ x_2 \\ s_3 \end{array} \left[\begin{array}{c|cccccc|c} z & x_1 & x_2 & s_1 & s_2 & s_3 & \text{RHS} \\ \hline 1 & -4 & 0 & 0 & 3 & 0 & 480 \\ 0 & 2.5 & 0 & 1 & -0.5 & 0 & 40 \\ 0 & 0.5 & 1 & 0 & 0.5 & 0 & 80 \\ 0 & 1 & 0 & 0 & 0 & 1 & 35 \end{array} \right]$$

We can see that x_1 is a valid entering variable, as it has a negative reduced cost (-4). We can again place the minimum ratio test values on the right-hand-side of the matrix to obtain:

$$(5.32) \quad \begin{array}{c} z \\ s_1 \\ x_2 \\ s_3 \end{array} \left[\begin{array}{c|cccccc|c} z & x_1 & x_2 & s_1 & s_2 & s_3 & \text{RHS} \\ \hline 1 & -4 & 0 & 0 & 3 & 0 & 480 \\ 0 & \boxed{2.5} & 0 & 1 & -0.5 & 0 & 40 \\ 0 & 0.5 & 1 & 0 & 0.5 & 0 & 80 \\ 0 & 1 & 0 & 0 & 0 & 1 & 35 \end{array} \right] \quad \begin{array}{c} \text{MRT } (x_1) \\ 16 \\ 160 \\ 35 \end{array}$$

We now pivot on the element we have boxed to obtain the new tableau¹:

$$(5.33) \quad \begin{array}{c} z \\ x_1 \\ x_2 \\ s_3 \end{array} \left[\begin{array}{c|cccccc|c} z & x_1 & x_2 & s_1 & s_2 & s_3 & \text{RHS} \\ \hline 1 & 0 & 0 & 1.6 & 2.2 & 0 & 544 \\ 0 & 1 & 0 & 0.4 & -0.2 & 0 & 16 \\ 0 & 0 & 1 & -0.2 & 0.6 & 0 & 72 \\ 0 & 0 & 0 & -0.4 & 0.2 & 1 & 19 \end{array} \right]$$

All the reduced costs of the non-basic variables (s_1 and s_2) are positive and so this is the optimal solution to the linear programming problem. We can also see that this solution agrees with our previous computations on the Toy Maker Problem.

Unit 7

Transportation and Assignment problems

Objectives

By the end of this unit you will be able to:

- formulate special linear programming problems using the transportation model.
- define a balanced transportation problem
- develop an initial solution of a transportation problem using the Northwest Corner Rule
- use the Stepping Stone method to find an optimal solution of a transportation problem
- formulate special linear programming problems using the assignment model
- solve assignment problems with the Hungarian method.

Introduction

In this unit we extend the theory of linear programming to two special linear programming problems, the **Transportation** and **Assignment Problems**. Both of these problems can be solved by the simplex algorithm, but the process would result in very large simplex tableaux and numerous simplex iterations.

Because of the special characteristics of each problem, however, alternative solution methods requiring significantly less mathematical manipulation have been developed.

The Transportation problem

The general transportation problem is concerned with determining an optimal strategy for distributing a commodity from a group of supply centres, such as factories, called *sources*, to various receiving centers, such as warehouses, called *destinations*, in such a way as to minimise total distribution costs.

Each source is able to supply a fixed number of units of the product, usually called the *capacity* or *availability*, and each destination has a fixed demand, often called the *requirement*.

Transportation models can also be used when a firm is trying to decide where to locate a new facility. Good financial decisions concerning facility location also attempt to minimize total transportation and production costs for the entire system.

4.3.1 Setting up a Transportation problem

To illustrate how to set up a transportation problem we consider the following example;

Example 4.1

A concrete company transports concrete from three plants, 1, 2 and 3, to three construction sites, A, B and C.

The plants are able to supply the following numbers of tons per week:

<i>Plant</i>	<i>Supply (capacity)</i>
<i>1</i>	<i>300</i>
<i>2</i>	<i>300</i>
<i>3</i>	<i>100</i>

The requirements of the sites, in number of tons per week, are:

<i>Construction site</i>	<i>Demand (requirement)</i>
<i>A</i>	<i>200</i>
<i>B</i>	<i>200</i>
<i>C</i>	<i>300</i>

The cost of transporting 1 ton of concrete from each plant to each site is shown in the figure 8 in Emalangen per ton.

For computational purposes it is convenient to put all the above information into a table, as in the simplex method. In this table each row represents a source and each column represents a destination.

		<i>Sites</i>			<i>Supply (Availability)</i>
<i>Plants</i>	<i>To</i>	<i>A</i>	<i>B</i>	<i>C</i>	
	<i>From</i>				
<i>1</i>		<i>4</i>	<i>3</i>	<i>8</i>	<i>300</i>
<i>2</i>		<i>7</i>	<i>5</i>	<i>9</i>	<i>300</i>
<i>3</i>		<i>4</i>	<i>5</i>	<i>5</i>	<i>100</i>
	<i>Demand (requirement)</i>	<i>200</i>	<i>200</i>	<i>300</i>	

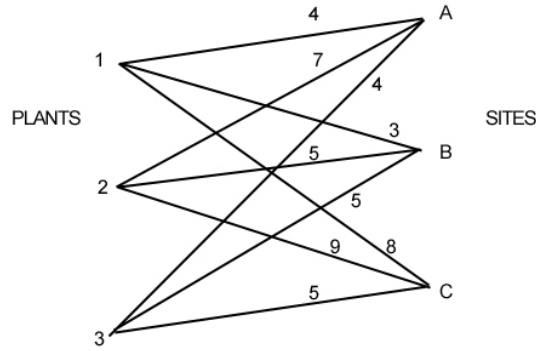


Figure 8: Constructing a transportation problem

4.3.2 Mathematical model of a transportation problem

Before we discuss the solution of transportation problems we will introduce the notation used to describe the transportation problem and show that it can be formulated as a linear programming problem.

We use the following notation;

- x_{ij} = the number of units to be distributed from
source i to destination j
($i = 1, 2, \dots, m; j = 1, 2, \dots, n$);
- s_i = supply from source i ;
- d_j = demand at destination j ;
- c_{ij} = cost per unit distributed from
source i to destination j

With respect to Example 4.1 the decision variables x_{ij} are the numbers of tons transported from plant i (where $i = 1, 2, 3$) to each site j (where $j = A, B, C$)

A basic assumption is that the distribution costs of units from source i to destination j is directly proportional to the number of units distributed. A typical **cost and requirements table** has the form shown on Table 4.

Let Z be total distribution costs from all the m sources to the n destinations. In example 4.1 each term in the objective function Z represents the total cost of tonnage transported on one route. For example, in the route $2 \rightarrow C$, the term is $9x_{2C}$, that is:

$$(\text{Cost per ton} = 9) \times (\text{number of tons transported} = x_{2C})$$

	<i>Destination</i>				<i>Supply</i>
	1	2	...	n	
1	c_{11}	c_{12}	...	c_{1n}	s_1
2	c_{21}	c_{22}	...	c_{2n}	s_2
<i>Source</i> $\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
m	c_{m1}	c_{m2}	...	c_{mn}	s_m
<i>Demand</i>	d_1	d_2	...	d_n	

Table 4: *Cost and requirements table*

Hence the objective function is:

$$\begin{aligned}
Z &= 4x_{1A} + 3x_{1B} + 8x_{1C} \\
&+ 7x_{2A} + 5x_{2B} + 9x_{2C} \\
&+ 4x_{3A} + 5x_{3B} + 5x_{3C}
\end{aligned}$$

Notice that in this problem the total supply is $300 + 300 + 200 = 700$ and the total demand is $200 + 200 + 300 = 700$. Thus

$$\text{Total supply} = \text{total demand}.$$

In mathematical form this expressed as

$$\sum_{i=1}^m s_i = \sum_{j=1}^n d_j \quad (47)$$

This is called a **balanced problem**. In this unit our discussion shall be restricted to the balanced problems.

In a balanced problem all the products that can be supplied are used to meet the demand. There are no slacks and so all constraints are *equalities* rather than *inequalities* as was the case in the previous unit.

The formulation of this problem as a linear programming problem is presented as

$$\text{Minimise } Z = \sum_{i=1}^m \sum_{j=1}^n c_{ij} x_{ij}, \quad (48)$$

subject to

$$\sum_{j=1}^n x_{ij} = s_i, \quad \text{for } i = 1, 2, \dots, m \quad (49)$$

$$\sum_{i=1}^m x_{ij} = d_j, \quad \text{for } j = 1, 2, \dots, n \quad (50)$$

and

$$x_{ij} \geq 0, \text{ for all } i \text{ and } j.$$

Any linear programming problem that fits this special formulation is of the transportation type, regardless of its physical context. For many applications, the supply and demand quantities in the model will have integer values and implementation will require that the distribution quantities also be integers. Fortunately, the unit coefficients of the unknown variables in the constraints guarantee an optimal solution with only integer values.

4.3.3 Initial solution - Northwest Corner Rule

The initial basic feasible solution can be obtained by using one of several methods. We will consider only the **North West corner rule** of developing an initial solution. Other methods can be found in standard texts on linear programming.

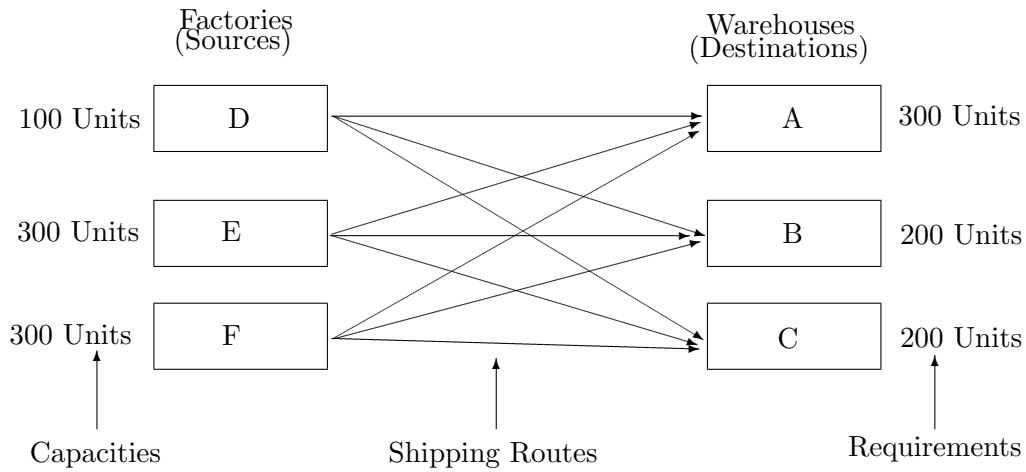
The procedure for constructing an initial basic feasible solution selects the basic variables one at a time. The North West corner rule begins with an allocation at the top left-hand corner of the tableau and proceeds systematically along either a row or a column and make allocations to subsequent cells until the bottom right-hand corner is reached, by which time enough allocations will have been made to constitute an initial solution.

The procedure for constructing an initial solution using the North West Corner rule is as follows:

NORTH WEST CORNER RULE

1. Start by selecting the cell in the most “North-West” corner of the table.
2. Assign the maximum amount to this cell that is allowable based on the requirements and the capacity constraints.
3. Exhaust the capacity from each row before moving down to another row.
4. Exhaust the requirement from each column before moving right to another column.
5. Check to make sure that the capacity and requirements are met.

Let us begin with an example dealing with Executive Furniture corporation, which manufactures office desks at three locations: D, E and F. The firm distributes the desks through regional warehouses located in A, B and C (see the Network format diagram below)



It is assumed that the production costs per desk are identical at each factory. The only relevant costs are those of shipping from each source to each destination. The costs are shown in Table 5

From \ To	A	B	C
D	\$5	\$4	\$3
E	\$8	\$4	\$3
F	\$9	\$7	\$5

Table 5: Transportation Costs per desk for Executive Furniture Corp.

We proceed to construct a transportation table and label its various components as show in Table 6.

We can now use the Northwest corner rule to find an initial feasible solution to the problem. We start in the upper left hand cell and allocate units to shipping routes as follows:

From \ To	A	B	C	Capacity
D	5	4	3	100
E	8	4	3	300
F	9	7	5	300
Requirements	300	200	200	700

Table 6: Transportation Table for Executive Furniture Corporation

1. Exhaust the supply (factory capacity) of each row before moving down to the next row.
2. Exhaust the demand (warehouse) requirements of each column before moving to the next column to the right.
3. Check that all supply and demand requirements are met.

The initial shipping assignments are given in Table 7

From \ To	A	B	C	Factory Capacity
D	100			100
E	200	100		300
F		100	200	300
Warehouse Requirements	300	200	200	700

Table 7: Initial Solution of the North West corner Rule

This initial solution can also be presented together with the costs per unit as shown in the Table 8.

We can compute the cost of this shipping assignment as follows;

Therefore, the initial feasible solution for this problem is \$4200.

Assignment Problem

The **assignment problem** refers to the class of linear programming problems that involve determining the most efficient assignment of

- people to projects
- salespeople to territories
- contracts to bidders
- jobs to machines, etc.

The objective is most often to minimize total costs or total time of performing the tasks at hand.

One important characteristic of assignment problems is that only one job or worker is assigned to one machine or project. An example is the problem of a taxi company with 4 taxis and 4 passengers. Which taxi should collect which passenger in order to minimize costs?

Each assignment problem has associated with it a table, or matrix. Generally, the rows contain the objects or people we wish to assign, and the columns comprise the tasks or things we want them assigned to. The numbers in the table are the costs associated with each particular assignment.

An assignment problem can be viewed as a transportation problem in which

- the capacity from each source (or person to be assigned) is 1 and
- the demand at each destination (or job to be done) is 1.

As an illustration of the assignment problem, let us consider the case of a Fix-It-Shop, which has just received three new rush projects to repair: (1) a radio, (2) a toaster oven, and (3) a broken coffee table. Three repair persons, each with different talents and abilities, are available to do the jobs. The owner of the shop estimates what it will cost in wages to assign each of the workers to each of the three projects. The costs which are shown in Table 18 differ because the owner believes that each worker will differ in speed and skill on these quite varied jobs.

Table 19 summarizes all six assignment options. The table also shows that the least-cost solution would be to assign Cooper to project 1, Brown to project 2, and Adams to project 3, at a total cost of \$25.

The owner's objective is to assign the three projects to the workers in a way that will result in the lowest cost to the shop. Note that the assignment of people to projects must be on a one-to-one basis; each project will be assigned exclusively to one worker only.

PERSON	PROJECT		
	1	2	3
Adams	\$11	\$14	\$6
Brown	8	10	11
Cooper	9	12	7

Table 18: Repair costs of the Fix-It-Shop assignment problem

PROJECT ASSIGNMENT			LABOUR COSTS (\$)	TOTAL COSTS (\$)
1	2	3		
Adams	Brown	Cooper	11 + 10 + 7	28
Adams	Cooper	Brown	11 + 12 + 11	34
Brown	Adams	Cooper	8 + 14 + 7	29
Brown	Cooper	Adams	8 + 12 + 6	26
Cooper	Adams	Brown	9 + 14 + 11	34
Cooper	Brown	Adams	9 + 10 + 6	25

Table 19: Assignment alternatives and Costs of Fix-It-Shop assignment problem

Special algorithms exist to solve assignment problems. The most common is probably the **Hungarian** solution method. The Hungarian method of assignment provides us with an efficient means of finding the optimal solution without having to make a direct comparison of every assignment option. It operates on a principle of matrix reduction, which means that by subtracting and adding appropriate numbers in the cost table or matrix, we can reduce the problem to a matrix of *opportunity costs*. Opportunity costs show the relative penalties associated with assigning any person to a project as opposed to making the best or least-cost assignment. We would like to make assignments such that the opportunity cost for each assignment is zero.

The steps involved in the Hungarian method are outlined below.

THE HUNGARIAN METHOD

1. *Find the opportunity cost table by*
 - (a) Subtracting the smallest number in each row of the original cost table or matrix from every number in that row.
 - (b) Then subtracting the smallest number in each column of the table obtained in part (a) from every number in that column.
2. *Test the table resulting from step 1 to see whether an optimal assignment can be made.* The procedure is to draw the minimum number of vertical and horizontal straight lines necessary to cover all zeros in the table. If the number of lines equals either the number of rows or columns, an optimal assignment can be made. If the number of lines is less than the number of rows or columns, we proceed to step 3.
3. *Revise the present opportunity cost table.* This is done by subtracting the smallest number not covered by a line from every other uncovered number. This same smallest number is also added to any number(s) lying at the intersection of the horizontal and vertical lines. We then return to step 2 and continue the cycle until an optimal assignment is possible.

Let us now apply the three steps to the Fix-It-Shop assignment example.

The original cost table for the problem is given in Table 20

PERSON	PROJECT		
	1	2	3
Adams	11	14	6
Brown	8	10	11
Cooper	9	12	7

Table 20: Initial Table

PERSON	PROJECT		
	1	2	3
Adams	5	8	0
Brown	0	2	3
Cooper	2	5	0

Table 21: Row reduction (part a)

After the row reduction (Step 1 part a) we get the cost Table 21.

Taking the costs in Table 21 and subtracting the the smallest number in each column from each number in that column results in the total opportunity costs given in Table 22. This step is the column reduction of Step 1 part (b)

If we draw vertical and horizontal straight lines (Step 2) to cover all the zeros in Table 22 we get Table 23. Since the number of lines is less than the number of rows or columns an optimal assignment cannot be made.

Since Table 23 doesn't give an optimal solution we revise the table. This is accomplished by subtracting the smallest number not covered by a line from all numbers not covered by

PERSON	PROJECT		
	1	2	3
Adams	5	6	0
Brown	0	0	3
Cooper	2	3	0

Table 22: Column Reduction (Step 1 part b)

PERSON	PROJECT		
	1	2	3
Adams	5	6	~~0~~
Brown	~~0~~	~~0~~	~~3~~
Cooper	2	3	~~0~~

Table 23: Testing for an optimal solution

a straight line. This same smallest number is then added to every number (including zeros) *lying in the intersection* on any two lines. The smallest uncovered number in Table 23 is 2, so this value is subtracted from each of the four uncovered numbers. A 2 is also added to the number that is covered by the intersecting horizontal and vertical lines. The results of this step are shown in Table 24

To test now for an optimal assignment, we return to Step 2 and find the minimum number of lines necessary to cover all zeros in the revised opportunity cost table. Because it requires three lines to cover the zeros (see Table 25), an optimal assignment can be made.

PERSON	PROJECT		
	1	2	3
Adams	3	4	0
Brown	0	0	5
Cooper	0	1	0

Table 24: Revised opportunity cost table

PERSON	PROJECT		
	1	2	3
Adams	~~3~~	4	~~0~~
Brown	~~0~~	~~0~~	~~5~~
Cooper	~~0~~	1	~~0~~

Table 25: Optimality test on the revised table

Finally, we make the allocation. Note that only one assignment will be made from each row or column. We use this fact to proceed to making the final allocation as follows:

- Find a row or column with only one zero cell.
- Make the assignment corresponding to that zero cell.
- Eliminate that row and column from the table.
- Continue until all the assignments have been made.

For our Fix-It-Shop problem these steps are summarized in Table 26.

FIRST ASSIGNMENT				SECOND ASSIGNMENT				THIRD ASSIGNMENT			
	1	2	3		1	2	3		1	2	3
Adams	3	4	0	Adams	~~3~~	~~4~~	~~0~~	Adams	~~3~~	~~4~~	~~0~~
Brown	0	0	5	Brown	0	0	5	Brown	0	0	5
Cooper	0	1	0	Cooper	0	1	0	Cooper	0	1	0

Table 26: Making the final assignment

To interpret the table we recall that our objective was to minimize costs, there is only one assignment that Adams can go to where the opportunity costs are \$0. That is to assign Adams Project 3. If Adams gets assigned to Project 3, then there is only one project left where the opportunity cost is \$0 for Cooper. Therefore Cooper gets assigned to Project 1. This leaves Brown being assigned to Project 2, where the opportunity costs are \$0.

The optimal allocation is to assign Adams to Project 3, Brown to Project 2, and Cooper to Project 1. The total labour cost of this assignment are computed from the original cost table (see Table 18). They are as follows:

ASSIGNMENT	COST (\$)
Adams to Project 3	6
Brown to Project 2	10
Cooper to Project 1	9
Total cost	25

Unit 8

Introduction to CPM / PERT Techniques

Applications of CPM / PERT

Basic Steps in PERT / CPM

Frame work of PERT/CPM

Network Diagram Representation

Rules for Drawing Network Diagrams

Common Errors in Drawing Networks

Advantages and Disadvantages

Critical Path in Network Analysis

Introduction to CPM / PERT Techniques

CPM/PERT or Network Analysis as the technique is sometimes called, developed along two parallel streams, one industrial and the other military.

CPM (Critical Path Method) was the discovery of M.R.Walker of E.I.Du Pont de Nemours & Co. and J.E.Kelly of Remington Rand, circa 1957. The computation was designed for the UNIVAC-I computer. The first test was made in 1958, when CPM was applied to the construction of a new chemical plant. In March 1959, the method was applied to maintenance shut-down at the Du Pont works in Louisville, Kentucky. Unproductive time was reduced from 125 to 93 hours.

PERT (Project Evaluation and Review Technique) was devised in 1958 for the POLARIS missile program by the Program Evaluation Branch of the Special Projects office of the U.S.Navy, helped by the Lockheed Missile Systems division and the Consultant firm of Booz-Allen & Hamilton. The calculations were so arranged so that they could be carried out on the IBM Naval Ordinance Research Computer (NORC) at Dahlgren, Virginia.

The methods are essentially **network-oriented techniques** using the same principle. PERT and CPM are basically time-oriented methods in the sense that they both lead to determination of a time schedule for the project. The significant difference between two approaches is that the time estimates for the different activities in CPM were assumed to be **deterministic** while in PERT these are described **probabilistically**. These techniques are referred as **project scheduling** techniques.

In **CPM** activities are shown as a network of precedence relationships using activity-on-node network construction

- Single estimate of activity time
- Deterministic activity times

USED IN: Production management - for the jobs of repetitive in nature where the activity time estimates can be predicted with considerable certainty due to the existence of past experience.

In **PERT** activities are shown as a network of precedence relationships using activity-on-arrow network construction

- Multiple time estimates
- Probabilistic activity times

USED IN: Project management - for non-repetitive jobs (research and development work), where the time and cost estimates tend to be quite uncertain. This technique uses probabilistic time estimates.

Benefits of PERT/CPM

- Useful at many stages of project management
- Mathematically simple

- Give critical path and slack time
- Provide project documentation
- Useful in monitoring costs

Limitations of PERT/CPM

- Clearly defined, independent and stable activities
- Specified precedence relationships
- Over emphasis on critical paths

Applications of CPM / PERT

These methods have been applied to a wide variety of problems in industries and have found acceptance even in government organizations. These include

- Construction of a dam or a canal system in a region
- Construction of a building or highway
- Maintenance or overhaul of airplanes or oil refinery
- Space flight
- Cost control of a project using PERT / COST
- Designing a prototype of a machine
- Development of supersonic planes

Basic Steps in PERT / CPM

Project scheduling by PERT / CPM consists of four main steps

1. Planning

- The planning phase is started by splitting the total project in to small projects. These smaller projects in turn are divided into activities and are analyzed by the department or section.
- The relationship of each activity with respect to other activities are defined and established and the corresponding responsibilities and the authority are also stated.
- Thus the possibility of overlooking any task necessary for the completion of the project is reduced substantially.

2. Scheduling

- The ultimate objective of the scheduling phase is to prepare a time chart showing the start and finish times for each activity as well as its relationship to other activities of the project.
- Moreover the schedule must pinpoint the critical path activities which require special attention if the project is to be completed in time.
- For non-critical activities, the schedule must show the amount of slack or float times which can be used advantageously when such activities are delayed or when limited resources are to be utilized effectively.

3. Allocation of resources

- Allocation of resources is performed to achieve the desired objective. A resource is a physical variable such as labour, finance, equipment and space which will impose a limitation on time for the project.
- When resources are limited and conflicting, demands are made for the same type of resources a systematic method for allocation of resources become essential.
- Resource allocation usually incurs a compromise and the choice of this compromise depends on the judgment of managers.

4. Controlling

- The final phase in project management is controlling. Critical path methods facilitate the application of the principle of management by expectation to identify areas that are critical to the completion of the project.
- By having progress reports from time to time and updating the network continuously, a better financial as well as technical control over the project is exercised.
- Arrow diagrams and time charts are used for making periodic progress reports. If required, a new course of action is determined for the remaining portion of the project.

The Framework for PERT and CPM

Essentially, there are six steps which are common to both the techniques. The procedure is listed below:

- I. Define the Project and all of its significant activities or tasks. The Project (made up of several tasks) should have only a single start activity and a single finish activity.
- II. Develop the relationships among the activities. Decide which activities must precede and which must follow others.
- III. Draw the "Network" connecting all the activities. Each Activity should have unique event numbers. Dummy arrows are used where required to avoid giving the same numbering to two activities.
- IV. Assign time and/or cost estimates to each activity
- V. Compute the longest time path through the network. This is called the critical path.

VI. Use the Network to help plan, schedule, and monitor and control the project.

The Key Concept used by CPM/PERT is that a small set of activities, which make up the longest path through the activity network control the entire project. If these "critical" activities could be identified and assigned to responsible persons, management resources could be optimally used by concentrating on the few activities which determine the fate of the entire project.

Non-critical activities can be replanned, rescheduled and resources for them can be reallocated flexibly, without affecting the whole project.

Five useful questions to ask when preparing an activity network are:

- Is this a Start Activity?
- Is this a Finish Activity?
- What Activity Precedes this?
- What Activity Follows this?
- What Activity is Concurrent with this?

Network Diagram Representation

In a network representation of a project certain definitions are used

1. Activity

Any individual operation which utilizes resources and has an end and a beginning is called activity. An arrow is commonly used to represent an activity with its head indicating the direction of progress in the project. These are classified into four categories

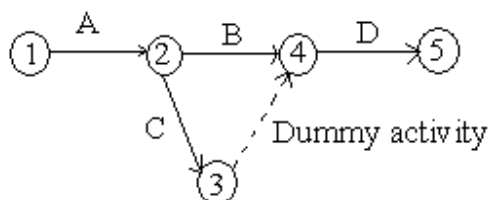
1. **Predecessor activity** – Activities that must be completed immediately prior to the start of another activity are called predecessor activities.

2. **Successor activity** – Activities that cannot be started until one or more of other activities are completed but immediately succeed them are called successor activities.
3. **Concurrent activity** – Activities which can be accomplished concurrently are known as concurrent activities. It may be noted that an activity can be a predecessor or a successor to an event or it may be concurrent with one or more of other activities.
4. **Dummy activity** – An activity which does not consume any kind of resource but merely depicts the technological dependence is called a dummy activity.

The dummy activity is inserted in the network to clarify the activity pattern in the following two situations

- To make activities with common starting and finishing points distinguishable
- To identify and maintain the proper precedence relationship between activities that is not connected by events.

For example, consider a situation where A and B are concurrent activities. C is dependent on A and D is dependent on A and B both. Such a situation can be handled by using a dummy activity as shown in the figure.



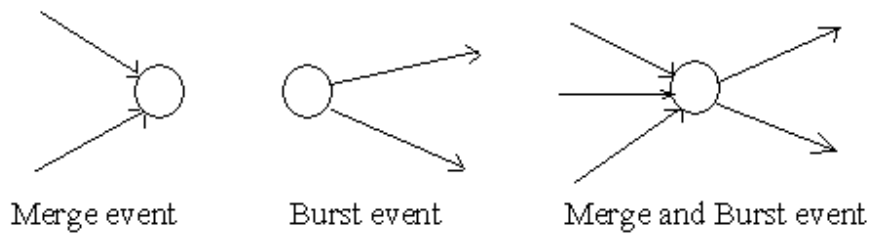
2. Event

An event represents a point in time signifying the completion of some activities and the beginning of new ones. This is usually represented by a circle in a network which is also called a node or connector.

The events are classified in to three categories

1. **Merge event** – When more than one activity comes and joins an event such an event is known as merge event.

2. **Burst event** – When more than one activity leaves an event such an event is known as burst event.
3. **Merge and Burst event** – An activity may be merge and burst event at the same time as with respect to some activities it can be a merge event and with respect to some other activities it may be a burst event.



3. Sequencing

The first prerequisite in the development of network is to maintain the precedence relationships. In order to make a network, the following points should be taken into considerations

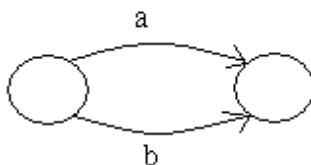
- What job or jobs precede it?
- What job or jobs could run concurrently?
- What job or jobs follow it?
- What controls the start and finish of a job?

Since all further calculations are based on the network, it is necessary that a network be drawn with full care.

Rules for Drawing Network Diagram

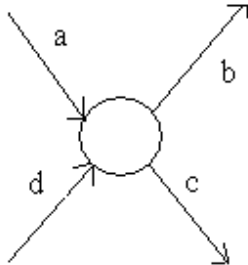
Rule 1

Each activity is represented by one and only one arrow in the network



Rule 2

No two activities can be identified by the same end events



Rule 3

In order to ensure the correct precedence relationship in the arrow diagram, following questions must be checked whenever any activity is added to the network

- What activity must be completed immediately before this activity can start?
- What activities must follow this activity?
- What activities must occur simultaneously with this activity?

In case of large network, it is essential that certain good habits be practiced to draw an easy to follow network

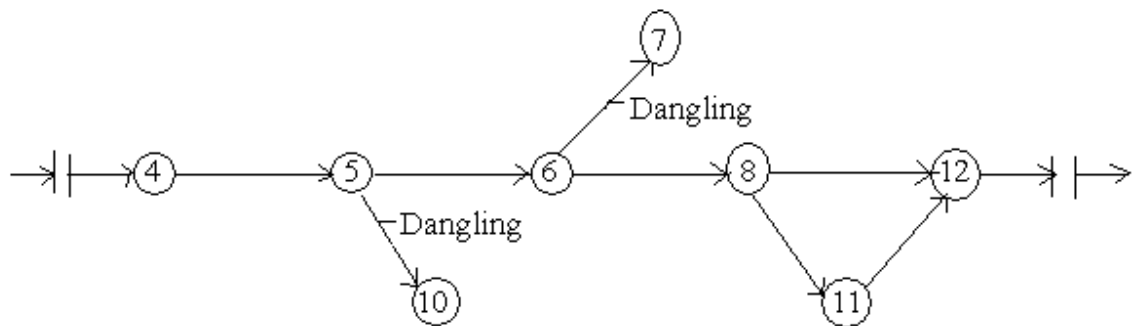
- Try to avoid arrows which cross each other
- Use straight arrows
- Do not attempt to represent duration of activity by its arrow length
- Use arrows from left to right. Avoid mixing two directions, vertical and standing arrows may be used if necessary.
- Use dummies freely in rough draft but final network should not have any redundant dummies.
- The network has only one entry point called start event and one point of emergence called the end event.

Common Errors in Drawing Networks

The three types of errors are most commonly observed in drawing network diagrams

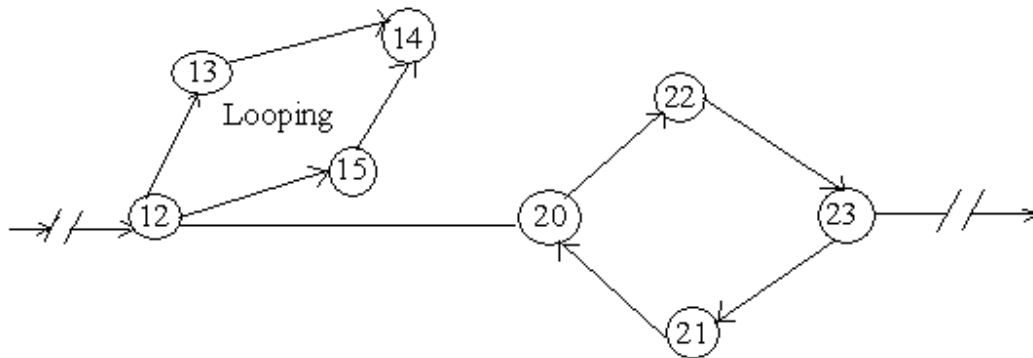
1. Dangling

To disconnect an activity before the completion of all activities in a network diagram is known as dangling. As shown in the figure activities (5 – 10) and (6 – 7) are not the last activities in the network. So the diagram is wrong and indicates the error of dangling



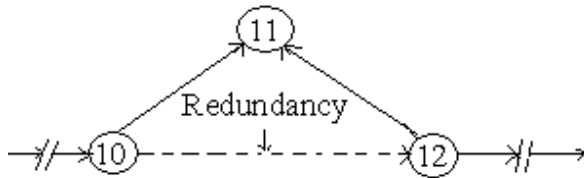
2. Looping or Cycling

Looping error is also known as cycling error in a network diagram. Drawing an endless loop in a network is known as error of looping as shown in the following figure.



3. Redundancy

Unnecessarily inserting the dummy activity in network logic is known as the error of redundancy as shown in the following diagram



Advantages and Disadvantages

PERT/CPM has the following advantages

- A PERT/CPM chart explicitly defines and makes visible dependencies (precedence relationships) between the elements,
- PERT/CPM facilitates identification of the critical path and makes this visible,
- PERT/CPM facilitates identification of early start, late start, and slack for each activity,
- PERT/CPM provides for potentially reduced project duration due to better understanding of dependencies leading to improved overlapping of activities and tasks where feasible.

PERT/CPM has the following disadvantages:

- There can be potentially hundreds or thousands of activities and individual dependency relationships,
- The network charts tend to be large and unwieldy requiring several pages to print and requiring special size paper,
- The lack of a timeframe on most PERT/CPM charts makes it harder to show status although colours can help (e.g., specific colour for completed nodes),
- When the PERT/CPM charts become unwieldy, they are no longer used to manage the project.

Critical Path in Network Analysis

Basic Scheduling Computations

The notations used are

(i, j) = Activity with tail event i and head event j

E_i = Earliest occurrence time of event i

L_j = Latest allowable occurrence time of event j

D_{ij} = Estimated completion time of activity (i, j)

$(Es)_{ij}$ = Earliest starting time of activity (i, j)

$(Ef)_{ij}$ = Earliest finishing time of activity (i, j)

$(Ls)_{ij}$ = Latest starting time of activity (i, j)

$(Lf)_{ij}$ = Latest finishing time of activity (i, j)

The procedure is as follows

1. Determination of Earliest time (E_j): Forward Pass computation

- **Step 1**

The computation begins from the start node and move towards the end node. For easiness, the forward pass computation starts by assuming the earliest occurrence time of zero for the initial project event.

- **Step 2**

- i. Earliest starting time of activity (i, j) is the earliest event time of the tail end event i.e. $(Es)_{ij} = E_i$
- ii. Earliest finish time of activity (i, j) is the earliest starting time + the activity time i.e. $(Ef)_{ij} = (Es)_{ij} + D_{ij}$ or $(Ef)_{ij} = E_i + D_{ij}$

- iii. Earliest event time for event j is the maximum of the earliest finish times of all activities ending in to that event i.e. $E_j = \max [(Ef)_{ij}]$ for all immediate predecessor of (i, j) or $E_j = \max [E_i + D_{ij}]$

2. Backward Pass computation (for latest allowable time)

- **Step 1**

For ending event assume $E = L$. Remember that all E's have been computed by forward pass computations.

- **Step 2**

Latest finish time for activity (i, j) is equal to the latest event time of event j i.e. $(Lf)_{ij} = L_j$

- **Step 3**

Latest starting time of activity (i, j) = the latest completion time of (i, j) – the activity time or $(Ls)_{ij} = (Lf)_{ij} - D_{ij}$ or $(Ls)_{ij} = L_j - D_{ij}$

- **Step 4**

Latest event time for event 'i' is the minimum of the latest start time of all activities originating from that event i.e. $L_i = \min [(Ls)_{ij}]$ for all immediate successor of (i, j) $= \min [(Lf)_{ij} - D_{ij}] = \min [L_j - D_{ij}]$

3. Determination of floats and slack times

There are three kinds of floats

- **Total float** – The amount of time by which the completion of an activity could be delayed beyond the earliest expected completion time without affecting the overall project duration time.

Mathematically

$$(Tf)_{ij} = (\text{Latest start} - \text{Earliest start}) \text{ for activity } (i - j)$$

$$(Tf)_{ij} = (Ls)_{ij} - (Es)_{ij} \text{ or } (Tf)_{ij} = (L_j - D_{ij}) - E_i$$

- **Free float** – The time by which the completion of an activity can be delayed beyond the earliest finish time without affecting the earliest start of a subsequent activity.

Mathematically

$$(Ff)_{ij} = (\text{Earliest time for event } j - \text{Earliest time for event } i) - \text{Activity time for } (i,$$

j)

$$(Ff)_{ij} = (E_j - E_i) - D_{ij}$$

- **Independent float** – The amount of time by which the start of an activity can be delayed without effecting the earliest start time of any immediately following activities, assuming that the preceding activity has finished at its latest finish time.

Mathematically

$$(If)_{ij} = (E_j - L_i) - D_{ij}$$

The negative independent float is always taken as zero.

- **Event slack** - It is defined as the difference between the latest event and earliest event times.

Mathematically

$$\text{Head event slack} = L_j - E_j, \text{ Tail event slack} = L_i - E_i$$

4. Determination of critical path

- **Critical event** – The events with zero slack times are called critical events. In other words the event i is said to be critical if $E_i = L_i$
- **Critical activity** – The activities with zero total float are known as critical activities. In other words an activity is said to be critical if a delay in its start will cause a further delay in the completion date of the entire project.
- **Critical path** – The sequence of critical activities in a network is called critical path. The critical path is the longest path in the network from the starting event to ending event and defines the minimum time required to complete the project.

Exercise

1. What is PERT and CPM?
2. What are the advantages of using PERT/CPM?
3. Mention the applications of PERT/CPM
4. Explain the following terms
 - a. Earliest time
 - b. Latest time
 - c. Total activity slack
 - d. Event slack
 - e. Critical path
5. Explain the CPM in network analysis.
6. What are the rules for drawing network diagram? Also mention the common errors that occur in drawing networks.
7. What is the difference between PERT and CPM/

8. What are the uses of PERT and CPM?
9. Explain the basic steps in PERT/CPM techniques.
10. Write the framework of PERT/CPM.

Worked Examples on CPM

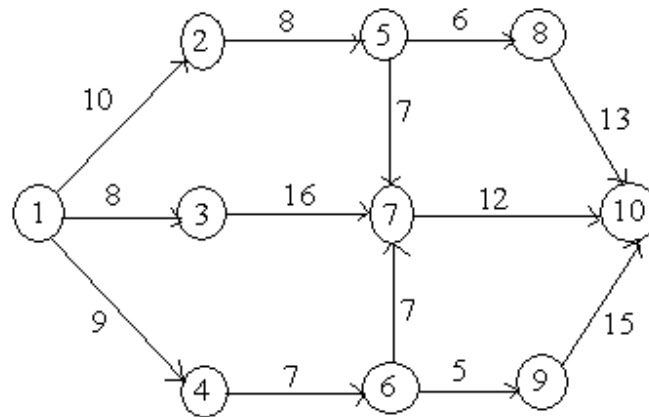
PERT

Worked Examples

Worked Examples on CPM

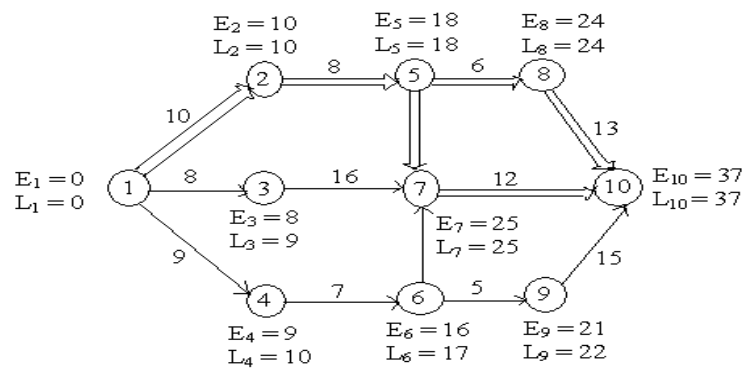
Example 1

Determine the early start and late start in respect of all node points and identify critical path for the following network.



Solution

Calculation of E and L for each node is shown in the network



Activity(i, j)	Normal Time (D _{ij})	Earliest Time		Latest Time		Float Time (L _i - D _{ij}) - E _i
		Start (E _i)	Finish (E _i + D _{ij})	Start (L _i - D _{ij})	Finish (L _i)	
(1, 2)	10	0	10	0	10	0
(1, 3)	8	0	8	1	9	1
(1, 4)	9	0	9	1	10	1
(2, 5)	8	10	18	10	18	0
(4, 6)	7	9	16	10	17	1

(3, 7)	16	8	24	9	25	1
(5, 7)	7	18	25	18	25	0
(6, 7)	7	16	23	18	25	2
(5, 8)	6	18	24	18	24	0
(6, 9)	5	16	21	17	22	1
(7, 10)	12	25	37	25	37	0
(8, 10)	13	24	37	24	37	0
(9, 10)	15	21	36	22	37	1

Network Analysis Table

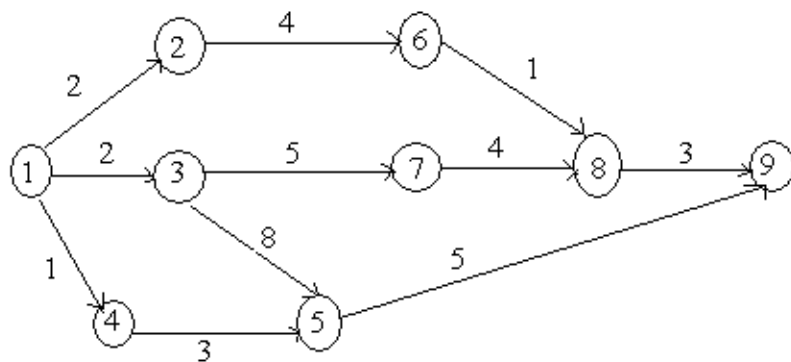
From the table, the critical nodes are (1, 2), (2, 5), (5, 7), (5, 8), (7, 10) and (8, 10)

From the table, there are two possible critical paths

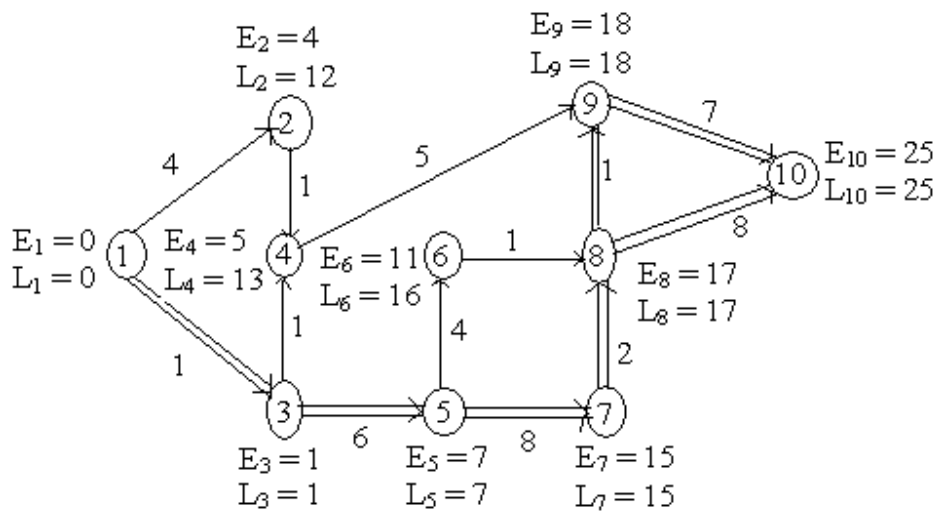
- i. $1 \rightarrow 2 \rightarrow 5 \rightarrow 8 \rightarrow 10$
- ii. $1 \rightarrow 2 \rightarrow 5 \rightarrow 7 \rightarrow 10$

Example 2

Find the critical path and calculate the slack time for the following network



Solution



The two critical paths are

- i. $1 \rightarrow 3 \rightarrow 5 \rightarrow 7 \rightarrow 8 \rightarrow 9 \rightarrow 10$
- ii. $1 \rightarrow 3 \rightarrow 5 \rightarrow 7 \rightarrow 8 \rightarrow 10$

Project Evaluation and Review Technique (PERT)

The main objective in the analysis through PERT is to find out the completion for a particular event within specified date. The PERT approach takes into account the uncertainties. The three time values are associated with each activity

1. **Optimistic time** – It is the shortest possible time in which the activity can be finished. It assumes that every thing goes very well. This is denoted by t_0 .
2. **Most likely time** – It is the estimate of the normal time the activity would take. This assumes normal delays. If a graph is plotted in the time of completion and the frequency of completion in that time period, then most likely time will represent the highest frequency of occurrence. This is denoted by t_m .
3. **Pessimistic time** – It represents the longest time the activity could take if everything goes wrong. As in optimistic estimate, this value may be such that

only one in hundred or one in twenty will take time longer than this value. This is denoted by t_p .

In PERT calculation, all values are used to obtain the percent expected value.

1. **Expected time** – It is the average time an activity will take if it were to be repeated on large number of times and is based on the assumption that the activity time follows Beta distribution, this is given by

$$t_e = (t_o + 4 t_m + t_p) / 6$$

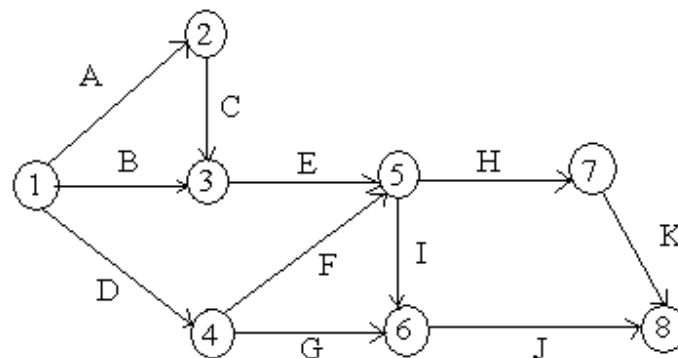
2. The **variance** for the activity is given by

$$\sigma^2 = [(t_p - t_o) / 6]^2$$

Worked Examples

Example 1

For the project



Task:	A	B	C	D	E	F	G	H	I	J	K
Least time:	4	5	8	2	4	6	8	5	3	5	6

Greatest time: 8 10 12 7 10 15 16 9 7 11 13

Most likely time: 5 7 11 3 7 9 12 6 5 8 9

Find the earliest and latest expected time to each event and also critical path in the network.

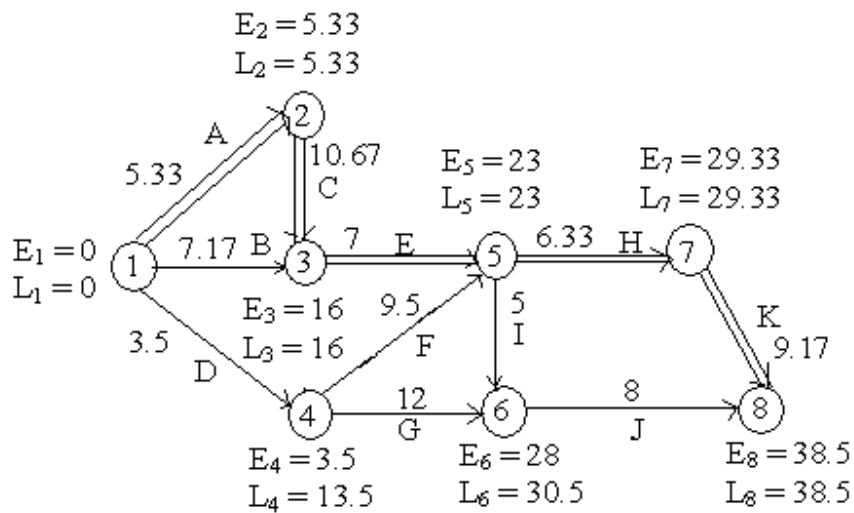
Solution

Task	Least time(t_0)	Greatest time (t_p)	Most likely time (t_m)	Expected time ($t_0 + t_p + 4t_m$)/6
A	4	8	5	5.33
B	5	10	7	7.17
C	8	12	11	10.67
D	2	7	3	3.5
E	4	10	7	7
F	6	15	9	9.5
G	8	16	12	12
H	5	9	6	6.33
I	3	7	5	5
J	5	11	8	8
K	6	13	9	9.17

Task	Expected time (t_e)	Start		Finish		Total float
		Earliest	Latest	Earliest	Latest	
A	5.33	0	0	5.33	5.33	0
B	7.17	0	8.83	7.17	16	8.83
C	10.67	5.33	5.33	16	16	0
D	3.5	0	10	3.5	13.5	10
E	7	16	16	23	23	0

F	9.5	3.5	13.5	13	23	10
G	12	3.5	18.5	15.5	30.5	15
H	6.33	23	23	29.33	29.33	0
I	5	23	25.5	28	30.5	2.5
J	8	28	30.5	36	38.5	2.5
K	9.17	29.33	29.33	31.5	38.5	0

The network is



The critical path is A → C → E → H → K

Choice Paper

MATCB 2.1

Block - III

Marks : 20 (SEE : 25; IA : 05)

Linear Algebra

Syllabus

Unit	Topic
9	Matrix: definition, order, symmetric and skew symmetric matrices
10	Determinant of a matrix, elementary properties of determinants, inverse of a matrix
11	Elementary concept of a vector space, linear dependence and independence of vectors, basis of a vector space, row space, column space, solution of system of linear equations, Cramer's rule.
12	Eigen values and Eigen vectors of matrices, Cayley Hamilton Theorem, Diagonalization of matrices.

Unit IX

Course Structure

- Introduction
 - Objectives
 - Matrix: definition, order
 - Symmetric and skew-symmetric matrices
-

Introduction

The matrix has a long history of application in solving linear equations. They were known as arrays until the 1800 's. The term “matrix” (Latin for “womb”, derived from mater—mother) was coined by James Joseph Sylvester in 1850 , who understood a matrix as an object giving rise to a number of determinants today called minors, that is to say, determinants of smaller matrices that are derived from the original one by removing columns and rows. An English mathematician named Cullis was the first to use modern bracket notation for matrices in 1913 and he simultaneously demonstrated the first significant use of the notation $A = a_{ij}$ to represent a matrix where a_{ij} refers to the element found in the i th row and the j th column. Matrices can be used to compactly write and work with multiple linear equations, referred to as a system of linear equations, simultaneously. Matrices and matrix multiplication reveal their essential features when related to linear transformations, also known as linear maps.

Objectives

After reading this unit you will be able to:

- describe the parts of a matrix and what they represent
- add, subtract and multiply matrices
- distinguish the symmetric and skew-symmetric matrices
- come across various examples of matrices along with their applications

Matrix: Definition

In mathematics, a matrix (plural matrices) is a rectangular array of numbers, symbols, or expressions, arranged in rows and columns. Matrices are commonly written in box brackets. Unless specified, we will consider the entries of matrix to be real or complex numbers. The horizontal and vertical lines of entries in a matrix are called rows and columns, respectively. The size of a matrix is defined by the number of rows and columns that it contains. A matrix with m rows and n columns is called an $m \times n$ matrix, while m and n are called its dimensions. The dimensions of the following matrix is 3×2 , because there are three rows and two columns.

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

The individual items (numbers, symbols or expressions) in a matrix are called its elements or entries. The elements in a row together form a row vector and the elements in a column together forms a column vector. Provided that they are the same size (have the same number of rows and the same number of columns), two matrices can be added or subtracted element by element. The rule for matrix multiplication, however, is that two matrices can be multiplied only when the number of columns in the first equals the number of rows in the second. Any matrix can be multiplied element-wise by a scalar from its associated field.

Matrices which have a single row are called row vectors, and those which have a single column are called column vectors. A matrix which has the same number of rows and columns is called a square matrix. In some contexts, such as computer algebra programs, it is useful to consider a matrix with no rows or no columns, called an empty matrix.

Operations of Matrices

Matrix addition, subtraction, and scalar multiplication are types of operations that can be applied to modify matrices. There are a number of operations that can be applied to modify matrices, such as matrix addition, subtraction, and scalar multiplication. These form the basic techniques to work with matrices. Matrix addition and subtraction requires both the matrices to be of equal dimensions. That is, if A and B are both of $m \times n$ order, then only matrix addition and subtraction are defined. Suppose,

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$$

Then, the matrix addition and subtraction are defined as

$$\begin{aligned} A + B &= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{bmatrix} \\ \text{and } A - B &= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} - \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} - b_{11} & a_{12} - b_{12} \\ a_{21} - b_{21} & a_{22} - b_{22} \end{bmatrix} \end{aligned}$$

The matrix addition and subtraction of any $n \times n$ matrix are analogously defined. Next, we define matrix multiplication. The first condition that needs to be satisfied for any two matrices $A_{m \times n}$ and $B_{p \times q}$ to get multiplied is $n = p$. Then the resulting matrix AB is of order $m \times q$. Let us illustrate a matrix multiplication of 2×3 and 3×2 matrices

$$\begin{bmatrix} 1 & 2 & 1 \\ 2 & 0 & 3 \end{bmatrix} \begin{bmatrix} 3 & 2 \\ 2 & 1 \\ 0 & 3 \end{bmatrix} = \begin{bmatrix} 1 \times 3 + 2 \times 2 + 1 \times 0 & 1 \times 2 + 2 \times 1 + 1 \times 3 \\ 2 \times 3 + 0 \times 2 + 3 \times 0 & 2 \times 2 + 0 \times 1 + 3 \times 3 \end{bmatrix} \\ = \begin{bmatrix} 7 & 7 \\ 6 & 13 \end{bmatrix}$$

The next thing that we come across is the multiplication of a matrix by a scalar, that is a real number, say k . The scalar multiplication is defined as

$$k.A = k \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \\ = \begin{bmatrix} k.a_{11} & k.a_{12} \\ k.a_{21} & k.a_{22} \end{bmatrix}$$

Identity Matrix

We know that whenever a real number, say a , is multiplied by 1, we always get a . That is, $a \times 1 = 1 \times a = a$. The same idea is extendible in case of matrices, particularly those matrices $A_{m \times n}$, where $m = n$ (such matrices are called square matrices). Then the identity matrix of order $n \times n$ is defined as

$$I_{n \times n} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

such that for any matrix $A_{n \times n}$, we will have,

$$A_{n \times n} \times I_{n \times n} = I_{n \times n} \times A_{n \times n} = A_{n \times n}$$

Inverse of a Matrix

Let $A_{n \times n}$ be a matrix. A matrix $B_{n \times n}$ is said to be the inverse of the matrix A if it satisfies the following condition:

$$A \times B = B \times A = I$$

where, I is the identity matrix of order n . We will learn more about inverse of a matrix in the next unit.

Transpose of a matrix

Let A be an $m \times n$ matrix defined as

$$A = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & x_{m3} & \dots & x_{mn} \end{bmatrix}$$

Then the transpose of A is denoted by A^T and defined as

$$A^T = \begin{bmatrix} x_{11} & x_{21} & x_{31} & \dots & x_{n1} \\ x_{12} & x_{22} & x_{32} & \dots & x_{n2} \\ \dots & \dots & \dots & \dots & \dots \\ x_{1n} & x_{2n} & x_{3n} & \dots & x_{nm} \end{bmatrix}$$

Note that the resulting matrix is of order $n \times m$.

Exercise 1. 1. Find the transpose of the matrices

i.

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

ii.

$$\begin{bmatrix} 1 & 3 \\ -1 & 0 \\ 7 & 10 \end{bmatrix}$$

iii.

$$\begin{bmatrix} 3 & 5 \\ 8 & 9 \end{bmatrix}$$

2. Let v be an $n \times 1$ matrix. Prove that vv^T is a symmetric matrix.

3. If A and B are two $n \times n$ symmetric matrices, then show that $A + B$ is also symmetric.

Few properties of transpose are:

- $(A^T)^T = A$
- $(A \pm B)^T = A^T \pm B^T$
- $(kA)^T = kA^T$
- $(AB)^T = A^T B^T$

for any two matrices A and B and a constant k .

Symmetric and Skew-Symmetric Matrices

Let A be an $m \times n$ matrix. Then A is said to be a symmetric matrix if $A = A^T$. Suppose,

$$A = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & x_{m3} & \dots & x_{mn} \end{bmatrix}$$

Then, $A = A^T$ implies that,

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & x_{m3} & \dots & x_{mn} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{21} & x_{31} & \dots & x_{n1} \\ x_{12} & x_{22} & x_{32} & \dots & x_{n2} \\ \dots & \dots & \dots & \dots & \dots \\ x_{1n} & x_{2n} & x_{3n} & \dots & x_{nn} \end{bmatrix}$$

Then from the above equality, we can immediately conclude the following:

1. $m = n$
2. $x_{ij} = x_{ji}$ for all $1 \leq i, j \leq n$

Example 1. i. For all n , the identity matrix is symmetric.

ii.

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 7 \\ 3 & 7 & 1 \end{bmatrix}$$

are two symmetric matrices.

iii. The matrices

$$\begin{bmatrix} 1 & 5 & 8 \\ 2 & 7 & 4 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 5 & 8 \\ 2 & 7 & 4 \\ 4 & 7 & 0 \end{bmatrix}$$

are not symmetric.

Skew-Symmetric Matrices

In a similar way, we can define skew-symmetric matrices. Let A be an $m \times n$ matrix. Then A is said to be a skew-symmetric matrix if $A = -A^T$. Suppose,

$$A = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & x_{m3} & \dots & x_{mn} \end{bmatrix}$$

Then, $A = -A^T$ implies that,

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & x_{m3} & \dots & x_{mn} \end{bmatrix} = \begin{bmatrix} -x_{11} & -x_{21} & -x_{31} & \dots & -x_{n1} \\ -x_{12} & -x_{22} & -x_{32} & \dots & -x_{n2} \\ \dots & \dots & \dots & \dots & \dots \\ -x_{1n} & -x_{2n} & -x_{3n} & \dots & -x_{nm} \end{bmatrix}$$

Then from the above equality, we can immediately conclude the following:

1. $m = n$
2. $x_{ij} = -x_{ji}$ for all $1 \leq i, j \leq n$ and $i \neq j$.
3. $x_{ii} = -x_{ii}$ for all $1 \leq i \leq n$, which implies that $x_{ii} = 0, \forall i$.

Thus, any skew-symmetric matrix looks like

$$\begin{bmatrix} 0 & x_{12} & x_{13} & x_{14} & \dots & x_{1n} \\ -x_{12} & 0 & x_{23} & x_{24} & \dots & x_{2n} \\ -x_{13} & -x_{23} & 0 & x_{34} & \dots & x_{3n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ -x_{1n} & -x_{2n} & -x_{3n} & -x_{4n} & \dots & 0 \end{bmatrix}$$

Theorem 0.1. For any matrix A with real entries, $A + A^T$ is symmetric and $A - A^T$ is skew-symmetric.

Proof. Let A be a matrix and let $B = A + A^T$ and $C = A - A^T$. Then we see that,

$$\begin{aligned} B^T &= (A + A^T)^T \\ &= A^T + (A^T)^T \\ &= A^T + A \\ &= A + A^T \\ &= B \end{aligned}$$

Similarly,

$$\begin{aligned} C^T &= (A - A^T)^T \\ &= A^T - (A^T)^T \\ &= A^T - A \\ &= -(A - A^T) \\ &= -C \end{aligned}$$

Hence, B is symmetric and C is skew-symmetric. □

Theorem 0.2. Any square matrix can be written as the sum of a symmetric and a skew symmetric matrix.

Proof. Let A be a given matrix. Then A can be written as,

$$A = \frac{1}{2}(A + A^T) + \frac{1}{2}(A - A^T)$$

From the previous theorem, we can say that $\frac{1}{2}(A + A^T)$ is symmetric and $\frac{1}{2}(A - A^T)$ is skew-symmetric. Hence the theorem. $\square$

Exercise 2. Let A and B be two $n \times n$ skew-symmetric matrices.

1. Prove that $A + B$ is skew-symmetric.
2. Prove that cA is skew-symmetric for any scalar c .
3. Let P be any $m \times n$ matrix. Prove that $P^T A P$ is skew-symmetric.
4. Prove that, if $AB = -BA$, then AB is a skew-symmetric matrix.

Unit X

Course Structure

- Introduction
 - Objectives
 - Determinant of a matrix, elementary properties of determinants
 - Inverse of a matrix
 - Normal form of a matrix, rank of a matrix
-

Introduction

Determinant of a square matrix is a real number that is assigned to every square matrix. It is a scalar property of the matrix, which can be thought of as the volume enclosed by the row vectors of the matrix. Note that determinant is not defined for any arbitrary matrix. Now, Determinants are mathematical objects that are very useful in the analysis and solution of systems of linear equations. Determinants also have wide applications in engineering, science, economics and social science as well. In this unit, we will mainly study determinant of a matrix and its basic properties and few applications.

Objectives

After reading this unit, you will be able to:

- find the determinant of matrices (of lower order)
- know the properties of determinants
- learn few applications of determinant such as finding inverse of a matrix
- know about cofactors and minors of a matrix
- find the rank of a matrix

Determinant of a Matrix

Let A be a square matrix with real entries. Then, the determinant of A is denoted by $|A|$, or $\det A$. Finding $|A|$ depends upon the dimension of the matrix. For 2×2 matrix

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

the determinant is defined as

$$|A| = a_{11}a_{22} - a_{12}a_{21}$$

Also, the determinant of a 3×3 matrix is

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

In general, the determinant of any $n \times n$ matrix is calculated as

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

where C_{ij} are referred to as the **cofactors** of A and are computed as

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

The term M_{ij} is known as the **Minor Matrix**, which is the matrix obtained by eliminating the i th row and j th column of the matrix A . So to find the determinant of e.g. a 4×4 matrix, you end up calculating a bunch of 3×3 matrix determinants which is much easier. Let us illustrate it for a 4×4 matrix.

Let

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

We intend to find the determinant of A . For that, we will first find the cofactors. We will set up a checkerboard of positive and negative signs as follows: for i th row and j th column, we write the sign as $(-1)^{i+j}$ as a superscript of the corresponding entries.

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

Next we will pick a row or column to expand on. It is easier to choose the row or column having

maximum zeros, since it will make the calculations easier. We are using column 3.

$$\begin{aligned}
 |A| &= (1)(3) \begin{vmatrix} -3 & 1 & 4 \\ -2 & 1 & 1 \\ -1 & 3 & -2 \end{vmatrix} + (-1)(0) \begin{vmatrix} 2 & -1 & 0 \\ -2 & 1 & 1 \\ -1 & 3 & -2 \end{vmatrix} + (1)(4) \begin{vmatrix} 2 & -1 & 0 \\ -3 & 1 & 4 \\ -1 & 3 & -2 \end{vmatrix} + (-1)(0) \begin{vmatrix} 2 & -1 & 0 \\ -3 & 1 & 4 \\ -2 & 1 & 1 \end{vmatrix} \\
 &= 3 \begin{vmatrix} -3 & 1 & 4 \\ -2 & 1 & 1 \\ -1 & 3 & -2 \end{vmatrix} + 4 \begin{vmatrix} 2 & -1 & 0 \\ -3 & 1 & 4 \\ -1 & 3 & -2 \end{vmatrix} \\
 &= 3(-10) + 4(-18) \\
 &= 102
 \end{aligned}$$

Exercise 3. Find the determinant of the following matrices:

1.

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

2.

$$\begin{bmatrix} 1 & 5 & 7 \\ 5 & 2 & -6 \\ 2 & 1 & 0 \end{bmatrix}$$

3.

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

Let us now discuss a few important properties of determinant. Let A be an $n \times n$ square matrix. Then,

- If two rows of A are equal or, any row of A is expressible as a linear combination of other rows, then $|A| = 0$. This is true for columns as well.
- $|A| = |A^T|$
- The value of the determinant remains unchanged if both rows and columns are interchanged.
- If any two rows (or columns) of a determinant are interchanged, then sign of determinant changes.
- If each element of a row (or a column) of a determinant is multiplied by a constant k , then its value gets multiplied by k . In other words, $|kA| = k^n |A|$.
- If some or all elements of a row or column of a determinant are expressed as the sum of two (or more) terms, then the determinant can be expressed as the sum of two (or more) determinants.
- If all the elements of a row (or a column) are zero, then the determinant is zero.

Let us work out some examples using the properties of determinant.

Example 2. Let us find the value of the determinant

$$\begin{vmatrix} \sin^2 x & \cos^2 x & 1 \\ \cos^2 x & \sin^2 x & 1 \\ -10 & 12 & 2 \end{vmatrix}$$

By $C_1 = C_1 + C_2$, the above determinant transforms to

$$\begin{vmatrix} 1 & \cos^2 x & 1 \\ 1 & \sin^2 x & 1 \\ 2 & 12 & 2 \end{vmatrix} = 0$$

since the columns C_1 and C_2 are identical.

Example 3. Solve for x

$$\begin{vmatrix} x & 2 & -1 \\ 2 & 5 & x \\ -1 & 2 & x \end{vmatrix} = 0$$

By $R_3 = R_3 - R_2$, the L.H.S. gets reduced to

$$\begin{vmatrix} x & 2 & -1 \\ 2 & 5 & x \\ -3 & -3 & 0 \end{vmatrix} = 0$$

or, $x(5 \times 0 + 3x) - 2(2 \times 0 + 3x) - 1(-6 + 15) = 0$

or, $3x^2 - 6x - 9 = 0$

or, $x^2 - 2x - 3 = 0$

or, $(x - 3)(x + 1) = 0$

or, $x = 3, -1$

Example 4. We show that

$$\begin{vmatrix} 115 & 106 & 97 \\ 10 & 1 & -8 \\ 106 & 97 & 88 \end{vmatrix} = 0$$

Operating $C_2 = C_2 - \frac{1}{2}(C_1 + C_3)$, we get,

$$\begin{vmatrix} 115 & 0 & 97 \\ 10 & 0 & -8 \\ 106 & 0 & 88 \end{vmatrix} = 0$$

Exercise 4. 1. Without expanding, prove that,

$$\begin{vmatrix} 1 & a & a^2 - bc \\ 1 & b & b^2 - ca \\ 1 & c & c^2 - ab \end{vmatrix} = 0$$

2. Evaluate

$$\begin{vmatrix} 1 & \log_x y & \log_x z \\ \log_y x & 1 & \log_y z \\ \log_z x & \log_z y & 1 \end{vmatrix}$$

where x, y, z are positive.

Inverse of a Matrix

As mentioned in the previous unit, an $n \times n$ matrix is said to be invertible if there exists another $n \times n$ matrix B such that

$$AB = BA = I$$

where I is the identity matrix of order n .

We can also define invertible matrix as follows:

Definition 1. An $n \times n$ matrix A is said to be *invertible* or *non-singular* if $|A| \neq 0$. The matrix whose determinant is zero is called *singular* matrix. The set of all non-singular matrices of order n and entries from a set X is called the General Linear Group of order n and denoted by $GL(n, X)$. If X is the set of real numbers, then it is denoted by $GL(n, \mathbb{R})$.

But why do we need to find the inverse of a matrix in the first place? To understand this, let's see the following equation:

$$ax = b$$

where a and b are real numbers, and we intend to find x . What we do is simply divide both sides by a and get the solution as

$$x = \frac{b}{a}$$

provided that $a \neq 0$. But what happens if we replace the real numbers a, b, x by real matrices?

$$AX = B$$

where A, B, X are matrices. We could not do it like below

$$X = \frac{B}{A}$$

since the division of matrix is as such not defined. This is where, the matrix inverse comes into play. What we do is multiply both sides by the inverse of A , provided the inverse exists, that is, $|A| \neq 0$ (which is analogous to the condition $a \neq 0$ in the previous case). What we get in that case is,

$$\begin{aligned} A^{-1}AX &= A^{-1}B \\ \text{or, } IX &= A^{-1}B \\ \text{or, } X &= A^{-1}B \end{aligned}$$

and hence, we are done.

How to find the inverse of a given matrix?

Suppose we are given an $n \times n$ matrix A as:

$$A = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & x_{24} & \dots & x_{2n} \\ x_{31} & x_{32} & x_{33} & x_{34} & \dots & x_{3n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} & \dots & x_{nn} \end{bmatrix}$$

First check whether $|A| = 0$ or not. If so, then the inverse of A does not exist and we are done. If not, then proceed to the next step. Find out the cofactor for each term of the matrix. Let C_{ij} be the cofactor of each term a_{ij} of the matrix A . Then the inverse is given by

$$A^{-1} = \frac{1}{|A|} \begin{bmatrix} C_{11} & C_{12} & C_{13} & C_{14} & \dots & C_{1n} \\ C_{21} & C_{22} & C_{23} & C_{24} & \dots & C_{2n} \\ C_{31} & C_{32} & C_{33} & C_{34} & \dots & C_{3n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ C_{nn} & C_{n2} & C_{n3} & C_{n4} & \dots & C_{nn} \end{bmatrix}^T$$

Note that the transpose of the cofactor matrix is called *adjoint* of the matrix.

Illustration

Let A be a 3×3 matrix given below:

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 5 \\ 5 & 6 & 0 \end{bmatrix}$$

First of all find the determinant of A by the said method. We see that $|A| = 5$. Then, we find the cofactor of each term. For example, let us find C_{11} . For this, first exclude the first row and first column from the matrix. Then find the determinant of the remaining matrix and multiply it by $(-1)^{1+1} = (-1)^2 = 1$, that is,

$$\begin{aligned} C_{11} &= 1 \cdot \begin{vmatrix} 1 & 5 \\ 6 & 0 \end{vmatrix} \\ &= 1 \cdot 0 - 5 \cdot 6 \\ &= -30 \end{aligned}$$

Similarly, we find $C_{12} = 25$, $C_{13} = -5$, $C_{21} = 18$, $C_{22} = -15$, $C_{23} = 4$, $C_{31} = 7$, $C_{32} = -5$, $C_{33} = 1$. Hence the adjoint is given by,

$$C = \begin{bmatrix} -30 & 18 & 7 \\ 25 & -15 & -5 \\ -5 & -4 & 1 \end{bmatrix}$$

Hence, the inverse is

$$\begin{aligned} A^{-1} &= \frac{1}{5} \begin{bmatrix} -30 & 18 & 7 \\ 25 & -15 & -5 \\ -5 & -4 & 1 \end{bmatrix} \\ &= \begin{bmatrix} -6 & 18/5 & 7/5 \\ 5 & -3 & -1 \\ -1 & -4/5 & 1/5 \end{bmatrix} \end{aligned}$$

Exercise 5. Find the inverse of the following matrices(if exists):

1.

$$\begin{bmatrix} 3 & 4 \\ 6 & 8 \end{bmatrix}$$

2.

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

Unit XI

Course Structure

- Introduction
 - Objectives
 - Elementary concept of a vector space
 - Linear dependence and independence of vectors, basis of a vector space, row space, column space
 - Solution of system of linear equations, Cramer's rule
-

Introduction

Consider arrows in a fixed plane starting at one fixed point, say the origin. Given any two such arrows, v and w , the parallelogram spanned by these two arrows contains one diagonal arrow that starts at the origin, too. This new arrow is called the sum of the two arrows and is denoted $v + w$. In the special case of two arrows on the same line, their sum is the arrow on this line whose length is the sum or the difference of the lengths, depending on whether the arrows have the same direction. Another operation that can be done with arrows is scaling: given any positive real number a , the arrow that has the same direction as v , but is dilated or shrunk by multiplying its length by a , is called multiplication of v by a . It is denoted av . When a is negative, av is defined as the arrow pointing in the opposite direction, instead.

Objectives

After studying this unit, you will be able to:

- know the definition of vector spaces
- learn various standard vector spaces
- visualize linearly independent and dependent vectors
- find out basis of a vector space
- solve system of linear equations independently and by Cramer's Rule

Vector Space

As the name suggests, a vector space is a non-empty set V , say, with two binary compositions, one external between the elements of V and the associated field F , and one internal composition between two elements of V , defined as follows:

1. for $x, y \in V$, we have, $x + y \in V$ (closure property)
2. for $x, y, z \in V$, we have, $(x + y) + z = x + (y + z)$ (associativity)
3. for all $v \in V$, we have, $v + 0 = 0 + v = v$ (identity property)
4. for all $v \in V$, there exists $-v \in V$ such that, $v + (-v) = (-v) + v = 0$ (additive inverse property)
5. for $x \in V$ and $s \in F$, we will always have $sx \in V$
6. for $x \in V$ and $r, s \in F$, we have, $r(sx) = (rs)x$
7. for $x \in V$ and $r, s \in F$, we have, $(r + s)x = rx + sx$
8. for $x, y \in V, r \in F$, we have, $r(x + y) = rx + ry$
9. for $x \in V$, we always have $1.x = x.1 = x$, where 1 is the multiplicative identity in V

Then, $(V, +, \cdot)$ forms a vector space over the field of scalars F . A vector space always has at least one element, the zero vector. Hence the set $(\{0\}, +, \cdot)$, forms a vector space, called the trivial space.

Example 5. 1. The set of real numbers $\mathbb{R}$ is a vector space over the field $\mathbb{R}$ with respect to addition and multiplication.

2. The set of n -tuples of real numbers, $\mathbb{R}^n$, where n is a positive integer, is a vector space over the field $\mathbb{R}$.
3. The set of complex numbers $\mathbb{C}$ forms a vector space over both the fields $\mathbb{R}$ and $\mathbb{C}$.
4. The set of all matrices form a vector space over the field $\mathbb{R}$ and $\mathbb{C}$.

In the study of any algebraic structure, it is of interest to examine subsets that possess the same structure as the set under consideration. We define the apt substructure of a vector space as follows:

Definition 2. A subset W of a vector space V over the field F , is said to be a subspace of V if W forms a vector space over the field F with respect to the operations of addition and multiplication defined over V .

In any vector space V , the zero set $\{0\}$ and V itself are always a vector subspace of V . These are called the trivial subspaces of V .

Example 6. 1. The subset $\{(x, y, z) \in \mathbb{R}^3 : x = y + z\}$ is a subspace of $\mathbb{R}^3$ while the subset $\{(x, y, z) \in \mathbb{R}^3 : x = y + z + 1\}$ is not a subspace of $\mathbb{R}^3$.

2. Check whether the subset $\{(x, y, z) \in \mathbb{R}^3 : xyz = 0\}$ is a subspace of $\mathbb{R}^3$.

Example 7. Let $u, v \in V$, where V is a vector space over a field F . Define the set W as

$$W = \{ru + sv : r, s \in F\}$$

Check that W forms a vector space over F . This set W is said to be *spanned* by the vectors u and v and is denoted by $W = L(u, v)$.

In a similar way, the set

$$W = \{c_1v_1 + c_2v_2 + \cdots + c_nv_n : c_1, c_2, \dots, c_n \in F\}$$

where $v_1, v_2, \dots, v_n \in V$ forms a vector subspace of V .

Definition 3. Let V be a vector space and let $v_1, v_2, \dots, v_n \in V$, then we define a *linear combination* of $v_1, v_2, \dots, v_n \in V$ as

$$v = c_1v_1 + c_2v_2 + \cdots + c_nv_n$$

where $c_1, c_2, \dots, c_n \in F$.

Definition 4. The subset W of V is called the *Linear Span* of $v_1, v_2, \dots, v_n$ which is represented as $LS(v_1, v_2, \dots, v_n)$.

Linear Dependence and Independence

Consider the vector space $\mathbb{R}^2$ and a vector say $(1, 2)$ in $\mathbb{R}^2$. Consider two other vectors $(3, 6)$ and $(2, 2)$. We see that the vector $(3, 6)$ can be written as $(3, 6) = 3(1, 2)$, that is, $(3, 6)$ lies on the line joining $(0, 0)$ and $(1, 2)$. But this is not the case with $(2, 2)$. Then $(1, 2)$ and $(3, 6)$ are said to be linearly dependent and $(1, 2)$ and $(2, 2)$ are said to be linearly independent. We formally define linear dependence and independence as follows:

Definition 5. The vectors $v_1, v_2, \dots, v_n$ in a vector space V are said to be linearly independent if

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

can be satisfied only when $c_i = 0, \forall i$. If $v_1, v_2, \dots, v_n$ are not linearly independent, then they are said to be linearly dependent. In other words, if $v_1, v_2, \dots, v_n$ are linearly dependent, then any equation as

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

can be satisfied when at least one of c_i is non-zero.

For the above example, the given vectors can be written as

$$(3, 6) + (-3)(1, 2) = 0$$

We can also say that $(3, 6)$ is a linear combination of the set of vectors $\{(1, 2)\}$.

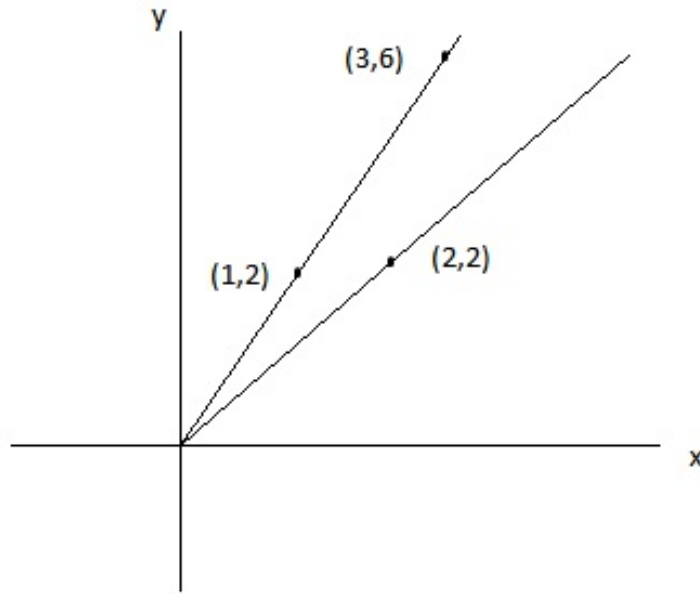


Figure 9: Representation of vectors in the xy -plane

Now, let $\{v_1, v_2, \dots, v_n\}$ are a set of linearly dependent vectors. Then for any equation

$$c_1v_1 + c_2v_2 + \dots + c_nv_n = 0$$

we must have at least one value of c , say $c_k \neq 0$. Then the above equation can be written as

$$\begin{aligned} c_1v_1 + c_2v_2 + \dots + c_kv_k + \dots + c_nv_n &= 0 \\ \text{or, } c_1v_1 + c_2v_2 + \dots + c_nv_n &= -c_kv_k \\ \text{or, } -\frac{c_1}{c_k}v_1 - \frac{c_2}{c_k}v_2 - \dots - \frac{c_n}{c_k}v_n &= v_k \end{aligned}$$

From the last line, we can say that v_k is a linear combination of the set of vectors $\{v_1, v_2, \dots, v_{k-1}, v_{k+1}, \dots, v_n\}$. Thus, if you have a set of linearly dependent vectors, then you can remove the vector(s) that have a dependence and not change the possible things that the other vectors sum to. We thus have

$$\text{LS}\{v_1, v_2, \dots, v_n\} = \text{LS}\{v_1, v_2, \dots, v_{k-1}, v_{k+1}, \dots, v_n\}$$

Now, consider the vector space V . Then of course V spans itself. Then we choose an element say v which can be written as the linear combination of some other elements of V . Since deletion of a linearly dependent element does not affect the linear span, so we delete v from V . Again we select an element from $V \setminus \{v\}$ which can be written as the linear combination of some elements of $V \setminus \{v\}$. We continue the process and obtain a minimal set that spans the whole set V . Let the set be $\{v_1, v_2, \dots\}$. Of course this set is linearly independent.

Now, consider an element v_1 of a vector space V . Then the linear span of v_1 is $\{av_1 : a \in F\}$ is of course a subspace of V . If $LS(v_1) = V$, then we stop here and say that v_1 spans the whole set V . If not, then we can find an element, say $v_2 \notin LS(v_1)$. Thus v_1, v_2 are linearly independent and again $LS(v_1, v_2)$ is a subspace of V . Again if $LS(v_1, v_2) = V$ then we say that V is spanned by $\{v_1, v_2\}$. If not, then we continue the process and get a linearly independent set $\{v_1, v_2, \dots\}$ that spans V . Such a set is called the *basis* of V . We define the basis of a vector space V as

Definition 6. Let V be a vector space. Then a set of elements of V $\{v_1, v_2, \dots\}$ is said to be a basis of V if

- it spans V , that is, every element of V can be written as a linear combination of the elements of the set
- it is linearly independent

Note that, a basis is the maximal linearly independent set and minimal spanning set of V . That is, if we add an element in the basis, then it no longer remains linearly independent. Also, if we exclude an element from the basis, then it no longer spans V . The number of elements in the basis of a vector space is called the **dimension** of V . If the basis of V contains infinite number of elements, then the dimension of V is infinite. Otherwise, the dimension of V is finite.

Some results related to the basis and dimension of a vector space

Let V be a vector space and let $S = \{v_1, v_2, \dots, v_n\}$ be a basis of V . Then we have the following results:

1. Every element of V can be written uniquely as a linear combination of the elements of S .

Proof. Let $v \in V$. Then there exists $c_1, c_2, \dots, c_n$ such that

$$v = c_1v_1 + c_2v_2 + \dots + c_nv_n$$

Again, let v has another representation as

$$v = d_1v_1 + d_2v_2 + \dots + d_nv_n$$

Then, subtracting the two, we get

$$0 = (c_1 - d_1)v_1 + (c_2 - d_2)v_2 + \dots + (c_n - d_n)v_n$$

Since $\{v_1, v_2, \dots\}$ is linearly independent, so we will get, $c_i = d_i, \forall i$. Hence proved. $\square$

This result is also true for infinite dimensional vector spaces.

2. Any subset of V with more than n elements is dependent.
3. Any subset of V with fewer than n elements cannot span V .
4. Any linearly independent set having exactly $\dim V$ number of elements is a basis of V .

5. If W is a subspace of V such that $\dim V = \dim W$, then $V = W$.

Example 8. 1. $\mathbb{R}$ is a one-dimensional vector space over the field $\mathbb{R}$ with basis $\{1\}$. What happens if we replace the field by the set of all rational numbers $\mathbb{Q}$?

2. $\mathbb{R}^n$ is an n -dimensional vector space over the field $\mathbb{R}$ with basis $\{(1, 0, \dots, 0), (0, 1, \dots, 0), \dots, (0, 0, \dots, 1)\}$.

3. $\mathbb{C}$ is a one-dimensional vector space over the field $\mathbb{C}$ whose basis is $\{1\}$. It is also a 2-dimensional vector space if we replace the field by the field of real numbers $\mathbb{R}$. In that case, the basis is $\{1, i\}$.

4. The set of all polynomials of degree less than or equal to n is a vector space over the field $\mathbb{R}$. The basis in that case is $\{1, x, x^2, \dots, x^n\}$. (verify!)

5. Consider the vector space $\mathbb{R}^2$. Is the set $\{(1, 1), (1, 0)\}$ linearly independent? Is it a basis of $\mathbb{R}^2$?

6. The matrices

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

is a basis of the set of all 2×2 matrices over the field $\mathbb{R}$. Similarly, we can find the basis of the set of all $n \times n$ matrices over the field $\mathbb{R}$. (Find it)

7. Let A be an $m \times n$ matrix with real entries. Then the set spanned by the row vectors (or linearly independent row vectors) of A is called the row space of A . It is a subspace of $\mathbb{R}^n$. We can similarly define the column space of A which would be a subspace of $\mathbb{R}^m$.

System of Linear Equations

Consider the system of linear equations

$$\begin{aligned} x + y &= 1 \\ x - y &= 0 \end{aligned}$$

These two are a set of linear equations in $\mathbb{R}^2$. Any solution of these two equations are straight lines in the xy -plane. Solution of this system is the point where these two lines intersect. For the given system, $(1/2, 1/2)$ is the unique solution. Now, what happens if the second equation is replaced by $x + y = 4$? Then the equations represent two parallel straight lines which never intersect. Hence such system will have no solution. If instead, we have the second equation as $2x + 2y = 2$, then clearly, it is a linear combination of the first equation. Hence in such case the equation has infinitely many solutions.

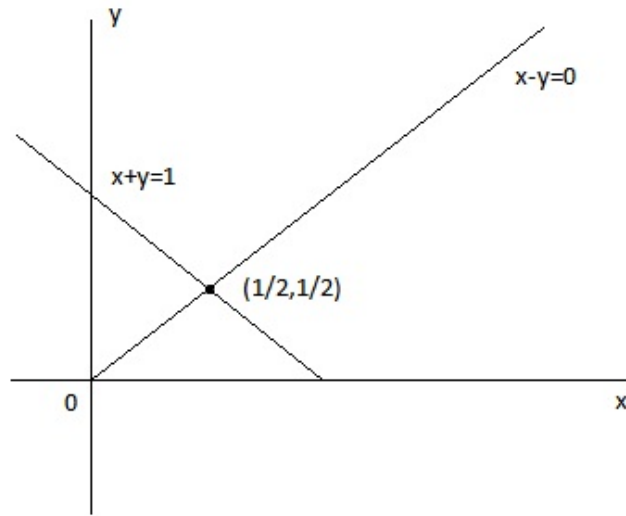


Figure 10: Geometrical Interpretation of the solution of the system of linear equations

General Form

The most general form of a system of linear equations is

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

which can be written in the matrix form as

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

or,

$$Ax = b$$

where

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

When the elements $b_i = 0, \forall i$, then the system is called homogeneous. Otherwise, it is called non-homogeneous system of linear equations. Notice that, $(0, 0, \dots, 0)$ is always a solution of the

homogeneous system. This is called the trivial solution. Hence, if the solution set is S_h , then it is always non-empty. Now, if u, v are two solutions of the homogeneous system, then $cu + dv$ is always a solution of it, for real constants c, d (verify). Hence, the solution set S_h forms a vector space over the field $\mathbb{R}$. If $\{x_1, x_2, \dots, x_n\}$ is the basis of S_h , then the most general solution of the homogeneous system is $c_1x_1 + c_2x_2 + \dots + c_nx_n$, where c_i are real constants. If x_1, x_0 are solutions of the homogeneous and non-homogeneous systems respectively, then check that $x_1 + x_0$ is also a solution of the non-homogeneous system. Thus, the most general solution of the non-homogeneous system is $c_1x_1 + c_2x_2 + \dots + c_nx_n + x_0$, where c_i are real constants.

Solution of System of Linear Equations

A system of linear equations can be solved in many ways. One of the ways is by finding the inverse of the corresponding coefficient matrix, if it exists. The solution in such case, is given by

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

Illustration

Let a system of linear equation be given as

$$\begin{aligned} 2x + y &= 1 \\ 3x - y &= 0 \end{aligned}$$

The corresponding coefficient matrix is

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

Hence, the solution is given by

$$\begin{aligned} \begin{bmatrix} x \\ y \end{bmatrix} &= \begin{bmatrix} 2 & 1 \\ 3 & -1 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ &= -5 \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} -5 \\ 0 \end{bmatrix} \end{aligned}$$

which is the required solution.

Note that the above process can only be used when the number of equations is equal to the number of unknowns, that is, the coefficient matrix is a square matrix.

The solution of the above system can also be found by elimination. For example, adding the two equations yield

$$\begin{aligned} 5x &= 1 \\ \text{or, } x &= \frac{1}{5} \end{aligned}$$

Putting the value of x in the second equation yields

$$y = \frac{3}{5}$$

Notice that the solutions obtained by the two methods are different. In fact, it is not mandatory for system of linear equations to have unique solution. They can have more than one solution by virtue of the description given in the previous section.

Cramer's Rule

Cramer's Rule is an explicit way of finding the solution of a system of linear equations whose number of equations and number of unknowns are same. Instead of solving the entire system of equations, you can use Cramer's to solve for just one single variable. Let's illustrate it with an example.

Illustration

Let the following be a linear system:

$$\begin{aligned} 2x + y + z &= 3 \\ x - y - z &= 0 \\ x + 2y + z &= 0 \end{aligned}$$

First we find the determinant of the coefficient matrix which is

$$D = \begin{vmatrix} 2 & 1 & 1 \\ 1 & -1 & -1 \\ 1 & 2 & 1 \end{vmatrix} = 3 \text{ (verify)}$$

Define D_1, D_2, D_3 as the determinant of the coefficient matrix in which the first, second and third columns respectively are replaced by the matrix b . That is,

$$D_1 = \begin{vmatrix} 3 & 1 & 1 \\ 0 & -1 & -1 \\ 0 & 2 & 1 \end{vmatrix} = 3, \quad D_2 = \begin{vmatrix} 2 & 3 & 1 \\ 1 & 0 & -1 \\ 1 & 0 & 1 \end{vmatrix} = -6, \quad D_3 = \begin{vmatrix} 2 & 1 & 3 \\ 1 & -1 & 0 \\ 1 & 2 & 0 \end{vmatrix} = 9$$

The solutions are given by

$$\begin{aligned} x &= \frac{D_1}{D} = 1 \\ y &= \frac{D_2}{D} = -2 \\ z &= \frac{D_3}{D} = 3 \end{aligned}$$

There are other ways to find the solution of linear systems in which the number of equations and the number of unknowns are not same. But in this unit, we restrict ourselves to the linear system having same number of equations as the number of unknowns.

Exercise 6. Find the solution of the following systems:

a.

$$\begin{aligned}2x + y + z &= 1 \\ x - y + 4z &= 0 \\ x + 2y - 2z &= 3\end{aligned}$$

b.

$$\begin{aligned}3x - y &= 7 \\ -5x + 4y &= -2\end{aligned}$$

c.

$$\begin{aligned}-6x + 8y &= 17 \\ 13x - 2y &= -4\end{aligned}$$

d.

$$\begin{aligned}2x - y + 6z &= 10 \\ -3x + 4y - 5z &= 11 \\ 8x - 7y - 9z &= 12\end{aligned}$$

e.

$$\begin{aligned}2x - 3y &= 4 \\ -x + 4y - z &= 11 \\ 6x - 5y + 2z &= -3\end{aligned}$$

Unit XII

Course Structure

- Introduction
 - Objectives
 - Eigen values and Eigen vectors of matrices
 - Cayley Hamilton Theorem
 - Diagonalization of matrices
-

Introduction

Consider the matrix

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

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

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

Geometrically speaking, when we take a particular vector (v_1, v_2) of the xy -plane and operate the matrix on it, we the resulting vector is a scalar multiple of the original one. That is, the resulting vector is either a contracted or expanded form of the original vector depending on the value of c . For example, if we take the vector $(1, 1)$, then the resulting vector will be $(3, 3) = 3(1, 1)$. That is, the particular vector is expanding to thrice its original value. On the other hand, if we operate the matrix over the vector $(0, 1)$, then the resulting vector $(2, 3)$ is not on the line joining $(0, 1)$ and $(2, 3)$. The vector $(1, 1)$ is called an eigen vector and 3 is the corresponding eigen value. $(0, 1)$ is not an eigen vector. We are now in a position to formally define eigen values and eigen vectors of a matrix.

Objectives

After reading this unit, you will be able to:

- find the eigen values and eigen vectors of a matrix
- learn the Cayley Hamilton theorem

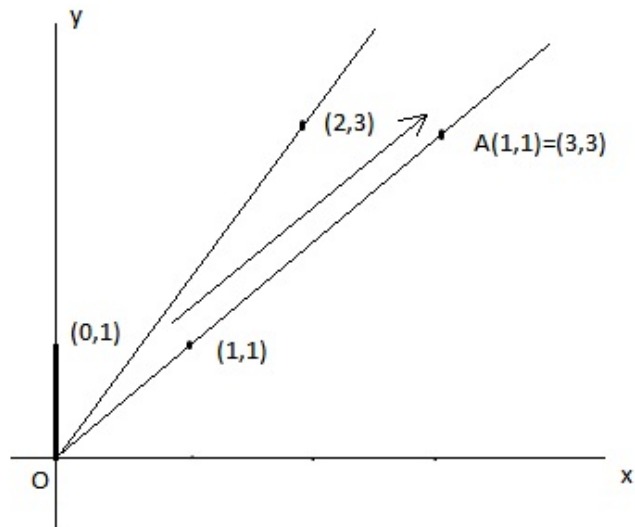


Figure 11: Eigen Values and Eigen Vectors Geometrically

- find the characteristic polynomial of a matrix
- find the eigen values from the characteristic polynomial
- define diagonalizability of a matrix
- find out when a matrix will be called diagonalizable

Eigen values and Eigen vectors of a matrix

Let A be an $n \times n$ matrix. The number c is called an eigen value of A if there exists a non-zero vector v such that

$$Av = cv$$

In such a case v is called the eigen vector corresponding to the eigen value c .

Computation of Eigen values and eigen vectors

The equation

$$Av = cv$$

can be re-written as

$$(A - cI)v = 0$$

Now, for non-trivial solution of the above equation, $A - cI$ must not be invertible, because in that case, we would have

$$\begin{aligned} (A - cI)^{-1}(A - cI)v &= (A - cI)^{-1}0 \\ \text{or, } v &= 0 \end{aligned}$$

So $\det(A - cI) = 0$. We call $f(c) = \det(A - cI)$ as the characteristic polynomial of A . Solving this for c , we get the eigen values. Hence, the roots of the characteristic polynomial of A gives the eigen values of A . Also the power to which the factor of any eigen value is raised, is called its algebraic multiplicity, that is, if λ is an eigen value of A , and d is the highest power to which $(x - \lambda)$ is raised, then d is the algebraic multiplicity of λ .

To find the eigen vectors of A , we need to simply solve the system of linear equations given by

$$(A - cI)v = 0$$

Let's illustrate it by an example.

Illustration

Let the given matrix be

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

Let c be an eigen value of A and (v_1, v_2) be the corresponding eigen vector. Then

$$(A - cI)v = 0$$

will give non-trivial solution if $\det(A - cI) = 0$. Solving this for c will give us all the eigen values. So,

$$\begin{aligned} \begin{vmatrix} 1-c & 2 \\ 2 & 4-c \end{vmatrix} &= 0 \\ \text{or, } (1-c)(4-c) - 4 &= 0 \\ \text{or, } c(c-5) &= 0 \\ \text{or } c &= 0, 5 \end{aligned}$$

Hence, 5 and 0 are the eigen values of A . Now, we will find the corresponding eigen vectors. For $c = 0$,

$$\begin{aligned} \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} &= 0 \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \end{aligned}$$

This gives us the homogeneous system of linear equations

$$\begin{aligned} v_1 + 2v_2 &= 0 \\ 2v_1 + 4v_2 &= 0 \end{aligned}$$

Solving, we get $v_1 = -2v_2$. If we take $v_2 = k$, then the eigen vector is given by $(-2k, k) = k(-2, 1)$, where k is any real constant. Thus any element of the subspace $S = \{k(-2, 1) : k \in \mathbb{R}\}$ is an eigen vector of A corresponding to the eigen value 0. Taking $k = 1$ we get a particular eigen vector $(-2, 1)$

corresponding to the eigen value 0. We can also say that $(-2, 1)$ spans the set S . This set is called the *eigen space* of A corresponding to the eigen value 0.

For $c = 5$, we will have

$$\begin{aligned} \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} &= 5 \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \\ &= \begin{bmatrix} 5v_1 \\ 5v_2 \end{bmatrix} \end{aligned}$$

which gives us the homogeneous system

$$\begin{aligned} -4v_1 + 2v_2 &= 0 \\ 2v_1 - v_2 &= 0 \end{aligned}$$

Solving, we get $2v_1 = v_2$. If we take $v_1 = k$, then the eigen vectors will be given by $(k, 2k) = k(1, 2)$ for any real constant k . Thus the eigen space, say T corresponding to the eigen value 5 is given by $\{k(1, 2) : k \in \mathbb{R}\}$. Putting $k = 1$, a particular eigen vector corresponding to the eigen value 5 is $(1, 2)$.

Remark: The dimension of the eigen space corresponding to the eigen value c of a matrix A is called its geometric multiplicity.

Exercise 7. Find the eigen values and eigen vectors of the following matrices:

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}, \begin{bmatrix} 3 & 6 & -8 \\ 0 & 0 & 6 \\ 0 & 0 & 2 \end{bmatrix}$$

Theorem 0.3. If $v_1, v_2, \dots, v_k$ are eigen vectors of a matrix A , corresponding to its distinct eigen values $c_1, c_2, \dots, c_k$, then the set $\{v_1, v_2, \dots, v_k\}$ is linearly independent.

Proof. Let us consider the equation

$$r_1v_1 + r_2v_2 + \dots + r_kv_k = 0$$

for real constants $r_1, r_2, \dots, r_k$. We have to show that these constants are zero. For any positive $i \leq k$, if we operate the matrix $(A - c_iI)$ on both sides of the equation □

Theorem 0.4. The determinant of an $n \times n$ matrix is 0 if and only if 0 is an eigen value of A .

Proof. Let 0 be an eigen value of A . Then

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

which yields $\det A = 0$ for non-trivial solution. Hence we are done.

Conversely, let $\det A = 0$. This implies that

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

which implies that the equation

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

has non-trivial solution. Hence, 0 is an eigen value of A . □

Remark: Note that the zero vector can never be an eigen vector of any matrix.

Definition 7. Two $n \times n$ matrices A and B are said to be similar if there exists an invertible matrix P such that

$$A = PBP^{-1}$$

Two similar matrices always have the same characteristic polynomial. So they have the same eigen values.

Diagonalizability

An $n \times n$ matrix A is said to be diagonalizable if it has n linearly independent eigen vectors. If A has n distinct eigen values then the corresponding eigen vectors are linearly independent. So, the existence of n distinct eigen values guarantees the diagonalizability of A . If A does not have distinct eigen values then we check the dimension of the eigen spaces. For example, if λ is a repeated root of the characteristic polynomial of A , then we consider the eigen space corresponding to λ . If the geometric multiplicity of every eigen value of A is equal to its algebraic multiplicity, then A is diagonalizable. Let's illustrate it with an example.

Illustration

Let us consider the matrix

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

First we find the eigen values of A by

$$\det(A - cI) = 0$$

which gives the characteristic polynomial A as below,

$$c^2 = 0$$

and hence, the only eigen value of A is 0 with algebraic multiplicity 2. We will consider the eigen space of 0. Let us find the eigen vector of A corresponding to 0. For that, consider the system of equations

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

which gives $v_2 = 0$ and v_1 , undetermined. So the eigen space is the set $\{k(1, 0) : k \in \mathbb{R}\}$. Thus the eigen space is spanned by only one element and hence, the geometric multiplicity of 0 is $1 < 2$, the algebraic multiplicity of 0. Hence, A is not diagonalizable.

Exercise 8. Show that the following matrices are not diagonalizable:

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

Let us have another example of a matrix which is diagonalizable. Consider the matrix

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

We find the eigen values of A by the equation

$$\det(A - cI) = 0$$

which gives us the characteristic polynomial of A as

$$(2 - c)^2(1 - c) = 0$$

Thus, the eigen values of A are 1 and 2 of algebraic multiplicities 1 and 2 respectively. Next we find the corresponding eigen spaces.

For $c = 1$, let $v = (v_1, v_2, v_3)$ be the eigen vector of A . Then we solve the system of linear equations as

$$(A - 1I)v = 0$$

which gives

$$\begin{aligned} 2v_1 + 0v_2 + 0v_3 &= 0 \\ v_1 + v_2 + v_3 &= 0 \\ -v_1 + 0v_2 + 0v_3 &= 0 \end{aligned}$$

Solving, we get $v_1 = 0, v_3 = -v_2$. Thus, if we take $v_2 = k$, for some real constant k , then the eigen space will be $\{k(0, 1, -1) : k \in \mathbb{R}\}$. Thus, the eigen space is spanned by $(0, 1, -1)$, and hence the geometric multiplicity of 1 is 1.

For $c = 2$, let $v = (v_1, v_2, v_3)$ be the eigen vector of A . Then we solve the system of linear equations as

$$(A - 2I)v = 0$$

which gives

$$\begin{aligned} 0v_1 + 0v_2 + 0v_3 &= 0 \\ v_1 + 0v_2 + v_3 &= 0 \\ -v_1 + 0v_2 - v_3 &= 0 \end{aligned}$$

solving, we get $v_3 = -v_1$ and v_2 remains undetermined. So, if we take $v_1 = k$ and $v_2 = l$, then we get the eigen space as $\{(k, l, -k) : k, l \in \mathbb{R}\} = \{k(1, 0, -1) + l(0, 1, 0) : k, l \in \mathbb{R}\}$. Hence, the eigen space corresponding to 2 is spanned by the vectors $(1, 0, -1)$ and $(0, 1, 0)$, hence the geometric multiplicity of 2 is 2.

Exercise 9. Are the following matrices diagonalizable?

$$\begin{bmatrix} 2 & 4 & 6 \\ 0 & 2 & 2 \\ 0 & 0 & 4 \end{bmatrix}, \quad \begin{bmatrix} 2 & 0 & 0 \\ 2 & 6 & 0 \\ 3 & 2 & 1 \end{bmatrix}$$

Cayley Hamilton Theorem

Cayley Hamilton theorem is an important tool to find the inverse and powers of a matrix in an easier way. We will learn to do a few of such applications. First let us state the theorem as follows:

Theorem 0.5. *Let A be an $n \times n$ matrix and $f(t)$ be its characteristic polynomial. Then $f(A)$ is the $n \times n$ zero matrix.*

Let us verify the Cayley-Hamilton theorem for the matrix

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

First find the characteristic polynomial of A . Verify that, it is $c^2 - 5c + 6 = 0$. Now we will find A^2 .

$$\begin{aligned} A^2 &= \begin{bmatrix} 6 & -2 \\ 6 & -1 \end{bmatrix} \begin{bmatrix} 6 & -2 \\ 6 & -1 \end{bmatrix} \\ &= \begin{bmatrix} 24 & -10 \\ 30 & -11 \end{bmatrix} \end{aligned}$$

So,

$$\begin{aligned} A^2 - 5A + 6I &= \begin{bmatrix} 24 & -10 \\ 30 & -11 \end{bmatrix} - 5 \begin{bmatrix} 6 & -2 \\ 6 & -1 \end{bmatrix} + 6 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 24 & -10 \\ 30 & -11 \end{bmatrix} - \begin{bmatrix} 30 & -10 \\ 30 & -5 \end{bmatrix} + \begin{bmatrix} 6 & 0 \\ 0 & 6 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \end{aligned}$$

Hence, A satisfies its characteristic polynomial.

Exercise 10. Verify Cayley-Hamilton theorem for the following matrices:

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

Applications of the Cayley-Hamilton theorem

Example 9. Let us find the inverse of

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

The characteristic equation of A is $f(t) = t^2 - 4t + 3$. Then by Cayley-Hamilton theorem,

$$A^2 - 4A + 3I = 0$$

Multiplying both sides by A^{-1} , we get

$$A^{-1}(A^2 - 4A + 3I) = A^{-1}.0$$

which gives

$$\begin{aligned}
 A - 4I + 3A^{-1} &= 0 \\
 A^{-1} &= \frac{1}{3}(4I - A) \\
 &= \frac{1}{3} \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix} - \frac{1}{3} \begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix} \\
 &= \begin{bmatrix} 1 & -2/3 \\ 0 & 1/3 \end{bmatrix}
 \end{aligned}$$

Exercise 11. a. Calculate and simplify the expression

$$-T^3 + 4T^2 + 5T - 2I$$

where, I is the 3×3 identity matrix and

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

b. Find the inverse of the matrix A using Cayley-Hamilton theorem, where

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 9 & 2 & 0 \\ 5 & 0 & 3 \end{bmatrix}$$

c. Find the inverse of the matrix B using Cayley-Hamilton theorem, where

$$B = \begin{bmatrix} 7 & 2 & -2 \\ -6 & -1 & 2 \\ 6 & 2 & -1 \end{bmatrix}$$

References

1. Topics in Algebra, I. N. Herstein
2. Linear Algebra, K.Hoffman and R. Kunze
3. Linear Algebra, B. C. Chatterjee

Choice Based Paper

MATCB 2.1

Block - IV

Marks : 25 (SSE : 20; IA : 05)

Dynamical Systems

Syllabus

• Unit 13 •

Linearization of dynamical systems: Two, three and higher dimension. Population growth. Lotka-Volterra system.

• Unit 14 •

Stability: Asymptotic stability (Hartman's theorem), Global stability (Liapunov's second method).

• Unit 15 •

Limit set, attractors, periodic orbits, limit cycles. Bendixon criterion, Dulac criterion, Poincare-Bendixon Theorem. Floquet's theorem.

• Unit 16 •

Stability and bifurcation: Routh-Hurwitz criterion for nonlinear systems. Saddle-Node, transcritical and pitchfork bifurcations. Hopf- bifurcation.

Unit 13

Linearization of Differential Equation Models

1 Motivation

We cannot solve most nonlinear models, so we often instead try to get an overall feel for the way the model behaves: we sometimes talk about looking at the **qualitative dynamics** of a system. Equilibrium points— steady states of the system— are an important feature that we look for. Many systems settle into a equilibrium state after some time, so they might tell us about the long-term behavior of the system.

Equilibrium points can be stable or unstable: put loosely, if you start near an equilibrium you might, over time, move closer (stable equilibrium) or away (unstable equilibrium) from the equilibrium. Physicists often draw pictures that look like hills and valleys: if you were to put a ball on a hill top and give it a push, it would roll down either side of the hill. If you were to put a ball at the bottom of a valley and push it, it would fall back to the bottom of the valley.

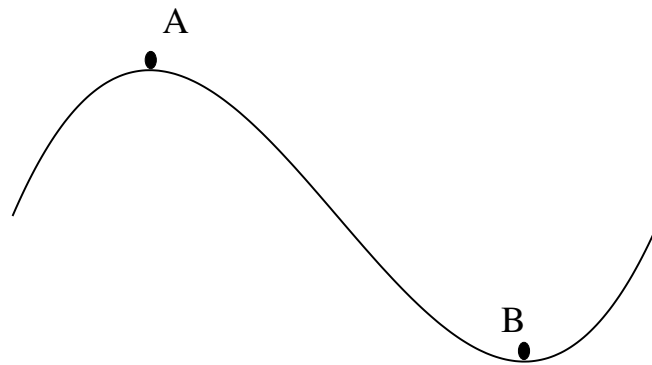


Figure 1: An example of stability: both A and B are equilibrium points, at the top of a hill and at the bottom of a valley. If the ball at point A is pushed in either direction, it will roll away down the hill. If the ball at point B is pushed a small amount in either direction, it will roll back to its initial point.

Mathematicians have many different definitions of ‘stability’, but we won’t worry too much about such details, except to say that we often distinguish between **local** and **global** stability. In the example above, point B is **locally stable** but not **globally stable**. If you only push the ball a short distance away, it will roll back to point B. If you push the ball far enough (i.e. beyond point A), it will not roll back to point B. More detailed information on stability can be found in books

on nonlinear differential equations or dynamical systems (for instance S. H. Strogatz's 'Nonlinear Dynamics and Chaos').

Linearization can be used to give important information about how the system behaves in the neighborhood of equilibrium points. Typically we learn whether the point is stable or unstable, as well as something about how the system approaches (or moves away from) the equilibrium point.

The basic idea is that (in most circumstances) one can approximate the nonlinear differential equations that govern the behavior of the system by linear differential equations. We can solve the resulting set of linear ODEs, whereas we cannot, in general, solve a set of nonlinear differential equations.

2 How to Linearize a Model

We shall illustrate the linearization process using the SIR model with births and deaths in a population of fixed size.

$$\dot{S} = \mu N - \beta SI/N - \mu S \quad (1)$$

$$\dot{I} = \beta SI/N - (\gamma + \mu)I \quad (2)$$

$$\dot{R} = \gamma I - \mu R, \quad (3)$$

with $S + I + R = N$. Since we assume that the population is closed, we can always calculate the value of R if we know S and I . Therefore we need only focus on the first two equations for S and I .

We denote an equilibrium point by attaching asterisks to the state variables: e.g. an equilibrium point of the SIR model may be written as (S^*, I^*) . Because an equilibrium point means that the values of S and I (and R) remain constant, this means that $dS/dt = dI/dt = 0$ when $(S, I) = (S^*, I^*)$.

If we imagine that both S and I are close to the equilibrium point, then the differences $S - S^*$, which we denote by x_1 , and $I - I^*$, which we denote by x_2 , will be small. An important point is that terms involving products of x_1 and x_2 (e.g. the quadratic terms x_1^2, x_2^2 or $x_1 x_2$) are much smaller still and so, to a very good approximation, can be ignored.

We can differentiate our expression for x_1 : $dx_1/dt = dS/dt - dS^*/dt$. Since S^* is a constant, we have $dx_1/dt = dS/dt$ and, using a similar argument, $dx_2/dt = dI/dt$. So we now have

$$\dot{x}_1 = \mu N - \beta SI/N - \mu S \quad (4)$$

$$\dot{x}_2 = \beta SI/N - (\gamma + \mu)I. \quad (5)$$

But these equations are in terms of the original variables, S and I .

There are two ways in which we can then obtain the linearization. One is a calculus-free method, the other uses the idea of Taylor series from calculus.

2.1 Non-calculus method:

We rewrite the previous equations in terms of x_1 and x_2 using $S = S^* + x_1$ and $I = I^* + x_2$.

$$\dot{x}_1 = \mu N - \beta(S^* + x_1)(I^* + x_2)/N - \mu(S^* + x_1) \quad (6)$$

$$\dot{x}_2 = \beta(S^* + x_1)(I^* + x_2)/N - (\gamma + \mu)(I^* + x_2). \quad (7)$$

If we multiply out the brackets that appear in the infection term, we have

$$\beta SI/N = (\beta/N)(S^* + x_1)(I^* + x_2) \quad (8)$$

$$= (\beta/N)(S^*I^* + S^*x_2 + I^*x_1 + x_1x_2) \quad (9)$$

$$\approx (\beta/N)(S^*I^* + S^*x_2 + I^*x_1). \quad (10)$$

Where, to reach the last step we made use of the fact that we noted above: the product x_1x_2 is very small indeed and so can be ignored.

This leads to the following:

$$\dot{x}_1 = \mu N - \beta S^*I^*/N - \mu S^* + \beta(S^*x_2 + I^*x_1)/N - \mu x_1 \quad (11)$$

$$\dot{x}_2 = \beta S^*I^*/N - (\gamma + \mu)I^* + \beta(S^*x_2 + I^*x_1)/N - (\gamma + \mu)x_2. \quad (12)$$

Since (S^*, I^*) is an equilibrium point, the original model equations tell us that $\mu N - \beta S^*I^*/N - \mu S^* = 0$ and $\beta S^*I^*/N - (\gamma + \mu)I^* = 0$. This allows us to cancel some terms in these equations and we are left with the following linearized equations:

$$\dot{x}_1 = (\beta/N)(S^*x_2 + I^*x_1) - \mu x_1 \quad (13)$$

$$\dot{x}_2 = (\beta/N)(S^*x_2 + I^*x_1) - (\gamma + \mu)x_2. \quad (14)$$

2.2 Calculus method:

By using a Taylor series expansion, we can arrive a little more quickly at the linearization.

As a shorthand, we write the right hand side of the dS/dt equation as $f(S, I)$ (e.g. $f(S, I) = \mu N - \beta SI/N - \mu S$) and the right hand side of the dI/dt equation as $g(S, I)$. We then expand about the point (S^*, I^*) to give

$$\frac{dS}{dt} = f(S^*, I^*) + (S - S^*)\frac{\partial f}{\partial S} + (I - I^*)\frac{\partial f}{\partial I} + (\text{higher order terms}) \quad (15)$$

$$\frac{dI}{dt} = g(S^*, I^*) + (S - S^*)\frac{\partial g}{\partial S} + (I - I^*)\frac{\partial g}{\partial I} + (\text{higher order terms}). \quad (16)$$

Here, both partial derivatives are evaluated at the point (S^*, I^*) . (In case you aren't familiar with partial derivatives: when you work out $\partial f/\partial S$, you imagine that I is a constant. For the SIR model, $\partial f/\partial S = \beta I/N - \mu$ and $\partial f/\partial I = \beta S/N$.)

Since (S^*, I^*) is an equilibrium point, we have that $f(S^*, I^*)$ equals zero, since $dS/dt = 0$ at this point and $g(S^*, I^*) = 0$ since $dI/dt = 0$. Remembering that $x_1 = S - S^*$ and $x_2 = I - I^*$ and that

$dx_1/dt = dS/dt$ and $dx_2/dt = dI/dt$ we have

$$\frac{dx_1}{dt} = x_1 \frac{\partial f}{\partial S} + x_2 \frac{\partial f}{\partial I} \quad (17)$$

$$\frac{dx_2}{dt} = x_1 \frac{\partial g}{\partial S} + x_2 \frac{\partial g}{\partial I}, \quad (18)$$

where we have again ignored the higher order terms since they are of much smaller size. For the SIR model, this becomes

$$\dot{x}_1 = (\beta I^*/N - \mu)x_1 - (\beta S^*/N)x_2 \quad (19)$$

$$\dot{x}_2 = (\beta I^*/N)x_1 + (\beta S^*/N - \gamma - \mu)x_2. \quad (20)$$

These are the same equations that we had before. Once you are familiar with the process, it's very easy to obtain the linearized equations in this way.

2.3 Matrix Notation for the Linearization

We can write linearizations in matrix form:

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} \frac{\partial f}{\partial S} & \frac{\partial f}{\partial I} \\ \frac{\partial g}{\partial S} & \frac{\partial g}{\partial I} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad (21)$$

or in shorthand

$$\dot{\mathbf{x}} = J\mathbf{x}, \quad (22)$$

where J is the so-called Jacobian matrix, whose entries are the partial derivatives of the right hand sides of the differential equations describing the model, taken with respect to the different state variables of the model (e.g. S and I). We often write the entries of J as

$$J = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}. \quad (23)$$

We can do this linearization process for a model with any number of state variables: if there are n state variables, we get an n -dimensional set of coupled linear differential equations. (Bear in mind that for the SIR model there are three state variables, S , I and R , but our assumption that $S + I + R = N$ leads to our only having to consider a two dimensional system.)

3 What do we do with the linearization?

There is a well-developed theory for solving linear differential equations such as (22). We can only cover the briefest points here: for more information, find a book on differential equations or an introductory mathematical biology text. (You might start with chapter 5 of S. H. Strogatz's 'Nonlinear Dynamics and Chaos').

3.1 One Dimensional Case

It's perhaps simplest to start with the corresponding one-dimensional equation:

$$\dot{x} = \lambda x. \quad (24)$$

This equation has solution

$$x(t) = ce^{\lambda t}, \quad (25)$$

where c is the initial value of x (i.e. the value taken by x when $t = 0$). This equation describes exponential growth or decay.

If λ is greater than zero, then points move away from $x = 0$. Remembering that $x = 0$ corresponds to the equilibrium point, we see that non-zero points move away from the equilibrium as time passes: the equilibrium is unstable. If λ is less than zero, points move towards $x = 0$: the equilibrium is stable. If $\lambda = 0$, points neither move towards nor away from the equilibrium.

The sign of λ tells us about the stability of the equilibrium, and the size of λ tells us something about how quickly points move away from or towards the equilibrium. When we have a stable equilibrium, we sometimes talk about the **relaxation time**, which is defined to be $-1/\lambda$. This is the time taken for the distance between the point and the origin to decrease by a factor $1/e \approx 0.368$.

We can summarize the behaviors in the following figures, in which arrows denote the directions in which points move.



Figure 2: The left panel illustrates an unstable equilibrium, the right panel a stable equilibrium.

3.2 Two Dimensional Case

By analogy with the one dimensional case, we try a solution of the form

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} e^{\lambda t}, \quad (26)$$

where λ , v_1 and v_2 are constants. In shorthand notation we have

$$\mathbf{x}(t) = \mathbf{v}e^{\lambda t}. \quad (27)$$

If we differentiate $\mathbf{x}(t)$ directly, we see that its derivative is given by

$$\dot{\mathbf{x}}(t) = \lambda \mathbf{v}e^{\lambda t}. \quad (28)$$

But we know that $\mathbf{x}(t)$ has to satisfy $d\mathbf{x}/dt = J\mathbf{x}$, so we have

$$\dot{\mathbf{x}} = J(\mathbf{v}e^{\lambda t}). \quad (29)$$

$$= (J\mathbf{v})e^{\lambda t}. \quad (30)$$

Comparing (28) and (29) we see that $\mathbf{v}$ and λ must satisfy

$$J\mathbf{v} = \lambda\mathbf{v}. \quad (31)$$

In linear algebra, such vectors $\mathbf{v}$ are known as **eigenvectors** of the matrix J and the constants λ are known as **eigenvalues** of the matrix. Because these are important properties of matrices, there is quite a large theory devoted to calculating their values. (If you need to know more, a good place to start looking is a linear algebra textbook.) In the two dimensional case, the matrix J typically has two independent eigenvectors and two eigenvalues (although the two eigenvalues can be equal).

Because we have a two dimensional set of linear differential equations, the general solution is, in most cases ¹, given by the sum of two terms of the form (27): $\mathbf{x}(t) = c_1\mathbf{v}_1e^{\lambda_1 t} + c_2\mathbf{v}_2e^{\lambda_2 t}$. (The fact that one can add solutions together in this way is a fundamental property of linear systems, and one which makes them nice to work with.) The constants c_1 and c_2 are determined by the initial conditions. Notice that this form of the solution shows that the eigenvectors define two directions along which there may be exponential growth or decay. We will see this in action later on.

3.2.1 Finding the Eigenvalues and Eigenvectors in the 2D Case

We now need to find the eigenvalues and eigenvectors. Rearranging (31) gives us

$$(J - \lambda I)\mathbf{v} = 0, \quad (32)$$

where I is the two dimensional **identity matrix**

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \quad (33)$$

Theory from linear algebra shows that such an equation can only have a non-trivial solution (i.e. one in which $\mathbf{v}$ is not equal to zero) if the **determinant** of the matrix $J - \lambda I$ is equal to zero.

For a 2×2 matrix with entries a_{ij} (e.g. as in equation (23)), the determinant of the matrix is given by the product of the two entries on the leading diagonal minus the product of the two entries on the other diagonal: $a_{11}a_{22} - a_{12}a_{21}$.

Using this definition, we see that in order for λ to be an eigenvalue of the matrix J , it must satisfy

$$(a_{11} - \lambda)(a_{22} - \lambda) - a_{12}a_{21} = 0. \quad (34)$$

(Notice that the ‘ $-\lambda$ ’ terms arise because we have $J - \lambda I$, and the matrix I has entries equal to one in the (1,1) and (2,2) positions.)

Equation (34) is a quadratic equation for λ : notice that this gives us (in most cases) two possible answers, λ_1 and λ_2 . If we had an n -dimensional system, we would get a polynomial of degree n . When n is anything other than a small number, such equations become difficult (or impossible) to solve. But in the 2D case we can get an explicit expression for the eigenvalues. At worst, we can always use the ‘quadratic formula’, but we can do a little better than this.

¹The general solution can be more complex if the matrix J does not have a complete set of linearly independent eigenvectors. In such cases, the general solution has terms of the form $p(t)\mathbf{v}e^{\lambda t}$, where $p(t)$ is a polynomial in t .

Multiplying out (34) gives

$$\lambda^2 - (a_{11} + a_{22})\lambda + (a_{11}a_{22} - a_{12}a_{21}) = 0. \quad (35)$$

We see that the linear term involves $a_{11} + a_{22}$, the sum of the entries on the leading diagonal of J . This sum is called the **trace** of the matrix J , $\text{tr } J$. The constant term of this quadratic is just the determinant of the matrix J , $\det J$. So we can rewrite (35) succinctly as

$$\lambda^2 - \text{tr } J \lambda + \det J = 0. \quad (36)$$

Since we know that the numbers λ_1 and λ_2 satisfy the quadratic equation for λ , we know that we can factor this quadratic in the following way

$$(\lambda - \lambda_1)(\lambda - \lambda_2) = 0. \quad (37)$$

Multiplying out this equation, we get

$$\lambda^2 - (\lambda_1 + \lambda_2)\lambda + \lambda_1\lambda_2 = 0. \quad (38)$$

Comparing (36) and (38) shows that

$$\lambda_1 + \lambda_2 = \text{tr } J \quad (39)$$

$$\lambda_1\lambda_2 = \det J. \quad (40)$$

We can immediately write down the sum of the eigenvalues and their product in terms of the trace and determinant of J : quantities that we can easily work out from the entries of J .

The formula for the solution of a quadratic equation, applied to (36) gives

$$\lambda = -\frac{1}{2} \text{tr } J \pm \frac{1}{2} \sqrt{(\text{tr } J)^2 - 4 \det J}. \quad (41)$$

Notice that an important difference between the 1D and 2D cases is that we can now have **complex** eigenvalues. This happens if the term inside the square root is negative, i.e. if $(\text{tr } J)^2 < 4 \det J$.

If we have a complex eigenvalue, which we can write in terms of its real and imaginary parts as $\lambda_1 = \rho + i\omega$, then it is fairly straightforward to show that $\lambda_2 = \rho - i\omega$ is also an eigenvalue. These eigenvalues have equal real parts but their imaginary parts have opposite signs: this is called a **complex conjugate pair**. There is a standard result that says if you have a polynomial whose coefficients are real, then complex eigenvalues always come in such complex conjugate pairs. (Notice that this result applies to our situation: the entries of the Jacobian matrix are real numbers.)

Notice that in the complex eigenvalue case, $\exp(\lambda t)$ is of the form $\exp\{(\rho \pm i\omega)t\}$, which can be multiplied out to give $\exp(\rho t) \exp(\pm i\omega t)$. The magnitude of $\exp(\lambda t)$ is given by the term $\exp(\rho t)$: the real part of the eigenvalue, ρ , describes whether one has growth or decay. In the decay case, the relaxation time is equal to $-1/\rho$. Since one can write

$$\exp(i\omega t) = \cos(\omega t) + i \sin(\omega t), \quad (42)$$

we see that the term $\exp(i\omega t)$ describes oscillatory behavior with constant amplitude, with (angular) frequency ω . The period of the oscillation is given by $2\pi/\omega$.

3.3 Complete Characterization of Behavior in 2D Case

We now make use of the above analysis of the eigenvalues in 2D to characterize all possible cases. We make use of the fact that $\det J = \lambda_1 \lambda_2$ and $\text{tr } J = \lambda_1 + \lambda_2$ to find the signs of the eigenvalues (or of their real parts in the complex case), as it is these quantities that determine stability. We will leave discussion of so-called ‘borderline’ cases until the end: these borderline cases involve ones in which either both eigenvalues are equal or one (or more) eigenvalues are equal (or have real part equal) to zero.

- **Both eigenvalues are real:** $(\text{tr } J)^2 > 4 \det J$

- $\text{tr } J > 0$

This means that the sum of the eigenvalues is positive.

There are two possible cases:

- $\det J > 0$

The product of the eigenvalues is also positive. This means that both λ_1 and λ_2 must be positive. In terms of the differential equation, we have exponential growth in the directions represented by both $\mathbf{v}_1$ and $\mathbf{v}_2$. We call this situation an **unstable node**. If we order the eigenvalues so that $\lambda_1 > \lambda_2 > 0$, then we see that the growth is faster in the direction defined by $\mathbf{v}_1$ and so trajectories starting close to the equilibrium point tend to move away from the equilibrium in this direction.

- $\det J < 0$

The product of the eigenvalues is negative. This means that we must have one positive (call it λ_1) and one negative (λ_2) eigenvalue. We have exponential growth in the direction represented by λ_1 but exponential decay in the direction represented by λ_2 . This is called a saddle: points move towards the equilibrium in one direction, but away from the equilibrium in the other.

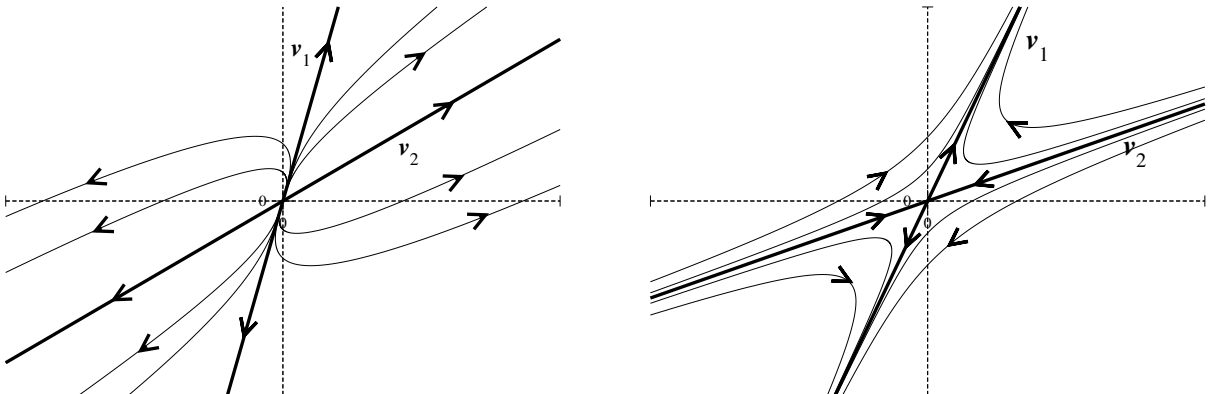


Figure 3: Left panel illustrates a stable node, with $\lambda_1 > \lambda_2 > 0$. Notice that trajectories leave the equilibrium in the direction of $\mathbf{v}_1$. Right panel illustrates a saddle, with $\lambda_1 > 0 > \lambda_2$. Notice that trajectories approach the equilibrium in the direction of $\mathbf{v}_2$, but leave in the direction of $\mathbf{v}_1$.

- $\text{tr } J < 0$

This means that the sum of the eigenvalues is negative.

There are two possible cases:

- $\det J > 0$

The product of the eigenvalues is positive. This means that both λ_1 and λ_2 must be negative. In terms of the differential equation, we have exponential decay in the directions represented by both $\mathbf{v}_1$ and $\mathbf{v}_2$. We call this situation a **stable node**. If we order the eigenvalues so that $\lambda_1 > \lambda_2$ (i.e. λ_2 is more negative), we see that the contraction in the $\mathbf{v}_2$ direction occurs more quickly than in the $\mathbf{v}_1$ direction. This means that trajectories approach the equilibrium along the line defined by the vector $\mathbf{v}_1$.

- $\det J < 0$

The product of the eigenvalues is negative. This means that we must have one positive (call it λ_1) and one negative (λ_2) eigenvalue. As in the situation above, this corresponds to a **saddle**.

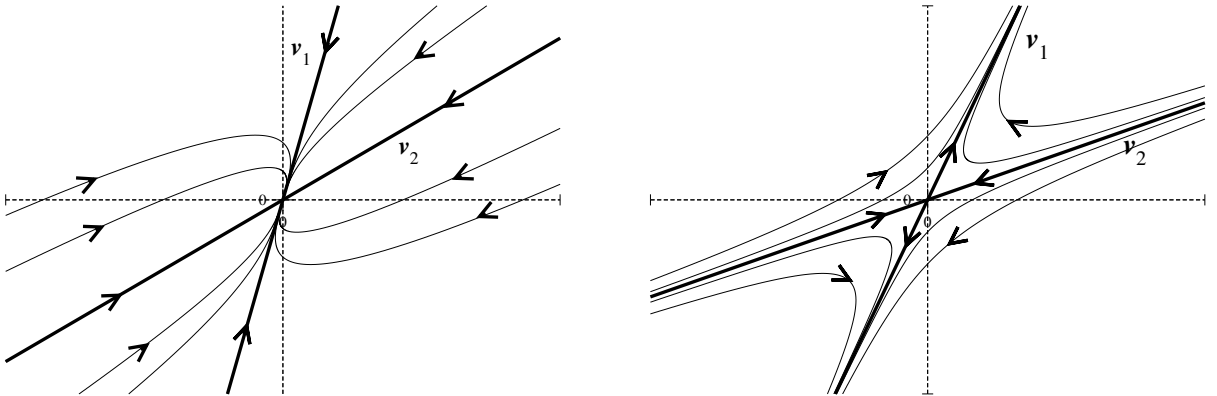


Figure 4: Left panel illustrates a stable node, with $0 > \lambda_1 > \lambda_2$. Notice that trajectories approach the equilibrium in the direction of $\mathbf{v}_1$. Right panel illustrates a saddle.

- **Eigenvalues are Complex:** $(\text{tr } J)^2 < 4 \det J$

In this case $\text{tr } J$ equals the sum of the real parts of the two eigenvalues. But we know that these eigenvalues occur as a complex conjugate pair and so have equal real parts. The trace of J tells us whether the real parts of the eigenvalues are positive or negative.

- $\text{tr } J > 0$

The eigenvalues have positive real part, so we have exponential growth. Points move away from the equilibrium point, but do so in an oscillatory manner as discussed above. We call this an **unstable spiral**. Notice that the spiral may be elliptical in nature.

- $\text{tr } J < 0$

The eigenvalues have negative real part, so we have exponential decay. Points move towards the equilibrium point, but do so in an oscillatory manner as discussed above. We call this a **stable spiral**.

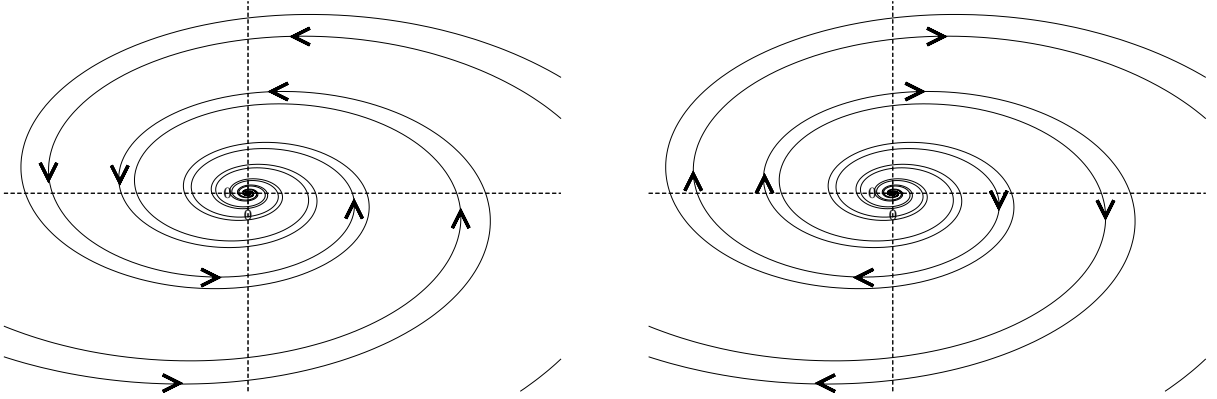


Figure 5: Stable (left panel) and unstable (right panel) spirals.

3.4 Borderline Cases

As mentioned above, borderline cases involve repeated eigenvalues (in which case we have $4 \det J = (\text{tr } J)^2$), situations in which one or more eigenvalues are zero ($\det J = 0$ and $\text{tr } J \leq 0$) or have zero real part ($\det J \geq 0$, $\text{tr } J = 0$).

If one or more eigenvalues is equal to zero, or have real parts equal to zero, then there is no motion in the direction(s) defined by the corresponding eigenvector(s). An important example is the **center**, when there is a pair of eigenvalues that are purely imaginary (recall that the signs of their imaginary parts must be opposite). The resulting motion involves points moving around on ellipses: there is no net motion towards or away from the equilibrium.

In the case when there is one zero eigenvalue and one real non-zero eigenvalue, points just move parallel to the direction defined by the eigenvector corresponding to the non-zero eigenvalue. They either move towards or away from the line described by the other eigenvector, depending on whether the sign of the eigenvalue is negative or positive. In this case, there is an entire line of equilibrium points.

The other borderline cases correspond to repeated eigenvalues, and can be viewed as limiting cases of nodes. If there are repeated eigenvalues, there may or may not be linearly independent eigenvectors. If there are, then the resulting equilibrium point is described as a **stable star** or an **unstable star** (depending on the sign of the eigenvalue). This is just like a node, but where the two rates of contraction or expansion are equal. The second possibility is that there is only one eigenvector. The resulting equilibrium is then known as a **type II node**.

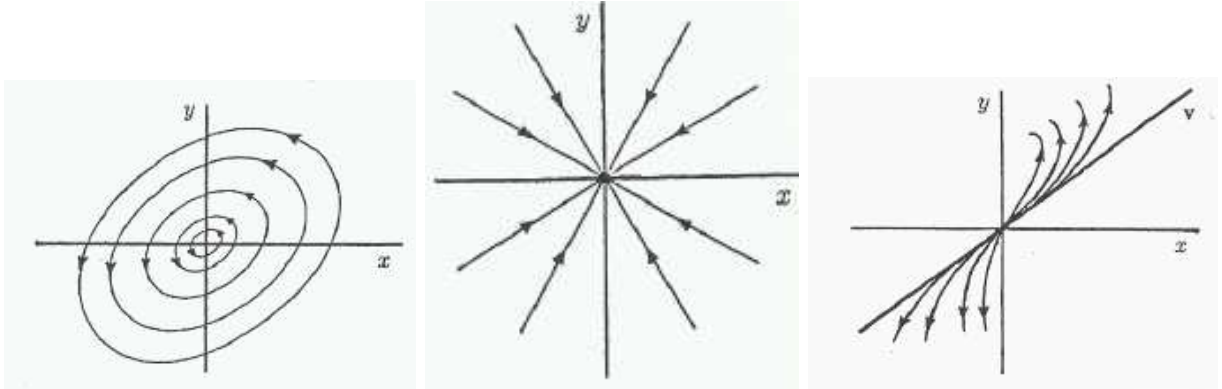


Figure 6: Borderline cases: the center, the star (stable star shown) and type II node (unstable shown). These figures were taken from Appendix A of J.D. Murray's 'Mathematical Biology'.

4 Relationship Between Nonlinear and Linearized Behavior

Some important results tell us when we can expect the behavior of the linearization to give a qualitatively correct picture of the behavior of the full nonlinear system in the neighborhood of an equilibrium point.

Provided that we are not in the borderline cases mentioned above paragraph (i.e. the linearization describes a saddle, a node or a spiral) then the picture we get from the linearized model gives a very good description of the behavior of the nonlinear model, at least in the neighborhood of the equilibrium. In the borderline cases (e.g. center, star, type II node), the nonlinear terms will change the picture.

The linearization correctly characterizes the **stability** of different directions provided that the corresponding eigenvalue is not zero or has real part equal to zero. In particular, if the eigenvalue with largest real part is non-zero, our linearized analysis will correctly predict if the equilibrium is stable or is unstable. Our linearization is uninformative regarding stability if this eigenvalue (the **dominant eigenvalue**) has zero real part: in such cases, nonlinear terms determine stability.

Taken together, these statements explain why we didn't give so much attention to the borderline cases. For instance, a stable star in the linear picture will correspond to a stable equilibrium in the nonlinear model, but the nonlinearity will mean that it may exhibit node-like or spiral-like behavior. A center has neutral stability in the linear picture: nonlinear terms can lead to stable or unstable behavior.

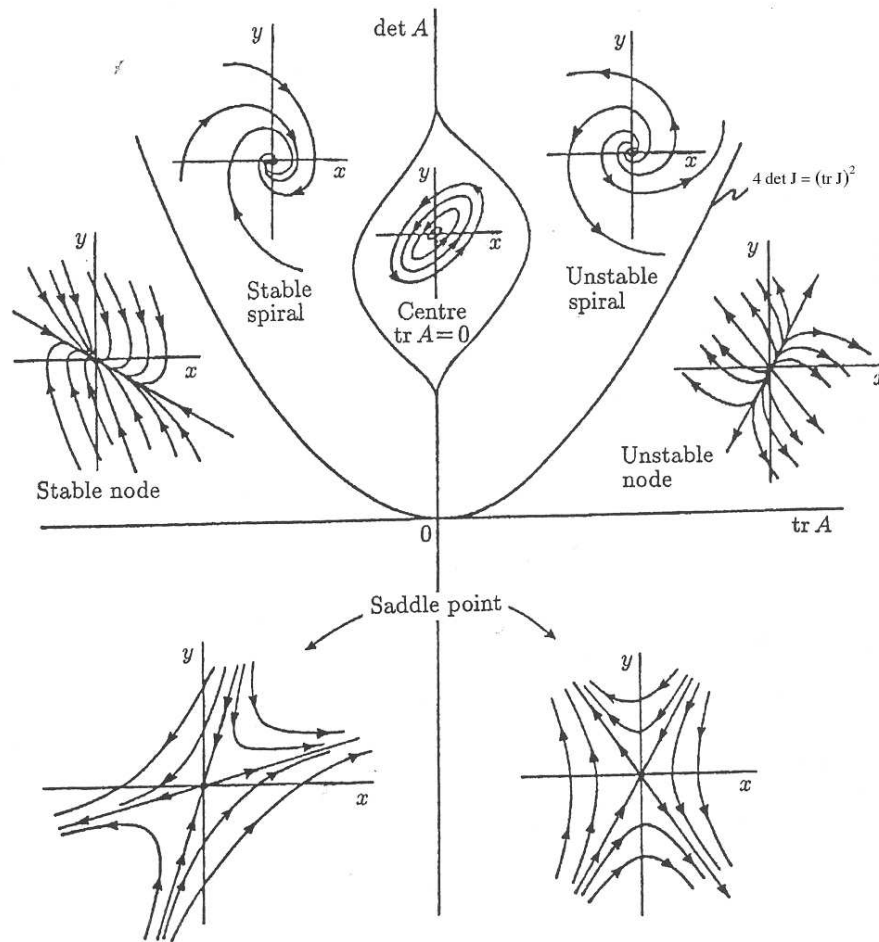


Figure A.2. Summary diagram showing how $\text{tr } A$ and $\det A$, where A is the linearisation matrix given by (A.4), determine the type of phase plane singularity for (A.1). Here $\det A = f_x g_y - f_y g_x$, $\text{tr } A = f_x + g_y$, where the partial derivatives are evaluated at the singularities, the solutions of $f(x, y) = g(x, y) = 0$.

Figure 7: Summary of the two dimensional case: dynamical behavior in terms of the determinant and trace of the Jacobian matrix. This figure was taken from Appendix A of J.D. Murray's 'Mathematical Biology'.

Unit 14

Lyapunov Stability

Whereas I/O stability is concerned with the effect of inputs on outputs, Lyapunov stability deals with unforced systems:

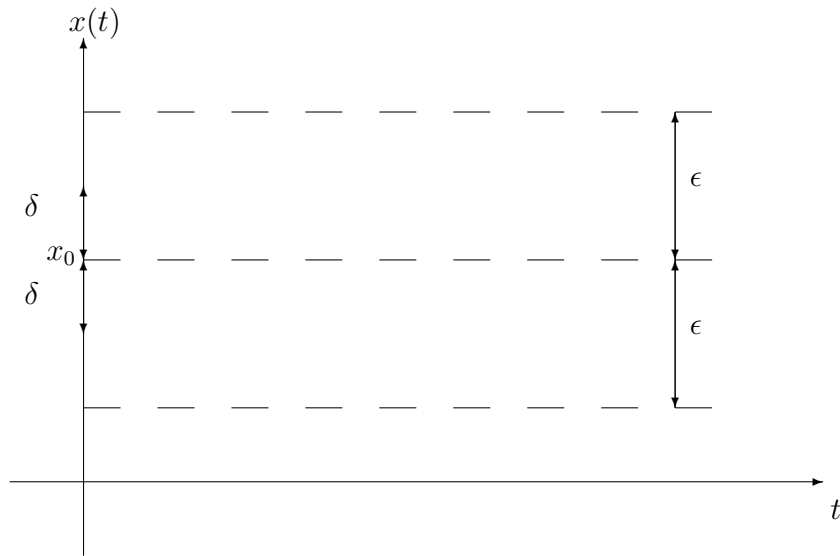
$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t) \quad (1)$$

where $\mathbf{x} \in R^n$, $t \in R^+$, and $\mathbf{f} : R^n \times R^+ \rightarrow R^n$. The system in (1) is autonomous (time invariant) if $\mathbf{f} = \mathbf{f}(\mathbf{x})$. Otherwise, it is nonautonomous (time varying).

If $\mathbf{f}(\mathbf{x}_0, t) = \mathbf{0}$, then $\mathbf{x}_0$ is an equilibrium. Then, the unique solution of Eq. (1) with $\mathbf{x}(0) = \mathbf{x}_0$ is $\mathbf{x}(t) = \mathbf{x}_0$.

Definition. The equilibrium $\mathbf{x}_0$ is *stable* (or L-stable or stable in the sense of Lyapunov) if for $\epsilon > 0$ there exists $\delta > 0$ such that

$$\|\mathbf{x}(0) - \mathbf{x}_0\| < \delta \Rightarrow \|\mathbf{x}(t) - \mathbf{x}_0\| < \epsilon$$



The equilibrium $\mathbf{x}_0$ is *unstable* if it is not stable.

The equilibrium $\mathbf{x}_0$ is *asymptotically stable* if it is stable and there exists $\delta > 0$ such that

$$\|\mathbf{x}(0) - \mathbf{x}_0\| < \delta \Rightarrow \mathbf{x}(t) \rightarrow \mathbf{x}_0 \text{ as } t \rightarrow \infty$$

The equilibrium $\mathbf{x}_0$ is *globally asymptotically stable* if it is stable and

$$\mathbf{x}(t) \rightarrow \mathbf{x}_0 \text{ as } t \rightarrow \infty$$

Consider the linear time-invariant system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}, \quad \mathbf{A} \in R^{n \times n} \quad (2)$$

The eigenvalues of $\mathbf{A}$ are λ_i , $i = 1, \dots, n$. The solution of (2) is

$$\mathbf{x}(t) = \exp(\mathbf{A}t)\mathbf{x}(0) \quad (3)$$

where

$$\exp(\mathbf{A}t) = \sum_{k=0}^{\infty} \frac{1}{k!} (\mathbf{A}t)^k = \mathbf{1} + \mathbf{A}t + \frac{1}{2}\mathbf{A}^2t^2 + \dots$$

which converges for $t \in [0, \infty]$. If $\mathbf{A}$ has distinct eigenvalues λ_i , and eigenvectors $\mathbf{e}_i$, then we have the following eigendecomposition:

$$\begin{aligned} \mathbf{A} &= \mathbf{E}\mathbf{\Lambda}\mathbf{E}^{-1} \\ \mathbf{E} &= \text{row}\{\mathbf{e}_i\} \\ \mathbf{\Lambda} &= \text{diag}\{\lambda_i\} \end{aligned}$$

In this case, it is not hard to show that

$$\begin{aligned} \mathbf{A}^k &= \mathbf{E}\mathbf{\Lambda}^k\mathbf{E}^{-1} \\ \mathbf{\Lambda}^k &= \text{diag}\{\lambda_i^k\} \\ \exp(\mathbf{A}t) &= \mathbf{E} \exp(\mathbf{\Lambda}t) \mathbf{E}^{-1} \\ \exp(\mathbf{\Lambda}t) &= \text{diag}\{\exp(\lambda_i t)\} \end{aligned}$$

We have the following stability results for the system in Eq. (2) and its equilibrium $\mathbf{x}_0 = \mathbf{0}$.

- (a) The system is stable if $\text{Re}\{\lambda_i\} \leq 0$, $i = 1 \dots n$, and there are no repeated eigenvalues on the imaginary axis.
- (b) The system is unstable if there is at least one λ_i with $\text{Re}\{\lambda_i\} > 0$.
- (c) The linear system is (globally) asymptotically stable if $\text{Re}\{\lambda_i\} < 0$, $i = 1 \dots n$.

2.1 Lyapunov's Linearization (or First) Method

Consider the system $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ with $\mathbf{f}(\mathbf{x}_0) = \mathbf{0}$. Let $\mathbf{x}(t) = \mathbf{x}_0 + \delta\mathbf{x}(t)$. Then,

$$\begin{aligned} \dot{\mathbf{x}} = \delta\dot{\mathbf{x}} &= \mathbf{f}(\mathbf{x}_0 + \delta\mathbf{x}) \\ &= \mathbf{f}(\mathbf{x}_0) + \left. \frac{\partial \mathbf{f}}{\partial \mathbf{x}^T} \right|_{\mathbf{x}=\mathbf{x}_0} \delta\mathbf{x} + \mathcal{O}(\|\delta\mathbf{x}\|^2) \end{aligned}$$

Neglecting higher order terms, we have the linearization

$$\delta\dot{\mathbf{x}} = \mathbf{A}\delta\mathbf{x}, \quad \mathbf{A} = \left. \frac{\partial \mathbf{f}}{\partial \mathbf{x}^T} \right|_{\mathbf{x}=\mathbf{x}_0} \quad (4)$$

If the equilibrium $\delta\mathbf{x} = \mathbf{0}$ of (4) is:

- (a) asymptotically stable then $\mathbf{x}_0$ is an asymptotically stable equilibrium of the nonlinear system;
- (b) unstable then $\mathbf{x}_0$ is an unstable equilibrium of the nonlinear system;
- (c) stable then no conclusion can be drawn about the nonlinear system.

2.2 Lyapunov's Direct or Second Method

Consider the system in Eq. (1) with $\mathbf{x} = \mathbf{x}_0$ an equilibrium. We assume that $\mathbf{x}_0 = \mathbf{0}$ which can be accomplished with a change of coordinates.

A function $V(\mathbf{x}, t)$ is C^1 if it is continuously differentiable. It is a *locally positive-definite function* [lpdf] if $V(\mathbf{0}, t) \equiv 0$ and there exists $r > 0$ such that $V(\mathbf{x}, t) > 0, \forall \mathbf{x} \neq \mathbf{0} \text{ s.t. } \|\mathbf{x}\| < r$. It is a *positive-definite function* [pdf] if $V(\mathbf{0}, t) \equiv 0$ and $V(\mathbf{x}, t) > 0, \forall \mathbf{x} \in R^n, \mathbf{x} \neq \mathbf{0}$.

It is *radially unbounded* if

$$V(\mathbf{x}, t) \rightarrow \infty \text{ as } \|\mathbf{x}\| \rightarrow \infty$$

It is *locally negative definite* if $-V$ is a lpdf.

It is *negative definite* if $-V$ is a pdf.

Let $B_r = \{\mathbf{x} \in R^n \mid \|\mathbf{x}\| < r\}$.

Examples

$V(x) = 1 - \cos x$ is a lpdf but it is not radially unbounded.

$V(x) = x^2$ is a pdf and it is radially unbounded.

$V(\mathbf{x}) = \mathbf{x}^T \mathbf{P} \mathbf{x}$ is a pdf and radially unbounded if $\mathbf{P}$ is a symmetric positive-definite matrix.

Theorem 1. The equilibrium $\mathbf{x}_0 = \mathbf{0}$ of (1) is stable if there exists a C^1 lpdf $V(\mathbf{x}, t)$ and $r > 0$ such that

$$\dot{V}(\mathbf{x}, t) \leq 0, \quad \forall t \geq 0, \quad \forall \mathbf{x} \in B_r$$

where $\dot{V}$ is evaluated along the trajectories of (1), *i.e.*,

$$\dot{V} = \frac{\partial V}{\partial t} + \frac{\partial V}{\partial \mathbf{x}^T} \mathbf{f}(\mathbf{x}, t)$$

V is called a Lyapunov function.

Example. Consider the equation of the simple pendulum

$$\ddot{\theta} + \sin \theta = 0$$

Let $x_1 = \theta$, $x_2 = \dot{\theta}$ so that

$$\begin{aligned}\dot{x}_1 &= x_2 \\ \dot{x}_2 &= -\sin x_1\end{aligned}$$

If we adopt $V(x_1, x_2) = (1 - \cos x_1) + \frac{1}{2}x_2^2$ as a Lyapunov function we have

$$\begin{aligned}\dot{V} &= \dot{x}_1 \sin x_1 + x_2 \dot{x}_2 \\ &= x_2 \sin x_1 - x_2 \sin x_1 \\ &= 0\end{aligned}$$

Hence $\mathbf{x} = \mathbf{0}$ is stable.

Example. Consider Euler's equation $\mathbf{I}\dot{\boldsymbol{\omega}} + \boldsymbol{\omega}^\times \mathbf{I}\boldsymbol{\omega} = \mathbf{0}$ where $\mathbf{I} = \mathbf{I}^T > \mathbf{0}$. Letting $V = \frac{1}{2}\boldsymbol{\omega}^T \mathbf{I}\boldsymbol{\omega}$, we have

$$\begin{aligned}\dot{V} &= \boldsymbol{\omega}^T \mathbf{I}\dot{\boldsymbol{\omega}} \\ &= -\boldsymbol{\omega}^T \boldsymbol{\omega}^\times \mathbf{I}\boldsymbol{\omega} \\ &= 0\end{aligned}$$

Hence, $\boldsymbol{\omega} = \mathbf{0}$ is a stable equilibrium.

Theorem 2. The equilibrium $\mathbf{x}_0 = \mathbf{0}$ of Eq. (1) is asymptotically stable if there exists a C^1 lpdf V such that $-\dot{V}$ is a lpdf.

Theorem 3. The equilibrium $\mathbf{x}_0 = \mathbf{0}$ of Eq. (1) is globally asymptotically stable if there exists a C^1 pdf V such that V is radially unbounded and $-\dot{V}$ is a pdf.

Note: Theorems 1-3 give sufficient conditions for stability.

Example. Consider the system

$$\dot{x} = -ax, \quad a > 0$$

Letting $V = \frac{1}{2}x^2 > 0$, $x \neq 0$, we have $\dot{V} = x\dot{x} = -ax^2 < 0$, $x \neq 0$. Therefore $x = 0$ is a globally asymptotically stable equilibrium.

Example. Consider the system

$$\begin{aligned} \mathbf{M}\ddot{\mathbf{q}} + \mathbf{K}\mathbf{q} &= \mathbf{u}(t), \quad \mathbf{u} = -\mathbf{D}\dot{\mathbf{q}} \\ \mathbf{M} &= \mathbf{M}^T > \mathbf{O}, \mathbf{K} = \mathbf{K}^T > \mathbf{O}, \mathbf{D} = \mathbf{D}^T > \mathbf{O} \end{aligned}$$

Letting $\mathbf{x}_1 = \mathbf{q}$, $\mathbf{x}_2 = \dot{\mathbf{q}}$, we have

$$\begin{aligned} \dot{\mathbf{x}}_1 &= \mathbf{x}_2 \\ \dot{\mathbf{x}}_2 &= -\mathbf{M}^{-1}\mathbf{K}\mathbf{x}_1 - \mathbf{M}^{-1}\mathbf{D}\mathbf{x}_2 \end{aligned}$$

Let

$$V = \frac{1}{2}\dot{\mathbf{q}}^T\mathbf{M}\dot{\mathbf{q}} + \frac{1}{2}\mathbf{q}^T\mathbf{K}\mathbf{q} = \frac{1}{2}[\mathbf{x}_1^T \quad \mathbf{x}_2^T] \begin{bmatrix} \mathbf{K} & \mathbf{O} \\ \mathbf{O} & \mathbf{M} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} > 0 \quad (\mathbf{x} \neq \mathbf{0})$$

which is a radially unbounded pdf. Then,

$$\dot{V} = \dot{\mathbf{q}}^T(\mathbf{M}\ddot{\mathbf{q}} + \mathbf{K}\mathbf{q}) = -\dot{\mathbf{q}}^T\mathbf{D}\dot{\mathbf{q}} = -[\mathbf{x}_1^T \quad \mathbf{x}_2^T] \begin{bmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \leq 0$$

Hence, $\mathbf{x} = \mathbf{0}$ is a stable equilibrium. It is in fact globally asymptotically stable. How do we show this when $\dot{V}$ does not contain $\mathbf{q}$?

Unit 15

A periodic solution or orbit corresponds to a special type of solution for a dynamical system, namely one which repeats itself in time. Many dynamical systems exhibit periodic orbits: the motion of planets around the Sun, the small oscillations of a simple pendulum and an oscillator circuit. Consider a system defined on the domain $D \subset \mathbb{R}^2$

$$\begin{aligned}\dot{x} &= f(x), \quad x \in \mathbb{R}^n \\ x(0) &= x_0.\end{aligned}\tag{23.1}$$

Definition 23.0.4 *A non-trivial solution $x(t, x_0)$ of (23.1) is said to be periodic if there exists a time $T > 0$ such that*

$$x(t + T, x_0) = x(t, x_0) \quad \forall t \geq 0.\tag{23.2}$$

Solutions of (23.1) with initial condition belonging to the equilibrium set correspond to trivial solutions that satisfy (23.2) for all $t \geq 0$ and are ruled out from the definition of a periodic solution. Periodic solutions correspond to closed trajectories in the phase portrait. Such a trajectory is called a *periodic orbit* or a *closed orbit*. Further, $x(t)$ is said to be bounded if there exists a constant $\mu > 0$ such that $\|x(t, x_0)\| \leq \mu$ for all $t \geq 0$.

Consider the simple harmonic oscillator given by the equation

$$\begin{aligned}\dot{x}_1 &= x_2 \\ \dot{x}_2 &= -x_1\end{aligned}\tag{23.3}$$

The phase portrait of this system shows continuum of closed orbits and hence it can be concluded that the system has periodic solutions. In Figure 23.1, the closed orbits correspond to two different initial conditions. It can be observed that a linear oscillator, apart from the fact that it cannot be practically realized, suffers from two major drawbacks:

- The amplitude of oscillation depends on the initial condition.

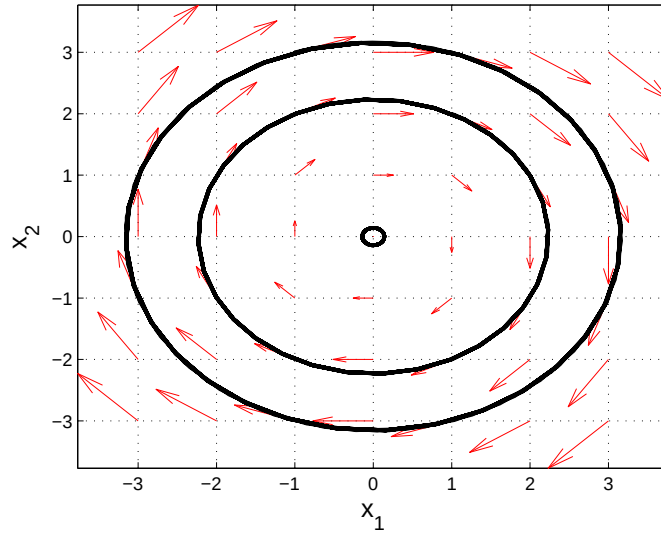


Figure 23.1: Vector field plot of linear harmonic oscillator

- The frequency of oscillation is structurally unstable.

It is thus natural to seek an oscillator that can be practically realized, is structurally stable and whose amplitude of oscillation is invariant with respect to the initial condition. One such oscillator is the Van der Pol oscillator given by

$$\begin{aligned}\dot{x}_1 &= x_2 \\ \dot{x}_2 &= -x_1 + \epsilon(1 - x_1^2)x_2\end{aligned}\tag{23.4}$$

where $\epsilon > 0$. A typical vector-field plot along with the closed orbit is depicted in Figure 23.2. Observe that the Van der Pol oscillator (23.4) exhibits a unique periodic solution. Further, all non-trivial trajectories originating within the closed orbit converge to it, and all trajectories originating outside the periodic solution also converge to it. Such a closed orbit is termed stable and in later chapters we will in a precise manner, define stable closed orbits. An isolated periodic orbit is referred to by the special name *limit cycle*. Further, a limit cycle is compact and invariant.

We end this lecture with the notion of attracting set. It helps to know what is meant by the neighbourhood of a set. An ϵ -neighbourhood N of a set $M \subset D$ is $N \triangleq \{x \in D : \|x - y\| < \epsilon, y \in M\}$ for some $\epsilon > 0$.

Definition 23.0.5 A closed invariant set $M \subset D$ is an attracting set of (23.1) if there exists a neighbourhood N of M such that, for all $x \in N$, $\phi(t, x) \in N \quad \forall t \geq 0$ and $\phi(t, x) \longrightarrow M$ as $t \longrightarrow \infty$.

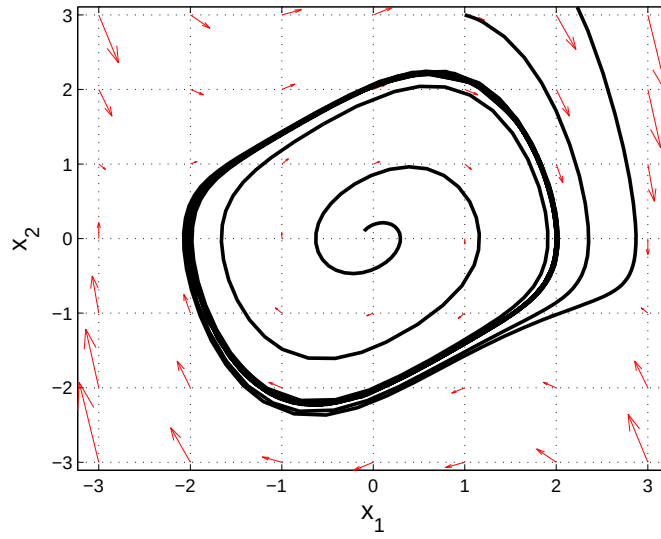


Figure 23.2: Vector field plot of a Van der Pol oscillator with $\epsilon = 0.5$

Definition 23.0.6 *A set $M \subset D$ is an attractor of (23.1) if M is an attracting set of D and contains a dense orbit.*

The limit cycle of the Van der Pol Oscillator (23.2) is an example of an attractor.

Periodic solutions on $\mathbb{R}^2$

When confined to $\mathbb{R}^2$, sufficient conditions exist to *guarantee* or *rule out* the existence of periodic solutions. These theorems are based on the concept of bounded trajectories and positive invariance of a set. We now have the requisite tools to understand the following Lemma.

Existence of periodic solutions

Lemma 24.0.7 The Poincaré-Bendixson criterion

Let P be a closed, bounded subset of $\mathbb{R}^2$. Poincaré-Bendixson criteria states that P contains a periodic solution of (23.1) if it satisfies the following conditions:

- 1. P contains no equilibrium points or contains only unstable equilibrium points (unstable node/focus).*
- 2. P is positive invariant with respect to the trajectories of (23.1).*

Note that the above criteria is only a *sufficient* condition and further it does not say anything about the uniqueness of the periodic solutions nor about the number of periodic solutions. Thus it cannot be used to rule-out the existence of limit cycles. While the first criteria is easy to check, the second one requires the test of positive invariance. For nonlinear systems, one way is to characterize the region enclosed by a simple closed curve defined by $V(x) = c, c > 0$ where, V is a smooth function. Differentiating V along the trajectory of (23.1), we obtain

$$\dot{V} = \frac{\partial V}{\partial x} f(x) = \left\langle \frac{\partial V}{\partial x}, f(x) \right\rangle$$

where, $\frac{\partial V}{\partial x}$ is the gradient of V and $f(x)$ the tangent vector at x . Recall that the gradient of a scalar function is a vector field which points in the direction of the greatest rate of increase of the scalar function. Thus the direction of $\frac{\partial V}{\partial x}$ is as shown in Figure 24.1. The inner product formula $\langle \frac{\partial V}{\partial x}, f(x) \rangle = \cos \theta \|\frac{\partial V}{\partial x}\| \|f(x)\|$ allows us to determine the

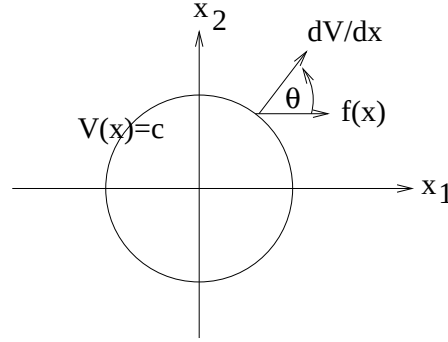


Figure 24.1: Gradient of V and the tangent vector at a point x on the closed curve $V(x) = c$

direction of $f(x)$ with respect to the gradient of V . We have

- $\dot{V} \leq 0$ whenever $f(x)$ points inwards at a point x on the curve $V(x) = c$, implying trajectories are trapped inside the set $P \triangleq \{x \in \mathbb{R}^2 : V(x) \leq c\}$.
- $\dot{V} \geq 0$ whenever $f(x)$ points outwards at a point x on the curve $V(x) = c$, implying trajectories leave the set P .
- $\dot{V} = 0$ whenever $f(x)$ is tangent to the curve $V(x) = c$ at x .

For the region bounded by the curve $V(x) = c$ to be positively invariant the tangent vector should point inwards or be tangential to $V(x) = c$ at all points. The test for the existence of periodic solutions using Poincaré Bendixson's criteria is applied to the following examples.

Example 24.0.8 Consider the system given by

$$\begin{aligned} \dot{x}_1 &= x_1 + x_2 - x_1(x_1^2 + x_2^2) \\ \dot{x}_2 &= -2x_1 + x_2 - x_2(x_1^2 + x_2^2) \end{aligned} \quad (24.1)$$

It can be seen that $(0,0)$ is the unique equilibrium point of this system. Further, the linearized model about the origin is given by

$$\dot{x} = \begin{bmatrix} 1 & 1 \\ -2 & 1 \end{bmatrix} x$$

and has the eigen spectrum $\{1 \pm i\sqrt{2}\}$ and thus the origin is an unstable focus of the nonlinear system. Let $V(x) = x_1^2 + x_2^2 = c$. Then

$$\begin{aligned}\dot{V} &= \frac{\partial V}{\partial x} \dot{x} \\ &= 2c - 2c^2 - 2x_1x_2 \\ &\leq 3c - 2c^2\end{aligned}$$

where we have used the inequality $|2x_1x_2| \leq x_1^2 + x_2^2$. Hence the region $P \triangleq \{x \in \mathbb{R}^2 : V(x) \leq c\}$ is positive invariant provided $c \geq 3/2$. By Poincaré Bendixson's criteria, P contains periodic solutions. The phase-portrait in Figure 24.2 reveals a stable limit cycle.

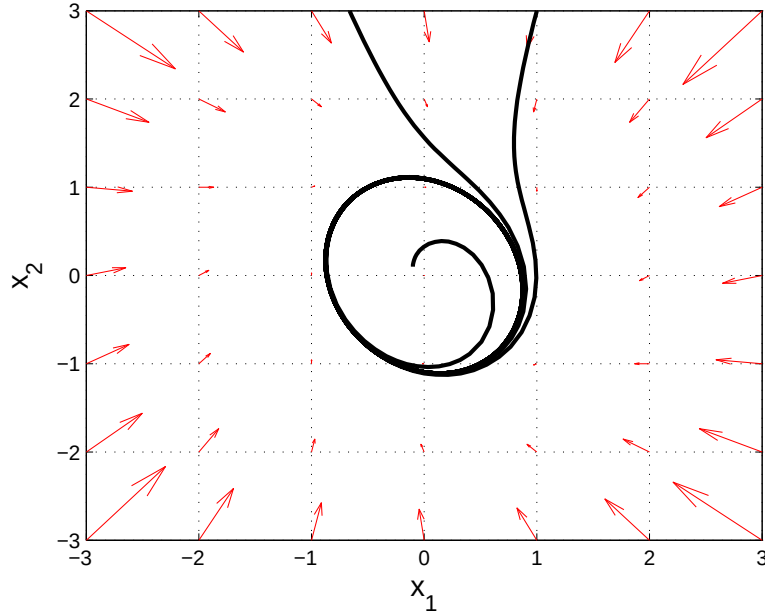


Figure 24.2: A stable limit cycle

Example 24.0.9 Consider the system given by

$$\begin{aligned}\dot{x}_1 &= -x_1 + x_2 \\ \dot{x}_2 &= -x_1 - x_2\end{aligned}\tag{24.2}$$

Consider the region P bounded by the curve $V(x) = x_1^2 + x_2^2 = c$. The derivative of $V(x)$ along the trajectories of the system yields

$$\dot{V} = 2x_1(-x_1 + x_2) + 2x_2(-x_1 - x_2) = -2c < 0.$$

Hence P is positive invariant and the second condition of the Poincaré Bendixson criterion is satisfied. The system has an unique equilibrium point at $(0,0)$ and is a stable focus, thus violating the first condition. Hence the criterion is not applicable. In fact it can be seen that all trajectory entering P will converge to this stable focus. This example brings out the importance of the set P not containing any stable equilibrium points even if it is positively invariant.

Example 24.0.10 Consider the linear oscillator given by

$$\begin{aligned}\dot{x}_1 &= x_2 \\ \dot{x}_2 &= -x_1\end{aligned}$$

The system has a center at $(0,0)$. Hence we consider a region in the form of an annular ring about the origin given by $P = \{x \in \mathbb{R}^2 : c_1 \leq V(x) \leq c_2, c_1, c_2 > 0\}$ where, $V(x) = x_1^2 + x_2^2$. This choice of P excludes the non-hyperbolic equilibrium from it. Differentiating $V(x)$ along the trajectories of the system yields $\dot{V} = 0$. Hence the vector $f(x)$ is tangential to all points of the curve $V(x) = c$ implying the annular ring is positively invariant, establishing the existence of periodic solution inside it. Recall, the linear oscillator exhibits a continuum of periodic solutions.

Analytical expressions for closed-orbits

In certain cases the periodic solution of a system can be obtained in an analytic form as show by the following examples.

Example 25.0.11 *Considering again the linear oscillator (23.1), let us assume $V(x)$ of the form $V(x) = h_1(x_1) + h_2(x_2)$. Differentiating $V(x)$ along the trajectories of (23.1), we require*

$$\begin{aligned}\dot{V} &= \frac{\partial h_1}{\partial x_1}x_2 + \frac{\partial h_2}{\partial x_2}(-x_1) = 0 \\ \Rightarrow \frac{1}{x_1} \frac{\partial h_1}{\partial x_1} &= \frac{1}{x_2} \frac{\partial h_2}{\partial x_2} = c\end{aligned}$$

where, c is a constant. Therefore $\frac{\partial h_1}{\partial x_1} = cx_1$ and $\frac{\partial h_2}{\partial x_2} = cx_2$. Integrating, we obtain $h_1 = cx_1^2$ and $h_2 = cx_2^2$. Hence we see that periodic solutions are of the form $x_1^2 + x_2^2 = c$.

Example 25.0.12 *Consider the simple pendulum given by*

$$\begin{aligned}\dot{x}_1 &= x_2 \\ \dot{x}_2 &= -\frac{g}{l} \sin x_1\end{aligned}$$

Continuing with the same $V(x)$ as in the previous example, we have

$$\dot{V} = \frac{\partial h_1}{\partial x_1}x_2 + \frac{\partial h_2}{\partial x_2}\left(-\frac{g}{l} \sin x_1\right) = 0$$

which implies $\frac{\partial h_1}{\partial x_1} = c\frac{g}{l} \sin x_1$ and $\frac{\partial h_2}{\partial x_2} = cx_2$. On integrating, we obtain $h_1 = -c\frac{g}{l} \cos x_1$, $h_2 = cx_2^2$. Thus, $V = -\frac{g}{l} \cos x_1 + x_2^2 = a$ constant. The expression for V correspond to closed-orbits of constant energy.

We discuss conditions under which the existence of periodic solutions in a plane are ruled out.

Non-existence of periodic solutions

We need the following definitions before stating the main result.

Definition 26.0.13 *A curve $\gamma : [a, b] \rightarrow \mathbb{R}$ is said to be closed if $\gamma(a) = \gamma(b)$. Further the curve γ is said to be a simple closed curve if γ is one-to-one on the interval $[a, b)$.*

The region in a plane about which the non-existence of periodic solutions need to be verified should possess the following property.

Definition 26.0.14 *A region D is said to be simply connected if for every simple closed curve γ in D , the inner region of γ is also a subset of D .*

An annular region is not simply connected, while the inner region of a circle is simply connected. We are now in a position to interpret the following criteria.

Lemma 26.0.15 Bendixson's criteria *If on a simply connected region D of $\mathbb{R}^2$, $\operatorname{div} f \triangleq \frac{\partial f_1}{\partial x_1} + \frac{\partial f_2}{\partial x_2}$ is not identically equal to zero and does not change sign then the system (23.1) has no periodic solutions lying entirely within D .*

Proof:

On any orbit of (23.1), the following holds

$$\frac{dx_2}{dx_1} = \frac{f_2}{f_1}$$

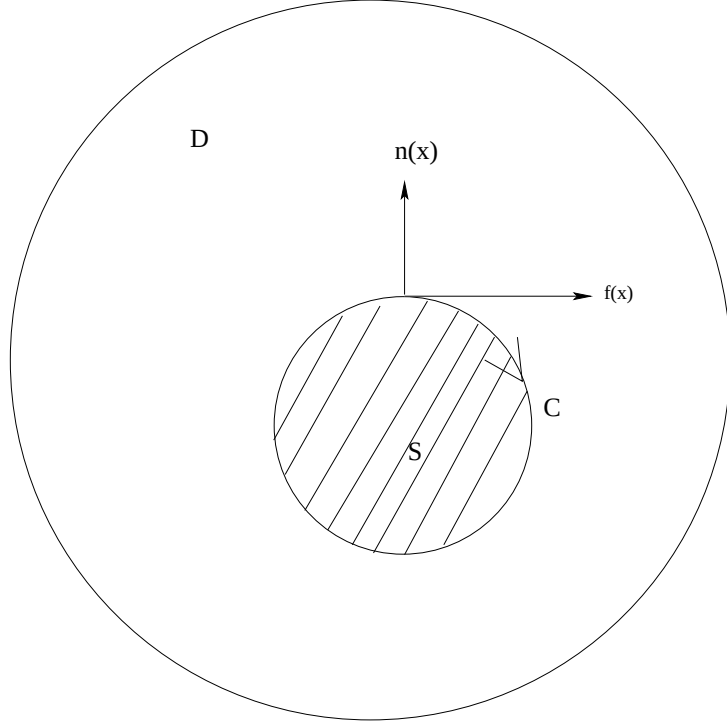


Figure 26.1: various sets in the proof of Bendixson's theorem

Therefore on any closed orbit C (see Figure 26.1) of (23.1),

$$\oint_C \langle f(x), n(x) \rangle dC = \oint_C f_2 dx_1 - f_1 dx_2 = 0$$

where, $n(x)$ is the outward normal to $f(x)$ at $x \in C$. Applying *Green's theorem*¹,

$$\oint_C f_2 dx_1 - f_1 dx_2 = 0 = \iint_S \left(\frac{\partial f_1}{\partial x_1} + \frac{\partial f_2}{\partial x_2} \right) dx_1 dx_2 = 0$$

where S is the region enclosed by C . From the last identity, we can conclude that $\text{div}(f) = 0$ or changes sign, the negation of which is the hypothesis of the criteria. $\square$

Example 26.0.16 Consider the system given by

$$\begin{aligned} \dot{x}_1 &= x_2 + x_1 x_2^2 \\ \dot{x}_2 &= -x_1 + x_1^2 x_2 \end{aligned} \tag{26.1}$$

We have $\text{div} f = \frac{\partial f_1}{\partial x_1} + \frac{\partial f_2}{\partial x_2} = x_2^2 + x_1^2 > 0$, $\forall x \neq 0$. By Bendixson's criteria the system has no periodic solutions in the entire plane.

¹ $\oint_C M(x_1, x_2) dx_1 + N(x_1, x_2) dx_2 = 0 = \iint_D \left(\frac{\partial N}{\partial x_1} - \frac{\partial M}{\partial x_2} \right) dA$, where D is the region enclosed by the closed curve C .

Unit 16

Introduction

For bifurcations and maps we are following early and late chapters in the lovely book by Steven Strogatz (1994) called ‘Non-linear Dynamics and Chaos’ and the extremely clear book by Richard A. Holmgren (1994) called ‘A First Course in Discrete Dynamical Systems’.

1 Bifurcations of one-dimensional dynamical systems

We have seen in some previous lectures that there are nice relations between dynamical systems (we explored Hamiltonian systems) and mappings (Poincaré maps). We also saw an example of a Hamiltonian system with number of fixed points that are dependent on a free parameter and so there was a *bifurcation* in the number of fixed points. Here we look at one dimensional dynamical systems on the real line and maps from the real line back to itself.

1.1 Saddle-node bifurcation

Let us look at dynamical systems that are derived from maps of the real line back onto the real line

$$\dot{x} = f(x)$$

For example

$$\dot{x} = r + x^2 \tag{1}$$

- When $f(x) > 0$, then x will increase in time.
- When $f(x) < 0$, then x decreases in time.
- When $f(x) = 0$ then we have a fixed point.

It is straightforward to classify fixed points from a plot of the function $y = f(x)$. Fixed points are where the function crosses the x axis. On the plot x increases in time (goes to the right) where the function is above the x -axis and decreases in time (goes to the left) where the function is below the x -axis. It is also easy to see whether a fixed point is stable by looking at the direction of motion on either side of the fixed point.

For the dynamical system given by equation 1 (see Figure 1)

- When $r < 0$ there are two fixed points, one stable and one unstable. As r approaches zero, the two fixed points move together.

- When $r = 0$, we can say the fixed point is half stable because even a small change in r will make it disappear.
- For $r > 0$, there are no fixed points.

We can also look at the stability of a fixed point by expanding about the fixed point. We take x^* a fixed point and expand about it with $x = x^* + y$.

$$f(x^*) = 0 + f'(x^*)y + \dots$$

The dynamical system near the fixed point

$$\dot{y} = f'(x^*)y$$

This has solution

$$y \propto \exp(f'(x^*)t)$$

so if $f'(x^*) > 0$ then the fixed point is unstable. If $f'(x^*) < 0$ then the fixed point is stable. If $f'(x^*) = 0$ then we need to expand to higher order to see how the dynamics behaves near the fixed point. If $f''(x^*) \neq 0$ then we know that $f(x^*)$ is a local maxima or minima. We can call this point either half unstable or half stable if a small change in a parameter like r changes its nature.

The dynamical system of equation 1 displays (as a function of r) a *bifurcation*. Going from positive r to negative r we say that a *saddle-node bifurcation* occurs at $r = 0$ with two fixed points appearing. The situation is illustrated in Figure 1.

The system

$$\dot{x} = r - x^2$$

also displays a saddle-node bifurcation but the bifurcation diagram is flipped both vertically and horizontally.

In a sense

$$\dot{x} = r + x^2 \quad \dot{x} = r - x^2$$

are representative of *all* saddle-node bifurcations. Near a fixed point we can locally expand in a Taylor series. If the Taylor series contains a constant and a quadratic term, then by rescaling variables, we can transform the equations of motion near the fixed point to look like one of these two equations. These two models are sometimes call *normal forms* for the saddle-node bifurcation.

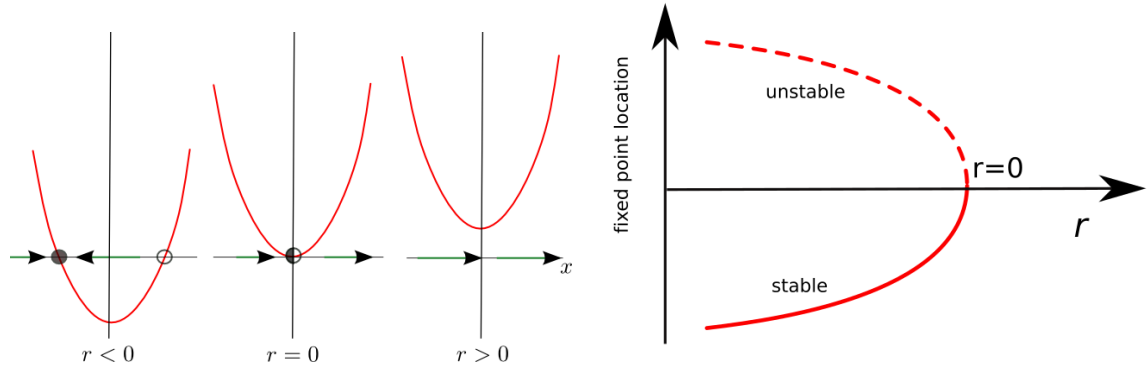


Figure 1: a) Fixed points for the dynamical system $\dot{x} = r + x^2$. When $r < 0$ the leftmost fixed point is stable and an attractor and the rightmost fixed point is unstable. When $r > 0$ there are no fixed points. The value of r determines whether there are stable or unstable fixed points. b) Bifurcation diagram for the same system. This is known as a *saddle-node* bifurcation.

1.1.1 Example of a saddle-node bifurcation

Let us look at the system

$$\dot{x} = r - x - e^{-x}$$

The fixed points are given by $r - x - e^{-x} = 0$ but we can't solve this equation. We can think about it graphically by considering where a line with slope $-x$ crosses the function e^{-x} (see Figure 2). We expect that we will have a bifurcation when the line $y = r - x$ is tangent to the line e^{-x} . When the line $y = r - x$ intersects e^{-x} at two points then there are two fixed points. When the line $y = r - x$ does not intersect e^{-x} at all then there are no fixed points. When the line $y = r - x$ intersects e^{-x} at a single point, then the line is tangent to the curve and we have a bifurcation point. Let us label r_c the value of r at the bifurcation point. For $r > r_c$, there are two fixed points and for $r < r_c$ there are none.

Let us now solve for r_c . The bifurcation must take place when the two curves are just touching. At this point the tangent to the curve $y = e^{-x}$ is equal to that of the line $y = r - x$. The tangent to $y = e^{-x}$ is $-e^{-x}$. We set this equal to -1 and solve for x , finding that the bifurcation happens where $e^{-x} = 1$ or at $x_c = 0$. What r value is required for this to be a fixed point? $r - 0 - 1 = 0$, so $r_c = 1$. The bifurcation happens at $r_c = 1, x_c = 0$.

Let us expand about $x = 0$. We need to expand the exponential $e^{-x} \sim 1 - x + x^2/2..$ and our function becomes

$$\dot{x} = r - x - 1 + x - \frac{x^2}{2}.. = (r - 1) - \frac{x^2}{2}..$$

which shows in another way that there is a saddle-node bifurcation at $r = 1$; see Figure 3.

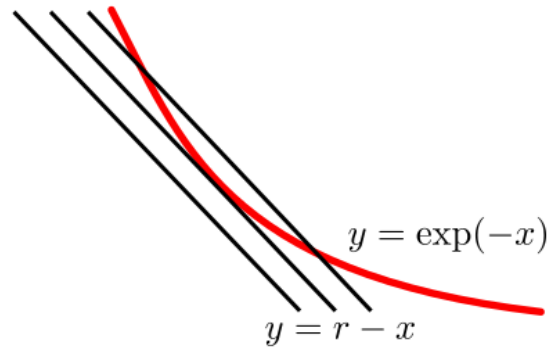


Figure 2: The system $\dot{x} = r - x - e^{-x}$ has a bifurcation where the curve $y = r - x$ is tangent to the curve $y = e^{-x}$.

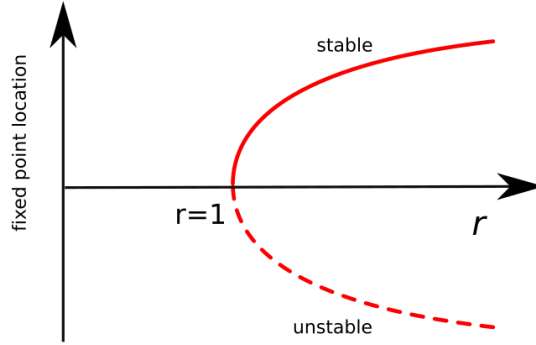


Figure 3: The system $\dot{x} = r - x - e^{-x}$ has a saddle node bifurcation with two fixed points at $r > r_c$ with $r_c = 1$. This bifurcation diagram has the same shape as that for $\dot{x} = r - x^2$, except shifted in r by 1.

1.1.2 Another example

Consider the system

$$\dot{x} = rx - \sin x$$

with $rx - \sin x$ plotted in Figure 4. The system has a number of bifurcations but we will focus on the one that occurs with $r \sim 0.13$ at $x \sim 7.7$. We can see from the red line in the figure that when the function $rx - \sin(x)$ is above zero (near $x = 7.7$) there is no nearby fixed point. For larger r there are two fixed points and the leftmost one is stable and the rightmost one unstable. So it's a saddle-node bifurcation.

Let us define a function

$$f(x, r) = rx - \sin x$$

and consider where the new fixed point first appears. For a particular $r = r_c$ value the curve just touches the x axis, producing a fixed point at x^* with

$$f(x^*, r_c) = 0 \quad \frac{\partial f}{\partial x}(x^*, r_c) = 0$$

Let us expand to first order in r about r_c and to second order in x about x^* .

$$f(x, r) = f(x^*, r_c) + \frac{\partial f}{\partial r}(x^*, r_c)(r - r_c) + \frac{\partial f}{\partial x}(x^*, r_c)(x - x^*) + \frac{\partial^2 f}{\partial x^2}(x^*, r_c) \frac{(x - x^*)^2}{2}$$

The first term is zero (x^* is a fixed point for $r = r_c$). The third term is zero because the curve is tangent to the x axis at r_c, x^* . We evaluate the other two terms finding

$$f(x, r) = x^*(r - r_c) + \sin x^* \frac{(x - x^*)^2}{2}$$

If we define new variables $X = x - x^*$ and $R = x^*(r - r_c)$ we can rewrite this as

$$f(X, R) = R + \frac{\sin x^*}{2} X^2$$

and we recognize that this is in the form of a saddle-node bifurcation. A further rescaling of X can put it in the form $R + X^2$.

As any curve with a minimum dips below the x axis, locally it can be approximated by a quadratic function. We can shift r and x and rescale both. So any local minimum or maximum as it crosses the x axis can be made to look like a saddle-node bifurcation.

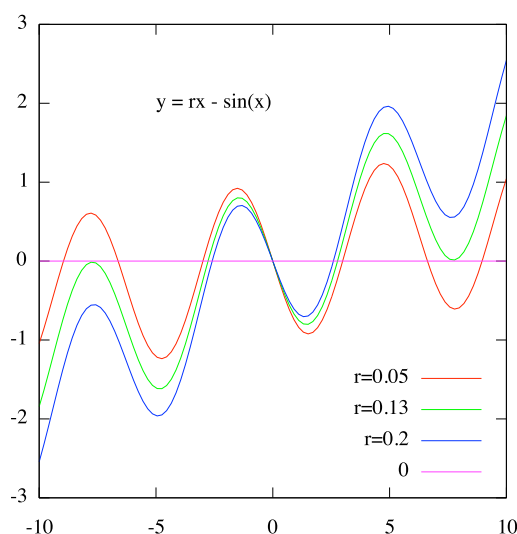


Figure 4: Saddle node bifurcation of $\dot{x} = rx - \sin x$. Plotted is $y = rx - \sin x$. There are saddle-node bifurcations at $r \sim 0.13$ and $x \sim \pm 7.5$. These can be seen with the green line touching the x axis at $x \sim \pm 7.5$.

1.2 Pitchfork bifurcations

Consider

$$\dot{x} = rx - x^3 \tag{2}$$

Depending on r there are either 3 roots or 1 root, so there are either 3 or 1 fixed points. The middle point is unstable (see figure 5). Solve $rx - x^3 = x(r - x^2) = 0$ finding fixed

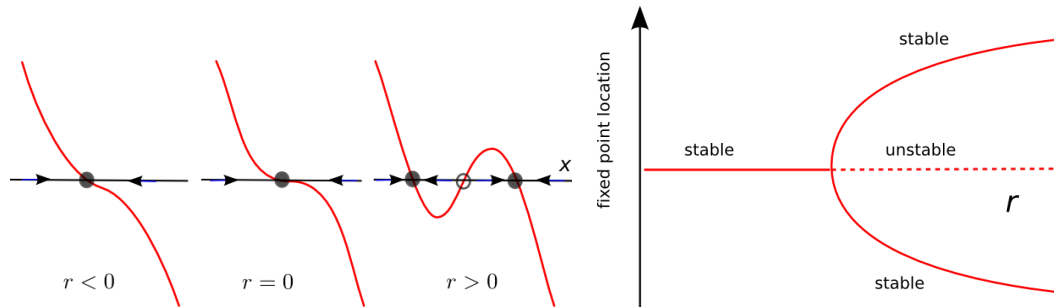


Figure 5: a) Fixed points for the dynamical system $\dot{x} = rx - x^3$. When $r < 0$ there is a single stable fixed point. When $r > 0$ there are three fixed points and the center one is unstable. The value of r determines the number of fixed points. b) Bifurcation diagram for the same system. This is known as a *supercritical pitch-fork bifurcation*.

points at $x^* = 0, \pm\sqrt{r}$.

When the sign of x^3 is flipped then the bifurcation diagram is inverted. For $r < 0$ there are three fixed points with the outer two unstable. For $r > 0$ there is a single unstable fixed point. The bifurcation is called a *subcritical pitchfork* bifurcation.

The system is vastly unstable, but can be made stable at large x with the addition of higher order terms (in x), and giving large amplitude stable branches and some hysteresis in its behavior as r is varied (see Figure 6b).

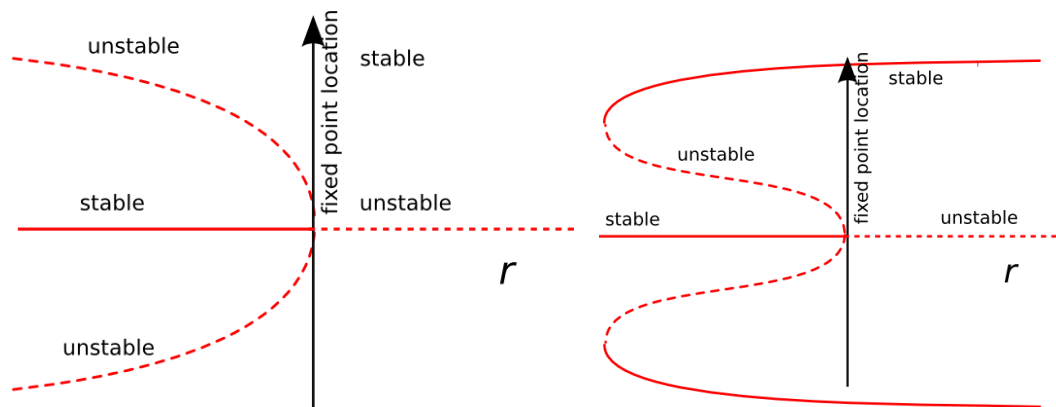


Figure 6: a) Bifurcation diagram for the dynamical system $\dot{x} = rx + x^3$. This is known as a *sub-critical pitch-fork bifurcation*. b) Bifurcation diagram for the dynamical system $\dot{x} = rx + x^3 - x^5$ that has stable fixed points at large x . This system exhibits hysteresis. Consider a system starting on the top right at high r . As r decreases the system follows the top stable curve and only jumps to the x -axis at large negative r . A system that starts on the left must reside on the stable curve on the x -axis. The system only jumps to high or low x when $r \sim 0$. The two trajectories are not the same. The system with decreasing r must reach a much lower r value before it makes the transition than the system with increasing r .

1.3 Trans-critical bifurcations

We consider the system

$$\dot{x} = rx - x^2$$

There is a fixed point at the origin in all cases.

- If $r < 0$ there is an unstable fixed point at $x = -r$ and the one at the origin is stable.
- If $r > 0$ there is an unstable fixed point at the origin and an unstable one at $x = r$.
- If $r = 0$ then there is a mixed type of fixed point at the origin. We can say that the fixed points swap locations at the origin.

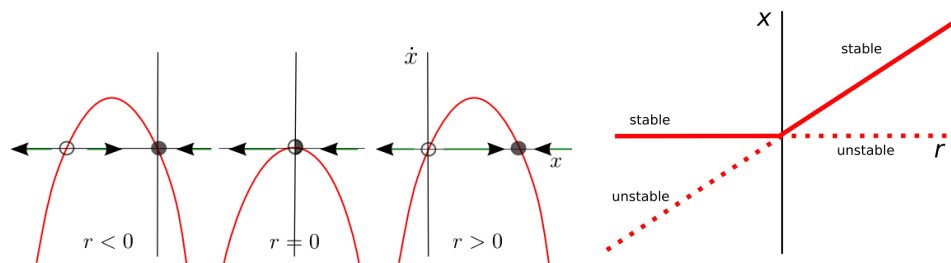


Figure 7: a) Fixed points for the dynamical system $\dot{x} = rx - x^2$. When $r < 0$ there are two fixed points. The leftmost is unstable and the rightmost is stable. When $r > 0$ there are also two fixed points but they have swapped locations. The bifurcation is known as a trans-critical bifurcation.

**POST GRADUATE DEGREE PROGRAMME (CBCS)
IN
MATHEMATICS**

**SEMESTER II
SELF LEARNING MATERIAL**

PAPER - MATC 2.2

**Real Analysis - II
Complex Analysis - II
Functional Analysis - II**



**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, 2019

Directorate of Open and Distance Learning, University of Kalyani
Published by the Directorate of Open and Distance Learning,
University of Kalyani, Kalyani-741235, West Bengal

All rights reserved. No part of this work should be reproduced in any form without the permission in writing from the Directorate of Open and Distance Learning, University of Kalyani.

Director's Note

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

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

Self Learning Materials (SLMs) have been published by the Directorate of Open and Distance Learning, University of Kalyani, Kalyani-741235, West Bengal and all the copyright reserved for University of Kalyani. No part of this work should be reproduced in any form without permission in writing from the appropriate authority of the University of Kalyani.

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

Director
Directorate of Open and Distance Learning
University of Kalyani

Contents

Sl. No.	Blocks	Units	Page
1	Real Analysis - II	I	2-13
		II	14-20
		III	21-34
		IV	35-40
		V	41-44
2	Complex Analysis - II	VII	47-56
		VIII	57-58
		IX	59-64
		X	65-72
3	Functional Analysis – II	XI	75-80
		XII	81-86
		XIII	87-96
		XIV	97-104
		XV	105-110
		XVI	111-117

Core Paper

MATC 2.2

Block - I

Marks : 32 (SEE : 25; IA : 07)

Real Analysis - II

Syllabus

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

Unit I

Course Structure

- Introduction
 - Objectives
 - Definition of the Lebesgue outer measure on the power set of $\mathbb{R}$
 - Countable subadditivity, Carathodory's definition of the Lebesgue measure and basic properties
 - Measurability of an interval (finite or infinite), Countable additivity
-

Introduction

A measure on a set is a systematic way to assign a number to each suitable subset of that set, intuitively interpreted as its size. In this sense, a measure is a generalization of the concepts of length, area, and volume. A measure is a function that assigns a non-negative real number or $+\infty$ to (certain) subsets of a set X . We will formally define measure in the following sections.

Objectives

After reading this unit, you will be able to:

- define the extended real line
- define measure of a set
- determine the measurability of a set
- find the measure of a set

1 Measure Theory

We will define a *set function*¹ on some collection of sets and taking values in the non-negative extended real line. But before that, we need to define the extended real numbers.

¹a function whose domain set contains sets

1.1 Extended Real Line

When we add $\{-\infty, +\infty\}$ to $\mathbb{R}$ we get what is called the extended real line. The operations on $\mathbb{R}$ can be extended to the extended real numbers as follows:

- $\infty + \infty = \infty$
- $\infty \cdot \infty = +\infty$
- $-\infty - \infty = -\infty$
- $\infty \cdot (-\infty) = (-\infty) \cdot \infty = -\infty$
- $a + \infty = +\infty + a = +\infty, a \neq -\infty$
- $a - \infty = -\infty + a = -\infty, a \neq +\infty$
- $a \cdot (\pm\infty) = \pm\infty \cdot a = \pm\infty, a \in (0, +\infty]$
- $a \cdot (\pm\infty) = \pm\infty \cdot a = \mp\infty, a \in [-\infty, 0)$
- $\frac{a}{\pm\infty} = 0, a \in \mathbb{R}$
- $\frac{\pm\infty}{a} = \pm\infty, a \in (0, +\infty)$
- $\frac{\pm\infty}{a} = \mp\infty, a \in (-\infty, 0)$

The operations such as $0 \cdot (\pm\infty)$, $\frac{\pm\infty}{\pm\infty}$, $\infty - \infty$ are undefined (called indeterminate forms). We will now define the outer measure of a set.

1.2 Outer measure

Let $I = [a, b]$, $b > a$, be an interval. Then the length of I is defined as $l(I) = b - a$. Now, consider the power set of real numbers, $\mathcal{P}(\mathbb{R})$. Then the Lebesgue outer measure, or outer measure is a function m^* from $\mathcal{P}(\mathbb{R})$ to $[0, +\infty]$ defined as

$$m^*(E) = \inf \left\{ \sum_{k=1}^{\infty} l(I_k) : \{I_k\} \text{ is the countable collection of closed open intervals such that } E \subseteq \bigcup_{k=1}^{\infty} I_k \right\}$$

Clearly, $m^*(E) \geq 0$ for all $E \subset \mathbb{R}$. Also, the definition will not be affected if any other interval is taken in place of closed open intervals. We also have the following theorem:

Theorem 1.1. 1. $m^*(\emptyset) = 0$
 2. $m^*(A) \leq m^*(B)$, if $A \subseteq B$

Proof. Left as an exercise.

□

Theorem 1.2. Outer measure is translation invariant, that is, for any set E , $m^*(E) = m^*(E + x)$, where, $E + x = \{y + x : y \in E\}$, $x \in \mathbb{R}$.

Proof. Let $\epsilon > 0$. Then, there exists a countable collection $\{I_n\}$ of intervals such that $E \subseteq \cup_n I_n$ and

$$m^*(E) \leq \sum_n l(I_n) \leq m^*(E) + \epsilon \quad (1.1)$$

Clearly, $E + x \subseteq \cup(I_n + x)$. So, for each ϵ , $m^*(E + x) \leq \sum_n l(I_n + x) = \sum_n l(I_n) \leq m^*(E) + \epsilon$, which implies,

$$m^*(E + x) \leq m^*(E) \quad (1.2)$$

Again, $E = (E + x) - x$. So, by (1.2), we get

$$m^*(E) \leq m^*(E + x) \quad (1.3)$$

By (1.2) and (1.3), we get

$$m^*(E) = m^*(E + x)$$

□

Theorem 1.3. The outer measure of an interval is equal to its length.

Proof. Case I: Suppose that I is a closed interval, $[a, b]$, say. Then, for each $\epsilon > 0$, we have

$$m^*([a, b]) \leq m^*([a, b + \epsilon]) \leq b - a + \epsilon$$

so,

$$m^*(I) \leq b - a \quad (1.4)$$

Again, for each $\epsilon > 0$, I may be covered by a collection of intervals $\{I_n\}$, such that

$$m^*(I) \geq \sum_{n=1}^{\infty} l(I_n) - \epsilon \quad (1.5)$$

where $I_n = [a_n, b_n)$, say. For each n , let $I' = (a_n - \epsilon/2^n, b_n)$. Then $\cup_{n=1}^{\infty} I'_n \supseteq I$, so by Heine-Borel theorem, a finite subcollection of I'_n , say $J_1, J_2, \dots, J_N$, where $J_k = (c_k, d_k)$, covers I . Then we may suppose that no J_k is contained in any other, we have, supposing that $c_1 < c_2 < \dots < c_N$,

$$d_N - c_1 = \sum_{k=1}^N (d_k - c_k) - \sum_{k=1}^{N-1} (d_k - c_{k+1}) < \sum_{k=1}^N l(J_k)$$

So we have

$$\begin{aligned} m^*(I) &\geq \sum_{n=1}^{\infty} l(I_n) - \epsilon \\ &\geq \sum_{n=1}^{\infty} l(I'_n) - 2\epsilon \\ &\geq \sum_{k=1}^N l(J_k) - 2\epsilon \\ &\geq d_N - c_1 - 2\epsilon \\ &> b - a - 2\epsilon \\ &> l(I) - 2\epsilon \end{aligned}$$

So we get

$$m^*(I) \geq b - a \quad (1.6)$$

By (1.4) and (1.6), we have

$$m^*(I) = b - a = l(I)$$

Case II: Suppose that $I = (a, b]$ and $a > -\infty$. If $a = b$, then its a singleton set and the next theorem gives the result. If $a < b$, suppose that $0 < \epsilon < b - a$ and write $I' = [a + \epsilon, b]$. Then

$$m^*(I) \geq m^*(I') = l(I) - \epsilon \quad (1.7)$$

But, $I \subseteq I'' = [a, b + \epsilon)$, so

$$m^*(I) \leq l(I'') = l(I) + \epsilon \quad (1.8)$$

Since (1.7) and (1.8) are true for all small ϵ , so $m^*(I) = l(I)$. We can consider similarly, the cases $I = (a, b)$ and $I = [a, b)$.

Case III: Let I be the infinite interval $I = (-\infty, a]$. Then, for any $M > 0$, there exists k such that the finite interval I_M , where $I_M = [k, k + M)$, is contained in I . So,

$$m^*(I) > M$$

and hence $m^*(I) = \infty = l(I)$. The proof of the other three cases is similar. $\square$

Obviously, the outer measure of any single point a is zero. More generally, we have the following theorem:

Theorem 1.4. If $C = \{c_1, c_2, \dots\}$ is a countable subset of I , the $m^*(C) = 0$.

Proof. Consider for each c_i , an open interval J_i of length $3^{-i}\epsilon$, centred at c_i , for any arbitrary $\epsilon > 0$. Thus,

$$\begin{aligned} m^*(C) &\leq \sum_{i=1}^{\infty} J_i \\ &= \sum_{i=1}^{\infty} 3^{-i}\epsilon \\ &= \epsilon \sum_{i=1}^{\infty} 3^{-i} \\ &= \epsilon \cdot \frac{1}{2} \\ &< \epsilon \end{aligned}$$

So, $0 \leq m^*(C) < \epsilon$, for every $\epsilon > 0$. So, $m^*(C) = 0$. Hence, the outer measure of any countable set is zero. $\square$

Corollary 1.4.1. 1. Outer measure of any finite set is zero.

2. The outer measures of the sets $\mathbb{N}, \mathbb{Q}, \mathbb{Z}$ are all zero.

The next theorem asserts that the outer measure is countably additive. We have the following theorem:

Theorem 1.5. For any sequence of sets $\{E_i\}$, we have

$$m^*(\cup_{i=1}^{\infty} E_i) \leq \sum_{i=1}^{\infty} m^*(E_i).$$

Proof. For each i , and any $\epsilon > 0$, there exists a sequence of intervals $\{I_{i,j}\}_j$ such that $E_i \subseteq \cup_{j=1}^{\infty} I_{i,j}$ and

$$m^*(E_i) \leq m^*(\cup_{j=1}^{\infty} I_{i,j}) \leq \sum_{j=1}^{\infty} l(I_{i,j}) \leq m^*(E_i) + \epsilon/2^i$$

That is,

$$m^*(E_i) \geq \sum_{j=1}^{\infty} l(I_{i,j}) - \epsilon/2^i$$

Also, $\cup_{i=1}^{\infty} E_i \subseteq \cup_{i=1}^{\infty} \cup_{j=1}^{\infty} I_{i,j}$. So,

$$m^*(\cup_{i=1}^{\infty} E_i) \leq \sum_{i,j=1}^{\infty} l(I_{i,j}) \leq \sum_{i=1}^{\infty} m^*(E_i) + \epsilon$$

Since $\epsilon > 0$ is arbitrary, so

$$m^*(\cup_{i=1}^{\infty} E_i) \leq \sum_{i=1}^{\infty} m^*(E_i)$$

□

Exercise 1.1. 1. If A and B are two disjoint sets, then show that

$$m^*(A \cup B) = m^*(A) + m^*(B)$$

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

1.3 Measurability of a set

Before going directly into the definition of measure of a set, we need to know about the concept of algebra and σ -algebra on a set.

Definition 1.6. Let X be any set and let Σ be a non-empty collection of subsets of X . Then Σ is called a σ -algebra if the following conditions are satisfied:

1. $X, \emptyset$ are in Σ
2. If A is in Σ , then $X \setminus A$ is in Σ
3. If $A, B \in \Sigma$, then $A \cup B$ is in Σ .

Definition 1.7. Let X be a set and Σ be a non-empty collection of subsets of X . Then Σ is called a σ -algebra if the following conditions are satisfied:

1. $X, \emptyset$ are in Σ
2. If A is in Σ , then $X \setminus A$ is in Σ
3. If A_n is a sequence of elements in Σ , then $\cup_n A_n$ is in Σ .

Example 1.8. 1. The power set of X always forms an algebra as well as σ -algebra on X .

2. The collection $\{X, \emptyset\}$ forms an algebra as well as a σ -algebra on X .

3. Let $X = \{1, 2, 3, 4\}$. Then the collection $\{\{1, 2, 3, 4\}, \emptyset, \{1\}, \{2, 3, 4\}\}$ is an algebra as well as a σ -algebra on X .

We will now define measurability of a set due to Caratheodory as follows:

Definition 1.9. A set E is Lebesgue measurable, or simply, measurable, if for each subset A of $\mathbb{R}$, we have

$$m^*(A) = m^*(A \cap E) + m^*(A \cap (\mathbb{R} \setminus E)) \quad (1.9)$$

Notice that, for all $A \subset \mathbb{R}$, we have

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

So by the sub-additive property of m^* , we will have

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

So, the equation (1.9) gets reduced to

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

That is, a set E will be called measurable if the above equation is satisfied for every subset A of $\mathbb{R}$. For such a set, the measure is defined as $m(E) = m^*(E)$.

Theorem 1.10. The set for which the outer measure is zero, is measurable.

Proof. Let E be a subset of $\mathbb{R}$ such that $m^*(E) = 0$. Then we need to show that

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

Now, since $A \cap E \subseteq E$, so

$$m^*(A \cap E) \leq m^*(E) = 0$$

Since the outer measure is a non-negative function, the above equation implies that

$$m^*(A \cap E) = 0$$

Also since $A \cap (\mathbb{R} \setminus E) \subseteq A$, so by the monotonicity of the outer measure function, we get

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

Which implies that,

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

Hence the result. □

Corollary 1.10.1. 1. All countable sets are measurable with measure zero. Can we get an uncountable set of measure zero?

2. The empty set $\emptyset$ is measurable and $m(\emptyset) = 0$.

3. Any subset of a set of outer measure (or only measure) zero is measurable.

Theorem 1.11. If E is measurable, then $\mathbb{R} \setminus E$ is measurable.

Proof. Let E be measurable. Then, for any arbitrary set A ,

$$\begin{aligned} m^*(A) &\geq m^*(A \cap E) + m^*(A \cap (\mathbb{R} \setminus E)) \\ &= m^*(A \cap (\mathbb{R} \setminus E)) + m^*(A \cap (\mathbb{R} \setminus (\mathbb{R} \setminus E))) \quad [\text{since for any set } A, \mathbb{R} \setminus (\mathbb{R} \setminus A) = A] \end{aligned}$$

So, by the definition of measurability, $\mathbb{R} \setminus E$ is measurable. $\square$

Theorem 1.12. Let A and B be two measurable sets. Then $A \cup B$ and $A \cap B$ are measurable.

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

$$\begin{aligned} m^*(E) &\geq m^*(E \cap A) + m^*(E \cap (\mathbb{R} \setminus A)) \\ &= m^*(E \cap A) + m^*(E \cap (\mathbb{R} \setminus A) \cap B) + m^*(E \cap (\mathbb{R} \setminus A) \cap (\mathbb{R} \setminus B)) \end{aligned} \quad (1.10)$$

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

$$m^*(E \cap (\mathbb{R} \setminus A) \cap (\mathbb{R} \setminus B)) \leq m^*((\mathbb{R} \setminus A) \cap (\mathbb{R} \setminus B)) = m^*(\mathbb{R} \setminus (A \cup B)) \quad (1.11)$$

Also,

$$m^*(E \cap (\mathbb{R} \setminus A) \cap B) \leq m^*(E \cap (A \cup B)) \quad (1.12)$$

Using (1.11) and (1.12) in (1.10), we get,

$$m^*(E) \geq m^*(E \cap (A \cup B)) + m^*(\mathbb{R} \setminus (A \cup B))$$

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

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

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

Corollary 1.12.1. If A and B are measurable, then $A \setminus B$ is measurable.

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

Theorem 1.13. If E is measurable, then $E + x$ is measurable for any real x .

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

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

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

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

which is the definition of measurability. Hence the result. $\square$

Theorem 1.14. Every interval is measurable.

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

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

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

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

For each k , let

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

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

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

Since ϵ is arbitrary, so we get

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

Hence the result. $\square$

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

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

We immediately conclude that $\mathbb{R}$ is measurable, since $\mathbb{R} = (-\infty, \infty)$ and for any $a \in \mathbb{R}$, we have

$$(-\infty, \infty) = (-\infty, a) \cup (a, \infty)$$

Since the union of two measurable sets is measurable, so $\mathbb{R}$ is measurable. We could also prove the measurability of $\mathbb{R}$ by considering it as the compliment of $\emptyset$, since it is measurable.

Let $\mathcal{M}$ denote the set of all measurable sets in $\mathbb{R}$. Then, we can summarize a few properties of it as follows:

1. $\mathbb{R}$, and $\emptyset$ are in $\mathcal{M}$
2. If $A \in \mathcal{M}$, then $\mathbb{R} \setminus A \in \mathcal{M}$
3. If $A, B \in \mathcal{M}$, then $A \cup B \in \mathcal{M}$.

Thus, we can say that $\mathcal{M}$ forms an algebra on $\mathbb{R}$. In fact, we will soon be able to say that it also forms a σ -algebra on $\mathbb{R}$. For this, we need the following theorem:

Theorem 1.15. If $\{E_i\}$ is any sequence of measurable sets, then $\cup E_i$ and $\cap E_i$ are measurable.

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

Lemma 1.16. Let $\{E_i\}_{i=1}^n$ be a finite collection of disjoint measurable sets. If $A \subseteq \mathbb{R}$, then

$$m^*(\cup_i (A \cap E_i)) = m^*(A \cap (\cup_i E_i)) = \sum_i m^*(A \cap E_i)$$

In particular, if $A = \mathbb{R}$, then $m(\cup_i E_i) = \sum_i m(E_i)$.

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

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

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

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

Hence the result. □

We will now prove the original theorem for countable sets.

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

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

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

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

since $\mathbb{R} \setminus E = \mathbb{R} \setminus (\cup_{i=1}^n H_i)$. Letting $n \rightarrow \infty$, we have

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

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

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

Hence E is measurable. Again, since $\cap_i E_i = \mathbb{R} \setminus (\cup_i (\mathbb{R} \setminus E_i))$, so $\cap_i E_i$ is also measurable. Hence the proof is complete. $\square$

In view of the above theorem, we can say that every open set is measurable since every open set can be written as the countable union of disjoint open intervals. Since the complement of every measurable set is measurable, so every closed set is measurable. We will further discuss about these in the following unit.

Now, we summarize the results as follows:

1. $\mathbb{R}$, and $\emptyset$ are in $\mathcal{M}$
2. If $A \in \mathcal{M}$, then $\mathbb{R} \setminus A \in \mathcal{M}$
3. If $\{E_i\}$ be a sequence of sets in $\mathcal{M}$, then $\cup_i E_i \in \mathcal{M}$.

Hence, $\mathcal{M}$ forms a σ -algebra on $\mathbb{R}$.

Theorem 1.17. If A and B are measurable, such that $A \subseteq B$, then $m(A) \leq m(B)$. Further, if $m(A) < \infty$, then $m(B \setminus A) = m(B) - m(A)$.

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

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

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

$$\begin{aligned} m(B) &= m(B \setminus A) + m(A) \\ &\geq m(A) \end{aligned} \tag{1.13}$$

Hence,

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

Also, by (1.13), we get, since $m(A) < \infty$, so

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

$\square$

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

We will conclude this unit giving two definitions of limit supremum and infimum of sets as follows:

Definition 1.18. Let $\{A_n\}$ be a sequence of sets. Using union and intersection, we define the limit superior and limit inferior as follows:

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

and

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

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

Theorem 1.19. Let $\{A_i\}$ be a sequence of measurable sets. Then

1. if $A_1 \subseteq A_2 \subseteq \dots$, we have

$$m(\lim A_i) = \lim m(A_i)$$

2. if $A_1 \supseteq A_2 \supseteq \dots$, and $m(A_i) < \infty$ for each i , then

$$m(\lim A_i) = \lim m(A_i)$$

Proof. 1. We write

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

Then clearly,

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

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

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

So,

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

which was required to be proved.

2. We have, $A_1 \setminus A_1 \subseteq A_1 \setminus A_2 \subseteq A_1 \setminus A_3 \subseteq \dots$. So, by the previous subpart, we have

$$\begin{aligned}
m(\lim(A_1 \setminus A_i)) &= \lim(m(A_1 \setminus A_i)) \\
&= m(A_1) - \lim m(A_i)
\end{aligned} \tag{1.14}$$

But,

$$\begin{aligned}
\lim(A_1 \setminus A_i) &= \bigcup_{i=1}^{\infty} (A_1 \setminus A_i) \\
&= A_1 \setminus \bigcap_{i=1}^{\infty} A_i \\
&= A_1 \setminus \lim A_i
\end{aligned}$$

So, taking the measures on both sides, the result follows from (1.14) and since $m(A_1) < \infty$. $\square$

Exercise 1.2. 1. Show that Cantor set has measure zero.

2. For $k > 0$ and $A \subseteq \mathbb{R}$, let $kA = \{x : k^{-1}x \in A\}$. Show that

- i. $m^*(kA) = km^*(A)$,
- ii. A is measurable iff kA is measurable.

3. For $A \subseteq \mathbb{R}$, let $-A = \{x : -x \in A\}$. Show that

- i. $m^*(A) = m^*(-A)$,
- ii. A is measurable iff $-A$ is measurable.

Unit II

Course Structure

- Introduction
 - Objectives
 - Characterizations of measurable sets by open sets, G sets
 - Characterizations of measurable sets by closed sets and F sets
 - Measurability of Borel sets, Existence of non-measurable sets.
-

Introduction

Open sets and closed sets are the sets we are most acquainted to work with. So, if we can understand measurability in terms of open and closed sets, we can get more insight into the structure of a measurable set. In this unit, that is what we are going to do. Approximating a measurable set by open set, closed set, G_δ set, F_σ sets and compact sets. Another important aspect of the study of measurable sets is the existence of non-measurable sets. The question that we come across while studying measurable sets is that, whether all sets are measurable? Most of the sets that we are acquainted with are measurable. So we might think that all sets are measurable. But that is not so. And in this unit, we will turn our attention towards the non-measurable sets.

Objectives

After reading this unit, you will be able to:

- learn the approximation of measurable set by open sets
- learn the approximation of measurable set by closed sets
- learn the approximation of measurable set by G_δ sets
- learn the approximation of measurable set by F_σ sets
- learn what are Borel sets
- learn about the measurability of the Borel sets
- learn about the existence of non-measurable sets

2 Characterization by open and closed sets

We begin this section by the following theorem:

Theorem 2.1. For every $\epsilon > 0$, there exists an open set $U \subseteq \mathbb{R}$ such that $m(U) \leq \epsilon$ and U contains the set $\mathbb{Q}$ or rational numbers.

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

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

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

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

Hence the result. □

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

Theorem 2.2. If $S \subseteq \mathbb{R}$, then

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

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

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

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

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

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

We will now use open sets to give a nice characterization of measurability.

Theorem 2.3. If $S \subseteq \mathbb{R}$ is Lebesgue measurable if and only if for every $\epsilon > 0$, there exists an open set U containing S such that $m^*(U \setminus S) \leq \epsilon$

Proof. Let S be measurable, and let $\epsilon > 0$. For each $n \in \mathbb{N}$, let $S_n = S \cap [-n, n]$, and let U_n be an open set containing S_n , so that

$$m(U_n) < m(S_n) + \frac{\epsilon}{2^n}$$

Let $U = \bigcup_{n \in \mathbb{N}} U_n$. Then U is an open set containing S and $U \setminus S \subseteq \bigcup_{n \in \mathbb{N}} (U_n - S_n)$. So,

$$m(U \setminus S) \leq \sum_{n \in \mathbb{N}} m(U_n - S_n) \leq \sum_{n \in \mathbb{N}} \frac{\epsilon}{2^n} = \epsilon$$

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

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

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

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

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

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

Theorem 2.4. A set S is measurable if and only if there exists a G_δ set G containing S such that $m^*(G \setminus S) = 0$.

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

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

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

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

Note that the complement of any open set is a closed set. So the characterization of measurability by closed set is closely linked with that by open sets with a slight variation as we shall see.

Theorem 2.5. A set $S \subseteq \mathbb{R}$ is measurable if and only if for every $\epsilon > 0$, there exists a closed set F contained in S such that $m^*(S \setminus F) \leq \epsilon$.

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

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

Theorem 2.6. A set S is measurable if and only if there exists an F_σ set F contained in S such that $m^*(S \setminus F) = 0$.

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

Proof. Left as an exercise. □

From all the above theorems, we conclude the following corollary:

Corollary 2.6.1. Let $S \subseteq \mathbb{R}$. Then S is Lebesgue measurable if and only if for every $\epsilon > 0$ there exists a closed set F and an open set U such that $F \subseteq S \subseteq U$ and $m(U \setminus F) < \epsilon$.

We can also define measure with respect to closed sets.

Definition 2.7. If $S \subseteq \mathbb{R}$, the Lebesgue inner measure of S is defined as

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

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

Theorem 2.8. Let $S \subseteq \mathbb{R}$, and suppose that $m^*(S) < \infty$. Then S is measurable if and only if

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

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

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

Then

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

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

Theorem 2.9. Every compact subset of $\mathbb{R}$ is measurable.

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

We will now approximate measurable sets of finite measure by compact sets.

Theorem 2.10. A set $S \subseteq \mathbb{R}$ is measurable with $m^*(S) < \infty$, if and only if, for every $\epsilon > 0$ there exists a compact set K contained in S and $m^*(S \setminus K) < \epsilon$.

Proof. Let S be measurable and $n \in \mathbb{N}$ and define E_n by

$$E_n = E \cap [-n, n]$$

Since the equation $\{E_n\}$ increases to E , it follows that the sequence $\{m(E_n)\}$ increases to $m(E) < \infty$. □

2.1 Borel Sets

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

Theorem 2.11. Let X be a set, and let $\mathcal{C}$ be any collection of σ -algebras on X . Then the intersection $\bigcap \mathcal{C}$ is also a σ -algebra on X .

Proof. Since $\emptyset \in \mathcal{M}$ for every $\mathcal{M} \in \mathcal{C}$, it follows that $\emptyset \in \bigcap \mathcal{C}$. Next, if $S \in \bigcap \mathcal{C}$ then $S \in \mathcal{M}$ for every $\mathcal{M} \in \mathcal{C}$. Thus, $X \setminus S \in \mathcal{M}$ for every $\mathcal{M} \in \mathcal{C}$, and hence $X \setminus S \in \bigcap \mathcal{C}$. Finally, if $\{S_n\}$ is a sequence in $\bigcap \mathcal{C}$, then each $\mathcal{M} \in \mathcal{C}$ must contain the entire sequence $\{S_n\}$. It follows that $\bigcup_{n \in \mathbb{N}} S_n \in \mathcal{M}$ for each $\mathcal{M} \in \mathcal{C}$, and hence $\bigcup_{n \in \mathbb{N}} S_n \in \bigcap \mathcal{C}$. Hence $\bigcap \mathcal{C}$ forms a σ -algebra on X . $\square$

Definition 2.12. Let X be a set, and let $\mathcal{C}$ be any collection of subsets of X . The σ -algebra generated by $\mathcal{C}$ is the intersection of all σ -algebras on X that contain $\mathcal{C}$.

It is important for this definition that there is always at least one σ -algebra containing $\mathcal{C}$, namely the power set $\mathcal{P}(X)$ of all subsets of X .

Definition 2.13. The Borel algebra $\mathcal{B}$ is the σ -algebra on $\mathbb{R}$ generated by the collection of all open sets. A set $B \subseteq \mathbb{R}$ is called a Borel set if $B \in \mathcal{B}$.

By definition, every open set is a Borel set. Moreover, since the Borel sets are a σ -algebra, the complement of any Borel set is a Borel set, and any countable union of Borel sets is a Borel set.

Theorem 2.14. Every Borel set is measurable.

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

Theorem 2.15. Every open set, closed set, F_σ set, or G_δ set is a Borel set.

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

Theorem 2.16. The Borel algebra is generated by the collection of all open intervals.

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

We see that open sets, closed sets, F_σ sets and G_δ sets are among the simplest of the Borel sets.

2.2 Non-measurable Sets

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

Definition 2.17. Let C be a collection of non-empty sets. Then we can choose a member from each set in that collection. In other words, there exists a function f defined on C with the property that, for each set S in the collection, $f(S)$ is a member of S .

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

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

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

Theorem 2.18. There exists a non-measurable set.

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

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

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

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

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

$$x = x_\alpha + r_n$$

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

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

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

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

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

Theorem 2.19. If T is a measurable set with positive measure. Define

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

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

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

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

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

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

and so that

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

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

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

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

Unit III

Course Structure

- Measurable functions : Definition on a measurable set in $\mathbb{R}$ and basic properties,
 - Simple functions, Sequences of measurable functions,
 - Measurable functions as the limits of sequences of simple functions,
 - Lusin's theorem on restricted continuity of measurable functions,
 - Egoroff's theorem,
 - Convergence in measure
-

Introduction

Measurable functions in measure theory are analogous to continuous functions in topology. A continuous function pulls back open sets to open sets, while a measurable function pulls back measurable sets to measurable sets. We begin this unit defining measurable functions and discuss a few examples of measurable and non-measurable sets. The most common example of measurable functions are the simple functions. So, we will learn to approximate any measurable function with sequence of simple functions.

Objectives

After reading this unit, you will be able to

- differentiate between measurable and non-measurable sets
- learn some examples of measurable and non-measurable sets
- learn about the sequence of measurable functions and their properties
- learn to approximate measurable functions with sequence of simple functions
- learn about convergence in measure
- learn Lusin's theorem and Egoroff's theorem with their applications
- learn about the convergence in measure

3 Measurable Functions

Definition 3.1. An extended real valued function f defined on a Lebesgue measurable set E is said to be Lebesgue measurable function if for each $a \in \mathbb{R}$, the set $\{x : f(x) > a\}$ is measurable.

In practice, the domain of definition of f will usually be either $\mathbb{R}$ or $\mathbb{R} \setminus F$, where $m(F) = 0$.

Theorem 3.2. The following statements are equivalent:

- (a.) f is a measurable function,
- (b.) $\forall a \in \mathbb{R}$, the set $\{x : f(x) \geq a\}$,
- (c.) $\forall a \in \mathbb{R}$, the set $\{x : f(x) < a\}$,
- (d.) $\forall a \in \mathbb{R}$, the set $\{x : f(x) \leq a\}$.

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

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

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

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

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

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

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

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

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

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

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

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

We also have an equivalent theorem as

Theorem 3.3. An extended real function f is measurable if and only if for every extended real number a , the set $\{x : f(x) = a\}$ is measurable.

Proof. For every finite a ,

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

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

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

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

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

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

Let us see an example.

Example 3.4. Let us define a function f on the closed interval $[0, 1]$ as

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

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

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

In each case, the resulting sets are measurable.

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

Definition 3.5. Let E be a subset of $\mathbb{R}$. Then the function $f : \mathbb{R} \mapsto \mathbb{R}$ is called the indicator function of E if

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

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

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

Theorem 3.6. Indicator function of a set E is measurable if and only if E is measurable.

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

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

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

Theorem 3.7. Any constant function is measurable.

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

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

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

Theorem 3.8. Every continuous function is measurable.

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

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

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

Theorem 3.9. Let f and g be two measurable functions. Then:

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

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

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

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

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

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

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

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

Hence,

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

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

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

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

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

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

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

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

5. Finally

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

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

Note 3.10. The results hold for extended real numbers except when $f + g$ is not defined. For example when $f = \infty$ and $g = -\infty$ or vice versa. Similarly for the case $f - g$.

3.1 Sequence of Measurable Functions

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

Theorem 3.11. Let $\{f_n\}$ be a sequence of measurable functions defined on the same measurable set. Then

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

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

$$\{x : \sup_{1 \leq i \leq n} f_i(x) > a\} = \bigcup_{i=1}^n \{x : f_i(x) > a\},$$

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

2. We know that

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

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

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

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

Hence, we arrive at the desired result.

4. We know that

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

and so $\inf f_n$ is measurable.

5. Since

$$\limsup f_n = \inf_{i \geq n} (\sup f_i)$$

a measurable function by the previous subparts.

6. Since

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

and so the required result. □

Definition 3.12. A property is said to hold almost everywhere(abbreviated as a.e), if it holds everywhere except on a set of measure zero.

- Exercise 3.13.**
1. If f is a measurable function and $f = g$ a.e, then show that g is also measurable.
 2. Let $\{f_n\}$ be a sequence of measurable functions converging a.e to a function f . Show that f is measurable.
 3. If f is a measurable function, then show that f^+ and f^- are measurable.

3.2 Simple Functions

Definition 3.14. A function f is said to be simple if it takes only a finite number of values.

Note that these values must be finite. Writing them as a_i , $1 \leq i \leq N$, and letting

$$A_i = \{x \in X : f(x) = a_i\},$$

we can write

$$f = \sum_{i=1}^N a_i \chi_{A_i}$$

where χ_{A_i} is the characteristic function of A_i . Notice that each set A_i are disjoint. Also, note that f is measurable if and only if all the sets A_i are measurable. We now move on to prove the Simple Function Approximation Lemma as follows:

Lemma 3.15. Let f be a Lebesgue measurable function on a Lebesgue Measurable set E . If f is bounded on E , then for all $\epsilon > 0$, there exist simple functions ϕ_ϵ and ψ_ϵ on E such that

1. $\phi_\epsilon(x) \leq f(x) \leq \psi_\epsilon(x)$ for all $x \in E$.
2. $0 \leq \phi_\epsilon(x) - \psi_\epsilon(x) < \epsilon$ for all $x \in E$.

Proof. Let f be a bounded Lebesgue measurable function on a Lebesgue Measurable set E . Then we have $m, M \in \mathbb{R}$ such that

$$m \leq f(x) \leq M, \forall x \in E$$

Consider the interval $[m, M]$. Let $P = \{m = x_0, x_1, \dots, x_n = M\}$ be a partition of $[m, M]$ with $m = x_0 < x_1 < \dots < x_n = M$ with the property that $x_k - x_{k-1} < \epsilon$ for each $k \in \{1, 2, \dots, n\}$ and also for each k , define

$$E_k = f^{-1}([x_{k-1}, x_k])$$

Then each E_k is Lebesgue measurable and the collection of sets $\{E_k\}$ are mutually disjoint. For each $k \in \{1, 2, \dots, n\}$, define the functions ϕ_k and ψ_k as

$$\phi_k(x) = x_{k-1} \quad \text{and} \quad \psi_k(x) = x_k$$

Then define the functions ϕ_ϵ and ψ_ϵ by

$$\begin{aligned} \phi_\epsilon(x) &= \phi_1(x) = x_0 \quad \text{if } x \in E_1 \\ &= \phi_2(x) = x_1 \quad \text{if } x \in E_2 \\ &\dots\dots\dots \\ &= \phi_{n-1}(x) = x_{n-1} \quad \text{if } x \in E_n \end{aligned}$$

$$\begin{aligned}
\psi_\epsilon(x) &= \psi_1(x) = x_1 \text{ if } x \in E_1 \\
&= \psi_2(x) = x_2 \text{ if } x \in E_2 \\
&\dots\dots\dots \\
&= \psi_n(x) = x_n \text{ if } x \in E_n
\end{aligned}$$

Then ϕ_ϵ and ψ_ϵ are simple functions for all $x \in E$. Also, we have

$$m \leq f(x) \leq M, \forall x \in E$$

Furthermore, for each $k \in \{1, 2, \dots, n\}$, we have,

$$\psi_k - \phi_k = x_k - x_{k-1} < \epsilon \forall x \in E_k$$

and so

$$0 \leq \phi_\epsilon(x) - \psi_\epsilon(x) < \epsilon \forall x \in E$$

□

The simple function approximation lemma tells us that for any Lebesgue measurable function f on a Lebesgue measurable set E that is bounded, we can find step functions which are arbitrarily close together and such that f is between them.

We now prove the next result - The Simple Function Approximation Theorem as follows:

Theorem 3.16. Let f be an extended Lebesgue measurable function defined on a measurable set E . Then there exists a sequence $\{\phi_n\}$ of simple functions on E such that

1. $\{\phi_n\}$ converges pointwise to $f(x)$ on E .
2. $|\phi_n(x)| \leq |f(x)|$ for all $x \in E$ and for all $n \in \mathbb{N}$.

Furthermore, if $f(x) \geq 0$, then there exists a sequence of pointwise increasing simple functions with these properties.

Proof. First let $f(x) \geq 0$ for all $x \in E$. For each $n \in \mathbb{N}$, let f_n be the function defined by

$$f_n(x) = \min\{f(x), n\}$$

Since f is Lebesgue measurable, each f_n is Lebesgue measurable(verify). We show that $\{f_n\}$ converges pointwise to $f(x)$. Let $x \in E$. If $f(x)$ is finite, then for some $N \in \mathbb{N}$, we have $f(x) < N$. So, $f_N(x) = \min\{f(x), N\} = f(x)$, and for each $n \geq N$, $f_n(x) = f(x)$, so $\{f_n\}$ converges to $f(x)$. Also, if $f(x) = \infty$, then for all $n \in \mathbb{N}$, $n < f(x)$. So, $f_n(x) = \min\{f(x), n\} = n$ and so $\{f_n\}$ converges to $f(x) = \infty$.

Now, note that for each $n \in \mathbb{N}$ we have that f_n is Lebesgue measurable function defined on a Lebesgue measurable set E , and f_n is bounded on E . So for each $n \in \mathbb{N}$, let $\epsilon_n = \frac{1}{n} > 0$. Then there exists step functions ϕ_n and ψ_n such that

$$\phi_n(x) \leq f(x) \leq \psi_n(x), \forall x \in E$$

and

$$0 \leq \psi_n(x) - \phi_n(x) < \epsilon_n = \frac{1}{n}$$

Hence, for each $n \in \mathbb{N}$ and for all $x \in E$,

$$0 \leq f_n(x) - \phi_n(x) < \epsilon_n = \frac{1}{n}$$

Therefore,

$$f_n(x) - \frac{1}{n} < \phi_n(x) < f_n(x)$$

Taking pointwise limit on both sides of the above inequality, we see that $\{\phi_n(x)\}$ converges pointwise to $f(x)$ on E .

Further since $f(x) \geq 0$, let $\phi'_n(x) = \max\{\phi_1(x), \phi_2(x), \dots, \phi_n(x)\}$. The maximum of a finite set of step functions is a step function, so each ϕ'_n is a step function. Furthermore,

$$\phi'_n(x) \leq \phi'_{n+1}(x), \forall x \in E$$

So, $\{\phi'_n\}$ is an increasing sequence of step functions that converges pointwise to $f(x)$ on E .

Now, suppose that f is any real-valued Lebesgue measurable function defined on a Lebesgue measurable set E . Let

$$f^+(x) = \max\{f(x), 0\} \text{ and } f^-(x) = \max\{-f(x), 0\}$$

Then, both $f^+(x)$ and $f^-(x)$ are non-negative on E . So, by the previous case, there exist sequences of increasing step functions $\{\phi_n^{(1)}(x)\}$ and $\{\phi_n^{(2)}(x)\}$ that converge pointwise to $f^+(x)$ and $f^-(x)$ respectively. We note that

$$f(x) = f^+(x) - f^-(x), \forall x \in E$$

So, for each $n \in \mathbb{N}$, let

$$\phi_n(x) = \phi_n^{(1)}(x) - \phi_n^{(2)}(x)$$

Then $\{\phi_n(x)\}$ is an increasing sequence of step functions that converges pointwise to $f(x)$ on E . Also, for $x \in E$ and all $n \in \mathbb{N}$,

$$|\phi_n(x)| = |\phi_n^{(1)}(x) - \phi_n^{(2)}(x)| = \phi_n^{(1)}(x) - \phi_n^{(2)}(x) \leq f^+(x) - f^-(x) = |f(x)|$$

This completes the proof. □

3.3 Lusin's and Egoroff's Theorems

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

1. Every measurable set of finite measure is nearly a finite union of intervals.

2. Every measurable function is nearly continuous.
3. Every convergent sequence of functions is nearly convergent.

The second and third principles are known as Lusin's theorem and Egoroff's theorem respectively. We will first prove the Egoroff's theorem and Lusin's theorem thereafter.

Theorem 3.17. Let E be a measurable set with $m(E) < \infty$. Is $\{f_n\}$ is a sequence of measurable functions that converges pointwise to a function $f(x)$ on E , then for all $\epsilon > 0$, there exists a closed set $F \subseteq E$ such that $m(E \setminus F) < \epsilon$ and $\{f_n(x)\}$ converges to $f(x)$ uniformly on F .

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

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

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

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

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

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

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

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

We will now use these to prove the main theorem.

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

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

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

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

Now, consider

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

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

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

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

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

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

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

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

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

Theorem 3.18. Let f be a simple function defined on a measurable set E . Then for each $\epsilon > 0$, there is a continuous function g on $\mathbb{R}$ and a closed set F contained in E for which

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

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

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

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

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

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

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

Theorem 3.19. (Lusin's Theorem) Let f be a real-valued measurable function on E . Then for each $\epsilon > 0$, there is a continuous function g on $\mathbb{R}$ and a closed set F contained in E for which

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

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

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

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

$$\begin{aligned}
m(E \setminus F) &= m(E \setminus \cap_{n=0}^{\infty} F_n) \\
&= m(\cup_{n=0}^{\infty} (E \setminus F_n)) \\
&= m((E \setminus F_0) \cup (\cup_{n=1}^{\infty} (E \setminus F_n))) \\
&\leq \frac{\epsilon}{2} + \sum_{n=1}^{\infty} \frac{\epsilon}{2^{n+1}} = \epsilon
\end{aligned}$$

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

3.4 Convergence in Measure

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

Definition 3.20. Let $\{f_n\}$ be a sequence of measurable functions and f , a measurable function. Then the sequence $\{f_n\}$ converges in measure to f , provided for each $\epsilon > 0$,

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

Theorem 3.21. If a sequence of measurable functions converges in measure, then the limit function is unique a.e.

Proof. Let $f_n \rightarrow f$ in measure and $f_n \rightarrow g$ in measure. Now, since

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

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

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

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

Definition 3.22. A sequence of functions is said to be **fundamental** with respect to a particular kind of convergence if it forms a Cauchy sequence in that sense. Thus a sequence $\{f_n\}$ is fundamental in measure if for any $\epsilon > 0$,

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

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

Theorem 3.23. If $\{f_n\}$ is a sequence of measurable functions which is fundamental in measure, then there exists a measurable function f such that $f_n \rightarrow f$ in measure.

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

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

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

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

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

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

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

$$m(\limsup E_k) \leq m\left(\bigcup_{k=m}^{\infty} E_k\right) \leq \sum_{k=m}^{\infty} \frac{1}{2^k} = \frac{1}{2^{m-1}}. \quad (3.1)$$

So $\{f_{n_k}\}$ converges a.e to some measurable function f . Also from (3.1), we have that $\{f_{n_k}\}$ is uniformly fundamental in $\mathbb{R} \setminus \bigcup_{k=m}^{\infty} E_k$, for each m . So, $f_{n_k} \rightarrow f$ uniformly on $\mathbb{R} \setminus \bigcup_{k=m}^{\infty} E_k$, and hence, for every positive ϵ ,

$$m(\{x : |f_{n_k}(x) - f(x)| > \epsilon/2\}) \rightarrow 0 \text{ as } k \rightarrow \infty. \quad (3.2)$$

But,

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

If n and n_k are sufficiently large, the measure of the first set on the right is arbitrarily small, as $\{f_n\}$ is fundamental in measure. But the second set has been shown to have arbitrarily small measure by (3.2) and the result follows. $\square$

Corollary 3.23.1. Let $f_n \rightarrow f$ in measure, where each f_n and f are measurable functions. Then there exists a subsequence $\{f_{n_l}\}$ such that $f_{n_l} \rightarrow f$ a.e.

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

An analogue to Fatou's Lemma, holds with convergence a.e. replaced by convergence in measure.

Theorem 3.24. Let $\{f_n\}$ be a sequence of non-negative measurable functions and let f be a measurable function such that $f_n \rightarrow f$ in measure; then

$$\int f dm \leq \liminf \int f_n dm$$

Proof. Left as an exercise. $\square$

Exercise 3.25. 1. Show that if $f_n \rightarrow f$ in measure and $g_n \rightarrow g$ in measure, then $f_n + g_n \rightarrow f + g$ in measure.

2. Show that $f_n \rightarrow f$ in measure and α is any real number, then $\alpha f_n \rightarrow \alpha f$ in measure.

3. Show that $f_n \rightarrow f$ in measure, then $|f_n| \rightarrow |f|$ in measure.

Unit IV

Course Structure

- Integrals of non-negative simple functions.
 - The integral of non-negative measurable functions on arbitrary measurable sets in $\mathbb{R}$ using integrals of non-negative simple functions.
 - Monotone convergence theorem and Fatous lemma.
-

Introduction

Lebesgue integration is an alternative way of defining the integral in terms of measure theory that is used to integrate a much broader class of functions than the Riemann integral or even the Riemann-Stieltjes integral. The idea behind the Lebesgue integral is that instead of approximating the total area by dividing it into vertical strips, one approximates the total area by dividing it into horizontal strips. This corresponds to asking "for each y -value, how many x -values produce this value?" as opposed to asking "for each x -value, what y -value does it produce?"

Because the Lebesgue integral is defined in a way that does not depend on the structure of , it is able to integrate many functions that cannot be integrated otherwise. For example, consider the following function:

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

The function is not Riemann Integrable as we have seen. But we will see that it is Lebesgue integrable.

Objectives

After reading this unit, you will be able to:

- define Lebesgue integration
- know the difference between Lebesgue integral and Riemann integral
- integrate certain functions by the definition of Lebesgue which were not Riemann Integrable
- integrate the simple functions
- learn the Monotone Convergence theorem and Fatou's Lemma

4 Lebesgue Integration

Let a non-negative simple function ϕ has the following representation:

$$\phi(x) = \sum_{j=1}^n a_j \chi_{E_j}$$

where $a_j \in \mathbb{R}$ and χ_{E_j} is the characteristic function of a set E_j in $\mathbb{R}$. Then the integral of ϕ with respect to x is defined as the extended real number

$$\int \phi dx = \sum_{j=1}^n a_j m(E_j)$$

We employ the convention that $0.(\pm\infty) = 0$. So the integral of the function identically zero is equal to 0 irrespective of value of the measure of the sets E_j . The integral is well-defined since all the a_j are non-negative and so we do not encounter meaningless expressions as $\infty - \infty$.

Lemma 4.1. If ϕ and ψ are simple functions, then

1.

$$\int c\phi dx = c \int \phi dx \text{ for } c \geq 0$$

2.

$$\int (\phi + \psi) dx = \int \phi dx + \int \psi dx$$

3.

$$\int \phi dx = \sum_{j=1}^n a_j m(A \cap E_j), \text{ for any measurable set } A$$

4.

$$\int_{A \cup B} \phi dx = \int_A \phi dx + \int_B \phi dx$$

for any disjoint measurable sets A and B .

Proof. Left as an exercise. □

Definition 4.2. Let f be any non-negative measurable function. Then the integral of f is defined as

$$\int f dx = \sup \int \phi dx$$

where the supremum is taken over all the measurable simple functions ϕ such that $\phi \leq f$.

Definition 4.3. For any measurable set E , and any non-negative measurable function f ,

$$\int_E f dx = \int f \chi_E dx$$

is the integral of f over E . If E is an interval, say $[a, b]$, then we use the notation $\int_a^b f dx$.

To avoid confusion, we denote the Riemann integral as

$$\mathbf{R} \int_a^b f dx$$

Theorem 4.4. Let f and g be two non-negative measurable functions.

1. If $f \leq g$, then $\int f dx \leq \int g dx$.
2. If A is a measurable set and $f \leq g$ on A , then $\int_A f dx \leq \int_A g dx$.
3. If $a \geq 0$, then $\int a f dx = a \int f dx$.
4. If A and B are measurable sets and $A \supseteq B$, then $\int_A f dx \geq \int_B f dx$.

Proof. 1. We have,

$$\int f dx = \sup \int \phi dx$$

where the supremum is taken over all the measurable simple functions ϕ such that $\phi \leq f$. So, for these simple measurable functions ϕ , we have, $\phi \leq g$. Thus we obviously have

$$\int g dx \geq \sup \int \phi dx = \int f dx$$

2. Since $f \leq g$ on a measurable set A , hence by the definition of the integration of a non-negative measurable function on a measurable set A , we have

$$\begin{aligned} \int_A f dx &= \int f \chi_A dx \\ &\leq \int g \chi_A dx \quad [\text{by the previous case}] \\ &= \int_A g dx \end{aligned}$$

3. Now, if $a = 0$, then the result is obvious. So, let $a > 0$. A function ϕ is simple measurable such that $\phi \leq a f$ if and only if $\phi = a \psi$ such that ψ is simple and $\psi \leq f$. In that case,

$$\int \phi dx = a \int \psi dx$$

by the integration properties of simple functions. So,

$$\int a f dx = \sup \int \phi dx = a \sup \int \psi dx = a \int f dx.$$

4. Since $A \supseteq B$, so $\chi_A \geq \chi_B$. Thus we also have, since $f \geq 0$,

$$\chi_A f \geq \chi_B f$$

Thus, by the subpart 1, we get the desired result. □

We will now prove the Monotone Convergence Theorem and Fatou's Lemma.

Theorem 4.5. (Fatou's Lemma) Let $\{f_n\}$ be a sequence of non-negative measurable functions. Then,

$$\liminf \int f_n dx \geq \int \liminf f_n dx$$

Proof. Let $f = \liminf f_n$. Then f is a non-negative measurable function. From the definition of integration of a non-negative measurable function, the result follows for each simple measurable function ϕ with $\phi \leq f$ we have

$$\int \phi dx \leq \liminf \int f_n dx$$

Case I When $\int \phi dx = \infty$. Then from the definition of the integration of simple measurable function, for some measurable set A , we have $m(A) = \infty$ and $\phi > a > 0$ on A . We write $g_k(x) = \inf_{j \geq k} f_j(x)$, and $A_n = \{x : g_k(x) > a, \forall k \geq n\}$, a measurable set. Then $A_n \subseteq A_{n+1}$, each n . But, for each x , $\{g_k(x)\}$ is monotonically increasing and

$$\lim_{k \rightarrow \infty} g_k(x) = f(x) \geq \phi(x)$$

So, $A \subseteq \bigcup_{n=1}^{\infty} A_n$. Hence, $\lim(A_n) = \infty$. But for each n ,

$$\int f_n dx \geq \int g_n dx > am(A_n)$$

So, $\liminf \int f_n dx = \infty$ and the result is true.

Case II When $\int \phi dx < \infty$. Write $B = \{x : \phi(x) > 0\}$. Then $m(B) < \infty$. Let M be the largest value of ϕ , and if $0 < \epsilon < 1$, write $B_n = \{x : g_k(x) > (1 - \epsilon)\phi(x), k \geq n\}$, where g_k is as defined above. Then the sets B_n are measurable, $B_n \subseteq B_{n+1}$ for each n , and $\bigcup_{n=1}^{\infty} (B \setminus B_n) = \emptyset$. As $m(B) < \infty$, so there exists N such that $m(B \setminus B_n) < \epsilon$ for all $n \geq N$. So, if $n \geq N$

$$\begin{aligned} \int g_n dx &\geq \int_{B_n} g_n dx \\ &\geq (1 - \epsilon) \int_{B_n} \phi dx \\ &= (1 - \epsilon) \left(\int_B \phi dx - \int_{B \setminus B_n} \phi dx \right) \\ &\geq (1 - \epsilon) \int_B \phi dx - \int_{B \setminus B_n} \phi dx \\ &\geq \int_B \phi dx - \epsilon \int_B \phi dx - \epsilon M \end{aligned}$$

Since ϵ is arbitrary, $\liminf \int g_n dx \geq \int \phi dx$, and since $f_n \geq g_n$, we get the desired result.

□

Theorem 4.6. (Monotone Convergence Theorem) If $\{f_n\}$ is a monotonically increasing sequence of non-negative measurable functions converging to a function f , then

$$\int f dx = \lim \int f_n dx.$$

Proof. We have by Fatou's Lemma,

$$\int f dx = \int \liminf f_n dx \leq \liminf \int f_n dx$$

By hypothesis, we have, $f \geq f_n$. So, we have by the property of integration of non-negative measurable functions,

$$\int f dx \geq \int f_n dx$$

and hence,

$$\int f dx \geq \limsup \int f_n dx$$

Combining, we have the desired result. □

Theorem 4.7. Let f and g be two non-negative measurable functions. Then

$$\int (f + g) dx = \int f dx + \int g dx$$

Proof. We first consider the theorem for simple measurable functions ϕ and ψ . Let the values of ϕ be $a_1, a_2, \dots, a_n$ on the sets $A_1, A_2, \dots, A_n$ and let the values of ψ be $b_1, b_2, \dots, b_m$ on the sets $B_1, B_2, \dots, B_m$. Then the simple function $\phi + \psi$ has the value $a_j + b_j$ on the measurable set $A_i \cap B_j$. So, from the properties of the integral of simple measurable functions we can say that

$$\int_{A_i \cap B_j} (\phi + \psi) dx = \int_{A_i \cap B_j} \phi dx + \int_{A_i \cap B_j} \psi dx$$

But the union of the disjoint sets $A_i \cap B_j$ is $\mathbb{R}$. So we have by the properties of integration of simple measurable functions,

$$\int (\phi + \psi) dx = \int \phi dx + \int \psi dx$$

Let f and g be any non-negative measurable functions. Let $\{\phi_n\}, \{\psi_n\}$ be sequences of measurable simple functions ϕ_n converging to f and ψ_n converging to g . Then $\phi_n + \psi_n$ converges to $f + g$. But, by the properties of the integral of simple measurable functions, we have

$$\int (\phi_n + \psi_n) dx = \int \phi_n dx + \int \psi_n dx$$

Letting the limit as $n \rightarrow \infty$, we get the desired result. □

Theorem 4.8. Let $\{f_n\}$ be a sequence of non-negative measurable functions. Then

$$\int \sum_{n=1}^{\infty} f_n dx = \sum_{n=1}^{\infty} \int f_n dx$$

Proof. By induction, we can show that the previous theorem applies to the sum of n functions. So, if

$$S_n = \sum_{i=1}^n f_i$$

then

$$\int S_n dx = \sum_{i=1}^n \int f_i dx.$$

But, $\{S_n\}$ is a monotonically increasing sequence of functions converging to $f = \sum_{i=1}^{\infty} f_i$. So, the result follows from Monotone Convergence Theorem. $\square$

Exercise 4.9. 1. Show that if f is a non-negative measurable function, then

$$f = 0 \text{ a.e. iff } \int f dx = 0$$

Unit V

Course Structure

- The integral of Measurable functions and basic properties, Absolute character of the integral
 - Dominated convergence theorem, Inclusion of the Riemann integral
 - Riesz-Fischer theorem on the completeness of the space of Lebesgue integrable functions.
-

Introduction

The definition of the integral will now be extended to a wide class of real-valued measurable functions, not necessarily non-negative. The strength of the two main convergence theorems that will be proved in this unit, shows that the definition is the appropriate one.

Objective

After reading this unit, you will be able to

- define the integration of a measurable function
- learn their basic properties
- apply the results in various related circumstances
- learn the Lebesgue Dominated Convergence theorem

5 General Integral

Definition 5.1. If $f(x)$ is any real function, then

$$f^+(x) = \max\{f(x), 0\}, \quad f^-(x) = \max\{-f(x), 0\},$$

are said to be the positive and negative parts of f , respectively.

Theorem 5.2. If $f(x)$ is any real function, then

1. $f = f^+ - f^-$; $|f| = f^+ + f^-$; $f^+, f^- \geq 0$
2. f is measurable if, and only if, f^+ and f^- are both measurable.

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

We will now define the integration of a general measurable function.

Definition 5.3. If f is a measurable function and $\int f^+ dx < \infty$, $\int f^- dx < \infty$, we say that f is integrable, and its integral is given by

$$\int f dx = \int f^+ dx - \int f^- dx.$$

Clearly, a measurable function f is integrable if, and only if, $|f|$ is so, and then

$$\int |f| dx = \int f^+ dx + \int f^- dx.$$

Definition 5.4. If E is a measurable set, f is a measurable function, and $\chi_E f$ is integrable, we say that f is integrable over E , and its integral is given by

$$\int_E f dx = \int \chi_E f dx.$$

The notation $f \in L(E)$ is then sometimes used.

Definition 5.5. If f is a measurable function such that at least one of $\int f^+ dx$, $\int f^- dx$ is finite, then

$$\int f dx = \int f^+ dx - \int f^- dx.$$

Theorem 5.6. Let f and g be integrable functions. Then

1. af is integrable, and $\int af dx = a \int f dx$.
2. $f = g$ is integrable, and $\int (f + g) dx = \int f dx + \int g dx$.
3. If $f = 0$, a.e., then $\int f dx = 0$.
4. If $f \leq g$ a.e., then $\int f dx \leq \int g dx$.
5. If A and B are disjoint measurable sets, then

$$\int_A f dx + \int_B f dx = \int_{A \cup B} f dx.$$

Proof. 1. Suppose that $a \geq 0$. Then

$$(af)^+ = af^+, \quad (af)^- = af^-.$$

So,

$$\int (af)^+ dx < \infty, \quad \int (af)^- dx < \infty.$$

So af is integrable and

$$\int af dx = \int af^+ dx - \int af^- dx = a \int f dx.$$

Suppose that $a = -1$. Then

$$(-f)^+ = f^-, \quad (-f)^- = f^+.$$

So, $-f$ is integrable and

$$\int (-f) dx = \int f^- dx - \int f^+ dx = - \int f dx.$$

But for $a < 0$, $af = -|a|f$, so

$$\int af dx = - \int |a|f dx = -|a| \int f dx = a \int f dx$$

and hence the result follows.

2. We have

$$\begin{aligned} (f+g)^+ &\leq f^+ + g^+, \\ (f+g)^- &\leq f^- + g^-. \end{aligned}$$

So $f+g$ is integrable. Also,

$$(f+g)^+ - (f+g)^- = f+g = f^+ + g^+ - f^- - g^-,$$

so

$$(f+g)^+ + f^- + g^- = (f+g)^- + f^+ + g^+.$$

Applying a previous theorem and rearranging the terms, we get the desired result.

3. $f^+ = 0$, a.e., and $f^- = 0$, a.e, so $\int f^+ dx = \int f^- dx = 0$.

4. $g = f + (g - f)$, so

$$\int f dx = \int f dx + \int (g-f)^+ - \int (g-f)^- dx.$$

But, $(g-f)^- = 0$, a.e, and so the result follows from 3.

5. We have

$$\chi_{A \cup B} = \chi_A + \chi_B,$$

so by definition, and 2, the result follows.

□

Theorem 5.7. (Lebesgue Dominated Convergence Theorem) Let $\{f_n\}$ be a sequence of measurable functions such that $|f_n| \leq g$, where g is integrable, and let $\lim f_n = f$ a.e. Then f is integrable and

$$\lim \int f_n dx = \int f dx.$$

Proof. Since for each n , $|f_n| \leq g$, we have $|f| \leq g$ a.e and so f_n and f are integrable. Also, $\{g + f_n\}$ is a sequence of non-negative measurable functions, so by Fatou's Lemma

$$\liminf \int (g + f_n) dx \geq \int \liminf (g + f_n) dx.$$

So, $\int g dx + \liminf \int f_n dx \geq \int g dx + \int f dx$. But, $\int g dx$ is finite so

$$\liminf \int f_n dx \geq \int f dx.$$

Again, $\{g - f_n\}$ is also a sequence of non-negative measurable functions, so

$$\liminf \int (g - f_n) dx \leq \int \liminf (g - f_n) dx.$$

So, $\int g dx - \limsup \int f_n dx \geq \int g dx - \int f dx$. So, $\limsup \int f_n dx \leq \int f dx \leq \liminf \int f_n dx$. So, the result follows. $\square$

Example 5.8. With the same hypothesis as the above theorem, we have $\lim \int |f_n - f| dx = 0$. In fact, $|f_n - f| \leq 2g$, for each n . So applying the Lebesgue Dominated Convergence theorem, we get the desired result.

Theorem 5.9. Let $\{f_n\}$ be a sequence of integrable functions such that

$$\sum_{n=1}^{\infty} \int |f_n| dx < \infty.$$

Then the series $\sum_{n=1}^{\infty} f_n(x)$ converges a.e., its sum $f(x)$ is integrable and

$$\int f dx = \sum_{n=1}^{\infty} \int f_n dx.$$

Proof. Let $\phi(x) = \sum_{i=1}^{\infty} |f_i|$. Then by the given condition, $\int \phi dx < \infty$, so ϕ is finite-valued a.e. $\square$

- Exercise 5.10.**
1. Show that if f and g are measurable, $|f| \leq |g|$ a.e., and g is integrable, then f is integrable.
 2. Show that if f is an integrable function, then $|\int f dx| \leq \int |f| dx$.
 3. Show that if f is integrable, then f is finite-valued a.e.

References

1. Principles of Mathematical Analysis, W. Rudin
2. Real Analysis, H.L. Royden
3. Real Analysis, B.K. Lahiri and K. C. Ray
4. Theory of Integrals of a Real Variable, I. P. Natanson (Vol. I and II)

Block - II
Marks : 32 (SEE : 25; IA : 07)

Complex Analysis - II

Syllabus

Unit	Topic
7	Contour integration. Conformal mapping, Bilinear transformation. Idea of analytic continuation.
8	Multivalued functions branch point. Idea of winding number.
9	Zeros of an analytic function. Singularities and their classification. Limit points of zeros and poles. Riemanns theorem. Weierstrass-Casorati theorem.
10	Theory of residues. Argument principle. Rouches theorem. Maximum modulus theorem. Schwarz lemma. Behaviour of a function at the point at infinity.

Unit VII

Course Structure

- Contour Integration
 - Conformal mapping, Bilinear transformation.
 - Analytic continuation.
-

6 Contour Integration

Definition 6.1. A set of points $z=(x,y)$ in a complex plane is said to be an arc if

$$x = x(t), \quad y = t(t) \quad a \leq t \leq b$$

where $x(t)$ and $y(t)$ are continuous functions of the real parameter t . This definition establishes a continuous mapping of the interval $a \leq t \leq b$ into the z -plane; and the image points are ordered according to increasing values of t . The points of an arc C by means of the equation

$$\begin{aligned} z &= z(t) \\ &= x(t) + iy(t) \end{aligned}$$

Definition 6.2. The arc C is called a simple arc, or a Jordan arc, if it does not cross itself; that is, C is said to be simple if $z(t_1) \neq z(t_2)$ when $t_1 \neq t_2$. And when the arc is simple except at the end points, that is, $z(a) = z(b)$, then C is a simple closed curve.

Example 6.3. The unit circle $z = e^{i\theta}$, $0 \leq \theta \leq 2\pi$ is a simple closed curve about the origin, oriented in the anti-clockwise sense while $z = e^{-i\theta}$ describes a simple closed curve in the clockwise sense.

Example 6.4. The points on the arc

$$z = e^{i2\theta} \quad (0 \leq \theta \leq 2\pi)$$

are the same as those making up the arcs in the previous example. The arc here differs, however, from each of those arcs since the circle is traversed twice in the anticlockwise sense.

We now turn to integrals of complex-valued function f of the complex variable z . Such an integral is defined in terms of the values $f(z)$ along a given contour C , extending from a point $z = z_1$ to a point $z = z_2$ in the complex plane. It is therefore, a line integral; and its value depends on the contour C as well as on the function f . It is written as $\int_C f(z)dz$ or $\int_{z_1}^{z_2} f(z)dz$, the latter notation often being used when the value of the integral is independent of the choice of the contour taken

between the two fixed points.
Let the equation

$$z = z(t) \quad a \leq t \leq b \quad (6.1)$$

represents a contour C , extending from a point $z_1 = z(a)$ to a point $z_2 = z(b)$. Let the function f be piecewise continuous on C . We define the line integral, or *contour integral* of f along C as follows:

$$\int_C f(z)dz = \int_a^b f[z(t)]z'(t)dt \quad (6.2)$$

Note that, since C is a contour, $z'(t)$ is also piecewise continuous on the interval $a \leq t \leq b$; so the existence of the integral (8) is ensured.

It follows immediately from the definition (8) that for any constant z_0 ,

$$\int_C z_0 f(z)dz = z_0 \int_C f(z)dz$$

and,

$$\int_C [f(z) + g(z)]dz = \int_C f(z)dz + \int_C g(z)dz$$

Associated with the contour C used in the integral (8), is the contour $-C$, consisting of the set of points but with the order reversed so that the new contour extends from the point z_2 to z_1 . The contour $-C$ has the parametric representation

$$z = z(-t) \quad -b \leq t \leq -a$$

so, we have

$$\begin{aligned} \int_{-C} f(z)dz &= \int_{-b}^{-a} f[z(-t)] \frac{d}{dt} z(-t) dt \\ &= - \int_{-b}^{-a} f[z(-t)] z'(-t) dt \end{aligned}$$

where $z'(-t)$ denotes the derivative of $z(t)$ with respect to t , evaluated at $-t$. Making the substitution $\tau = -t$ in the last integral, we obtain

$$\int_{-C} f(z)dz = - \int_a^b f[z(\tau)] z'(\tau) d\tau$$

which is the same as

$$\int_{-C} f(z)dz = - \int_C f(z)dz \quad (6.3)$$

Now, consider a path C , denoted as $z = z(t)$, $a \leq t \leq b$, that consists of a contour C_1 from z_1 to z_2 followed by a contour C_2 from z_2 to z_3 , the initial point of C_2 being the final point of C_1 . There is a value c of t , $a < c < b$, such that $z(c) = z_2$. Also,

$$C = C_1 + C_2$$

Also, we can represent C_1 as

$$z = z(t) \quad a \leq t \leq c$$

and C_2 as

$$z = z(t) \quad c \leq t \leq b$$

By a rule of the integrals of the functions $w(t)$, we have,

$$\int_a^b f[z(t)]z'(t)dt = \int_a^c f[z(t)]z'(t)dt + \int_c^b f[z(t)]z'(t)dt$$

which means that,

$$\int_C f(z)dz = \int_{C_1} f(z)dz + \int_{C_2} f(z)dz$$

Sometimes the contour C is called the *sum* of its legs C_1 and C_2 and is denoted by $C_1 + C_2$.

Definite integrals in calculus can be interpreted as areas, and they have other interpretations as well. Except in special cases, no corresponding helpful interpretation, geometric or physical, is available for integrals in the complex plane. Let us look into a few examples for contour integrals.

Example 6.5. Let us find the value of the integral

$$I = \int_C \bar{z}dz$$

where C is the right-hand half

$$z = e^{2i\theta} \quad \left(-\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}\right)$$

of the circle $|z| = 2$, from $z = -2i$ to $z = 2i$. By the definition above, we have

$$I = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} 2e^{-i\theta}(2e^{i\theta})'d\theta;$$

and since $\overline{e^{i\theta}} = e^{-i\theta}$ and $(e^{i\theta})' = ie^{i\theta}$, this means that

$$I = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} 2e^{-i\theta}2ie^{i\theta}d\theta = 4 \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} d\theta.$$

Note that when a point z is on the circle $|z| = 2$, it follows that $z\bar{z} = 4$, or $\bar{z} = 4/z$. Hence the result $I = 4\pi i$ can also be written as

$$\int_C \frac{dz}{z} = \pi i$$

Example 6.6. Let C_1 denote the contour OAB which joins the origin $O(0, 0)$ to $A(0, 1)$ and $B(1, 1)$; and C_2 be the contour joining B to O . Suppose we are to find the integral

$$\int_C f(z)dz$$

where, $C = C_1 + C_2$ and $f(x, y) = y - x - i3x^2$. Then,

$$\int_C f(z)dz = \int_{C_1} f(z)dz + \int_{C_2} f(z)dz$$

Now,

$$\int_{C_1} f(z)dz = \int_{OA} f(z)dz + \int_{AB} f(z)dz \quad (6.4)$$

The leg OA can be represented as $z = 0 + iy$, $0 \leq y \leq 1$; and since $x = 0$ at points on that leg, the values of f there vary with the parameter y according to the equation $f(z) = y$, $0 \leq y \leq 1$. As a result,

$$\begin{aligned} \int_{OA} f(z)dz &= \int_0^1 yidy \\ &= i \int_0^1 ydy \\ &= \frac{i}{2} \end{aligned}$$

On the leg AB , $z = x + i$, $0 \leq x \leq 1$. So,

$$\begin{aligned} \int_{AB} f(z)dz &= \int_0^1 (1 - x - i3x^2)dx \\ &= \int_0^1 (1 - x)dx - 3i \int_0^1 x^2dx \\ &= \frac{1}{2} - i \end{aligned}$$

So, by equation (10), we have,

$$\int_{C_1} f(z)dz = \frac{1 - i}{2}$$

Again, the leg BO has parametric representation $z = x + ix$, $0 \leq x \leq 1$. So,

$$\begin{aligned} \int_{C_2} f(z)dz &= \int_1^0 -i3x^2(1 + i)dx \\ &= 3(1 - i) \int_1^0 x^2dx \\ &= i - 1 \end{aligned}$$

Then, the integrals of f along the paths C_1 and C_2 have different values even though those paths have the same initial and the same final points. Now, we have

$$\int_C f(z)dz = \frac{i - 1}{2}$$

6.1 Conformal Mappings

Let C be a smooth arc represented by the equation

$$z = z(t) \quad a \leq t \leq b \quad (6.5)$$

and let $f(z)$ be a function defined at all points z on C . The equation

$$w = f(z(t)) \quad a \leq t \leq b \quad (6.6)$$

is a parametric representation of the image γ of C under the transformation $w = f(z)$.

Suppose that C passes through a point $z_0 = z(t_0)$, at which f is analytic and that $f'(t_0) \neq 0$. According to the chain rule of differentiation, we get

$$w'(t_0) = f'(z_0)z'(t_0) \quad (6.7)$$

and this gives

$$\arg w'(t_0) = \arg f'(z_0) + \arg z'(t_0) \quad (6.8)$$

Now, let ψ_0 be a particular value of $\arg f'(z_0)$, and let θ_0 be the angle of inclination of a directed line tangent to γ at z_0 . Then from equation (10) we get that

$$\phi_0 = \psi_0 + \theta_0 \quad (6.9)$$

which is the value of $\arg w'(t_0)$ and is hence the angle of inclination of a directed line tangent to γ at the point $w_0 = f(z_0)$. Hence the angle of inclination of the directed line at w_0 differs from the angle of inclination of the directed line at z_0 by the angle of rotation

$$\psi_0 = \arg f'(z_0) \quad (6.10)$$

Now, let C_1 and C_2 be two smooth arcs passing through z_0 , and let θ_1 and θ_2 be angles of inclination of directed lines tangent to C_1 and C_2 respectively, at z_0 . We know from the preceding paragraph that the quantities

$$\begin{aligned} \phi_1 &= \psi_0 + \theta_1 \\ \phi_2 &= \psi_0 + \theta_2 \end{aligned}$$

are the angles of inclination of directed lines tangent to the image curves γ_1 and γ_2 , respectively, at the point $w_0 = f(z_0)$. Thus, $\phi_2 - \phi_1 = \theta_2 - \theta_1$, that is, the angle $\phi_2 - \phi_1$ from γ_1 to γ_2 is the same in magnitude and sense as the angle $\theta_2 - \theta_1$ from C_1 to C_2 .

Because of this angle preserving property, a transformation $w = f(z)$ is said to be *conformal* at a point if f is analytic there and $f'(z_0) \neq 0$.

Example 6.7. The mapping $w = e^z$ is conformal throughout the entire complex plane since $(e^z) = e^z \neq 0$ for each z .

A mapping that preserves the magnitude of the angle between two smooth arcs but not necessarily the sense is called an *isogonal mapping*.

Example 6.8. The transformation $w = \bar{z}$, which is a reflection in the real axis, is isogonal but not conformal.

Suppose that f is not a constant function and is analytic at a point z_0 . If, in addition, $f'(z_0) = 0$, then z_0 is called the *critical point* of the transformation f .

Example 6.9. The point $z = 0$ is a critical point of the transformation $w = 1 + z^2$.

We now come across an important example of conformal maps of the unit disk.

Example 6.10. Consider a map $f : D \mapsto D$ of the form

$$f(z) = \frac{a - z}{1 - \bar{a}z}$$

where D is the open unit disk and a is a point inside D . We note that f is analytic in the unit disk since

$$\left| \frac{1}{\bar{a}} \right| = \frac{1}{|a|} > 1$$

since $|a| < 1$. Now,

$$f'(z) = -\frac{1 + |a|^2}{(1 - \bar{a}z)^2} \neq 0$$

in D . Hence f is conformal in D . Also, we can check that,

$$(f \circ f)(z) = z$$

This shows that f is a bijective conformal map from D onto D .

The above example is in fact a conformal map moving the origin to the point a . In fact, there are many such bijective conformal maps from the unit circle onto itself. We will discuss them shortly. For now, let us see the following definition:

Definition 6.11. Let U and V be any two subsets of the complex plane. Then U and V are said to be conformally equivalent if there exists a bijective analytic map $f : U \rightarrow V$. An important fact is that, given such a function f , its inverse is automatically analytic.

We have the following theorem in this context:

Theorem 6.12. If $f : U \rightarrow V$ is analytic and injective, then $f'(z_0) \neq 0$ for all $z_0 \in U$. In particular, the inverse of f defined on its range is analytic, and thus the inverse of a conformal map is also analytic.

6.1.1 The Unit Disk and the Upper Half Plane

The upper half plane, denoted by $\mathbb{H}$, consists of the complex numbers with positive imaginary part, that is

$$\mathbb{H} = \{z \in \mathbb{C} : \text{Im}(z) > 0\}$$

A remarkable fact, which at first seems surprising, is that the unbounded set $\mathbb{H}$ is conformally equivalent to the unit disc. Moreover, an explicit formula giving this equivalence exists. Indeed, let

$$F(z) = \frac{i - z}{i + z}, \quad G(w) = i \frac{1 - w}{1 + w}$$

Theorem 6.13. The map $F : \mathbb{H} \rightarrow D$ is a conformal map with inverse $F : D \rightarrow \mathbb{H}$.

Proof. First we observe that both the maps are analytic in their respective domains. Then, notice that any point on the upper half plane is closer to i than to $-i$. So, $|F(z)| < 1$ and F maps $\mathbb{H}$ into D . To prove that G maps into the upper half-plane, we must compute $\text{Im}(G(w))$ for $w \in D$. For this, let $w = u + iv$. Then

$$\begin{aligned} \text{Im}(G(w)) &= \text{Re} \left(\frac{1 - u - iv}{1 + u + iv} \right) \\ &= \text{Re} \left(\frac{(1 - u - iv)(1 + u - iv)}{(1 + u)^2 + v^2} \right) \\ &= \frac{1 - u^2 - v^2}{(1 + u)^2 + v^2} > 0 \end{aligned}$$

since $|w| < 1$. Hence G maps the D to the upper half plane. Finally,

$$\begin{aligned} F(G(w)) &= \frac{i - i \frac{1-w}{1+w}}{i + i \frac{1-w}{1+w}} \\ &= \frac{1 + w - 1 + w}{1 + w + 1 - w} \\ &= w \end{aligned}$$

We can similarly show that $G(F(z)) = z$. This proves the theorem. $\square$

An interesting aspect of these functions is their behaviour on the boundaries of our open sets. Observe that F is analytic everywhere on $\mathbb{C}$ except at $z = -i$, and in particular it is continuous everywhere on the boundary of $\mathbb{H}$, namely the real line. If we take $z = x$ real, then the distance from x to i is the same as the distance from i to $-i$, therefore $|F(x)| = 1$. Thus F maps $\mathbb{R}$ onto the boundary of D .

Example 6.14. Translations and dilations provide simple examples. If $h \in \mathbb{C}$, the translation $z \mapsto z + h$ is a conformal map from $\mathbb{C}$ to itself whose inverse is $w \mapsto w - h$.

For any non-zero complex number c , the map $z \mapsto cz$ is a map from $\mathbb{C}$ to itself, whose inverse is $w \mapsto c^{-1}w$. If $|c| = 1$, then $c = e^{i\psi}$ for some real ψ , then the map is a rotation by angle ψ . If $c > 0$ then the map is a dilation. Finally, if $c < 0$ the map consists of a dilation by the factor $|c|$, followed by a rotation of π .

Example 6.15. The map

$$f(z) = \frac{1+z}{1-z}$$

takes the upper half disc $\{z \in \mathbb{C} : |z| < 1 \text{ and } \text{Im}(z) > 0\}$ conformally to the first quadrant $\{w = u + iv : u > 0 \text{ and } v > 0\}$. Indeed, if $z = x + iy$, then

$$f(z) = \frac{1 - (x^2 + y^2)}{(1 - x)^2 + y^2} + i \frac{2y}{(1 - x)^2 + y^2}$$

The inverse map, given by

$$g(w) = \frac{w - 1}{w + 1}$$

is clearly analytic in the first quadrant. Moreover, for all w in the first quadrant, $|w + 1| > |w - 1|$, since the distance of w to -1 is greater than the distance of w from 1 . Thus g maps into the disc D . An easy calculation also shows that the imaginary part of $g(w)$ is positive whenever w is in the first quadrant. So we conclude that f is conformal since g is the inverse of f .

To examine the action of f on the boundary, note that if $z = e^{i\theta}$ belongs to the upper half circle, then

$$\begin{aligned} f(z) &= \frac{1 + e^{i\theta}}{1 - e^{i\theta}} \\ &= \frac{e^{-i\theta/2} + e^{i\theta/2}}{e^{-i\theta/2} - e^{i\theta/2}} \\ &= \frac{i}{\tan(\theta/2)} \end{aligned}$$

As θ travels from 0 to π , we see that $f(e^{i\theta})$ travels along the imaginary axis from infinity to 0 . Moreover, if $z = x$ is real, then

$$f(z) = \frac{1+x}{1-x}$$

is also real. So, we see that f is actually a bijection from $(-1, 1)$ to the positive real axis, with $f(x)$ increasing from 0 to infinity as x travels from -1 to 1 . Note also that $f(0) = 1$.

6.2 Bilinear Transformations

Bilinear Transformations are a special kind of conformal maps. We formally define them as follows:

Definition 6.16. A mapping of the form

$$S(z) = \frac{az + b}{cz + d}$$

is called a linear fractional transformation. If a, b, c , and d also satisfy $ad - bc \neq 0$, then S is called a Möbius Transformation.

Example 6.17. Let S be a Möbius Transformation on the upper half plane $\mathbb{H}$ where,

$$S(z) = \frac{az + b}{cz + d}$$

where $ad - bc > 0$. Then S maps $\mathbb{H}$ conformally onto itself.

If S is a Möbius Transformation, then

$$S^{-1}(z) = \frac{dz - b}{-cz + a}$$

satisfies $S(S^{-1}(z)) = S^{-1}(S(z)) = z$. Then, S^{-1} is the inverse of S . Now, if S and T are both linear fractional transformations, then it follows that $S \circ T$ is also so. S can be defined on the extended complex plane $\mathbb{C}_\infty$ with $S(\infty) = a/c$ and $S(-d/c) = \infty$. Since S has inverse, it maps $\mathbb{C}_\infty$ to $\mathbb{C}_\infty$. If $S(z) = z + a$, it is called Translation; if $S(z) = az$ with $a \neq 0$, it is called dilation; if $S(z) = \exp^{i\theta}$, it is called rotation; if $S(z) = 1/z$, it is called inversion.

Theorem 6.18. If S is a Möbius Transformation, then S is the composition of translations, dilations, rotations, and inversions.

Proof. First, let us suppose that $c = 0$. Hence, $S(z) = \frac{a}{d}z + \frac{b}{d}$. Hence, $S_1(z) = \frac{a}{d}z$, then $S_2 \circ S_1 = S$ and we are done.

Now, let $c \neq 0$ and put $S_1(z) = z + \frac{d}{c}$, $S_2(z) = \frac{1}{z}$, $S_3(z) = \frac{ad-bc}{c^2}z$, $S_4(z) = z + \frac{a}{c}$. Then,

$$S \equiv S_4 \circ S_3 \circ S_2 \circ S_1$$

Hence the result. □

Now let us investigate about the fixed points of S . Fixed points of S are precisely the roots of the equation

$$\begin{aligned} S(z) &= z \\ \implies cz^2 + (d - a)z - b &= 0 \end{aligned}$$

Hence, a Möbius Transformation can have at most two fixed points unless it is the identity transformation.

Now, let S_1 be a Möbius Transformation and let a, b, c be distinct points in $\mathbb{C}_\infty$ with, $S_1(a) = \alpha$, $S_1(b) = \beta$, $S_1(c) = \gamma$. Let S_2 be another Möbius Transformation with the same property. Then, $S_2^{-1} \circ S_1$ has a, b, c as fixed points. So, $S_2^{-1} \circ S_1$ is identically equal to the identity transformation. Thus, $S_1 \equiv S_2$. Hence, a Möbius Transformation is uniquely determined by its action on three distinct points in $\mathbb{C}_\infty$.

Let z_1, z_2, z_3 be points in $\mathbb{C}_\infty$. Define $S : \mathbb{C}_\infty \mapsto \mathbb{C}_\infty$ by

$$\begin{aligned} S(z) &= \left(\frac{z - z_3}{z - z_4} \right) \left(\frac{z_2 - z_4}{z_2 - z_3} \right) \text{ if } z_2, z_3, z_4 \in \mathbb{C} \\ &= \frac{z - z_3}{z - z_4} \text{ if } z_2 = \infty \\ &= \frac{z_2 - z_4}{z - z_4} \text{ if } z_3 = \infty \\ &= \frac{z - z_3}{z_2 - z_3} \text{ if } z_4 = \infty \end{aligned}$$

In all the cases $S(z_2) = 1$, $S(z_3) = 0$, and $S(z_4) = \infty$ and S is the only transformation having the property.

Definition 6.19. If $z_1 \in \mathbb{C}_\infty$, then (z_1, z_2, z_3, z_4) , called the *cross ratio* of z_1, z_2, z_3 and z_4 , is the image of z_1 under the unique Möbius Transformation that takes z_2 to 1, z_3 to 0 and z_4 to ∞ .

6.3 Analytic Continuation

Our main aim in this section is whether a given domain of an analytic function can be extended to a larger one. We now formally define analytic continuation of a function analytic in a domain D .

Definition 6.20. Let the function f be analytic in a domain D . If there exists another function g analytic in a domain D_1 such that

1. $D \cap D_1 \neq \emptyset$,
2. $f(z) = g(z), \forall z \in D \cap D_1$,

then we say that g is a direct analytic continuation of f and vice-versa.

Theorem 6.21. (Uniqueness of analytic continuation) Let f be analytic in domain D . If f_1, f_2 be two direct analytic continuations of f in the domain D_1 , then $f_1(z) = f_2(z), \forall z \in D_1$.

Proof. By hypothesis, $f(z) = f_1(z) = f_2(z), \forall z \in D \cap D_1$, where $D \cap D_1 \neq \emptyset$. Hence, f_1, f_2 which are analytic in D_1 coincide in the subdomain $D \cap D_1$. Hence, by the interior uniqueness theorem for analytic functions we have, $f_1(z) = f_2(z), \forall z \in D_1$. $\square$

The function F defined by

$$\begin{aligned} F(z) &= f(z), \quad z \in D \\ &= f_1(z), \quad z \in D_1 \end{aligned}$$

is analytic in the union domain $D \cup D_1$.

Example 6.22. Let $f(z) = \sum_{n=0}^{\infty} z^n$ and $g(z) = \frac{1}{1-z}$. Then, f is defined and analytic only in the disc $D : |z| < 1$ and $f(z) = \frac{1}{1-z}$ in D . f is not defined when $|z| \geq 1$. But the function g is analytic in the domain $D_1 : \mathbb{C} - 1$. Hence, $f(z) = g(z) \forall z \in D \cap D_1 (= D)$. Hence, g is a direct analytic continuation of f from D to D_1 .

Unit VIII

Course Structure

- Multivalued functions
 - Branch cuts, branch point.
-

7 Multivalued Functions

A *multi-valued function* $f(z)$ has two or more distinct values for each value of z . Important multivalued functions are $\sqrt{z}$, $z^{1/n}$, $\sqrt{(z-a)(z-b)}$, $\log z$, z^α , for any non-integral α , and the inverse trigonometric functions. In order to use the multi-valued functions, we must use a single branch to ensure single-valuedness. This is done by introducing cuts in the convenient positions of the complex plane, which are commonly called the *branch cuts*. In order to illustrate this, let us use the simplest example, say $w = \sqrt{z}$. If $z = re^{i\theta}$, then the square root has two branches: $w_1 = r^{1/2}e^{i\theta/2}$ and $w_2 = -r^{1/2}e^{i\theta/2}$.

Let us go around $z = 0$. If we start at some point $z_0 = re^{i\theta}$ with the branch w_1 , and go along any closed curve around $z = 0$, then r remains the same while θ changes to $\theta + 2\pi$. As z moves, w_1 changes continuously, and when z returns to the original point, w_1 has changed to w_2 . Going around $z = 0$ again changes w_2 again to w_1 . Any other path which excludes $z = 0$ has ordinary single-valued behaviour. The point $z = 0$, in this case acts as a singularity called *branch point* for the function and is very different from pole. The function $\sqrt{z}$ is double valued in any region which includes $z = 0$ as an interior point. If z goes on a circuit around $z = 0$, w changes to $-w$. If we stay in a region that does not include the branch point, then while z moves around w_1 does not change to w_2 . This is achieved by restricting the movement of z to a region R which excludes the branch point. For this purpose, we just introduce a "cut" in the complex plane around the branch point and extending to infinity. This is known as the *branch cut*. The function remains single-valued as long as z does not cross over the branch cut. If it crosses the cut, then w moves on to the next branch.

Example 7.1. The function $\log z$ is the inverse of the exponential. So, if

$$z = e^w$$

then,

$$w = \log z$$

The logarithm has infinite number of branches. We can write from $z = e^w$,

$$\begin{aligned} z &= re^{i\theta} \\ &= re^{i\theta + i2n\pi} \\ &= e^{\ln r + i\theta + i2n\pi} \end{aligned}$$

where, $n = 0, \pm 1, \pm 2, \dots$, any integer. So, the logarithm is

$$\log z = \ln r + i\theta + i2n\pi$$

Each time z moves in a closed curve around the branch point 0, θ increases by 2π and we go from one branch to another, and this can go on forever. θ goes through multiples of 2π and the infinite set of values of logarithm differ by $2n\pi i$. We can get a single-valued function by picking a cut, such as the negative real axis

$$\log z = \ln r + i\theta, \quad -\pi \leq \theta < \pi$$

At two points z_1, z_2 , on either side of the cut,

$$\log z_1 = \ln r - i\pi$$

$$\log z_2 = \ln r + i\pi$$

Unit IX

Course Structure

- Zeros of an analytic function
 - Singularities and their classification
 - Limit point of zeros and poles
 - Characteristics of the singularities
-

8 Zeros of an analytic function

If a function f is analytic at a point α , then α is called a zero of f if $f(\alpha) = 0$. Recall that if a function f is analytic at a point α then f has a Taylor series expansion

$$f(z) = \sum_{n=0}^{\infty} a_n (z - \alpha)^n,$$

where

$$a_n = \frac{f^{(n)}(\alpha)}{n!}.$$

If α is a zero of f then $a_0 = 0$. Other than a_0 , if $a_1 = a_2 = \cdots = a_m = 0$, then α is called a zero of f of order m . Zeros of order one and two are called simple and double zero respectively.

Theorem 8.1. Let α be a zero of f of order m . Then, near α , $f(z) = (z - \alpha)^m \phi(z)$, where ϕ is analytic at α and $\phi(\alpha) \neq 0$.

Proof. Since f is analytic at α and α is a zero of f of order m , so near α , f has a Taylor series expansion

$$\begin{aligned} f(z) &= a_m (z - \alpha)^m + a_{m+1} (z - \alpha)^{m+1} + \cdots \\ &= (z - \alpha)^m (a_m + a_{m+1} (z - \alpha) + \cdots), \end{aligned}$$

where $a_m \neq 0$. Now, let

$$\phi(z) = a_m + a_{m+1} (z - \alpha) + \cdots.$$

Then ϕ is analytic at α with $\phi(\alpha) = a_m \neq 0$. □

8.1 Singularities and their classification

For a function f , a point α is called a regular point of f if f is analytic at α . The points where f is not analytic are called the *singular points* or *singularities* of f . Let a point α is a singular point of f . If f is analytic in a deleted neighbourhood of α , then α is called an *isolated singularity* of f . If α is not an isolated singularity of f , then it is called *non-isolated singularity* of f .

Let α be an isolated singularity of f . Then we have a positive real number r , such that f has a *Laurent series* expansion of the form

$$f(z) = \sum_{n=0}^{\infty} a_n(z - \alpha)^n + \sum_{n=1}^{\infty} b_n(z - \alpha)^{-n}$$

valid in $0 < |z - \alpha| < r$. The sum

$$\sum_{n=0}^{\infty} a_n(z - \alpha)^n$$

is called the analytic part and the sum

$$\sum_{n=1}^{\infty} b_n(z - \alpha)^{-n}$$

is called the *principal part* of f at the isolated singularity α . If the principal part is a terminating series, then the point α is called the pole of f . α is called the pole of f of order m if $b_m \neq 0$ and $b_{m+1} = b_{m+2} = \dots = 0$. Poles of order one and two are called *simple pole* and *double pole* respectively. If the principal part is non-terminating, then α is called an *essential singularity* of f . And if all the coefficients b_n of the principal part are zero then α is called removable singularity of f .

Theorem 8.2. α is a pole of a function f of order m if and only if f can be written in the form

$$f(z) = \frac{\phi(z)}{(z - \alpha)^m}$$

near α , where, ϕ is analytic at α and $\phi(\alpha) \neq 0$.

Proof. If α is a pole of order m , then

$$f(z) = \sum_{n=0}^{\infty} a_n(z - \alpha)^n + \sum_{n=1}^m b_n(z - \alpha)^{-n},$$

where $b_m \neq 0$. This gives,

$$\begin{aligned} f(z) &= \left(\sum_{n=0}^{\infty} a_n(z - \alpha)^n + b_1(z - \alpha)^{m-1} + \dots + b_m \right) (z - \alpha)^{-m} \\ &= \frac{\phi(z)}{(z - \alpha)^m} \end{aligned}$$

where

$$\phi(z) = \sum_{n=0}^{\infty} a_n(z - \alpha)^n + b_1(z - \alpha)^{m-1} + \cdots + b_m$$

and $\phi(\alpha) = b_m \neq 0$

. Conversely, if

$$f(z) = \frac{\phi(z)}{(z - \alpha)^m},$$

where ϕ satisfies the given criteria, then, due to analyticity of ϕ at α ,

$$\begin{aligned} \phi(z) &= \frac{\sum_{n=0}^{\infty} a_n(z - \alpha)^n}{(z - \alpha)^m} \\ &= \frac{a_0}{(z - \alpha)^m} + \frac{a_1}{(z - \alpha)^{m-1}} + \cdots + a_m + \sum_{n=m+1}^{\infty} a_n(z - \alpha)^n. \end{aligned}$$

Hence, α is a pole of f of order m , since $a_0 \neq 0$. □

8.2 Limit points of Zeros and poles

Theorem 8.3. If α is a pole of f , then $f(z) \rightarrow \infty$ as $z \rightarrow \alpha$.

Proof. Let α be a pole of f of order m . Then,

$$f(z) = \frac{\phi(z)}{(z - \alpha)^m},$$

where, ϕ is analytic at α and $\phi(\alpha) \neq 0$. So,

$$|f(z)| = \frac{|\phi(z)|}{|z - \alpha|^m}.$$

Since ϕ is continuous at α , so for

$$\epsilon = \frac{1}{2}|\phi(\alpha)|,$$

we can find a $\delta > 0$, such that

$$|\phi(z) - \phi(\alpha)| < \epsilon$$

for $|z - \alpha| < \delta$. Hence,

$$\begin{aligned} |\phi(z)| &= |\phi(z) - \phi(\alpha) + \phi(\alpha)| \\ &\geq |\phi(\alpha)| - \frac{1}{2}|\phi(\alpha)| \\ &= \frac{1}{2}|\phi(\alpha)|, \end{aligned}$$

for $|z - \alpha| < \delta$. Hence,

$$|f(z)| > \frac{\frac{1}{2}|\phi(\alpha)|}{|z - \alpha|^m}$$

for $0 < |z - \alpha| < \delta$. Let $M > 0$ be a large number. Then $|f(z)| > M$ if

$$\begin{aligned} \frac{\frac{1}{2}|\phi(\alpha)|}{|z - \alpha|^m} &> M, \\ \text{or, } 0 < |z - \alpha|^m &< \frac{\frac{1}{2}|\phi(\alpha)|}{M} \\ \text{or, } 0 < |z - \alpha| &< \frac{\frac{1}{2}|\phi(\alpha)|^{\frac{1}{m}}}{M}. \end{aligned}$$

This shows that $f(z) \rightarrow \infty$ as $z \rightarrow \alpha$. □

Theorem 8.4. Limit point of the zeros of a function f which is not identically equal to zero is an essential singularity of the function.

Proof. Let α be a limit point of the zeros of f . Then, by definition, every neighbourhood of α contains infinitely many zeros of f . If possible, let f be analytic at α . Then f is continuous at the point α . So, for any $\epsilon > 0$, there exists a $\delta > 0$ such that $|f(z) - f(\alpha)| < \epsilon$ for $|z - \alpha| < \delta$. But, there is an infinite number of zeros of f in $|z - \alpha| < \delta$. For these zeros, we must have, $|f(\alpha)| < \epsilon$, that is, $f(\alpha) = 0$. Hence, α is a zero of f . This is impossible unless f is identically equal to zero since zeros are isolated points. Hence f is not analytic at α .

Now, if possible, let α be a pole of f . Then, $f(z) \rightarrow \infty$ as $z \rightarrow \alpha$, that is, given any number $M > 0$, we can find a $\delta > 0$ such that $|f(z)| > M$ is valid in $0 < |z - \alpha| < \delta$. But the deleted neighbourhood $0 < |z - \alpha| < \delta$ has infinite number of zeros of f so that $|f(z)| > M$ is not true for infinite number of points in $0 < |z - \alpha| < \delta$. Hence α cannot be a pole of f . Thus f has an essential singularity at α . □

Theorem 8.5. The limit point of the poles of a function f is a non-isolated essential singularity of f .

Proof. Let α be a limit point of the poles of f . Since every neighbourhood of α contains infinite number of poles of f , α cannot be a regular point of f . Hence α is a singularity of f which is non-isolated. Hence the result. □

8.3 Riemann's Theorem On Removable Singularity

Theorem 8.6. If a function f is bounded and analytic in a deleted neighbourhood $0 < |z - \alpha| < \delta$, of a point α , then either f is analytic at α or α is a removable singularity of f .

Proof. Since f is analytic in $0 < |z - \alpha| < \delta$, so there it has a Laurent expansion of the form

$$f(z) = \sum_{n=0}^{\infty} a_n(z - \alpha)^n + \sum_{n=1}^{\infty} b_n(z - \alpha)^{-n} \quad (8.1)$$

If $\gamma : |z - \alpha| = \rho$, $\rho < \delta$ then the coefficients b_n are given by

$$b_n = \frac{1}{2\pi i} \int_{\gamma} \frac{f(z)dz}{(z - \alpha)^{-n+1}} \quad (8.2)$$

for $n = 1, 2, \dots$. Since f is bounded in the deleted neighbourhood, there exists a $M > 0$ such that $|f(z)| \leq M$ in $0 < |z - \alpha| < \delta$. Hence,

$$\begin{aligned} |b_n| &= \frac{1}{2\pi} \left| \int_{\gamma} \frac{f(z)dz}{(z - \alpha)^{-n+1}} \right| \\ &\leq \frac{M}{2\pi} \rho^{n-1} 2\pi\rho \\ &= M\rho^n \end{aligned}$$

Since b'_n s are constants and ρ can be chosen arbitrarily small, we conclude $b_n = 0$, $n = 1, 2, \dots$. Hence, equation (1) reduces to

$$f(z) = \sum_{n=0}^{\infty} a_n(z - \alpha)^n$$

in $0 < |z - \alpha| < \delta$. If $f(\alpha) = a_0$, this power series representation of f is actually valid in $|z - \alpha| < \delta$, and in that case, f is analytic at α . Otherwise f can be made analytic at α by letting $f(\alpha) = a_0$. The point α is then a removable singularity of f . $\square$

8.4 Casorati-Weierstrass Theorem

Theorem 8.7. Let α be an isolated essential singularity of a function f and let z_0 be any given complex number. Then, for any $\epsilon > 0$, the inequality $|f(z) - z_0| < \epsilon$ is satisfied at some point z in each deleted neighbourhood of α .

Proof. Since α is isolated, there exists a deleted neighbourhood $0 < |z - \alpha| < \delta$ of α where f is analytic. If possible, let

$$|f(z) - z_0| < \epsilon \tag{8.3}$$

be not satisfied at any point in that deleted neighbourhood. Then $|f(z) - z_0| \geq \epsilon$ for all z in $0 < |z - \alpha| < \delta$. Let

$$g(z) = \frac{1}{f(z) - z_0},$$

in $0 < |z - \alpha| < \delta$. Then

$$|g(z)| = \left| \frac{1}{f(z) - z_0} \right| \leq \frac{1}{\epsilon}$$

in $0 < |z - \alpha| < \delta$. Hence g is bounded and analytic in $0 < |z - \alpha| < \delta$. By Riemann's Theorem on removable singularity, α is a removable singularity of g . Let $g(\alpha)$ be defined such that g is analytic at α . Since f cannot be a constant function, g is also not a constant function. Since g is analytic at α , it has a Taylor series representation at α . Hence, either $g(\alpha) \neq 0$ or g has a zero at α . So, its reciprocal

$$\frac{1}{g(z)} = f(z) - z_0$$

is either analytic at α or has a pole there which contradicts the hypothesis that α is an essential singularity of f . Hence the result. $\square$

Corollary 8.7.1. If α is an isolated essential singularity of f , then for arbitrary positive numbers δ and M , there is a point z in $0 < |z - \alpha| < \delta$ at which $|f(z)| > M$.

Hence from all the above theorems , we can conclude that

1. If $f(z)$ tends to a finite limit as z tends to α , then α is either a regular point or a removable singularity of f .
2. If $f(z) \rightarrow \infty$ as $z \rightarrow \alpha$, then α is a pole of f .
3. If $f(z)$ does not tend to any definite limit, finite or infinite, then α is an essential singularity of f .

Unit X

Course Structure

- Theory of Residues, Argument Principle
 - Rouché's Theorem
 - Maximum Modulus Theorem, Schwarz Lemma
 - Behaviour of a function at the point at infinity
-

Introduction

The inspiration behind this section is a desire for an answer to the following question: if f has an isolated singularity at $z = \alpha$, what are the possible values for $\int_{\gamma} f$ where γ is a simple closed curve not passing through α ? If the singularity is removable, then clearly the integral will be zero. If $z = \alpha$ is a pole or an essential singularity, then the answer is not always zero. We investigate the problem by introducing the theory of residues.

9 Theory of Residues

Definition 9.1. Let f has an isolated singularity at $z = \alpha$ and let

$$f(z) = \sum_{n=0}^{\infty} a_n(z - \alpha)^n + \sum_{n=1}^{\infty} b_n(z - \alpha)^{-n}$$

be its Laurent expansion about $z = \alpha$. Then the residue of f at $z = \alpha$ is the coefficient b_1 . We denote it as $\text{Res}(f; \alpha) = b_1$.

We know that

$$b_n = \frac{1}{2\pi i} \int_{\gamma} \frac{f(z)dz}{(z - \alpha)^{-n+1}}$$

where γ is any simple closed curve lying in a deleted neighbourhood $0 < |z - \alpha| < r$ of α and $\alpha \in \text{int}\gamma$. So for $n = 1$, we have

$$b_1 = \frac{1}{2\pi i} \int_{\gamma} f(z)dz$$

Hence,

$$\int_{\gamma} f(z)dz = 2\pi i b_1$$

Result 9.2. If α be a pole of f of order m , then

$$Res(f; \alpha) = \frac{1}{(m-1)!} \lim_{z \rightarrow \alpha} \frac{d^{m-1}}{dz^{m-1}} (z - \alpha)^m f(z)$$

Theorem 9.3. (Cauchy's Residue Theorem) If f is analytic within and on a simple closed curve γ except for a finite number of singular points $a_1, a_2, \dots, a_n$ within γ , then

$$\int_{\gamma} f(z) dz = 2\pi i \sum_{k=1}^n Res(f; a_k)$$

Proof. Since there are finite number of singularities, they must be isolated. Round each singularity a_k , $k = 1, 2, \dots, n$, we draw a circle γ_k with radius so small that these n circles do not intersect each other and they all lie within γ . Then f becomes analytic within and on the multiply connected region bounded by $\gamma, \gamma_1, \dots, \gamma_n$. According to Cauchy Goursat Theorem, extended to such regions, we have,

$$\begin{aligned} \int_{\gamma} f(z) dz &= \sum_{k=1}^n \int_{\gamma_k} f(z) dz \\ &= 2\pi i \sum_{k=1}^n Res(f; a_k) \end{aligned}$$

all the integration taken in the positive sense. □

The theory of residues can be used to evaluate improper integrals. We consider the following cases:

1. When we evaluate the integral of the form

$$\int_0^{2\pi} f(\cos \theta, \sin \theta) d\theta,$$

we use the transformation $z = e^{i\theta}$ and choose the contour C as the unit circle $|z| = 1$. Then,

$$\cos \theta = \frac{1}{2} \left(z + \frac{1}{z} \right), \quad \sin \theta = \frac{1}{2i} \left(z - \frac{1}{z} \right)$$

and,

$$dz = ie^{i\theta} d\theta$$

Example 9.4. Let

$$\begin{aligned} I &= \int_0^{2\pi} \frac{\cos^2 3\theta}{5 - 4 \cos 2\theta} d\theta \\ &= \frac{1}{2} \int_0^{2\pi} \frac{1 + \cos 6\theta}{5 - 4 \cos 2\theta} d\theta \\ &= \text{Real part of } \frac{1}{2} \int_0^{2\pi} \frac{1 + e^{i6\theta}}{5 - 4 \cos 2\theta} d\theta \end{aligned} \tag{9.1}$$

Let $z = e^{i\theta}$. Then, $dz = ie^{i\theta}d\theta$. Also we have,

$$\begin{aligned}\cos 2\theta &= \cos^2 \theta - \sin^2 \theta \\ &= \frac{1}{4} \left(z + \frac{1}{z} \right)^2 + \frac{1}{4} \left(z - \frac{1}{z} \right)^2 \\ &= \frac{1}{2} \left(z^2 + \frac{1}{z^2} \right)\end{aligned}$$

We take $C : |z| = 1$. Then we will get

$$\begin{aligned}I_1 &= \frac{1}{2} \int_C \frac{1 + z^6}{5 - 2 \left(z^2 + \frac{1}{z^2} \right)} \frac{dz}{iz} \\ &= -\frac{1}{2i} \int_C \frac{z(1 + z^6)}{(2z^2 - 1)(z^2 - 2)} dz\end{aligned}\tag{9.2}$$

Let

$$f(z) = \frac{z(1 + z^6)}{(2z^2 - 1)(z^2 - 2)}$$

Then the poles of f are at $z = \pm \frac{1}{\sqrt{2}}$ and $z = \pm \sqrt{2}$, of which, $z = \pm \frac{1}{\sqrt{2}}$ lie inside C . Now,

$$\text{Res} \left(f; \frac{1}{\sqrt{2}} \right) = -\frac{3}{16}$$

and,

$$\text{Res} \left(f; -\frac{1}{\sqrt{2}} \right) = -\frac{3}{16}$$

Hence, from residue theorem we get from equation (5),

$$I_1 = -\frac{1}{2i} 2\pi i \left(-\frac{3}{8} \right) = \frac{3\pi}{8}$$

Hence, we get,

$$I = \text{Real part of } \frac{3\pi}{8} = \frac{3\pi}{8}$$

2. When we want to evaluate integrals of the form

$$\int_{-\infty}^{\infty} f(x) dx$$

where, f is a rational function of the real variable x . The improper integral converges if

- (a) the degree of the denominator of f exceeds that of the numerator by more than one, and
- (b) the denominator of f does not vanish on the real line.

We assume that f satisfies both the conditions. Also we have the following lemma:

Lemma 9.5. If $f(z) \rightarrow 0$ uniformly as $z \rightarrow \infty$, then

$$\lim_{R \rightarrow \infty} \int_{C_1} f(z) dz = 0$$

where C_1 is the semi-circle $z = Re^{i\theta}$, $0 \leq \theta \leq \pi$.

If

$$f(z) = \frac{a_0 + a_1z + \cdots + a_nz^n}{b_0 + b_1z + \cdots + b_mz^m}, \quad a_n \neq 0, \quad b_m \neq 0$$

and $m > n + 1$, then $zf(z) \rightarrow 0$ uniformly as $z \rightarrow \infty$. Hence we can apply the above lemma for a function of this type.

Example 9.6. Let

$$I = \int_{-\infty}^{\infty} \frac{x^2}{(x^2 + 1)(x^2 + 4)} dx$$

We take

$$f(z) = \frac{z^2}{z^2 + 1)(z^2 + 4)}$$

and consider

$$\int_C f(z) dz$$

where, $C = C_1 + C_2$, $C_1 : z = Re^{i\theta}$, $0 \leq \theta \leq \pi$ and $C_2 : [-R, R]$, for some real R , sufficiently large. Poles of f are at $z = \pm i$ and $z = \pm 2i$ of which i and $2i$ lie within C . Now,

$$\text{Res}(f; i) = \frac{i}{6}$$

and,

$$\text{Res}(f; 2i) = -\frac{i}{3}$$

By Residue theorem,

$$\begin{aligned} \int_C f(z) dz &= 2\pi i (\text{Res}(f; i) + \text{Res}(f; 2i)) \\ \text{or, } \int_{C_1} f(z) dz + \int_{C_2} f(z) dz &= \frac{i}{6} - \frac{i}{3} \\ \text{or } \int_{C_1} f(z) dz + \int_{-R}^R \frac{x^2}{(x^2 + 1)(x^2 + 4)} dx &= \frac{\pi}{3} \end{aligned} \quad (9.3)$$

Now, since the degree of the denominator of f exceeds that of the numerator by more than one, $zf(z) \rightarrow 0$ uniformly as $z \rightarrow \infty$. Hence,

$$\lim_{R \rightarrow \infty} \int_{C_1} f(z) dz = 0$$

Taking $R \rightarrow \infty$ in Equation (6), we get,

$$\int_{-\infty}^{\infty} \frac{x^2}{(x^2 + 1)(x^2 + 4)} dx = \frac{\pi}{3}$$

9.1 Argument Principle

Suppose f is analytic and has a zero of order m at $z = a$. So, $f(z) = (z - a)^m g(z)$, where g is analytic at a and $g(a) \neq 0$. Hence,

$$\frac{f'(z)}{f(z)} = \frac{m}{z - a} + \frac{g'(z)}{g(z)} \quad (9.4)$$

and g'/g is analytic near $z = a$ since $g(z) \neq 0$. Now suppose, f has a pole of order m at a . Then, $f(z) = (z - a)^{-m} g(z)$ where g is analytic at a and $g(a) \neq 0$. This gives

$$\frac{f'(z)}{f(z)} = -\frac{m}{z - a} + \frac{g'(z)}{g(z)} \quad (9.5)$$

Again g'/g is analytic near $z = a$.

Definition 9.7. If G is open and f is a function defined and analytic in G except for poles, then f is a meromorphic function on G .

Theorem 9.8. Given a function f , meromorphic in a region G , suppose γ is a simple closed curve in G and $f(z) \neq 0$ on γ and f is analytic on γ . If, N and P denote the number of zeros and poles respectively of f within γ , multiplicities counted accordingly, then,

$$\frac{1}{2\pi i} \int_{\gamma} \frac{f'(z)}{f(z)} dz = N - P$$

Proof. Since f is analytic on γ , and $f(z) \neq 0$ on γ , the function f'/f is also analytic on γ . Also, a point within γ which is neither a pole or zero of f is a regular point of f'/f . Thus, the only singularities of f'/f within γ are the zeros and poles of f within γ . Let a be a zero of f of order n within γ . Then, by equation (4)

$$\frac{f'(z)}{f(z)} = \frac{n}{z - a} + \frac{g'(z)}{g(z)}$$

Hence, a is a simple pole of f'/f with residue n . Next let, b be a pole of f of order n . Then, by equation (5),

$$\frac{f'(z)}{f(z)} = -\frac{n}{z - b} + \frac{g'(z)}{g(z)}$$

Hence, b is a simple pole of f'/f with residue $-n$. Now, let $a_1, a_2, \dots, a_k$ be the zeros of f with respective orders $p_1, p_2, \dots, p_k$ within γ and let $b_1, b_2, \dots, b_l$ be the poles of f with respective orders $q_1, q_2, \dots, q_l$ within γ . Then the only singularities of f'/f within γ are the points $a_1, a_2, \dots, a_k, b_1, b_2, \dots, b_l$ with residues $p_1, p_2, \dots, p_k, -q_1, -q_2, \dots, -q_l$. By Residue Theorem,

$$\begin{aligned} \frac{1}{2\pi i} \int_{\gamma} \frac{f'(z)}{f(z)} dz &= (p_1 + p_2 + \dots + p_k) - (q_1 + q_2 + \dots + q_l) \\ &= N - P \end{aligned}$$

□

Corollary 9.8.1. Let a function f be analytic within and on γ and $f(z) \neq 0$ on γ . If, N denote the number of zeros of f within γ , multiplicities being counted, then,

$$\frac{1}{2\pi i} \int_{\gamma} \frac{f'(z)}{f(z)} dz = N$$

The argument principle establishes a striking relationship between the number of zeros and the number of poles of f in a domain D and how f maps the boundary δD of the domain, namely, the number of times the image of δD winds around the origin.

Example 9.9. Let $f(z) = z^2 - 1$, on $C : |z - 1| = 1$. Then

$$w = f(z) = |z|^2 e^{2i\theta} - 1$$

The change of argument of w on C is 2π

9.2 Rouché's Theorem

Theorem 9.10. If the functions f and g are analytic within and on a simple closed curve γ and if $|g(z)| < |f(z)|$ on γ , then f and $f + g$ have the same number of zeros inside γ .

Proof. For $t \in [0, 1]$, define

$$f_t(z) = f(z) + tg(z)$$

so that $f_0 = f$ and $f_1 = f + g$. Let n_t denote the number of zeros of f_t inside γ counted with multiplicities, so that in particular, n_t is an integer. The condition $|f(z)| \geq |g(z)|$ for $z \in \gamma$ clearly implies that f_t has no zeros on γ , and the argument principle implies

$$n_t = \frac{1}{2\pi i} \int_{\gamma} \frac{f'_t(z)}{f_t(z)} dz$$

To prove that n_t is constant, it suffices to show that it is a continuous function of t . Then we could argue that if n_t were not constant, the intermediate value theorem would guarantee the existence of some $t_0 \in [0, 1]$ with n_{t_0} not integral, contradicting the fact that $n_t \in \mathbb{Z}$ for all t .

To prove the continuity of n_t , we observe that $f'_t(z)/f_t(z)$ is jointly continuous for $t \in [0, 1]$ and $z \in \gamma$. This joint continuity follows from the fact that it holds for both the numerator and denominator, and our assumptions guarantee that $f_t(z)$ does not vanish on γ . Hence, n_t is integer-valued and continuous, and it must be constant. We conclude that $n_0 = n_1$, which is Rouché's theorem. $\square$

9.3 Maximum Modulus Theorem

We now come to an important geometric property of analytic functions. A mapping is said to be **open** if it maps open sets to open sets.

Theorem 9.11. (Open Mapping Theorem) If f is holomorphic and non-constant in a region G , then f is open.

Proof. Let w_0 belong to the image of f , say $w_0 = f(z_0)$. We must prove that all points w near w_0 also belong to the image of f .

Define $g(z) = f(z) - w$ and write

$$\begin{aligned} g(z) &= (f(z) - w_0) + (w_0 - w) \\ &= F(z) + G(z). \end{aligned}$$

Now choose $\delta > 0$ such that the disc $|z - z_0| \leq \delta$ is contained in G and $f(z) \neq w_0$ on the circle $|z - z_0| = \delta$. We then select $\epsilon > 0$ so that we have $|f(z) - w_0| \geq \epsilon$ on the circle. Now if $|w - w_0| < \epsilon$ we have $|F(z)| > |G(z)|$ on the circle, and by Rouché's theorem, we conclude that $g = F + G$ has a zero inside the circle since F has one. □

We use this theorem to find the size of an analytic function.

Theorem 9.12. If f is a non-constant analytic function in a region G , then f cannot attain a maximum in G .

Proof. Suppose that f did attain a maximum at z_0 . Since f is analytic, it is an open mapping, and thus, if $D \subset G$ is a small disc centred at z_0 , its image $f(D)$ is open and contains $f(z_0)$. This proves that there are points in $z \in D$ such that $|f(z)| > |f(z_0)|$, a contradiction. □

Corollary 9.12.1. Suppose that G is a region with compact closure $\bar{G}$. If f is analytic on G and continuous on $\bar{G}$ then

$$\sup_{z \in G} |f(z)| \leq \sup_{z \in \bar{G} - G} |f(z)|$$

9.4 Schwarz Lemma

Theorem 9.13. Let $D = \{z : |z| < 1\}$ and suppose that f is analytic on D with

1. $|f(z)| \leq 1$
2. $f(0) = 0$

Then $|f'(0)| \leq 1$ and $|f(z)| \leq |z|$ for all z in the disk D . Moreover, if $|f'(0)| = 1$ or if $|f(z)| = |z|$ for some $z \neq 0$, then there exists a constant c , $|c| = 1$, such that $f(w) = cw$ for all w in D .

Proof. Define $g : D \rightarrow \mathbb{C}$ by

$$\begin{aligned} g(z) &= \frac{f(z)}{z}, z \neq 0 \\ &= f'(z), z = 0 \end{aligned} \tag{9.6}$$

Then, g is analytic in D . Using Maximum Modulus Theorem,

$$|g(z)| \leq \frac{1}{r}$$

for $|z| \leq r$ and $0 < r < 1$. Letting r approach 1, we get,

$$|g(z)| \leq 1$$

for all z in D . That is,

$$|f(z)| \leq |z|$$

and $|f'(0)| = |g(0)| \leq 1$.

If $|f(z)| = |z|$ for some $z \neq 0$, in D , or $|f'(0)| = 1$, then $|g|$ assumes its maximum value inside D . Thus, by Maximum Modulus Theorem, $g(z) \equiv c$, for some constant c with $|c| = 1$. Hence, $f(z) = cz$ and hence the result. $\square$

9.5 Behaviour of a function at the point at infinity

If a function is analytic at all large values of z , we can describe its behaviour at infinity. If a function is analytic for all large values of z , we consider $F(z) = f(1/z)$, which is now analytic in a deleted neighbourhood of the origin. We say that f has a **pole at infinity** if F has a pole at the origin. Similarly, we can speak of f having an **essential singularity at infinity**, or a removable singularity (hence analytic) at infinity in terms of the corresponding behaviour of F at 0. A meromorphic function in a complex plane that is either analytic at infinity or has a pole at infinity is said to be **meromorphic in the extended complex plane**.

References

1. Theory of functions of a Complex Variable, A. I. Markushevich
2. Complex Variables and Applications, R. V. Churchill and J. W Brown
3. Functions of One Complex Variable, J. B. Conway
4. Complex Analysis, L. V. Ahlfors
5. An Introduction to the Theory of Functions of a Complex Variable, E. T. Copson

Core Paper

MATC 2.2

Block - III

Marks : 36 (SEE : 30; IA : 06)

Functional Analysis - II

Syllabus

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

Unit XI

Course Structure

- Introduction
 - Objectives
 - Linear Operators on Normed Linear Spaces
-

10 Introduction

We are familiar with the real-valued functions on the real line $\mathbb{R}$. In case of these functions, the domain is a subset of $\mathbb{R}$. Here, we will deal with more general functions, general spaces, such as Normed spaces, and mappings on them. In case of vector spaces and, in particular, normed spaces, a mapping is called an operator. The operators that preserve two algebraic operations of vector spaces, is of special interest to us.

Objectives

After reading this unit, you will be able to

- define a linear operator on a Normed space
- give examples of linear operators on normed spaces
- learn properties related to them

11 Linear Operators on Normed Spaces

Definition 11.1. A map $T : X \rightarrow Y$ is called a linear operator if for all $x, y \in X$ and a scalar α , we have

$$\begin{aligned}T(x + y) &= T(x) + T(y) \\T(\alpha x) &= \alpha T(x)\end{aligned}$$

where, X is a vector space and Y lies in a vector space over the same field. The definition can be equivalently given as

$$T(\alpha x + y) = \alpha T(x) + T(y)$$

Definition 11.2. The Null space $\mathcal{N}(T)$ of T is the set of all $x \in X$ such that $T(x) = 0$. The Null space of T is also called the Kernel of T .

Now, taking $\alpha = 0$ in the definition of T , we get that

$$T(0) = 0$$

Definition 11.3. The range space $\mathcal{R}(T)$ is the set

$$\mathcal{R}(T) = \{T(x) \in Y : x \in X\}$$

From the form of a linear operator, we can say that T is a homomorphism of a vector space (its domain) into another vector space, that is, T preserves the two operations of vector space. This property makes linear operators important.

Example 11.4. Let X be a vector space.

1. The identity operator $I : X \rightarrow X$, defined as

$$I(x) = x$$

for all $x \in X$. This is a linear operator on X .

2. The zero operator $0 : X \rightarrow X$ defined as

$$0(x) = 0$$

for all $x \in X$ is a linear operator on X .

3. Let X be the space of all polynomial on $[a, b]$. Then the operator $T : X \rightarrow X$ defined as

$$T(x(t)) = x'(t)$$

for every polynomial $x(t) \in X$, is linear.

4. A linear operator T on $C[a, b]$ into itself can be defined as

$$T(x(t)) = \int_a^t x(\xi) d\xi$$

where $t \in [a, b]$. We can define another linear operator on $C[a, b]$ as

$$T'(x(t)) = tx(t)$$

The following are certain elementary results which are always applicable:

Theorem 11.5. Let T be a linear operator from X to Y .

1. The range $\mathcal{R}(T)$ is a vector space.
2. If $\dim X = n < \infty$, then $\dim \mathcal{R}(T) \leq n$.

3. The null space $\mathcal{N}(T)$ is a vector space.

Proof. 1. Let $y_1, y_2 \in \mathcal{R}(T)$. Then, there exists $x_1, x_2 \in X(T)$ such that $T(x_1) = y_1$ and $T(x_2) = y_2$. Then for any scalars α and β , we will always have

$$T(\alpha x_1 + x_2) = \alpha T(x_1) + T(x_2) = \alpha y_1 + y_2$$

This shows that $\alpha y_1 + y_2 \in \mathcal{R}(T)$ for some scalar α . Since y_1 and y_2 are arbitrary, so $\mathcal{R}(T)$ is a vector space (over the same field as that of X).

2. We choose $n + 1$ elements $y_1, y_2, \dots, y_{n+1}$ of $\mathcal{R}(T)$ in an arbitrary fashion. Then we will have $T(x_1) = y_1, T(x_2) = y_2, \dots, T(x_{n+1}) = y_{n+1}$ for some $x_1, x_2, \dots, x_{n+1} \in \mathcal{D}(T)$. Now, since $\dim X(T) = n$, the set $\{x_1, x_2, \dots, x_{n+1}\}$ is linearly dependent. Thus,

$$a_1 x_1 + a_2 x_2 + \dots + a_{n+1} x_{n+1} = 0$$

for some scalars $a_1, a_2, \dots, a_{n+1}$, not all zero. Since T is linear and $T(0) = 0$, application of T on both sides gives

$$T(a_1 x_1 + a_2 x_2 + \dots + a_{n+1} x_{n+1}) = a_1 y_1 + a_2 y_2 + \dots + a_{n+1} y_{n+1} = 0$$

This shows that $\{y_1, y_2, \dots, y_{n+1}\}$ is linearly dependent set because all of a_i 's are not zero. Since this subset was chosen in an arbitrary fashion, we conclude that $\mathcal{R}(T)$ has no linearly independent subsets of $n + 1$ or more elements. This means that $\dim \mathcal{R}(T) \leq n$.

3. We take any $x_1, x_2 \in \mathcal{N}(T)$. Then, $T(x_1) = T(x_2) = 0$. Since T is linear, for any scalar α , we have

$$T(\alpha x_1 + x_2) = \alpha T(x_1) + T(x_2) = 0$$

Hence $\alpha x_1 + x_2 \in \mathcal{N}(T)$. Thus, $\mathcal{N}(T)$ is a vector space. □

In fact, $\mathcal{N}(T)$ is a vector subspace of X .

In the process of proving part 2 of the above theorem, we have proved the following theorem

Theorem 11.6. Any linear operator on a vector space preserves linear dependence.

We will now turn to the inverse of linear operators. A mapping $: X \rightarrow Y$ is called injective if different points in the domain have different images, that is, if for any $x, y \in X$,

$$x \neq y \implies T(x) \neq T(y)$$

In such a case, there exists a mapping

$$T^{-1} : \mathcal{R}(T) \rightarrow X$$

as $y \mapsto x$, where $y = T(x)$. This map T^{-1} is called the inverse map for T .

By the definition of T^{-1} , we clearly have

$$\begin{aligned} T^{-1}T(x) &= x, \quad x \in X \\ TT^{-1}(y) &= y, \quad y \in \mathcal{R}(T) \end{aligned}$$

In case of the linear operators on vector spaces, we have the following result for the existence of inverse of linear operators.

Theorem 11.7. Let X and Y be vector spaces, both real or complex and let $T : X \rightarrow Y$ be a linear operator with range $\mathcal{R}(T) \subset Y$. Then

1. The inverse $T^{-1} : \mathcal{R}(T) \rightarrow X$ exists if and only if

$$T(x) = 0 \implies x = 0$$

2. If T^{-1} exists, it is a linear operator.
3. If $\dim X = n < \infty$ and T^{-1} exists, then

$$\dim \mathcal{R}(T) = \dim X$$

Proof. 1. Suppose that $T(x) = 0$ implies $x = 0$. Let $T(x_1) = T(x_2)$ for some $x_1, x_2 \in X$. Since T is linear,

$$T(x_1 - x_2) = T(x_1) - T(x_2) = 0,$$

so that by hypothesis, $x_1 - x_2 = 0$ which implies that $x_1 = x_2$. Hence T is injective and T^{-1} exists.

Conversely, if T^{-1} exists, then T is injective. Taking $x_2 = 0$, we have by the injectivity of T ,

$$T(x_1) = T(0) = 0 \implies x_1 = 0.$$

Hence 1 is proved.

2. We assume that T^{-1} exists and show that its linear. The domain of T^{-1} is $\mathcal{R}(T)$ and its a vector space. We consider $x_1, x_2 \in X$ and their images

$$y_1 = T(x_1), \text{ and } y_2 = T(x_2).$$

Then

$$x_1 = T^{-1}(y_1), \text{ and } x_2 = T^{-1}(y_2).$$

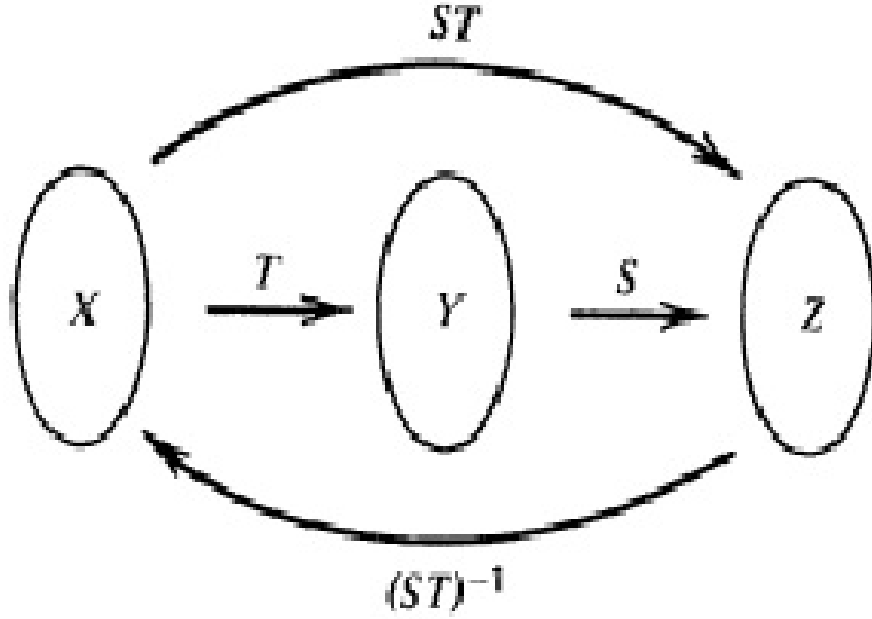
Since T is linear, for some scalar α , we have

$$\alpha y_1 + y_2 = \alpha T(x_1) + T(x_2) = T(\alpha x_1 + x_2).$$

Since $x_i = T^{-1}(y_i)$, we have by the above equation,

$$T^{-1}(\alpha y_1 + y_2) = \alpha T^{-1}(y_1) + T^{-1}(y_2).$$

which shows that T^{-1} is linear.



3. We have $\dim \mathcal{R}(T) \leq \dim X$. Applying the same theorem to T^{-1} , we get

$$\dim X \leq \dim \mathcal{R}(T).$$

Combining, we get the desired result. □

We will now state a very useful formula for the inverse of the composite of linear operators.

Lemma 11.8. Let $T : X \rightarrow Y$ and $S : Y \rightarrow Z$ be bijective linear operators, where X, Y, Z are vector spaces. Then the inverse $(ST)^{-1} : Z \rightarrow X$ of the product of ST exists, and

$$(ST)^{-1} = T^{-1}S^{-1}.$$

Proof. The operator $ST : X \rightarrow Z$ is bijective, so that $(ST)^{-1}$ exists. We thus have

$$ST(ST)^{-1} = I_Z$$

where I_Z is the identity operator on Z . Applying S^{-1} and using $S^{-1}S = I_Y$ (the identity operator on Y), we obtain

$$S^{-1}ST(ST)^{-1} = T(ST)^{-1} = S^{-1}I_Z = S^{-1}.$$

Applying T^{-1} and using $T^{-1}T = I_X$, we obtain

$$T^{-1}T(ST)^{-1} = (ST)^{-1} = T^{-1}S^{-1}.$$

Hence the result. □

11.1 Continuous functions

We are familiar with the notion of continuous functions on any arbitrary metric space. Since we are going to need the idea of continuity in the following unit, let us define continuity explicitly on a metric space.

Definition 11.9. Let (X, d_1) and (Y, d_2) be two metric spaces and let $f : X \rightarrow Y$ be a function. f is said to be continuous on X , if for all $a \in X$ and $\epsilon > 0$, there exists a $\delta(\epsilon, a) > 0$ such that

$$d_2(f(x), f(a)) < \epsilon, \text{ whenever } d_1(x, a) < \delta.$$

This idea can also be extended to any normed linear spaces as follows.

Definition 11.10. Let $T : X \rightarrow Y$ be any operator, not necessarily linear, where X and Y are normed spaces. Then T is said to be continuous at a point x_0 in X if for every $\epsilon > 0$ there is a $\delta > 0$ such that

$$\|T(x) - T(x_0)\| < \epsilon \text{ for all } x \in X \text{ such that } \|x - x_0\| < \delta.$$

T is continuous if T is continuous at every point of X .

We will deal with these in detail in the next unit.

- Exercise 11.11.**
1. If the product (the composite) of two linear operators exists, show that its linear.
 2. Let $T : X \rightarrow Y$ be a linear operator whose inverse exists. If $\{x_1, x_2, \dots, x_n\}$ is linearly independent in X then show that the set $\{T(x_1), T(x_2), \dots, T(x_n)\}$ is linearly independent in Y .

Unit XII

Course Structure

- Introduction
 - Objectives
 - Bounded Linear operators
 - Continuity
-

12 Introduction

In this section we will introduce the concept of bounded and continuous linear operators on a normed linear space (NLS). We will straightaway start with the definition of bounded linear operators and then move to its relationship with continuity.

Objectives

After reading this unit, you will be able to

- define bounded linear operators
- learn about their various characterizations
- learn various examples of bounded linear operators
- learn about boundedness in finite dimensional normed spaces
- learn the relationship between continuity and boundedness

12.1 Bounded Linear Operators

Definition 12.1. Let $(X, ||\cdot||_1)$ and $(Y, ||\cdot||_2)$ be two normed spaces and $T : X \rightarrow Y$ be a linear operator. Then T is said to be bounded if there exists a real number c such that for all $x \in X$, we have

$$||T(x)||_1 \leq c||x||_2$$

For simplicity we will henceforth denote both the norms by the same symbol $||\cdot||$. The definition of bounded linear operators show that such functions map bounded sets in X onto bounded sets in Y .

A valid question is what should be the minimum value of c such that the inequality

$$||T(x)|| \leq c||x||$$

is satisfied, or equivalently for $||x|| \neq 0$,

$$\frac{||T(x)||}{||x||} \leq c$$

is satisfied for all non-zero $x \in X$. Of course the value of c has to be at least as big as the supremum of the expression on the left hand side of the above equation. This supremum is termed as the 'norm' of the operator T , denoted by $||T||$. Thus

$$||T|| = \sup_{x \in X, x \neq 0} \frac{||T(x)||}{||x||}.$$

If $X = \{0\}$, then we define $||T|| = 0$.

Thus, taking $c = ||T||$, we get

$$||T(x)|| \leq ||T||||x||.$$

Lemma 12.2. Let T be a bounded linear operator. Then:

1. An alternative formula for the norm of T is

$$||T|| = \sup_{x \in X, ||x||=1} ||T(x)||.$$

2. The definition of $||T||$ given previously, satisfies all the properties of the norm.

Proof. 1. We write $||x|| = a$ and set $y = (1/a)x$, where $x \neq 0$. Then $||y|| = ||x||/a = 1$, and since T is linear, so

$$||T|| = \sup_{x \in X, x \neq 0} \frac{1}{a} ||T(x)|| = \sup_{x \in X, x \neq 0} \left\| T \left(\frac{1}{a} x \right) \right\| = \sup_{y \in X, ||y||=1} ||T(y)||$$

Since y is dummy, writing x for y , we get the desired result.

2. Left as exercise. □

Example 12.3. 1. The identity operator $I : X \rightarrow X$ on a NLS $X \neq \{0\}$ is bounded and has norm $||I|| = 1$.

2. The zero operator $0 : X \rightarrow Y$ is bounded and has norm $||0|| = 0$.

3. Let X be the NLS of all polynomials on $J = [0, 1]$ with norm given $\|x\| = \max |x(t)|$, $t \in J$. Let us define a linear operator T on X as

$$T(x(t)) = x'(t).$$

The prime denotes the differentiation with respect to t . This operator is linear but not bounded. Let $x_n(t) = t^n$, where $n \in \mathbb{N}$. Then $\|x_n\| = 1$ and

$$T(x_n(t)) = x'_n(t) = nt^{n-1}$$

so that $\|T(x_n)\| = n$ and $\|T(x_n)\|/\|x_n\| = n$. Since $n \in \mathbb{N}$ is arbitrary, this shows that there is no fixed number c such that $\|T(x_n)\|/\|x_n\| \leq c$. Hence T is not bounded.

Lemma 12.4. Let $\{e_1, e_2, \dots, e_n\}$ be a linearly independent set of vectors in a normed space X (of any dimension). Then there is a number $c > 0$ such that for every choice of scalars $\xi_1, \xi_2, \dots, \xi_n$, we have

$$\|\xi_1 e_1 + \xi_2 e_2 + \dots + \xi_n e_n\| \geq c(|\xi_1| + |\xi_2| + \dots + |\xi_n|).$$

We will use the above lemma to prove the following theorem.

Theorem 12.5. If a normed space X is finite dimensional, then every linear operator on X is bounded.

Proof. Let $\dim X = n$ and $\{e_1, e_2, \dots, e_n\}$ a basis for X . Let $x \in X$. Then

$$x = \sum_{i=1}^n \xi_i e_i$$

for some unique scalars ξ_i . Consider any linear operator T on X . Since T is linear

$$\|T(x)\| = \left\| \sum_{i=1}^n \xi_i e_i \right\| \leq \sum_{i=1}^n |\xi_i| \|T(e_i)\| \leq \max_i \|T(e_i)\| \sum_{i=1}^n |\xi_i|$$

Then by the previous lemma, we get a $c > 0$ such that

$$\sum_{i=1}^n |\xi_i| \leq \frac{1}{c} \left\| \sum_{i=1}^n \xi_i e_i \right\| = \frac{1}{c} \|x\|.$$

Together,

$$\|T(x)\| \leq \gamma \|x\|, \text{ where } \gamma = \frac{1}{c} \max_i \|T(e_i)\|.$$

Hence T is bounded. □

We will now relate continuity to operators. Operators are mappings, so that the definition of continuity applies to them. It is a fundamental fact that for a linear operator, continuity and boundedness become equivalent concepts. This is given by the following theorem.

Theorem 12.6. Let $T : X \rightarrow Y$ be a linear operator where X and Y are NLS. Then

1. T is continuous if and only if T is bounded.
2. If T is continuous at a single point, it is continuous.

Proof. 1. For $T = 0$ the result is trivial. Let $T \neq 0$. We assume T to be bounded and consider $x_0 \in X$. Let $\epsilon > 0$ be given. Then, since T is linear, for every $x_0 \in X$ such that $\|x - x_0\| < \delta$, where $\delta = \frac{\epsilon}{\|T\|}$, we obtain

$$\|T(x) - T(x_0)\| = \|T(x - x_0)\| \leq \|T\| \|x - x_0\| < \|T\| \delta = \epsilon.$$

Since $x_0 \in X$ was arbitrary, this shows that T is continuous.

Conversely, assume that T is continuous at an arbitrary $x_0 \in X$. Then, given $\epsilon > 0$, there is a $\delta > 0$ such that

$$\|T(x) - T(x_0)\| \leq \epsilon \text{ for all } x \in X \text{ such that } \|x - x_0\| \leq \delta.$$

We now take any $y \neq 0$ in X and set

$$x = x_0 + \frac{\delta}{\|y\|} y.$$

Then

$$x - x_0 = \frac{\delta}{\|y\|} y.$$

Hence $\|x - x_0\| = \delta$, so that by continuity of T and since T is linear, we have

$$\epsilon \geq \|T(x) - T(x_0)\| = \|T(x - x_0)\| = \left\| T \left(\frac{\delta}{\|y\|} y \right) \right\| = \frac{\delta}{\|y\|} \|T(y)\|$$

Thus,

$$\|T(y)\| \leq \frac{\epsilon}{\delta} \|y\|.$$

This can be written as

$$\|T(y)\| \leq c \|y\|$$

where $c = \epsilon/\delta$, and shows that T is bounded.

2. Continuity of T at a point implies boundedness of T by the second part of the proof of 1, which in turn implies continuity of T by 1.

□

Theorem 12.7. Let T be a bounded linear operator. Then:

1. $x_n \rightarrow x$ when $x_n, x \in X$ implies $T(x_n) \rightarrow T(x)$.
2. The null space $\mathcal{N}(T)$ is closed.

Proof. 1. The result follows from the first part of the previous theorem and the sequential criterion of continuity.

2. For every $x \in \overline{\mathcal{N}(T)}$ there is a sequence $\{x_n\}$ in $\mathcal{N}(T)$ such that $x_n \rightarrow x$. Hence $T(x_n) \rightarrow T(x)$ by part 1 of this theorem. Also $T(x) = 0$ since $T(x_n) = 0$ so $x \in \mathcal{N}(T)$. Since $x \in \overline{\mathcal{N}(T)}$ was arbitrary, $\mathcal{N}(T)$ is closed. $\square$

It is worth noting that the range of a bounded linear operator may not be closed. We have another useful formula given in the form of the following theorem.

Theorem 12.8. Let $T_2 : X \rightarrow Y$ and $T_1 : Y \rightarrow Z$ and $T : X \rightarrow X$ be bounded linear operators, where X, Y, Z are NLS. Then

1. $\|T_1 T_2\| \leq \|T_1\| \|T_2\|$, and
2. $\|T^n\| \leq \|T\|^n$, for $n \in \mathbb{N}$.

Proof. The proof is easy and has been left as an exercise. $\square$

We now give some useful definitions in series.

Definition 12.9. Two operators T_1 and T_2 are said to be equal, written as

$$T_1 = T_2$$

if they have the same domain and if $T_1(x) = T_2(x)$ for all $x \in X$, which is the common domain for both.

Definition 12.10. Let $T : X \rightarrow Y$ be an operator and let $B \subset X$. Then the restriction of T to B

$$T|_B : B \rightarrow Y$$

is defined as $T|_B(x) = T(x)$ for $x \in B$.

An extension of T to $Z \supset X$ is an operator

$$\tilde{T} : Z \rightarrow Y$$

such that $\tilde{T}|_X = T$, that is $\tilde{T}(x) = T(x)$ for all $x \in X$.

If X is a proper subset of Z , then T can have many extensions. But we are more interested to study those operators which preserve certain basic properties, for example, linearity or boundedness. The following theorem is an important tool to extend an operator T to the closure of X such that the extended operator is again bounded and linear, and even has the same norm.

Theorem 12.11. Let $T : X \rightarrow Y$ be a bounded linear operator, where X lies in a normed space Z and Y is a Banach space. Then T has an extension

$$\tilde{T} : \overline{X} \rightarrow Y$$

where $\tilde{T}$ is a bounded linear operator of norm $\|\tilde{T}\| = \|T\|$.

Proof. We consider any $x \in \overline{X}$. Then there is a sequence $\{x_n\}$ in X such that $x_n \rightarrow x$. Since T is linear and bounded, we have

$$\|T(x_n) - T(x_m)\| = \|T(x_n - x_m)\| \leq \|T\| \|x_n - x_m\|.$$

Since $\{x_n\}$ is convergent, so $\{T(x_n)\}$ is Cauchy. Since Y is complete, so $\{T(x_n)\}$ converges to some $y \in Y$, say. We define $\tilde{T}$ as

$$\tilde{T}(x) = y.$$

We show that this definition is independent of the particular choice of a sequence in X converging to x . Suppose that $x_n \rightarrow x$ and $z_n \rightarrow x$. Then $v_m \rightarrow x$ where, $\{v_m\}$ is the sequence $\{x_1, z_1, x_2, z_2, \dots\}$. Hence $\{T(v_m)\}$ converges and the two subsequences $\{T(x_n)\}$ and $\{T(z_n)\}$ must have the same limit. This proves that $\tilde{T}$ is uniquely defined at every point $x \in \overline{X}$.

Clearly $\tilde{T}$ is linear and $\tilde{T}(x) = T(x)$ for every $x \in X$, so that $\tilde{T}$ is an extension of T . We now use

$$\|T(x_n)\| \leq \|T\| \|x_n\|$$

and let $n \rightarrow \infty$. Then $T(x_n) \rightarrow y = \tilde{T}(x)$. Since norm is a continuous function, we thus obtain

$$\|\tilde{T}(x)\| \leq \|T\| \|x\|.$$

Hence $\tilde{T}$ is bounded and $\|\tilde{T}\| \leq \|T\|$. Obviously $\|\tilde{T}\| \geq \|T\|$ because the norm, being defined by a supremum, cannot decrease in an extension. Combining, we have

$$\|\tilde{T}\| = \|T\|.$$

Hence proved. □

Exercise 12.12. 1. Let $T : X \rightarrow Y$ be a linear operator, where X and Y are normed spaces. Show that T is bounded if and only if T maps bounded sets in X into bounded sets in Y .

2. If $T \neq 0$ is a bounded linear operator, show that for any $x \in X$ such that $\|x\| < 1$, we have $\|T(x)\| < \|T\|$.

3. Let T be a bounded linear operator from a normed space X onto a normed space Y . If there is a positive b such that

$$\|T(x)\| \geq b\|x\|,$$

for all $x \in X$, then show that $T^{-1} : Y \rightarrow X$ exists and is bounded.

Unit XIII

Course Structure

- Introduction
 - Objectives
 - Linear functionals
 - Hahn-Banach theorem, simple applications.
 - Normed conjugate space and separability of the space
 - Uniform boundedness principle, simple application.
-

13 Introduction

Functional analysis was initially the analysis of functionals. A functional is an operator whose range lies on the real line $\mathbb{R}$ or in the complex plane $\mathbb{C}$. We introduce the concept of functionals in this unit and thereby will see its various properties.

Objectives

After reading this unit, you will be able to

- define linear and bounded linear functionals
- learn certain properties of them
- see some examples of them
- learn the uniform boundedness theorem
- learn its applications

13.1 Linear Functionals

Definition 13.1. A linear functional f is a linear operator with domain in a vector space X and range in the scalar field K of X , thus

$$f : X \rightarrow K,$$

where $K = \mathbb{R}$ if X is real and $K = \mathbb{C}$ if X is complex.

Definition 13.2. A bounded linear functional f is a bounded linear operator with range in the scalar field of the normed space X in which the domain X lies. Thus there exists a real number c such that for all $x \in X$,

$$|f(x)| \leq c\|x\|.$$

Furthermore, the norm of f is

$$\|f\| = \sup_{x \in X, x \neq 0} \frac{|f(x)|}{\|x\|}$$

or,

$$\|f\| = \sup_{x \in X, \|x\|=1} |f(x)|.$$

Thus as in the case of bounded linear operators, we also have in this case

$$|f(x)| \leq \|f\|\|x\|,$$

Also as in the case of bounded linear operators we have the following theorem

Theorem 13.3. A linear functional f with domain X is continuous if and only if f is bounded.

Example 13.4. 1. The norm $\|\cdot\| : X \rightarrow \mathbb{R}$ on a NLS $(X, \|\cdot\|)$ is a functional on X which is not linear.

2. The familiar dot product with one factor kept fixed defines a functional $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ by means of

$$f(x) = x \cdot a = \xi_1 a_1 + \xi_2 a_2 + \xi_3 a_3,$$

where $a = (a_1, a_2, a_3) \in \mathbb{R}^3$ is fixed. f is linear and bounded. In fact,

$$|f(x)| = |x \cdot a| \leq \|x\| \|a\|,$$

so that $\|f\| \leq \|a\|$ follows immediately from the previous literature if we take the supremum over all x of norm one. On the other hand, by taking $x = a$ we obtain

$$\|f\| \geq \frac{|f(a)|}{\|a\|} = \frac{\|a\|^2}{\|a\|} = \|a\|.$$

Hence the norm of f is $\|f\| = \|a\|$.

3. On the space of continuous functions $C[a, b]$, choose a fixed $t_0 \in J = [a, b]$ and set

$$f_1(x) = x(t_0), \quad x \in C[a, b].$$

f_1 is linear. f_1 is bounded and has norm $\|f_1\| = 1$. In fact, we have

$$|f_1(x)| = |x(t_0)| \leq \|x\|,$$

and this implies that $\|f_1\| \leq 1$. On the other hand, for $x_0 = 1$ we have $\|x_0\| = 1$ and obtain

$$\|f_1\| \geq |f_1(x_0)| = 1.$$

It is of basic importance that the set of all linear functionals defined on a vector space X can itself be made into a vector space. This space is denoted by X^* and is called the algebraic dual space of X . Its algebraic operations of vector space are defined in a natural way as follows. The sum $f_1 + f_2$ of two functionals f_1 and f_2 is the functional s whose value at every $x \in X$ is

$$s(x) = (f_1 + f_2)(x) = f_1(x) + f_2(x);$$

and the product αf of a scalar α and a functional f is the functional p whose value at $x \in X$ is

$$p(x) = (\alpha f)(x) = \alpha f(x).$$

This agrees with the usual way of adding functions and multiplying them by constants.

We may go a step further and consider the algebraic dual $(X^*)^*$ of X^* , whose elements are the linear functionals defined on X^* . We denote $(X^*)^*$ by X^{**} and call it the second algebraic dual space of X .

We can obtain a $g \in X^{**}$, which is a linear functional defined on X^* , by choosing a fixed $x \in X$ and setting

$$g(f) = g_x(f) = f(x), \quad x \in X \text{ fixed, } f \in X^* \text{ variable.}$$

g_x as defined above, is linear since

$$\begin{aligned} g_x(\alpha f_1 + \beta f_2) &= (\alpha f_1 + \beta f_2)(x) \\ &= \alpha f_1(x) + \beta f_2(x) \\ &= \alpha g_x(f_1) + \beta g_x(f_2). \end{aligned}$$

Hence g_x is an element of X^{**} , by the definition of X^{**} .

To each $x \in X$, there corresponds a $g_x \in X^{**}$. This defines a mapping

$$C : X \rightarrow X^{**}, \quad x \mapsto g_x.$$

C is called the canonical mapping of X into X^{**} .

C is linear since its domain is a vector space and we have

$$\begin{aligned} (C(\alpha x + \beta y))(f) &= g_{\alpha x + \beta y}(f) \\ &= f(\alpha x + \beta y) \\ &= \alpha f(x) + \beta f(y) \\ &= \alpha g_x(f) + \beta g_y(f) \\ &= \alpha(Cx)(f) + \beta(Cy)(f). \end{aligned}$$

There are various other properties related to the dual spaces which is omitted in this section. We will next move on to the Hahn-Banach Theorem.

13.2 Hahn-Banach Theorem

The Hahn-Banach theorem is an extension theorem for linear functionals. This theorem guarantees that a normed space is richly supplied with bounded linear functionals and makes possible an adequate theory of dual spaces, which is an essential part of the general theory of normed spaces. The theorem was given H. Hahn (1927), rediscovered in its present more general form by S. Banach (1929) and generalized to complex vector spaces by H. F. Bohnenblust and A. Sobczyk (1938). Before starting with the theorem, let us first see a few preliminary results which are required to prove it.

Theorem 13.5. (Zorn's Lemma) Let $M \neq \emptyset$ be a partially ordered set. Suppose that every chain $C \subset M$ has an upper bound. Then, M has at least one maximal element.

In the Hahn-Banach theorem, the object to be extended is a linear functional f which is defined on a subspace Z of a vector space X and has a certain boundedness property which will be formulated in terms of a sublinear functional. By definition, this is a real-valued functional p on a vector space X which is subadditive, that is,

$$p(x + y) \leq p(x) + p(y), \quad \forall x, y \in X,$$

and positive-homogeneous, that is,

$$p(\alpha x) = \alpha p(x), \quad \forall \alpha \geq 0 \text{ in } \mathbb{R} \text{ and } x \in X.$$

Theorem 13.6. (Hahn-Banach Theorem) Let X be a real vector space and p a sub linear functional on X . Furthermore, let f be a linear functional which is defined on a subspace Z of X and satisfies

$$f(x) \leq p(x) \quad \forall x \in Z.$$

Then f has a linear extension $\tilde{f}$ from Z to X satisfying

$$\tilde{f}(x) \leq p(x) \quad \forall x \in X,$$

that is, $\tilde{f}$ is a linear functional on X , satisfies the above theorem and

$$\tilde{f}(x) = f(x), \quad \forall x \in Z.$$

Proof. Let E be the set of all linear extensions g of f satisfying

$$g(x) \leq p(x)$$

on their respective domains. Clearly, $E \neq \emptyset$ since $f \in E$. On E , we can define a partial ordering by

$$g \leq h \implies h \text{ is an extension of } g,$$

that is, by definition, $\mathcal{D}(h) \supset \mathcal{D}(g)$, where, $\mathcal{D}(g)$ and $\mathcal{D}(h)$ denote the domains of g and h respectively, and $h(x) = g(x)$ for all $x \in \mathcal{D}(g)$. For any chain $C \subset E$, we define $\hat{g}$ as

$$\hat{g}(x) = g(x) \text{ if } x \in \mathcal{D}(g)$$

for $g \in C$. $\hat{g}$ is a linear functional, the domain being

$$\mathcal{D}(\hat{g}) = \bigcup_{g \in C} \mathcal{D}(g),$$

which is a vector space since C is a chain. The definition of $\hat{g}$ is unambiguous. Since for $x \in \mathcal{D}(g_1) \cap \mathcal{D}(g_2)$ with $g_1, g_2 \in C$, we have $g_1(x) = g_2(x)$ since C is a chain, so that

$$g_1 \leq g_2 \text{ or } g_2 \leq g_1.$$

Clearly, $g \leq \hat{g}$ for all $g \in C$. Hence $\hat{g}$ is an upper bound of C . Since $C \subset E$ was arbitrary, Zorn's lemma thus implies that E has a maximal element $\tilde{f}$. By the definition of E , this is a linear extension of f which satisfies

$$\tilde{f}(x) \leq p(x), \quad x \in \mathcal{D}(\tilde{f}).$$

This representation is unique. In fact, $y + \alpha y_1 = \tilde{y} + \beta y_1$ with $\tilde{y} \in \mathcal{D}(\tilde{f})$ implies $y - \tilde{y} = (\beta - \alpha)y_1$, where, $y - \tilde{y} \in \mathcal{D}(\tilde{f})$ whereas, $y_1 \notin \mathcal{D}(\tilde{f})$, so that, the only solution is $y - \tilde{y} = 0$ and $\beta - \alpha = 0$.

A functional g_1 on Y_1 is defined by

$$g_1(y + \alpha y_1) = \tilde{f}(y) + \alpha c \quad (13.1)$$

where c is any real constant. g_1 is linear. Moreover, for $\alpha = 0$, we have, $g_1(y) = \tilde{f}(y)$. Hence g_1 is a proper extension of $\tilde{f}$, that is, an extension such that $\mathcal{D}(\tilde{f})$ is a proper subset of $\mathcal{D}(g_1)$. So, we can prove that, $g_1 \in E$ by showing that

$$g_1(x) \leq p(x), \quad x \in \mathcal{D}(g_1), \quad (13.2)$$

this will contradict the maximality of $\tilde{f}$, so that $\mathcal{D}(\tilde{f}) \neq X$ is false and so, $\mathcal{D}(\tilde{f}) = X$.

We will now show that g_1 , with a suitable c , satisfies equation (13.2). Let us consider $y, z \in \mathcal{D}(\tilde{f})$. By the properties of $\tilde{f}$ and p , we get

$$\begin{aligned} \tilde{f}(y) - \tilde{f}(z) &= \tilde{f}(y - z) \\ &\leq p(y - z) \\ &= p(y + y_1 - y_1 - z) \\ &\leq p(y + y_1) + p(-y_1 - z) \end{aligned}$$

Rearranging, we get

$$-p(-y_1 - z) - \tilde{f}(z) \leq p(y + y_1) - \tilde{f}(y)$$

where y_1 is fixed. Since y does not appear on the left and z not on the right, the inequality continues to hold if we take the supremum over $z \in \mathcal{D}(\tilde{f})$ on the left (call it m_0) and the infimum over $y \in \mathcal{D}(\tilde{f})$ on the right, and call it m_1 . Then, $m_0 \leq m_1$ and for a c with $m_0 \leq c \leq m_1$ we have, from the above equation

$$-p(-y_1 - z) - \tilde{f}(z) \leq c, \quad \forall z \in \mathcal{D}(\tilde{f}) \quad (13.3)$$

and

$$c \leq p(y + y_1) - \tilde{f}(y), \quad \forall y \in \mathcal{D}(\tilde{f}) \quad (13.4)$$

For negative α in equation (13.1), we use equation (13.3) with $\alpha^{-1}y$ in place of z , and get

$$-p\left(-y_1 - \frac{1}{\alpha}y\right) - \tilde{f}\left(\frac{1}{\alpha}y\right) \leq c, \quad \forall z \in \mathcal{D}(\tilde{f}).$$

Multiplying by $-\alpha > 0$ gives

$$\alpha p\left(-y_1 - \frac{1}{\alpha}y\right) + \tilde{f}(y) \leq -\alpha c.$$

From this and (13.1), using $y + \alpha y_1 = x$, we obtain

$$\begin{aligned} g_1(x) &= \tilde{f}(y) + \alpha c \\ &\leq -\alpha p\left(-y_1 - \frac{1}{\alpha}y\right) \\ &= p(\alpha y_1 + y) \\ &= p(x). \end{aligned}$$

For $\alpha = 0$, we have $x \in \mathcal{D}(\tilde{f})$ and we have nothing to prove. For $\alpha > 0$ we use equation (13.4) with y replaced by $\alpha^{-1}y$ to get

$$c \leq p\left(\frac{1}{\alpha}y + y_1\right) - \tilde{f}\left(\frac{1}{\alpha}y\right).$$

Multiplying by $\alpha > 0$ gives

$$\alpha c \leq \alpha p\left(\frac{1}{\alpha}y + y_1\right) - \tilde{f}(y) = p(x) - \tilde{f}(y).$$

From this and (13.1), we get

$$g_1(x) = \tilde{f}(y) + \alpha c \leq p(x).$$

Hence the theorem. □

This was Hahn-Banach theorem in a real vector space. We will now state the theorem for any a NLS.

Theorem 13.7. Let f be a bounded linear functional on a subspace Z of a normed space X . Then there exists a bounded linear functional $\tilde{f}$ on X which is an extension of f to X and has the same norm,

$$\|\tilde{f}\|_X = \|f\|_Z,$$

where,

$$\|\tilde{f}\|_X = \sup_{x \in X, \|x\|=1} |\tilde{f}(x)|, \quad \|f\|_Z = \sup_{x \in Z, \|x\|=1} |f(x)|.$$

Theorem 13.8. Let X be a normed space and let $x_0 \neq 0$ be any element of X . Then there exists a bounded linear functional $\tilde{f}$ on X such that

$$\|\tilde{f}\| = 1, \quad \tilde{f}(x_0) = \|x_0\|.$$

Proof. We consider the subspace Z of X consisting of all elements $x = \alpha x_0$, where α is a scalar. On Z , we define a linear functional f by

$$f(x) = f(\alpha x_0) = \alpha \|x_0\|. \quad (13.5)$$

f is bounded and has norm $\|f\| = 1$ because

$$|f(x)| = |f(\alpha x_0)| = |\alpha| \|x_0\| = \|\alpha x_0\| = \|x\|.$$

The previous theorem implies that f has a linear extension $\tilde{f}$ from Z to X , of norm $\|\tilde{f}\| = \|f\| = 1$. From equation (13.5), we see that, $\tilde{f}(x_0) = \|x_0\|$. $\square$

Corollary 13.8.1. For every x in a normed space X , we have

$$\|x\| = \sup_{f \in X^*, f \neq 0} \frac{|f(x)|}{\|f\|}.$$

Hence, if x_0 is such that $f(x_0) = 0$ for all $f \in X^*$, then $x_0 = 0$.

Proof. From the previous theorem, we have, writing x for x_0 ,

$$\sup_{f \in X^*, f \neq 0} \frac{|f(x)|}{\|f\|} \geq \frac{|\tilde{f}(x)|}{\|f\|} = \frac{\|x\|}{1} = \|x\|,$$

and from $|f(x)| \leq \|f\| \|x\|$, we obtain,

$$\sup_{f \in X^*, f \neq 0} \frac{|f(x)|}{\|f\|} \leq \|x\|.$$

Combining, we get the desired result. $\square$

13.3 Uniform Boundedness Theorem

The uniform boundedness theorem is one of the corner stones of functional analysis in normed spaces, which characterize some of the most important properties of Banach spaces which normed spaces in general may not have. We are already aware of the terms nowhere dense sets, first category sets and second category sets in a metric space. Let us state the Baire Category theorem and then we will move on to the deeper results.

Theorem 13.9. (Baire Category Theorem) If a metric space $X \neq \emptyset$ is complete, then it is a second category set in itself. Hence, if $X \neq \emptyset$ is complete, and

$$X = \bigcup_{k=1}^{\infty} A_k,$$

where A_k are closed sets, then at least one A_k contains a non-empty open subset.

We will now obtain the uniform boundedness theorem.

Theorem 13.10. (Uniform Boundedness Theorem) Let $\{T_n\}$ be a sequence of bounded linear operators $T_n : X \rightarrow Y$ from a Banach space X into a normed space Y such that $\{\|T_n(x)\|\}$ is bounded for every $x \in X$, say

$$\|T_n(x)\| \leq c_x, \quad n = 1, 2, \dots,$$

where c_x is a real number. Then the sequence of the norms $\|T_n\|$ is bounded, that is, there is a c such that

$$\|T_n(x)\| \leq c, \quad n = 1, 2, \dots$$

Proof. For every $k \in \mathbb{N}$, let $A_k \subset X$ be the set of all x such that

$$\|T_n(x)\| \leq k, \quad n = 1, 2, \dots$$

Then A_k is closed. Indeed, for any $x \in \overline{A_k}$, there is a sequence $\{x_j\}$ in A_k converging to x . This means that for every fixed n we have

$$\|T_n(x_j)\| \leq k,$$

and obtain $\|T_n(x)\| \leq k$ since T_n is continuous and so is the norm. Hence $x \in A_k$, and so A_k is closed. Since

$$\|T_n(x)\| \leq c_x, \quad n = 1, 2, \dots,$$

so each $x \in X$ belongs to some A_k . Hence

$$X = \bigcup_{k=1}^{\infty} A_k.$$

Since X is complete, Baire's theorem implies that some A_k contains an open ball, say

$$B_0 = B(x_0, r) \subset A_{k_0}.$$

Let $x \in X$ be non-zero arbitrary element. We set

$$z = x_0 + \gamma x, \quad \gamma = \frac{r}{2\|x\|}.$$

Then $\|z - x_0\| < r$, so that $z \in B_0$. We thus have $\|T_n(z)\| \leq k_0$ for all n . Also, $\|T_n(x_0)\| \leq k_0$ since $x_0 \in B_0$. We also have,

$$x = \frac{1}{\gamma}(z - x_0).$$

This yields for all n ,

$$\begin{aligned} \|T_n(x)\| &= \frac{1}{\gamma} \|T_n(z - x_0)\| \\ &\leq \frac{1}{\gamma} (\|T_n(z)\| + \|T_n(x_0)\|) \\ &\leq \frac{4}{r} \|x\| k_0. \end{aligned}$$

Hence for all n ,

$$\|T_n\| = \sup_{\|x\|=1} \|T_n(x)\| \leq \frac{4}{r} k_0.$$

Taking $c = \frac{4}{r} k_0$, we get the desired result. □

Let us see certain applications of the theorem.

Theorem 13.11. The normed space X of all polynomials with norm defined by

$$||x|| = \max_j |a_j|$$

where $a_1, a_1, \dots$ are the coefficients of x , is not complete.

Proof. We construct a sequence of bounded linear operators on X , that satisfies the conditions given in the uniform boundedness theorem, which does not satisfy its conclusions, so that way, X can't be complete. We may write a polynomial $x \neq 0$ of degree N_x in the form

$$x(t) = \sum_{j=0}^{\infty} a_j t^j,$$

where, $a_j = 0$ for $j > N_x$. As a sequence of operators on X we take the sequence of functionals $T_n = f_n$ defined by

$$T_n(0) = f_n(0) = 0, \quad T_n(x) = f_n(x) = a_0 + a_1 + \dots + a_{n-1}.$$

f_n is linear. f_n is bounded since $|a_j| \leq ||x||$, so that $|f_n(x)| \leq n||x||$. Furthermore, for each fixed $x \in X$, the sequence $\{|f_n(x)|\}$ satisfies the inequality

$$|f_n(x)| \leq (N_x + 1) \max_j |a_j| = c_x$$

since a polynomial of degree N_x has $N_x + 1$ coefficients. We will now show that there is no c such that $||T_n|| = ||f_n|| \leq c$ for all n . Let us take

$$x(t) = 1 + t + t^2 + \dots + t^n.$$

Then $||x|| = 1$ and

$$f_n(x) = 1 + 1 + \dots + 1 = n = n||x||.$$

Hence,

$$||f_n|| \geq |f_n(x)|/||x|| = n,$$

so that $\{||f_n||\}$ is unbounded. Thus, by Uniform Boundedness theorem, the space X is not complete. $\square$

- Exercise 13.12.**
1. Show that a norm on a vector space X is a sub linear functional on X .
 2. Show that a sublinear functional p satisfies $p(0) = 0$ and $p(-x) \geq -p(x)$.
 3. If p is a sub linear functional on a vector space X , show that $M = \{x \mid p(x) \leq \gamma, \gamma > 0 \text{ fixed}\}$ is a convex set.
 4. If p and q are sublinear functionals on a vector space X and a and b are positive constants, then show that $ap + bq$ is sunlinear in X .
 5. Find all the nowhere (nw) dense sets in a discrete metric space.
 6. Show that a subset M of a metric space X is nw dense X if and only if $X \setminus \overline{M}$ is dense in X .
 7. Let X be a Banach space and Y a normed space and T_n are bounded linear operators from X to Y . Also, if $\{T_n(x)\}$ is Cauchy in Y for every $x \in X$, then show that $\{\|T_n\|\}$ is bounded.

Unit XIV

Course Structure

- Introduction
 - Objectives
 - Inner product space.
 - Some preliminary results
 - Bessel's inequality and its generalisation
-

Introduction

In a normed space we can add vectors and multiply vectors by scalars, just as in elementary vector algebra. Furthermore, the norm on such a space generalizes the elementary concept of the length of a vector. However, what is still missing in a general normed space, and what we would like to have if possible, is an analogue of the familiar dot product

$$a.b = a_1b_1 + a_2b_2 + a_3b_3$$

and resulting formula like

$$|a| = \sqrt{a.a}$$

Here, in this unit, we will generalise these and further ideas to arbitrary vector spaces leading further to inner product spaces and few preliminary properties related to these.

Objectives

After reading this unit, you will be able to

- define an inner product space
- learn certain examples of inner product spaces
- learn certain preliminary definitions and inequalities concerning inner product spaces
- apply them in various appropriate situations
- deduce important formulae and generalisations
- learn Bessel's inequality and its generalisations

14 Inner Product Spaces

An inner product space is a vector space X with an inner product defined on it. Let us formally define an inner product on a vector space X .

Definition 14.1. Let X be a vector space over a field, say F . Then any function

$$\langle, \rangle : X \times X \rightarrow F$$

is said to be an inner product space if it satisfies the following conditions:

$$(IP1) \quad \langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle.$$

$$(IP2) \quad \langle \alpha x, y \rangle = \alpha \langle x, y \rangle.$$

$$(IP3) \quad \langle x, y \rangle = \overline{\langle y, x \rangle}.$$

$$(IP4) \quad \langle x, x \rangle \geq 0 \text{ and } \langle x, x \rangle = 0 \text{ if and only if } x = 0.$$

for all $x, y \in X$ and $\alpha \in F$.

If X is a real vector space, then (IP3) simply is

$$\langle x, y \rangle = \langle y, x \rangle.$$

The above conditions from (IP1)-(IP3) can be equivalently written as

$$1. \quad \langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle.$$

$$2. \quad \langle x, \alpha y \rangle = \overline{\alpha} \langle x, y \rangle.$$

$$3. \quad \langle x, \alpha y + \beta z \rangle = \overline{\alpha} \langle x, y \rangle + \overline{\beta} \langle x, z \rangle.$$

for all $x, y, z \in X$ and $\alpha, \beta \in F$.

One can easily verify that our most familiar dot product, defined at the beginning of this unit, on the space $\mathbb{R}^3$, over the field $\mathbb{R}$ forms an inner product space.

Now, an inner product space on X defines a norm on X given by

$$||x|| = \sqrt{\langle x, x \rangle},$$

and similarly, a metric as

$$d(x, y) = ||x - y|| = \sqrt{\langle x - y, x - y \rangle}.$$

Hence inner product spaces are normed spaces. But the converse is not necessarily true in general. We will consider this in a while. First, let us see the parallelogram equality. It says that the norm on an inner product space X satisfies the parallelogram equality which is

$$||x + y||^2 + ||x - y||^2 = 2(||x||^2 + ||y||^2),$$

for all $x, y \in X$. The idea generalises the elementary concept of length of a vector.

Example 14.2. 1. The space $\mathbb{R}^n$ is an inner product with the inner product defined as

$$\langle x, y \rangle = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$$

where $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_n)$ are two elements of $\mathbb{R}^n$.

2. The unitary space $\mathbb{C}^n$ is an inner product space where the inner product is defined as

$$\langle x, y \rangle = x_1 \bar{y}_1 + x_2 \bar{y}_2 + \cdots + x_n \bar{y}_n$$

where $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_n)$ are two elements of $\mathbb{C}^n$.

3. The l^2 space forms an inner product space with the inner product defined as

$$\langle x, y \rangle = \sum_{i=1}^{\infty} x_i \bar{y}_i.$$

The convergence of the series follows from the Cauchy-Schwarz inequality and the fact that $x, y \in l^2$, by assumption.

4. The space l^p , $p \neq 2$ is not an inner product space since the norm does not satisfy the parallelogram equality. Let us take, $x = (1, 1, 0, 0, \dots)$ and $y = (1, -1, 0, 0, \dots)$ in l^p and calculate

$$\|x\| = \|y\| = 2^{1/p}.$$

And,

$$\|x + y\| = \|x - y\| = 2.$$

Hence, the parallelogram equality is not satisfied. Hence, we get a NLS which is not an inner product space.

5. The space $C[a, b]$ is not an inner product space. We will show that the norm defined by

$$\|x\| = \max_{t \in [a, b]} |x(t)|$$

can't form an inner product since it does not satisfy the parallelogram equality. Let us take $x(t) = 1$ and $y(t) = \frac{t-a}{b-a}$, we have obviously $\|x\| = 1$ and $\|y\| = 1$ and

$$\begin{aligned} x(t) + y(t) &= 1 + \frac{t-a}{b-a} \\ x(t) - y(t) &= 1 - \frac{t-a}{b-a}. \end{aligned}$$

Hence, $\|x + y\| = 2$ and $\|x - y\| = 1$ and

$$\|x + y\|^2 + \|x - y\|^2 = 5,$$

but

$$2(\|x\|^2 + \|y\|^2) = 4.$$

Thus, we are done.

Definition 14.3. An element $x \in X$ is said to be orthogonal to an element $y \in X$ if

$$\langle x, y \rangle = 0.$$

If x and y are orthogonal, then we give the notation as $x \perp y$. Similarly, two subsets A and B are called orthogonal if $a \perp b$ for all $a \in A$ and $b \in B$.

We mention the following fact for inner product spaces. If X is a real inner product space, then for every $x, y \in X$,

$$\langle x, y \rangle = \frac{1}{4}(\|x + y\|^2 - \|x - y\|^2)$$

and for a complex inner product space, we have

$$\begin{aligned} \operatorname{Re}\langle x, y \rangle &= \frac{1}{4}(\|x + y\|^2 - \|x - y\|^2) \\ \operatorname{Im}\langle x, y \rangle &= \frac{1}{4}(\|x + iy\|^2 - \|x - iy\|^2) \end{aligned}$$

The above formula is called the polarization identity.

Theorem 14.4. If $x \perp y$ in a real inner product space X , then show that

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2.$$

Proof. Since $x \perp y$, so we have $\langle x, y \rangle = 0$. Thus,

$$\begin{aligned} \|x + y\|^2 &= \langle x + y, x + y \rangle \\ &= \langle x, x \rangle + 2\langle x, y \rangle + \langle y, y \rangle \\ &= \|x\|^2 + \|y\|^2. \end{aligned}$$

Hence proved. □

Theorem 14.5. (Schwarz Inequality) Let X be an inner product space. Then, for all $x, y \in X$, we have,

$$|\langle x, y \rangle| \leq \|x\| \|y\|.$$

The equality occurs if and only if the set $\{x, y\}$ is a linearly dependent set.

Proof. If $y = 0$, then the result is trivially true. Let $y \neq 0$. For every scalar α , we have

$$\begin{aligned} 0 \leq \|x - \alpha y\|^2 &= \langle x - \alpha y, x - \alpha y \rangle \\ &= \langle x, x \rangle - \bar{\alpha}\langle x, y \rangle - \alpha[\langle y, x \rangle - \bar{\alpha}\langle y, y \rangle]. \end{aligned}$$

We see that the expression in the bracket is zero if we choose

$$\bar{\alpha} = \frac{\langle y, x \rangle}{\langle y, y \rangle}.$$

The remaining inequality is

$$0 \leq \langle x, x \rangle - \frac{\langle y, x \rangle}{\langle y, y \rangle} \langle x, y \rangle = \|x\|^2 - \frac{|\langle x, y \rangle|^2}{\|y\|^2};$$

here we used $\langle y, x \rangle = \overline{\langle x, y \rangle}$. Multiplying by $\|y\|^2$ and simplifying, we get the required inequality.

Equality holds in the derivation if and only if $y = 0$ or $\|x - \alpha y\|^2 = 0$, hence $x - \alpha y = 0$, so that $x = \alpha y$, which shows the linear dependence. $\square$

Theorem 14.6. (Triangle Inequality) Let X be an inner product space. Then, for all $x, y \in X$, we have,

$$\|x + y\| \leq \|x\| + \|y\|.$$

The equality occurs if and only if $y = 0$ or $x = cy$ for some $c \in \mathbb{R}$ and $c \geq 0$.

Proof. We have

$$\begin{aligned} \|x + y\|^2 &= \langle x + y, x + y \rangle \\ &= \|x\|^2 + \langle x, y \rangle + \langle y, x \rangle + \|y\|^2. \end{aligned}$$

By Schwarz inequality,

$$|\langle x, y \rangle| = |\langle y, x \rangle| \leq \|x\| \|y\|.$$

By the triangle inequality for numbers we thus obtain

$$\begin{aligned} \|x + y\|^2 &\leq \|x\|^2 + 2|\langle x, y \rangle| + \|y\|^2 \\ &\leq \|x\|^2 + 2\|x\| \|y\| + \|y\|^2 \\ &= (\|x\| + \|y\|)^2. \end{aligned}$$

Taking square roots on both sides, we get the desired result.

Equality holds if and only if

$$\langle x, y \rangle + \langle y, x \rangle = 2\|x\| \|y\|.$$

The left hand side is $2\operatorname{Re}\langle x, y \rangle$, where, Re denotes the real part. From this, and Schwarz inequality, we get

$$\operatorname{Re}\langle x, y \rangle = \|x\| \|y\| \geq |\langle x, y \rangle|.$$

Since the real part of a complex number cannot exceed the absolute value, we must have equality, which implies linear dependence by part 1, say $y = 0$ or $x = cy$. We show that c is real and greater than equals to 0. From the above equation, with the equality sign, we have $\operatorname{Re}\langle x, y \rangle = |\langle x, y \rangle|$. But if the real part of a complex number equals the absolute value, the imaginary part must be zero. Hence, $\langle x, y \rangle = \operatorname{Re}\langle x, y \rangle \geq 0$, by the above equation and $c \geq 0$ follows from

$$0 \leq \langle x, y \rangle = \langle cy, y \rangle = c\|y\|^2.$$

$\square$

Theorem 14.7. The inner product is continuous, that is, if $x_n \rightarrow x$ and $y_n \rightarrow y$, then $\langle x_n, y_n \rangle \rightarrow \langle x, y \rangle$.

Proof. Subtracting and adding a term, using the triangle inequality for numbers and finally, the Schwarz inequality, we obtain,

$$\begin{aligned} |\langle x_n, y_n \rangle - \langle x, y \rangle| &= |\langle x_n, y_n \rangle - \langle x_n, y \rangle + \langle x_n, y \rangle - \langle x, y \rangle| \\ &\leq |\langle x_n, y_n - y \rangle| + |\langle x_n - x, y \rangle| \\ &\leq \|x_n\| \|y_n - y\| + \|x_n - x\| \|y\| \rightarrow 0 \end{aligned}$$

since $x_n \rightarrow x$ and $y_n \rightarrow y$ as $n \rightarrow \infty$. Hence the result. $\square$

Definition 14.8. A set of vectors $\{x_1, x_2, \dots, x_n\}$ in an inner product space X is said to be orthogonal if

$$\langle x_i, x_j \rangle = 0, \quad i \neq j.$$

And, a set of vectors $\{x_1, x_2, \dots, x_n\}$ in an inner product space is said to be orthonormal if

$$\begin{aligned} \langle x_i, x_j \rangle &= 0, \quad i \neq j. \\ &= 1, \quad i = j. \end{aligned}$$

Example 14.9. The set of vectors $\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$ is an orthonormal set in $\mathbb{R}^3$.

Theorem 14.10. An orthogonal set of non-null vectors in an inner product space V is linearly independent.

Proof. Left as an exercise. $\square$

Corollary 14.10.1. An orthonormal set of vectors of an inner product space is linearly independent.

Definition 14.11. Let z be a fixed non-null vector in an inner product space X . Then, for a non-null vector x in X , there exists a unique real number c such that $x - cz$ is orthogonal to z . This c is determined as $\langle x - cz, z \rangle = 0$. Hence, $\langle x, z \rangle = c \langle z, z \rangle$. So,

$$c = \frac{\langle x, z \rangle}{\langle z, z \rangle}.$$

This c is called the scalar component of x along z and cz is said to be the projection of x upon z .

Theorem 14.12. (Bessel's Inequality) Let $\{x_1, x_2, \dots, x_n\}$ be an orthonormal set of vectors in an inner product space X , then for any vector x in X ,

$$\|x\|^2 \geq c_1^2 + c_2^2 + \dots + c_n^2$$

where, c_i are the scalar component of x along x_i .

Proof. For all i , $c_i = \langle x, x_i \rangle$, since $\langle x_i, x_i \rangle = 1$. Now, $x - c_1x_1 - c_2x_2 - \dots - c_nx_n$ is orthogonal to each x_i since

$$\langle x - c_1x_1 - c_2x_2 - \dots - c_nx_n, x_i \rangle = \langle x, x_i \rangle - c_i \langle x_i, x_i \rangle = 0.$$

It follows that $x - c_1x_1 - c_2x_2 - \dots - c_nx_n$ is orthogonal to $c_1x_1 + c_2x_2 + \dots + c_nx_n$. Thus,

$$x = (x - c_1x_1 - c_2x_2 - \dots - c_nx_n) + (c_1x_1 + c_2x_2 + \dots + c_nx_n).$$

By Pythagoras theorem,

$$\begin{aligned} \|x\|^2 &= \|x - c_1x_1 - c_2x_2 - \cdots - c_nx_n\|^2 + \|c_1x_1 + c_2x_2 + \cdots + c_nx_n\|^2 \\ &\geq \|c_1x_1 + c_2x_2 + \cdots + c_nx_n\|^2. \end{aligned}$$

But,

$$\begin{aligned} \|c_1x_1 + c_2x_2 + \cdots + c_nx_n\|^2 &= \langle c_1x_1 + c_2x_2 + \cdots + c_nx_n, c_1x_1 + c_2x_2 + \cdots + c_nx_n \rangle \\ &= c_1^2 + c_2^2 + \cdots + c_n^2, \end{aligned}$$

since $\{x_1, x_2, \dots, x_n\}$ is orthonormal. So, we get,

$$\|x\|^2 \geq c_1^2 + c_2^2 + \cdots + c_n^2.$$

Hence proved. □

The equality occurs when

$$\|x - c_1x_1 - c_2x_2 - \cdots - c_nx_n\|^2 = 0,$$

that is, when $x = c_1x_1 + c_2x_2 + \cdots + c_nx_n$, that is, when $x = L\{x_1, x_2, \dots, x_n\}$.

Theorem 14.13. Let $\{x_1, x_2, \dots, x_n\}$ be an orthogonal set of vectors in an inner product space X . Then any vector $x \in L\{x_1, x_2, \dots, x_n\}$ has a unique representation

$$x = c_1x_1 + c_2x_2 + \cdots + c_nx_n,$$

where, c_i is the scalar component of x along x_i .

Theorem 14.14. (Parseval's Theorem) Let $\{x_1, x_2, \dots, x_n\}$ be an orthonormal basis of an inner product space X , then for any vector $x \in X$,

$$\|x\|^2 = c_1^2 + c_2^2 + \cdots + c_n^2.$$

where, c_i is the scalar component of x along x_i .

Proof. Since $\{x_1, x_2, \dots, x_n\}$ is a basis of X , so for any vector $x \in X$, we will have

$$x = c_1x_1 + c_2x_2 + \cdots + c_nx_n,$$

where, c_i is the scalar component of x along x_i . So,

$$\begin{aligned} \|x\|^2 &= \langle x, x \rangle \\ &= \langle c_1x_1 + c_2x_2 + \cdots + c_nx_n, c_1x_1 + c_2x_2 + \cdots + c_nx_n \rangle \\ &= c_1^2 + c_2^2 + \cdots + c_n^2, \end{aligned}$$

since $\{x_1, x_2, \dots, x_n\}$ is orthonormal. Hence the proof. □

Exercise 14.15. 1. Show that every inner product space is a NLS.

2. If in an inner product space, $\langle x, u \rangle = \langle x, v \rangle$ for all x , then show that $u = v$.
3. If an inner product space X is real, show that the condition $\|x\| = \|y\|$ implies that $\langle x+y, x-y \rangle = 0$.
4. Show that a set of orthogonal vectors is linearly independent.
5. Show that in an inner product space X , $x \perp y$ if and only if $\|x + \alpha y\| \geq \|x\|$ for all scalars α .
6. Show that in an inner product space X , $x \perp y$ if and only if $\|x + \alpha y\| = \|x - \alpha y\|$ for all scalars α .

Unit XV

Course Structure

- Introduction
 - Objectives
 - Hilbert's space
 - Orthogonal complement
 - Projection theorem
-

Introduction

In this unit, we will study the Hilbert spaces, which is nothing but the analogous of Banach spaces in an inner product space. We will study related concepts.

Objectives

After reading this unit, you will be able to

- define Hilbert spaces
- give examples of Hilbert spaces
- apply the related concepts in various cases
- define orthogonal complement
- learn the projection theorem

15 Hilbert Spaces

In the preceding unit, we have learnt about inner product spaces and the metric defined by the inner product. Hilbert space is none other than a complete inner product space (complete in the metric defined by the inner product). We have seen various examples of inner product spaces in the preceding unit. Let us see whether they form a Hilbert space or not.

Example 15.1. 1. The Euclidean space $\mathbb{R}^n$ is a Hilbert space with the inner product defined as

$$\langle x, y \rangle = x_1y_1 + x_2y_2 + \cdots + x_ny_n$$

where $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_n)$. We have

$$\|x\| = \sqrt{\langle x, x \rangle} = (x_1^2 + x_2^2 + \cdots + x_n^2)^{1/2}$$

and from this the Euclidean metric defined by

$$d(x, y) = \|x - y\| = \sqrt{\langle x - y, x - y \rangle} = [(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2]^{1/2}.$$

We already know that $\mathbb{R}^n$ with the above metric is complete. So, $\mathbb{R}^n$ is a Hilbert space with the given inner product.

2. Similarly, we can show that $\mathbb{C}^n$ is a Hilbert space with the inner product as defined in the preceding unit.
3. The space l^2 is a Hilbert space with the inner product

$$\langle x, y \rangle = \sum_{j=1}^{\infty} x_j \bar{y}_j.$$

4. In the preceding unit, we have shown that the space l^p , $p \neq 2$ is not an inner product space and hence can't be a Hilbert space.
5. Also, the space $C[a, b]$ is not an inner product space and hence can't be a Hilbert space as we have previously seen.

Definition 15.2. An isomorphism T of an inner product space X onto an inner product space Y over the same field is a bijective linear operator $T : X \rightarrow Y$ which preserves the inner product, that is, for all $x, y \in X$,

$$\langle T(x), T(y) \rangle = \langle x, y \rangle,$$

where we denoted inner products on X and Y by the same symbol. Y is then called isomorphic with X , and X and Y are called isomorphic inner product spaces. Note that the bijectivity and linearity guarantees that T is a vector space isomorphism of X onto Y , so that T preserves the whole structure of inner product space. T is also an isometry of X onto Y since the distance is determined by the inner products on X and Y .

Definition 15.3. A subspace A of an inner product space X is defined to be a vector subspace of X taken with the inner product on X restricted to $A \times A$. If X is a Hilbert space, then note that, A need not be so since A need not be complete.

In fact, we have the following theorem in this direction.

Theorem 15.4. Let A be a subspace of a Hilbert space X . Then

1. A is complete if and only if A is closed in X .

2. If A is finite dimensional, then A is complete.
3. If X is complete, so is A . More generally, every subset of a separable inner product space is separable.

Proof. The proof is simple and is left as an exercise. $\square$

We have seen in the preceding unit, that in an inner product space, $x_n \rightarrow x$ and $y_n \rightarrow y$ implies that $\langle x_n, y_n \rangle \rightarrow \langle x, y \rangle$. As a first application of this lemma, let us prove that every inner product space can be completed. The completion is a Hilbert space and is unique except for isomorphisms.

The theorem about the completion of an inner product space can now be stated as follows.

Theorem 15.5. For any inner product space X there exists a Hilbert space H and an isomorphism A from X onto a dense subspace $W \subset H$. The space H is unique except for isomorphisms.

15.1 Orthogonal Complements and Direct Sums

Definition 15.6. The segment joining two given elements x and y of a vector space X is defined to be the set of all $z \in X$ of the form

$$z = ax + (1 - a)y,$$

where $a \in \mathbb{R}$ and $0 \leq a \leq 1$. A subset M of X is said to be convex if for every $x, y \in M$ the segment joining x and y is contained in M .

In a metric space X , the distance δ from an element $x \in X$ to a non-empty subset $M \subset X$ is defined as

$$\delta = \inf_{m \in M} d(x, m).$$

The equivalent definition for a normed space is

$$\delta = \inf_{m \in M} \|x - m\|,$$

for a non-empty set $M \subset X$. Now, it is important to know whether there exists a $y \in M$ for which $\|x - y\| = \delta$. If such an element exists, is it unique? The answer is given in the following theorem.

Theorem 15.7. Let X be an inner product space and $M \neq \emptyset$ a convex subset which is complete (in the metric induced by the inner product). Then for every given $x \in X$, there exists a unique $y \in M$ such that

$$\delta = \inf_{m \in M} \|x - m\| = \|x - y\|.$$

Proof. By the definition of an infimum, there is a sequence $\{y_n\}$ in M such that

$$\delta_n = \|x - y_n\|, \quad \delta_n \rightarrow \delta.$$

We show that $\{y_n\}$ is Cauchy. Writing $y_n - x = v_n$, we have, $\|v_n\| = \delta_n$ and

$$\begin{aligned} \|v_n + v_m\| &= \|y_n + y_m - 2x\| \\ &= 2\left\|\frac{1}{2}(y_n + y_m) - x\right\| \\ &\geq 2\delta. \end{aligned}$$

because M is convex, so that $\frac{1}{2}(y_n + y_m) \in M$. Furthermore, we have $y_n - y_m = v_n - v_m$. Hence, by parallelogram equality,

$$\begin{aligned} \|y_n - y_m\|^2 &= \|v_n - v_m\|^2 \\ &= -\|v_n + v_m\|^2 + 2(\|v_n\|^2 + \|v_m\|^2) \\ &\leq -(2\delta)^2 + 2(\delta_n^2 + \delta_m^2), \end{aligned}$$

which implies that $\{y_n\}$ is Cauchy since $\delta_n \rightarrow \delta$. Since M is complete, $\{y_n\}$ converges, say $y_n \rightarrow y \in M$. Since $y \in M$, we have $\|x - y\| \geq \delta$. Also,

$$\begin{aligned} \|x - y\| &\leq \|x - y_n\| + \|y_n - y\| \\ &= \delta_n + \|y_n - y\| \rightarrow \delta. \end{aligned}$$

This shows that $\|x - y\| = \delta$.

We will now show that this y is unique. Let $y \in M$ and $z \in M$ be such that

$$\|x - y\| = \delta, \quad \|x - z\| = \delta.$$

We show that $y = z$. By the parallelogram equality,

$$\begin{aligned} \|y - z\|^2 &= \|(y - x) - (z - x)\|^2 \\ &= 2\|y - x\|^2 + 2\|z - x\|^2 - \|(y - x) + (z - x)\|^2 \\ &= 2\delta^2 + 2\delta^2 - 2\left\|\frac{1}{2}(y + z) - x\right\|^2. \end{aligned}$$

The element $\frac{1}{2}(y + z) \in M$, so that

$$\left\|\frac{1}{2}(y + z) - x\right\| \geq \delta.$$

This implies that the right-hand side is less than or equal to $2\delta^2 + 2\delta^2 - 4\delta^2 = 0$. Hence we have the inequality $\|y - z\| \leq 0$. Clearly, $\|y - z\| \geq 0$. Combining, we have $\|y - z\| = 0$, which implies $y = z$. $\square$

Turning from arbitrary convex sets to subspaces, we obtain a lemma which generalizes the familiar idea of elementary geometry that the unique point Y in a given subspace Y closest to a given x is found by "dropping a perpendicular from x to Y ."

Theorem 15.8. In the previous theorem, let M be a complete subspace Y and $x \in X$ is fixed. Then $z = x - y$ is orthogonal to Y .

Proof. Let $z \notin Y$. Then there must be some $s \in Y$ such that

$$\langle z, s \rangle = \beta \neq 0.$$

Clearly, $s \neq 0$, since otherwise $\langle z, s \rangle = 0$. Further, for any scalar α ,

$$\begin{aligned} \|z - \alpha s\|^2 &= \langle z - \alpha s, z - \alpha s \rangle \\ &= \langle z, z \rangle - \bar{\alpha} \langle z, s \rangle - \alpha [\langle s, z \rangle - \bar{\alpha} \langle s, s \rangle] \\ &= \langle z, z \rangle - \bar{\alpha} \beta - \alpha [\bar{\beta} - \bar{\alpha} \langle s, s \rangle]. \end{aligned}$$

The expression in the brackets is zero if we choose

$$\bar{\alpha} = \frac{\bar{\beta}}{\langle s, s \rangle}.$$

From the previous theorem, we have $\|z\| = \|x - y\| = \delta$, so that our equation now yields

$$\|z - \alpha s\|^2 = \|z\|^2 - \frac{|\beta|^2}{\langle s, s \rangle} < \delta^2.$$

But this is impossible because we have

$$z - \alpha s = x - t, \quad \text{where } t = y + \alpha s \in Y,$$

so that $\|z - \alpha s\| \geq \delta$ by the definition of δ . Hence, our assumption is false and $z \perp Y$. $\square$

Definition 15.9. (Direct Sum) A vector space X is said to be the direct sum of two subspaces Y and Z of X , written

$$X = Y \oplus Z,$$

if each $x \in X$ has a unique representation

$$x = y + z, \quad y \in Y, \quad z \in Z.$$

Then Z is called an algebraic complement of Y in X and vice versa, and Y, Z is called a complementary pair of subspaces in X .

Similarly, in the case of a general Hilbert space H , the main interest concerns representations of H as a direct sum of a closed subspace Y and its orthogonal complement

$$Y^\perp = \{z \in H \mid z \perp Y\},$$

which is the set of all vectors orthogonal to Y . This gives our main result in this section, which is sometimes called the projection theorem, for reasons to be explained after the proof.

Theorem 15.10. Let Y be any closed subspace of a Hilbert space H . Then

$$H = Y \oplus Z,$$

where, $Z = Y^\perp$.

Proof. Since H is complete and Y is closed, Y is complete. Since Y is convex, by the previous theorem and lemma, for every $x \in H$ there is a $y \in Y$ such that

$$x = y + z,$$

$z \in Z = Y^\perp$. To prove the uniqueness, we assume that

$$x = y + z = y_1 + z_1,$$

where, $y, y_1 \in Y$ and $z, z_1 \in Z$. Then, $y - y_1 = z_1 - z$. Since $y - y_1 \in Y$, whereas $z - z_1 \in Z = Y^\perp$, we see that $y - y_1 \in Y \cap Y^\perp = \{0\}$. This implies that $y = y_1$. Hence also $z = z_1$. $\square$

y in the above theorem is called the orthogonal projection of x on Y (or the projection of x on Y). If we take $\mathbb{R}^2$ and project any point $a = (\xi, \eta)$ on the x -axis, which then plays the role of Y , the projection is $y = (\xi, 0)$.

From the above theorem, we can define a mapping

$$P : H \rightarrow Y, \quad x \mapsto y = P(x).$$

Then P is called the (orthogonal) projection (or projection operator) of H onto Y . Obviously, P is a bounded linear operator. P maps H onto Y , Y onto itself and $Z = Y^\perp$ onto $\{0\}$, and is idempotent, that is

$$P^2 = P.$$

Thus, for every $x \in H$,

$$P^2(x) = P(P(x)) = P(x).$$

Hence, $P|_Y$ is the identity operator on Y .

Exercise 15.11. 1. Show that $y \perp x_n$ and $x_n \rightarrow x$ together imply $x \perp y$.

2. Show that for a sequence $\{x_n\}$ in an inner product space the conditions $\|x_n\| \rightarrow \|x\|$ and $\langle x_n, x \rangle \rightarrow \langle x, x \rangle$ imply $x_n \rightarrow x$.

3. Show that for any set $M \neq \emptyset$, the set $M^\perp$ is a vector space.

4. Show that $M^\perp$ is closed.

5. If Y is a closed subspace of a Hilbert space H , then show that $Y = Y^{\perp\perp}$.

6. Consider a subset M of $\mathbb{R}^2$. Find $M^\perp$ if M is

(a) $\{x\}$ where $x = (\xi_1, \xi_2) \neq 0$,

(b) a linearly independent set $\{x_1, x_2\} \subset \mathbb{R}^2$.

7. Let A and $B \supset A$ be non-empty subsets of an inner product space X . Then show that

(a) $A \subset A^\perp$,

(b) $B^\perp \subset A^\perp$,

(c) $A^{\perp\perp\perp} = A^\perp$.

8. Show that a subspace Y of a Hilbert space H is closed in H if and only if $Y = Y^{\perp\perp}$.

9. Let $M \neq \emptyset$ is a subset of a Hilbert space H , show that $M^{\perp\perp}$ is the smallest closed subspace of H which contains M .

Unit XVI

Course Structure

- Introduction
 - Objectives
 - Riesz Representation theorem
 - Convergence of series corresponding to orthogonal sequence
 - Fourier coefficient
-

Introduction

Orthogonality of elements as defined in the preceding units plays a basic role in inner product and Hilbert spaces. Of particular interest are sets whose elements are orthogonal in pairs. In this unit, we will mainly focus on the concepts of orthonormal sets and sequences and their explicit applications.

Objectives

After reading this unit, you will be able to

- learn the Riesz Representation theorem
- apply the Riesz Representation theorem
- learn the concepts of orthogonal sets and sequences
- learn their properties
- give examples of each
- learn about the series related to the orthonormal sequences and sets
- apply them in different circumstances

16 Riesz Representation Theorem

Let us start with the Riesz Representation theorem.

Theorem 16.1. Every bounded linear functional f on a Hilbert space H can be represented in terms of the inner product, namely,

$$f(x) = \langle x, z \rangle \quad (16.1)$$

where z depends on f , is uniquely determined by f and has norm

$$\|z\| = \|f\|. \quad (16.2)$$

Proof. If $f = 0$ then (16.1) and (16.2) hold if we take $z = 0$. Let $f \neq 0$. First of all, $z \neq 0$ since otherwise $f = 0$. Second, $\langle x, z \rangle = 0$ for all x for which $f(x) = 0$, that is, for all x in the null space $\mathcal{N}(f)$ of f . Hence $z \perp \mathcal{N}(f)$. This suggests that we consider $\mathcal{N}(f)$ and its orthogonal complement $\mathcal{N}(f)^\perp$. $\mathcal{N}(f)$ is a vector space which is closed. Further, $f \neq 0$ implies $\mathcal{N}(f) \neq H$, so that $\mathcal{N}(f)^\perp \neq \{0\}$ by the projection theorem. Hence $\mathcal{N}(f)^\perp$ contains a $z_0 \neq 0$. We set

$$v = f(x)z_0 - f(z_0)x$$

where $x \in H$ is arbitrary. Applying f , we obtain

$$f(v) = f(x)f(z_0) - f(z_0)f(x) = 0.$$

This shows that $v \in \mathcal{N}(f)$. Since $z_0 \perp \mathcal{N}(f)$, we have

$$\begin{aligned} 0 &= \langle v, z_0 \rangle \\ &= \langle f(x)z_0 - f(z_0)x, z_0 \rangle \\ &= f(x)\langle z_0, z_0 \rangle - f(z_0)\langle x, z_0 \rangle. \end{aligned}$$

Noting that $\langle z_0, z_0 \rangle = \|z_0\|^2 \neq 0$, we can solve for $f(x)$. The result is

$$f(x) = \frac{f(z_0)}{\langle z_0, z_0 \rangle} \langle x, z_0 \rangle.$$

This can be written in the form (16.1), where

$$z = \frac{\overline{f(z_0)}}{\langle z_0, z_0 \rangle} z_0.$$

Since $x \in H$ was arbitrary, (16.1) is proved.

We will now prove that z in (16.1) is unique. Suppose for all $x \in H$,

$$f(x) = \langle x, z_1 \rangle = \langle x, z_2 \rangle.$$

Then, $\langle x, z_1 - z_2 \rangle = 0$ for all x . Choosing the particular $x = z_1 - z_2$, we have

$$\langle x, z_1 - z_2 \rangle = \langle z_1 - z_2, z_1 - z_2 \rangle = \|z_1 - z_2\|^2 = 0,$$

which implies that $z_1 = z_2$. Hence the uniqueness.

We finally prove the second part of the theorem. If $f = 0$, then $z = 0$ and (16.2) holds. Let $f \neq 0$. Then $z \neq 0$. From (16.1), with $x = z$ we obtain

$$||z||^2 = \langle z, z \rangle = f(z) \leq ||f|| ||z||.$$

This implies

$$||f|| = \sup_{||x||=1} |\langle x, z \rangle| \leq ||z||.$$

□

16.1 Convergence of series corresponding to orthogonal sequence

We will first recapitulate the definitions of orthonormal sets and sequences in more detail.

Definition 16.2. An orthogonal set M in an inner product space X is a subset $M \subset X$ whose elements are pairwise orthogonal. An orthonormal set $M \subset X$ is an orthogonal set in X whose elements have norm 1, that is, for all $x, y \in M$,

$$\begin{aligned} \langle x, y \rangle &= 0, \quad x \neq y \\ &= 1, \quad x = y. \end{aligned}$$

If an orthogonal or orthonormal set M is countable, we can arrange it in a sequence $\{x_n\}$ and call it an orthogonal or orthonormal sequence, respectively.

More generally, an indexed set $\{x_\alpha\}$, $\alpha \in I$ is called orthogonal if $x_\alpha \perp x_\beta$ for all $\alpha, \beta \in I$, with $\alpha \neq \beta$. The family is called orthonormal if it is orthogonal and all x_α have norm 1, so that for all $\alpha, \beta \in I$, we have

$$\begin{aligned} \langle x_\alpha, x_\beta \rangle = \delta_{\alpha\beta} &= 0, \quad \alpha \neq \beta, \\ &= 1, \quad \alpha = \beta. \end{aligned}$$

Let us consider simple properties of orthogonal elements. For orthogonal elements x, y , we always have

$$||x + y||^2 = ||x||^2 + ||y||^2.$$

This is called the Pythagorean relation. More generally, if we have an orthogonal set $\{x_1, x_2, \dots, x_n\}$ then

$$||x_1^2 + x_2^2 + \dots + x_n^2||^2 = ||x_1||^2 + \dots + ||x_n||^2.$$

We have already seen that an orthonormal set is linearly independent.

Example 16.3. 1. In the space $\mathbb{R}^3$, the three unit vectors $\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$ forms an orthonormal set.

2. In the l^2 space, an orthonormal set is the sequence $\{e_n\}$ such that e_n is the sequence of numbers whose n th term is one and the other terms are zero.

A great advantage of orthonormal sequences over arbitrary linearly independent sequences is the following. If we know that a given x can be represented as a linear combination of some elements of an orthonormal sequence, then the orthonormality makes the actual determination of the coefficients very easy. In fact, if $\{e_1, e_2, \dots\}$ is an orthonormal sequence in an inner product space X and we have $x \in L(\{e_1, e_2, \dots, e_n\})$, for some fixed n . Then, we have

$$x = \sum_{k=1}^n \alpha_k e_k,$$

and if we take the inner product by a fixed e_j , we get

$$\langle x, e_j \rangle = \alpha_j.$$

Thus, we get

$$x = \sum_{k=1}^n \langle x, e_k \rangle e_k.$$

This shows that the determination of the unknown coefficients is simple.

If we consider some $x \in X$, which is not necessarily in $Y_n = L(\{e_1, e_2, \dots, e_n\})$, we can define $y \in Y_n$ as

$$y = \sum_{k=1}^n \langle y, e_k \rangle e_k,$$

and then define z as

$$x = y + z,$$

or, $z = x - y$. We want to show that $z \perp y$. Note that every $y \in Y_n$ can be represented as

$$y = \sum_{k=1}^n \alpha_k e_k,$$

where $\alpha_k = \langle y, e_k \rangle$. By orthonormality, we have

$$\|y\|^2 = \left\langle \sum \langle x, e_m \rangle e_m, \sum \langle x, e_k \rangle e_k \right\rangle = \sum |\langle x, e_k \rangle|^2.$$

Using this, we will now show that $z \perp y$.

$$\begin{aligned} \langle z, y \rangle &= \langle x - y, y \rangle \\ &= \langle x, y \rangle - \langle y, y \rangle \\ &= \left\langle x, \sum \langle x, e_k \rangle e_k \right\rangle - \|y\|^2 \\ &= \sum \langle x, e_k \rangle \overline{\langle x, e_k \rangle} - \sum |\langle x, e_k \rangle|^2 \\ &= 0. \end{aligned}$$

Hence, by Pythagorean relation, we get

$$\|x\|^2 = \|y\|^2 + \|z\|^2.$$

Thus,

$$||y||^2 = ||x||^2 - \sum |\langle x, e_k \rangle|^2.$$

Since $||z|| \geq 0$, we have for every $n = 1, 2, \dots$

$$\sum |\langle x, e_k \rangle|^2 \leq ||x||^2.$$

These sums have nonnegative terms, so that they form a monotone increasing sequence. This sequence converges because it is bounded by $||x||^2$. This is the sequence of the partial sums of an infinite series, which thus converges. Thus we get the Bessel' Inequality, which we have done earlier.

Theorem 16.4. Let $\{e_k\}$ be an orthonormal sequence in an inner product space X . Then for every $x \in X$, we have

$$\sum_{k=1}^{\infty} |\langle x, e_k \rangle|^2 \leq ||x||^2. \quad (16.3)$$

The inner products in (16.3) are called the Fourier coefficients of x with respect to the orthonormal sequence $\{e_k\}$.

Note that if X is finite dimensional, then every orthonormal set in X must be finite because it is linearly independent. Hence the series in (16.3) must be finite.

We have seen that orthonormal sequences are very convenient to work with. The remaining practical problem is how to obtain an orthonormal sequence if an arbitrary linearly independent sequence is given. This is accomplished by a constructive procedure, the GramSchmidt process for orthonormalizing a linearly independent sequence $\{x_n\}$ in an inner product space. The resulting orthonormal sequence $\{e_j\}$ has the property that for every n ,

$$\text{LS}(\{e_1, \dots, e_n\}) = \text{LS}(\{x_1, \dots, x_n\}).$$

Given any orthonormal sequence $\{e_k\}$ in a Hilbert space H , we may consider series of the form

$$\sum_{k=1}^{\infty} \alpha_k e_k$$

where α_k are any scalars. We will discuss the convergence of such series. We have the following theorem in this direction.

Theorem 16.5. Let $\{e_k\}$ be an orthonormal sequence in a Hilbert space H , then

1. the series

$$\sum_{k=1}^{\infty} \alpha_k e_k$$

converges if and only if the series

$$\sum_{k=1}^{\infty} |\alpha_k|^2.$$

2. if the series converges, then the coefficients α_k are the Fourier coefficients $\langle x, e_k \rangle$, where x is the sum of the series. Hence, we can write

$$x = \sum_{k=1}^{\infty} \langle x, e_k \rangle e_k.$$

3. For any $x \in H$, the given series with $\alpha_k = \langle x, e_k \rangle$ converges.

Proof. 1. Let $s_n = \alpha_1 e_1 + \cdots + \alpha_n e_n$ and $\sigma_n = |\alpha_1|^2 + \cdots + |\alpha_n|^2$. By virtue of orthonormality, for any m and $n > m$, we have

$$\begin{aligned} \|s_n - s_m\|^2 &= \|\alpha_{m+1} e_{m+1} + \cdots + \alpha_n e_n\|^2 \\ &= |\alpha_{m+1}|^2 + \cdots + |\alpha_n|^2 \\ &= \sigma_n - \sigma_m. \end{aligned}$$

Hence $\{s_n\}$ is Cauchy in H if and only if $\{\sigma_n\}$ is Cauchy in $\mathbb{R}$. Since H and $\mathbb{R}$ are complete, so the first part is proved.

2. Taking the inner product of s_n and e_j and using the orthonormality, we have

$$\langle s_n, e_j \rangle = \alpha_j, \quad j = 1, 2, \dots, k \quad (k \leq n \text{ and fixed}).$$

By the assumption that $s_n \rightarrow x$ since the inner product is continuous, we have

$$\alpha_j = \langle s_n, e_j \rangle \rightarrow \langle x, e_j \rangle, \quad (j \leq k).$$

Here we take $k(\leq n)$ as large as we please because $n \rightarrow \infty$, so that we have $\alpha_j = \langle x, e_j \rangle$ for every $j = 1, 2, \dots$

3. From the Bessel's Inequality, we see that the series

$$\sum_{k=1}^{\infty} |\langle x, e_k \rangle|^2$$

converges. from this, and 1 we conclude that 3 must hold. □

If an orthonormal family $\{e_k\}$, $k \in I$, in an inner product space X is uncountable, we can still form the Fourier coefficients $\langle x, e_k \rangle$ for an $x \in X$, where $k \in I$. We have the following lemma.

Lemma 16.6. Any x in an inner product space X can have at most countably many nonzero Fourier coefficients $\langle x, e_k \rangle$ with respect to the orthonormal family $\{e_k\}$, $k \in I$ in X .

Hence for any fixed $x \in H$, we can associate a series similar to that given in the previous theorem

$$\sum_{k=1}^{\infty} \langle x, e_k \rangle$$

and we can arrange the e_k with $\langle x, e_k \rangle \neq 0$ in a sequence $\{e_1, e_2, \dots\}$ so that the above equation takes the form of

$$x = \sum_{k=1}^{\infty} \langle x, e_k \rangle e_k.$$

Convergence follows from the previous Theorem. The sum of the series does not depend on the order in which those e_k are arranged in a sequence.

Exercise 16.7. 1. Show that an inner product space of finite dimension n has a basis $\{e_1, e_2, \dots, e_n\}$ of orthonormal vectors.

2. Show that in a Hilbert space H , convergence of $\sum \|x_j\|$ implies convergence of $\sum x_j$.

3. Let $\{e_k\}$ be an orthonormal sequence in a Hilbert space H . Show that for every $x \in H$, the vector

$$y = \sum_{k=1}^{\infty} \langle x, e_k \rangle e_k$$

exists in H and $x - y$ is orthogonal to every e_k

References

1. Functional Analysis, G. Bachman and L. Narici
2. Functional Analysis, W. Rudin
3. Functional Analysis, B. K. Lahiri
4. Functional Analysis, B. V. Limaye
5. Functional Analysis, A. E. Taylor

**POST GRADUATE DEGREE PROGRAMME (CBCS)
IN
MATHEMATICS**

**SEMESTER II
SELF LEARNING MATERIAL**

PAPER - MATC 2.3

**Mechanics – II (Classical Mechanics)
Abstract algebra - II
Operations Research - II**



**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, 2019

Directorate of Open and Distance Learning, University of Kalyani
Published by the Directorate of Open and Distance Learning,
University of Kalyani, Kalyani-741235, West Bengal

All rights reserved. No part of this work should be reproduced in any form without the permission in writing from the Directorate of Open and Distance Learning, University of Kalyani.

Director's Note

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

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

Self Learning Materials (SLMs) have been published by the Directorate of Open and Distance Learning, University of Kalyani, Kalyani-741235, West Bengal and all the copyright reserved for University of Kalyani. No part of this work should be reproduced in any form without permission in writing from the appropriate authority of the University of Kalyani.

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

Director
Directorate of Open and Distance Learning
University of Kalyani

CONTENTS

Sl. No.	Block	Unit	Page
1	Mechanics – II (Classical Mechanics)	1	2 – 10
		2	11 – 21
		3	22 – 29
		4	30 – 46
		5	47 – 53
2	Abstract Algebra - II	6	55 – 59
		7	60 – 65
		8	66 – 70
		9	71 – 78
		10	79 – 84
3	Operations Research - II	11	86 – 95
		12	96 – 104
		13	105 – 110
		14	111 – 116
		15	117 – 121
		16	122 – 130

Core Paper

MATC 2.3

Block - I

Marks : 32 (SSE : 25; IA : 07)

Mechanics - II (Classical Mechanics)

Syllabus

• Unit 1 •

Lagrangian Formulation : Generalised coordinates. Holonomic and nonholonomic systems. Scleronomic and rheonomic systems. D'Alembert's principle. Lagrange's equations. Energy equation for conservative fields. Cyclic (ignorable) coordinates. Generalised potential.

• Unit 2 •

Moving Coordinate System : Coordinate systems with relative translational motions. Rotating coordinate systems. The Coriolis force. Motion on the earth. Effect of Coriolis force on a freely falling particle. Euler's theorem. Euler's equations of motion for a rigid body. Eulerian angles.

• Unit 3 •

Variational Principle : Calculus of variations and its applications in shortest distance, minimum surface of revolution, Brachistochrone problem, geodesic. Hamilton's principle. Lagrange's undetermined multipliers. Hamilton's equations of motion.

• Unit 4 •

Canonical Transformations : Canonical coordinates and canonical transformations. Poincaré theorem. Lagrange's and Poisson's brackets and their variance under canonical transformations, Hamilton's equations of motion in Poisson's bracket. Jacobi's identity. Hamilton-Jacobi equation.

• Unit 5 •

Small Oscillations : General case of coupled oscillations. Eigen vectors and Eigen frequencies. Orthogonality of Eigen vectors. Normal coordinates. Two-body problem.

Unit 1

Course Structure

- Lagrangian Formulation : Generalised coordinates. Holonomic and nonholonomic systems. Scleronomic and rheonomic systems. D'Alembert's principle. Lagrange's equations. Energy equation for conservative fields. Cyclic (ignorable) coordinates. Generalised potential.

Introduction

Lagrangian mechanics is a reformulation of classical mechanics, introduced by the Italian-French mathematician and astronomer *Joseph-Louis Lagrange*. In Lagrangian mechanics, the trajectory of a system of particles is derived by solving Lagrange equation which treat constraints explicitly as extra equations or directly by judicious choice of generalised coordinates. In this case, a mathematical function called the *Lagrangian* is a function of the generalised coordinates, their first derivatives, and time, and contains the information about the dynamics of the system. Although Lagrangian formulation reduce to Newton's laws, they are characterized not only by the relative ease with which many problems can be formulated and solved but by their relationship in both theory and application to such advanced fields as quantum mechanics, statistical mechanics and electrodynamics.

Objective

The objective of this unit is to presents Lagrangian dynamics, an advance formalism for studying various problems in mechanics. Lagrangian techniques can provide a much cleaner way of solving some physical systems than Newtonian mechanics, in particular the inclusion of constraints on the motion. In this unit, we will learn the basics of Lagrangian mechanics. It is worthwhile to note that the Lagrangian approach is not independent of Newton's second law; the derivation of D'Alembert's principle, the critical step in developing Lagrangian mechanics, relies directly on Newton's second law.

1 Lagrangian Formulation

1.1 Some Basic Definitions

Definition 1.1. Degrees of freedom: The minimum number of independent variables required to specify the position and the configuration of a dynamical system which are compatible with the given constraints is called *degrees of freedom* (DOF) of the system.

- A particle moving freely in space requires 3 coordinates, e.g (x, y, z) to specify its position. Thus the number of degrees of freedom is 3.

- A system consisting of N particles moving freely in space requires $3N$ coordinates to specify its motion. Thus the number of degrees of freedom is $3N$.

- A system consisting of N particles moving in space subject to k independent constraints relation requires $3N - k$ coordinates to specify its motion. Thus the number of degrees of freedom is $3N - k$.

- The number of degrees of freedom of a rigid body consisting of N particles is 6, which independent of N (Why?).

Definition 1.2. Generalised coordinates: The number of independent coordinates sufficient to completely specify the configuration of a dynamical system is known as *generalised coordinate*. These generalised coordinates are denoted by $q_1, q_2, q_3, \dots, q_n$ such that all cartesian coordinates are known functions of the q_i variables, i.e.,

$$\begin{aligned} x_1 &= x_1(q_1, q_2, \dots, q_n, t) \\ x_2 &= x_2(q_1, q_2, \dots, q_n, t) \\ &\vdots \\ x_i &= x_i(q_1, q_2, \dots, q_n, t) \\ &\vdots \\ x_{3N} &= x_{3N}(q_1, q_2, \dots, q_n, t) \end{aligned} \tag{1.1}$$

Definition 1.3. Holonomic and non-holonomic system: Let $q_1, q_2, q_3, \dots, q_n$ denote the generalized coordinates describing a system and let t denote the time. If all the constraints of the system can be expressed as equations having of the form

$$f(q_1, q_2, \dots, q_n, t) = 0$$

or their equivalent, then the system is said to be *holonomic*. Otherwise the system is said to be *non-holonomic* system. A non-holonomic system can be expressed as

$$f(q_1, q_2, \dots, q_n, \dot{q}_1, \dot{q}_2, \dots, \dot{q}_n, t) = 0$$

Definition 1.4. Scleronomic and rheonomic system: In many mechanical systems of importance the time t does not enter explicitly in the system of equations (??). Such systems are sometimes called *scleronomic*. If the time t does enter explicitly, then the system is called *rheonomic*.

- The generalised coordinates do not necessarily belong to any conventional coordinate system. They can be any curvilinear coordinates as well as different non-dimensional parameters such as angles or their combinations.

- For holonomic systems, the number of generalised coordinates is exactly equal to the number of degrees of freedom of the system.

- For nonholonomic systems, the number of generalise coordinates required is larger than the number of degrees of freedom.

- The total time derivative of any generalised coordinate is called the generalised coordinate velocity or the generalised velocity. It is denoted by $\dot{q}_i$, that is,

$$\dot{q} = \frac{dq_i}{dt}, \quad i = 1, 2, \dots, n$$

1.2 D'Alembert's Principle

D'Alembert's principle, also known as the Lagrange–d'Alembert principle, is a statement of the fundamental classical laws of motion. The principle states that the sum of the differences between the forces acting on a system of mass particles and the time derivatives of the momenta of the system itself projected onto any virtual displacement consistent with the constraints of the system is zero. Thus, in symbols d'Alembert's principle is written as following,

$$\sum_{j=1}^N (\mathbf{F}_j^{(a)} - \dot{\mathbf{p}}_j) \cdot \delta \mathbf{r}_j = 0$$

where $\mathbf{F}_j^{(a)}$ is the total applied force on j -th particle, $\mathbf{p}_j$ the linear momentum of j -th particle and $\delta \mathbf{r}_j$ refers to the infinitesimal change of the coordinate of the j -th particle.

Remarks:

- Unlike Newton's $3N$ equations of motion, D'Alembert's principle is just one equation of motion.
- D'Alembert's principle does not involve the forces of constraints in any way. So it is sufficient to specify all the applied forces only.
- It is so general that even a knowledge of the constraint relations is not explicitly required except for determining $\delta \mathbf{r}_j$.
- Its validity extends to all rheonomic and scleronomic systems that are either holonomic, or homogeneous nonholonomic. All such systems for which D'Alembert's principle is valid may be called *D'Alembertian* system.

1.3 Lagrange's Equations of motion

Consider a rheonomic, holonomic system having N particles and k holonomic constraints. Thus it has $n = 3N - K$ degrees of freedom. Choose n generalised coordinates $(q_1, q_2, \dots, q_n)$. These are related to cartesian coordinates through (1.1). Therefore, the cartesian velocity components are

$$\dot{x}_i = \frac{dx_i}{dt} = \frac{\partial x_i}{\partial t} + \sum_{j=1}^n \frac{\partial x_i}{\partial q_j} \dot{q}_j \quad (1.2)$$

Equation (1.2) gives

$$\frac{\partial \dot{x}_i}{\partial \dot{q}_j} = \frac{\partial x_i}{\partial q_j} \quad (1.3)$$

Next consider

$$\begin{aligned}
\frac{d}{dt} \left(\frac{\partial x_i}{\partial q_j} \right) &= \frac{\partial}{\partial t} \left(\frac{\partial x_i}{\partial q_j} \right) + \frac{\partial}{\partial q_l} \left(\frac{\partial x_i}{\partial q_j} \right) \dot{q}_l \\
&= \frac{\partial}{\partial q_j} \left(\frac{\partial x_i}{\partial t} + \frac{\partial x_i}{\partial q_l} \dot{q}_l \right) \\
&= \frac{\partial \dot{x}_i}{\partial q_j}
\end{aligned} \tag{1.4}$$

Now since any infinitesimal virtual displacement is given by

$$\delta x_i = \frac{\partial x_i}{\partial q_j} \delta q_j \tag{1.5}$$

we can begin to transform D'Alembert's principle for this system, which at present is of the cartesian form, that is,

$$\left(m_i \frac{d^2 x_i}{dt^2} - F_i^{(a)} \right) \delta x_i = 0 \tag{1.6}$$

to a form expressed totally in terms of the generalised coordinates. Consider the first term on the LHS of Eq.(1.6):

$$\begin{aligned}
m_i \frac{d^2 x_i}{dt^2} \delta x_i &= m_i \left(\frac{d^2 x_i}{dt^2} \right) \left(\frac{\partial x_i}{\partial q_j} \right) \delta q_j \\
&= \frac{d}{dt} \left(m_i \frac{dx_i}{dt} \frac{\partial x_i}{\partial q_j} \right) \delta q_j - m_i \left(\frac{dx_i}{dt} \right) \frac{d}{dt} \left(\frac{\partial x_i}{\partial q_j} \right) \delta q_j \\
&= \frac{d}{dt} \left(m_i \dot{x}_i \frac{\partial \dot{x}_i}{\partial \dot{q}_j} \right) \delta q_j - \left(m_i \dot{x}_i \frac{\partial \dot{x}_i}{\partial q_j} \right) \delta q_j \\
&= \frac{d}{dt} \left[\frac{\partial}{\partial \dot{q}_j} \left(\frac{1}{2} m_i \dot{x}_i^2 \right) \right] \delta q_j - \frac{\partial}{\partial q_j} \left(\frac{1}{2} m_i \dot{x}_i^2 \right) \delta q_j \\
&= \left[\frac{d}{dt} \left(\frac{\partial T}{\partial \dot{q}_j} \right) - \frac{\partial T}{\partial q_j} \right] \delta q_j
\end{aligned} \tag{1.7}$$

where $T = \frac{1}{2} m_i \dot{x}_i^2$ is the total kinetic energy of the system. The second term on the LHS of D'Alembert's equation (1.6) is

$$\begin{aligned}
-F_i^{(a)} \delta x_i &= -F_i^{(a)} \left(\frac{\partial x_i}{\partial q_j} \right) \delta q_j \\
&= -Q_j \delta q_j
\end{aligned} \tag{1.8}$$

which defines

$$Q_j = F_i^{(a)} \left(\frac{\partial x_i}{\partial q_j} \right) \tag{1.9}$$

as the j -th component of the generalised force. Combining Eqs.(1.6),(1.7) and (1.8), D'Alembert's equation becomes

$$\left[\frac{d}{dt} \left(\frac{\partial T}{\partial \dot{q}_j} \right) - \frac{\partial T}{\partial q_j} - Q_j \right] \delta q_j = 0 \tag{1.10}$$

Since the q_j coordinates are independent, the values of all the δq_j 's can be arbitrary. Therefore, the above equality can hold good if and only if the individual coefficient of each δq_j is separately zero. This implies

$$\frac{d}{dt} \left(\frac{\partial T}{\partial \dot{q}_j} \right) - \frac{\partial T}{\partial q_j} - Q_j = 0 \quad (1.11)$$

for every generalised coordinate q_j and its generalised velocity $\dot{q}_j$. These $n = 3N - k$ differential equations are called Lagrange's equations of motion of the second kind.

Remarks: Lagrange's equations are n second-order ordinary, linear, coupled differential equations. A general solution of these equations involves $2n$ arbitrary constants of integration. The values of these $2n$ constants can only be determined if we know the state of the system (the value of all generalised coordinates and velocities) at some instant of time.

Some important types of Q_j 's:

Case 1: $Q_j = Q_j(q_1, q_2, \dots, q_n, t)$

Here Q_j 's are functions of the q_i 's and t and not of the generalised velocities. In such cases there exists an *ordinary potential energy function* $V(q_1, q_2, \dots, q_n, t)$ such that by definition

$$Q_j = -\frac{\partial V}{\partial q_j}, \quad j = 1, \dots, n \quad (1.12)$$

Lagrange's equations then becomes

$$\frac{d}{dt} \left(\frac{\partial T}{\partial \dot{q}_j} \right) - \frac{\partial T}{\partial q_j} + \frac{\partial V}{\partial q_j} = 0 \quad (1.13)$$

or we can write,

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}_j} \right) - \frac{\partial L}{\partial q_j} = 0 \quad (1.14)$$

where

$$L = T - V \quad (1.15)$$

$L = L(q_1, \dots, q_n, \dot{q}_1, \dots, \dot{q}_n)$ is called by definition the *Lagrangian* of the system. The n equations in Eq.(1.14) are called the *Euler-Lagrange equations of motion*. These are a system of second order coupled linear differential equations in the q_i variables.

Case 2: $Q_j = Q_j(q_1, q_2, \dots, q_n, \dot{q}_1, \dots, \dot{q}_n, t)$, also derivable from a velocity dependent potential energy function in the following way.

When the Q_j 's are derivable from a potential energy function U satisfying the equations by definition,

$$Q_j = \frac{d}{dt} \left(\frac{\partial U}{\partial \dot{q}_j} \right) - \frac{\partial U}{\partial q_j}, \quad j = 1, \dots, n \quad (1.16)$$

where $U = U(q_1, \dots, q_n, \dot{q}_1, \dots, \dot{q}_n, t)$ is called a *generalised potential energy function*, then we get from Lagrange's equations of motion (Eq.(1.11)),

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}_j} \right) - \frac{\partial L}{\partial q_j} = 0, \quad j = 1, \dots, n \quad (1.17)$$

where

$$L = T - U \quad (1.18)$$

is also called the *Lagrangian* of the system as before.

Case 3: Existence of nonpotential forces

It is possible that here exist some nonpotential forces Q'_j apart from the potential force component. In this case Euler-Lagrange's equations take the following most general form

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}_j} \right) - \frac{\partial L}{\partial q_j} = Q'_j, \quad j = 1, \dots, n \quad (1.19)$$

Here $Q'_j = Q'_j(q_1, \dots, q_n, \dot{q}_1, \dots, \dot{q}_n, t)$ in general, and $L = T - V$ or $L = T - U$ as the case may be.

1.4 Total energy

Let us consider a system which satisfies Lagrange's equations in the following form

$$\frac{d}{dt} \left(\frac{\partial T}{\partial \dot{q}_j} \right) - \frac{\partial T}{\partial q_j} = Q_j \quad (1.20)$$

where in general,

$$Q_j = -\frac{\partial V}{\partial q_j} + Q'_j \quad (1.21)$$

V being the ordinary potential energy function and Q'_j the j -th component of the generalised non-potential forces. By definition, the total energy of the system (E) is the sum of its ordinary potential energy and kinetic energy. We want to determine its time dependence. Now, since $T = T(q_1, \dots, q_n, \dot{q}_1, \dots, \dot{q}_n, t)$,

$$\begin{aligned} \frac{dT}{dt} &= \frac{\partial T}{\partial t} + \frac{\partial T}{\partial q_j} \dot{q}_j + \frac{\partial T}{\partial \dot{q}_j} \ddot{q}_j \\ &= \frac{\partial T}{\partial t} + \frac{\partial T}{\partial q_j} \dot{q}_j + \frac{d}{dt} \left(\frac{\partial T}{\partial \dot{q}_j} \dot{q}_j \right) - \frac{d}{dt} \left(\frac{\partial T}{\partial \dot{q}_j} \right) \dot{q}_j \\ &= \frac{\partial T}{\partial t} + \frac{d}{dt} \left(\frac{\partial T}{\partial \dot{q}_j} \dot{q}_j \right) + \left(\frac{\partial V}{\partial q_j} - Q'_j \right) \dot{q}_j \end{aligned} \quad (1.22)$$

using Euler-Lagrange's equation in the last step. Now the kinetic energy function is $T = \frac{1}{2} m_i \dot{x}_i^2$ where for a generally rheonomic system the Cartesian coordinates and velocities are given by

$$x_i = x_i(q_1, q_2, \dots, q_n, t) \quad \text{and} \quad \dot{x}_i = \frac{\partial x_i}{\partial t} + \frac{\partial x_i}{\partial q_j} \dot{q}_j$$

This gives

$$\begin{aligned}
T &= \frac{1}{2}m_i \left(\frac{\partial x_i}{\partial t} + \frac{\partial x_i}{\partial q_j} \dot{q}_j \right)^2 \\
&= \frac{1}{2}m_i \left(\frac{\partial x_i}{\partial t} \right)^2 + m_i \left(\frac{\partial x_i}{\partial t} \right) \left(\frac{\partial x_i}{\partial q_j} \right) \dot{q}_j + \frac{1}{2} \left(\frac{\partial x_i}{\partial q_j} \right) \left(\frac{\partial x_i}{\partial q_l} \right) \dot{q}_j \dot{q}_l \\
&= T_0 + T_1 + T_2
\end{aligned} \tag{1.23}$$

Now substituting (1.23) in (1.22) and using Euler's theorem on homogeneous functions,

$$\begin{aligned}
\frac{dT}{dt} &= \frac{\partial T}{\partial t} + \frac{d}{dt}(T_1 + 2T_2) + \left(\frac{\partial V}{\partial q_j} - Q'_j \right) \dot{q}_j \\
&= 2\frac{dT}{dt} - \frac{d}{dt}(2T_0 + T_1) + \frac{\partial T}{\partial t} + \left(\frac{dV}{dt} - \frac{\partial V}{\partial t} \right) - Q'_j \dot{q}_j
\end{aligned} \tag{1.24}$$

Since by definition $E = T + V$, we have for the time rate of variation of total energy

$$\frac{dE}{dt} = \frac{d}{dt}(T + V) = \frac{d}{dt}(2T_0 + T_1) - \frac{\partial T}{\partial t} + \frac{\partial V}{\partial t} + Q'_j \dot{q}_j \tag{1.25}$$

This is the most general result for rheonomic systems. Some special cases are considered below.

Case (i): Scleronomic Systems

For all such systems, $\frac{\partial T}{\partial t} = 0$, $T_0 = 0$ and $T_1 = 0$. Thus we have

$$\frac{dE}{dt} = \frac{\partial V}{\partial t} + Q'_j \dot{q}_j$$

Suppose V does not explicitly depend on time, then

$$\frac{dE}{dt} = Q'_j \dot{q}_j = P'$$

where P' is the power associated with the nonpotential forces. This equation tells us that nonconservation of total energy is directly associated with the existence of nonpotential forces Q'_j 's, even through the Lagrangian $L = T - V$ does not have any explicit dependence on time in this case.

Case (ii): Conservative Systems

Here we consider systems which are scleronomic with no nonpotential forces. We also assume further that the ordinary potential energy function does not explicitly depend on time. It is only then

$$\frac{dE}{dt} = 0 \quad \text{or} \quad E \equiv T + V = E_0$$

or in words, throughout the motion, the energy of the system is conserved. Such systems are by definition called *conservative systems*. This integral of motion for any conservative system is called the *energy integral*.

1.5 Cyclic (ignorable) coordinates

The Lagrangian of any physical system is generally expected to have explicit dependence on all the generalised coordinates q_i , all the generalised velocities $\dot{q}_i$, and time t , that is,

$$L = L(q_1, \dots, q_n, \dot{q}_1, \dots, \dot{q}_n, t)$$

where n is the total number of generalised coordinates. Due to some reason if some of the generalised coordinates do not appear explicitly in the expression for the Lagrangian, these coordinates are by definition called *cyclic* or *ignorable* coordinates. Any change in these coordinates cannot affect the Lagrangian.

Example: The Lagrangian for a projectile, moving under the earth's approximately constant field of gravity is given by

$$L = \frac{1}{2}mv^2 - mgz = \frac{1}{2}m(\dot{x}^2 + \dot{y}^2 + \dot{z}^2) - mgz$$

where g is the constant acceleration due to gravity. Here (x, y, z) is the set of the Cartesian coordinates of the projectile of mass m , z being the vertical component (upward positive). Obviously, in this example, x and y are the cyclic coordinates.

Theorem: *In absence of any nonpotential forces, the generalised momentum corresponding to any cyclic coordinates is conserved quantity.*

Proof: Let q_i be a cyclic coordinate, then by its definition

$$\frac{\partial L}{\partial q_i} = 0$$

Euler-Lagrange's equation for the same coordinate reduces to

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}_i} \right) = Q'_i \equiv 0$$

Therefore,

$$p_i = \frac{\partial L}{\partial \dot{q}_i} \text{ is a constant of motion}$$

1.6 Generalised potential

By definition from Eq.(1.16), of the generalised force corresponding to the j -th generalised coordinate

$$\begin{aligned} Q_j &= \frac{d}{dt} \left(\frac{\partial U}{\partial \dot{q}_j} \right) - \frac{\partial U}{\partial q_j} \\ &= \frac{\partial^2 U}{\partial \dot{q}_i \partial \dot{q}_j} \ddot{q}_i + \text{terms that do not contain } \ddot{q}_i \end{aligned} \quad (1.26)$$

For a large number of realistic situations, Q_j 's do not explicitly depend on the generalised accelerations, hence for such systems, $\frac{\partial^2 U}{\partial \dot{q}_i \partial \dot{q}_j}$ must vanish. Thus for such systems the generalised potential energy can be given by

$$U = V_i \dot{q}_i + V \quad (1.27)$$

where V_i and the ordinary potential energy function V are all functions of q_i 's and t only. We then have

$$\begin{aligned} Q_j &= \frac{d}{dt} \left(\frac{\partial U}{\partial \dot{q}_j} \right) - \frac{\partial U}{\partial q_j} \\ &= \frac{dV_j}{dt} - \frac{\partial V_i}{\partial q_j} \dot{q}_i - \frac{\partial V}{\partial q_j} \\ &= -\frac{\partial V}{\partial q_j} + \frac{\partial V_j}{\partial t} + \left(\frac{\partial V_j}{\partial q_i} - \frac{\partial V_i}{\partial q_j} \right) \dot{q}_i \end{aligned} \quad (1.28)$$

Suppose further that the coefficients V_j do not depend explicitly on time,

$$Q_j = -\frac{\partial V}{\partial q_j} + \left(\frac{\partial V_j}{\partial q_i} - \frac{\partial V_i}{\partial q_j} \right) \dot{q}_i \quad (1.29)$$

that is, under circumstances, the generalised force can be thought of as a sum of an ordinary type of potential force $\frac{\partial V}{\partial q_j}$ and a gyroscopic force $Q_j = A_{ij} \dot{q}_i$, since the second term in the RHS is antisymmetric with respect to i and j . Thus all such gyroscopic forces can be included in the definition of a generalised potential energy function U , leaving only the dissipative forces to be included in the category of the nonpotential forces Q_j 's.

Summary

Suggestive Reading

1. N.C. Rana, P. S. Joag, *Classical Mechanics*, Mc Graw Hill Education (India) Pvt. Ltd.

Unit 2

Course Structure

- Moving Coordinate System : Coordinate systems with relative translational motions. Rotating coordinate systems. The Coriolis force. Motion on the earth. Effect of Coriolis force on a freely falling particle. Euler's theorem. Euler's equations of motion for a rigid body. Eulerian angles.

Introduction

We have seen earlier that the equations of motion due to Newton, D'Alembert, or Lagrange are valid only if the forces or the Lagrangian is formulated in some inertial frame of reference. However, many systems in nature are found to be naturally rotating or accelerating and it may be more convenient to use the coordinates that directly refer to such non-inertial frames. For example, the earth is rotating about an axis passing through its geographical north pole. So any reference frame that is firmly attached to earth is rotating (non-inertial) frame. Therefore, it is necessary to develop methods for writing down the Lagrangian or the Newtonian equations of motion using the coordinates of a rotating frame. Taking the rotating frames as an example, we have in this unit discussed in detail the consequences of changing over from an inertial to a noninertial frame.

Objective

The objective of this unit is to presents equation of motion in moving coordinate system, an advance formalism for studying various problems in mechanics. Moving coordinate systems are important because, no material body is at absolute rest. As we know, even galaxies are not stationary. Therefore, a coordinate frame at absolute rest is hypothetical, hypothesized by Newton, where his laws of motion hold. In reality, we have the moving frames, prime example being Earth itself. We therefore need to know how the Newton's laws operate in a moving frame like a rotating frame (e.g. Earth). Studying motion in moving coordinate system can provide a much cleaner way of understanding some practical systems than Newtonian, D'Alembert or Lagrange mechanics. In this unit, we will learn the basics of rotating frame of reference.

2 Moving Coordinate Systems

2.1 Coordinate systems with relative translational motion

When a body experiences motion, the motion itself takes place with respect to a coordinate system. The question arises as to whether the coordinate system in reference to which motion takes place is itself accelerating, is at rest, or is moving at constant velocity. Thus we can have an *inertial* coordinate system that is defined to be one in which Newton's first law of motion applies. That is, it is one that, although it could be moving at constant velocity, it does not experience an acceleration. We

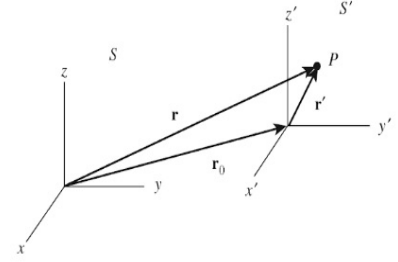
can also have a *noninertial* or accelerated coordinate system in which Newton's first law of motion does not apply; rather, because the reference frame is accelerated, inertial or fictitious forces arise, as will be discussed shortly. Here, we consider translational motion of bodies with respect to reference frames. One frame is the inertial reference frame or the S frame of reference with coordinates (x, y, z) , and the other is a non-inertial reference frame or the S' frame of reference with primed coordinates (x', y', z') . In the following figure, we consider a particle of mass m located at point P whose position is $\mathbf{r}'$ with respect to the S' frame.

We see that the S' frame is located at position $\mathbf{r}_0$ with respect to the S frame. Thus, we can find the particle's position with respect to the S frame as

$$\mathbf{r} = \mathbf{r}_0 + \mathbf{r}'. \quad (2.1)$$

The particle's velocity as seen from the S frame is obtained by taking the time derivatives of this, or

$$\mathbf{v} = \mathbf{v}_0 + \mathbf{v}'. \quad (2.2)$$



where $\mathbf{v}_0$ is the velocity of the S' frame with respect to the S frame, and $\mathbf{v}'$ is the velocity of the particle with respect to the S' frame. Taking the time derivative of (2.2) we have the particle's acceleration with respect to the S frame is the sum of the accelerations of the S' frame with respect to the S frame, and the acceleration of the particle with respect to the S' frame; that is,

$$\mathbf{a} = \mathbf{a}_0 + \mathbf{a}'. \quad (2.3)$$

Because the S frame is an inertial frame, Newton's second law of motion holds, so that we can write

$$\mathbf{f} = m\mathbf{a}, \quad (2.4)$$

which is what an observer in the S frame sees. From (2.3), an observer in the S' frame will see

$$\mathbf{f}' = m\mathbf{a}' = m\mathbf{a} - m\mathbf{a}_0. \quad (2.5)$$

The term $m\mathbf{a}_0$ stems from the fact that the S' frame is an accelerated coordinate system. Thus, the S' frame observer has to add an extra term, the so-called *fictitious* force, $\mathbf{f}_{fict} \equiv -m\mathbf{a}_0$, in order to find agreement with the inertial frame of reference observer. In the event that S' frame moves at constant velocity, $\mathbf{a}_0 \rightarrow 0$ and $\mathbf{f}' = \mathbf{f}$, without mention being made to the fictitious force. Newton's second law, in this case, becomes identical in each frame.

2.2 Rotating coordinate systems

Let there be a fixed frame S with $(\hat{i}, \hat{j}, \hat{k})$ as the fixed unit vectors forming a rectangular cartesian triad and the frame S' with triad $(\hat{i}', \hat{j}', \hat{k}')$, originally coincident with S but rotating with velocity $\vec{\omega} (= \hat{n}\omega)$ about their common origin O .

Physically any vector $\vec{G}$ is same in both the frames. But the components of $\vec{G}$ in S and S' are different because the unit vectors in both frames are pointing differently. Hence

$$G|_{fixed} = G_1\hat{i} + G_2\hat{j} + G_3\hat{k} = G'_1\hat{i}' + G'_2\hat{j}' + G'_3\hat{k}' = G|_{rot} \quad (2.6)$$

The time derivatives of $\vec{G}_{fixed}$ and $\vec{G}_{rot}$ must also be equal, because they are also one and the same physical vector. Since the unit vector $(\hat{i}, \hat{j}, \hat{k})$ of the fixed frame S do not change with time, we can write

$$\left[\frac{d\vec{G}}{dt} \right]_{fixed} = \frac{dG_1}{dt}\hat{i} + \frac{dG_2}{dt}\hat{j} + \frac{dG_3}{dt}\hat{k} \quad (2.7)$$

which must be equal to $(d[\vec{G}_{rot}]/dt)$ given by,

$$\left[\frac{d\vec{G}}{dt} \right]_{fixed} = \frac{d}{dt}[\vec{G}_{rot}] = \frac{dG'_1}{dt}\hat{i}' + \frac{dG'_2}{dt}\hat{j}' + \frac{dG'_3}{dt}\hat{k}' + G'_1\frac{d\hat{i}'}{dt} + G'_2\frac{d\hat{j}'}{dt} + G'_3\frac{d\hat{k}'}{dt} \quad (2.8)$$

One may denote the first three terms by $[d\vec{G}]/dt|_{rot}$ because this is precisely the time derivative of $\vec{G}$ as measured by an observer in the rotating frame S' .

An arbitrary small rotation $\Delta\theta$ about some direction $\hat{n}$ over an time interval Δt passing through the origin brings about a change in any vector $\vec{A}$ given by

$$\frac{\Delta\vec{A}}{\Delta t} = \hat{n} \frac{\Delta\theta}{\Delta t} \times \vec{A}$$

Taking the limit $\Delta t \rightarrow 0$,

$$\frac{d\vec{A}}{dt} = \omega\hat{n} \times \vec{A} \quad (2.9)$$

Now using Eq.(2.9), Eq.(2.8) may be written as

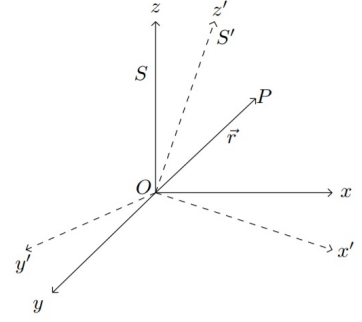
$$\begin{aligned} \left[\frac{d\vec{G}}{dt} \right]_{fixed} &= \left[\frac{d\vec{G}}{dt} \right]_{rot} + \vec{\omega} \times (G'_1\hat{i}' + G'_2\hat{j}' + G'_3\hat{k}') \\ &= \left[\frac{d\vec{G}}{dt} \right]_{rot} + \vec{\omega} \times \vec{G} \end{aligned} \quad (2.10)$$

Thus we have an operator identity

$$\left[\frac{d}{dt} \right]_{fixed} = \left[\frac{d}{dt} \right]_{rot} + \vec{\omega} \times \quad (2.11)$$

Putting $\vec{G} = \vec{r}$ in Eq.(2.10), the position vector of a particle, we get

$$\begin{aligned} \left[\frac{d\vec{r}}{dt} \right]_{fixed} &= \left[\frac{d\vec{r}}{dt} \right]_{rot} + \vec{\omega} \times \vec{r} \\ \Rightarrow [\vec{v}]_{fixed} &= [\vec{v}]_{rot} + \vec{\omega} \times \vec{r} \end{aligned} \quad (2.12)$$



Denoting $[\vec{v}]_{fixed} = \vec{v}_0$ and $[\vec{v}]_{rot} = \vec{v}$ we write

$$\vec{v}_0 = \vec{v} + \vec{\omega} \times \vec{r} \quad (2.13)$$

Again putting $\vec{G} = \vec{v}_0$ in Eq.(2.10),

$$\begin{aligned} \left[\frac{d\vec{v}_0}{dt} \right]_{fixed} &= \left[\frac{d\vec{v}_0}{dt} \right]_{rot} + \vec{\omega} \times \vec{v}_0 \\ &= \frac{d}{dt}(\vec{v} + \vec{\omega} \times \vec{r}) + \vec{\omega} \times (\vec{v} + \vec{\omega} \times \vec{r}) \\ &= \frac{d\vec{v}}{dt} + \dot{\vec{\omega}} \times \vec{r} + 2\vec{\omega} \times \vec{v} + \vec{\omega} \times (\vec{\omega} \times \vec{r}) \end{aligned}$$

where $d\vec{v}/dt$ is the acceleration as measured in the S' frame. If we are to consider the motion of a particle of mass m , its equations of motion from the point of view of an observer who is at rest with respect to S' can be expressed as

$$m \frac{d\vec{v}}{dt} = m \frac{dv_0}{dt} + m\vec{r} \times \dot{\vec{\omega}} + 2m\vec{v} \times \vec{\omega} + m(\vec{\omega} \times \vec{r}) \times \vec{\omega} \quad (2.14)$$

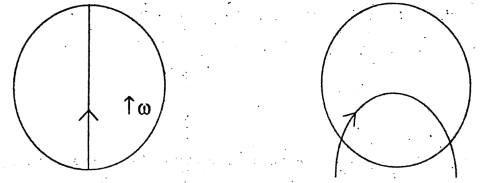
The first term on RHS of Eq.(2.14) corresponds to the product of mass and the inertial acceleration in the fixed frame, that is, the actual forces applied on the system, or the true external forces. Other terms correspond to the pseudoforces or the fictitious or inertial forces that arise due to the fact that the rotating frame is non-inertial. They are called, from left to right, *Euler force* ($m\vec{r} \times \dot{\vec{\omega}}$), *Coriolis force* ($2m\vec{v} \times \vec{\omega}$) and *centrifugal force* ($m(\vec{\omega} \times \vec{r}) \times \vec{\omega}$) respectively. These forces appear to exist only in a rotating frame of reference.

Note:

- The Euler force can be neglected for a uniform rotation $\frac{d\vec{\omega}}{dt} = 0$
- Coriolis term implies that it is always perpendicular to the direction of velocity and so it can never change the speed of a particle.

2.3 Motion of Earth

Consider a flat rotating disc. For a particle moving (See Fig.1) across a disc under no forces, an inertial observer (i.e., the one who is in a fixed frame) will see it moving across a straight line according to Newton's law. However, in view of the fact that the disc is rotating, an observer stationed on the disc will view the particle taking a curved track due to the Coriolis force operating in a direction perpendicular to the motion of the particle (See Fig.2). Note that the effect of the Coriolis force is to bend the path of the particle to the right in the Northern Hemisphere and to the left in the Southern Hemisphere.



Assuming ω to be constant, Eq.(2.14) may be rewritten as

$$m \left(\frac{d^2 \vec{r}}{dt^2} \right)_{fixed} = m \left(\frac{d^2 \vec{r}}{dt^2} \right)_{rot} + 2m\vec{\omega} \times \left(\frac{d\vec{r}}{dt} \right)_{rot} + m\vec{\omega} \times (\vec{\omega} \times \vec{r}) \quad (2.15)$$

For a particle moving under gravitation (i.e., the attractive force guided by Newton's law of gravitation) and also subjected to an additional force $\vec{F}$, the equation of motion is

$$m \left(\frac{d^2 \vec{r}}{dt^2} \right)_{fixed} = m\vec{g} + \vec{F} \quad (2.16)$$

Thus from Eqs.(2.15) and (2.16), we find

$$\left(\frac{d^2 \vec{r}}{dt^2} \right)_{rot} = m\vec{g} + \vec{F} - 2m\vec{\omega} \times \left(\frac{d\vec{r}}{dt} \right)_{rot} - m\vec{\omega} \times (\vec{\omega} \times \vec{r}) \quad (2.17)$$

In the laboratory when we measure the acceleration due to gravity what we determine is actually $\vec{g}^*$, the effective gravitational acceleration given by

$$\vec{g}^* = \vec{g} - \vec{\omega} \times (\vec{\omega} \times \vec{r}) \quad (2.18)$$

The right-hand-side of Eq.(2.18) is a combination of gravitational and centrifugal forces. Note that the horizontal and vertical components of $\vec{g}^*$ are $g_{hor}^* = \omega^2 r \sin \theta \cos \theta$ and $g_{ver}^* = g - \omega^2 r \sin \theta$. At the pole $g^* = g$ while on the equator $g^* = g - \omega^2 r$. Substituting Eq.(2.18) in Eq.(2.17) and writing $\vec{g}$ in place of $\vec{g}^*$ we get,

$$m \frac{d^2 \vec{r}}{dt^2} = m \vec{g} + \vec{F} - 2m \vec{\omega} \times \frac{d\vec{r}}{dt} \quad (2.19)$$

where the suffix 'rot' has been dropped. The components of $\vec{\omega}$ being $(0, \omega \sin \theta, \omega \cos \theta)$, where θ is the angle between the direction of earth's axis and $\vec{g}^*$, we find the Coriolis force to be given by

$$2m \vec{\omega} \times \frac{d\vec{r}}{dt} = 2m\omega (\dot{y} \cos \theta - \dot{z} \sin \theta, -\dot{x} \cos \theta, \dot{x} \sin \theta) \quad (2.20)$$

where $d\vec{r}/dt \equiv (\dot{x}, \dot{y}, \dot{z})$.

2.4 Foucault's pendulum

A useful device for observing the effects of the Coriolis force is the so called Foucault's pendulum. It is perfectly symmetric and designed to swing freely in any direction. Because of its symmetric nature, the periods of oscillation of a Foucault's pendulum in all directions are equal.

Neglecting the vertical components of the Coriolis force for it is very small compared to g , the equations of motion in the x and y directions are

$$\begin{aligned} \ddot{x} &= -\frac{g}{l}x + 2\omega\dot{y} \cos \theta \\ \ddot{y} &= -\frac{g}{l}y + 2\omega\dot{x} \cos \theta \end{aligned} \quad (2.21)$$

where we have used Eq.(2.19) and Eq.(2.20). One can see that the right-hand-sides of the Eq.(2.21) carry the contributions from the Coriolis acceleration in addition to the usual ones of the simple harmonic motion. In writing down Eq.(2.21) we have also assumed that for small amplitude, the motion is nearly horizontal, i.e., $\dot{z} \approx 0$.

A straightforward way to solve Eq.(2.21) is to set $r = x + iy$ resulting in

$$\ddot{r} + 2ik\dot{r} + \omega_0^2 r = 0 \quad (2.22)$$

where $k = \omega \cos\theta$, $\omega_0^2 = g/l$. For $r = Ae^{\lambda t}$ where A and λ are constants we find $\lambda^2 + 2ik\lambda + \omega_0^2 = 0$ whose solution are

$$\lambda = -ik \pm i\omega'$$

where $\omega'^2 = \omega_0^2 + k^2$. The solution of Eq.(2.22) can thus be written as

$$r = C_1 e^{-i(k-\omega')t} + C_2 e^{-i(k+\omega')t} \quad (2.23)$$

An interesting particular case of Eq.(2.23) corresponds to the choice of the integration constants $C_1 = C_2 = \frac{1}{2}a$:

$$\begin{aligned} r &= a e^{-ikt} \cos \omega' t \\ \text{i.e., } x &= a \cos kt \cos \omega' t, \quad y = -a \sin kt \cos \omega' t \end{aligned} \quad (2.24)$$

The interpretation of Eq.(2.24) is that initially (i.e., at $t = 0$) the oscillation is in the x -direction. However, with passage of time, the amplitude of the y -coordinate grows at the expense of the x -coordinate that dampens. Overall the solution depicts an oscillation (of amplitude a) in a plane that is rotating with an angular velocity $-k$.

Example: Find the deflection of a freely falling body from the vertical caused by earth's rotation.

Solution: If a particle is dropped from rest of a height h above the ground then the motion is described by

$$x = 0, \quad y = 0, \quad z = h - \frac{1}{2}gt^2$$

where the effect of the Coriolis force is neglected. To find the effect of the Coriolis force to first order in ω , we substitute for $\frac{d\vec{r}}{dt}$ in Eq.(2.19) the zero-order expressions namely

$$\dot{x} = 0, \quad \dot{y} = 0, \quad \dot{z} = -gt$$

Then using Eq.(2.20), Eq.(2.19) reads component wise

$$m\ddot{x} = 2m\omega g t \sin \theta, \quad m\ddot{y} = 0, \quad m\ddot{z} = -gt$$

We thus find that the particle will hit the ground $z = 0$ (i.e., $t = \sqrt{\frac{2h}{g}}$ from $z = h - \frac{1}{2}gt^2$) in the eastern direction at a distance of

$$\begin{aligned} x &= \omega g \frac{1}{3} \sin \theta \left(\frac{2h}{g} \right)^{3/2} \\ &= \frac{1}{3} \omega \left(\frac{8h^3}{g} \right)^{1/2} \sin \theta \end{aligned}$$

where we have integrated the above expression for $\ddot{x}$ and used $x = 0$, $\dot{x} = 0$ at $t = 0$. We conclude that there will be an easterly deviation from the vertical of amount $\frac{1}{3}\omega \sin \theta \sqrt{\frac{8h^3}{g}}$.

2.5 Rigid Body

A rigid body is regarded as a system of many (at least three) non-collinear particles whose positions relative to one another remain fixed. i.e., distance between any two of them remains constant throughout the motion. The internal forces holding the particles at fixed distances from one another are known as forces of constraint. These forces of constraint obey the Newton's third law of motion

2.6 Euler's theorem

Statement: A rotation of a rigid body about a fixed point of the body is equivalent to a rotation about a line which passes through the point.

The line referred to is called the *instantaneous axis of rotation*.

2.7 Euler's equations of motion for a rigid body

A set of equations governing the rotation of a rigid body referred to its own axis are known as Euler's equations of motion of a rigid body with one point fixed.

- Obtain the Euler's equations of motion of a rigid body when one point of the body remains fixed.

Consider a rigid of which one point is fixed. Hence translational motion of the body is absent and the body rotates about an axis passes through the fixed point. The rotation of the body takes place under the action of torque acting on it. Thus the equation of the rotational motion of the body in a fixed frame is given by

$$\begin{aligned} \text{Torque} &= \text{rate of change of angular momentum} \\ \Rightarrow \vec{N} &= \left(\frac{d\vec{L}}{dt} \right)_{fixed} \end{aligned}$$

However, we know

$$\left(\frac{d}{dt} \right)_{fixed} = \left(\frac{d}{dt} \right)_{rot} + \vec{\omega} \times$$

Therefore the equation of motion of the rigid body becomes

$$N = \left(\frac{d\vec{L}}{dt} \right)_{fixed} = \left(\frac{d\vec{L}}{dt} \right)_{rot} + \vec{\omega} \times \vec{L} \quad (2.25)$$

where ω is the angular velocity of the body and L is the angular momentum and is given by

$$\vec{L} = I\vec{\omega} \quad (2.26)$$

I is the moment of inertia tensor and is constant with respect to the body frame of reference. We choose the principal axis of the body with respect to which the off diagonal elements of the moment of inertia tensor vanish and only the diagonal elements remain in the expression for $\vec{I}$. If I_1, I_2, I_3 are the principal moments of inertia then we have

$$I = \begin{bmatrix} I_1 & 0 & 0 \\ 0 & I_2 & 0 \\ 0 & 0 & I_3 \end{bmatrix}$$

In this case the expression for angular momentum Eq.(2.26) becomes

$$L = I_1\omega_x\hat{i} + I_2\omega_y\hat{j} + I_3\omega_z\hat{k}, \quad (2.27)$$

Differentiating with respect to time t in the rotating frame we get

$$\left(\frac{dL}{dt}\right)_{rot} = I_1\dot{\omega}_x\hat{i} + I_2\dot{\omega}_y\hat{j} + I_3\dot{\omega}_z\hat{k} \quad (2.28)$$

Also we find the value of

$$\begin{aligned} \vec{\omega} \times \vec{L} &= \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ \omega_x & \omega_y & \omega_z \\ I_1\omega_x & I_2\omega_y & I_3\omega_z \end{vmatrix} \\ &= \hat{i}(I_3 - I_2)\omega_y\omega_z - \hat{j}(I_3 - I_1)\omega_x\omega_z + \hat{k}(I_2 - I_1)\omega_x\omega_y \end{aligned}$$

Hence the equation of rotational motion of the rigid body becomes

$$\begin{aligned} N &= \left(\frac{d\vec{L}}{dt}\right)_{fixed} = \left(\frac{d\vec{L}}{dt}\right)_{rot} + \vec{\omega} \times \vec{L} \\ &= \hat{i}[I_1\dot{\omega}_x + (I_3 - I_2)\omega_y\omega_z] + \hat{j}[I_2\dot{\omega}_y + (I_1 - I_3)\omega_x\omega_z] + \hat{k}[I_3\dot{\omega}_z + (I_2 - I_1)\omega_x\omega_y] \end{aligned}$$

If torque is expressed in term of its components as

$$\vec{N} = N_x\hat{i} + N_y\hat{j} + N_z\hat{k} \quad (2.29)$$

then on equating the corresponding components on both sides of the above equations we obtain

$$\begin{aligned} N_x &= I_1\dot{\omega}_x + (I_3 - I_2)\omega_y\omega_z, \\ N_y &= I_2\dot{\omega}_y + (I_1 - I_3)\omega_x\omega_z, \\ N_z &= I_3\dot{\omega}_z + (I_2 - I_1)\omega_x\omega_y, \end{aligned} \quad (2.30)$$

These are the required Euler's equations of motion of the rigid body with one point fixed.

Note : In the case of torque free motion of a rigid body, equations (2.30) reduce to

$$\begin{aligned} I_1\dot{\omega}_x &= (I_2 - I_3)\omega_y\omega_z, \\ I_2\dot{\omega}_y &= (I_3 - I_1)\omega_x\omega_z, \\ I_3\dot{\omega}_z &= (I_1 - I_2)\omega_x\omega_y, \end{aligned} \quad (2.31)$$

Example: If the rigid body with one point fixed rotates about the principal axis of the body, then show that (i) kinetic energy of the body and (ii) the magnitude of the angular momentum are constants throughout the motion.

Solution: The kinetic energy of a rigid body with one point fixed is given by

$$T = \frac{1}{2} (I_{xx}\omega_x^2 + I_{yy}\omega_y^2 + I_{zz}\omega_z^2 + 2I_{xy}\omega_x\omega_y + 2I_{yz}\omega_y\omega_z + 2I_{zx}\omega_z\omega_x)$$

where I_{xx}, I_{yy}, I_{zz} and I_{xy}, I_{yz}, I_{zx} are the moments of inertia and product of inertia about the coordinate axes respectively. If the body rotates about the principal axis of the body then the products of inertia tensors are zero and the moments of inertia tensors are constants. In this case the kinetic energy becomes

$$T = \frac{1}{2} (I_1\omega_x^2 + I_2\omega_y^2 + I_3\omega_z^2)$$

where $I_1 = I_{xx}, I_2 = I_{yy}, I_3 = I_{zz}$. In the absence of torque the Euler's equations of motion of the rigid body are given by

$$\begin{aligned} I_1\dot{\omega}_x + (I_3 - I_2)\omega_y\omega_z &= 0 \\ I_2\dot{\omega}_y + (I_1 - I_3)\omega_x\omega_z &= 0 \\ I_3\dot{\omega}_z + (I_2 - I_1)\omega_x\omega_y &= 0 \end{aligned} \tag{2.32}$$

(i) Multiply each above equation by $\omega_x, \omega_y, \omega_z$ respectively and adding we get

$$I_1\omega_x\dot{\omega}_x + I_2\omega_y\dot{\omega}_y + I_3\omega_z\dot{\omega}_z = 0$$

We write this as

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} (I_1\omega_x^2 + I_2\omega_y^2 + I_3\omega_z^2) &= 0 \\ \Rightarrow \frac{dT}{dt} &= 0 \\ \Rightarrow T &= \frac{1}{2} (I_1\omega_x^2 + I_2\omega_y^2 + I_3\omega_z^2) = \text{constant} \end{aligned}$$

(ii) Now we claim that the magnitude of the angular momentum is constant.

The moment of inertia tensor with respect to the principal axis of the body, is given by

$$\vec{I} = \begin{bmatrix} I_1 & 0 & 0 \\ 0 & I_2 & 0 \\ 0 & 0 & I_3 \end{bmatrix}$$

where I_1, I_2, I_3 are constant with respect to the principal axis of the body. Hence the expression for angular momentum becomes

$$\begin{aligned} L &= I_1\omega_x\hat{i} + I_2\omega_y\hat{j} + I_3\omega_z\hat{k} \\ \Rightarrow L^2 &= L \cdot L = I_1^2\omega_x^2 + I_2^2\omega_y^2 + I_3^2\omega_z^2 \end{aligned}$$

Multiply each equation of (2.32) by $I_1\omega_x$, $I_2\omega_y$, $I_3\omega_z$ respectively and adding we get

$$I_1^2\omega_x\dot{\omega}_x + I_2^2\omega_y\dot{\omega}_y + I_3^2\omega_z\dot{\omega}_z = 0$$

Thus we write

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} (I_1^2\omega_x^2 + I_2^2\omega_y^2 + I_3^2\omega_z^2) &= 0 \\ \Rightarrow \frac{d}{dt} (L^2) &= 0 \\ \Rightarrow L^2 &= \text{constant} \end{aligned}$$

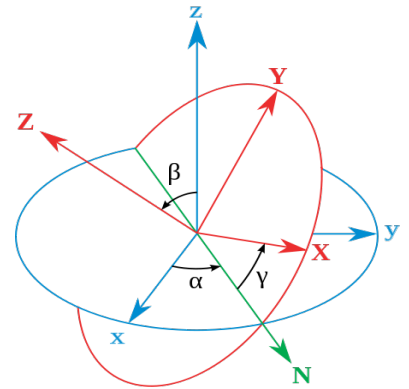
This shows that the magnitude of the angular momentum of the rigid body is constant.

2.8 Eulerian angles

The Euler angles are three angles introduced by Leonhard Euler to describe the orientation of a rigid body with respect to a fixed coordinate system. They can also represent the orientation of a mobile frame of reference in physics or the orientation of a general basis in 3-dimensional linear algebra. Any orientation can be achieved by composing three elemental rotations, i.e. rotations about the axes of a coordinate system. Euler angles can be defined by three of these rotations. They can also be defined by elemental geometry and the geometrical definition demonstrates that three rotations are always sufficient to reach any frame.

Euler angles are typically denoted as α, β, γ or ϕ, θ, ψ . Different authors may use different sets of rotation axes to define Euler angles, or different names for the same angles. Therefore, any discussion employing Euler angles should always be preceded by their definition.

The axes of the original frame are denoted as x, y, z and the axes of the rotated frame as X, Y, Z . The geometrical definition (sometimes referred to as static) begins by defining the line of nodes as the intersection of the planes xy and XY (it can also be defined as the common perpendicular to the axes z and Z and then written as the vector product $N = z \times Z$). Using it, the three Euler angles can be defined as follows:



- α (or φ) is the angle between the x -axis and the N axis (x -convention - it could also be defined between y and N , called y -convention).

- β (or θ) is the angle between the z axis and the Z axis.

- γ (or ψ) is the angle between the N axis and the X axis (x -convention).

Euler angles between two reference frames are defined only if both frames have the same handedness.

Summery

In order to go from an inertial to a noninertial frame, the initial position vectors r_i , velocities v_i in the Lagrangian scheme and r_i , v_i and a_i the acceleration in the Newtonian scheme are to be expressed in terms of the respective quantities in the noninertial frame. For frames rotating about a fixed pint, the position vectors like any other vector in the two frames are identical, except that the components of the vectors along the respective coordinate axes (Cartesian) would be different. But if we consider any given vector, say $\mathbf{G}$, and then take the time derivative of the respective components in the two frames rather than the time derivative of the vector as a whole, the answer would differ by a term $\omega \times \mathbf{G}$. It is this extra term that finally brings about differences in the expressions for the acceleration in the two frames. The effect of the Coriolis force due to rotation of the earth is studied.

Suggestive Reading

1. Javier E. Hasbun, *Classical Mechanics with MATLAB Applications*, Jones and Bartlett Publishers

Unit 3

Course Structure

- Variational Principle : Calculus of variations and its applications in shortest distance, minimum surface of revolution, Brachistochrone problem, geodesic. Hamilton's principle. Lagrange's undetermined multipliers. Hamilton's equations of motion.

Introduction

The Euler-Lagrangian approach to classical mechanics stems from a deep philosophical belief that the laws of nature are based on a principle of economy. That is, the physical universe follows paths through space and time that are based on extrema principles. The standard Lagrangian L is defined as the difference between the kinetic and potential energy, that is $L = K - U$. The laws of classical mechanics can be expressed in terms of Hamilton's variational principle which states that the motion of the system between the initial time t_1 and final time t_2 follows a path that minimizes the scalar action integral S defined as the time integral of the Lagrangian.

$$S = \int_{t_1}^{t_2} L dt$$

Objective

The calculus of variations provides the mathematics required to determine the path that minimizes the action integral. This variational approach is both elegant and beautiful, and has withstood the rigors of experimental confirmation. In fact, not only is it an exceedingly powerful alternative approach to the intuitive Newtonian approach to mechanics, but Hamilton's variational principle now is recognized to be more fundamental than Newton's Laws of Motion. The Lagrangian and Hamiltonian variational approaches to mechanics are the only approaches that can handle the Theory of Relativity, statistical mechanics, and the dichotomy of philosophical approaches to quantum physics

3 Variational Principle

Consider first the problem in an essentially one-dimensional form: We have a function $f(x, y, \dot{y})$ defined on a path $y = y(x)$ between two values x_1 and x_2 , where $\dot{y}$ is the derivative of y with respect to x . We wish to find a particular path $y(x)$ such that the line integral J of the function f between x_1 and x_2 ,

$$\dot{y} \equiv \frac{dy}{dx}, \quad J = \int_{x_1}^{x_2} f(x, y, \dot{y}) dx, \quad (3.1)$$

has a stationary value relative to paths differing infinitesimally from the correct function $y(x)$. The variable x here plays the role of the parameter t , and we consider only such varied paths for which

$y(x_1) = y_1$, $y(x_2) = y_2$ (cf.Fig.3.1). Since J must have a stationary value for the correct path relative to any neighbouring path, the variation must be zero relative some particular set of neighbouring paths labelled by an infinitesimal parameter α . For any such parametric family of curves, J in Eq.(3.1) is also a function of α :

$$J(\alpha) = \int_{x_1}^{x_2} f(x, y(x, \alpha), \dot{y}(x, \alpha)) dx. \quad (3.2)$$

and the condition for obtaining a stationary point is the familiar one that

$$\left(\frac{dJ}{d\alpha} \right)_{\alpha=0} = 0 \quad (3.3)$$

But the usual methods of differentiating under the integral sign, we find that

$$\frac{dJ}{d\alpha} = \int_{x_1}^{x_2} \left(\frac{\partial f}{\partial y} \frac{\partial y}{\partial \alpha} + \frac{\partial f}{\partial \dot{y}} \frac{\partial \dot{y}}{\partial \alpha} \right) dx \quad (3.4)$$

Consider the second of these integrals,

$$\int_{x_1}^{x_2} \frac{\partial f}{\partial \dot{y}} \frac{\partial \dot{y}}{\partial \alpha} dx = \int_{x_1}^{x_2} \frac{\partial f}{\partial \dot{y}} \frac{\partial^2 y}{\partial x \partial \alpha} dx \quad (3.5)$$

Integrating by parts, the integral becomes

$$\int_{x_1}^{x_2} \frac{\partial f}{\partial \dot{y}} \frac{\partial^2 y}{\partial x \partial \alpha} dx = \left. \frac{\partial f}{\partial \dot{y}} \frac{\partial y}{\partial \alpha} \right|_{x_1}^{x_2} - \int_{x_1}^{x_2} \frac{d}{dx} \left(\frac{\partial f}{\partial \dot{y}} \right) \frac{\partial y}{\partial \alpha} dx. \quad (3.6)$$

The conditions on all the varied curves are that they pass through the points (x_1, y_1) , (x_2, y_2) , and hence the partial derivative of y with respect to α at x_1 and x_2 must vanish. Therefore, the first term of Eq.(3.6) vanishes and Eq.(3.5) reduces to

$$\frac{dJ}{d\alpha} = \int_{x_1}^{x_2} \left(\frac{\partial f}{\partial y} - \frac{d}{dx} \frac{\partial f}{\partial \dot{y}} \right) \frac{\partial y}{\partial \alpha} dx$$

Now using *fundamental lemma* of the calculus of variations, it therefore follows that J can have a stationary value only if

$$\frac{\partial f}{\partial y} - \frac{d}{dx} \left(\frac{\partial f}{\partial \dot{y}} \right) = 0 \quad (3.7)$$

This Eq.(3.7) is known as *Euler's equation*. Some simple examples of the application of Eq.(3.7) may now be considered.

3.1 Shortest distance between two points in a plane

An element of length in a plane is

$$ds = \sqrt{dx^2 + dy^2}$$

and the total length of any curve going between points 1 and 2 is

$$I = \int_1^2 ds = \int_{x_1}^{x_2} \sqrt{1 + \left(\frac{dy}{dx}\right)^2}$$

The condition that the curve be the shortest path is that I be a minimum. This is an example of the extremum problem as expressed by Eq., with

$$f = \sqrt{1 + \dot{y}^2}$$

Substituting in Eq. with

$$\frac{\partial f}{\partial y} = 0, \quad \frac{\partial f}{\partial \dot{y}} = \frac{\dot{y}}{\sqrt{1 + \dot{y}^2}}$$

we have

$$\frac{d}{dx} \left(\frac{\dot{y}}{\sqrt{1 + \dot{y}^2}} \right) = 0 \Rightarrow \frac{\dot{y}}{\sqrt{1 + \dot{y}^2}} = c$$

where c is constant. This solution can be valid only if $\dot{y} = a$, where a is a constant related to c by

$$a = \frac{c}{\sqrt{1 - c^2}}$$

But this is clearly the equation of a straight line, $y = ax + b$, where b is another constant of integration. Strictly speaking, the straight line has only been proved to be an extremum path, but for this problem it is obviously also a minimum. The constants of integration, a and b , are determined by the condition that the curve pass through two end points, (x_1, y_1) , (x_2, y_2) .

In a similar fashion we can obtain the shortest distance between two points on a sphere, by setting up the arc length on the surface of the sphere in terms of the angle coordinates of position on the sphere. In general, curves that give the shortest distance between two points on a given surface are called the *geodesics* of the surface.

3.2 Minimum surface of revolution

Suppose we form a surface of revolution by taking some curve passing between two fixed end points (x_1, y_1) and (x_2, y_2) defining the xy - plane, and revolving it about the y - axis (cf. Fig.1). The problem then is to find that curve for which the surface area is minimum. The area of a strip of the surface is $2\pi x ds = 2\pi x \sqrt{1 + \dot{y}^2} dx$, and total area is

$$2\pi \int_1^2 x \sqrt{1 + \dot{y}^2} dx.$$

The extremum of this integral is again given by Eq. where

$$f = x \sqrt{1 + \dot{y}^2} \quad \text{and} \quad \frac{\partial f}{\partial y} = 0, \quad \frac{\partial f}{\partial \dot{y}} = \frac{x \dot{y}}{\sqrt{1 + \dot{y}^2}}$$

Eq. becomes in this case

$$\frac{d}{dx} \left(\frac{x\dot{y}}{\sqrt{1+\dot{y}^2}} \right) = 0 \Rightarrow \frac{x\dot{y}}{\sqrt{1+\dot{y}^2}} = c$$

where a is some constant of integration clearly smaller than the minimum value of x . Squaring the above equation and factoring terms, we have

$$\dot{y}^2(x^2 - a^2) = a^2,$$

or solving

$$\frac{dy}{dx} = \frac{a}{\sqrt{x^2 - a^2}}.$$

The general solution of this differential equation, in light of the nature of a , is

$$\begin{aligned} y &= a \int \frac{dx}{\sqrt{x^2 - a^2}} + b = a \cosh^{-1} \frac{x}{a} + b \\ \Rightarrow x &= a \cosh \frac{y - b}{a} \end{aligned}$$

which is the equation of catenary. Again the two constants of integration, a and b , are determined in principle by the requirements that the curve pass through the two given end points, as shown in Fig.3.3.

Curves satisfying the preceding equation all scale as x/a and y/a with one independent parameter b/a . This suggests that when the solutions are examined in detail they turn out to be a great deal more complicated than these considerations suggest. For some pairs of end points, unique constants of integration a and b can be found. But for other end points, it is possible to draw two different catenary curves through the end points, while for additional cases no possible values can be found for a and b .

3.3 Brachistochrone problem

This well-known problem is to find the curve joining two points (x_0, y_0) and (x_1, y_1) , along which a particle falling from rest under the influence of gravity travels from the higher to the lower point in the least time.

If v is the speed along the curve, then the time required to fall an arc length ds is ds/v , and the problem is to find a minimum of the integral

$$t_{12} = \int_{x_0}^{x_1} \frac{ds}{v}.$$

If y is measured down from the initial point of release, the conservation theorem for the energy of the particle can be written as

$$\frac{1}{2}mv^2 = mgy \Rightarrow v = \sqrt{2gy}$$

Then the expression for t_{12} becomes

$$t_{12} = \int_{x_0}^{x_1} \frac{\sqrt{1 + \dot{y}^2}}{\sqrt{2gy}} dx,$$

and f is defined as

$$f = \sqrt{\frac{1 + \dot{y}^2}{2gy}}$$

Now the Euler's Eq.(3.7) becomes

$$\begin{aligned} & -\frac{f}{2y} - \frac{d}{dx} \left(\frac{f\dot{y}}{1 + \dot{y}^2} \right) = 0 \\ \Rightarrow & -\frac{f}{2y} - \frac{df}{dx} \frac{\dot{y}}{1 + \dot{y}^2} + \frac{2f\dot{y}^2\ddot{y}}{(1 + \dot{y}^2)^2} - \frac{f\ddot{y}}{1 + \dot{y}^2} = 0 \\ \Rightarrow & -\frac{f}{2y} - \frac{df}{dx} \frac{\dot{y}}{1 + \dot{y}^2} - \frac{f(1 - \dot{y}^2)\ddot{y}}{(1 + \dot{y}^2)^2} = 0 \end{aligned}$$

Now,

$$\begin{aligned} \frac{df}{dx} &= \frac{\partial f}{\partial y} \dot{y} + \frac{\partial f}{\partial \dot{y}} \ddot{y} \\ &= -\frac{f}{2y} \dot{y} + \frac{f\dot{y}\ddot{y}}{1 + \dot{y}^2} \end{aligned}$$

So we then have for our Euler equation

$$\begin{aligned} & -\frac{f}{2y} \left(1 - \frac{\dot{y}^2}{1 + \dot{y}^2} \right) - \frac{f(\dot{y}^2 + (1 - \dot{y}^2))\ddot{y}}{(1 + \dot{y}^2)^2} = 0 \\ \Rightarrow & -\frac{f}{2y(1 + \dot{y}^2)} - \frac{f\ddot{y}}{(1 + \dot{y}^2)^2} = 0 \end{aligned} \tag{3.8}$$

Since $f \neq 0$, so it can be eliminated. If we multiply across by some factors, we find

$$1 + \dot{y}^2 + 2y\ddot{y} = 0$$

This differential equation can be solved by making some substitutions. First, define $u = \sqrt{y}\dot{y}$. Then we may rewrite the above equation as

$$\begin{aligned} & 1 + 2\sqrt{y} \frac{du}{dx} = 0 \\ \Rightarrow & 2 \frac{du}{dx} = -\frac{1}{\sqrt{y}} = -\frac{\dot{y}}{u} \\ \Rightarrow & -\frac{du^2}{dx} = -2u \frac{du}{dx} = \dot{y} \end{aligned}$$

convenience) which is now a perfect differential. Integrating, we find (where we choose the form of the constant of integration for later

$$\begin{aligned}
 -u^2 &= y + 2a \\
 \Rightarrow y\dot{y}^2 &= -(y + 2a) \\
 \Rightarrow \dot{y} &= \pm \sqrt{\frac{-(y + 2a)}{y}} \\
 \Rightarrow \frac{y \, dy}{\sqrt{-y(y + 2a)}} &= dx
 \end{aligned}$$

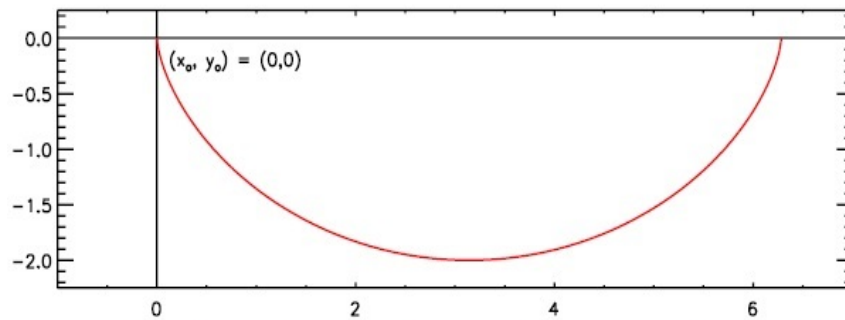
This kind of differential is not that unusual and can be solved by trigonometric substitution. Let $y = -a(1 - \cos \theta)$. Then we have $dy = -a \sin \theta \, d\theta$ and

$$\begin{aligned}
 \frac{a(1 - \cos \theta) \, a \sin \theta \, d\theta}{[a^2(1 - \cos \theta)(1 + \cos \theta)]^{1/2}} &= \pm dx \\
 \Rightarrow \frac{a(1 - \cos \theta) \, a \sin \theta \, d\theta}{[a^2(1 - \cos^2 \theta)]^{1/2}} &= \pm dx \\
 \Rightarrow a(1 - \cos \theta) \, d\theta &= \pm dx
 \end{aligned}$$

which is easily integrated. Our solution is therefore

$$\begin{aligned}
 x &= \pm a(\theta - \sin \theta) \\
 y &= -a(1 - \cos \theta)
 \end{aligned}$$

which is the final parametric solution for the curve that minimizes the travel time. The $\pm$ sign is chosen based on whether $x_1 > 0$ or $x_1 < 0$. The results describe a *cycloid*, the curve that is traced out by a single point on the edge of a rolling disk. The constant a is determined by the position of the endpoint. The brachistochrone problem is famous in the history of mathematics, for it was analysis of this problem by John Bernoulli that led to the formal foundation of the calculus of variations.



Cycloid solution to brachistochrone problem, for $x_1 > 0$ and $a = 1$.

3.4 Hamilton's equations of motion

In the former units, we were concerned exclusively with Lagrangian dynamics. For a holonomic system described by a set of n generalised coordinates, there exists an equivalent formulation of its

dynamics, a formulation in terms of its Hamilton's function or the Hamiltonian. Just like the Lagrangian, the Hamiltonian of a system can be used to obtain the equations of motion of the system.

• **Note:** The chief disadvantage of Lagrange's equations of motion is that they are second order total differential equations in generalised coordinates. Hamilton's equations of motion, on the other hand, are first order total differential equations in generalised coordinates and generalised momenta.

Theorem: Let a function $F(u_1, u_2, \dots, u_n)$ have an explicit dependence on the n independent variables $u_1, u_2, \dots, u_n$. Let the function F be transformed to another function $G = G(v_1, v_2, \dots, v_n)$ expressed in terms of a new set of n independent variables $v_1, v_2, \dots, v_n$, where these new variables are connected to the old variables by a given set of relations

$$v_i = \frac{\partial F}{\partial u_i}, \quad i = 1, \dots, n \quad (3.9)$$

and the form of G is given by $G(v_1, v_2, \dots, v_n) = u_i v_i - F(u_1, u_2, \dots, u_n)$. then the variables $u_1, \dots, u_n$ satisfy the dual transformation, namely, the relations

$$u_i = \frac{\partial G}{\partial v_i}, \quad i = 1, \dots, n \quad \text{and} \quad F(u_1, \dots, u_n) = u_i v_i - G(v_1, \dots, v_n) \quad (3.10)$$

This duality of transformation between the two functions $F(u_1, \dots, u_n)$ and $G(v_1, \dots, v_n)$ and also between the two sets of variables given by the Eqs.(3.9) and (3.10) are referred to as *Legendre's dual transformation*.

Now suppose that there is a further set of m independent passive variables $w_1, \dots, w_m$ which are present in both F and G . Then there would be some extra conditions for Legendre's dual transformation to be satisfied. These conditions are

$$\frac{\partial F}{\partial w_i} = -\frac{\partial G}{\partial w_i}, \quad i = 1, \dots, m \quad (3.11)$$

3.5 Hamilton's function and Hamilton's equations of motion

Let us now apply Legendre's dual transformation to the Lagrangian of a system $L(q_1, \dots, q_n, \dot{q}_1, \dots, \dot{q}_n, t)$ with $\dot{q}_i$ as the active variables and q_i and t as the passive variables. The dual variables of $\dot{q}_i$, $i = 1, \dots, n$ are given by the generalised momenta

$$p_i = \frac{\partial L}{\partial \dot{q}_i}, \quad i = 1, \dots, n \quad (3.12)$$

Hence the dual function of the Lagrangian L is

$$H = p_i \dot{q}_i - L(q, \dot{q}, t) \quad (3.13)$$

where $H = H(q_1, \dots, q_n, p_1, \dots, p_n, t) = H(q, p, t)$ is called *Hamiltonian* of the system. The dual transformation of Eq.(3.12) is

$$\dot{q}_i = \frac{\partial H}{\partial p_i}, \quad i = 1, \dots, n \quad (3.14)$$

and the Eqs.(3.11) for the passive variables takes the form

$$\frac{\partial L}{\partial t} = -\frac{\partial H}{\partial t} \quad \text{and} \quad \frac{\partial L}{\partial q_i} = -\frac{\partial H}{\partial q_i} \quad (3.15)$$

We know that provided (i) there are no on-potential forces, (ii) the system is holonomic and (iii) Euler-Lagrange's equations of motion are valid, one can write

$$\frac{\partial L}{\partial q_i} = \frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}_i} \right) = \dot{p}_i$$

Substituting in Eq.(3.15) we get

$$\dot{p}_i = -\frac{\partial H}{\partial q_i}, \quad i = 1, \dots, n \quad (3.16)$$

The set of Eqs.(3.14) and (3.16) together called the *Hamilton's equations of motion* or the canonical equations for motion.

Some Properties of the Hamiltonian and of Hamilton's equations of motion:

1. If the Lagrangian does not have any explicit dependence on time, the Hamiltonian also does not depend explicitly on time.

2. Consider

$$\begin{aligned} \frac{dH}{dt} &= \frac{d}{dt} H(q_1, \dots, q_n, p_1, \dots, p_n, t) \\ &= \frac{\partial H}{\partial t} + \frac{\partial H}{\partial q_i} \dot{q}_i + \frac{\partial H}{\partial p_i} \dot{p}_i \\ &= \frac{\partial H}{\partial t} - \dot{p}_i \dot{q}_i + \dot{p}_i \dot{q}_i \\ &= \frac{\partial H}{\partial t} \end{aligned}$$

where we have used Hamilton's equations of motion. Thus if the Lagrangian or equivalently the Hamiltonian does not explicitly depend on time, the Hamiltonian is a constant of motion.

Suggestive Reading

1. N.C. Rana, P.S. Joag, *Classical Mechanics*, Mc. Graw. Hill.

Unit 4

Course structure

Canonical Transformations : Canonical transformations and Generating functions, Different types of generating functions, Poisson's brackets and their invariance under canonical transformations, Illustrative problems

Introduction

In classical mechanics, there is no unique prescription for one to choose the generalized coordinates for a problem. As long as the coordinates and the corresponding momenta span the entire phase space, it becomes an acceptable set. However, it turns out in practice that some choices are better than some others as they make a given problem simpler while still preserving the form of Hamilton's equations. Going over from one set of chosen coordinates and momenta to another set which satisfy Hamilton's equations is done by **canonical transformation**.

For instance, if we consider the central force problem in two dimensions and choose Cartesian coordinate, the potential is $-\frac{k}{\sqrt{x^2 + y^2}}$. However, if we choose (r, θ) coordinates,

the potential is $-\frac{k}{r}$ and θ is cyclic, which greatly simplifies the problem. The number of cyclic coordinates in a problem may depend on the choice of generalized coordinates. A cyclic coordinate results in a constant of motion which is its conjugate momentum. The example above where we replace one set of coordinates by another set is known as **point transformation**.

In the Hamiltonian formalism, the coordinates and momenta are given equal status and the dynamics occurs in what we know as the **phase space**. In dealing with phase space dynamics, we need point transformations in the phase space where the new coordinates (Q_i) and the new momenta (P_i) are functions of old coordinates (p_i) and old momenta (p_i):

$$\begin{aligned}Q_i &= Q_i(\{q_j\}, \{p_j\}, t) \\P_i &= P_i(\{q_j\}, \{p_j\}, t)\end{aligned}$$

Such transformations are called **contact transformations**.

However, in Hamiltonian mechanics, only those transformations are of interest for which the quantities Q_i and P_i are canonically conjugate pairs. This means that there exists some Hamiltonian $K(Q, P, t)$ with respect to which P and Q satisfy Hamilton's equations:

$$\begin{aligned}\dot{Q}_i &= \frac{\partial K}{\partial P_i} \\ \dot{P}_i &= -\frac{\partial K}{\partial Q_i}\end{aligned}$$

Note that Q and P are not specific to a particular mechanical system but are common to all systems having the same degree of freedom. We may, for instance, obtain P and Q for a plane harmonic oscillator and use the same set to solve Kepler's problem. Such transformations, which preserve the form of Hamilton's equations, are known as **canonical transformation**. Though the terms contact transformation is used synonymously with the term canonical transformation, not all contact transformations are canonical, as the following example shows. Let us consider a free particle system with the Hamiltonian $\mathcal{H} = \frac{p^2}{2m}$. Since q is cyclic, the conjugate momentum p is a constant of motion, $\dot{p} = 0$, and we have $\dot{q} = \frac{p}{m}$. Consider now, a contact transformation

$$P = pt; \quad Q = qt \tag{1}$$

Does a Hamiltonian $K(P, Q)$ exist for which P and Q satisfy the Hamilton's equations:

$$\begin{aligned}\dot{P} &= -\frac{\partial K}{\partial Q} \\ \dot{Q} &= \frac{\partial K}{\partial P}\end{aligned}$$

Existence of such a Hamiltonian would imply $\frac{\partial \dot{P}}{\partial P} = -\frac{\partial \dot{Q}}{\partial Q}$ as each of the expressions equals $-\frac{\partial^2 K}{\partial P \partial Q}$. However, Since $\frac{\partial \dot{P}}{\partial P} = \frac{\partial \dot{Q}}{\partial Q} = \frac{1}{t}$, this condition is not satisfied and the modified Hamiltonian K does not exist. Thus the contact transformations (1) are non-canonical, they do not preserve the volume of the phase space.

It may be noted that this does not imply that one cannot find an equation of motion using these variables. Indeed, since p is constant, $dP/dt = p = P/t$, which gives $P = kt$, where k is a constant. Likewise, we have,

$$\frac{d}{dt} \left(\frac{Q}{t} \right) = \frac{dq}{dt} = \frac{p}{m} = \frac{Pt}{m} = \frac{k}{m}$$

so that $Q = \frac{k}{m}t^2 + C$, where C is a constant.

To illustrate a canonical transformation, consider the free particle Hamiltonian again for which $\dot{p} = 0$ and $\dot{q} = p/m$, as before. Consider a linear transformation of (p, q) to a new set (P, Q) , which is given by

$$\begin{aligned} P &= \alpha p + \beta q \\ Q &= \gamma p + \delta q \end{aligned} \quad (2)$$

If the matrix of transformation is non-singular, i.e. if $\Delta = \begin{vmatrix} \alpha & \beta \\ \gamma & \delta \end{vmatrix} \neq 0$, then the above transformation is invertible and we have

$$\begin{aligned} p &= \frac{1}{\Delta}(\delta P - \beta Q) \\ q &= \frac{1}{\Delta}(-\gamma P + \alpha Q) \end{aligned} \quad (3)$$

It can be checked that

$$\begin{aligned} \dot{P} &= \alpha \dot{p} + \beta \dot{q} = \beta \dot{q} = \beta \frac{p}{m} = \frac{\beta}{m\Delta}(\delta P - \beta Q) \\ \dot{Q} &= \gamma \dot{p} + \delta \dot{q} = \delta \dot{q} = \delta \frac{p}{m} = \frac{\delta}{m\Delta}(\delta P - \beta Q) \end{aligned} \quad (4)$$

Is there a Hamiltonian K for which P and Q are canonical? It is seen that

$$\begin{aligned} \frac{\partial \dot{P}}{\partial P} &= \frac{\beta\delta}{m\Delta} \\ \frac{\partial \dot{Q}}{\partial Q} &= -\frac{\beta\delta}{m\Delta} \end{aligned}$$

so that K exists. To determine K note that $\dot{P} = -\frac{\partial K}{\partial Q}$ which gives

$$K = - \int \dot{P} dQ = -\frac{\beta\delta}{m\Delta} PQ + \frac{\beta^2}{m\Delta} \frac{Q^2}{2} + f(P)$$

and $\dot{Q} = \frac{\partial K}{\partial P}$ gives

$$K = - \int \dot{Q} dP = -\frac{\beta\delta}{m\Delta} PQ + \frac{\delta^2}{m\Delta} \frac{P^2}{2} + f(Q)$$

These two expressions give

$$K = \frac{1}{2m\Delta}(\delta P - \beta Q)^2$$

Except for the factor Δ , this is just the old Hamiltonian written in terms of the new variables. For $\Delta = 1$, the two Hamiltonians are identical and the transformation corresponds

to a point transformation of Lagrangian mechanics.

Consider a second example where, once again we take the free particle Hamiltonian of the previous example. Consider a transformation

$$P = p \cos q, \quad Q = p \sin q \quad (5)$$

The transformation is invertible and it is easily seen that

$$p = \sqrt{P^2 + Q^2}, \quad q = \tan^{-1} \frac{Q}{P} \quad (6)$$

Using the same algebra as for the previous example, we get

$$\dot{P} = \dot{p} \cos q - \dot{q} p \sin q = -\frac{p^2}{m} \sin q = -\frac{Q}{m} \sqrt{P^2 + Q^2}$$

and

$$\dot{Q} = \frac{P}{m} \sqrt{P^2 + Q^2}$$

It can be seen that $\frac{\partial \dot{P}}{\partial P} = -\frac{\partial \dot{Q}}{\partial Q}$, showing that the Hamiltonian exists. It is easy to show that the Hamiltonian is given by $K = \frac{1}{3m} \sqrt{P^2 + Q^2}$.

However, instead of considering the free particle Hamiltonian, consider the Hamiltonian of a particle in a uniform gravitational field.

$$\mathcal{H} = \frac{p^2}{2m} + mgq$$

In this case $\dot{p} = -mg$ and $\dot{q} = \frac{p}{m}$. We had seen that the transformation is invertible. In this case, we get

$$\begin{aligned} \dot{P} &= \dot{p} \cos q - \dot{q} p \sin q \\ &= -mg \cos q - \frac{p^2}{m} \sin q \\ &= -mg \frac{P}{\sqrt{P^2 + Q^2}} - \frac{1}{m} \frac{Q}{\sqrt{P^2 + Q^2}} \end{aligned}$$

and

$$\begin{aligned} \dot{Q} &= \dot{p} \sin q + \dot{q} p \cos q \\ &= -mg \frac{Q}{\sqrt{P^2 + Q^2}} + \frac{p^2}{m} \cos q \\ &= -mg \frac{Q}{\sqrt{P^2 + Q^2}} - \frac{1}{m} P \sqrt{P^2 + Q^2} \end{aligned}$$

One can check $\frac{\partial \dot{P}}{\partial P} \neq -\frac{\partial \dot{Q}}{\partial Q}$, so that the transformation is not canonical with respect to the given Hamiltonian. This shows that (5) is not a valid canonical transformation. This is because, in order that a transformation may be canonical, it must satisfy Hamilton's equation for all systems having the same number of degrees of freedom.

2 Canonical Transformation & Generating Function

We start with an observation that Hamilton's equations of motion can be derived from the Hamilton's principle by writing action written in the form $S = \int \mathcal{L} dt = \int (p\dot{q} - \mathcal{H}) dt$ and optimizing the action. In deriving the Euler Lagrange equation, we had optimized action subject to the condition that $\delta q(t_i) = \delta q(t_f) = 0$, i.e. there is no variation at the end points. We have

$$\delta S = \delta \int_{t_0}^{t_f} (p_i \dot{q}_i - \mathcal{H}) dt = 0$$

with implied summation on repeated index and we have denoted the initial time as t_0 so as not to confuse with the index i . The variation in the action is then given by

$$\begin{aligned} \delta S &= \int_{t_0}^{t_f} \left[\dot{q}_i \delta p_i + p_i \delta \dot{q}_i - \frac{\partial \mathcal{H}}{\partial q_i} \delta q_i - \frac{\partial \mathcal{H}}{\partial p_i} \delta p_i \right] dt \\ &= \int_{t_0}^{t_f} \left[\left(\dot{q}_i - \frac{\partial \mathcal{H}}{\partial p_i} \right) \delta p_i + \frac{d}{dt} (p_i \delta q_i) - \dot{p}_i \delta q_i - \frac{\partial \mathcal{H}}{\partial q_i} \delta q_i \right] dt \\ &= \int_{t_0}^{t_f} \left[\left(\dot{q}_i - \frac{\partial \mathcal{H}}{\partial p_i} \right) \delta p_i - \left(\dot{p}_i + \frac{\partial \mathcal{H}}{\partial q_i} \right) \delta q_i \right] dt + [p_i \delta q_i]_{t_0}^{t_f} \end{aligned}$$

We have seen earlier that setting $\delta S = 0$ for all variations of q_i gave us Euler Lagrange equation and we expect that setting $\delta S = 0$ for all variations of q_i and p_i will lead to Hamilton's equations. However, there is a subtle difference between the two cases. In the Lagrangian formulation q_i and $\dot{q}_i$ were not independent variables and hence we only needed variation with respect to q_i . In the Hamiltonian formalism, however, we need to vary both q_i and p_i . The last term of the above equation vanishes if $\delta q_i(t_0) = \delta q_i(t_f) = 0$ and it does not requires the corresponding relation for vanishing of the momentum coordinates. Thus, we have an additional degree of freedom and we could define the action as

$$S = \int (p_i \dot{q}_i - \mathcal{H} + \frac{dF(q, p)}{dt}) dt$$

so that for all variations of p_i and q_i , the Hamilton's equations would be valid.

The variational principle requires the function $p\dot{q} - \mathcal{H}$ to satisfy the Euler Lagrange equations. Thus we have

$$\begin{aligned} \frac{d}{dt} \left(\frac{\partial}{\partial \dot{q}} (p\dot{q} - \mathcal{H}) \right) - \frac{\partial}{\partial q} (p\dot{q} - \mathcal{H}) &= 0 \\ \frac{d}{dt} \left(\frac{\partial}{\partial \dot{p}} (p\dot{q} - \mathcal{H}) \right) - \frac{\partial}{\partial p} (p\dot{q} - \mathcal{H}) &= 0 \end{aligned}$$

Since $\mathcal{H} = \mathcal{H}(p, q)$, we get from above, $\frac{d}{dt}p + \frac{\partial \mathcal{H}}{\partial q} = 0$, i.e. $\dot{p} = -\frac{\partial \mathcal{H}}{\partial q}$ and likewise, since $p\dot{q} - \mathcal{H}$ no dependence on $\dot{p}$, $-\dot{q} + \frac{\partial \mathcal{H}}{\partial p} = 0$, i.e., $\dot{q} = \frac{\partial \mathcal{H}}{\partial p}$. The variational principle satisfied by the old pair of variables (q, p) is

$$\delta S = \delta \int (p\dot{q} - \mathcal{H})dt = 0$$

with no variation at the end points. With respect to the new variables, one then must have

$$\delta S' = \delta \int (P\dot{Q} - \mathcal{H})dt = 0$$

These two forms of S and S' are completely equivalent if the two integrands differ either by a scale factor or by total time differential of a function F , i.e. if

$$\sum_i P_i \dot{Q}_i - K = \lambda \sum_i (p_i \dot{q}_i - \mathcal{H}) \quad (7)$$

known as the **scale transformation** or,

$$\sum_i P_i \dot{Q}_i - K + \frac{dF}{dt} = \sum_i p_i \dot{q}_i - \mathcal{H} \quad (8)$$

known as the **canonical transformation** or by a combination of both

$$\sum_i P_i \dot{Q}_i - K + \frac{dF}{dt} = \lambda \sum_i (p_i \dot{q}_i - \mathcal{H}) \quad (9)$$

which is usually referred to as **extended canonical transformation**. If we rewrite (8) as

$$\sum_i p_i dq_i - \mathcal{H}dt - \sum_i P_i dQ_i + Kdt = dF \quad (10)$$

we notice that the difference between the two differential forms must be an exact differential. Analogous statement can be made regarding (9) as well. It turns out that this condition of exact differentiability is both necessary and sufficient condition for a transformation to be canonical.

2.1 Scale Transformation

Scale transformation is achieved by multiplying the old coordinates and momenta with some scale factors to get the corresponding new quantities, $P_i = \nu p_i$ and $Q_i = \mu q_i$. The scale factors can be different for momenta and the coordinates but must be the same for all coordinates (and all momenta). P and Q are obtainable from the new Hamiltonian K through Hamilton's equations

$$\dot{P}_i = -\frac{\partial K}{\partial Q_i} \quad \dot{Q}_i = \frac{\partial K}{\partial P_i}$$

Thus we have, using $P_i = \nu p_i$ and $Q_i = \mu q_i$,

$$\nu \dot{p}_i = -\frac{\partial K}{\partial \mu q_i}$$

which gives

$$\dot{p}_i = -\frac{1}{\mu\nu} \frac{\partial K}{\partial q_i}$$

Comparing this with the equation $\dot{p}_i = \frac{\partial \mathcal{H}}{\partial q_i}$, we get $K = \mu\nu \mathcal{H}$. This gives

$$\sum_i P_i \dot{Q}_i - K(P, Q, t) = \mu\nu \sum_i (p_i \dot{q}_i - \mathcal{H}(p, q, t))$$

Comparing this with (7) we get $\lambda = \mu\nu$.

2.2 Generating Function

The function F in (8) or (9) is, in principle, a function of p_i, q_i, P_i, Q_i and t . It is known as the **generating function** of canonical transformation. We will illustrate it with a couple of examples. *We will use Einstein summation convention according to which a repeated index in a single term implies a sum, i.e. $f_i g_i = \sum_i f_i g_i$.*

Example : Identity Transformation:

Consider a transformation $F = q_i P_i - Q_i P_i$ in which the dependence on time is implicit. We have

$$\begin{aligned} P_i Q_i - K + \frac{dF}{dt} &= P_i Q_i - K + \dot{q}_i P_i + q_i \dot{P}_i - \dot{Q}_i P_i - Q_i \dot{P}_i \\ &= -K + (q_i - Q_i) \dot{P}_i + P_i \dot{q}_i \\ &\equiv p_i \dot{q}_i - \mathcal{H} \end{aligned}$$

The equation is satisfied by identity transformation $p_i = P_i, q_i = Q_i$ and $K = \mathcal{H}$.

Example: Swapping Coordinates and Momenta

We can generalize the above example a consider a generating function of the following type (summation convention used)

$$F = f_i(q_1, q_2, \dots, q_n; t) P_i - Q_i P_i$$

where f_i is a function of the arguments $q_1, q_2, \dots, q_n$ and t . We then have

$$\begin{aligned} P_i \dot{Q}_i - K + \frac{dF}{dt} &= P_i \dot{Q}_i - K + \left(\frac{\partial f_i}{\partial q_j} \dot{q}_j P_i + \frac{\partial f_i}{\partial t} P_i \right) + f_i \dot{P}_i - (\dot{Q}_i P_i + Q_i \dot{P}_i) \\ &= -K + (f_i - Q_i) \dot{P}_i + \frac{\partial f_i}{\partial q_j} \dot{q}_j P_i + \frac{\partial f_i}{\partial t} P_i \\ &\equiv p_i \dot{q}_i - \mathcal{H} \end{aligned}$$

The set of equations above can be satisfied by taking $Q_i = f_i$, $K = \mathcal{H} + \frac{\partial f_i}{\partial t} P_i$ and requiring $p_i \dot{q}_i = \frac{\partial f_i}{\partial q_j} \dot{q}_j P_i$ which can be rearranged as follows:

$$p_i \dot{q}_i = \frac{\partial f_i}{\partial q_j} \dot{q}_j P_i = \frac{\partial f_j}{\partial q_i} \dot{q}_i P_j$$

where in the last term we have interchanged the sum over i and j . This gives

$$p_i = \sum_j \frac{\partial f_j}{\partial q_i} P_j$$

where we have explicitly reintroduced the sum for clarity. These are n equations which must be inverted to get P_i . It may be noted that the choice of F is not unique corresponding to a particular canonical transformation. For instance, if we add any $g(t)$ to F it does not affect the action integral, because $\frac{dF}{dt} \rightarrow \frac{dF}{dt} + \frac{dg}{dt}$ does not change the Lagrangian equations.

Let us illustrate how to find generator. Let $K(Q, P, t) = \mathcal{H}(q, p, t)$. Plugging this into (8), we get

$$\frac{dF}{dt} = \sum_i (p_i \dot{q}_i - P_i \dot{Q}_i)$$

This is not very interesting because there is no time dependence. The simplest way to satisfy it is

$$F = F(q, Q)$$

We have

$$\begin{aligned} \frac{dF}{dt} &= \sum_i \frac{\partial F}{\partial q_i} \dot{q}_i + \frac{\partial F}{\partial Q_i} \dot{Q}_i \\ &= \sum_i (p_i \dot{q}_i - P_i \dot{Q}_i) \end{aligned}$$

This gives with $\frac{\partial F}{\partial q_i} = p_i$; $\frac{\partial F}{\partial Q_i} = -P_i$. A trivial way to satisfy the equations is to take $F(q_i, Q_i) = \sum_i q_i Q_i$ so that

$$\begin{aligned} \frac{\partial F}{\partial q_i} &= Q_i = p_i \\ \frac{\partial F}{\partial Q_i} &= q_i = -P_i \end{aligned}$$

which simply requires us to swap the coordinates and the momenta in the Hamiltonian formalism (but for a sign).

3 Different types of Generating Functions

We have remarked that the generating function F is a function of the old coordinates (and momenta) q, p , the new coordinates Q, P and time t . However, since Q and P themselves are functions of q and p (and vice-versa), only two of the four variables are needed in the description of the generating function. We need one each from the old pair and the new pair to describe the generating function, in addition to time. There are four possibilities and depending on the combination that we choose, the generating functions are classified as type 1, 2, 3 or 4.

3.1 Type 1 Generator

Type 1 generator, denoted by F_1 is a function of q_i and Q_i . We denote

$$F = F_1(q_i, Q_i, t) \quad (11)$$

Since $F_1 = F_1(q, Q, t)$, we can write

$$\begin{aligned} \frac{dF}{dt} &= \frac{\partial F_1}{\partial q_i} \dot{q}_i + \frac{\partial F_1}{\partial Q_i} \dot{Q}_i + \frac{\partial F_1}{\partial t} \\ &\equiv \sum_i p_i \dot{q}_i - \sum_i P_i \dot{Q}_i + K - \mathcal{H} \end{aligned} \quad (12)$$

where the last expression is from (8). We can compare the last two expressions for $\frac{dF_1}{dt}$ to get

$$\begin{aligned} p_i &= \frac{\partial F_1}{\partial q_i} \\ P_i &= -\frac{\partial F_1}{\partial Q_i} \\ K &= \mathcal{H} + \frac{\partial F_1}{\partial t} \end{aligned} \quad (13)$$

Note that (12) can be rewritten as

$$dF_1 = p_i dq_i - P_i dQ_i + (K - \mathcal{H})dt \quad (14)$$

If the Hamiltonian has no explicit time dependence, the generating function will also not have an explicit time dependence and in such a case $K = \mathcal{H}$. In such a situation (8) can be written as (summation convention used)

$$\frac{dF}{dt} = p_i \dot{q}_i - P_i \dot{Q}_i$$

For instantaneous transformation, we can write the above as

$$dF = p_i \delta q_i - P_i \delta Q_i$$

Since the left hand side is a perfect differential, so is the right hand side. In order that this may be so, we require

$$\frac{\partial p_i}{\partial Q_i} = \frac{\partial P_i}{\partial q_i} \quad (15)$$

When the functional dependence are given, one can always use (15) to test whether the transformation is canonical.

As shown earlier, one of the simplest examples of a F_1 type transformation is $F_1(q_i, Q_i) = q_i Q_i$, which gives the swapping transformation $p_i = Q_i$ and $P_i = -q_i$. Let us see what it does to the Hamiltonian of the Harmonic oscillator for which $\mathcal{H} = \frac{p^2}{2m} + \frac{1}{2}kq^2$. Since the Hamiltonian has no explicit time dependence, we have $K = \mathcal{H}$. In such a case

$$K = \frac{Q^2}{2m} + \frac{1}{2}kP^2$$

The new equations of motion are

$$\begin{aligned} \dot{Q} &= \frac{\partial K}{\partial P} = kP \\ \dot{P} &= -\frac{\partial K}{\partial Q} = -\frac{Q}{m} \end{aligned}$$

Thus we get

$$\ddot{Q} = k\ddot{P} = -\frac{Q}{m}$$

which has the solution $Q = Q_0 \cos(\omega t + \delta)$. We can obtain P from the equation $P = \frac{\dot{Q}}{k}$ which gives $P = -\frac{Q_0\omega}{k} \sin(\omega t + \delta)$. Since this is equal to $-q$, we can identify $q_0 = Q_0\omega/k$. We then can write

$$q = q_0 \cos(\omega t + \delta')$$

where $\delta' - \delta = \pi/2$.

3.2 Type 2 Generator

Type 2 generators, denoted by $F_2(q, P, t)$ can be obtained from the type 1 generator $F_1(q, Q, t)$ by means of a Legendre transformation. Since we have, in view of (13)

$$P_i = -\frac{\partial F_1}{\partial Q_i}$$

and we wish to replace the variable Q_i by the new variable P_i , we define

$$F_2(q_i, P_i, t) = F_1(q_i, Q_i, t) + \sum_i P_i Q_i \quad (16)$$

Note that starting with the equation (14) for F_1 can be written (use summation convention) by observing

$$\frac{dF_1}{dt} = p_i \dot{q}_i - P_i \dot{Q}_i + K - \mathcal{H} \quad (17)$$

Using (16), we have

$$\begin{aligned} \frac{dF_2}{dt} &= \frac{dF_1}{dt} + P_i \dot{Q}_i + \dot{P}_i Q_i \\ &= p_i \dot{q}_i + \dot{P}_i Q_i + (K - \mathcal{H}) \end{aligned} \quad (18)$$

Since F_2 depends on q and P , we can write

$$\frac{dF_2}{dt} = \frac{\partial F_2}{\partial q_i} \dot{q}_i + \frac{\partial F_2}{\partial P_i} \dot{P}_i + \frac{\partial F_2}{\partial t} dt$$

Comparing it with (17), we get

$$\begin{aligned} \frac{\partial F_2}{\partial q_i} &= p_i \\ \frac{\partial F_2}{\partial P_i} &= Q_i \\ \frac{\partial F_2}{\partial t} &= K - \mathcal{H} \end{aligned} \quad (19)$$

3.3 Type 3 Generator

In a similar way as above, we can define $F_3(p, Q, t)$ through a Legendre transform on F_1 . Referring to (13), we have, since the variable q_i is being replaced by p_i , and, $p_i = \frac{\partial F_1}{\partial q_i}$

$$F_3(p, Q, t) = F_1(q, Q, t) - p_i q_i$$

We then have

$$\begin{aligned} \frac{dF_3}{dt} &= \frac{dF_1}{dt} - p_i \dot{q}_i - q_i \dot{p}_i \\ &= -q_i \dot{p}_i - P_i \dot{Q}_i + (K - \mathcal{H}) \end{aligned} \quad (20)$$

On the other hand, since F_3 depends on (p_i, Q_i, t) , we can write,

$$\frac{dF_3}{dt} = \frac{\partial F_3}{\partial p_i} \dot{p}_i + \frac{\partial F_3}{\partial Q_i} \dot{Q}_i + \frac{\partial F_3}{\partial t}$$

Comparing this with (20) we get

$$\begin{aligned} \frac{\partial F_3}{\partial Q_i} &= -P_i \\ \frac{\partial F_3}{\partial p_i} &= -q_i \\ \frac{\partial F_3}{\partial t} &= K - \mathcal{H} \end{aligned} \quad (21)$$

3.4 Type 4 Generator

Finally we would like to define $F_4(p, P, t)$. In this case it is better to start with F_2 or F_3 and do a Legendre transformation. Starting with F_3 , we have to replace Q_i with P_i . In view of (21), we have

$$F_4(p, P, t) = F_3(p, Q, t) - Q_i P_i$$

We then have

$$\begin{aligned} \frac{dF_4}{dt} &= \frac{dF_3}{dt} + Q_i \dot{P}_i + P_i \dot{Q}_i \\ &= -q_i \dot{p}_i + Q_i \dot{P}_i + (K - \mathcal{H}) \end{aligned}$$

Since $F_4 = F_4(p, P, t)$, we have

$$\frac{dF_4}{dt} = \frac{\partial F_4}{\partial p_i} \dot{p}_i + \frac{\partial F_4}{\partial P_i} \dot{P}_i + \frac{\partial F_4}{\partial t}$$

Comparing with the preceding relation, we have we get

$$\begin{aligned} \frac{\partial F_4}{\partial p_i} &= -q_i \\ \frac{\partial F_4}{\partial P_i} &= Q_i \\ \frac{\partial F_4}{\partial t} &= K - \mathcal{H} \end{aligned} \tag{22}$$

3.5 Summary of the different types of Generators

The following table gives a summary of relationships for different types of generators.

Generators	Derivatives	Relationship
$F_1(q, Q, t)$	$p_i = \frac{\partial F_1}{\partial q_i}; P_i = -\frac{\partial F_1}{\partial Q_i}$	$\frac{\partial p_i}{\partial Q_j} = -\frac{\partial P_j}{\partial q_i}$
$F_2(q, P, t)$	$p_i = \frac{\partial F_2}{\partial q_i}; Q_i = \frac{\partial F_2}{\partial P_i}$	$\frac{\partial p_i}{\partial P_j} = \frac{\partial Q_j}{\partial q_i}$
$F_3(p, Q, t)$	$P_i = -\frac{\partial F_3}{\partial Q_i}; q_i = -\frac{\partial F_3}{\partial p_i}$	$\frac{\partial P_i}{\partial p_j} = \frac{\partial q_j}{\partial Q_i}$
$F_4(p, P, t)$	$q_i = -\frac{\partial F_4}{\partial p_i}; Q_i = \frac{\partial F_4}{\partial P_i}$	$\frac{\partial Q_i}{\partial p_j} = -\frac{\partial q_j}{\partial P_i}$

4 Invariance of Poisson's Bracket

A canonical transformation preserves the Poisson's bracket. This property can be used to verify whether a transformation is indeed canonical. Let $(q, p) \rightarrow (Q, P)$ be a transformation. The set (q, p) satisfies Poisson's bracket relationship, i.e.

$$\{q_i, p_j\} = \delta_{ij} \quad (23)$$

What we need to show is that Poisson's bracket calculated using either basis is the same, i.e.,

$$\{Q_i, P_j\}_{QP} = \{Q_i, P_j\}_{qp} \quad (24)$$

Clearly, since $\{Q_i\}, \{P_i\}$ are canonically conjugate pairs, they satisfy

$$\{Q_i, P_j\}_{QP} = \delta_{ij} \quad (25)$$

To prove (24), we note

$$\begin{aligned} \{Q_i, P_j\}_{qp} &= \sum_k \left[\frac{\partial Q_i}{\partial q_k} \cdot \frac{\partial P_j}{\partial p_k} - \frac{\partial Q_i}{\partial p_k} \cdot \frac{\partial P_j}{\partial q_k} \right] \\ &= \sum_k \left[\frac{\partial Q_i}{\partial q_k} \cdot \frac{\partial q_k}{\partial Q_j} - \frac{\partial Q_i}{\partial p_k} \cdot \left(-\frac{\partial p_k}{\partial Q_j} \right) \right] = \frac{\partial Q_i}{\partial Q_j} = \delta_{ij} \end{aligned} \quad (26)$$

In proving the above, we have used (15).

In a similar way, we can prove that the Poisson's brackets of the old variables are preserved in the new basis,

$$\{q_i, p_j\}_{qp} = \{q_i, p_j\}_{QP}$$

In general, for any two dynamical variables X and Y , one can show that the Poisson's bracket relationship remains the same in either basis

$$\{X, Y\}_{QP} = \{X, Y\}_{qp} \quad (27)$$

5 Illustrative Problems

Example 1: Harmonic Oscillator The Hamiltonian for the Harmonic oscillator is $\mathcal{H}(p, q) = \frac{p^2}{2m} + \frac{1}{2}m\omega^2 q^2$. Let us carry out a transformation

$$\begin{aligned} p &= f(P) \cos Q \\ q &= \frac{f(P)}{m\omega} \sin Q \end{aligned} \quad (28)$$

This converts the Hamiltonian to a function of P alone with $K = \mathcal{H} = \frac{f(P)^2}{2m}$. Note that the Hamiltonian is cyclic in Q so that P is constant. The problem is to find $f(P)$ such

that the transformation is canonical.

Let us try an F_1 type transformation for which, as per eqn. (13)

$$\begin{aligned} p &= \frac{\partial F_1}{\partial q} \\ P &= -\frac{\partial F_1}{\partial Q} \\ K &= \mathcal{H} + \frac{\partial F_1}{\partial t} = \mathcal{H} \end{aligned} \tag{29}$$

the last equality because the Hamiltonian is time independent. From (28), we get,

$$p = m\omega q \cot Q = \frac{\partial F_1}{\partial q}$$

Integrating, we get

$$F_1 = \frac{m\omega q^2}{2} \cot Q$$

The momenta conjugate to Q is then given by

$$\begin{aligned} P &= -\frac{\partial F_1}{\partial Q} \\ &= \frac{m\omega q^2}{2} \operatorname{cosec}^2 Q \\ &= \frac{p^2}{2m\omega} \frac{1}{\cos^2 Q} \end{aligned}$$

which gives

$$p = \sqrt{2Pm\omega} \cos Q$$

so that $f(P) = \sqrt{2Pm\omega}$. This in turn gives the new Hamiltonian to be $K = \frac{f(P)^2}{2m} = \omega P$.

Thus the transformation (28) is equivalent to

$$\begin{aligned} p &= \sqrt{2Pm\omega} \cos Q \\ q &= \sqrt{\frac{2}{m\omega}} \sin Q \end{aligned}$$

Inverting, we get

$$\begin{aligned} P &= \frac{1}{2m\omega} (q^2 m^2 \omega^2 + p^2) \\ Q &= \tan^{-1} \left(\frac{m\omega q}{p} \right) \end{aligned}$$

Note that

$$\begin{aligned}\frac{\partial Q}{\partial q} &= \frac{m\omega p}{p^2 + m^2\omega^2 q^2} \\ \frac{\partial P}{\partial p} &= \frac{p}{m\omega} \\ \frac{\partial Q}{\partial p} &= -\frac{m\omega q}{p^2 + m^2\omega^2 q^2} \\ \frac{\partial P}{\partial q} &= qm\omega\end{aligned}$$

which shows that

$$\{Q, P\}_{qp} = 1$$

i.e. as expected, the transformation preserves Poisson's bracket.

As the energy is constant we have $P = \frac{E}{\omega}$. The Hamilton's equations for the new variables then gives

$$\dot{Q} = \frac{\partial K}{\partial P} = \omega$$

yielding $Q = \omega t + \alpha$ The original problem is now solved,

$$\begin{aligned}p &= \sqrt{2mE} \cos Q = \sqrt{2mE} \cos(\omega t + \alpha) \\ q &= \sqrt{\frac{2P}{m\omega}} \sin Q = \sqrt{\frac{2E}{m\omega^2}} \sin(\omega t + \alpha)\end{aligned}$$

Example 2:

Consider the Hamiltonian $\mathcal{H} = \frac{1}{2}(p^2 + q^2)$. Determine if the transformation $Q = \frac{1}{2}(q^2 + p^2)$ and $P = -\tan^{-1}(q/p)$ is canonical. If so, find a generating function of type F_1 .

In order that the given transformation is canonical, $p\delta q - P\delta Q = dF$ must be an exact differential. Note that

$$\begin{aligned}p\delta q - P\delta Q &= p\delta q + \tan^{-1} \frac{q}{p} \delta \left(\frac{q^2 + p^2}{2} \right) \\ &= \left(p + q \tan^{-1} \frac{q}{p} \right) \delta q + p \tan^{-1} \frac{q}{p} \delta p\end{aligned}$$

In order that the above is an exact differential,

$$\frac{\partial}{\partial p} \left(p + q \tan^{-1} \frac{q}{p} \right) = \frac{\partial}{\partial q} \left(p \tan^{-1} \frac{q}{p} \right)$$

It can be easily seen that both sides of the above condition equals $\frac{p^2}{p^2 + q^2}$ so that the transformation is canonical.

The Poisson's brackets can be shown to be preserved:

$$\begin{aligned}\{Q, P\} &= \frac{\partial}{\partial q} \left(\frac{1}{2}(q^2 + p^2) \right) \frac{\partial}{\partial p} \left(-\tan^{-1} \frac{q}{p} \right) - \frac{\partial}{\partial p} \left(\frac{1}{2}(q^2 + p^2) \right) \frac{\partial}{\partial q} \left(-\tan^{-1} \frac{q}{p} \right) \\ &= q \frac{(q/p^2)}{1 + q^2/p^2} - p \left(-\frac{1/p}{1 + q^2/p^2} \right) = 1\end{aligned}$$

We can obtain F by integrating either of the factors. For instance, a bit of exercise in integration gives

$$\begin{aligned}F &= \int \left(p + q \tan^{-1} \frac{q}{p} \right) dq \\ &= \frac{pq}{2} + \frac{q^2 + p^2}{2} \tan^{-1} \frac{q}{p}\end{aligned}$$

However, the above form cannot be a generating function as it has no dependence on the new variables. Since we wish it to be of form F_1 , we need to eliminate p . This is done by substitution $p = \sqrt{2Q - q^2}$ so that $\tan^{-1} \frac{q}{p} = \tan^{-1} \frac{q}{\sqrt{2Q - q^2}} = \sin^{-1} \frac{q}{\sqrt{2Q}}$. With these substitutions, we get

$$F = Q \sin^{-1} \frac{q}{\sqrt{2Q}} + \frac{1}{2} q \sqrt{2Q - q^2}$$

The Hamiltonian $K = \frac{1}{2}(q^2 + p^2) = Q$, so that $\dot{Q} = \frac{\partial K}{\partial P} = 0$, i.e. Q is constant, and $\dot{P} = -\frac{\partial K}{\partial Q} = -1$. Since Q is constant, so is the Hamiltonian and P decreases with time.

Example 3:

Prove that the transformation $Q = \ln \frac{\sin p}{q}$ and $P = q \cot p$ is canonical and derive all the different forms of generating functions.

We have, for the given transformation

$$\delta Q = \cot p \delta p - \frac{1}{q} \delta q$$

Thus

$$\begin{aligned}p \delta q - P \delta Q &= p \delta q - q \cot p \left(\cot p \delta p - \frac{1}{q} \delta q \right) \\ &= (p + \cot p) \delta q - q \cot^2 p \delta p\end{aligned}$$

This is an exact differential if

$$\frac{\partial}{\partial p}(p + \cot p) = \frac{\partial}{\partial q}(-q \cot^2 p)$$

It can be seen that both sides equal $-\cot^2 p$ and hence the transformation is canonical. It is also easy to see that Poisson's bracket is preserved.

The structure of generating function is

$$F = \int (p + \cot p) dq = pq + q \cot p$$

(We can also integrate the other factor over p and show that both the forms are the same if the constant integration is taken to be zero). We will now express this in terms of different forms of the generating function. Note that for the given transformation, $\sin p = qe^Q$ which gives $\cos p = \sqrt{1 - q^2 e^{2Q}}$ and $q \cot p = \sqrt{e^{-2Q} - q^2}$. Substituting these, we get

$$F_1 = q \cos^{-1} \sqrt{1 - q^2 e^{2Q}} + \sqrt{e^{-2Q} - q^2}$$

Suppose we wish to find the generator form F_2 . In such a case, we need to replace p by P , which is quite straightforward. We can find F_2 by a Legendre transformation. Since $P = -\frac{\partial F_1}{\partial Q}$, we have $F_2(q, P) = F_1(q, Q) + PQ$. Thus

$$F_2(q, P) = F_1(q, Q) = q \sin^{-1}(qe^Q) + \frac{1}{e^Q} \sqrt{1 - q^2 e^{2Q}} + PQ$$

However, as F_2 must be expressed as a function of q and P , we must accordingly express the variables. We had shown that $P = \frac{\sqrt{1 - q^2 e^{2Q}}}{e^Q}$, which gives $e^Q = \frac{1}{\sqrt{P^2 + q^2}}$. Thus

$$F_2(q, P) = q \sin^{-1} \left(\frac{q}{\sqrt{P^2 + q^2}} \right) + P - \frac{P}{2} \ln(P^2 + q^2)$$

Unit 5

Small Oscillations

1 Introduction

As an example of the use of the Lagrangian, we will examine the problem of small oscillations about a stable equilibrium point. The description of motion about a stable equilibrium is one of the most important problems in physics. This is true for both classical and quantum mechanics. Although there are other means of solving this problem, the Lagrangian method results in equations of motion that emphasizes symmetries of the system.

1.1 Formulation of the Problem

We assume that the forces acting on the system are conservative in which the potential depends only on the spatial coordinates; it is independent of the time and velocity. In addition, the transformations defining the generalized coordinates are assumed to not depend on the time explicitly. Hence, the constraints are time independent. For the system to be in equilibrium, the generalized force must be zero

$$Q_i = \left. \frac{\partial V}{\partial q_i} \right|_{q_i=q_{0i}} = 0 \quad (1)$$

where q_{0i} is the equilibrium point. Furthermore, the equilibrium point is stable if the second derivative of potential is positive; the potential is concave up at the equilibrium point.

We will be interested in small displacement about the equilibrium. Therefore, we will write the generalized coordinates in the following form $q_i = q_{0i} + \eta_i$ where η_i is the displacement from the equilibrium point. Assuming the displacements are small, the potential can be expanded about q_{0i} as follows

$$V = V_0 + \sum_i \left. \frac{\partial V}{\partial q_i} \right|_{q_i=q_{0i}} \eta_i + \frac{1}{2} \sum_{ij} \left. \frac{\partial^2 V}{\partial q_i \partial q_j} \right|_{q_i=q_{0i}} \eta_i \eta_j + \cdots \quad (2)$$

The first term in the expansion is a constant, since the potential is only defined to a constant, we can ignore this term. The second term is zero, since we are expanding about the minimum. Therefore, the potential in this approximation is

$$V \approx \frac{1}{2} \sum_{ij} \left. \frac{\partial^2 V}{\partial q_i \partial q_j} \right|_{q_i=q_{0i}} \eta_i \eta_j = \frac{1}{2} \sum_{ij} V_{ij} \eta_i \eta_j \quad (3)$$

where we assume that the displacements are small such that higher order terms are negligible. The V_{ij} in this equation are the equivalent spring constant for the one-dimensional problem. In addition, we note that the V_{ij} are symmetric as can be seen from the definition.

The kinetic energy, as shown earlier, can be written as the sum of three terms

$$T = \frac{1}{2} \sum_k m_k \left(\sum_j \frac{\partial \vec{r}_k}{\partial q_j} \dot{q}_j + \frac{\partial \vec{r}_k}{\partial t} \right)^2 = M_0 + \sum_j M_j \dot{q}_j + \frac{1}{2} \sum_{ij} M_{ij} \dot{q}_j \dot{q}_i \quad (4)$$

where

$$M_0 = \frac{1}{2} \sum_k m_k \left(\frac{\partial \vec{r}_k}{\partial t} \right)^2 \quad M_j = \sum_k m_k \frac{\partial \vec{r}_k}{\partial t} \frac{\partial \vec{r}_k}{\partial q_j} \quad M_{ij} = \sum_k m_k \frac{\partial \vec{r}_k}{\partial q_j} \frac{\partial \vec{r}_k}{\partial q_i} \quad (5)$$

Since we assume that the transformations are independent of time, only the M_{ij} terms are non-zero; this is a natural system. In addition, since M_{ij} in general is a function of the coordinates, we expand the M_{ij} about the equilibrium point, since we are assuming small displacements about it

$$M_{ij}(q) = M_{ij}(q_0) + \sum_k \left(\frac{\partial M_{ij}}{\partial q_k} \right)_{q_0} \eta_k + \dots \quad (6)$$

We keep only the lowest order term in this expansion, since the kinetic energy is already quadratic in the coordinates (velocities). Thus, the kinetic energy becomes

$$T = \frac{1}{2} \sum_{ij} M_{ij} \dot{q}_i \dot{q}_j = \frac{1}{2} \sum_{ij} T_{ij} \dot{\eta}_i \dot{\eta}_j \quad (7)$$

where the relation $q_i = q_{0i} + \eta_i$ is used to derive $\dot{q}_i = \dot{\eta}_i$ and T_{ij} denotes $M_{ij}(q_0)$. Note that the T_{ij} terms are symmetric.

The Lagrangian for small oscillations about an equilibrium point is

$$L = \frac{1}{2} \sum_{ij} T_{ij} \dot{\eta}_i \dot{\eta}_j - \frac{1}{2} \sum_{ij} V_{ij} \eta_i \eta_j \quad (8)$$

where T_{ij} describes the inertia of the system, and V_{ij} the stiffness¹. In Cartesian coordinates, and many other case, the kinetic energy term can be reduced to $\frac{1}{2} T_i \dot{\eta}_i^2$, the very familiar form of the kinetic energy. To calculate the equations of motion, we apply the Euler-Lagrange equations to the Lagrangian

$$\sum_j [T_{ij} \ddot{\eta}_j + V_{ij} \eta_j] = 0 \quad (9)$$

Keep in mind that the M_{ij} and V_{ij} are independent of the coordinates and time, and they are symmetric.

1.2 The Eigenvalues and Eigenvectors

The solutions to the equations of motion are of the form

$$\eta_j = C a_j e^{-i\omega t} \Rightarrow \sum_j [(V_{ij} - \omega^2 T_{ij}) a_j] = 0 \quad (10)$$

where C is an overall scale factor, a_i the relative amplitude of the i th term, and the second equation comes from substituting the solutions into the equations of motion. Notice that this is a matrix equation, representing matrix multiplication between an $n \times n$ matrix and an n -dimensional column

¹The stiffness does not have to be constrained to be positive. It can have negative or zero terms, but for small oscillations, the terms are positive.

vector. This set of equations can be solved for the allowed values of ω by setting the determinant of the coefficients of a_i equal to zero²

$$\det |\mathbf{V} - \omega^2 \mathbf{T}| = 0 \quad (11)$$

where we have written the equation explicitly in matrix form.

Through the manipulations we have performed, we have arrived at an eigenvalue problem

$$\mathbf{V}\mathbf{a} = \lambda \mathbf{T}\mathbf{a} \quad (12)$$

where $\mathbf{V}\mathbf{a}$ is shown to be a multiple of $\mathbf{T}\mathbf{a}$. First we will show that the eigenvalues (λ) are real and positive, and the eigenvectors ($\mathbf{a}$) are orthogonal. Start by taking the complex conjugate of Eq. 12

$$\mathbf{a}_j^\dagger \mathbf{V} = \lambda_j^* \mathbf{a}_j^\dagger \mathbf{T} \quad (13)$$

where we have used the fact that $\mathbf{V}$ and $\mathbf{T}$ are real and symmetric, and in addition, j corresponds to the j th eigenvalue and eigenvector. Next, multiply Eq. 13 by $\mathbf{a}_i$ from the right

$$\mathbf{a}_j^\dagger \mathbf{V}\mathbf{a}_i = \lambda_j^* \mathbf{a}_j^\dagger \mathbf{T}\mathbf{a}_i \quad (14)$$

For the i th eigenvalue and eigenvector, multiply Eq. 12 from the left by $\mathbf{a}_j^\dagger$

$$\mathbf{a}_j^\dagger \mathbf{V}\mathbf{a}_i = \lambda_i \mathbf{a}_j^\dagger \mathbf{T}\mathbf{a}_i \quad (15)$$

Next, subtract the two equations

$$0 = (\lambda_i - \lambda_j^*) \mathbf{a}_j^\dagger \mathbf{T}\mathbf{a}_i \quad (16)$$

If we assume that the λ_i are distinct and we consider the case $i \neq j$, then $\mathbf{a}_j^\dagger \mathbf{T}\mathbf{a}_i = 0$. This states that the amplitudes are orthogonal to each other with respect to $\mathbf{T}$.

Let's now consider the case where $i = j$. First notice that $\mathbf{a}_i^\dagger \mathbf{T}\mathbf{a}_i = (\mathbf{a}_i^\dagger \mathbf{T}\mathbf{a}_i)^\dagger$ implies that the matrix product is real. Next, write $\mathbf{a}_i = \boldsymbol{\alpha}_i + i\boldsymbol{\beta}_i$, where we have written $\mathbf{a}_i$ in terms of its real and imaginary components. Then, we get

$$\mathbf{a}_i^\dagger \mathbf{T}\mathbf{a}_i = \boldsymbol{\alpha}_i^T \mathbf{T}\boldsymbol{\alpha}_i + \boldsymbol{\beta}_i^T \mathbf{T}\boldsymbol{\beta}_i + i(\boldsymbol{\alpha}_i^T \mathbf{T}\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^T \mathbf{T}\boldsymbol{\alpha}_i) \quad (17)$$

Since the left hand side of this expression has already been shown to be real, the imaginary term is zero, thus

$$\mathbf{a}_i^\dagger \mathbf{T}\mathbf{a}_i = \boldsymbol{\alpha}_i^T \mathbf{T}\boldsymbol{\alpha}_i + \boldsymbol{\beta}_i^T \mathbf{T}\boldsymbol{\beta}_i \quad (18)$$

Starting with Eq. 7, we see that in matrix notation the kinetic energy is

$$T = \frac{1}{2} \dot{\boldsymbol{\eta}}^T \mathbf{T} \dot{\boldsymbol{\eta}} \quad (19)$$

Since $\eta \propto a$ and the phase of the amplitude is arbitrary, Eq. 18 represents twice the kinetic energy. For a real velocity, the kinetic energy must be positive. Therefore, the term $\mathbf{a}_j^\dagger \mathbf{T}\mathbf{a}_i$ must be positive (we already showed it was real), therefore $\lambda_i = \lambda_i^*$; λ_i is real.

²If the coefficient is invertible, then the determinant is non-zero. In this case, the amplitudes would all be zero, since I can multiply both sides of the equation by the inverse. Therefore to have a non-trivial solution, the determinant of the coefficients is zero.

Since $\mathbf{a}_i^\dagger \mathbf{T} \mathbf{a}_i$ is real, we can write $\mathbf{a}_i$ as

$$\mathbf{a}_i = N \boldsymbol{\alpha}_i e^{i\phi} \quad (20)$$

where N is an overall normalization factor, and ϕ is a real phase. The two factors can be absorbed into C of Eq. 10 making the $\mathbf{a}_i$ real.

Starting with the eigenvalue equation (Eq. 12) and multiplying from the left by $\mathbf{a}_i^T$, we can solve for the individual eigenvalues

$$\lambda_i = \frac{\mathbf{a}_i^T \mathbf{V} \mathbf{a}_i}{\mathbf{a}_i^T \mathbf{T} \mathbf{a}_i} \quad (21)$$

The denominator is proportional to the kinetic energy, thus is must be positive. The numerator is the potential energy, which must be positive for a stable equilibrium. Therefore, λ_i must be positive and the frequencies are therefore real.

To remove the arbitrary normalization and phase from $\mathbf{a}_i$, we require $\mathbf{a}_i^T \mathbf{T} \mathbf{a}_i = 1$. Since we have an equation for each degree of freedom, we can combine this relation with $\mathbf{a}_j^T \mathbf{T} \mathbf{a}_i = 0$, and write a matrix $\mathbf{A}$ whose columns are the individual $\mathbf{a}_i$. This leads to the relation

$$\mathbf{A}^T \mathbf{T} \mathbf{A} = \mathbf{1} \quad (22)$$

which is a congruence transformation that diagonalizes the matrix $\mathbf{T}$. We take this further by noting that Eq. 12 can be written as

$$\sum_j V_{ij} a_{jk} = \sum_k \sum_j T_{ij} a_{jl} \lambda_{lk} \quad \Rightarrow \quad \mathbf{V} \mathbf{A} = \mathbf{T} \mathbf{A} \boldsymbol{\lambda} \quad (23)$$

where $\boldsymbol{\lambda}$ is a diagonal matrix composed of the eigenvalues. If we multiply from the left by $\mathbf{A}^T$, we diagonalize $\mathbf{V}$

$$\mathbf{A}^T \mathbf{V} \mathbf{A} = \mathbf{A}^T \mathbf{T} \mathbf{A} \boldsymbol{\lambda} = \boldsymbol{\lambda} \quad (24)$$

Therefore, the problem of small oscillation is reduced to solving the following matrix problem

$$\mathbf{A}^T \mathbf{T} \mathbf{A} = \mathbf{1} \quad \mathbf{A}^T \mathbf{V} \mathbf{A} = \boldsymbol{\lambda} \quad (25)$$

That is, we must find a transformation that diagonalizes the matrix $\mathbf{T}$. In this form, the Lagrangian is written as

$$L = \sum_i \left(\frac{1}{2} \dot{\xi}_i^2 - \frac{1}{2} \omega_i^2 \xi_i^2 \right) \quad (26)$$

1.3 Multiple Roots

Let's assume that two of the eigenvalues are equal $\lambda_1 = \lambda_2$. From the formalism given above, we expect that one of the eigenvectors will be orthogonal to the remaining eigenvectors, but there is no guarantee that they will be orthogonal to each other; recall that we wish to create the matrix $\mathbf{A}$, which must be orthogonal in order to diagonalize the $\mathbf{T}$ and $\mathbf{V}$ matrices. To do this, we can write one of the eigenvector as the sum of the two

$$\mathbf{a}'_2 = c_1 \mathbf{a}_1 + c_2 \mathbf{a}_2 \quad (27)$$

and apply the orthogonality relation

$$\mathbf{a}_2^T \mathbf{T} \mathbf{a}_1 = c_1 + c_2 \mathbf{a}_2^T \mathbf{T} \mathbf{a}_1 = 0 \quad \Rightarrow \quad \frac{c_1}{c_2} = -\mathbf{a}_2^T \mathbf{T} \mathbf{a}_1 \quad (28)$$

This procedure can be generalized for n repeated roots (eigenvalues).

1.4 Example

Let's consider a triple pendulum as shown in Fig 1. We would like to determine the eigenfrequencies and eigenvectors (normal modes) of this oscillator for small oscillations.

We start by writing the Lagrangian in rectangular coordinates

$$L = \frac{1}{2}m_1(\dot{x}_1^2 + \dot{y}_1^2) + \frac{1}{2}m_2(\dot{x}_2^2 + \dot{y}_2^2) + \frac{1}{2}m_3(\dot{x}_3^2 + \dot{y}_3^2) - m_1gy_1 - m_2gy_2 - m_3gy_3 \quad (29)$$

where all the coordinates are measured from a coordinate system fixed to the pivot of the top pendulum as shown in the figure. Next we convert to a set of coordinates given by the angular displacement from the equilibrium point. The coordinates of the top pendulum are transformed as follows

$$\left. \begin{aligned} x_1 &= \ell_1 \sin \theta_1 \\ y_1 &= -\ell_1 \cos \theta_1 \end{aligned} \right\} \Rightarrow \left\{ \begin{aligned} \dot{x}_1 &= \ell_1 \dot{\theta}_1 \cos \theta_1 \\ \dot{y}_1 &= -\ell_1 \dot{\theta}_1 \sin \theta_1 \end{aligned} \right. \quad (30)$$

The kinetic and potential energies for m_1 are

$$T_1 = \frac{1}{2}m_1\ell_1^2\dot{\theta}_1^2 \quad V_1 = -m_1g\ell_1 \cos \theta_1 \approx \frac{1}{2}m_1g\ell_1\theta_1^2 \quad (31)$$

where the approximation comes from expanding the potential in a Taylor series

$$V \approx \frac{1}{2} \left(\frac{\partial^2 V}{\partial \theta_1^2} \right)_{\theta_1=0} \theta_1^2 \quad (32)$$

Next, we consider the kinetic and potential energies associated with m_2 . The transformation equations are

$$\left. \begin{aligned} (x_2 - x_1) &= \ell_2 \sin \theta_2 \\ (y_2 - y_1) &= -\ell_2 \cos \theta_2 \end{aligned} \right\} \Rightarrow \left\{ \begin{aligned} x_2 &= \ell_1 \sin \theta_1 + \ell_2 \sin \theta_2 \\ y_2 &= -\ell_1 \cos \theta_1 - \ell_2 \cos \theta_2 \end{aligned} \right\} \Rightarrow \left\{ \begin{aligned} \dot{x}_2 &= \ell_1 \dot{\theta}_1 \cos \theta_1 + \ell_2 \dot{\theta}_2 \cos \theta_2 \\ \dot{y}_2 &= \ell_1 \dot{\theta}_1 \sin \theta_1 + \ell_2 \dot{\theta}_2 \sin \theta_2 \end{aligned} \right. \quad (33)$$

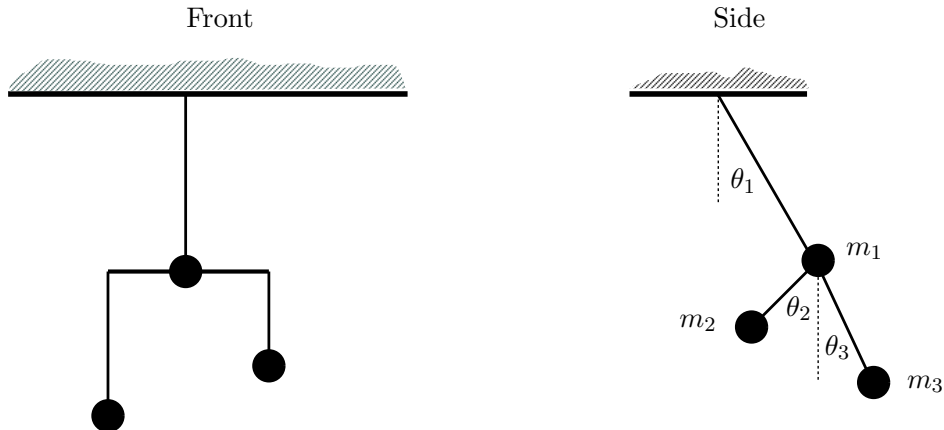


Figure 1: Triple pendulum for example.

Applying the transformations to the m_2 terms, the kinetic and potential energies are

$$T_2 = \frac{1}{2}m_2 \left[\ell_1^2 \dot{\theta}_1^2 + \ell_2^2 \dot{\theta}_2^2 + 2\ell_1 \ell_2 \dot{\theta}_1 \dot{\theta}_2 (\sin \theta_1 \sin \theta_2 + \cos \theta_1 \cos \theta_2) \right] \quad (34)$$

$$\approx \frac{1}{2}m_2 \left[\ell_1^2 \dot{\theta}_1^2 + \ell_2^2 \dot{\theta}_2^2 + 2\ell_1 \ell_2 \dot{\theta}_1 \dot{\theta}_2 \right] \quad (35)$$

$$V_2 = -m_2 g (\ell_1 \cos \theta_1 + \ell_2 \cos \theta_2) \approx \frac{1}{2}m_2 g (\ell_1 \theta_1^2 + \ell_2 \theta_2^2) \quad (36)$$

The kinetic and potential energies associated with m_3 are arrived at in a manner similar to m_2 . These energies are

$$T_3 \approx \frac{1}{2}m_3 \left[\ell_1^2 \dot{\theta}_1^2 + \ell_2^2 \dot{\theta}_3^2 + 2\ell_1 \ell_3 \dot{\theta}_1 \dot{\theta}_3 \right] \quad (37)$$

$$V_2 \approx -\frac{1}{2}m_3 g (\ell_1 \theta_1^2 + \ell_3 \theta_3^2) \quad (38)$$

Combining the kinetic and potential energy terms together, the Lagrangian becomes

$$L = \frac{1}{2}(m_1 + m_2 + m_3)\ell_1^2 \dot{\theta}_1^2 + \frac{1}{2}m_2 \ell_2^2 \dot{\theta}_2^2 + m_2 \ell_1 \ell_2 \dot{\theta}_1 \dot{\theta}_2 + \frac{1}{2}m_3 \ell_3^2 \dot{\theta}_3^2 + m_3 \ell_1 \ell_3 \dot{\theta}_1 \dot{\theta}_3 \\ - \frac{1}{2}(m_1 + m_2 + m_3)g\ell_1 \theta_1^2 - \frac{1}{2}m_2 g \ell_2 \theta_2^2 - \frac{1}{2}m_3 g \ell_3 \theta_3^2 \quad (39)$$

To determine the eigenvalues, we write the kinetic and potential energies in matrix form. The mass matrix for the kinetic energy is

$$[T_{ij}] = \begin{pmatrix} (m_1 + m_2 + m_3)\ell_1^2 & m_2 \ell_1 \ell_2 & m_3 \ell_1 \ell_3 \\ m_2 \ell_1 \ell_2 & m_2 \ell_2^2 & 0 \\ m_3 \ell_1 \ell_3 & 0 & m_3 \ell_3^2 \end{pmatrix} \quad (40)$$

and the stiffness matrix is

$$[V_{ij}] = \begin{pmatrix} (m_1 + m_2 + m_3)g\ell_1 & 0 & 0 \\ 0 & m_2 g \ell_2 & 0 \\ 0 & 0 & m_3 g \ell_3 \end{pmatrix} \quad (41)$$

Next, we solve the eigenvalue equation

$$\mathbf{V}\mathbf{a} = \lambda \mathbf{T}\mathbf{a} \quad \Rightarrow \quad (\mathbf{V} - \lambda \mathbf{T})\mathbf{a} = 0 \quad \text{where} \quad \lambda = \omega^2 \quad (42)$$

The eigenvalues are given by setting the determinant of the coefficients of $\mathbf{a}$ equal to zero

$$\begin{vmatrix} (m_1 + m_2 + m_3)(g\ell_1 - \lambda\ell_1^2) & -m_2 \ell_1 \ell_2 \lambda & -m_3 \ell_1 \ell_3 \lambda \\ -m_2 \ell_1 \ell_2 \lambda & m_2(g\ell_2 - \lambda\ell_2^2) & 0 \\ -m_3 \ell_1 \ell_3 \lambda & 0 & m_3(g\ell_3 - \lambda\ell_3^2) \end{vmatrix} = 0 \quad (43)$$

Since the algebra becomes extremely complicated for arbitrary masses and lengths, we will take all the masses and length to be equal, in this case the determinant is

$$\begin{vmatrix} 3m(g\ell - \lambda\ell^2) & -m\ell^2 \lambda & -m\ell^2 \lambda \\ -m\ell^2 \lambda & m(g\ell - \lambda\ell^2) & 0 \\ -m\ell^2 \lambda & 0 & m(g\ell - \lambda\ell^2) \end{vmatrix} = 0 \quad (44)$$

and the eigenfrequencies are

$$\omega^2 = \left\{ \frac{g}{\ell}, (3 + \sqrt{6})\frac{g}{\ell}, (3 - \sqrt{6})\frac{g}{\ell} \right\} \quad (45)$$

Next, we calculate the motion of the normal modes. This is done by calculating the eigenvectors $\mathbf{a}$ for each eigenfrequency. We start with the the eigenfrequency $\omega^2 = g/\ell$

$$\begin{pmatrix} 0 & -m\ell^2 & -mg\ell \\ -mg\ell & 0 & 0 \\ -mg\ell & 0 & 0 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = 0 \quad \Rightarrow \quad \begin{pmatrix} 0 \\ a_2 \\ -a_2 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} \quad (46)$$

We follow the same procedure for the two remaining eigenfrequencies and find the eigenvalues to be

$$\omega^2 = (3 + \sqrt{6})\frac{g}{\ell} \quad \Rightarrow \quad -mg\ell \begin{pmatrix} 3(2 + \sqrt{6}) & (3 + \sqrt{6}) & (3 + \sqrt{6}) \\ (3 + \sqrt{6}) & (2 + \sqrt{6}) & 0 \\ (3 + \sqrt{6}) & 0 & (2 + \sqrt{6}) \end{pmatrix} \begin{pmatrix} 1 \\ a_2 \\ a_3 \end{pmatrix} \quad \Rightarrow \quad \frac{1}{2} \begin{pmatrix} 1 \\ a_- \\ a_- \end{pmatrix} \quad (47)$$

where

$$a_- = -\frac{3 + \sqrt{6}}{2 + \sqrt{6}} = -\sqrt{\frac{3}{2}} \approx -1.22474 \quad (48)$$

Finally, repeat for the remaining eigenfrequency and we get

$$\omega^2 = (3 - \sqrt{6})\frac{g}{\ell} \quad \Rightarrow \quad -mg\ell \begin{pmatrix} 3(2 - \sqrt{6}) & (3 - \sqrt{6}) & (3 - \sqrt{6}) \\ (3 - \sqrt{6}) & (2 - \sqrt{6}) & 0 \\ (3 - \sqrt{6}) & 0 & (2 - \sqrt{6}) \end{pmatrix} \begin{pmatrix} 1 \\ a_2 \\ a_3 \end{pmatrix} \quad \Rightarrow \quad \frac{1}{2} \begin{pmatrix} 1 \\ a_+ \\ a_+ \end{pmatrix} \quad (49)$$

where

$$a_+ = -\frac{3 - \sqrt{6}}{2 - \sqrt{6}} = \sqrt{\frac{3}{2}} \approx +1.22474 \quad (50)$$

What we have arrived at is the following description of the motion: The mode associated with the frequency $\omega^2 = g/\ell$ has the upper mass stationary and the lower masses oscillating with the same amplitude, but π radians out of phase. The frequency $\omega^2 = (3 + \sqrt{6})g/\ell$ corresponds to the two lower mass in phase and same amplitude, but π radians out of phase with the upper mass which has a smaller amplitude. The final mode corresponds to the two lower masses in phase with the upper mass, but the upper mass has a smaller amplitude.

Core Paper

MATC 2.3

Block - II

Marks : 32 (SEE : 25; IA : 07)

Abstract Algebra - II

Syllabus

• Unit 6 •

Preliminaries: Review of earlier related concepts-Rings, integral domains, fields and their simple properties.

• Unit 7 •

Classification of rings, their definitions and characterization theorem with examples and counter examples. Polynomial rings, division algorithm.

• Unit 8 •

Ideals in Rings: Definitions, classifications with related theorems examples and counter examples.

• Unit 9 •

Domains in Rings: Classification, definitions and related theories with example and counter examples, irreducible polynomials, Eisenstein's criterion for irreducibility.

• Unit 10 •

Fields: Definition and simple properties.

Unit VI

Course Structure

- Rings and properties
 - Integral Domains and properties
 - Fields and properties
-

4 Rings

We earlier read that a non-empty set R with two binary operations '+' and 'o' is said to be a **ring** if it satisfies the following:

1. R is an abelian group with respect to '+'
2. R is a semi-group with respect to 'o'
3. Multiplication distributes over addition, i.e.,

$$(a) \ a(b+c)=ab+ac, \ a,b,c \in R$$

$$(b) \ (b+c)a=ba+ca, \ a,b,c \in R$$

Recall that R is called a ring with *unit element* if there exists an element $e \in R$ such that $ae = a = ea, \forall a \in R$. R is called *commutative only* if it is commutative with respect to 'o'.

Example 4.1. 1. $R = \mathbb{Z}$, is the ring of integers for usual addition and multiplication. It has 1 as unit element and is commutative.

2. $R = n\mathbb{Z}, n \in \mathbb{N}$ is also a ring for usual addition and multiplication. But it has no unit, and is commutative.
3. $R = \mathbb{Z}_n, n \in \mathbb{N}$, the additive abelian group of residue classes modulo n . It is a ring, known to be *ring of residue classes modulo n* . It has $\bar{1}$ as unit element and it is commutative.
4. Let R be the set of all 2×2 matrices with real entries. R forms a ring with respect to usual matrix addition and multiplication. It is not commutative.

Proposition 4.2. Let R be a ring. Then

1. $a0=0=0a, a \in R$
2. $a(-b)=(-a)b=-(ab), a, b \in R$.
3. $(-a)(-b)=ab, a, b \in R$.

The proof of this proposition follows directly from the definition of rings.

4.1 Integral Domains

We know that in a ring R , a is called left zero divisor of b if $ab = 0$. And b is called a right zero divisor, where $a, b \in R$. A commutative ring which has no zero divisors is called an *integral domain*.

Example 4.3. 1. The rings $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$ of integers, rational, real or complex numbers are all integral domains.

2. $\mathbb{Z}_5$ is an integral domain.

3. The ring $C[0,1]$ of real valued continuous functions on $[0,1]$ is not an integral domain because if

$$\begin{aligned} f(t) &= 0, \quad 0 \leq t \leq 1/2 \\ &= t - \frac{1}{2}, \quad 1/2 \leq t \leq 1 \end{aligned}$$

and

$$\begin{aligned} g(t) &= \frac{1}{2} - t, \quad 0 \leq t \leq 1/2 \\ &= 0, \quad 1/2 \leq t \leq 1 \end{aligned}$$

then both $f(t)$ and $g(t)$ belongs to $C[0, 1]$, and both of them are not identically equal to 0. But,

$$fg \equiv 0$$

which shows that $C[0, 1]$ has zero divisors.

4. $\mathbb{Z}_6$ is not an integral domain since $\bar{2}$ and $\bar{3}$ are both non-zero elements; although, we have,

$$\bar{2}.\bar{3} = \bar{6} = \bar{0}$$

showing that $\mathbb{Z}_6$ has zero divisors.

Observe that, in the definition of ring, it is only a semi-group with respect to ' $\circ$ '.

4.2 Fields

Definition 4.4. A commutative ring R with unit element $1 \neq 0$ in which every non-zero element has an inverse with respect to multiplication, is called a field.

In this case, $(R, \circ)$ forms a commutative group which was not the case for a general ring.

Example 4.5. 1. The ring of rational, real or complex numbers are all fields.

2. The ring $\mathbb{Z}_p$, where p is a prime, forms a field with respect to addition and multiplication modulo p . We have seen that $\mathbb{Z}_p$ is a commutative ring with unit element $\bar{1}$. Now, for any non-zero $\bar{a} \in \mathbb{Z}_p$, a is prime to p . So, there exist some λ, μ in $\mathbb{Z}$ such that

$$\lambda a + \mu p = 1$$

Consequently,

$$\bar{\lambda}.\bar{a} = \bar{1}$$

Hence, $\bar{a}$ has an inverse in $\mathbb{Z}_p$ with respect to $\circ$. Hence, the non-zero elements of $\mathbb{Z}_p$ forms a commutative group with respect to $\circ$. It is a finite field.

Definition 4.6. A ring R with unit element in which every non-zero element has an inverse with respect to multiplication is called skew field or division ring.

By definition, every division ring is a field.

Definition 4.7. Let R be a ring with unit element 1. An element $a \in R$ is called a regular element or unit in R if there exists another element b such that $ab = ba = 1$.

Example 4.8. 1. In a field or a skew field, every non-zero element is regular.

2. In the ring of integers $\mathbb{Z}$, the only regular elements are ± 1 .

3. In the ring $\mathbb{Z}_n$ of integers modulo n , the regular elements are the prime residue classes modulo n .

Proposition 4.9. Every field is an integral domain.

Proof. Let R be a field. Suppose $a \in R$, $b \in R$, $a \neq 0$, $b \neq 0$ and $ab = 0$. Since R is a field, and $b \neq 0$, b has an inverse $b' \in R$ such that $b.b' = b'.b = 1$. Now, $(ab)b' = 0.b' = 0$, but, $a(bb') = a.1 = a$. Hence, $a = 0$, which is a contradiction. Thus $ab \neq 0$ showing that R is not an integral domain. $\square$

Corollary 4.9.1. If R is a ring with unit element, and $b \in R$, which is regular, then b is not a zero divisor.

Corollary 4.9.2. $\mathbb{Z}_n$ is not a field, when n is not a prime.

Proof. If n is not a prime, $n = m_1.m_2$, where m_1, m_2 are proper divisors. Hence,

$$\overline{m_1}.\overline{m_2} = \overline{m_1m_2} = \overline{n} = \overline{0},$$

where, $\overline{m_1} \neq 0$, $\overline{m_2} \neq 0$. Hence, $\mathbb{Z}_n$ is not an integral domain. Hence, it is not a field. $\square$

Proposition 4.10. Let R be a finite integral domain. Then R is a field.

Proof. We shall first show that R has a unit element. Let $R = \{a_1, a_2, \dots, a_n\}$ be the distinct elements of R and let a be any non-zero element of R . Then the elements $aa_1, aa_2, aa_3, \dots, aa_n$ are all distinct points of R , because if $aa_i = aa_j$, then $a(a_i - a_j) = 0$ and since R is an integral domain and $a \neq 0$, so $a_i = a_j$. Hence, the set $\{aa_1, aa_2, aa_3, \dots, aa_n\}$ coincide with R . In particular, $aa_k = a$ for some k . We shall show that a_k is the unit element of R . Let a_j be any element of R . Then $a_j = aa_i$ for some i . Now,

$$\begin{aligned} a_j a_k &= a_k a_j \\ &= a_k (aa_i) \\ &= (a_k a) a_i \\ &= (aa_k) a_i \\ &= aa_i \\ &= a_j \end{aligned}$$

Thus, a_k is the unit element of R . The unit element is unique and we shall denote it by 1. Let $a \in R$, $a \neq 0$. Since $1 \in R$, $aa_l = a_l a = 1$, for some a_l . Thus a has an inverse with respect to multiplication showing that R is a field. $\square$

Corollary 4.10.1. For any prime p , $\mathbb{Z}_p$ is a field.

Proof. $\mathbb{Z}_p$ is an integral domain, because if $\bar{a}\bar{b} = 0$, that is, $\overline{ab} = \bar{0}$, then p divides ab . Since p is a prime, p divides either a or b , that is, $\bar{a} = \bar{0}$ or $\bar{b} = \bar{0}$. Hence, $\mathbb{Z}_p$ is an integral domain, and since it is a finite ring, so it is a field. $\square$

We now define the *characteristic* of a ring.

Definition 4.11. Let R be a ring. The *characteristic* of R is the smallest positive integer n if it exists such that $na = 0$ for all $a \in R$. If no such n exists, the characteristic of R is said to be zero. We denote the characteristic of R as $\text{char}R$.

Example 4.12. 1. If $R = \mathbb{Z}$, then $\text{char}R = 0$, because there exists no such positive integer m such that, $m\mathbb{Z} = 0$.

2. $\text{char}\mathbb{Z}_n = n$ since $n\bar{a} = 0$ for all $\bar{a} \in \mathbb{Z}_n$.

Proposition 4.13. Let R be a ring with unit element e . Then

1. $\text{char}R = n$, if n is the smallest positive integer such that $ne = 0$, and
2. $\text{char}R = 0$ if no such n exists.

Proof. 1. Suppose n is the smallest positive integer such that $ne = 0$. Then, for any $a \in R$,

$$\begin{aligned} na &= a + a + \cdots + a (n \text{ times}) \\ &= ae + ae + \cdots + ae (n \text{ times}) \\ &= a(e + e + \cdots + e) \\ &= a(ne) \\ &= a0 \\ &= 0 \end{aligned}$$

Also, for $0 < m < n$, $me \neq 0$. Hence, $\text{char}R = n$.

2. Suppose no such n exists. If possible, let there exists some $m > 0$, with $ma = 0$ for all $a \in R$. In particular, $me = 0$, $m > 0$ which contradicts the given hypothesis. Hence, our assumption is wrong and $\text{char}R$ has to be 0. $\square$

Proposition 4.14. Let R be an integral domain. Then $\text{char}R$ is either 0 or a prime.

Proof. Suppose $\text{char}R = n \neq 0$ and suppose n is not prime. Then $n = m_1.m_2$, where m_1 and m_2 are proper divisors of n . For any $a \in R$, $a \neq 0$, we have,

$$0 = na^2 = (m_1m_2)a^2 = (m_1a)(m_2a).$$

Since R is an integral domain, $m_1a = 0$ or $m_2a = 0$. Suppose $m_1a = 0$. Then we show that $m_1x = 0$ for any $x \in R$. Now, $m_1(xa) = (m_1x)a = x(m_1a) = x0 = 0$. Since $a \neq 0$, and R is an integral domain, $m_1x = 0$. Thus, $m_1x = 0$ for all x , $m_1 < n$. This contradicts the assumption that $\text{char}R = n$. Hence proved. $\square$

Corollary 4.14.1. The characteristic of a field is either 0 or a prime.

- Exercise 4.15.**
1. Give an example of a finite non-commutative ring. Also, give an example of an infinite non-commutative ring that does not have the unity.
 2. Give an example of a commutative ring without divisor of zero that is not an integral domain.
 3. A ring element is called idempotent if $a^2 = a$. Prove that the only idempotent elements in an integral domain are 0 and 1.
 4. Let R be the ring of real valued continuous functions on $[0, 1]$. Show that R has zero divisors.
 5. Let F be a field of order 2^n . Show that $\text{char } F = 2$.
 6. Find the characteristic of $\mathbb{Z}_4 \oplus 4\mathbb{Z}$.

Unit VII

Course Structure

- Classification of rings and properties
 - Polynomial rings, division algorithm
-

5 Ring Homomorphisms

A *ring homomorphism* ϕ from a ring R to S is a mapping from R to S that preserves the two ring operations, that is, for all $a, b \in R$,

1. $\phi(a + b) = \phi(a) + \phi(b)$
2. $\phi(ab) = \phi(a)\phi(b)$.

A ring homomorphism that is both one-to-one and onto is called a ring isomorphism.

Example 5.1. 1. For any positive integer n , the mapping $k \rightarrow k \bmod n$ is a ring homomorphism from $\mathbb{Z}$ onto $\mathbb{Z}_n$. This mapping is called the natural homomorphism from $\mathbb{Z}$ onto $\mathbb{Z}_n$.

2. The mapping $a + bi \rightarrow a - bi$ is a ring isomorphism from the complex numbers onto the complex numbers.
3. Let R be a commutative ring of characteristic 2. Then the mapping $a \rightarrow a^2$ is a ring homomorphism from R to R .

Theorem 5.2. (Properties of Ring Homomorphism) Let ϕ be a ring homomorphism from a ring R to a ring S . Let A be a subring of R .

1. For any $r \in R$ and any positive integer n , $\phi(nr) = n\phi(r)$ and $\phi(r^n) = (\phi(r))^n$.
2. $\phi(A) = \{\phi(a) | a \in A\}$ is a subring of S .
3. If R is commutative, then $\phi(R)$ is commutative.
4. If R has a unity 1 , $S \neq \{0\}$, and ϕ is onto, then $\phi(1)$ is the unity of S .
5. ϕ is an isomorphism if and only if ϕ is onto and $\ker \phi = \{r \in R | \phi(r) = 0\} = \{0\}$.
6. ϕ is an isomorphism from R onto S , then ϕ^{-1} is an isomorphism from S onto R .

Proof. Proofs are similar to those of groups. Left as exercise. □

Theorem 5.3. (First Isomorphism Theorem) Let ϕ be a ring homomorphism from R to S . Then the mapping from $R/\text{Ker}\phi$ to $\phi(R)$, given by $r + \text{Ker}\phi \rightarrow \phi(r)$, is an isomorphism.

Theorem 5.4. Let R be a ring with unity 1. The mapping $\phi : \mathbb{Z} \rightarrow R$ given by $n \rightarrow n.1$ is a ring homomorphism.

Proof. Since the multiplicative group property $a^{m+n} = a^m a^n$ translates to $(m+n)a = ma + na$ when the operation is addition, we have

$$\phi(m+n) = (m+n).1 = m.1 + n.1$$

So, ϕ preserves addition.

ϕ also follows multiplication. We know that $(m.a)(n.b) = (mn)(ab)$ for all integers. Thus,

$$\phi(mn) = (mn).1 = (mn).((1)(1)) = (m.1)(n.1) = \phi(m)\phi(n)$$

So, ϕ preserves multiplication as well. □

Corollary 5.4.1. If R is a ring with unity and characteristic of R is $n > 0$, then R contains a subring isomorphic to $\mathbb{Z}_n$. If the characteristic of R is 0, then R contains a subring isomorphic to $\mathbb{Z}$.

Proof. Let 1 be the unity of ring R and let $S = \{k.1 | k \in \mathbb{Z}\}$. The previous theorem shows that ϕ from $\mathbb{Z}$ to S given by $\phi(k) = k.1$ is a homomorphism, and by the First Isomorphism theorem for rings, we have

$$\mathbb{Z}/\text{Ker}\phi \simeq S.$$

Clearly, $\text{Ker}\phi = \langle n \rangle$, where n is the additive order of 1 and n is also the characteristic of R . So, R has characteristic n ,

$$S \simeq \mathbb{Z}/\langle n \rangle \simeq \mathbb{Z}_n.$$

When R has characteristic 0, then

$$S \simeq \mathbb{Z}/\langle 0 \rangle \simeq \mathbb{Z}.$$

□

Corollary 5.4.2. For any positive integer m , the mapping $\phi : \mathbb{Z} \rightarrow \mathbb{Z}_m$ given by $x \rightarrow x \bmod m$ is a ring homomorphism.

Proof. Left as exercise. □

Corollary 5.4.3. If F is a field of characteristic p , then F contains a subfield isomorphic to $\mathbb{Z}_p$. If F is a field of characteristic 0, then F contains a subfield isomorphic to $\mathbb{Q}$.

5.1 Polynomial rings

We are all familiar with expressions like, $x^2 + 4x + 1$. This is a polynomial with integer coefficients. We can factorise them, find their roots, etc. Likewise, here we will look into those polynomials with coefficients from commutative rings only.

Definition 5.5. Let R be a ring. A polynomial in x with coefficients from R is an expression of the type $f(x) = a_0 + a_1x + \cdots + a_nx^n$, $a_i \in R$, $i = 0, 1, \dots, n$.

Example 5.6. 1. $f(x) = x^4 + 4x^3 + 1$ is a polynomial with coefficients in $\mathbb{Z}$.

2. $f(x) = x^3 + (4 + i)x^2 + 5 + 6i$ is a polynomial with coefficients from the ring of Gaussian integers, that is, $\mathbb{Z}[i] = \{a + bi : a, b \in \mathbb{Z}\}$.

3. If $R = \{a + b\sqrt{-5} : a, b \in \mathbb{Z}\}$, then

$$f(x) = (3 + 6\sqrt{-5})x^7 + 5x^3 + (8 + 7\sqrt{-5})x + (1 + \sqrt{-5})$$

is a polynomial with coefficients from R .

We now formally define addition and multiplication between two polynomials from a set of polynomials with coefficients from R .

Definition 5.7. Let $f(x) = a_0 + a_1x + \cdots + a_nx^n$ and $g(x) = b_0 + b_1x + \cdots + b_mx^m$, where $a_n, b_m \neq 0$ be two polynomials over R . Then their sum and multiplication are defined as,

$$(f + g)(x) = (a_0 + b_0) + (a_1 + b_1)x + \cdots$$

and,

$$(fg)(x) = c_0 + c_1x + c_2x^2 + \cdots + c_{m+n}x^{m+n}$$

where, c_i is defined as

$$c_i = \sum_{k=0}^i a_k b_{i-k}$$

for $i = 0, 1, \dots, m + n$.

Then it is clear that addition is associative since addition in R is so. Also, the zero element is given by the polynomial

$$0(x) = 0 + 0x + 0x^2 + \cdots$$

The inverse element of any polynomial $f(x) = a_0 + a_1x + \cdots + a_nx^n$ is given by $-f(x) = -a_0 - a_1x - \cdots - a_nx^n$. by virtue of all these properties, we have the following theorem:

Theorem 5.8. The set of polynomials over the ring R forms a ring under addition and multiplication. This ring of polynomials is denoted by $R[x]$.

Example 5.9. The ring of polynomials with coefficients from $\mathbb{R}, \mathbb{C}$ are denoted by $\mathbb{R}[x], \mathbb{C}[x]$ respectively.

$R[x]$ may or may not have unit element. For example, consider the ring $2\mathbb{Z}$. It has no unit element. The polynomial ring over $2\mathbb{Z}$ also does not have unit element. Also see that $\mathbb{R}, \mathbb{C}$ both have unit element 1 and also their corresponding polynomials have so. So we see that the existence of unit element in R can be a possible requisite for $R[x]$ to contain so. We can see from the following theorem:

Theorem 5.10. Let R be a commutative ring with unit element 1. Then, $R[x]$ is also a commutative ring with unit element.

Proof. Since R is commutative, for any $a, b \in R$, $a.b = b.a$. So, for $f(x) = a_0 + a_1x + \cdots + a_nx^n$, and $g(x) = b_0 + b_1x + \cdots + b_mx^m$ in $R[x]$, we have

$$(fg)(x) = c_0 + c_1x + c_2x^2 + \cdots + c_{m+n}x^{m+n}$$

where, c_i is given as

$$\begin{aligned} c_i &= \sum_{k=0}^i a_k b_{i-k} \\ &= \sum_{k=0}^i b_{i-k} a_k \end{aligned}$$

since R is commutative. This shows that, $fg = gf$, for any polynomials $f(x)$ and $g(x)$ in $R[x]$. Also, if we define the polynomial $1(x) = 1 + 0x + 0x^2 + \cdots$, then we can easily check that, for any polynomial $f(x) \in R[x]$, we have $f.1 = f$. So, the above defined polynomial is the unit element in $R[x]$. $\square$

All the examples of polynomial rings given before are polynomial rings with unity. The next theorem that directly follows from the above is

Theorem 5.11. If R is an integral domain, then $R[x]$ is so.

Proof. Since, R is an integral domain, it is a commutative ring with unity. By the previous theorem, we can say that $R[x]$ is also so. We are left only to show that $R[x]$ does not contain any divisors of zero. For this, let us assume that $f(x) = a_0 + a_1x + \cdots + a_nx^n$, and $g(x) = b_0 + b_1x + \cdots + b_mx^m$ are two polynomials in $R[x]$, with $a_n, b_m \neq 0$. So, we have,

$$(fg)(x) = a_0b_0 + (a_1b_0 + b_1a_0)x + \cdots + a_nb_mx^{m+n}$$

Since R does not contain any zero divisor, and $a_n, b_m \neq 0$, so $a_nb_m \neq 0$. Hence, $fg \neq 0$. Thus, $R[x]$ does not contain any zero divisor. Hence the result. $\square$

Corollary 5.11.1. If R is a field, then $R[x]$ is an integral domain.

Definition 5.12. Let $f(x)$ be a non-zero polynomial in $R[x]$. Then the largest n for which the coefficient of x^n is non-zero, is called the *degree* of f . It is generally denoted by $\deg f$.

Theorem 5.13. Let R be any commutative ring, $f(x), g(x) \in R[x]$. Then $\deg(fg) \leq \deg(f) + \deg(g)$ and equality holds when R is an integral domain.

Proof. Let $f(x) = a_0 + a_1x + \cdots + a_nx^n$, $a_n \neq 0$, so that $\deg f = n$ and let $g(x) = b_0 + b_1x + \cdots + b_mx^m$, $b_m \neq 0$ with $\deg g = m$. Then, $f(x)g(x) = a_0b_0 + (a_0b_1 + a_1b_0)x + \cdots + a_nb_mx^{m+n}$. Thus, $\deg(fg) \leq m + n = \deg f + \deg g$. If R is an integral domain, then $a_nb_m \neq 0$. So, $\deg(fg) = m + n = \deg f + \deg g$. $\square$

Corollary 5.13.1. If F is a field, $\deg(fg) = \deg f + \deg g$ and in particular $\deg(fg) \geq \deg f$, as $\deg g \geq 0$.

Definition 5.14. If $f(x) \in R[x]$ is such that $\deg f = 0$, it is called a constant polynomial.

If $\deg f = 0$, then $f(x) = a_0 \in R$ and conversely, if $a \in R$, then a can be written as $a + 0x + 0x^2 + \cdots$ of degree 0. Thus the constant polynomials can be identified with elements of R and under this identification, R is a subring of $R[x]$.

5.2 Division Algorithm

Theorem 5.15. Let F be a field and let $f(x), g(x) \in F[x]$ with $g(x) \neq 0$. Then there exists unique polynomials $q(x)$ and $r(x)$ in $F[x]$ such that

$$f(x) = g(x)q(x) + r(x)$$

and either $r(x)=0$ or $\deg r(x) < \deg g(x)$.

Proof. If $f(x) = 0$ or $\deg f(x) < \deg g(x)$, we simply set $q(x) = 0$ and $r(x) = f(x)$. So, we assume that $n = \deg f(x) \geq \deg g(x) = m$. Let

$$\begin{aligned} f(x) &= a_n x^n + \cdots + a_0, \text{ and} \\ g(x) &= b_m x^m + \cdots + b_0 \end{aligned}$$

By, doing long division method, we let

$$f_1(x) = f(x) - a_n b_m^{-1} x^{n-m} g(x)$$

Then, $f_1(x) = 0$ or $\deg f_1(x) < \deg g(x)$. By induction hypothesis, there exists $q_1(x)$ and $r_1(x)$ in $F[x]$ such that

$$f_1(x) = g(x)q_1(x) + r_1(x)$$

where, $r_1(x) = 0$ or $\deg r_1(x) < \deg g(x)$. Thus,

$$\begin{aligned} f(x) &= a_n b_m^{-1} x^{n-m} g(x) + f_1(x) \\ &= a_n b_m^{-1} x^{n-m} g(x) + g(x)q_1(x) + r_1(x) \\ &= [a_n b_m^{-1} x^{n-m} + q_1(x)]g(x) + r_1(x) \end{aligned}$$

So the polynomials

$$\begin{aligned} q(x) &= a_n b_m^{-1} x^{n-m} + q_1(x), \text{ and} \\ r(x) &= r_1(x) \end{aligned}$$

have the desired properties.

To prove the uniqueness, let

$$\begin{aligned} f(x) &= g(x)q(x) + r(x) \text{ and} \\ f(x) &= g(x)\bar{q}(x) + \bar{r}(x) \end{aligned}$$

where, and either $r(x) = 0$ or $\deg r(x) < \deg g(x)$ and and either $\bar{r}(x) = 0$ or $\deg \bar{r}(x) < \deg g(x)$. Subtracting these two equations, we obtain

$$\begin{aligned} 0 &= g(x)[q(x) - \bar{q}(x)] + [r(x) - \bar{r}(x)] \\ \bar{r}(x) - r(x) &= g(x)[q(x) - \bar{q}(x)] \end{aligned}$$

Thus, $\bar{r}(x) - r(x) = 0$, or the degree of $\bar{r}(x) - r(x)$ is at least that of $g(x)$. Since the latter is impossible, we have $\bar{r}(x) = r(x)$ and $q(x) = \bar{q}(x)$ as well. $\square$

The polynomials $q(x)$ and $r(x)$ in the division algorithm are called *quotient* and *remainder* in the division.

5.3 Remainder Theorem

Corollary 5.15.1. Let F be a field, $a \in F$, and $f(x) \in F[x]$. Then $f(a)$ is the remainder in the division of $f(x)$ by $x - a$.

Proof. Left as exercise. □

5.4 Factor Theorem

Corollary 5.15.2. Let F be a field, $a \in F$, and $f(x) \in F[x]$. Then a is the zero of $f(x)$ if and only if $x - a$ is a factor of $f(x)$.

Proof. Left as exercise. □

Corollary 5.15.3. A polynomial of degree n over a field has at most n zeros, counting multiplicities.

Proof. We use induction over n . Clearly, a polynomial of degree 0 over a field has no zeros. Now, let $f(x)$ is a polynomial of degree n over a field and a is a zero of $f(x)$ of multiplicity k . Then

$$\begin{aligned} f(x) &= (x - a)^k q(x) \text{ and} \\ q(a) &\neq 0 \end{aligned}$$

and since

$$\begin{aligned} n &= \deg f(x) \\ &= \deg (x - a)^k q(x) \\ &= k + \deg q(x) \end{aligned}$$

we have $k \leq n$. If $f(x)$ has no zeros other than a , we are done. On the other hand, if $b \neq a$ and b is a zero of $f(x)$, then

$$0 = f(b) = (b - a)^k q(b)$$

so that b is also a zero of $q(x)$ with the same multiplicity as it has for $f(x)$. By the Second Principle of Induction, we know that $q(x)$ has at most $\deg q(x) = n - k$ zeros, counting multiplicity. Thus $f(x)$ has at most $k + n - k = n$ zeros, counting multiplicity. □

Exercise 5.16. 1. What are the units of the ring $\mathbb{Z}_7[x]$?

2. Give examples of two polynomials such that $\deg fg < \deg f + \deg g$.
3. If R and R' are two isomorphic rings, show that $R[x]$ and $R'[x]$ are isomorphic.
4. Show that for any R , $R[x]$ can never be a field.

Unit VIII

Course Structure

- Definition of Ideals
 - Classification of ideals
-

6 Ideals in rings

Let R be a ring and $I \subset R$. I is called a left(right) ideal of R if

1. $a - b \in I$ where $a, b \in I$.
2. For each $a \in I$ and $x \in R$, $xa \in I$ ($ax \in I$).

Thus an ideal is both a left and right ideal, and in a commutative ring, any left or right ideal is an ideal.

Example 6.1. 1. For any ring R , $\{0\}$ and R are ideals of R . These are called the *trivial ideals* of R .

2. For any positive integer n , the set $n\mathbb{Z} = \{nz : z \in \mathbb{Z}\}$ is an ideal of $\mathbb{Z}$.

3. Let R be a commutative ring with unit element and let $a_1, a_2, \dots, a_n \in R$. Then $I = \langle a_1, a_2, \dots, a_n \rangle = \{r_1a_1 + r_2a_2 + \dots + r_na_n : r_i \in R\}$ is an ideal of R called the *ideal generated by* $a_1, a_2, \dots, a_n$. This is the smallest ideal containing $a_1, a_2, \dots, a_n$.

Also, if I is an ideal of R , then I forms a normal subgroup of $(R, \circ)$ (check it).

Let R be a ring and A be an ideal of R . Since R is a group under addition and A is a normal subgroup of R , we may form the factor group $R/A = \{a + A : a \in R\}$. Now, we come to the question that whether it forms a ring of cosets. Yes, it is indeed possible if A is an ideal of R .

Theorem 6.2. (Existence of factor rings) Let R be a ring and A be a subring of R . The set of cosets is a ring under the operations

1. $(s+A) + (t+A) = s+t+A$,
2. $(s+A)(t+A) = st+A$

$s, t \in R$, if and only if A is an ideal of R .

Proof. Proof left as an exercise. □

Example 6.3. 1. $\mathbb{Z}/4\mathbb{Z} = \{0+4\mathbb{Z}, 1+4\mathbb{Z}, 2+4\mathbb{Z}, 3+4\mathbb{Z}\}$ is a factor ring under the two operations modulo 4.

2. $2\mathbb{Z}/6\mathbb{Z}$ is a non-commutative example of an ideal and factor ring.

Theorem 6.4. Every ideal of a ring R is the kernel of a ring homomorphism of R . In particular, an ideal A is the kernel of the mapping

$$r \longrightarrow r + A$$

from R to R/A .

The homomorphism given above is called the *natural* homomorphism for R to R/A . Let R be a commutative ring with unity and let $a \in R$. The set $\langle a \rangle = \{ra \mid r \in R\}$ is an ideal of R called the *principal ideal* generated by a .

Definition 6.5. A *prime ideal* A of a commutative ring R is a proper ideal of R such that $a, b \in R$ and $ab \in A$ imply either $a \in A$ or $b \in A$. A *maximal ideal* of a commutative ring R is a proper ideal of R such that whenever B is an ideal of R and $A \subseteq B \subseteq R$, then either $B = A$ or $B = R$.

Example 6.6. 1. Let n be an integer greater than 1. Then, in the ring of integers, the ideal $n\mathbb{Z}$ is prime if and only if n is prime.

2. $\langle 2 \rangle$ and $\langle 3 \rangle$ are maximal ideals of $\mathbb{Z}_{36}$.

3. $\langle x^2 + 1 \rangle$ is maximal in $\mathbb{R}[x]$.

Example 6.7. Since the mapping ϕ from $\mathbb{Z}[x]$ onto $\mathbb{Z}$ given by $\phi(f(x)) = f(0)$ is a ring homomorphism with $\text{Ker}\phi = \langle x \rangle$, we have

$$\mathbb{Z}[x]/\langle x \rangle \simeq \mathbb{Z}$$

Theorem 6.8. In a commutative ring R , let A be an ideal. If a unit element is in A , then $A=R$.

Proof. Let $u \in A$ be a unit element of R . Let $v \in R$ be such that $uv = 1$. Now, since A is an ideal of R , so for $u \in A$ and $v \in R$, we must have $uv \in A$, i.e., $1 \in A$.

Now, for every $r \in R$, we must have $r1 \in A$, i.e., $r \in A \forall r \in R$. Thus, $A = R$. $\square$

Theorem 6.9. Let R be a commutative ring with unity and let A be an ideal of R . Then, R/A is an integral domain if and only if A is prime.

Proof. Let R/A is an integral domain and $ab \in A$. Then, $(a + A)(b + A) = ab + A = A$, the zero element of the ring R/A . So, either $a + A = A$ or $b + A = A$, that is, either $a \in A$ or $b \in A$. Hence, A is prime.

We observe that R/A is a commutative ring with unity for any proper ideal A . Our task is only to show that R/A has no zero divisors. So, let A be prime ideal and $(a + A)(b + A) = 0 + A = A$. Then, $ab \in A$ and hence, $a \in A$ or $b \in A$. Thus, one of $a + A$ or $b + A$ is the zero coset in R/A . $\square$

Theorem 6.10. Let R be a commutative ring with unity and let A be an ideal of R . Then, R/A is a field if and only if A is maximal.

Proof. Let R/A be a field and B is an ideal of R that properly contains A . Let $b \in B$ but $b \notin A$. Then $b + A$ is a nonzero element of R/A and hence there exists an element $c + A$ such that $(b + A)(c + A) = 1 + A$, the multiplicative identity of R/A . Since $b \in B$, we have $bc \in B$. Since

$$1 + A = (b + A)(c + A) = bc + A,$$

we have $1 - bc \in A \subset B$. So, $1 = (1 - bc) + bc \in B$. So, $B = R$. Thus, A is maximal. Now let A is maximal and $b \in R$ but $b \notin A$. It suffices to show that $b + A$ has a multiplicative inverse. Consider $B = \{br + a | r \in R, a \in A\}$. This is an ideal in R that properly contains A . Since A is maximal, we must have $B = R$. Thus, $1 \in B$, and $1 = bca'$ (say), where $a' \in A$. Then

$$1 + A = bc + a' + A = bc + A = (b + A)(c + A).$$

□

Since $\mathbb{Z}$ is an integral domain but not a field, and since

$$\mathbb{Z}[x] / \langle x \rangle \simeq \mathbb{Z}$$

we must have from the previous theorems, ideal $\langle x \rangle$ is prime but not maximal.

Theorem 6.11. In a commutative ring with unity, every maximal ideal is a prime ideal.

But the converse of the theorem is not true in general. For example, $\langle 0 \rangle$ is a prime ideal in $\mathbb{Z}$ but not a maximal ideal.

6.1 Factorization

We assume throughout this section that R is an integral domain with unit element.

Definition 6.12. Let $a \in R$ and $b \in R$, $a \neq 0$. a is said to divide b if there exists $c \in R$ such that $b = ac$.

Example 6.13. 1. In $R = \mathbb{Z}$, 3 divides 15.

2. In $R = \mathbb{Z} + i\mathbb{Z} = \{a + ib | a, b \in \mathbb{Z}\}$, $(1+3i)$ divides 10 as $10 = (1+3i)(1-3i)$.

Definition 6.14. Two non-zero elements a and b are said to be *associates* if $a|b$ and $b|a$.

That is, a and b are associates if and only if $b = au$, for some unit $u \in R$.

Definition 6.15. Any element a in ring R is called *irreducible* if

1. a is not a unit,
2. the only divisors of a are units and associates of a

Example 6.16. 1. Let $R = \mathbb{Z}$ and $n > 1$. Then n is irreducible if and only if the only divisors of n are 1, -1, n , $-n$. Thus n is a prime integer.

2. In $R = \mathbb{Z}$, 5 and -5 are associates as $-5 = (-1) \cdot 5$.

3. Let $R = \{a + b\sqrt{-5} \mid a, b \in \mathbb{Z}\}$. Then, $1 + 2\sqrt{-5}$ is an irreducible element of R .

Definition 6.17. Let $p \in R$ which is not a unit, p is called a *prime* element if whenever $p \mid ab$, then either $p \mid a$ or $p \mid b$.

Theorem 6.18. Every prime element is irreducible.

Proof. Let p is prime, and let a be any divisor of p so that $p = ab$. Since p is a prime $p \mid a$ or $p \mid b$. If $p \mid a$, since we have $a \mid p$ we have a is an associate of p and b is a unit. Similarly, if $p \mid b$, then b is an associate of p and a is a unit. Thus the only divisors of p are units and associates of p . Hence, p is irreducible. $\square$

In general rings, the converse of the theorem is not true. But we know, in $R = \mathbb{Z}$, every irreducible element is prime.

Proposition 6.19. An element $p \in R$ is prime if and only if the ideal $\langle p \rangle$ is prime ideal.

Proof. Let p is a prime element and let $ab \in \langle p \rangle$. Then $ab = cp$ for some $c \in R$. Since, p is a prime, $p \mid a$ or $p \mid b$, i.e., $a \in P$ or $b \in P$. Hence, P is a prime ideal. Conversely, let P be a prime ideal and let $p \mid ab$. Then $ab = cp \in P$ and since P is a prime ideal, $a \in P$ or $b \in P$, i.e., $p \mid a$ or $p \mid b$. Hence, p is a prime element. $\square$

Definition 6.20. Let $a \in R$ and $b \in R$. An element $d \in R$ is called a greatest common divisor(gcd) of a and b if

1. $d \mid a$ and $d \mid b$,
2. whenever $d' \mid a$ and $d' \mid b$, then $d' \mid d$.

Example 6.21. In $R = \mathbb{Z}$, if $a = 9$, $b = -48$ then $d = 3$ is a gcd of a, b .

Example 6.22. Let R be an integral domain and let a and b are two elements in R .

1. b divides a if and only if $Ra \subset Rb$ and vice versa. The proof is obvious since if $b \mid a$ there exists some c in R , such that $a = bc$. Thus, for any x in R , we have

$$\begin{aligned} x &= ya \\ &= ybc \\ &= ycb \end{aligned}$$

Hence, $x \in Rb$. Thus, $Ra \subset Rb$. the converse part is left as an exercise.

2. a and b are associates if and only if $Ra = Rb$. Since a and b are associates, $a \mid b$ which yields $Rb \subset Ra$; and $b \mid a$ which gives $Ra \subset Rb$. Combining, we get the desired result. Thse converse part is left as exercise.

Example 6.23. Let R be an integral domain. If $a \in R$ such that Ra is a maximal ideal, then a is irreducible. Since Ra is maximal ideal, a is not a unit. If possible, let a be reducible. Then there exists b and c in R such that $a = bc$. Then b and c are proper divisors of a . Hence, by the previous example, $Ra \subset Rb$ and $Ra \subset Rc$. Which is impossible since Ra is a maximal ideal in R .

- Exercise 6.24.**
1. Show that the union of a chain $I_1 \subset I_2 \subset \dots$ of ideals of a ring R is an ideal of R
 2. Show that the product of an irreducible and a unit is irreducible in an integral domain. Is the result true in any arbitrary ring?
 3. Let D be an integral domain. Define $a\rho b$ if a and b are associates. Show that ' ρ ' is an equivalence relation on R .

Unit IX

Course Structure

- Domains in rings
 - Classification of domains
 - Irreducible Polynomials and Eisenstein's Criterion for irreducibility
-

7 Euclidean Domain

Definition 7.1. A *Euclidean domain* is an integral domain R in which there exists an integer valued function d on the non-zero elements of R , satisfying the following conditions:

1. $d(a) \geq 0$ for all non-zero $a \in R$.
2. $d(ab) \geq d(a)$, $a, b \in R$
3. For $a, b \in R$, $b \neq 0$ there exists $q, r \in R$ such that $a = bq + r$ with either $r = 0$ or $d(r) < d(b)$.

Example 7.2. 1. If $R = \mathbb{Z}$, then d can be defined on R as $d(a) = |a|$. Then check that d satisfies all the conditions of an Euclidean domain.

2. Let $R = \{a + ib : a, b \in \mathbb{Z}\}$ be the ring of Gaussian Integers. The if d is defined as $d(z) = |z|^2 = a^2 + b^2$, where, $z = a + ib$, then R forms a Euclidean Domain.
3. Let R be a field. Then, d can be defined as, $d(a) = a \cdot a^{-1}$. This definition is well defined since each element a has an inverse. d satisfies all the properties of an integral domain.

Theorem 7.3. Let R be an Euclidean Domain. Then, every ideal I of R is of the form $I = Ra$ for some $a \in R$.

Proof. If $I = 0$, then $a = 0$ and we are done. So, we take $I \neq 0$. Choose some $a \neq 0$ such that $d(a)$ is the least in R . Such an a exists by well-ordering principle of real numbers. We claim that $I = Ra$. Since $a \in R$, so $Ra \subset I$. Let $b \in I$. By the third condition of the definition of Euclidean domain, there exists q and r in R such that $b = aq + r$, where, either $r = 0$ or $d(r) < d(a)$. Now, $r = b - aq$ is in R . If $d(r) < d(a)$, then this contradicts the fact that a is the element with the least value of d . Then obviously $r = 0$. Thus, we get $b = aq$. Hence, $I \subset Ra$. Combining, we get $I = Ra$. $\square$

Theorem 7.4. Let R be a Euclidean Domain. Then any two elements a, b in R , have a gcd .

Proof. Let $I = Ra$ and $J = Rb$. Then $I + J$ is also an ideal of R . By the previous theorem, we have $I + J = Rd$ for some d in R . We claim that d is the gcd of a and b . We can see that $I = Ra \subset Rd$ and $J = Rb \subset Rd$. Hence, $d|a$ and $d|b$. Let d' be another element in R such that $d'|a$ and $d'|b$. So, $Ra \subset Rd'$ and $Rb \subset Rd'$. This implies that $Rd = I + J \subset Rd'$. Hence, $d'|d$. Hence, by definition, d is the required gcd. $\square$

Theorem 7.5. Let R be a Euclidean domain. Then, $a \in R$ is a unit if and only if $d(a) = d(1)$.

Proof. Let a be a unit in R . Then, $a.a^{-1} = 1$. By the second condition of d , $d(1) = d(a.a^{-1}) \geq d(a)$. Also, $d(a) = d(a.1) \geq d(1)$. So, combining, we get $d(a) = d(1)$. Conversely, let $d(a) = d(1)$. By the third condition of d , we have $1 = qa + r$, with either $r = 0$ or $d(r) < d(a) = d(1)$. But, $d(r) < d(1)$ is impossible since $d(r) = d(r.1) \geq d(1)$. Hence, $r = 0$. Thus, $aq = 1$. Hence, a is a unit in R . $\square$

Theorem 7.6. Let R be a Euclidean Domain and a, b be two elements in R . If a is a proper divisor of b , then $d(a) < d(b)$.

Proof. Since $a|b$, there exists some $c \in R$ such that $b = ac$. So we have, $d(b) \geq d(a)$. We will show that $d(b) \neq d(a)$. If possible, $d(a) = d(b)$. Then, for any $x \in Ra$ we have $d(x) \geq d(a) = d(b)$ which means that $d(b)$ is least in Ra . This implies that, $Ra = Rb$, which implies that a and b are associates which is a contradiction. Hence the result. $\square$

Theorem 7.7. Let R be an Euclidean Domain. Then any $a \in R$ which is not a unit can be expressed as a product of irreducible elements.

Proof. If a is irreducible, then there is nothing to prove. Otherwise, it has a proper divisor b , that is, $a = bc$. Then, $d(b) < d(a)$ and $d(c) < d(a)$. If b and c are irreducible, then we are done. If b (or c) is irreducible, then we can write $b = ef$, where $d(e) < d(b)$ and $d(f) < d(b)$. If this process is continued, then after a finite number of stages all the factors will be irreducible since the values of d is strictly reducing at each stage. Thus, after a finite number of steps, all the factors will be irreducible and a will be expressible as the product of irreducible elements. $\square$

Example 7.8. Let R be an Euclidean Domain and let

$$I = \{a : d(a) > d(1)\} \cup \{0\}$$

Then I does not form an ideal in R . In fact, if we consider the ring $\mathbb{Z}$, the only units in $\mathbb{Z}$ are ± 1 . Now, 2, 3 belong to I but $2 - 3$ does not belong to I . Hence, I is not an ideal in this case.

Example 7.9. Let R be a Euclidean Domain. If d' be another function defined as $d'(a) = d(a) + m$, for some positive integer m , with $d(1) + m \geq 0$. Check that R is a Euclidean Domain with respect to the function d' .

Example 7.10. If R is a Euclidean Domain and for a, b in R , if $a|b$ and $d(a) = d(b)$, then, a and b are associates.

7.1 Principal Ideal Domain

Definition 7.11. An integral domain R is called a *Principal ideal domain*(PID) if every ideal of R is a principal ideal, that is, for any ideal I in R , there exists an element a in R such that $I = \langle a \rangle = Ra$.

It is clear from theorem 7.7, we can say that every Euclidean Domain is a Principal Ideal Domain. But the converse is not necessarily true. For example, the ring $R = \{a + \frac{b}{2}(1 + \sqrt{-19}) : a, b \in \mathbb{Z}\}$ is a Principal ideal domain, but not a Euclidean domain.

Theorem 7.12. Any two elements in a PID R have gcd.

Proof. Similar as before. □

Theorem 7.13. Let R be a PID. Then every $a \in R$ can be expressed as a product of irreducible elements.

Proof. If $a \in R$ is irreducible, then there is nothing to prove. Otherwise, let $a = bc$, where b, c are proper divisors of a . If both b, c are irreducible, then we are done. Suppose, b is irreducible. Then we have

$$b = ef$$

where e, f are proper divisors of b . If we continue in this way, after a finite number of steps, all the factors will be irreducible for otherwise, there will be an infinite sequence of elements,

$$a_0 = a, a_1 = b, a_2 = e, \dots, a_n, \dots$$

such that a_{n+1} is a proper divisor of a_n . We will show that this is impossible.

Let such a sequence exists. Let $I_n = Ra_n$, so that we have an increasing sequence of ideals

$$I_0 \subset I_1 \subset I_2 \cdots I_n \cdots$$

Since a_{n+1} is a proper divisor of a_n , $I_n \neq I_{n+1}$. Let

$$I = \bigcup_{k=0}^{\infty} I_k$$

Then I is an ideal of R , because if $a, b \in I$, then $a \in I$ and $b \in I$, where either $I_r \subset I_s$ or $I_s \subset I_r$, so that $a - b \in I_r \cup I_s \subset I$, and $xa \in I_r \subset I$ for all $x \in R$. Since R is a principal ideal domain, $I = \langle d \rangle$ for some $d \in R$. Now, $d \in I_m$ for some m , so that $I = \langle d \rangle \subset I_m \subset I_{m+1} \subset \cdots \subset I$, i.e., $I_m = I_{m+1} = \cdots = I$, a contradiction. Hence proved. □

Example 7.14. The ring $R = \{m/n : m, n \in \mathbb{Z}, n \text{ odd}\}$ is a principal ideal domain.

7.2 Unique Factorisation Domain

Definition 7.15. An integral domain R is called a *factorisation domain* if every element $a \in R$, which is not a unit can be expressed as a product of irreducible elements.

Thus, the Euclidean domain and principal ideal domains are factorisation domains.

Definition 7.16. $a \in R$ is said to be expressible uniquely as a product of irreducible elements if whenever $a = p_1 p_2 \dots p_m = q_1 q_2 \dots q_n$, where p_i, q_j are irreducible, then $m = n$ and each $p_i = u_i q_i$, where u_i is a unit in some order.

If $a \in R$ be expressed as a product of irreducible elements, the expression need not be unique.

Definition 7.17. An integral domain R is said to be a *unique factorisation domain* if every $a \in R$ which is not a unit can be expressed uniquely as a product of irreducible elements.

Theorem 7.18. Let R be an integral domain in which

1. Every $a \in R$ which is non-unit can be expressed as a product of irreducible elements.
2. Every irreducible element is prime.

Then R is a ufd.

Proof. It is sufficient to show that factorisation is unique. Let $a = p_1 p_2 \dots p_m = q_1 q_2 \dots q_n$, where p_i, q_j are irreducible, and hence prime. Since $p_1 | a$, we have, $p_1 | q_1 q_2 \dots q_n$. So, $p_1 | q_j$ for some j . Without any loss of generality, let us assume that $p_1 | q_1$. Since q_1 is irreducible, and p_1 is not a unit, p_1 is an associate of q_1 , that is, $q_1 = u_1 p_1$, where u_1 is a unit. Thus, $p_1 p_2 \dots p_m = (u_1 p_1) q_2 \dots q_n$. Since R is an integral domain, we have, $p_2 p_3 \dots p_m = u_1 q_2 q_3 \dots q_n$. Repeating the same process for p_2 , we have, $q_2 = u_2 p_2$, where u_2 is a unit. Continuing this process we must have neither p_t nor q_t left after a finite number of steps, for otherwise, in either case, a unit will be expressible as a product of irreducible elements, which is impossible. Thus, $m = n$, and each $p_t = u_i q_i$, where u_i is a unit. $\square$

Corollary 7.18.1. Let R be a Euclidean domain or a PID. Then, R is a ufd.

Proof. It is sufficient to show that every irreducible element is prime. Let p be an irreducible element and $p | ab$. Consider the $\gcd(p, a)$ of p and a . It is either p or 1, and if $\gcd(p, a) = p$, then $p | a$. If $\gcd(p, a) = 1$, then $\lambda p + \mu a = 1$ for some $\lambda, \mu \in R$. Multiplying both sides by b , we have, $\lambda pb + \mu ab = b$. Since $p | ab$, it follows that $p | b$. Hence p is a prime. $\square$

There exists ufd's which are not PID's (or ED's).

Theorem 7.19. If R is a ufd, then any two elements of R has a gcd.

Theorem 7.20. If F is a field, then $F[x]$ is a principal ideal domain.

Proof. We know that $F[x]$ is an integral domain. Let I be an ideal in $F[x]$. If $I \neq \{0\}$, then among all the elements of I , let $g(x)$ be one of minimum degree. We will show that $I = \langle g(x) \rangle$. Since $g(x) \in I$, we have $\langle g(x) \rangle \subseteq I$. Now, let $f(x) \in I$. Then by division algorithm, we may write,

$$f(x) = g(x)q(x) + r(x)$$

and either $r(x)=0$ or $\deg r(x) < \deg g(x)$. Since $r(x) = f(x) - g(x)q(x) \in I$, the minimality of $\deg g(x)$ implies that the latter condition cannot hold. So, $r(x) = 0$ and hence $f(x) \in \langle g(x) \rangle$. This shows that $I \subseteq \langle g(x) \rangle$. $\square$

Theorem 7.21. Let F be a field, I a non-zero ideal in $F[x]$, and $g(x)$ an element of $F[x]$. Then $I = \langle g(x) \rangle$ if and only if $g(x)$ is a nonzero polynomial of minimum degree in I .

7.3 Irreducible Polynomials

Recollect all that we have learnt in the preceding units concerning polynomial rings. We will be defining an irreducible element in $R[x]$ for an integral domain R .

Definition 7.22. Let R be an integral domain. A polynomial $f(x) \in R[x]$ which is neither the zero polynomial nor a unit in $R[x]$ is said to be irreducible over R if, whenever $f(x)$ is expressed as $f(x) = g(x)h(x)$, where $h(x), g(x) \in R[x]$, then $g(x)$ and $h(x)$ are units in $R[x]$. A nonzero, nonunit element in $R[x]$ that is not irreducible over R is said to be reducible over R .

When R is a field, then it can be more conveniently said that a polynomial $f(x)$ is irreducible if it can't be expressed as the product of two polynomials of lower degree.

Example 7.23. The polynomial $f(x) = 2x^2 + 4$ is irreducible over $\mathbb{Q}$ but reducible over $\mathbb{Z}$, since $2x^2 + 4 = 2(x^2 + 2)$ and neither 2 nor $x^2 + 2$ is a unit in $\mathbb{Z}$.

In fact, we will soon see that any polynomial reducible over $\mathbb{Q}$ is reducible over $\mathbb{Z}$ but the converse is not true as we have seen in the above example.

Example 7.24. The polynomial $f(x) = 2x^2 + 4$ is irreducible over $\mathbb{R}$ but reducible over $\mathbb{C}$.

Example 7.25. The polynomial $f(x) = x^2 - 2$ is irreducible over $\mathbb{Q}$ but reducible over $\mathbb{R}$ since $x^2 - 2 = (x - \sqrt{2})(x + \sqrt{2})$.

In general it is not a very easy job to decide whether a given polynomial is irreducible over some given integral domain. So, we have certain theorems to help us decide that. Our first such theorem is

Theorem 7.26. Let F be a field. If $f(x) \in F[x]$ and $\deg f(x) = 2$, or 3, then $f(x)$ is reducible over F if and only if $f(x)$ has a zero in F .

Proof. Let $f(x) = g(x)h(x)$, where $h(x), g(x) \in F[x]$ and have degrees less than that of $f(x)$. Since $\deg f(x) = \deg g(x) + \deg h(x)$, and $\deg f(x)$ is 2 or 3, at least one of $g(x)$ and $h(x)$ has degree 1. Say $g(x) = ax + b$. Then of course $-a^{-1}b$ is a root of $g(x)$ and hence a zero of $f(x)$ too.

Conversely, suppose that $f(a) = 0$, where $a \in F$. Then, by the Factor Theorem, we know that $x - a$ is a factor of $f(x)$ and therefore $f(x)$ is reducible over F . $\square$

One may think whether the theorem is true for polynomials over degree 3. For this, let us consider the following example:

Example 7.27. Consider $f(x) = x^4 + 2x^2 + 1$ over $\mathbb{Q}[x]$. Then clearly, $x^4 + 2x^2 + 1 = (x^2 + 1)(x^2 + 1)$. Hence $f(x)$ is reducible without having zeros in $\mathbb{Q}$.

Of course if $f(x)$ has a root in $R[x]$ then it is reducible. But the converse is true only if the conditions in the above theorem hold.

We have another powerful theorem for checking irreducibility of polynomials in a UFD called the Eisenstein's Criterion. Let's check that out.

7.3.1 Eisenstein's criterion for irreducibility

Definition 7.28. The content of a nonzero polynomial $a_n x^n + \cdots + a_0$, where the a 's are the integers, is the greatest common divisor of the integers $a_n, \dots, a_0$.

A primitive polynomial is an element of $\mathbb{Z}[x]$ with content 1.

The product of two primitive polynomials is primitive.

Theorem 7.29. Let $f(x) \in \mathbb{Z}[x]$. If $f(x)$ is reducible over $\mathbb{Q}$, then it is reducible over $\mathbb{Z}$.

Proof. Let $f(x) = g(x)h(x)$, where $g(x)$ and $h(x) \in \mathbb{Q}[x]$. Clearly, we assume that $f(x)$ is a primitive because we can divide both $f(x)$ and $g(x)$ by the content of $f(x)$. Let a be the least common multiple of the denominators of the coefficients of $g(x)$, and b be the least common multiple of the denominators of the coefficients of $h(x)$. Then

$$abf(x) = ag(x) \cdot bh(x),$$

where $ag(x)$ and $bh(x) \in \mathbb{Z}[x]$. Let c_1 be the content of $ag(x)$ and c_2 be the content of $bh(x)$. Then

$$\begin{aligned} ag(x) &= c_1 g_1(x) \text{ and} \\ bh(x) &= c_2 h_1(x) \end{aligned}$$

where both $g_1(x)$ and $h_1(x)$ are primitive, and

$$abf(x) = c_1 c_2 g_1(x) h_1(x)$$

Since, $f(x)$ is primitive, the content of $abf(x)$ is ab . Also, since the product of two primitives is primitive, it follows that the content of $c_1 c_2 g_1(x) h_1(x)$ is $c_1 c_2$. Thus, $ab = c_1 c_2$ and $f(x) = g_1(x) h_1(x)$, where $g_1(x)$ and $h_1(x) \in \mathbb{Z}[x]$ and $\deg g_1(x) = \deg g(x)$ and $\deg h_1(x) = \deg h(x)$. $\square$

Theorem 7.30. Let

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0 \in \mathbb{Z}[x]$$

If there is a prime p such that p does not divide a_n , $p | a_{n-1}, \dots, p | a_0$ and p^2 does not divide a_0 , then $f(x)$ is irreducible over $\mathbb{Q}$.

This is Eisenstein's Criterion. This theorem can also be stated for any arbitrary UFD.

Proof. If $f(x)$ is reducible over $\mathbb{Q}$ then there exists $g(x)$ and $h(x) \in \mathbb{Z}[x]$ such that

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

$1 \leq \deg g(x)$ and $1 \leq \deg h(x) < n$. Say

$$\begin{aligned} g(x) &= b_r x^r + \cdots + b_0 \text{ and} \\ h(x) &= c_s x^s + \cdots + c_0. \end{aligned}$$

Then, since $p | a_0$, and p^2 does not divide a_0 , and $a_0 = b_0 c_0$, it follows that p divides one of b_0 and c_0 , but not the other. Let $p | b_0$ and not c_0 . Also, since p does not divide $a_n = b_r c_s$, we know that p does not divide b_r . So, there is a least integer t such that p does not divide b_t . Now, consider

$$a_t = b_t c_0 + b_{t-1} c_1 + \cdots + b_0 c_t.$$

By assumption, p divides a_t and by choice of t , every summand on the right after the first one is divisible by p . Clearly, this forces p to divide $b_t c_0$ as well. This is impossible, however, since p is prime and p divides neither b_t nor c_0 . $\square$

Corollary 7.30.1. For any prime p , the p -th cyclotomic polynomial

$$\phi_p(x) = \frac{x^p - 1}{x - 1} = x^{p-1} + x^{p-2} + \cdots + x + 1$$

is irreducible over $\mathbb{Q}$.

Example 7.31. The polynomial $3x^5 + 15x^4 - 20x^3 + 10x + 20$ is irreducible over $\mathbb{Q}$ because 5 does not divide 3 and 25 does not divide 20 but 5 divides 15, -20 , 10 and 20.

Theorem 7.32. Let F be a field and $p(x) \in F[x]$. Then $\langle p(x) \rangle$ is maximal ideal in $F[x]$ if and only if $p(x)$ is irreducible over F .

Proof. Let $\langle p(x) \rangle$ be maximal ideal in $F[x]$. Clearly, $p(x)$ is neither zero nor unit in $F[x]$, because neither $\{0\}$ nor $F[x]$ is a maximal ideal in $F[x]$. If $p(x) = g(x)h(x)$ is a factorization of $p(x)$ over F , then $\langle p(x) \rangle \subseteq \langle g(x) \rangle \subseteq F[x]$. Thus, $\langle p(x) \rangle = \langle g(x) \rangle$ or $F[x] = \langle g(x) \rangle$. In the first case, we must have $\deg p(x) = \deg g(x)$. In the second case, it follows that $\deg g(x) = 0$ and consequently, $\deg h(x) = \deg p(x)$. Thus $p(x)$ cannot be written as product of two polynomials in $F[x]$ of lower degree.

Now, let $p(x)$ is irreducible over F . Let I be any ideal of $F[x]$ such that $\langle p(x) \rangle \subseteq I \subseteq F[x]$. Since $F[x]$ is a principal ideal domain, we know that $I = \langle g(x) \rangle$ for some $g(x) \in F[x]$. So, $p(x) \in \langle g(x) \rangle$ and thus $p(x) = g(x)h(x)$, where $h(x) \in F[x]$. Since, $p(x)$ is irreducible over F , it follows that either $g(x)$ is a constant or $h(x)$ is a constant. In the first case, we have $I = F[x]$; in the second case, we have $\langle p(x) \rangle = \langle g(x) \rangle = I$. So, $\langle p(x) \rangle$ is maximal in $F[x]$. $\square$

Corollary 7.32.1. Let F be a field and $p(x)$ be an irreducible polynomial over F . Then, $F[x]/\langle p(x) \rangle$ is a field.

Proof. This easily follows. $\square$

Corollary 7.32.2. Let F be a field and let $p(x), a(x), b(x) \in F[x]$. If $p(x)$ is irreducible over F and $p(x) | a(x)b(x)$, then $p(x) | a(x)$ or $p(x) | b(x)$.

Proof. Since, $p(x)$ is irreducible, $F[x]/\langle p(x) \rangle$ is a field and, therefore, an integral domain. From one of the previous theorems, we know that $\langle p(x) \rangle$ is a prime ideal, and since $p(x)$ divides $a(x)b(x)$, we have $a(x)b(x) \in \langle p(x) \rangle$. Thus, $a(x) \in \langle p(x) \rangle$ or $b(x) \in \langle p(x) \rangle$. This means that $p(x) | a(x)$ or $p(x) | b(x)$. $\square$

Exercise 7.33. 1. Show that a subring of a PID need not be a PID.

2. If R is a PID, and $p \in R, p \neq 0$. Show that the following conditions are equivalent:

- a) p is a prime
- b) p is an irreducible element
- c) Rp is a prime ideal
- d) Rp is a maximal ideal

3. Show that in a PID, every non-zero prime ideal is maximal.

4. Show that every polynomial with real coefficients of odd degree is reducible.
5. Find all roots of $f(x) = x^3 + x^2 + x + 1$ in $\mathbb{Z}_5$.
6. Let $f(x) = x^2 + 8x - 2$. Show that it is irreducible over $\mathbb{Q}$. Is it irreducible over $\mathbb{R}$?
7. Test for irreducibility of the following polynomial over $\mathbb{Q}$:
 - a) $8x^3 + 6x^2 - 9x + 24$
 - b) $x^4 + 9x + 3$
 - c) $x^5 + 9x^4 + 12x^2 + 6$
8. Show that $x^2 + x + 4$ is irreducible over $\mathbb{Z}_{11}$.

Unit X

Course Structure

- Introduction
 - Objectives
 - Fields
 - Extension of fields
-

8 Field extensions

Definition 8.1. Let F be a field and K be another field containing F . Then K is called an extension of F and is denoted by the symbol K/F .

Example 8.2. 1. Let $F = \mathbb{Q}$, $K = \mathbb{R}$ and $L = \mathbb{C}$, the fields of rational, real, and complex numbers. Then, K and L are extensions of F .

2. Let K be a field of $\text{char } p \geq 0$. If $p = 0$, K contains a field F isomorphic to $\mathbb{Q}$, that is, K is an extension of F . If $p > 0$, K contains a field F isomorphic to $\mathbb{Z}_p$, that is, K is an extension of F .

Definition 8.3. Let K/F be an extension, $a_1, a_2, \dots, a_n \in K$. The smallest subfield of K containing F and $a_1, a_2, \dots, a_n$ is called the field generated by $a_1, a_2, \dots, a_n$ over F and is denoted by $F(a_1, a_2, \dots, a_n)$.

Theorem 8.4.

$$F(a_1, a_2, \dots, a_n) = \left\{ \frac{f(a_1, a_2, \dots, a_n)}{g(a_1, a_2, \dots, a_n)} : f, g \in F[x_1, x_2, \dots, x_n], g(a_1, a_2, \dots, a_n) \neq 0 \right\}$$

Proof. Let

$$L = \left\{ \frac{f(a_1, a_2, \dots, a_n)}{g(a_1, a_2, \dots, a_n)} : f, g \in F[x_1, x_2, \dots, x_n], g(a_1, a_2, \dots, a_n) \neq 0 \right\}$$

Then L is evidently a field for addition and multiplication induced from K . Also, $L \supset F$ as $a \in F$ can be represented as $a = a/1 \in L$. Let K be some other field containing F and $a_1, a_2, \dots, a_n$, then K contains $f(a_1, a_2, \dots, a_n)$ and also $\frac{f(a_1, a_2, \dots, a_n)}{g(a_1, a_2, \dots, a_n)}$, if $g(a_1, a_2, \dots, a_n) \neq 0$. Hence, $L \subset K$. Hence the result. $\square$

Definition 8.5. K/F is called a simple extension if $K = F(a)$.

Theorem 8.6. Let K/F be a simple extension with $K = F(a)$. Then, either

1. there does not exist any non-zero polynomial $g(x) \in F[x]$ with $g(a) = 0$, or
2. there exists a unique monic polynomial $f(x)$ of least degree with $f(a) = 0$.

Proof. Let $\phi : F[x] \mapsto F[a]$ be defined by $\phi(f(x)) = f(a)$, $f(x) \in F[x]$. Then ϕ is a ring homomorphism which is onto. Then we have two cases:

1. If $\ker \phi = 0$, there does not exist any non-zero polynomial $g(x) \in F[x]$ such that $g(a) = 0$.
2. If $\ker \phi \neq 0$, then since $F[x]$ is a PID, so there exists a polynomial $h(x) \in F[x]$ such that $\ker \phi = \langle h(x) \rangle$, $h(x)$ is unique if chosen to be monic. Now, since $h(x) \in \ker \phi$, so $h(a) = 0$. Let $f(x) \in F[x]$ with $f(a) = 0$. Then $f(x) \in \ker \phi$ and $f(x)$ is a multiple of $h(x)$, that is, $\deg f \geq \deg h$. Hence the second part is proved. □

We further see that $\ker \phi$ is a non-zero prime ideal, since $F[x]$ is a PID, hence $\ker \phi$ is a maximal ideal, that is, $F[x]/(\ker \phi)$ is isomorphic to $F[a]$, field. This shows that $F[a] = F(a)$.

Definition 8.7. Let K/F be an extension of $a \in K$. Then a is called *transcendental* over F if there does not exist any non-zero $f(x) \in F[x]$ with $f(a) = 0$. Otherwise, a is called *algebraic* over F and the unique monic polynomial $f(x) \in F[x]$ of least degree with $f(a) = 0$ is called the *minimum polynomial* of a . And, if every element of K is algebraic, then K/F is called an algebraic extension.

Example 8.8. 1. Let $F = \mathbb{R}$. Then $a = i = \sqrt{-1}$ is algebraic over F with minimal polynomial $x^2 + 1$.
 2. $F = \mathbb{Q}$. Then $a = e$ or π . Then a is transcendental over F .
 3. $F = \mathbb{Q}$ and $a = \sqrt{3}$. Then a is algebraic over F since the minimum polynomial is $x^2 - 3 \in \mathbb{Q}[x]$.

Theorem 8.9. Let $a \in K$ be algebraic over F and $f(x)$ be the minimum polynomial of a with degree n . Then $F(a)$ forms a vector space over F with dimension n .

Proof. We have,

$$F(a) = \frac{F[x]}{\langle f(x) \rangle} = F[a]$$

where

$$f(x) = x^n + c_1x^{n-1} + \cdots + c_n,$$

where, $c_i \in F$. We claim that $1, a, a^2, \dots, a^{n-1}$ is a basis of $F(a)$. They are linearly independent since otherwise a will satisfy a polynomial of degree less than n over F contradicting that $f(x)$ is the minimum polynomial of a . Since

$$a^n = -(c_1a^{n-1} + c_2a^{n-2} + \cdots + c_na)$$

This is an F -linear combination of $1, a, \dots, a^{n-1}$ by substituting for a^n . Similarly, any power of a can be expressed as a linear combination of $1, a, a^2, \dots, a^{n-1}$. Thus, $1, a, a^2, \dots, a^{n-1}$ generates $F[a] = F(a)$ over F . Hence, the result. □

Definition 8.10. Let K/F be an extension. It is called a *finite* extension if dimension of K over F , denoted as $\dim_F K < \infty$, or $[K : F] < \infty$.

Theorem 8.11. Let K/F be a finite extension. Then K/F is an algebraic extension.

Proof. Let $[K : F] = n$. Then, by the previous theorem, $1, a, a^2, \dots, a^n$ are linearly dependent over F . In particular, there exists $c_0, c_1, \dots, c_n$, not all zero, such that,

$$c_0 + c_1a + c_2a^2 + \dots + c_na^n = 0$$

Hence, if we define $f(x) = c_0 + c_1x + c_2x^2 + \dots + c_nx^n \equiv 0$. Hence, a is algebraic over F . Since every $a \in K$ is algebraic over F , K/F is an algebraic extension. $\square$

Theorem 8.12. Let $F \subset K \subset L$ be extensions such that K/F and L/K are finite. Then L/F is finite and

$$[L : F] = [L : K][K : F].$$

Proof. Let $\{e_1, e_2, \dots, e_n\}$ be a basis of K/F and $\{f_1, f_2, \dots, f_m\}$ be a basis of L/K . We claim that $\{e_if_j\}$, $1 \leq i \leq n$ and $1 \leq j \leq m$ is a basis of L/F . Let $a \in L$. Then,

$$a = \sum_{j=1}^m b_j f_j$$

where, $b_j \in K$. Now, for $b_j \in K$, we have

$$b_j = \sum_{i=1}^n c_{ij} e_i$$

where, $c_{ij} \in F$. Then,

$$a = \sum_{j=1}^m \sum_{i=1}^n c_{ij} e_i f_j$$

Thus showing that $\{e_if_j\}$ generates L over F . In order to prove the linear independence, let

$$\sum_{i,j} \mu_{ij} e_i f_j = 0,$$

where, $\mu_{ij} \in F$. Then,

$$\sum_j \left(\sum_i \mu_{ij} e_i \right) f_j = 0$$

Owing to the linear independence of $\{f_j\}$ over K ,

$$\sum_i \mu_{ij} e_i = 0$$

for all j . Since $\{e_i\}$ are linearly independent over F , so $\mu_{ij} = 0$ for all i, j . Hence, $\{e_if_j\}$ is a basis of L over F . Hence the result. $\square$

Corollary 8.12.1. Let K/F be a finite extension and $a \in K$ with minimum polynomial of degree n . Then n divides $[K : F]$.

Proof. Follows directly from the above theorem. $\square$

Corollary 8.12.2. Let K/F be an extension and $a, b \in K$ are algebraic over F . Then $a \pm b, ab, a/b, b \neq 0$ are all algebraic over F .

Theorem 8.13. Let K/F and L/K are algebraic extensions. Then L/F is an algebraic extension.

Proof. Since L/K is algebraic, every $a \in L$ satisfies the relation $a^n + c_1 a^{n-1} + \dots + c_n = 0$, $a_i \in K$. Then a is also algebraic over $F' = F(c_1, c_2, \dots, c_n)$ as the above relation is also a relation over the field F' . Since $c_i \in K$, they are all algebraic over F . Hence, F/F' is a finite extension. Now, $F'(a)/F'$ is a finite extension and hence $F'(a)/F$ is finite. In particular, a is algebraic over F . Hence, L/F is an algebraic extension. $\square$

8.1 Normal Extensions

We have seen that a polynomial $f(x) \in F[x]$ may not always have a root in F , but we can obtain a root of $f(x)$ in an extension field K/F . We see the following theorem:

Theorem 8.14. Let $f(x) \in F[x]$ be an irreducible polynomial. Then there exists an extension K/F which contains a root of $f(x)$.

Proof. Since $f(x)$ is irreducible, the ideal $I = \langle f(x) \rangle$ in $F[x]$ is a maximal ideal and hence $K = F[x]/I$ is a field. The map $F \rightarrow K$ given by $a \mapsto a + I$ is an isomorphism of F onto its image $F' \subset K$. Identifying F with $F' \subset K$, we can consider K as an extension of F . Also, $\alpha = x + I \in K$ is a root of $f(x)$ since $f(\alpha) = f(x + I) = f(x) + I = \bar{0}$ as $f(x) \in I$. Hence the result. $\square$

Corollary 8.14.1. Let $g(x) \in F[x]$ be any non-constant polynomial. Then there exists an extension K/F which contains a root of $g(x)$.

Proof. Consider an irreducible factor $f(x)$ of $g(x)$ and take the extension K/F which contains a root α of $f(x)$. Then α is a root of $g(x)$. $\square$

Now, if $g(x) \in F[x]$ is of degree n , then we know that $g(x)$ can have at most n roots in any extension K/F , counting a root of multiplicity k , as k roots. We now show that there exists a field K/F which contains all the n roots of $f(x)$ and in fact we can obtain a smallest extension which contains all the n roots.

Definition 8.15. Let $f(x) \in F[x]$ be of degree n . An extension K/F is called a splitting field of $f(x)$ if

1. K contains all the n roots $a_1, a_2, \dots, a_n$, and
2. $K = F(a_1, a_2, \dots, a_n)$.

Theorem 8.16. Let $f(x) \in F[x]$ be of degree n . Then $f(x)$ has a splitting field.

Proof. The proof is by induction on n . If $n = 1$, $f(x) = cx + d$, where $c \neq 0$. Then $a_1 = -d/c \in F$ is clearly a root of $f(x)$ and $K = F$. Let $n > 1$. Let $f_1(x)$ be an irreducible factor of $f(x)$ and let $L = F(a_1)$ be an extension such that $f_1(a_1) = 0$. Then $f(a_1) = 0$ and hence $f(x) = (x - a_1)g(x)$, $g(x) \in L[x]$. Since $\deg g(x) = n - 1$, by induction, there exists an extension K/L which contains $n - 1$ roots $a_2, a_3, \dots, a_n$ of $g(x)$ such that $K = L(a_2, a_3, \dots, a_n)$. Then $K = F(a_1, a_2, \dots, a_n)$ is the required splitting field of $f(x)$. $\square$

Let $f(x) \in F[x]$ and K/F be a splitting field of $f(x)$ with roots $a_1, a_2, \dots, a_n$. Then $f(x) = (x - a_1)(x - a_2) \dots (x - a_n)$ over $K[x]$.

Example 8.17. 1. The splitting field of $x^2 - 3$ is $\mathbb{Q}(\sqrt{3})$.

2. The splitting field of $x^3 - 2$ is $\mathbb{Q}(\omega, \sqrt[3]{2})$, where ω is a cube root of unity.

Definition 8.18. Let K/F and K'/F be extension fields. An isomorphism $\sigma : K \rightarrow K'$ of K onto K' is called an F -isomorphism if $\sigma|_F = Id$, where, Id is the identity mapping.

Theorem 8.19. Let $\sigma : F \rightarrow F'$ be an isomorphism of F onto F' , $f(x) \in F[x]$ an irreducible polynomial, $f(x) = \sum a_i x^i$, $\bar{f}(x) = \sum \bar{a}_i x^i$, where $\bar{a}_i = \sigma(a_i)$, the image polynomial over $F'[x]$. Let a, a' be respectively roots of $f(x)$ and $\bar{f}(x)$. Then σ can be extended to an isomorphism $\bar{\sigma} : F(a) \rightarrow F'(a')$ such that $\bar{\sigma}|_F = \sigma$.

Proof. Since σ is an isomorphism and $f(x) \in F[x]$ is irreducible, $\bar{f}(x) \in F'[x]$ is also irreducible, σ can be extended to an isomorphism $\sigma_1 : F[x] \rightarrow F'[x]$, where $\sigma_1(\sum b_i x^i) = \sum \bar{b}_i x^i$. Then $\sigma_1(f(x)) = \bar{f}(x)$ and this induces an isomorphism of the quotient rings

$$\bar{\sigma}_1 : \frac{F[x]}{\langle f(x) \rangle} \rightarrow \frac{F'[x]}{\langle \bar{f}(x) \rangle}$$

with $\bar{\sigma}_1(x + \langle f(x) \rangle) = x + \langle \bar{f}(x) \rangle$. Since for any root α of $f(x)$, $F(\alpha) \simeq \frac{F[x]}{\langle f(x) \rangle}$, we have an isomorphism $\bar{\sigma} : F(\alpha) \rightarrow F'(\alpha')$. Clearly $\bar{\sigma}(\alpha) = \alpha'$ and $\bar{\sigma}|_F = \sigma$. $\square$

Corollary 8.19.1. Let $f(x) \in F[x]$ be an irreducible polynomial and α, α' be roots of $f(x)$ in some extension. Then there exists an F -isomorphism $\bar{\sigma} : F(\alpha) \rightarrow F(\alpha')$ such that $\bar{\sigma}(\alpha) = \alpha'$.

Proof. Take $F = F'$ and $\sigma = Id$. $\square$

Theorem 8.20. Let $\sigma : F \rightarrow F'$ be an isomorphism from F onto F' , $f(x) \in F[x]$, $f(x) = \sum a_i x^i$, $\bar{f}(x) = \sigma(f(x)) = \sum \bar{a}_i x^i$, where $\bar{a}_i = \sigma(a_i)$.

Proof. The proof is by induction on $n = \deg f$. If $n = 0$, there is nothing to prove. Assume $n \geq 1$. Let $f_1(x)$ be an irreducible factor of $f(x)$, α a root of $f_1(x)$ and α' a root of $\sigma(f_1)$. By the previous theorem, σ can be extended to an isomorphism $\sigma_1 : F(\alpha) \rightarrow F'(\alpha')$ with $\sigma_1|_F = \sigma$.

Let $f(x) = (x - \alpha)g(x)$, $g(x) \in F(\alpha)[x]$, so that $\sigma(f(x)) = (x - \alpha')\sigma_1(g)$. Then K (respectively K') is a splitting field of $g(x)$ over $F(\alpha)$ (respectively $\sigma_1(g)$ over $F'(\alpha')$). By induction, σ_1 can be extended to an isomorphism $\bar{\sigma} : K \rightarrow K'$ of K onto K' . Clearly, $\bar{\sigma}|_F = \sigma_1|_F = \sigma$. $\square$

Corollary 8.20.1. Any two splitting fields of $f(x) \in F[x]$ are isomorphic.

Proof. Take $F = F'$ and $\sigma = Id$. $\square$

We have seen earlier that an extension K/F of F may contain some roots of $f(x)$ but not all. We will now deal with those extensions which have that property.

Definition 8.21. An extension K/F is called a normal extension if it is algebraic and for any irreducible polynomial $f(x) \in F[x]$ which has a root in K , $f(x)$ has all its roots in K .

Example 8.22. Let K/F be an extension of degree 2. Then it is normal. Let $\alpha \in K$ with minimum polynomial $f(x)$. Then $\deg f(x) \leq 2$ and if $\deg f(x) = 1$, $\alpha \in F$. If $\deg f(x) = 2$, then $f(x) = x^2 + ax + b$, $a, b \in F$. Then the other root is $\alpha - a$ as the sum of the roots is $-a$. Hence $\alpha - a \in K$, thus K/F is normal.

Definition 8.23. Two algebraic elements α and α' are said to be conjugate over F if there exists an F -isomorphism $\sigma : F(\alpha) \rightarrow F(\alpha')$ with $\sigma(\alpha) = \alpha'$.

Theorem 8.24. Two elements α and α' are conjugate over F iff they have the same minimum polynomial over F .

Proof. Let α and α' be conjugate over F and let $\sigma : F(\alpha) \rightarrow F(\alpha')$ with $\sigma(\alpha) = \alpha'$. If $f(x)$ is the minimum polynomial of α , then $\sigma(f(x)) = f(x)$ is the minimum polynomial of $\sigma(\alpha) = \alpha'$.

Conversely, assume that α and α' have the same minimum polynomial. Then using a previous theorem, taking $F = F'$ and $\sigma = Id$, we have an F -isomorphism $\bar{\sigma} : F(\alpha) \rightarrow F(\alpha')$ with $\bar{\sigma}(\alpha) = \alpha'$. Hence α and α' . $\square$

Exercise 8.25. 1. Find the degree of the following field extensions

i) $\mathbb{Q}(\sqrt[3]{2}, \sqrt{3})/\mathbb{Q}$

ii) $\mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{5})/\mathbb{Q}$

2. Show that $\mathbb{Q}(\sqrt{2}, \sqrt{3}) = \mathbb{Q}(\sqrt{2} + \sqrt{3})$.

3. Find the degree of the splitting field of $x^4 + 1 \in \mathbb{Q}[x]$ over $\mathbb{Q}$.

4. Find the splitting field of $x^4 - x^2 - 2$ over $\mathbb{Q}$.

5. Find all the conjugates of the following elements over $\mathbb{Q}$

i) $\sqrt{1 + \sqrt{2}}$

ii) $\sqrt{2} + i$

6. Show that if n is not an integer which is not a perfect square and $\alpha = a + b\sqrt{n}$, $a, b \in \mathbb{Q}$ is a root of a polynomial $f(x) \in \mathbb{Q}[x]$, then $a + b\sqrt{n}$ is also a root of $f(x)$.

Core Paper

MATC 2.3

Block - III

Marks : 36 (SSE : 30; IA : 06)

Operations Research - II

Syllabus

- Unit 11

Sensitivity Analysis : Changes in price vector of objective function, changes in resource requirement vector, addition of decision variable, addition of a constraint.

- Unit 12

Parametric Programming : Variation in price vector, Variation in requirement vector.

- Unit 13

Replacement and Maintenance Models : Failure mechanism of items, General replacement policies for gradual failure of items with constant money value and change of money value at a constant rate over the time period, Selection of best item

- Unit 14

Dynamic Programming (DP) : Basic features of DP problems, Bellman's principle of optimality, Multistage decision process with Forward and Backward recursive relations, DP approach to stage-coach problems.

- Unit 15

Non-Linear Programming (NLP) : Lagrange Function and Multipliers, Lagrange Multipliers methods for nonlinear programs with equality and inequality constraints.

- Unit 16

Separable programming, Piecewise linear approximation solution approach, Linear fractional programming.

Unit 11

Course Structure

- Sensitivity Analysis :
 - (i) Changes in price vector of objective function
 - (ii) Changes in resource requirement vector
 - (iii) Addition of decision variable
 - (iv) Addition of a constraint

Introduction

In an LP model, the coefficients (also known as parameters) such as: (i) profit (cost) contribution (c_j) per unit of a decision variable, x_j (ii) availability of a resources (b_i), and (iii) consumption of resource per unit of decision variables (a_{ij}), are assumed to be constant and known with certainty during a planning period. However, in real-world situations, these input parameters value may change due to dynamic nature of the business environment. Such changes in any of these parameters may raise doubt on the validity of the optimal solution of the given LP model. Thus, a decision-maker, in such situations, would like to know how changes in these parameters may affect the optimal solution and the range within which the optimal solution will remain unchanged.

Sensitivity analysis and parametric linear programming are the two techniques that are used to evaluate the effect on an optimal solution of any LP problem due to changes in its parameters.

- *Sensitivity analysis determines the sensitivity range (both lower and upper limit) within which the LP model parameters can vary (one at a time) without affecting the optimality of the current optimal solution.*
- *Parametric analysis is the study of measuring the effect on the optimal solution of the LP model due to changes at a time in more than one input parameters value outside the sensitivity range.*

However, while sensitivity analysis provides the sensitive ranges (both lower and upper limits) within which the LP model parameters can vary without changing the optimality of the current optimal solution; parametric linear programming provides information about changes in parameters value outside the sensitive range, as well as to changes in more than one parameter at a time.

Objective

After studying this chapter, you should be able to

- appreciate the significance of sensitivity analysis concept in managerial decision-making.

- perform sensitivity analysis on various parameters in an LP model without affecting the optimal solution.
- introduce a new variable and a constraint in the existing LP model with reformulation.

9 Sensitivity Analysis

The sensitivity analysis is referred to as post-optimality analysis because it does not begin until the optimal solution to the given LP model has been obtained. Different parametric changes in an LP problem discussed in this unit are:

- (i) Profit (or cost) per unit (c_j) associated with both basic and non-basic decision variables (i.e., coefficients in the objective function).
- (ii) Availability of resources (i.e., right-hand side constants, b_i in constraints).
- (iii) Consumption of resources per unit of decision variables x_j (i.e., coefficients of decision variables in the constraints, a_{ij}).
- (iv) Addition of a new variable to the existing list of variables in LP problem.
- (v) Addition of a new constraint to the existing list of constraints in the LP problem.

9.1 Change in Objective Function Coefficient (c_j)

Given an optimal basic feasible solution, suppose that the coefficient c_k of a variable x_k in the objective function is changed from c_k to $c_k + \Delta c_k$, where Δc_k represents the positive (or negative) amount of change in the value of c_k . In optimal simplex table, the feasibility of the solution remains unaffected due to changes in the coefficients, c_j of basic variables in the objective function. However, any change in these coefficients (c'_j s) only affect the optimality of the solution. Thus, such a change requires recomputing z_j values in $c_j - z_j$ row of the optimal simplex table because $c_k - z_k (c_k - c_B B^{-1} a_k)$ involves coefficients c_k . Obviously the effect of this change will be seen in the $c_j - z_j$ row of the optimal simplex table.

While recomputing $c_j - z_j$ values, the following two cases will arise

- (i) The new $c_j - z_j$ values satisfy optimality condition and the current optimal feasible solution remains unchanged. However, optimal value of objective function may change.
- (ii) The optimality condition is not satisfied. In such a case usual simplex method is used to obtain new optimal solution.

Case I: Change in the coefficient of a non-basic variable

If $c_j - z_j \leq 0$ for all non-basic variables in a maximization LP problem, then the current optimal feasible solution remains unchanged. Let c_k be the coefficient of a non-basic variable x_k in the objective function. Since c_k is the coefficient of non-basic variable x_k , therefore, it does not effect any of the c_j values listed in the ' c_B ' column of optimal simplex table associated with basic variables. Since the calculation of $z_j = c_B B^{-1} a_j$ values do not involve c_j , therefore changes in c_j do not change z_j values and hence $(c_j - z_j)$ values remain unchanged, except the $c_k - z_k$ value due to change in c_k .

If change in coefficient, c_k does not affect the feasibility of the optimal solution, then unit profit

of x_k can be lowered to any level without causing the optimal solution to change. But any increase in c_k value, beyond a certain level, (i.e. upper limit) should make this variable eligible to be a basic variable in the new solution mix. Obviously, $c_k - z_k$ will no longer then be negative.

To retain optimality of the current optimal solution for a change Δc_k in c_k , we must have $(c_k + \Delta c_k) - z_k \leq 0$ or $c_k + \Delta c_k \leq z_k$. Hence, for an LP problem with an objective function of the maximization type, the value of c_k may be increased up to the value of z_k , and decreased to negative infinity $(-\infty)$ without affecting the optimal solution.

Case II: Change in the coefficient of a basic variable

If $c_j - z_j \leq 0$ for all non-basic variables in a maximization LP problem, then the current optimal feasible solution remains unchanged. Let c_k be the coefficient of a non-basic variable x_k in the objective function. Since c_k is the coefficient of non-basic variable x_k , therefore, it does not effect any of the c_j values listed in the ' c'_B ' column of optimal simplex table associated with basic variables. Since the calculation of $z_j = c_B B^{-1} a_j$ values do not involve c_j , therefore changes in c_j do not change z_j values and hence $(c_j - z_j)$ values remain unchanged, except the $c_k - z_k$ value due to change in c_k .

If change in coefficient, c_k does not affect the feasibility of the optimal solution, then unit profit of x_k can be lowered to any level without causing the optimal solution to change. But any increase in c_k value, beyond a certain level, (i.e. upper limit) should make this variable eligible to be a basic variable in the new solution mix. Obviously, $c_k - z_k$ will no longer then be negative.

To retain optimality of the current optimal solution for a change Δc_k in c_k , we must have $(c_k + \Delta c_k) - z_k \leq 0$ or $c_k + \Delta c_k \leq z_k$. Hence, for an LP problem with an objective function of the maximization type, the value of c_k may be increased up to the value of z_k , and decreased to negative infinity $(-\infty)$ without affecting the optimal solution.

Example-1: A company wants to produce three products: A, B and C. The per unit profit on these products is Rs 4, Rs 6 and Rs 2, respectively. These products require two types of resources, manpower and raw material. The LP model formulated for determining the optimal product mix is as follows:

$$\begin{aligned} &\text{Maximize } Z = 4x_1 + 6x_2 + 2x_3 \\ &\text{subject to the constraints} \\ &(i) x_1 + x_2 + x_3 \leq 3 \text{ (Manpower required),} \\ &(ii) x_1 + 4x_2 + 7x_3 \leq 9 \text{ (Raw material available)} \end{aligned}$$

and $x_1, x_2, x_3 \geq 0$ where x_1, x_2 and x_3 = number of units of products A, B and C, respectively, to be produced.

(a) Find the optimal product mix and the corresponding profit of the company.

(b) Find the range of the profit contribution of product C (i.e. coefficient c_3 of variable x_3) in the

objective function, such that the current optimal product mix remains unchanged.

(c) Find the range of the profit contribution of product A (i.e. coefficient c_1 of variable x_1) in the objective function such that the current optimal product mix remains unchanged.

Solution: (a) Introducing the slack variables s_1 and s_2 in the constraints to convert LP problem into its standard form as follows:

$$\text{Maximize } Z = 4x_1 + 6x_2 + 2x_3 + 0s_1 + 0s_2$$

subject to the constraints

$$(i) x_1 + x_2 + x_3 + s_1 = 3, (ii) x_1 + 4x_2 + 7x_3 + s_2 = 9, \text{ and } x_1, x_2, x_3, s_1, s_2 \geq 0$$

The optimal solution obtained by applying the simplex method is shown in the following table.

$c_j \rightarrow$			4	6	2	0	0
Basic Variables Coefficient c_B	Basic Variables B	Basic Variables Value $b (= x_B)$	x_1	x_2	x_3	s_1	s_2
4	x_1	1	1	0	-1	4/3	-1/3
6	x_2	2	0	1	2	-1/3	1/3
$Z = 16$			4	6	8	10/3	2/3
			0	0	-6	-10/3	-2/3

Table - 1

The optimal solution is: $x_1 = 1$, $x_2 = 2$ and Max $Z = \text{Rs } 16$.

(b) Effect of change in the coefficient c_3 of non-basic variable x_3 for product C: In the variable x_3 is non-basic and its coefficient $c_3 = 2$ is not listed in the c_B column of the table. This means a further decrease in its profit contribution $c_3 (= \text{Rs } 2)$ per unit will have no effect on the current optimal product mix. But, if c_3 is increased beyond a certain value, the product may become profitable to be produced. Hence, there is only an upper limit on c_3 for which the current optimal product mix will be effected.

$c_j \rightarrow$			4	6	$2 + \Delta c_3$	0	0
Basic Variables Coefficient c_B	Basic Variables B	Basic Variables Value $b (= x_B)$	x_1	x_2	x_3	s_1	s_2
4	x_1	1	1	0	-1	4/3	-1/3
6	x_2	2	0	1	2	-1/3	1/3
$Z = 16$			4	6	8	10/3	2/3
			0	0	$\Delta c_3 - 6$	-10/3	-2/3

Table - 2

Now a change Δc_3 in c_3 will cause a change in z_3 and $c_3 - z_3$ value in the x_3 -column. The change in c_3 results in the modified simplex Table 2. For an optimal solution shown in Table 2 to remain unchanged, we must have: $\Delta c_3 - 6 \leq 0$ or $\Delta c_3 \leq 6$. Recalling that $c_3 = 2 + \Delta c_3$, or $\Delta c_3 = c_3 - 2$, after substituting this amount in the above inequality, we get $c_3 - 2 \leq 6$ or $c_3 \leq 8$. This implies that as long as the profit contribution per unit of product C is less than Rs 8 (i.e. change should not be more

than Rs 8) it is not profitable to produce it and therefore the current optimal solution will remain unchanged.

(b) Effect of change in the coefficient c_1 of basic variable x_1 for product A: In the optimal simplex Table 1, variable x_1 is the basic variable. This means that any decrease in its coefficient (profit contribution per unit) c_1 (= Rs 4) will make product A less profitable to produce and therefore the current optimal product mix will be affected. Also, an increase in the value of c_1 , beyond a certain limit, will make the product A much more profitable and may force the decision-maker to decide to only produce product A. Thus, in either case, the current optimal product mix will be affected and hence we need to know both the lower and upper limit on the value of c_1 , within which the current optimal solution will remain unchanged. For this reproduce Table 1 once again with unknown value of c_1 as shown in Table 3.

$c_j \rightarrow$			$4 + \Delta c_1$	6	2	0	0
Basic Variables Coefficient c_B	Basic Variables B	Basic Variables Value $b (= x_B)$	x_1	x_2	x_3	s_1	s_2
$4 + \Delta c_1$	x_1	1	1	0	-1	4/3	-1/3
6	x_2	2	0	1	2	-1/3	1/3
$Z = 12 + (4 + \Delta c_1)$			$4 + \Delta c_1$	6	$8 - \Delta c_1$	$10/3 + 4\Delta c_1/3$	$2/3 - \Delta c_1/3$
$c_j - z_j$			0	0	$\Delta c_1 - 6$	$-10/3 - 4\Delta c_1/3$	$\Delta c_1/3 - 2/3$

Table - 3

For the solution shown in Table 3 to remain optimal we must have all $c_j - z_j \leq 0$. That is

$$\begin{aligned}\Delta c_1 \leq 0 \text{ gives } \Delta c_1 &\leq 6 \\ -\frac{10}{3} - \frac{4\Delta c_1}{3} &\leq 0 \text{ gives } \Delta c_1 \geq -5/2 \\ \frac{\Delta c_1}{3} - \frac{2}{3} &\leq 0 \text{ gives } \Delta c_1 \leq 2\end{aligned}$$

Thus, the range of values within which c_1 may change without affecting the current optimal solution is:

$$-\frac{5}{2} \leq \Delta c_1 \leq 2 \text{ or } 4 - \frac{5}{2} \leq c_1 \leq 4 + 2, \text{ i.e. } \frac{3}{2} \leq c_1 \leq 6$$

From these calculations, it may be concluded that in the case of competitive pressure or in the absence of any competition, the decision-maker knows to what extent the prices can be adjusted without changing the optimal product mix.

9.2 Change in the availability of Resources (b_i)

Any change in the right-hand side of the constraints does not affect the optimality condition. Since in the determination of solution values, $(x_B = B^{-1}b)$ value of resource (i.e. b) is involved, therefore any change in b_i values affect the value of basic variables and value of objective function $Z (= c_B x_B)$. Thus, if resource b_k is changed to $b_k + \Delta b_k$, then the new values of resources becomes

$$\begin{aligned}b^* &= (b_1, b_2, \dots, b_k + \Delta b_k, \dots, b_m) \\ &= (b_1, b_2, \dots, b_m) + (0, 0, \dots, \Delta b_k, \dots, 0) = b + \Delta b_k\end{aligned}$$

The range of values within which Δb_k can vary without affecting the optimality of the current solution is determined as follows:

$$\begin{aligned}
\mathbf{x}_B^* &= \mathbf{B}^{-1}(\mathbf{b} + \Delta \mathbf{b}) = \mathbf{B}^{-1}\mathbf{b} + \mathbf{B}^{-1}(\Delta \mathbf{b}) \\
&= \mathbf{B}^{-1}\mathbf{b} + \mathbf{B}^{-1}(0, 0, \dots, \Delta b_k, \dots, 0) \\
&= \mathbf{x}_B + (\beta_1, \beta_2, \dots, \beta_k, \dots, \beta_m)(0, 0, \dots, \Delta b_k, \dots, 0) = \mathbf{x}_B + \beta_k(\Delta b_k) \\
x_{Bi}^* &= x_{Bi} + \beta_{ik}(\Delta b_k); \quad \text{for } i\text{-th basic variable,}
\end{aligned}$$

where β_{ik} is the (i, k) element of $\mathbf{B}^{-1}$ and is the k -th column vector of $\mathbf{B}^{-1}$. In order to maintain the feasibility of the solution at each iteration, the solution values, $\mathbf{x}_B$ must be non-negative. That is, we must have:

$$x_{Bi}^* = x_{Bi} + \beta_{ik}(\Delta b_k) \geq 0; i = 1, 2, \dots, m$$

Hence, the range of variation in b_k can be obtained solving the following system of inequalities

$$\text{Min} \left\{ \frac{-x_{Bi}}{\beta_{ik} < 0} \right\} \geq \Delta b_k \geq \text{Max} \left\{ \frac{-x_{Bi}}{\beta_{ik} > 0} \right\}$$

Example 2

Solve the following LP problem

$$\begin{aligned}
&\text{Maximize } Z = 4x_1 + 6x_2 + 2x_3 \\
&\text{subject to the constraints} \\
&(i) x_1 + x_2 + x_3 \leq 3, \\
&(ii) x_1 + 4x_2 + 7x_3 \leq 9
\end{aligned}$$

and $x_1, x_2, x_3 \geq 0$. Discuss the effect of discrete change in the availability of resources from $[3, 9]^T$ to $[9, 6]^T$.

Solution: The given LP problem in its standard form can be stated as follows:

$$\begin{aligned}
&\text{Maximize } Z = 4x_1 + 6x_2 + 2x_3 + 0.s_1 + 0.s_2 \\
&\text{subject to the constraints} \\
&(i) x_1 + x_2 + x_3 + s_1 = 3, \\
&(ii) x_1 + 4x_2 + 7x_3 + s_2 = 9
\end{aligned}$$

and $x_1, x_2, x_3, s_1, s_2 \geq 0$. Applying the simplex method, the optimal solution: $x_1 = 1, x_2 = 2$ and $\text{Max } Z = 16$. If new values of the right-hand side constants in the constraints are $[9, 6]^T$, then the new values of the basic variables ($x_B = \mathbf{B}^{-1}\mathbf{b}$) shown in Table 4 will become:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 4/3 & -1/3 \\ -1/3 & 1/3 \end{bmatrix} \begin{bmatrix} 9 \\ 6 \end{bmatrix} = \begin{bmatrix} 10 \\ -1 \end{bmatrix} \quad \text{or } x_1 = 10, \text{ and } x_2 = -1$$

$c_j \rightarrow$			4	6	2	0	0
Basic Variables Coefficient c_B	Basic Variables B	Basic Variables Value $b (= x_B)$	x_1	x_2	x_3	s_1	s_2
4	x_1	1	1	0	-1	4/3	-1/3
6	x_2	2	0	1	2	-1/3	1/3
$Z = 16$		z_j	4	6	8	10/3	2/3
		$c_j - z_j$	0	0	-6	-10/3	-2/3

Table - 4

Since the value of x_2 is negative, the optimal solution shown in Table 4 is infeasible. Apply dual simplex to remove this infeasibility. For this, reproducing Table 5 with new values of x_1 and x_2 as shown in Table 5.

$c_j \rightarrow$			4	6	2	0	0
Basic Variables Coefficient c_B	Basic Variables B	Basic Variables Value $b (= x_B)$	x_1	x_2	x_3	s_1	s_2
4	x_1	10	1	0	-1	4/3	-1/3
6	x_2	-1	0	1	2	-1/3	1/3
$Z = 34$		z_j	4	6	8	10/3	2/3
		$c_j - z_j$	0	0	-6	-10/3	-2/3

Table - 5

Since $x_2 = -1$, therefore x_2 -row becomes the key row and so x_2 is the outgoing variable. Calculating the minimum ratio to identify key column as follows.

$$\text{Min} \left\{ \frac{c_j - z_j}{y_{rj}}, y_{rj} < 0 \right\} = \left\{ \frac{-10/3}{-1/3} \right\} = 10 \quad (\text{column } s_1)$$

Hence column ' s_1 ' is the key column and variable s_1 will enter into the basis. Revise the simplex table using the following row operations:

$$R_2 \text{ (new)} \rightarrow R_2 \text{ (old)} \times -3 \quad R_1 \text{ (new)} \rightarrow R_1 \text{ (old)} - (4/3)R_2 \text{ (new)}$$

The new solution is shown in Table 6.

$c_j \rightarrow$			4	6	2	0	0
Basic Variables Coefficient c_B	Basic Variables B	Basic Variables Value $b (= x_B)$	x_1	x_2	x_3	s_1	s_2
4	x_1	6	1	4	7	0	1
0	s_1	3	0	-3	-6	1	-1
$Z = 24$		z_j	4	16	28	0	4
		$c_j - z_j$	0	-10	-26	0	-4

Table - 6

Since all $c_j - z_j \leq 0$ and all $b_i > 0$, the solution: $x_1 = 6, x_2 = 0, x_3 = 0$ and Max $Z = 24$ given in Table 6 is optimal.

9.3 Addition of decision variable and constraint

After solving an LP model, the decision-maker may recall that a particular resource constraint was overlooked during the model formulation or perhaps he may desire to know the effect of adding a new resource to enhance the objective function value. The addition of a constraint in the existing constraints will cause a simultaneous change in the objective function coefficients (c_j), as well as coefficients a_{ij} of a corresponding non-basic variable. Consequently, it will affect the optimality of the LP problem. This means that the new variable should enter into the basis only if it improves the value of the objective function.

Suppose that a new constraint

$$\mathbf{a}_{m+1,1}x_1 + \mathbf{a}_{m+1,2}x_2 + \dots + \mathbf{a}_{m+1,n}x_n \leq \mathbf{b}_{m+1}$$

is added to the system of original constraints $Ax = b, x \geq 0$, where b_{m+1} is positive, zero or negative. Then the following two cases may arise.

1. The optimal solution ($\mathbf{x}_B$) of the original problem satisfies the new constraint. If this is the case, then the solution remains feasible as well as optimal. This is because the new constraint either reduces or leaves unchanged the feasible region of the given LP problem.
2. The optimal solution ($\mathbf{x}_B$) of the original problem does not satisfy the new constraint. In this case the optimal solution to the modified LP problem should be re-obtained. We will illustrate this case by the following example.

Example 3: Consider the following LP problem

$$\text{Maximize } Z = 3x_1 + 5x_2$$

subject to the constraints

$$(i)3x_1 + 2x_2 \leq 18, (ii)x_1 + 2x_2 \leq 4, (iii)x_2 \leq 6 \text{ and } x_1, x_2 \geq 0$$

- (a) Suppose variable x_6 is added to the given LP problem. Then obtain an optimal solution to the resulting LP problem. It is given that the coefficients of x_6 in the constraint of the problem are 1, 1 and 1, and that its coefficient in the objective function is 2.
- (b) Discuss the effect on the optimal basic feasible solution by adding a new constraint $2x_1 + x_2 \leq 8$ to the given set of constraints.

Solution: The optimal solution: $x_1 = 2, x_2 = 6$ with Max $Z = 36$ to the given LP problem is shown in Table 7.

$c_j \rightarrow$			3	5	0	0	0
Basic Variables Coefficient c_B	Basic Variables B	Basic Variables Value $b (= x_B)$	x_1	x_2	s_1	s_2	s_3
3	x_1	2	1	0	1/3	0	-2/3
0	s_2	0	0	0	-2/3	1	4/3
5	x_2	6	0	1	0	0	1
$Z = 36$	$c_j - z_j$		0	0	-1	0	-3

Table - 7

(a) After adding the new variable, say x_6 in the given problem, the new LP problem becomes:

$$\text{Maximize } Z = 3x_1 + 5x_2 + 2x_6$$

subject to the constraints

$$(i) 3x_1 + 2x_2 + x_6 \leq 18, (ii) x_1 + 2x_2 + x_6 \leq 4, (iii) x_2 + x_6 \leq 6 \text{ and } x_1, x_2, x_6 \geq 0$$

Given that the column vector associated with variable, x_6 is $a_6 = (1, 1, 1)$, with the help of Table 7, we compute

$$y_6 = B^{-1} a_6 = \begin{bmatrix} 1/3 & 0 & -2/3 \\ -2/3 & 1 & 4/3 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1/3 \\ 5/3 \\ 1 \end{bmatrix}$$

Since in the current optimal solution, $c_B = (3, 0, 5)$, we have

$$c_6 - z_6 = c_6 - c_B y_6 = 2 - (3, 0, 5) \begin{bmatrix} -1/3 \\ 5/3 \\ 1 \end{bmatrix} = -2 (\leq 0)$$

As $c_6 - z_6 \leq 0$, the optimality of the current solution remains unaffected with the addition of x_6 .

(b) The optimal basic feasible solution given in Table 7 does not satisfy the additional constraint $2x_1 + x_2 \leq 8$. Thus, a new optimal solution is obtained by adding slack variable s_4 to this constraint. Then, it is written together with the entries in Table 7, as shown in Table 8.

$c_j \rightarrow$			3	5	0	0	0	0
Basic Variables Coefficient c_B	Basic Variables B	Basic Variables Value $b (= x_B)$	x_1	x_2	s_1	s_2	s_3	s_4
3	x_1	2	1	0	1/3	0	-2/3	0
0	s_2	0	0	0	-2/3	1	4/3	0
5	x_2	6	0	1	0	0	1	0
0	s_4	8	2	1	0	0	0	1
$Z = 36$			0	0	-1	0	-3	0
$c_j - z_j$								

Table - 8

In Table 8 it may be noted that the basis matrix B has been disturbed due to row 4. Thus, the coefficients in row 4 must become zero. This can be done by using the following row operations.

$$R_4 (\text{new}) \rightarrow R_4 (\text{old}) - 2R_1 - R_3$$

We get a new table as shown in Table 9.

$c_j \rightarrow$			3	5	0	0	0	0
Basic Variables Coefficient c_B	Basic Variables B	Basic Variables Value $b (= x_B)$	x_1	x_2	s_1	s_2	s_3	s_4
3	x_1	2	1	0	1/3	0	-2/3	0
0	s_2	0	0	0	-2/3	1	4/3	0
5	x_2	6	0	1	0	0	1	0
0	s_4	-2	0	0	-2/3	0	1/3	1 $\rightarrow$
$Z = 36$			0	0	-1	0	-3	0
$c_j - z_j$					$\uparrow$			

Table - 9

Since the solution given in Table 9 is optimal but not feasible, apply the dual simplex method to get an optimal basic feasible solution. Introduce s_1 into the basis and remove s_4 from the basis. The new solution is shown in Table 10.

$c_j \rightarrow$			3	5	0	0	0	0
Basic Variables Coefficient c_B	Basic Variables B	Basic Variables Value $b (= x_B)$	x_1	x_2	s_1	s_2	s_3	s_4
3	x_1	1	1	0	0	0	-1/2	1/2
0	s_2	2	0	0	0	1	1	-1
5	x_2	6	0	1	0	0	1	0
0	s_1	3	0	0	1	0	-1/2	-3/2
$Z = 33$		$c_j - z_j$	0	0	0	0	-7/2	-3/2

Table - 10

Since all $c_j - z_j \leq 0$, the solution shown in Table 10 is the optimal basic feasible solution: $x_1 = 1, x_2 = 6$ and Max $Z = 33$. It may be noted here that the additional constraint has decreased the optimal value of the objective function from 36 to 33.

Summary

Sensitivity analysis provides the sensitive ranges (both lower and upper limits) within which LP model parameters can vary without changing the optimality of the current solution. Formulae to conduct sensitivity analysis followed by numerical examples have been solved to illustrate how changes in the parameters specially, (i) the change in the objective function coefficients; right hand-side value of a constraint and (ii) addition of decision variable and constraint, affect its optimal solution.

Unit 12

Course Structure

- Parametric Programming : Variation in objective function coefficients, Variation in the availability of resources

Introduction

Once an LP model based on real-life data has been solved, the decision-maker desires to know how the solution will change if parameters, such as cost (or profit) c_j , availability of resources b_i and the technological coefficients a_{ij} are changed. We have already discussed the need to perform a sensitivity analysis in order to consider the impact of *discrete changes* in its parameters on optimal solution of LP model. In this unit, we will discuss another parameter variation analysis also called *parametric analysis* to find various basic feasible solutions of an LP model that become optimal one after the other, due to continuous variations in the parameters. Since LP model parameters change as a linear function of a single parameter, this technique is known as *linear parametric programming*.

The purpose of this analysis is to keep to a minimum the additional efforts required to take care of changes in the optimal solution due to variation in LP model parameters *over a range of variation*. In this chapter we will perform parametric analysis only for the following two parameters (evaluation of other parameters, over a range, is also possible but tend to be more complex.)

- (i) Variation in objective function coefficients, c_j
- (ii) Variation in resources availability (Right-hand side values), b_i

Let λ be the unknown (positive or negative) scalar parameter with which coefficients in the LP model vary. We start the analysis at optimal solution obtained at $\lambda = 0$. Then, using the optimality and feasibility conditions of the simplex method we determine the range of λ for which the optimal solution at $\lambda = 0$ remains unchanged. Let λ lies between 0 and λ_1 . This means $0 \leq \lambda \leq \lambda_1$ is the range of λ beyond which the current solution will become infeasible and/or non-optimal. Thus at $\lambda = \lambda_1$ a new solution is determined which remains optimal and feasible in other interval, say $\lambda_1 \leq \lambda \leq \lambda_2$. Again a new solution at $\lambda = \lambda_2$ is obtained. The process of determining the range of λ is repeated till a stage is reached beyond which the solution either does not change or exist.

Objective

After studying this chapter you should be able to

- appreciate the need of parametric analysis to find various basic feasible solutions of any LP problem, which become optimal one after the other due to continuous variations in the parameters

of LP problem.

- perform parametric analysis to study variation in the objective function coefficients and resources availability.
- take care of change in the optimal solution due to variation in LP model parameters over a range of variation.

10 Parametric Programming

10.1 Variation in the objective function coefficients

Let us define the parametric linear programming model as follows:

$$\begin{aligned} \text{Maximize } Z &= \sum_{j=1}^n (c_j + \lambda c'_j) x_j \\ \text{subject to the constraints} \\ \sum_{j=1}^n a_{ij} x_j &= b_i; \quad i = 1, 2, \dots, m \quad \text{and} \quad x_j \geq 0; \quad j = 1, 2, \dots, n \end{aligned}$$

where $\lambda c'_j$ represents predetermined variation in the parameter c and $\lambda \geq 0$ is a scalar parameter. Now the aim is to determine such consecutive values of λ at which the current optimal basic feasible solution tends to change with a change in the coefficients c_j . Such consecutive values of λ are called critical (range) values of λ and are measured from $\lambda = 0$. Thus, the given LP problem is initially solved by using simplex method at $\lambda = 0$. Since changes in cost coefficient c_j only affect the optimality of the current solution, therefore as λ changes only $c_j - z_j$ values are affected. Hence, for the perturbed LP problem let us calculate $c_j - z_j$ values corresponding to non-basic variable columns in the optimal simplex table as follows:

$$\begin{aligned} c_j(\lambda) - z_j(\lambda) &= c_j(\lambda) - \mathbf{c}_B(\lambda) \mathbf{B}^{-1} \mathbf{a}_j = (c_j + \lambda c'_j) - (\mathbf{c}_B + \lambda \mathbf{C}'_B) \mathbf{y}_j; \quad \mathbf{y}_j = \mathbf{B}^{-1} \mathbf{a}_j \\ &= (c_j - \mathbf{c}_B \mathbf{y}_j) + \lambda (c'_j - \mathbf{C}'_B \mathbf{y}_j) = (c_j - z_j) + \lambda (c'_j - z'_j); \quad z_j = \mathbf{c}_B \mathbf{y}_j \end{aligned}$$

For a solution to be optimal for all values of λ we must have $c_j(\lambda) - z_j(\lambda) \leq 0$ (maximization case) and $c_j(\lambda) - z_j(\lambda) \geq 0$ (minimization case). These inequalities, for a given solution, are also used for determining the range $\lambda_1 \leq \lambda \leq \lambda_2$, within which the current solution remains optimal as follows:

$$\lambda = \text{Min} \left\{ \frac{-(c_j - z_j)}{(c'_j - z'_j)} \right\}$$

where $c'_j - z'_j > 0$ for maximization and $c'_j - z'_j < 0$ for minimization.

Example: Consider the parametric linear programming problem:

$$\begin{aligned} \text{Maximize } Z &= (3-6\lambda)x_1 + (2-2\lambda)x_2 + (5+5\lambda)x_3 \\ \text{subject to the constraints} \\ (i) x_1 + 2x_2 + x_3 &\leq 430, \quad (ii) 3x_1 + 2x_3 \leq 460, \quad (iii) 3x_1 + 4x_2 \leq 420 \\ \text{and } x_1, x_2, x_3 &\geq 0 \end{aligned}$$

Perform the parametric analysis and identify all the critical values of the parameter λ .

Solution: The given parametric LP problem can be written in its standard form as:

$$\text{Maximize } Z = (3-6\lambda)x_1 + (2-2\lambda)x_2 + (5+5\lambda)x_3 + 0s_1 + 0s_2 + 0s_3$$

subject to the constraints

$$(i)x_1 + 2x_2 + x_3 + s_1 = 430, \quad (ii)3x_1 + 2x_3 + s_2 = 460, \quad (iii)3x_1 + 4x_2 + s_3 = 420$$

$$\text{and } x_1, x_2, x_3, s_1, s_2, s_3 \geq 0$$

According to the problem, we have:

$$c(\lambda) = c_j + \lambda c'_j = (3, 2, 5, 0, 0, 0) + \lambda(-6, -2, 5, 0, 0, 0)$$

Solving the given LP problem with $\lambda = 0$. The optimal solution at $\lambda = 0$ is shown in Table 2.1.

$c_j \rightarrow$			3	2	5	0	0	0
c_B	Basic Variables B	Solution Values $b (= x_B)$	x_1	x_2	x_3	s_1	s_2	s_3
2	x_2	100	-1/4	1	0	1/2	-1/4	0
5	x_3	230	3/2	0	1	0	1/2	0
0	s_3	20	2	0	0	-2	1	1
$Z = 1,350$		z_j	7	2	5	1	2	0
		$c_j - z_j$	-4	0	0	-1	-2	0

Table - 2.1

The optimal solution is: $x_1 = 0, x_2 = 100, x_3 = 230$ and $\text{Max } Z = 1,350$. In order to find the first critical (or range) value of λ in which the solution shown in Table 2.1 remains optimal, we first find $c'_j - z'_j$ values corresponding to non-basic variables x_1, s_1 and s_2 columns as follows:

$$\begin{aligned} c'_j - z'_j &= c'_j - c'_B y_j = (-6, 0, 0) - (-2, 5, 0) \begin{bmatrix} -1/4 & 1/2 & -1/4 \\ 3/2 & 0 & 1/2 \\ 2 & -2 & 1 \end{bmatrix}; \quad j = 1, 4, 5 \\ &= (-6, 0, 0) - \left[\frac{1}{2} + \frac{15}{2}, -1, \frac{1}{2} + \frac{5}{2} \right] = (-6, 0, 0) - (8, -1, 3) = (-14, 1, -3) \end{aligned}$$

For a maximization LP problem, the current solution will remain optimal provided all $c_j(\lambda) - z_j(\lambda) \leq 0$. Since $c'_4 - z'_4 > 0$, the first critical value of λ is given by:

$$\lambda_1 = \text{Min} \left\{ \frac{-(c_j - z_j)}{(c'_j - z'_j) > 0} \right\} = -\frac{(c_4 - z_4)}{c'_4 - z'_4} = -\frac{(-1)}{1} = 1$$

This means that for $\lambda_1 \in [0, 1]$, the solution given in Table 2.1 remains optimal. The objective function value in this interval is given by:

$$Z(\lambda) = Z + Z'(\lambda) = c_B x_B + \lambda c'_B x_B = 1,350 + 950\lambda$$

Now, for values of λ other than zero in the interval $[0, 1]$, we compute $c_j(\lambda) - z_j(\lambda)$ values for non-basic variables x_1, s_1 and s_2 as shown in Table 2.2.

$$\begin{aligned} c_1(\lambda) - z_1(\lambda) &= (c_1 - z_1) + \lambda(c'_1 - z'_1) = -4 - 14\lambda \leq 0 \quad \text{or } \lambda \geq 2/7 \\ c_4(\lambda) - z_4(\lambda) &= (c_4 - z_4) + \lambda(c'_4 - z'_4) = -1 - \lambda \leq 0 \quad \text{or } \lambda \geq 1 \\ c_5(\lambda) - z_5(\lambda) &= (c_5 - z_5) + \lambda(c'_5 - z'_5) = -2 - 3\lambda \leq 0 \quad \text{or } \lambda \geq -2/3 \end{aligned}$$

The optimal solution for any value of λ in the interval $[0, 1]$ is given in Table 2.2.

				$c'_j \rightarrow$	-6	-2	5	0	0	0
				$c_j \rightarrow$	3	2	5	0	0	0
c'_B	c_B	Basic Variables B	Solution Values $b (= x_B)$		x_1	x_2	x_3	s_1	s_2	s_3
2	2	x_2	100		-1/4	1	0	1/2	-1/4	0
5	5	x_3	230		3/2	0	1	0	1/2	0
0	0	s_3	20		2	0	0	-2	1	1
$Z(\lambda) = 1,350 + 950\lambda$				$c_j - z_j$	-4	0	0	-1	-2	0
				$c'_j - z'_j$	-14	0	0	1	-3	0
				$c_j(\lambda) - z_j(\lambda)$	-4 - 14λ	0	0	-1 + λ	-2 - 3λ	0

Table - 2.2

At $\lambda = 1$, $c_4(\lambda) - z_4(\lambda) = 0$ in the ' s_1 ' column. But for $\lambda > 1$, $c_4(\lambda) - z_4(\lambda) > 0$ for non-basic variable s_1 and hence the solution in Table 2.2 no longer remains optimal. We now enter variable s_1 in the solution to find new optimal solution. The new optimal solution shown in Table 2.3 is: $x_1 = 0$, $x_2 = 0$, $x_3 = 230$ and Max $Z = 2,300$.

				$c'_j \rightarrow$	-6	-2	5	0	0	0
				$c_j \rightarrow$	3	2	5	0	0	0
c'_B	c_B	Basic Variables B	Solution Values $b (= x_B)$		x_1	x_2	x_3	s_1	s_2	s_3
0	0	s_1	200		-1/2	2	0	1	-1/2	0
5	5	x_3	230		3/2	0	1	0	1/2	0
0	0	s_3	420		1	4	0	0	0	1
$Z(\lambda) = 2,300$				$c_j - z_j$	-9/2	2	0	0	-5/2	0
				$c'_j - z'_j$	-27/2	-2	0	0	-5/2	0
				$c_j(\lambda) - z_j(\lambda)$	-18	0	0	0	-5	0

Table - 2.3

The solution shown in Table 2.3 will be optimal if all $c_j(\lambda) - z_j(\lambda) \leq 0$, $j = 1, 2, 5$. To check the optimality we compute these values for the non-basic variables x_1 , x_2 and s_2 as follows:

$$c_1(\lambda) - z_1(\lambda) = (c_1 - z_1) + \lambda(c'_1 - z'_1) = -\frac{9}{2} - \frac{27}{2}\lambda \leq 0 \text{ or } \lambda \geq -\frac{1}{3}$$

$$c_2(\lambda) - z_2(\lambda) = (c_2 - z_2) + \lambda(c'_2 - z'_2) = -2 - 2\lambda \leq 0 \text{ or } \lambda \geq 1$$

$$c_5(\lambda) - z_5(\lambda) = (c_5 - z_5) + \lambda(c'_5 - z'_5) = -\frac{5}{2} - \frac{5}{2}\lambda \leq 0 \text{ or } \lambda \geq -1$$

Therefore, for $\lambda = 1$, the $c_j(\lambda) - z_j(\lambda) \leq 0$ for all non-basic variable columns and hence the solution in Table 2.3 is optimal: $x_1 = x_2 = 0$, $x_3 = 230$ and Max $Z = 2,300$.

For $\lambda \leq -2/3$, $c_j(\lambda) - z_j(\lambda)$ value for non-basic variable s_2 becomes positive and again solution shown in Table 2.3 no longer remains optimal. Entering variable s_2 in the basis to find new optimal solution. The variable s_2 will replace basic variable s_3 in the basis. The new optimal solution is shown in Table 2.4.

				$c'_j \rightarrow$	-6	-2	5	0	0	0
				$c_j \rightarrow$	6	2	5	0	0	0
c'_B	c_B	Basic Variables B	Solution Values $b (= x_B)$		x_1	x_2	x_3	s_1	s_2	s_3
-2	2	x_2	105		1/4	1	0	0	0	1/4
5	5	x_3	220		1/2	0	1	1	0	-1/2
0	0	s_2	20		2	0	0	-2	1	1
$Z = 1310 + 890\lambda$				$c_j - z_j$	0	0	0	-5	0	2
				$c'_j - z'_j$	-8	0	0	-5	0	3

Table - 2.4

Solution shown in Table 2.4 will be optimal only when $c_j(\lambda) - z_j(\lambda) \leq 0$; $j = 1, 4, 6$

$$\begin{aligned} c_1(\lambda) - z_1(\lambda) &= (c_1 - z_1) + \lambda(c'_1 - z'_1) = 0 - 8\lambda \leq 0 \text{ or } \lambda \geq 0 \\ c_4(\lambda) - z_4(\lambda) &= (c_4 - z_4) + \lambda(c'_4 - z'_4) = -5 - 5\lambda \leq 0 \text{ or } \lambda \geq -1 \\ c_6(\lambda) - z_6(\lambda) &= (c_6 - z_6) + \lambda(c'_6 - z'_6) = 2 + 3\lambda \leq 0 \text{ or } \lambda \geq 2/3 \end{aligned}$$

Thus for $-1 \leq \lambda \leq -2/3$, the optimal solution is: $x_1 = 0$, $x_2 = 105$, $x_3 = 220$ and $\text{Max } Z = 1310 + 890\lambda$. Hence Table 2.2 to 2.4 give family of optimal solutions for $-2/3 \leq \lambda \leq 1$, $\lambda \geq 1$ and $-1 \leq \lambda \leq -2/3$. The critical value of λ are $-2/3$ and 1 .

10.2 Variation in the availability of resources

Let us define the parametric linear programming model as follows:

$$\begin{aligned} \text{Maximize } Z &= \sum_{j=1}^n c_j x_j \\ \text{subject to the constraints} \\ \sum_{j=1}^n a_{ij} x_j &= b_i \pm \lambda b'_i, \quad i = 1, 2, \dots, m \\ \text{and } x_j &\geq 0, \quad j = 1, 2, \dots, n \end{aligned}$$

where $b_i \pm \lambda b'_i$ is the predetermined variation in resource values (right-hand side values), where $\lambda \geq 0$ is a scalar parameter. Now our aim is to find the range (or critical values) of λ so that the current optimal solution remains unchanged with a change in the right-hand side constants b_i , for all i .

Let $\mathbf{B}_0$ and $\mathbf{x}_{B_0} = \mathbf{B}_0^{-1} \mathbf{b}$ be the optimal basis and optimal basic feasible solution, respectively, of the original LP problem when it is solved at $\lambda = 0$.

If b is replaced by $\mathbf{b} + \lambda \mathbf{b}'$, then the optimality condition $c_j - z_j$ will not be affected. However, such changes will affect the value of the basic variables. The new values are given by:

$$\mathbf{x}_B(\lambda) = \mathbf{B}^{-1}(\mathbf{b} + \lambda \mathbf{b}') = \mathbf{B}^{-1} \mathbf{b} + \lambda \mathbf{B}^{-1} \mathbf{b}' = \mathbf{x}_B + \lambda \mathbf{x}'_B$$

Now as long as $\mathbf{x}_B(\lambda) \geq 0$, the current basis remains optimal. Thus, this criterion can be used to determine the range of λ , within which the solution remains optimal, as follows:

$$\lambda = \text{Min} \left\{ \frac{x_{Bi}}{-x'_{Bi}} \right\} = \frac{x_{Br}}{-x'_{Br}}$$

Let $\lambda = \lambda_1$. Then for $\lambda \in [0, \lambda_1]$, the current solution remains optimal and at this solution the value of the objective function is given by $Z(\lambda) = Z + \lambda Z' = \mathbf{c}_B \mathbf{x}_B + \lambda \mathbf{c}_B \mathbf{x}'_B$. At λ_1 the current basis $\mathbf{x}_B$ (righthand side) is replaced by $\mathbf{x}_B(\lambda) = \mathbf{B}^{-1}(\mathbf{b} + \lambda \mathbf{b}')$ and x_{Br} is removed from the basis by the usual simplex method. The process of finding the new range $[\lambda_1, \lambda_2]$ of values of λ is repeated, over which the new basis is optimal. The process is terminated when $x'_{Bi} = \mathbf{B}^{-1} \mathbf{b}' \geq 0$ for all i . This also

implies that the current basis is optimal for all values of λ greater than or equal to the last value of λ .

Example: Consider the linear programming problem

$$\begin{aligned} &\text{Maximize } Z = 3x_1 + 2x_2 + 5x_3 \\ &\text{subject to the constraints} \\ &(i) x_1 + 2x_2 + x_3 \leq 430 + \lambda \\ &(ii) 3x_1 + 2x_3 \leq 460 - 4\lambda \\ &(iii) x_1 + 4x_2 \leq 420 - 4\lambda \quad \text{and} \quad x_1, x_2, x_3 \geq 0 \end{aligned}$$

Determine the critical value (range) of λ for which the solution remains optimal basic feasible.

Solution : The given parametric LP problem can be written in its standard form as:

$$\begin{aligned} &\text{Maximize } Z = 3x_1 + 2x_2 + 5x_3 + 0s_1 + 0s_2 + 0s_3 \\ &\text{subject to the constraints} \\ &(i) x_1 + 2x_2 + x_3 + s_1 = 430 + \lambda \\ &(ii) 3x_1 + 2x_3 + s_2 = 460 - 4\lambda \\ &(iii) x_1 + 4x_2 + s_3 = 420 - 4\lambda \quad \text{and} \quad x_1, x_2, x_3, s_1, s_2, s_3 \geq 0 \end{aligned}$$

The optimal solution when $\lambda = 0$ is shown in Table 2.5.

$c_j \rightarrow$				3	2	5	0	0	0
c_B	Basic Variables B	Solution Values b' b		x_1	x_2	x_3	s_1	s_2	s_3
2	x_2	3/2	100	-1/4	1	0	1/2	-1/4	0
5	x_3	-2	230	3/2	0	1	0	1/2	0
0	s_3	-10	20	2	0	0	-2	1	1
$Z = 1,350$		z_j		7	2	5	1	2	0
		$c_j - z_j$		-4	0	0	-1	-2	0

Table - 2.5

In order to find the range in which the solution shown in Table 2.5 is optimal, we first calculate

$$\begin{aligned} \mathbf{x}'_B &= \mathbf{B}^{-1} \mathbf{b}' \\ \begin{bmatrix} x_2 \\ x_3 \\ s_3 \end{bmatrix} &= \begin{bmatrix} 1/2 & -1/4 & 0 \\ 0 & 1/2 & 0 \\ -2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -4 \\ -4 \end{bmatrix} = \begin{bmatrix} 3/2 \\ -2 \\ -10 \end{bmatrix} \end{aligned}$$

For a fixed λ , values of basic variables in Table 2.5 becomes: $x_2 = 100 + (3/2)\lambda$, $x_3 = 230 - 2\lambda$ and $s_3 = 20 - 10\lambda$.

The optimal solution shown in Table 2.5 will remain optimal as long as:

$$\begin{aligned} x_2 &= 100 + (3/2)\lambda \geq 0 \quad \text{or} \quad \lambda \leq -200/3 \\ x_3 &= 230 - 2\lambda \geq 0 \quad \text{or} \quad \lambda \leq 115 \\ s_3 &= 20 - 10\lambda \geq 0 \quad \text{or} \quad \lambda \leq 2 \end{aligned}$$

Consequently the solution in Table 2.5 will remain optimal between $-200/3$ and 2 , i.e. $-200/3 \leq \lambda \leq 2$. In particular for any $\lambda \in [0, 2]$, the objective function value and the right-hand side values are given by

$$Z(\lambda) = \mathbf{c}_B \mathbf{x}_B + \lambda \mathbf{c}_B \mathbf{x}'_B = (2, 5, 0) \begin{bmatrix} 100 \\ 230 \\ 20 \end{bmatrix} + \lambda (2, 5, 0) \begin{bmatrix} 3/2 \\ -2 \\ -10 \end{bmatrix} = 1,350 - 7\lambda$$

$$\mathbf{x}_B(\lambda) = \mathbf{x}_B + \lambda \mathbf{x}'_B = \begin{bmatrix} 100 \\ 230 \\ 20 \end{bmatrix} + \lambda \begin{bmatrix} 3/2 \\ -2 \\ -10 \end{bmatrix} = \begin{bmatrix} 100 + 3\lambda/2 \\ 230 - 2\lambda \\ 20 - 10\lambda \end{bmatrix}$$

Evidently for $\lambda > 2$, the new solution will be primal infeasible because $x_{B3}(= s_3)$ will become negative. The solution at $\lambda = 2$ is shown in Table 2.6.

$c_j \rightarrow$			3	2	5	0	0	0
$\mathbf{c}_B$	Basic Variables $\mathbf{B}$	Solution Values $\mathbf{b} (= \mathbf{x}_B)$	x_1	x_2	x_3	s_1	s_2	s_3
2	x_1	103	-1/4	1	0	1/2	-1/4	0
5	x_2	226	3/2	0	1	0	1/2	0
0	s_3	0	2	0	0	-2	1	1
$Z = 1,336$		z_j	7	2	5	1	2	0
		$c_j - z_j$	-4	0	0	-1	-2	0

Table - 2.6

For $\lambda > 2$, the basic variable s_3 becomes negative. Consequently solution becomes infeasible. Therefore dual simplex method is applied to find the new optimal solution. Remove s_3 (because $x_{B3} \leq 0$) from the basis. Determine the ratio $\{c_j - z_j\}/y_{rj}; y_{rj} < 0\} = 1/2$ (corresponds to s_1), and enter s_1 into the basis. The new solution is shown in Table 2.7.

$c_j \rightarrow$			3	2	5	0	0	0
$\mathbf{c}_B$	Basic Variables $\mathbf{B}$	Solution Values $\mathbf{b}'$ $\mathbf{b}$	x_1	x_2	x_3	s_1	s_2	s_3
2	x_2	-1 105	1/4	1	0	0	0	1/4
5	x_3	-2 203	3/2	0	1	0	1/2	0
0	s_1	5 -10	-1	0	0	1	-1/2	-1/2
$Z = 1,336$		z_j	8	2	5	0	5/2	1/2
		$c_j - z_j$	-5	0	0	0	-5/2	-1/2

Table - 2.7

In order to find the next critical value of λ in the interval $[2, \lambda_2]$ in which the solution shown in Table 2.6 remains optimal, we first find

$$\mathbf{x}'_B = \mathbf{B}^{-1} \mathbf{b}' = \begin{bmatrix} 0 & 0 & 1/4 \\ 0 & 1/2 & 0 \\ 0 & -1/2 & -1/2 \end{bmatrix} \begin{bmatrix} 1 \\ -4 \\ -4 \end{bmatrix} = \begin{bmatrix} -1 \\ -2 \\ 5 \end{bmatrix}$$

The solution shown in Table 2.7 will remain optimal as long as basic variable x_2, x_3 and s_1 remain non-negative, i.e.

$$x_2 = 105 - \lambda \geq 0 \text{ or } \lambda \leq 105; \quad x_3 = 230 - 2\lambda \geq 0 \text{ or } \lambda \leq 115; \quad s_1 = -10 + 5\lambda \geq 0 \text{ or } \lambda \geq 2$$

Thus the solution is optimal for all values of λ in the range $2 \leq \lambda \leq 105$.

For $\lambda \in [2, 105]$, the optimal objective function value and the right-hand side values are given by

$$Z(\lambda) = \mathbf{c}_B \mathbf{x}_B + \lambda \mathbf{c}_B \mathbf{x}'_B = [2, 5, 0] \begin{bmatrix} 105 \\ 230 \\ 10 \end{bmatrix} + \lambda [2, 5, 0] \begin{bmatrix} -1 \\ -2 \\ 5 \end{bmatrix} = 1360 - 12\lambda$$

$$\mathbf{x}_B(\lambda) = \mathbf{x}_B + \lambda \mathbf{x}'_B = \begin{bmatrix} 105 \\ 230 \\ 10 \end{bmatrix} + \lambda \begin{bmatrix} -1 \\ -2 \\ 5 \end{bmatrix} = \begin{bmatrix} 105 - \lambda \\ 230 - 2\lambda \\ 10 + 5\lambda \end{bmatrix}$$

Evidently, if $\lambda > 105$, the new solution will be primal infeasible because basic variable x_2 becomes negative. Hence, no optimal solution exists for all $\lambda \geq 105$.

For $\lambda \leq -200/3$, the basic variable in Table 2.7 becomes negative. Applying dual simplex method to find solution for $\lambda \leq -200/3$. Entering non-basic variable s_2 into the basis to replace basic variable x_2 . The new optimal solution at $\lambda = 105$ is shown in Table 2.8.

				$c_j \rightarrow$					
				3	2	5	0	0	0
c_B	Basic Variables B	Solution Values b'	b	x_1	x_2	x_3	s_1	s_2	s_3
0	s_2	-6	400	1	-4	0	-2	1	0
5	x_3	1	430	1	2	1	1	0	0
0	s_3	-4	420	1	4	0	0	0	1
$Z = 2,150$				5	10	5	5	0	0
				-2	-8	0	-5	0	0
				$c_j - z_j$					

Table - 2.8

The critical values of λ for which solution shown in Table 2.8 remains optimal are calculated as follows:

$$\begin{bmatrix} s_2 \\ x_3 \\ s_3 \end{bmatrix} = \mathbf{B}^{-1} \mathbf{b}' = \begin{bmatrix} -2 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -4 \\ -4 \end{bmatrix} = \begin{bmatrix} -6 \\ 1 \\ -4 \end{bmatrix}$$

Hence the basic solution in Table 2.8 will remain optimal provided

$$\begin{aligned} x_3 &= 430 + \lambda \geq 0 \quad \text{or} \quad \lambda \geq -430; \\ s_2 &= 400 - 6\lambda \geq 0 \quad \text{or} \quad \lambda \leq 200/3; \\ s_3 &= 420 - 4\lambda \geq 0 \quad \text{or} \quad \lambda \leq 105 \end{aligned}$$

This implies that the solution is optimal in the range $-430 \leq \lambda \leq 200/3$. For $\lambda < -430$, the basic variable x_3 in Table 2.8 becomes negative. As there is no negative entry in x_3 row of Table 2.8, the prime solution is infeasible. Hence, there exists no optimal solution to the problem for all $\lambda < -430$. Hence, Table 2.5, Table 2.7 and Table 2.8 gives range of λ values, $-200/3 \leq \lambda \leq 2$, $2 \leq \lambda \leq 105$ and $-430 \leq \lambda \leq 200/3$, respectively for which solution is optimal.

Summary

Parametric linear programming techniques are used to determine the effect of pre-determined continuous variation in the input data (or parameters c_j , b_i and a_{ij}), on the optimal solution of an LP problem. The parametric analysis aims at finding various basic solutions that become optimal one after the other due to continuous variations in the LP model parameters. These techniques reduce computational time required to obtain the changes in the optimal solution due to variation in LP model parameters over a range of variation.

Unit 13

Course Structure

- Replacement and Maintenance Models : Failure mechanism of items, General replacement policies for gradual failure of items with constant money value and change of money value at a constant rate over the time period, Selection of best item

Introduction

The problem of replacement is felt when the job performing units such as men, machines, equipments, parts, etc., become less effective or useless due to either sudden, or gradual deterioration in their efficiency, failure or breakdown. By replacing them with new ones at frequent intervals, maintenance and other overhead costs can be reduced. However, such replacements would increase the need of capital cost for new ones. Thus, the basic problem in such situations is to formulate a replacement policy in order to determine an age (or period) at which the replacement of the given machinery/equipment is most economical, keeping in view all possible alternatives. In this unit, we shall discuss the replacement policies in the context of the following three types of replacement situations:

(i) Items such as machines, vehicles, tyres, etc., whose efficiency deteriorates with age due to constant use and which need increased operating and maintenance costs. In such cases the deterioration level is predictable and is represented by (a) increased maintenance/operational cost, (b) its waste or scrap value and damage to item and safety risk.

(ii) Items such as light bulbs and tubes, electric motors, radio, television parts, etc., which do not give any indication of deterioration with time but fail all of a sudden and are rendered useless. Such cases require an anticipation of failures to specify the probability of failure for any future time period. With this probability distribution and the cost information, it is desired to formulate optimal replacement policy in order to balance the wasted life of an item, replaced before failure against the costs incurred when the item fails in service.

(iii) The existing working staff in an organization gradually reduces due to retirement, death, retrenchment and other reasons.

Objective

After studying this chapter you should be able to

- realize the need to study replacement and maintenance analysis techniques.

- make distinctions among various types of failures.
- apply replacement policy for items whose efficiency deteriorates with time and for items that fail completely.
- derive replacement policy for items whose running cost increases with time and those whose value of money either remains constant or the value of money changes with constant rate during the period of time.
- appreciate the use of replacement analysis in handling problems like, 'staffing problem' and 'equipment renewal problem', etc.

11 Replacement and Maintenance Models

11.1 Types of failure

The term 'failure' here will be discussed in the context of replacement decisions. There are two types of failures: (i) Gradual failure, and (ii) Sudden failure.

Gradual failure: Gradual failure is progressive in nature. That is, as the life of an item increases, its operational efficiency also deteriorates. This results in:

- increased running (maintenance and operating) costs
- decrease in its productivity
- decrease in the resale or salvage value

Mechanical items like pistons, rings, bearings, etc., and automobile tyres fall under this category.

Sudden Failure: This type of failure occurs in items after some period of desired service rather than deterioration while in service. The period of desired service is not constant but follows some frequency distribution which may be *progressive*, *retrogressive* or *random* in nature.

(a) *Progressive failure* If the probability of failure of an item increases with the increase in its life, then such a failure is called a progressive failure. For example, light bulbs and tubes fail progressively.

(b) *Retrogressive failure* If the probability of failure in the beginning of the life of an item is more but as time passes the chances of its failure become less, then such failure is said to be retrogressive.

(c) *Random failure* In this type of failure, the constant probability of failure is associated with items that fail from random causes such as physical shocks, not related to age. For example, vacuum tubes in air-born equipment have been found to fail at a rate independent of the age of the tube.

11.2 Replacement of items whose efficiency deteriorates with time

When operational efficiency of an item deteriorates with time (gradual failure), it is economical to replace the same with a new one. For example, the maintenance cost of a machine increases with time and a stage is reached when it may not be economical to allow the machine to continue in the system. Besides, there could be a number of alternative choices and one may like to compare the available alternatives on the basis of the running costs (average maintenance and operating costs) involved. In this section, we shall discuss various techniques for making such comparisons under different conditions. While making such comparisons it is assumed that suitable expressions for running costs are available.

Model I: Replacement Policy for Items Whose Running Cost Increases with Time Model I and Value of Money Remains Constant During a Period

Theorem: The cost of maintenance of a machine is given as a function increasing with time, whose scrap value is constant.

(a) If time is measured continuously then the average annual cost will be minimized by replacing the machine when the average cost to date becomes equal to the current maintenance cost.

(b) If time is measured in discrete units, then the average annual cost will be minimized by replacing the machine when the next period's maintenance cost becomes greater than the current average cost.

Example A firm is considering the replacement of a machine, whose cost price is Rs 12,200, and its scrap value is Rs 200. From experience the running (maintenance and operating) costs are found to be as follows:

Year	:	1	2	3	4	5	6	7	8
Running cost (Rs)	:	200	500	800	1,200	1,800	2,500	3,200	4,000

When should the machine be replaced?

Solution: We are given the running cost, $R(n)$, the scrap value $S = \text{Rs } 200$ and the cost of the machine, $C = \text{Rs } 12,200$. In order to determine the optimal time n when the machine should be replaced, we first calculate the average cost per year during the life of the machine, as shown in Table 3.1.

Year of Service n	Running Cost (Rs) $R(n)$	Cumulative Running Cost (Rs) $\Sigma R(n)$	Depreciation Cost (Rs) $C - S$	Total Cost (Rs) TC	Average Cost (Rs) ATC_n
(1)	(2)	(3)	(4)	(5) = (3) + (4)	(6) = (5)/(1)
1	200	200	12,000	12,200	12,000
2	500	700	12,000	12,700	6,350
3	800	1,500	12,000	13,500	4,500
4	1,200	2,700	12,000	14,700	3,675
5	1,800	4,500	12,000	16,500	3,300
6	2,500	7,000	12,000	19,000	3,167
7	3,200	10,200	12,000	22,200	3,171
8	4,000	14,200	12,000	26,200	3,275

Table - 3.1

The average cost per year, $ATC_n = \text{Rs } 3,167$ is minimum in the sixth year as shown in Table 17.1. Also the average cost, Rs 3,171 in seventh year is more than the cost in the sixth year. Hence, the machine should be replaced after every six years.

Example: The data collected in running a machine, the cost of which is Rs 60,000 are given below:

Year	:	1	2	3	4	5
Resale value (Rs)	:	42000	30000	20400	14400	9650
Cost of spare (Rs)	:	4000	4270	4880	5700	6800
Cost of labour (Rs)	:	14000	16000	18000	21000	25000

Solution: The costs of spares and labour, together determine the running (operational or maintenance) cost. Thus, the running costs and the resale price of the machine in successive years are as follows:

Year	:	1	2	3	4	5
Resale value (Rs)	:	42000	30000	20400	14400	9650
Running cost (Rs)	:	18000	20270	22880	26700	31800

The calculations of average running cost per year during the life of the machine are shown in Table 3.4.

Year of Service n	Running Cost (Rs) $R(n)$	Cumulative Running Cost (Rs) $\Sigma R(n)$	Resale Value (Rs) S	Depreciation Cost (Rs) $C - S$	Total Cost (Rs) TC	Average Cost (Rs) ATC_n
(1)	(2)	(3)	(4)	(5) = 60,000 - (4)	(6) = (3) + (5)	(7) = (6)/(1)
1	18,000	18,000	42,000	18,000	36,000	36,000.00
2	20,270	38,270	30,000	30,000	68,270	34,135.00
3	22,880	61,150	20,400	39,600	1,00,750	33,583.30
4	26,700	87,850	14,400	45,600	1,33,450	33,362.50
5	31,800	1,19,650	9,650	50,350	1,70,000	34,000.00

Table - 3.4

The average cost, $ATC_4 = \text{Rs } 33,362.50$ is the lowest during the fourth year as shown in Table 17.2. Hence, the machine should be replaced after every four years, otherwise the average cost per year for running the machine would start increasing.

Model II: Replacement Policy for Items Whose Running Cost Increases with Time but Value of Money Changes with Constant Rate During a Period

Value of money criterion: If the effect of the time-value of money is to be considered, then replacement decision must be based upon an equivalent annual cost. For example, if the interest rate on Rs 100 is 10 per cent per year, then the value of Rs 100 to be spent after one year will be Rs 110. This is also called value of money. Also, the value of money that decreases with constant rate is known as its depreciation ratio or discounted factor. The discounted value is the amount of money required to build up funds at compound interest that is sufficient to pay the required cost when due.

For example, if the interest rate on Rs 100 is r per cent per year, then the present value (or worth) of Rs 100 to be spent after n years will be:

$$d = \left(\frac{100}{100 + r} \right)^n$$

where d is the discount rate or depreciation value. After calculating depreciation value, we need to determine the critical age at which an item should be replaced so that the sum of all discounted costs is minimum.

Example: Let the value of the money be assumed to be 10 per cent per year and suppose that machine A is replaced after every three years, whereas machine B is replaced every six years. The yearly costs (in Rs) of both the machines are given below:

Year	:	1	2	3	4	5	6
Machine A	:	1000	200	400	1000	200	400
Machine B	:	1700	100	200	300	400	500

Determine which machine should be purchased.

Solution : The discounted cost (present worth) at 10 per cent rate per year for machine A for three years and B for six years is given in Tables 3.2 and 3.4, respectively.

Year	Discounted Cost at 10% Rate (Rs)		
	Cost	Present Worth	
1	1,000	1000×1	= 1,000
2	200	$200 \left(\frac{100}{100+10} \right)$	= 200×0.9091 = 181.82
3	400	$400 \left(\frac{100}{100+10} \right)^2$	= 400×0.8264 = 330.58
Total			Rs 1,512.40

Table - 3.2

Thus, the average yearly cost of machine A is $1,512.40/3 = \text{Rs } 504.13$.

Year	Discounted Cost at 10% Rate (Rs)		
	Cost	Present Worth	
1	1,700	$1,700 \times 1$	= 1,700.00
2	100	$100 \times (10/11)$	= 100×0.9091 = 90.91
3	200	$200 \times (10/11)^2$	= 200×0.8264 = 165.28
4	300	$300 \times (10/11)^3$	= 300×0.7513 = 225.39
5	400	$400 \times (10/11)^4$	= 400×0.6830 = 273.20
6	500	$500 \times (10/11)^5$	= 500×0.6209 = 310.45
Total			Rs 2,765.23

Table - 3.3

Thus, the average yearly cost of machine B is $2,765.23/6 = \text{Rs } 460.87$.

With the data on average yearly cost of both the machines, the apparent advantage is in purchasing machine B. But, the periods for which the costs are considered are different. Therefore, let us first

calculate the total present worth of machine A for six years.

$$\begin{aligned} \text{Total present worth} \\ = 1,000 + 2000 \cdot 0.9091 + 400 \times 0.8264 + 1,000 \times 0.7513 + 200 \times 0.6830 + 400 \times 0.6209 = \text{Rs } 2,648.64 \end{aligned}$$

This is less than the total present worth of machine B. Thus machine A should be purchased.

Example: A pipeline is due for repairs. The repair would cost Rs 10,000 and would last for three years. Alternately, a new pipeline can be laid at a cost of Rs 30,000, which would last for 10 years. Assuming the cost of capital to be 10 per cent and ignoring salvage value, which alternative should be chosen?

Solution: Consider the two types of pipelines for infinite replacement cycles of ten years for the new pipeline and three years for the existing pipeline.

Since the discount rate of money per year is 10 per cent, the present worth of the money to be spent over a period of one year is: $d = 100/(100 + 10) = 0.9091$

Let D_n be the discounted value of all future costs associated with a policy of replacing the equipment after n years. Then

$$\begin{aligned} D_n &= c + c \times d^n + c \times d^{2n} + \dots \\ &= c(1 + d^n + d^{2n} + \dots) = \frac{c}{1 - d^n} \quad (\text{sum of infinite GP}) \end{aligned}$$

where c is the initial cost.

Substituting the values of c , d 's and n for two types of pipelines, we get:

$$\begin{aligned} D_3 &= \frac{10000}{1 - (0.9091)^3} = \text{Rs } 4021 \text{ (approx), for existing pipeline} \\ \text{and } D_{10} &= \frac{30000}{1 - (0.9091)^{10}} = \text{Rs } 48820, \text{ for new pipeline} \end{aligned}$$

Since the value of $D_3 < D_{10}$, the existing pipeline should be continued. Alternatively, the comparison may be made over $3 \times 10 = 30$ years.

Summery

The problem of replacement is felt when the job performing units such as men, machines, equipments, parts, etc., become less effective or useless due to either sudden or gradual deterioration in their efficiency, failure or breakdown. By replacing them with new ones at frequent intervals, maintenance and other overhead costs can be reduced. However, such replacements would increase the need of capital cost for new ones. The basic problem in such situations is to formulate a replacement policy in order to determine an age (or period) at which the replacement of the given machinery/equipment is most economical, keeping in view all possible alternatives.

Unit 14

Course Structure

- Dynamic Programming (DP) : Basic features of DP problems, Bellman's principle of optimality, Multistage decision process with Forward and Backward recursive relations, DP approach to stage-coach problems.

Introduction

The decision-making process often involves several decisions to be taken at different times. For example, problems of inventory control, evaluation of investment opportunities, long-term corporate planning, and so on, require sequential decision-making. The mathematical technique of optimizing a sequence of interrelated decisions over a period of time is called *dynamic programming*. The dynamic programming approach uses the idea of recursion to solve a complex problem, broken into a series of interrelated (sequential) decision stages (also called *subproblems*) where the outcome of a decision at one stage affects the decision at each of the following stages. The word *dynamic* has been used because time is explicitly taken into consideration.

Dynamic programming (DP) differs from linear programming in two ways:

(i) In DP, there is no set procedure (algorithm) as in LP to solve any decision-problem. The DP technique allows to break the given problem into a sequence of smaller subproblems, which are then solved in a sequential order (stage).

(ii) LP approach provides one-time period (single stage) solution to a problem whereas DP approach is useful for decision-making over time and solves each subproblem optimally.

Objective

After studying this chapter, you should be able to

- make distinction between linear programming and dynamic programming approaches for solving a problem.
- develop recursive function based on Bellmain's principle of optimality to get an optimal solution of any multi-stage decision problem.
- understand various dynamic programming models and their applications in solving a decision-problem.
- solve an LP problem using the dynamic programming approach.

12 Dynamic Programming

12.0.1 DYNAMIC PROGRAMMING TERMINOLOGY

Stage: The dynamic programming problem can be decomposed or divided into a sequence of smaller subproblems called stages. At each stage there are a number of decision alternatives (courses of action) and a decision is made by selecting the most suitable alternative. Stages represent different time periods in the planning period. For example, in the replacement problem each year is a stage, in the salesman allocation problem each territory represents a stage.

State: Each stage in a dynamic programming problem is associated with a certain number of states. These states represent various conditions of the decision process at a stage. The variables that specify the condition of the decision process or describe the status of the system at a particular stage are called state variables. These variables provide information for analyzing the possible effects that the current decision could have upon future courses of action. At any stage of the decision-making process there could be a finite or infinite number of states. For example, a specific city is referred to as state variable, in any stage of the shortest route problem.

Return function: At each stage, a decision is made that can affect the state of the system at the next stage and help in arriving at the optimal solution at the current stage. Every decision that is made has its own worth or benefit associated and can be described in an algebraic equation form, called a return function. This return function, in general, depends upon the state variable as well as the decision made at a particular stage. An optimal policy or decision at a stage yields optimal (maximum or minimum) return for a given value of the state variable.

13 Developing Optimal Decision Policy

Dynamic programming is an approach in which the problem is broken down into a number of smaller subproblems called *stages*. These subproblems are then solved sequentially until the original problem is finally solved. A particular sequence of alternatives (courses of action) adopted by the decision-maker in a multistage decision problem is called a *policy*. The optimal policy, therefore, is the sequence of alternatives that achieves the decision-maker's objective. The solution of a dynamic programming problem is based upon *Bellman's principle of optimality* (recursive optimization technique), which states:

The optimal policy must be one such that, regardless of how a particular state is reached, all later decisions (choices) proceeding from that state must be optimal.

Based on this principle of optimality, an optimal policy is derived by solving one stage at a time, and then sequentially adding a series of one-stage-problems that are solved until the optimal solution of the initial problem is obtained. The solution procedure is based on a *backward induction process* and *forward induction process*. In the first process, the problem is solved by solving the problem in the last stage and working backwards towards the first stage, making optimal decisions at each stage of the problem. In the forward process is used to solve a problem by first solving the initial stage

of the problem and working towards the last stage, making an optimal decision at each stage of the problem.

14 The General Procedure

The procedure for solving a problem by using the dynamic programming approach can be summarized in the following steps:

Step 1: Identify the problem decision variables and specify the objective function to be optimized under certain limitations, if any.

Step 2: Decompose (or divide) the given problem into a number of smaller sub-problems (or stages). Identify the state variables at each stage and write down the transformation function as a function of the state variable and decision variable at the next stage.

Step 3: Write down a general recursive relationship for computing the optimal policy. Decide whether to follow the forward or the backward method for solving the problem.

Step 4: Construct appropriate tables to show the required values of the return function at each stage as shown in Table 4.1.

Step 5: Determine the overall optimal policy or decisions and its value at each stage. There may be more than one such optimal policy.

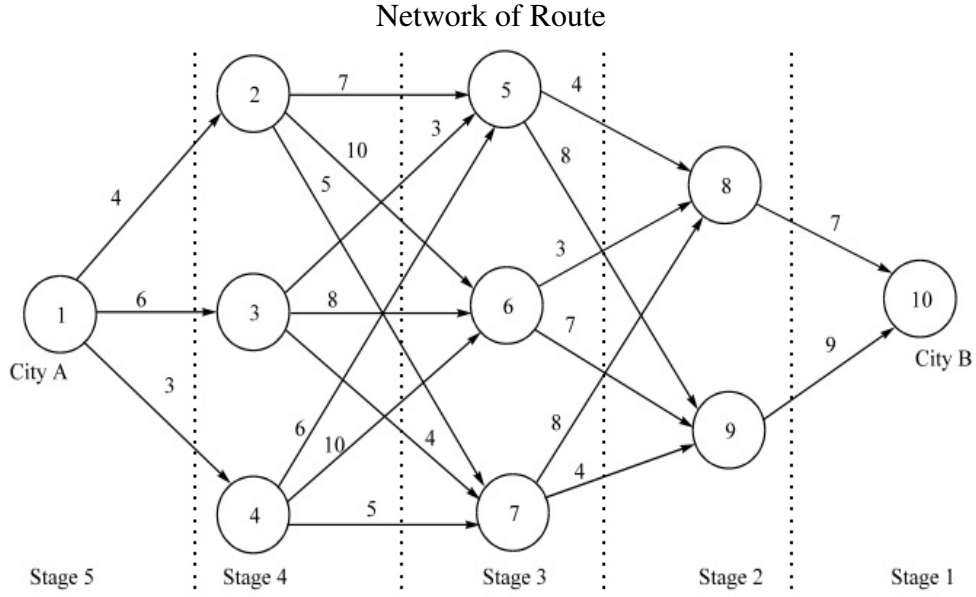
<i>Decision, $d_n \rightarrow$</i> <i>States, $s_n \downarrow$</i>	$f_n(s_n, d_n)$	<i>Optimal Return</i>	<i>Optimal Decision</i>
	d_n	$f_n^*(s_n)$	d_n^*

Table - 4.1

Shortest Route Problem

Example: A salesman located in a city A decided to travel to city B. He knew the distances of alternative routes from city A to city B. He then drew a highway network map as shown in the Fig. 22.3. The city of origin A, is city 1. The destination city B, is city 10. Other cities through which the salesman will have to pass through are numbered 2 to 9. The arrow representing routes between cities and distances in kilometers are indicated on each route. The salesman's problem is to find the shortest route that covers all the selected cities from A to B.

Solution: To solve the problem, we need to define problem stages, decision variables, state variables, return function and transition function. For this particular problem, the following definitions will be used to denote various the state and decision variables.



d_n = decision variables that define the immediate destinations when there are n ($n = 1, 2, 3, 4,$) stages to go.

s_n = state variables describe a specific city at any stage.

D_{s_n, d_n} = distance associated with the state variable, s_n , and the decision variable, d_n for the current n th stage.

$f_n(s_n, d_n)$ = minimum total distance for the last n stages, given that salesman is in state s_n and selects d_n as immediate destination.

$f_n^*(s_n)$ = optimal path (minimum distance) when the salesman is in state s_n with n more stages to go for reaching the final stage (destination).

We start calculating distances between a pair of cities from destination city 10 ($= x_1$) and work backwards $x_5 \rightarrow x_4 \rightarrow x_3 \rightarrow x_2 \rightarrow x_1$ to find the optimal path. The recursion relationship for this problem can be stated as follows:

$$f_n^*(s_n) = \min_{d_n} \{ D_{s_n, d_n} + f_{n-1}^*(d_n) \}; \quad n = 1, 2, 3, 4$$

where $f_{n-1}^*(d_n)$ is the optimal distance for the previous stages.

Working backward in stages from city B to city A, we determine the shortest distance to city B (node 10) in stage 1, from state $s_1 = 8$ (node 8) and state $s_1 = 9$ (node 9) in stage 2. Since the distances associated with entering stage 2 from state $s_1 = 8$ and $s_1 = 9$ are $D_{8,10} = 7$ and $D_{9,10} = 9$, respectively, the optimal value of $f_1^*(s_1)$ is the minimum value between $D_{8,10}$ and $D_{9,10}$. These results are shown in Table 4.2.

Decision, $d_1 \rightarrow$		$f_1(s_1, d_1) = D_{s_1, d_1}$	Minimum Distance	Optimal Decision
		10	$f_1^*(s_1)$	d_1
States, s_1	8	7	7	10
	9	9	9	10

Table - 4.2

We move backward to stage 3. Suppose that the salesman is at state $s_2 = 5$ (node 5). Here he has to decide whether he should go to either $d_2 = 8$ (node 8) or $d_2 = 9$ (node 9). For this he must evaluate two sums:

$$D_{5,8} + f_1^*(8) = 4 + 7 = 11 \text{ (to state } s_1 = 8)$$

$$D_{5,9} + f_1^*(9) = 8 + 9 = 17 \text{ (to state } s_1 = 9)$$

This distance function for travelling from state $s_2 = 5$ is the smallest of these two sums:

$$f_2(s_2) = \text{Min}_{d_2=8,9}\{11, 17\} = 11 \text{ (to state } s_1 = 8)$$

Similarly, the calculation of distance function for travelling from state $s_2 = 6$ and $s_2 = 7$ can be completed as follows:

$$\begin{aligned} \text{For state, } s_2 = 6 \quad f_2(6) &= \text{Min}_{d_2=8,9} \begin{cases} D_{6,8} + f_1^*(8) = 3 + 7 = 10 \\ D_{6,9} + f_1^*(9) = 7 + 9 = 16 \end{cases} \\ &= 10 \text{ (to state } s_1 = 8) \\ \text{For state, } s_2 = 7 \quad f_2(7) &= \text{Min}_{d_2=8,9} \begin{cases} D_{7,8} + f_1^*(8) = 8 + 7 = 15 \\ D_{7,9} + f_1^*(9) = 4 + 9 = 13 \end{cases} \\ &= 13 \text{ (to state } s_1 = 9) \end{aligned}$$

These results are entered into the two-stage table as shown in Table 4.3.

Decision, $d_2 \rightarrow$		$f_2(s_2, d_2) = D_{s_2, d_2} + f_1^*(d_2)$		Minimum Distance	Optimal Decision
		8	9	$f_2^*(s_2)$	d_2
States, s_2	5	11	17	11	8
	6	10	16	10	8
	7	15	13	13	9

Table - 4.3

The results that we obtain by continuing the same process for stages 4 and 5, are shown in Tables 4.4 and 4.5.

Table - 4.4 & 4.5

Decision, $d_3 \rightarrow$		$f_3(s_3, d_3) = D_{s_3, d_3} + f_2^*(d_3)$			Minimum Distance $f_3^*(s_3)$	Optimal Decision d_3
		5	6	7		
States, s_3	2	18	20	18	18	5 or 7
	3	14	18	17	14	5
	4	17	20	18	17	5

Decision, $d_4 \rightarrow$		$f_4(s_4, d_4) = D_{s_4, d_4} + f_3^*(d_4)$			Minimum Distance $f_4^*(s_4)$	Optimal Decision d_4
		2	3	4		
States, s_4	1	22	20	20	20	3 or 4

The above optimal results at various stages can be summarized as below:

$$\begin{array}{l}
 \text{Entering states (nodes)} \\
 \text{Sequence } \left\{ \begin{array}{l} 10 \quad 8 \quad 5 \quad 3 \quad 1 \\ 10 \quad 8 \quad 5 \quad 4 \quad 1 \end{array} \right. \\
 \text{Distances } \left\{ \begin{array}{l} 7 \quad 4 \quad 3 \quad 6 = 20 \\ 7 \quad 4 \quad 6 \quad 3 = 20 \end{array} \right.
 \end{array}$$

From the above, it is clear that there are two alternative shortest routes for this problem, both having a minimum distance of 20 kilometres.

Unit 15

Most mathematical techniques for solving nonlinear programming problems are very complex. In this module two of the more well known but simpler mathematical methods will be demonstrated—the substitution method and the method of Lagrange multipliers.

The Substitution Method

*In the **substitution method** the constraint equation is solved for one variable in terms of another and then substituted into the objective function.*

The least complex method for solving nonlinear programming problems is referred to as the **substitution method**. This method is restricted to models that contain only equality constraints, and typically only one of these. The method involves solving the constraint equation for one variable in terms of another. This new expression is then substituted into the objective function, effectively eliminating the constraint. In other words, a constrained optimization model is transformed into an unconstrained model.

For an example of the substitution method we will use a profit analysis model. This is a nonlinear model that we introduced in Chapter 10 of the text. The demand function is a constraint. The nonlinear programming model is formulated as

$$\begin{aligned} \text{maximize } Z &= vp - c_f - vc_v \\ \text{subject to} \\ v &= 1,500 - 24.6p \end{aligned}$$

The objective function in this model is nonlinear, because both v (volume) and p (price) are variables, and multiplying them (i.e., vp) creates a curvilinear relationship.

The constraint has already been solved for one variable (v) in terms of another (p); thus, we can substitute this expression directly into the objective function. This results in the following unconstrained function:

$$Z = 1,500p - 24.6p^2 - c_f - 1,500c_v + 24.6pc_v$$

By substituting the constant values for c_f (\$10,000) and c_v (\$8), we obtain

$$Z = 1,696.8p - 24.6p^2 - 22,000$$

Next, we solve this problem by differentiating the function Z and setting it equal to zero:

$$\begin{aligned} \frac{\partial Z}{\partial p} &= 1,696.8 - 49.2p \\ 0 &= 1,696.8 - 49.2p \\ 49.2p &= 1,696.8 \\ p &= \$34.49 \end{aligned}$$

We will present another example as a further illustration of a nonlinear programming problem and the substitution method. In this example the Beaver Creek Pottery Company produces bowls (x_1) and mugs (x_2). We will assume that the profit contribution for each product *declines* as the quantity of each item produced increases. Thus, for bowls the per unit profit contribution is now expressed according to the relationship

$$\$4 - 0.1x_1$$

For mugs, the profit contribution per unit is

$$\$5 - 0.2x_2$$

(These relationships express the fact that production costs for each product increase as the number of units sold increases.)

THE SUBSTITUTION METHOD

These profit relationships are on a per unit basis. Thus, the total profit contribution from each product is determined by multiplying these relationships by the number of units produced. For bowls, the profit contribution is

$$(4 - 0.1x_1)x_1$$

or

$$4x_1 - 0.1x_1^2$$

For mugs, the profit contribution is

$$(5 - 0.2x_2)x_2$$

or

$$5x_2 - 0.2x_2^2$$

Total profit is the sum of these two terms:

$$Z = \$4x_1 - 0.1x_1^2 + 5x_2 - 0.2x_2^2$$

In this model, we will consider only a labor constraint, and we will treat it as an equality rather than an inequality:

$$x_1 + 2x_2 = 40 \text{ hr.}$$

The complete nonlinear programming model is as follows:

$$\text{maximize } Z = \$4x_1 - 0.1x_1^2 + 5x_2 - 0.2x_2^2$$

subject to

$$x_1 + 2x_2 = 40$$

The first step in the substitution method is to solve the constraint equation for one variable in terms of another. We will arbitrarily decide to solve for x_1 as follows:

$$x_1 = 40 - 2x_2$$

Now wherever x_1 appears in the nonlinear objective function, we will substitute the expression $40 - 2x_2$:

$$\begin{aligned} Z &= 4(40 - 2x_2) - 0.1(40 - 2x_2)^2 + 5x_2 - 0.2x_2^2 \\ &= 160 - 8x_2 - 0.1(1,600 - 1,600x_2 + 4x_2^2) + 5x_2 - 0.2x_2^2 \\ &= 160 - 8x_2 - 160 + 160x_2 - 0.4x_2^2 + 5x_2 - 0.2x_2^2 \\ &= 13x_2 - 0.6x_2^2 \end{aligned}$$

This is an unconstrained optimization function, and we can solve it by differentiating it and setting it equal to zero:

$$\begin{aligned} \frac{\partial Z}{\partial x_2} &= 13 - 1.2x_2 \\ 0 &= 13 - 1.2x_2 \\ 1.2x_2 &= 13 \\ x_2 &= 10.8 \text{ mugs} \end{aligned}$$

To determine x_1 , we can substitute x_2 into the constraint equation:

$$\begin{aligned} x_1 + 2x_2 &= 40 \\ x_1 + 2(10.8) &= 40 \\ x_1 &= 18.4 \text{ bowls} \end{aligned}$$

Substituting the values of x_1 and x_2 into the original objective function gives the total profit, as follows:

$$\begin{aligned} Z &= \$4x_1 - 0.1x_1^2 + 5x_2 - 0.2x_2^2 \\ &= \$4(18.4) - 0.1(18.4)^2 + 5(10.8) - 0.2(10.8)^2 \\ &= \$70.42 \end{aligned}$$

Both of the examples presented in this section for solving nonlinear programming problems exhibit the limitations of this approach. The objective functions were not very complex (i.e., the highest order of a variable was a power of two in the second example), there were only two variables, and the single constraint in each example was an equation. This method becomes very difficult if the constraint becomes complex. An alternative solution approach that is not quite as restricted is the method of Lagrange multipliers.

The Method of Lagrange Multipliers

*In the **method of Lagrange multipliers**, constraints as multiples of a multiplier, λ , are subtracted from the objective function, which is then differentiated with respect to each variable and solved.*

The **method of Lagrange multipliers** is a general mathematical technique that can be used for solving constrained optimization problems consisting of a nonlinear objective function and one or more linear or nonlinear constraint equations. In this method, the constraints as multiples of a Lagrange multiplier, λ , are subtracted from the objective function.

To demonstrate this method, we will use our modified pottery company example developed in the preceding section. This model was formulated as

$$\begin{aligned} \text{maximize } Z &= 4x_1 - 0.1x_1^2 + 5x_2 - 0.2x_2^2 \\ \text{subject to} \\ x_1 + 2x_2 &= 40 \end{aligned}$$

The first step is to transform the nonlinear objective function into a *Lagrangian function*. This is accomplished by transforming the constraint equation as follows:

$$x_1 + 2x_2 - 40 = 0$$

Next, this expression is multiplied by λ , the *Lagrangian multiplier*, and subtracted from the objective function to form the Lagrangian function L :

$$L = 4x_1 - 0.1x_1^2 + 5x_2 - 0.2x_2^2 - \lambda(x_1 + 2x_2 - 40)$$

Because the constraint equation now equals zero, the subtraction of the constraint, multiplied by λ , from the objective function does not affect the value of the function. (We will explain the exact meaning of λ after a solution has been determined.)

Now we must determine the partial derivatives of the **Lagrangian function** with respect to each of the three variables, x_1 , x_2 , and λ :

$$\begin{aligned} \frac{\partial L}{\partial x_1} &= 4 - 0.2x_1 - \lambda \\ \frac{\partial L}{\partial x_2} &= 5 - 0.4x_2 - 2\lambda \\ \frac{\partial L}{\partial \lambda} &= -x_1 - 2x_2 + 40 \end{aligned}$$

*The **Lagrangian function** is differentiated with respect to each variable, and the resulting equations are solved simultaneously.*

THE METHOD OF LAGRANGE MULTIPLIERS

These three equations are all set equal to zero and solved simultaneously to determine the values of x_1 , x_2 , and λ :

$$\begin{aligned} 4 - 0.2x_1 - \lambda &= 0 \\ 5 - 0.4x_2 - 2\lambda &= 0 \\ -x_1 - 2x_2 + 40 &= 0 \end{aligned}$$

To solve these equations simultaneously, we multiply the first equation by -2 and add it to the second equation, which eliminates λ :

$$\begin{array}{r} -8 + 0.4x_1 + 2\lambda = 0 \\ 5 - 0.4x_2 - 2\lambda = 0 \\ \hline -3 + 0.4x_1 - 0.4x_2 = 0 \end{array}$$

This new equation and the original preceding third equation represent two equations with two unknowns (x_1 and x_2). We multiply the preceding third equation by 0.4 and add it to the new equation in order to eliminate x_1 :

$$\begin{array}{r} -0.4x_1 - 0.8x_2 + 16 = 0 \\ 0.4x_1 - 0.4x_2 - 3 = 0 \\ \hline -1.2x_2 + 13 = 0 \end{array}$$

The resulting equation is solved for x_2 as follows:

$$\begin{aligned} -1.2x_2 + 13 &= 0 \\ -1.2x_2 &= -13 \\ x_2 &= 10.8 \text{ mugs} \end{aligned}$$

Substituting this value back into the previous equations gives the values of both x_1 and λ :

$$\begin{aligned} -x_1 - 2x_2 + 40 &= 0 \\ -x_1 - 2(10.8) &= -40 \\ x_1 &= 18.3 \text{ bowls} \\ 5 - 0.4x_2 - 2\lambda &= 0 \\ 5 - 0.4(10.8) &= 2\lambda \\ \lambda &= 0.33 \end{aligned}$$

Substituting the values for x_1 and x_2 into the original objective function yields the total profit:

$$\begin{aligned} Z &= \$4x_1 - 0.1x_1^2 + 5x_2 - 0.2x_2^2 \\ &= \$4(18.3) - 0.1(18.3)^2 + 5(10.8) - 0.2(10.8)^2 \\ &= \$70.42 \end{aligned}$$

This result can also be obtained by using the Lagrangian function, L , and multiplier, λ :

$$\begin{aligned} L &= \$4x_1 - 0.1x_1^2 + 5x_2 - 0.2x_2^2 - \lambda(x_1 + 2x_2 - 40) \\ &= \$4(18.3) - 0.1(18.3)^2 + 5(10.8) - 0.2(10.8)^2 - 0.33(0) \\ &= \$70.42 \end{aligned}$$

To summarize, we have $x_1 = 18.3$ bowls, $x_2 = 10.8$ mugs, $Z = \$70.42$, and $\lambda = 0.33$. This is the same answer obtained previously using the substitution method. However, unlike the substitution method, the Lagrange multiplier approach can be used to solve nonlinear programming problems with more complex constraint equations and *inequality constraints*. In addition, it can encompass problems with more than two variables. We will not pursue any examples, though, that demonstrate the complexities involved. The

Lagrange multiplier method must be altered to compensate for inequality constraints and additional variables, and the resulting mathematics are very difficult. Even though the Lagrange multiplier method is more flexible than the substitution method, it is practical for solving only small problems. As the size of the problem increases, the mathematics become overwhelmingly difficult.

The Meaning of λ

*The **Lagrange multiplier**, λ , reflects a change in the objective function from a unit change in the right-hand-side value of a constraint.*

The **Lagrange multiplier**, λ , in nonlinear programming problems is analogous to the dual variables in a linear programming problem. It reflects the approximate change in the objective function resulting from a unit change in the quantity (right-hand-side) value of the constraint equation. For our example, we will increase the quantity value in the constraint equation from 40 to 41 hours of labor. (You will recall from the preceding calculations that $\lambda = 0.33$.)

$$x_1 + 2x_2 = 41 \quad \text{or} \quad x_1 + 2x_2 - 41 = 0$$

This constraint equation will result in the following Lagrangian function:

$$L = 4x_1 - 0.1x_1^2 + 5x_2 - 0.2x_2^2 - \lambda(x_1 + 2x_2 - 41)$$

Solving this problem the same way we solved the original model gives the following solution:

$$\begin{aligned} x_1 &= 18.7 \\ x_2 &= 11.2 \\ \lambda &= 0.27 \\ Z &= \$70.75 \end{aligned}$$

This value for Z is \$0.33 greater than the previous Z value of \$70.42. Thus, a one-unit increase in the right-hand side of the constraint equation results in a λ increase in the objective function. More specifically, a unit increase in a resource (labor) results in a \$0.33 increase in profit. Thus, we would be willing to pay \$0.33 for one additional hour of labor. This is the same interpretation as that given for a dual variable in linear programming.

In general, if λ is positive, the optimal objective function value will increase if the quantity (absolute) value in the constraint equation is increased, and it will decrease if the quantity (absolute) value is decreased. On the other hand, if λ is negative, the optimal objective function value will increase if the quantity (absolute) value is decreased, and it will decrease if the quantity (absolute) value is increased.

Unit 16

Course Structure

- Separable programming, Piecewise linear approximation solution approach, Linear fractional programming.

Introduction

Separable programming is one of the indirect methods used to solve a non-linear programming problem. Indirect methods solve an NLP problem by dealing with one or more linear problems that are extracted from the original problem.

Separable programming is useful in solving those NLP problems in which the objective function and constraints are separable. Sometimes, functions that are not separable can be made separable by using simplified approximation. Such approximation reduces the single variable non-linear function into piece-wise linear functions, as shown in Fig 6.1.

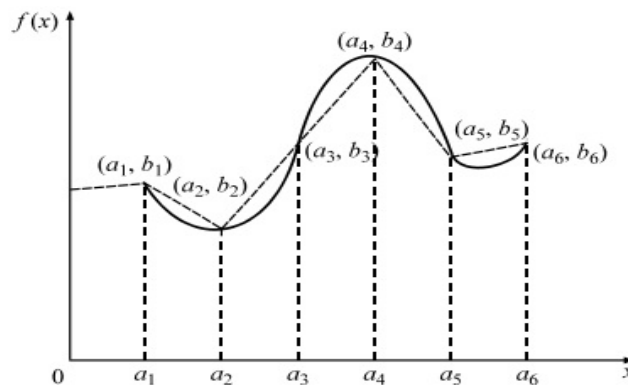


Fig 6.1 Linear approximation of a function

There is no particular method to determine the exact number of such piece-wise linear segments. Efforts should be made to have large number of linear functions (or segments) to reduce the chance of error in the approximation. However, such a number will increase the size of the problem and more computational time would obviously be required to obtain the optimal solution.

Separable Functions

A function $f(x_1, x_2, \dots, x_n)$ that can be expressed as the sum of n single-variable functions, $f_1(x_1)$, $f_2(x_2)$, $\dots$, $f_n(x_n)$ such that:

$$f(x_1, x_2, \dots, x_n) = f_1(x_1) + f_2(x_2) + \dots + f_n(x_n)$$

is said to be a separable function. For example, the linear function

$$h(x_1, x_2, \dots, x_n) = c_1x_1 + c_2x_2 + \dots + c_nx_n$$

(where c 's are constants) is a separable function. But the function

$$h(x_1, x_2, \dots, x_n) = x_1^2 + x_1 \cos(x_2 + x_3) + x_3 2^{x_2}$$

is not a separable function.

Reduction to separable form : A few non-linear functions are not directly separable, but can be separated by applying suitable substitutions. For example, in the function $y = x_1 \cdot x_2$, the non-separable term $x_1 \cdot x_2$ can be expressed in terms of two linear separable functions by taking log on both sides:

$$\log y = \log x_1 + \log x_2$$

The problem can then be stated as:

$$\begin{aligned} &\text{Maximize } Z = y \\ &\text{subject to the constraint} \\ &\log y = \log x_1 + \log x_2 \end{aligned}$$

This problem is separable. Since logarithmic function is undefined for non-positive values, substitution assumes that both x_1 and x_2 are positive.

If x_1 and x_2 assume zero values (i.e. $x_1, x_2 \geq 0$), then two new variables u_1 and u_2 are defined as follows:

$$\begin{aligned} u_1 &= x_1 + a_1 \quad \text{and} \quad u_2 = x_2 + a_2 \\ \text{or } x_1 &= u_1 - a_1 \quad \text{and} \quad x_2 = u_2 - a_2 \end{aligned}$$

where a_1 and a_2 are positive constants. This implies that both u_1 and u_2 are strictly positive. Now:

$$x_1 x_2 = (u_1 - a_1)(u_2 - a_2) = u_1 u_2 - a_1 u_2 - a_2 u_1 + a_1 a_2$$

Let $y = u_1 u_2$. The original problem is then stated as:

$$\begin{aligned} &\text{Maximize } Z = y - a_1 u_2 - a_2 u_1 + a_1 a_2 \\ &\text{subject to the constraint} \\ &\log y = \log u_1 + \log u_2; \quad u_1 \geq a_1, \quad u_2 \geq a_2 \end{aligned}$$

This problem is also separable. Few other functions that can also be expressed as separable functions using suitable substitution are:

$$e^{x_1+x_2}, x_1^{x_2}, (x_1)^{1/2}(x_2^2 + e^{x_2})^{-2} \text{ etc.}$$

Definitions

Separable programming problem : If the objective function of an NLP problem can be expressed as a linear combination of several different one-variable functions, of which some or all are non-linear, then such an NLP problem is called a *separable programming problem*.

Separable convex programming : It is the special case of separable programming in which separate functions are convex. Also, the non-linear function $f(x)$ is convex in case of minimization and concave in case of maximization.

For example, if $f(x)$ is the non-linear objective function, then for separable convex programming, it is expressed as:

$$f(x) = \sum_{j=1}^n f_j(x_j)$$

where all $f_j(x_j)$ are convex.

Illustration Let $f(x) = 9x_1^2 + 5x_2^2 - 5x_1 + 2x_2$. Then $f(x)$ is separated as:

$$f_1(x_1) = 9x_1^2 - 5x_1 \quad \text{and} \quad f_2(x_2) = 5x_2^2 + 2x_2$$

where both $f_1(x_1)$ and $f_2(x_2)$ are convex functions, such that $f(x) = f_1(x_1) + f_2(x_2)$.

Piece-Wise Linear Approximation of Non-linear Functions : In this section, we shall discuss piece-wise linear approximation method to reduce a separable convex (or concave) non-linear programming problem to a linear programming problem. Consider the following NLP problem:

$$\text{Optimize (Max or Min)} Z = \sum_{j=1}^n f_j(x_j)$$

subject to the constraints

$$\sum_{j=1}^n a_{ij}x_j = b_i; \quad i = 1, 2, \dots, m$$

$$\text{and } x_j \geq 0 \quad \text{for all } j$$

where $f_j(x_j)$ is the j th separable function to be approximated over a defined interval.

Define (a_k, b_k) for all $k = 1, 2, \dots, K$ as the k th breaking point joining a linear segment, which approximate the non-linear function $f(x)$, as shown in Fig.6.1. Further define W_k as the non-negative weight associated with the k th breaking point such that $\sum_{k=1}^K W_k = 1$.

Let us impose an additional condition (if necessary) so that all W_k and $W_k + 1$ are equated with zero to determine the weighted average of breaking points. This means that W_k and $W_k + 1$ will represent the weighted average of breaking point (a_k, b_k) and (a_{k+1}, b_{k+1}) respectively. Thus $f(x)$

is approximated as follows:

$$f(x) = \sum_{k=1}^K b_k W_k; \quad x = \sum_{k=1}^K a_k W_k$$

This approximation is valid, provided the following conditions hold good:

$$\begin{aligned} 0 &\leq W_1 \leq y_1 \\ 0 &\leq W_2 \leq y_1 + y_2 \\ 0 &\leq W_3 \leq y_2 + y_3 \\ &\dots \\ 0 &\leq W_{k-1} \leq y_{k-2} + y_{k-1} \\ 0 &\leq W_k \leq y_{k-1} \\ \text{and } \sum_{k=1}^K W_k &= 1; \quad \sum_{k=1}^{K-1} y_k = 1 \\ y_k &= 0 \text{ or } 1 \text{ for all } k. \end{aligned}$$

The variables for approximation are now W_k and y_k . The last constraints implies that if $y_k = 1$, then all other $y_k = 0$. Consequently immediately preceding constraints ensure that $0 \leq W_k \leq y_k = 1$ and $0 \leq W_{k+1} \leq y_k = 1$. This means all other constraints should give $W_k \leq 0$.

Mixed-Integer Approximation of Separable NLP Problem

The single-variable non-linear separable function $f(x)$, as defined earlier, can also be approximated by a piece-wise linear function, using mixed-integer programming. Let the number of breaking points for the j th variable, x_j , be equal to K_j and g_{jk} be its k th breaking value. Also, let w_{jk} be the weight associated with the k th breaking point of j th variable, x_j . Then the equivalent mixed integer programming problem is stated as:

$$\text{Optimize (Max or Min)} Z = \sum_{j=1}^n \sum_{k=1}^{K_j} f_j(a_{jk}) w_{jk}$$

subject to the constraints

$$\sum_{j=1}^n \sum_{k=1}^{K_j} g_{ij}(a_{jk}) w_{jk} \leq b_i; \quad i = 1, 2, \dots, m$$

$$0 \leq w_{j1} \leq y_{j1}$$

.....

$$0 \leq w_{jk} \leq y_{jk-1} + y_{jk}; \quad k = 2, 3, \dots, K_j - 1$$

$$\text{and } \sum_{k=1}^{K_j} w_{jk} = 1; \quad \sum_{k=1}^{K_j-1} y_{jk}$$

$$y_{jk} = 0 \text{ or } 1 \text{ for all } j \text{ and } k.$$

The approximation is valid only under following two conditions:

- (i) For each j , no more than two w_{jk} should appear in the basis. That is, no more than two w_{jk} are positive for each j .
- (ii) Two w_{jk} can be positive only if they are adjacent.

The simplex method can now be used to solve the above stated problem, along with additional constraints involving y_{jk} variables. Thus, the optimality criterion of the simplex method can be used to select the *entering variable* w_{jk} into the basis, only if it satisfies the above two conditions. Otherwise the variable w_{jk} , having the next best optimality indicator ($c_{jk} - z_{jk}$), is considered for entering the basis. The process is repeated until the optimality criterion is satisfied or until it is impossible to introduce a new w_{jk} without violating the restricted basis condition, whichever occurs first. The last simplex table provides the approximate optimal solution to the given problem.

Remarks

1. It is important to note that the *restricted basis method* yields only a local optimum, whereas mixed integer programming method guarantees a global optimum to the approximate problem.
2. The approximate solution obtained by using any of the two methods may not be feasible for the original NLP problem.
3. The solution space of approximate problem may have additional extreme points that do not exist in the solution space of the original problem. However, this depends on the degree of accuracy while obtaining linear approximation.

The Procedure

Step 1: Convert minimization objective function of the given NLP problem into that of maximization, with the usual method as discussed earlier.

Step 2: Examine whether the functions $f_j(x_j)$ and $g_{ij}(x_j)$ satisfy the concavity (convexity) conditions required for the maximization of NLP problem. If yes, then go to Step 3. Otherwise stop where $f(x)$ is to be approximated.

Step 3: Divide the interval $0 \leq x_j \leq t_j$ ($j = 1, 2, \dots, n$) into a number of breaking points a_{jk} ($k = 1, 2, \dots, K_j$) such that $a_{j1} = 0, a_{j1} < a_{j2} < \dots < a_{jK_j}$.

Step 4: For each point a_{jk} , compute piece-wise linear approximation $f_j(x_j)$ and $g_{ij}(x_j)$, for all i and j .

Step 5: Write down piece-wise linear approximation of the given NLP problem obtained from Step 4.

Step 6: Solve the resulting LP problem using *two-phase simplex method* treating w_{i1} ($i = 1, 2, \dots, m$) as artificial variables. The coefficients associated with these variables are assumed to be zero. This assumption yields optimal simplex table of Phase I and hence would be considered as the initial simplex table for Phase II.

Step 7: Obtain optimum solution of the original NLP problem by using the relations:

$$x_j^* = \sum_{k=1}^{K_j} a_{jk} w_{jk}; \quad j = 1, 2, \dots, n.$$

Example : Solve the following non-linear programming problem using separable programming algorithm.

$$\begin{aligned} \text{Max } Z &= 3x_1 + 2x_2 \\ \text{subject to the constraints} \\ g(x) &= 4x_1^2 + x_2^2 \leq 16 \\ \text{and } x_1, x_2 &\geq 0 \end{aligned}$$

Solution : The objective function is already of maximization form. Consider the separable functions:

$$\begin{aligned} f_1(x_1) &= 3x_1; & f_2(x_2) &= 2x_2 \\ g_{11}(x_1) &= 4x_1^2; & g_{12}(x_2) &= x_2^2 \end{aligned}$$

Since $f_1(x_1)$ and $f_2(x_2)$ are in linear form, can leave them in their present form. Further, it may be observed that these functions satisfy the concavity (convexity) conditions.

By inspection, constraints of the problem suggest the values of variables as: $x_1 \leq 2$ and $x_2 \leq 4$. Therefore, we take $t_1 = 4$ and $t_2 = 4$ as the upper limits for the variables x_1 and x_2 respectively. Thus, we divide the closed interval $[0, 4]$ into four subintervals of equal size for both x_1 and x_2 . It is important to note that the number of subintervals for x_1 and x_2 should be the same, but they need not be equal in size.

To obtain the approximate LP problem for the given NLP problem, divide the interval $0 \leq x_j \leq 4$ into five breaking points a_{jk} ($j = 1, 2; k = 1, 2, 3, 4, 5$) such that:

$$a_{j1} = 0, \quad a_{j1} \leq a_{j2} \leq a_{j3} \leq a_{j4} \leq a_{j5} = 4$$

For each point a_{jk} , compute the piece-wise linear approximation for each of $f_j(x_j)$ and $g_{1j}(x_j)$; $j = 1, 2$ as follows:

k	a_{jk}	$f_1(x_1 = a_{jk})$	$f_2(x_2 = a_{jk})$	$g_{11}(x_1 = a_{jk})$	$g_{12}(x_2 = a_{jk})$
1	0	0	0	0	0
2	1	3	2	4	1
3	2	6	4	16	4
4	3	9	6	36	9
5	4	12	8	64	16

Using this data, we have the following piece-wise linear approximation:

$$f_1(x_1) = 0 w_{11} + 3 w_{12} + 6 w_{13} + 9 w_{14} + 12 w_{15}$$

$$f_2(x_2) = 0 w_{21} + 2 w_{22} + 4 w_{23} + 6 w_{24} + 8 w_{25}$$

$$g_{11}(x_1) = 0 w_{11} + 4 w_{12} + 16 w_{13} + 36 w_{14} + 64 w_{15}$$

$$g_{12}(x_2) = 0 w_{21} + 1 w_{22} + 4 w_{23} + 9 w_{24} + 16 w_{25}$$

Using the data of Step 4, the approximating LP problem can now be stated as follows:

$$\text{Max } f(x) = (0 w_{11} + 3 w_{12} + 6 w_{13} + 9 w_{14} + 12 w_{15}) + (0 w_{21} + 2 w_{22} + 4 w_{23} + 6 w_{24} + 8 w_{25})$$

subject to the constraints

$$(0 w_{11} + 4 w_{12} + 16 w_{13} + 36 w_{14} + 64 w_{15}) + (0 w_{21} + w_{22} + 4 w_{23} + 9 w_{24} + 16 w_{25}) \leq 16$$

$$w_{11} + w_{12} + w_{13} + w_{14} + w_{15} = 1$$

$$w_{21} + w_{22} + w_{23} + w_{24} + w_{25} = 1$$

and $w_{jk} \geq 0$ for j and k

with the additional two restricted basis conditions: (i) for each j no more than two w_{jk} are positive, and (ii) if two w_{jk} are positive, they must correspond to adjacent points. To solve the LP problem formulated in Step 5 by using simplex method, introduce the slack variables $s_i (\geq 0)$ to convert inequality constraint into equation. For using *Phase II* of simplex method to solve the given LP problem, treat w_{11} and w_{21} as artificial variables whose coefficients in the objective function of reduced LP problem are zero. The initial simplex table for Phase II is shown in Table 6.1.

$c_j \rightarrow$			3	6	9	12	2	4	6	8	0	0	0
c_B	Basic Variables (B)	Solution Values $b(=x_B)$	w_{12}	w_{13}	w_{14}	w_{15}	w_{22}	w_{23}	w_{24}	w_{25}	s_1	w_{11}	w_{21}
0	s_1	16	4	16	36	64	1	4	9	16	1	0	0
0	w_{11}	1	1	1	1	1	0	0	0	0	0	1	0
0	w_{21}	1	0	0	0	0	1	1	1	①	0	0	1→
$f(x) = 0$	$c_j - z_j$		3	6	9	12	2	4	6	8 ↑	—	—	—

Table 6.1

In Table 6.1, the entries in $c_j - z_j$ row indicate that variable w_{15} must enter into the new solution and variable s_1 should leave the current solution. But this does not satisfy the additional Conditions (i) and (ii). The next best variable to enter the basis is therefore w_{25} and variable w_{21} leaves the basis. The new solution, so obtained, is shown in Table 6.2. In Table 6.2 out of eligible variables $w_{12}, w_{13}, w_{14}, w_{15}$ to enter the basis, we decide to enter variable w_{12} into the basis in view of the

additional conditions. The new solution after introducing variable w_{12} into the basis and dropping variable s_1 from the basis is shown in Table 6.3.

$c_j \rightarrow$			3	6	9	12	2	4	6	8	0	0
c_B	Basic Variables (B)	Solution Values $b(=x_B)$	w_{12}	w_{13}	w_{14}	w_{15}	w_{22}	w_{23}	w_{24}	w_{25}	s_1	w_{11}
0	s_1	0	④	16	36	64	-15	-12	-7	0	1	0 $\rightarrow$
0	w_{11}	1	1	1	1	1	0	0	0	0	0	1
8	w_{25}	1	0	0	0	0	1	1	1	1	0	0
$f(x) = 8$	$c_j - z_j$		3 ↑	6	9	12	6	4	2	—	—	—

Table 6.2

$c_j \rightarrow$			3	6	9	12	2	4	6	8	0	0
c_B	Basic Variables (B)	Solution Values $b(=x_B)$	w_{12}	w_{13}	w_{14}	w_{15}	w_{22}	w_{23}	w_{24}	w_{25}	s_1	w_{11}
3	w_{12}	0	1	4	9	16	-15/4	-3	-7/4	0	1/4	0
0	w_{11}	1	0	-3	-8	-15	15/4	3	⑦/4	0	-1/4	1 $\rightarrow$
8	w_{25}	1	0	0	0	0	1	1	1	1	0	0
$f(x) = 8$	$c_j - z_j$		—	-6	-18	-36	21/4	5 ↑	13/4	—	-3/4	—

Table 6.3

To get the next best solution, we need to introduce the variable w_{24} into the basis and drop variable w_{11} from the basis in the solution, as shown in Table 6.4. The new solution is shown in Table 6.4.

$c_j \rightarrow$			3	6	9	12	2	4	6	8	0
c_B	Basic Variables (B)	Solution Values $b(=x_B)$	w_{12}	w_{13}	w_{14}	w_{15}	w_{22}	w_{23}	w_{24}	w_{25}	s_1
3	w_{12}	1	1	1	1	1	0	0	0	0	0
6	w_{24}	4/7	0	-12/7	-32/7	-60/7	15/7	12/7	1	0	-1/7
8	w_{25}	3/7	0	12/7	32/7	60/7	-8/7	-5/7	0	1	1/7
$f(x) = 69/7$	$c_j - z_j$		—	-3/7	-22/7	-67/7	-12/7	-4/7	—	—	-9/14

Table 6.4

Since all $c_j - z_j \leq 0$, the optimal solution has been arrived at. The optimal solution shown in Table 6.4 is:

$$w_{12} = 1, w_{24} = 4/7, w_{25} = 3/7 \quad \text{and} \quad f(x) = 69/7$$

The optimal solution to the original NLP problem can be obtained by using the formula:

$$x_j = \sum_{k=1}^5 a_{jk} w_{jk}; \quad j = 1, 2$$

This gives

$$\begin{aligned} x_1 &= a_{11}w_{11} + a_{12}w_{12} + a_{13}w_{13} + a_{14}w_{14} + a_{15}w_{15} \\ &= 0(0) + 1(1) + 2(0) + 3(0) + 4(0) = 1 \\ x_1 &= a_{21}w_{21} + a_{22}w_{22} + a_{23}w_{23} + a_{24}w_{24} + a_{25}w_{25} \\ &= 0(0) + 1(0) + 2(0) + 3(4/7) + 4(3/7) = 24/7 \end{aligned}$$

Hence, the optimal solution to the given NLP problem is:

$$x_1 = 1, \quad x_2 = 24/7 \quad \text{and} \quad \text{Max } f(x) = 3 + 2(24/7) = 69/7.$$

Exercise : Use separable programming algorithm to solve the non-linear programming problem:

$$\begin{aligned} \text{Max } Z &= x_1 + x_2^2 \\ \text{subject to the constraints} \\ 3x_1 + 2x_2^2 &\leq 9 \\ x_1, x_2 &\geq 0. \end{aligned}$$

Reference

1. Sharma J.K. Operations Research:Theory and Applications.

**POST GRADUATE DEGREE PROGRAMME (CBCS)
IN
MATHEMATICS**

**SEMESTER II
SELF LEARNING MATERIAL**

**PAPER - MATA 2.4
(Applied Stream)**

**Mechanics of Fluids
Stochastic Processes**



**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, 2019

Directorate of Open and Distance Learning, University of Kalyani
Published by the Directorate of Open and Distance Learning,
University of Kalyani, Kalyani-741235, West Bengal

All rights reserved. No part of this work should be reproduced in any form without the permission in writing from the Directorate of Open and Distance Learning, University of Kalyani.

Director's Note

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

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

Self Learning Materials (SLMs) have been published by the Directorate of Open and Distance Learning, University of Kalyani, Kalyani-741235, West Bengal and all the copyright reserved for University of Kalyani. No part of this work should be reproduced in any form without permission in writing from the appropriate authority of the University of Kalyani.

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

Director
Directorate of Open and Distance Learning
University of Kalyani

CONTENTS

Sl. No.	Block	Unit	Page
1	Fluid Mechanics	I & II	3 – 48
		III & IV	49 – 79
		V & VI	80 – 101
		VII	102 – 120
		VIII & IX	121 – 144
		X	145 – 153
2	Stochastic Processes	XI, XII, XIII	155 – 169
		XIV	170 – 171
		XV	172 – 183
		XVI	184 – 189

Elective Paper

Applied Stream

MATA 2.4

Block - I

Marks : 60 (SSE : 50; IA : 10)

Mechanics of Fluids

Syllabus

• Unit 1 •

Kinematics : Real and ideal fluids. Streamlines and paths of particles. Steady and unsteady flows. Lagrange's and Euler's methods of description of fluid motion. Accelerations. Boundary surface. Irrotational and rotational motions.

• Unit 2 •

Equation of continuity. Equations of Motion : Lagrange's and Euler's equations of motion. Bernoulli's theorem. Cauchy's integrals. Impulsive action.

• Unit 3 •

Motion in Two Dimensions : Stream function. Sources, sinks and doublets. Images. Image of a source (sink) with regard to a plane and a sphere.

• Unit 4 •

Image of a doublet with regard to a sphere, Images in two dimensions. Milne-Thomson circle theorem. Blasius theorem.

• Unit 5 •

General Theory of Irrotational Motion : Flow and circulation. Cyclic and acyclic motions. Impulsive motion. Properties of irrotational motion. Kelvin's theorem of minimum kinetic energy.

● Unit 6 ●

Motion of a sphere. Liquid streaming past a fixed sphere. Equations of motion of a sphere.

● Unit 7 ●

Vortex Motion : Vortex motion and its simple properties. Motion due to circular and rectilinear vortices. Vortex pair and doublet. Karman vortex street.

● Unit 8 ●

Viscous Liquid Motion : Stress components in real fluid. Rate of strain quadric. Stress analysis in fluid motion.

● Unit 9 ●

Relation between stress and rate of strain. Navier-Stokes' equations.

● Unit 10 ●

Plane Poiseuille and Couette flow between two parallel plates.

UNIT-I & II

Basic Concepts and Definitions

(i) Let $\vec{q} = \hat{i}u + \hat{j}v + \hat{k}w$, then

$$|\vec{q}| = \sqrt{u^2 + v^2 + w^2} = q$$

D.C's are given by $l = \cos \alpha = \frac{u}{|\vec{q}|}$, $m = \cos \beta = \frac{v}{|\vec{q}|}$, $n = \cos \gamma = \frac{w}{|\vec{q}|}$

where l, m, n , are components of a unit vector i.e. $l^2 + m^2 + n^2 = 1$

(ii) $\vec{a} \cdot \vec{b} = ab \cos \theta$, $\vec{a} \times \vec{b} = ab \sin \theta \hat{n}$

(iii) $\nabla \phi = \hat{i} \frac{\partial \phi}{\partial x} + \hat{j} \frac{\partial \phi}{\partial y} + \hat{k} \frac{\partial \phi}{\partial z}$, where ϕ is a scalar and

$\nabla \equiv \hat{i} \frac{\partial}{\partial x} + \hat{j} \frac{\partial}{\partial y} + \hat{k} \frac{\partial}{\partial z}$ is a vector (operator)

(iv) $\text{div } \vec{q} = \nabla \cdot \vec{q} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}$, $\vec{q} = (u, v, w)$

If $\nabla \cdot \vec{q} = 0$, then $\vec{q}$ is said to be solenoidal vector.

(v) $d\vec{r} = \hat{i}dx + \hat{j}dy + \hat{k}dz$, $d\phi = \frac{\partial \phi}{\partial x} dx + \frac{\partial \phi}{\partial y} dy + \frac{\partial \phi}{\partial z} dz$

and

$$\nabla \phi = \hat{i} \frac{\partial \phi}{\partial x} + \hat{j} \frac{\partial \phi}{\partial y} + \hat{k} \frac{\partial \phi}{\partial z},$$

Therefore,

$$d\phi = (\nabla \phi) \cdot d\vec{r}$$

(vi) $\text{Curl } \vec{q} = \nabla \times \vec{q} = \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ u & v & w \end{vmatrix}$

$$= \hat{i} \left(\frac{\partial w}{\partial y} - \frac{\partial v}{\partial z} \right) + \hat{j} \left(\frac{\partial u}{\partial z} - \frac{\partial w}{\partial x} \right) + \hat{k} \left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \right)$$

(vii) (a) Gradient of a scalar is a vector.

(b) Divergence of a scalar and curl of a scalar are meaningless.

(c) Divergence of a vector is a scalar and curl of a vector is a vector.

$$(viii) \quad \nabla \cdot \nabla \phi = \nabla^2 \phi = \frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2}$$

where ∇^2 is Laplacian operator.

$$(ix) \quad \text{Curl grad } \phi = 0, \text{ div curl } \bar{q} = 0$$

$$(x) \quad \text{Curl curl } \bar{q} = \text{grad div } \bar{q} - \nabla^2 \bar{q}$$

i.e. $\nabla^2 \bar{q} = \text{grad div } \bar{q} - \text{curl curl } \bar{q}$

(xi) **Gauss's divergence theorem**

$$(a) \quad \int_S \bar{q} \cdot d\bar{S} = \int_V \text{div } \bar{q} \, dv$$

$$(b) \quad \int_S \hat{n} \times \bar{q} \, dS = \int_V \text{curl } \bar{q} \, dv$$

(xii) **Green's theorem**

$$(a) \quad \int_V \nabla \phi \cdot \nabla \psi \, dV = \int_S \phi \nabla \psi \cdot d\bar{S} - \int_V \phi \nabla^2 \psi \, dV$$

$$= \int_S \psi \nabla \phi \cdot d\bar{S} - \int_V \psi \nabla^2 \phi \, dV$$

$$(b) \quad \int_V (\phi \nabla^2 \psi - \psi \nabla^2 \phi) \, dV = \int_V \left(\phi \frac{\partial \psi}{\partial n} - \psi \frac{\partial \phi}{\partial n} \right) dS$$

$$(xiii) \quad \text{Stoke's theorem} \quad \int_C \bar{q} \cdot d\bar{r} = \int_S \text{curl } \bar{q} \cdot d\bar{S} = \int_S \text{curl } \bar{q} \cdot \hat{n} \, dS$$

(xiv) **Orthogonal curvilinear co-ordinates :**

Let there be three orthogonal families of surfaces

$$f_1(x, y, z) = \alpha, f_2(x, y, z) = \beta, f_3(x, y, z) = \gamma \quad (1)$$

where x, y, z are Cartesian co-ordinates of a point $P(x, y, z)$ in space. The surfaces

$$\alpha = \text{constant}, \beta = \text{constant}, \gamma = \text{constant} \quad (2)$$

form an orthogonal system in which every pair of surfaces is an orthogonal system. The values α, β, γ are called orthogonal curvilinear co-ordinates.

From three equations in (1), we can get

$$x = x(\alpha, \beta, \gamma), y = y(\alpha, \beta, \gamma), z = z(\alpha, \beta, \gamma)$$

The surfaces (2) are called co-ordinate surfaces.

Let $\bar{r}$ be the position vector of the point $P(x, y, z)$

$$\text{i.e. } \bar{r} = x\hat{i} + y\hat{j} + z\hat{k} = \bar{r}(\alpha, \beta, \gamma)$$

A tangent vector to the α -curve ($\beta = \text{constant}$, $\gamma = \text{constant}$) at P is $\frac{\partial \bar{r}}{\partial \alpha}$. A unit tangent vector is

$$\hat{e}_1 = \frac{\partial \bar{r} / \partial \alpha}{\left| \partial \bar{r} / \partial \alpha \right|}$$

or
$$\frac{\partial \bar{r}}{\partial \alpha} = h_1 \hat{e}_1$$

where
$$h_1 = \left| \frac{\partial \bar{r}}{\partial \alpha} \right| = \sqrt{\left(\frac{\partial x}{\partial \alpha} \right)^2 + \left(\frac{\partial y}{\partial \alpha} \right)^2 + \left(\frac{\partial z}{\partial \alpha} \right)^2}$$

Similarly, $\hat{e}_2, \hat{e}_3$ are unit vectors along β -curve and γ -curve respectively such that

$$\frac{\partial \bar{r}}{\partial \beta} = h_2 \hat{e}_2, \quad \frac{\partial \bar{r}}{\partial \gamma} = h_3 \hat{e}_3$$

Further,
$$\begin{aligned} d\bar{r} &= \frac{\partial \bar{r}}{\partial \alpha} d\alpha + \frac{\partial \bar{r}}{\partial \beta} d\beta + \frac{\partial \bar{r}}{\partial \gamma} d\gamma \\ &= h_1 d\alpha \hat{e}_1 + h_2 d\beta \hat{e}_2 + h_3 d\gamma \hat{e}_3 \end{aligned}$$

Therefore,

$$(ds)^2 = d\bar{r} \cdot d\bar{r} = h_1^2 d\alpha^2 + h_2^2 d\beta^2 + h_3^2 d\gamma^2$$

where $h_1 d\alpha$, $h_2 d\beta$, $h_3 d\gamma$ are arc lengths along α , β and γ curves.

In orthogonal curvilinear co-ordinates, we have the following results.

(i)
$$\text{grad } \phi = \left(\frac{1}{h_1} \frac{\partial \phi}{\partial \alpha}, \frac{1}{h_2} \frac{\partial \phi}{\partial \beta}, \frac{1}{h_3} \frac{\partial \phi}{\partial \gamma} \right)$$

(ii) If $\bar{q} = (q_1, q_2, q_3)$, then

$$\text{div } \bar{q} = \frac{1}{h_1 h_2 h_3} \left[\frac{\partial}{\partial \alpha} (h_2 h_3 q_1) + \frac{\partial}{\partial \beta} (h_3 h_1 q_2) + \frac{\partial}{\partial \gamma} (h_1 h_2 q_3) \right]$$

(iii) If $\text{curl } \bar{q} = \bar{\xi} = (\xi_1, \xi_2, \xi_3)$, then

$$\xi_1 = \frac{1}{h_2 h_3} \left[\frac{\partial}{\partial \beta} (h_3 q_3) - \frac{\partial}{\partial \gamma} (h_2 q_2) \right]$$

$$\xi_2 = \frac{1}{h_3 h_1} \left[\frac{\partial}{\partial \gamma} (h_1 q_1) - \frac{\partial}{\partial \alpha} (h_3 q_3) \right]$$

$$\xi_3 = \frac{1}{h_1 h_2} \left[\frac{\partial}{\partial \alpha} (h_2 q_2) - \frac{\partial}{\partial \beta} (h_1 q_1) \right]$$

$$(iv) \nabla^2 \phi = \frac{1}{h_1 h_2 h_3} \left[\frac{\partial}{\partial \alpha} \left(\frac{h_2 h_3}{h_1} \frac{\partial \phi}{\partial \alpha} \right) + \frac{\partial}{\partial \beta} \left(\frac{h_3 h_1}{h_2} \frac{\partial \phi}{\partial \beta} \right) + \frac{\partial}{\partial \gamma} \left(\frac{h_1 h_2}{h_3} \frac{\partial \phi}{\partial \gamma} \right) \right].$$

The Cartesian co-ordinate system (x, y, z) is the simplest of all orthogonal co-ordinate systems. In many problems involving vector field theory, it is convenient to work with other two most common orthogonal co-ordinates i.e. cylindrical polar co-ordinates and spherical polar co-ordinates denoted respectively by (r, θ , z) and (r, θ , ψ). For cylindrical co-ordinates, $h_1 = 1$, $h_2 = r$, $h_3 = 1$. For spherical co-ordinates, $h_1 = 1$, $h_2 = r$, $h_3 = r \sin \theta$.

1. Fluid Dynamics

Fluid dynamics is the science treating the study of fluids in motion. By the term fluid, we mean a substance that flows i.e. which is not a solid. Fluids may be divided into two categories

(i) liquids which are incompressible i.e. their volumes do not change when the pressure changes

(ii) gases which are compressible i.e. they undergo change in volume whenever the pressure changes. The term hydrodynamics is often applied to the science of moving incompressible fluids. However, there is no sharp distinctions between the three states of matter i.e. solid, liquid and gases.

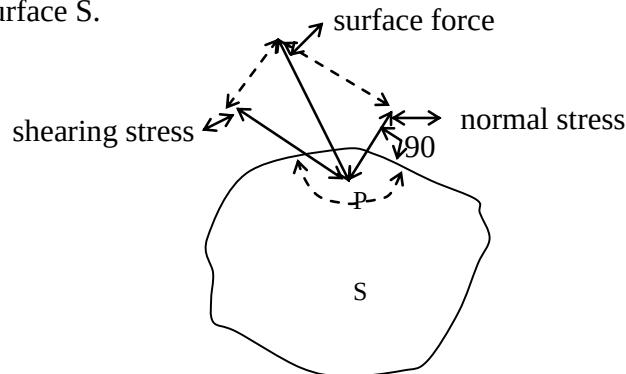
In **microscopic** view of fluids, matter is assumed to be composed of molecules which are in random relative motion under the action of intermolecular forces. In solids, spacing of the molecules is small, spacing persists even under strong molecular forces. In liquids, the spacing between molecules is greater even under weaker molecular forces and in gases, the gaps are even larger.

If we imagine that our microscope, with which we have observed the molecular structure of matter, has a variable focal length, we could change our observation of matter from the fine detailed microscopic viewpoint to a longer range **macroscopic** viewpoint in which we would not see the gaps between the molecules and the matter would appear to be continuously distributed. We shall take this macroscopic view of fluids in which physical quantities associated with the fluids within a given volume V are assumed to be distributed continuously and, within a sufficiently small volume δV , uniformly. This observation is known as **Continuum hypothesis**. It implies that at each point of a fluid, we can prescribe a unique velocity, a unique pressure, a unique density etc. Moreover, for a continuous or ideal fluid we can define a **fluid particle** as the fluid contained within an infinitesimal volume whose size is so small that it may be regarded as a geometrical point.

1.1. Stresses : Two types of forces act on a fluid element. One of them is **body force** and other is **surface force**. The body force is proportional to the

mass of the body on which it acts while the surface force is proportional to the surface area and acts on the boundary of the body.

Suppose $\vec{F}$ is the surface force acting on an elementary surface area dS at a point P of the surface S .



Let F_1 and F_2 be resolved parts of $\vec{F}$ in the directions of tangent and normal at P . The normal force per unit area is called the **normal stress** and is also called **pressure**. The tangential force per unit area is called the **shearing stress**.

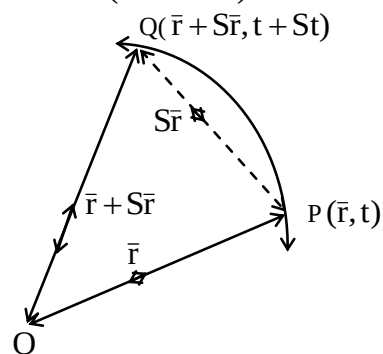
1.2. Viscosity : It is the internal friction between the particles of the fluid which offers resistance to the deformation of the fluid. The friction is in the form of tangential and shearing forces (stresses). Fluids with such property are called **viscous** or **real** fluids and those not having this property are called **inviscid** or **ideal** or **perfect fluids**.

Actually, all fluids are real, but in many cases, when the rates of variation of fluid velocity with distances are small, viscous effects may be ignored.

From the definition of body force and shearing stress, it is clear that body force per unit area at every point of surface of an ideal fluid acts along the normal to the surface at that point. Thus ideal fluid does not exert any shearing stress.

Thus, we conclude that viscosity of a fluid is that property by virtue of which it is able to offer resistance to shearing stress. It is a kind of molecular frictional resistance.

1.3. Velocity of Fluid at a Point : Suppose that at time t , a fluid particle is at the point P having position vector $\vec{r}$ (i.e. $\vec{OP} = \vec{r}$)



and at time $t + \delta t$ the same particle has reached at point Q having position vector $\vec{r} + \delta\vec{r}$. The particle velocity $\vec{q}$ at point P is

$$\vec{q} = \lim_{\delta t \rightarrow 0} \frac{(\vec{r} + \delta\vec{r}) - \vec{r}}{\delta t} = \lim_{\delta t \rightarrow 0} \frac{\delta\vec{r}}{\delta t} = \frac{d\vec{r}}{dt}$$

where the limit is assumed to exist uniquely. Clearly $\vec{q}$ is in general dependent on both $\vec{r}$ and t , so we may write

$$\vec{q} = \vec{q}(\vec{r}, t) = \vec{q}(x, y, z, t),$$

$$\vec{r} = x\hat{i} + y\hat{j} + z\hat{k} \text{ (P has co-ordinates (x, y, z))}$$

Suppose,

$$\vec{q} = u\hat{i} + v\hat{j} + w\hat{k}$$

and since

$$\vec{q} = \frac{d\vec{r}}{dt} = \frac{dx}{dt}\hat{i} + \frac{dy}{dt}\hat{j} + \frac{dz}{dt}\hat{k},$$

therefore

$$u = \frac{dx}{dt}, \quad v = \frac{dy}{dt}, \quad w = \frac{dz}{dt}.$$

1.4. Remarks. (i) A point where $\vec{q} = \vec{0}$, is called a stagnation point.

(ii) When the flow is such that the velocity at each point is independent of time i.e. the flow pattern is same at each instant, then the motion is termed as steady motion, otherwise it is unsteady.

1.5. Flux across any surface : The flux i.e. the rate of flow across any surface S is defined by the integral

$$\int_S \rho(\vec{q} \cdot \hat{n}) dS$$

where ρ is the density, $\vec{q}$ is the velocity of the fluid and $\hat{n}$ is the outward unit normal at any point of S.

Also, we define

$$\text{Flux} = \text{density} \times \text{normal velocity} \times \text{area of the surface.}$$

2. Eulerian and Lagrangian Methods (Local and Total range of change)

We have two methods for studying the general problem of fluid dynamics.

2.1. Eulerian Method : In this method, we fix a point in the space occupied by the fluid and observation is made of whatever changes of velocity, density pressure etc take place at that point. i.e. point is fixed and fluid particles are allowed to pass through it. If P(x, y, z) is the point under reference, then x, y, z do not depend upon the time parameter t, therefore $\dot{x}, \dot{y}, \dot{z}$ do not exist (dot denotes derivative w.r.t. time t).

Let $f(x, y, z, t)$ be a scalar function associated with some property of the fluid (e.g. its density) i.e. $f(x, y, z, t) = f(\bar{r}, t)$, where $\bar{r} = x\hat{i} + y\hat{j} + z\hat{k}$ is the position vector of the point P, then

$$\frac{\partial f}{\partial t} = \lim_{\delta t \rightarrow 0} \frac{f(\bar{r}, t + \delta t) - f(\bar{r}, t)}{\delta t} \quad (1)$$

Here, $\frac{\partial f}{\partial t}$ is called local time rate of change.

2.2. Lagrangian Method :- In this case, observations are made at each point and each instant, i.e., any particle of the fluid is selected and observation is made of its particular motion and it is pursued throughout its course.

Let a fluid particle be initially at the point (a, b, c). After lapse of time t, let the same fluid particle be at (x, y, z). It is obvious that x, y, z are functions of t. But since the particles which have initially different positions occupy different positions after the motion is allowed. Hence the co-ordinates of the final position i.e. (x, y, z) depend on (a, b, c) also. Thus

$$x = f_1(a, b, c, t), y = f_2(a, b, c, t), z = f_3(a, b, c, t).$$

For this case, if $f(x, y, z, t)$ be scalar function associated with the fluid, then

$$\frac{df}{dt} = \lim_{\delta t \rightarrow 0} \frac{f(\bar{r} + \delta\bar{r}, t + \delta t) - f(\bar{r}, t)}{\delta t} \quad (2)$$

where $\dot{x}, \dot{y}, \dot{z}$ exist.

Here $\frac{df}{dt}$ is called an individual time rate or total rate or particle rate of change.

Now, we establish the relation between these two time rates (1) & (2).

We have

$$f = f(x, y, z, t)$$

Therefore,

$$\begin{aligned} \frac{df}{dt} &= \frac{\partial f}{\partial x} \frac{dx}{dt} + \frac{\partial f}{\partial y} \frac{dy}{dt} + \frac{\partial f}{\partial z} \frac{dz}{dt} + \frac{\partial f}{\partial t} \\ &= \left(\frac{\partial f}{\partial x} \hat{i} + \frac{\partial f}{\partial y} \hat{j} + \frac{\partial f}{\partial z} \hat{k} \right) \cdot \left(\frac{dx}{dt} \hat{i} + \frac{dy}{dt} \hat{j} + \frac{dz}{dt} \hat{k} \right) + \frac{\partial f}{\partial t} \\ &= \nabla f \cdot \bar{q} + \frac{\partial f}{\partial t} \end{aligned}$$

where

$$\bar{q} = \frac{dx}{dt} \hat{i} + \frac{dy}{dt} \hat{j} + \frac{dz}{dt} \hat{k} = (u, v, w)$$

Thus

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + \bar{q} \cdot \nabla f \quad (3)$$

2.3. Remarks. (i) The relation

$$\begin{aligned} \frac{df}{dt} &= \frac{\partial f}{\partial t} + \bar{q} \cdot \nabla f \\ \Rightarrow \frac{df}{dt} &= \left(\frac{\partial}{\partial t} + \bar{q} \cdot \nabla \right) f \\ \Rightarrow \frac{d}{dt} &\equiv \frac{\partial}{\partial t} + \bar{q} \cdot \nabla \end{aligned}$$

The operator $\frac{d}{dt}$ (also denoted by $\frac{D}{Dt}$) is called Lagrangian operator or material derivative i.e. time rate of change in Lagrangian view. Sometimes, it is called ‘differentiation following the fluid’.

(ii) Similarly, for a vector function $\bar{F}(x, y, z, t)$ associated with some property of the fluid (e.g. its velocity, acceleration), we can show that

$$\frac{d\bar{F}}{dt} = \frac{\partial \bar{F}}{\partial t} + \bar{q} \cdot \nabla \bar{F}$$

Hence the relation (3) holds for both scalar and vector functions associated with the moving fluid.

(iii) The Eulerian method is sometimes also called the flux method.

(iv) Both Lagrangian and Eulerian methods were used by Euler for studying fluid dynamics.

(v) Lagrangian method resembles very much with the dynamics of a particle

(vi) The two methods are essentially equivalent, but depending upon the problem, one has to judge whether Lagrangian method is more useful or the Eulerian.

3. Streamlines, Pathlines and Streaklines

3.1. Streamlines : It is a curve drawn in the fluid such that the direction of the tangent to it at any point coincides with the direction of the fluid velocity vector $\bar{q}$ at that point. At any time t , let $\bar{q} = (u, v, w)$ be the velocity at each point $P(x, y, z)$ of the fluid. The direction ratios of the tangent to the curve at $P(x, y, z)$ are $d\bar{r} = (dx, dy, dz)$ since the tangent and the velocity at P have the same direction, therefore $\bar{q} \times d\bar{r} = \bar{0}$

$$\text{i.e. } (u\hat{i} + v\hat{j} + w\hat{k}) \times (dx\hat{i} + dy\hat{j} + dz\hat{k}) = \vec{0}$$

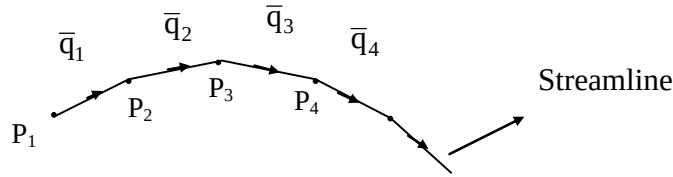
$$\text{i.e. } (vdy - wdy)\hat{i} + (wdx - udz)\hat{j} + (udy - vdx)\hat{k} = \vec{0}$$

$$\text{i.e. } vdz - wdy = 0 = wdx - udz = udy - vdx$$

$$\Rightarrow \frac{dx}{u} = \frac{dy}{v} = \frac{dz}{w}$$

These are the differential equations for the streamlines.

i.e. their solution gives the streamlines.



In the figure, if $\bar{q}_1, \bar{q}_2, \bar{q}_3, \dots$ denote the velocities at neighbouring points $P_1, P_2, P_3, \dots$, then the small straight line segments $P_1P_2, P_2P_3, P_3P_4, \dots$ collectively give the approximate form of the streamlines.

3.2. Pathlines: When the fluid motion is steady so that the pattern of flow does not vary with time, the paths of the fluid particles coincide with the streamlines. But in case of unsteady motion, the flow pattern varies with time and the paths of the particles do not coincide with the streamlines. However, the streamline through any point P does touch the pathline through P . Pathlines are the curves described by the fluid particles during their motion i.e. these are the paths of the particles.

The differential equations for pathlines are

$$\frac{d\vec{r}}{dt} = \bar{q} \text{ i.e. } \frac{dx}{dt} = u, \frac{dy}{dt} = v, \frac{dz}{dt} = w \quad (1)$$

where now (x, y, z) are the Cartesian co-ordinates of the fluid particle and not a fixed point of space. The equation of the pathline which passes through the point (x_0, y_0, z_0) , which is fixed in space, at time $t = 0$ say, is the solution of (1) which satisfy the initial condition that $x = x_0, y = y_0, z = z_0$ when $t = 0$. The solution gives a set of equations of the form

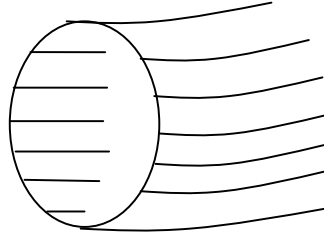
$$\left. \begin{aligned} x &= x(x_0, y_0, z_0, t) \\ y &= y(x_0, y_0, z_0, t) \\ z &= z(x_0, y_0, z_0, t) \end{aligned} \right\} \quad (2)$$

which, as t takes all values greater than zero, will trace out the required pathline.

3.3. Remarks : (i) Streamlines give the motion of each particle at a given instant whereas pathlines give the motion of a given particle at each instant.

We can make these observations by using a suspension of aluminium dust in the liquid.

- (ii) If we draw the streamlines through every point of a closed curve in the fluid, we obtain a **stream tube**. A stream tube of very small cross-section is called a **stream filament**.



- (iii) The components of velocity at right angle to the streamline is always zero. This shows that there is no flow across the streamlines. Thus, if we replace the boundary of stream tube by a rigid boundary, the flow is not affected. The principle of conservation of mass then gives that the flux across any cross-section of the stream tube should be the same.

3.4. Streaklines : In addition to streamlines and pathlines, it is useful for observational purpose to define a streakline. This is the curve of all fluid particles which at some time have coincided with a particular fixed point of space. Thus, a streakline is the locus of different particles passing through a fixed point. The streakline is observed when a neutrally buoyant marker fluid is continuously injected into the flow at a fixed point of space from time $\tau = -\infty$. The marker fluid may be smoke if the main flow involves a gas such as air, or a dye such as potassium permanganate (KMnO_4) if the main flow involves a liquid such as water.

If the co-ordinates of a particle of marker fluid are (x, y, z) at time t and the particle coincided with the injection point (x_0, y_0, z_0) at some time τ , where $\tau \leq t$, then the time-history (streakline) of this particle is obtained by solving the equations for a pathline, subject to the initial condition that $x = x_0, y = y_0, z = z_0$ at $t = \tau$. As τ takes all possible values in the range $-\infty \leq \tau \leq t$, the locations of all fluid particles on the streakline through (x_0, y_0, z_0) are obtained. Thus, the equation of the streakline at time t is given by

$$\left. \begin{aligned} x &= x(x_0, y_0, z_0, t, \tau) \\ y &= y(x_0, y_0, z_0, t, \tau) \\ z &= z(x_0, y_0, z_0, t, \tau) \end{aligned} \right\} (-\infty \leq \tau \leq t) \quad (2)$$

3.5. Remark: (i) For a steady flow, streaklines also coincide with streamlines and pathlines.

(ii) Streamlines, pathlines and streaklines are termed as **flowlines** for a fluid.

4. Velocity Potential

Suppose that $\bar{q} = u\hat{i} + v\hat{j} + w\hat{k}$ is the velocity at any time t at each point $P(x, y, z)$ of the fluid. Also suppose that the expression $u dx + v dy + w dz$ is an exact differential, say $-d\phi$.

Then, $-d\phi = u dx + v dy + w dz$

i.e. $-\left(\frac{\partial\phi}{\partial x}dx + \frac{\partial\phi}{\partial y}dy + \frac{\partial\phi}{\partial z}dz + \frac{\partial\phi}{\partial t}dt\right) = u dx + v dy + w dz$ where $\phi = \phi(x, y, z, t)$

is some scalar function, uniform throughout the entire field of flow.

Therefore,

$$u = -\frac{\partial\phi}{\partial x}, v = -\frac{\partial\phi}{\partial y}, w = -\frac{\partial\phi}{\partial z}, \frac{\partial\phi}{\partial t} = 0$$

But

$$\frac{\partial\phi}{\partial t} = 0 \Rightarrow \phi = \phi(x, y, z)$$

Hence

$$\bar{q} = u\hat{i} + v\hat{j} + w\hat{k} = -\left(\frac{\partial\phi}{\partial x}\hat{i} + \frac{\partial\phi}{\partial y}\hat{j} + \frac{\partial\phi}{\partial z}\hat{k}\right) = -\nabla\phi$$

where ϕ is termed as the **velocity potential** and the flow of such type is called flow of potential kind.

In the above definition, the negative sign in $\bar{q} = -\nabla\phi$ is a convention and it ensures that flow takes place from higher to lower potentials. The level surfaces $\phi(x, y, z, t) = \text{constant}$, are called **equipotentials** or **equipotential surfaces**.

4.1. Theorem : At all points of the field of flow the equipotentials (i.e. equipotential surfaces) are cut orthogonally by the streamlines.

Proof. If the fluid velocity at any time t be $\bar{q} = (u, v, w)$, then the equations of streamlines are

$$\frac{dx}{u} = \frac{dy}{v} = \frac{dz}{w} \quad (1)$$

The surfaces given by

$$\bar{q} \cdot d\bar{r} = 0 \text{ i.e. } udx + vdy + wdz = 0 \quad (2)$$

are such that the velocity is at right angles to the tangent planes. The curves (1) and the surfaces (2) cut each other orthogonally. Suppose that the expression on the left hand side of (2) is an exact differential, say, $-d\phi$, then

$$d\phi = udx + vdy + wdz \quad (3)$$

where ϕ is velocity potential.

The necessary and sufficient condition for the relations.

$$u = -\frac{\partial\phi}{\partial x}, v = -\frac{\partial\phi}{\partial y}, w = -\frac{\partial\phi}{\partial z}$$

i.e. $\bar{q} = -\nabla\phi$ to hold is

$$\text{curl } \bar{q} = \text{curl } (-\nabla\phi) = \bar{0} \quad (4)$$

The solution of (2) i.e. $d\phi = 0$ is

$$\phi(x, y, z) = \text{const} \quad (5)$$

The surfaces (5) are called equipotentials. Thus the equipotentials are cut orthogonally by the stream lines.

4.2. Note : When $\text{curl } \bar{q} = \bar{0}$, the flow is said to be irrotational or of potential kind, otherwise it is rotational. For irrotational flow, $\bar{q} = -\nabla\phi$.

4.3. Example. The velocity potential of a two dimensional flow is $\phi = c xy$. Find the stream lines

Solution. The stream lines are given by

$$\frac{dx}{u} = \frac{dy}{v} = \frac{dz}{w}$$

where $\bar{q} = (u, v, w)$

For an irrotational motion (i.e. motion of potential kind)

we have

$$\text{curl } \bar{q} = 0 = \text{curl } (-\nabla\phi)$$

i.e. $\bar{q} = -\nabla\phi$, where ϕ is the velocity potential.

From here,

$$(u, v, w) = -\left(\frac{\partial\phi}{\partial x}, \frac{\partial\phi}{\partial y}, \frac{\partial\phi}{\partial z}\right) = -(cy, cx, 0)$$

i.e. $u = -cy, v = -cx, w = 0$

Therefore, streamlines are

$$\frac{dx}{-cy} = \frac{dy}{-cx} = \frac{dz}{0}$$

i.e. $x \, dx - y \, dy = 0, dz = 0$

i.e. $x^2 - y^2 = a^2, z = K$

which are rectangular hyperbolae

4.4. Example. If the speed of fluid is everywhere the same, the streamlines are straight.

Solution. The streamlines are given by the differential equations.

$$\frac{dx}{u} = \frac{dy}{v} = \frac{dz}{w}$$

where u, v, w are constants. The solutions are

$$vx - uy = \text{constant}, vz - wy = \text{constant}$$

The intersection of these planes are necessarily straight lines. Hence the result.

To obtain the differential equations for streamlines in cylindrical and spherical coordinates

We know that the streamlines are obtained from the differential equations

$$\bar{q} \times d\bar{r} = \bar{0} \quad (1)$$

where $\bar{q}$ is the velocity vector and $\bar{r}$ is the position vector of a liquid particle.

If the motion is irrotational, then

$$\bar{q} = -\nabla\phi$$

Therefore, the differential equations (1) become

$$\nabla\phi \cdot d\bar{r} = \bar{0} \quad (2)$$

(i) In cylindrical co-ordinates (r, θ, z) , we have

$$d\bar{r} = (dr, r d\theta, dz)$$

and

$$\nabla\phi = \text{grad } \phi = \left(\frac{\partial\phi}{\partial r}, \frac{1}{r} \frac{\partial\phi}{\partial\theta}, \frac{\partial\phi}{\partial z} \right)$$

Thus, the different equations (2) become

$$\begin{aligned} & \left(\frac{\partial\phi}{\partial r}, \frac{1}{r} \frac{\partial\phi}{\partial\theta}, \frac{\partial\phi}{\partial z} \right) \times (dr, r d\theta, dz) = \bar{0} \\ \Rightarrow & \frac{dr}{\partial\phi/\partial r} = \frac{rd\theta}{1/r \cdot \partial\phi/\partial\theta} = \frac{dz}{\partial\phi/\partial z}. \end{aligned} \quad (3)$$

(ii) In spherical co-ordinates (r, θ, ψ) , we have

$$d\bar{r} = (dr, rd\theta, r \sin\theta d\psi)$$

and

$$\nabla\phi = \text{grad } \phi = \left(\frac{\partial\phi}{\partial r}, \frac{1}{r} \frac{\partial\phi}{\partial\theta}, \frac{1}{r \sin\theta} \frac{\partial\phi}{\partial\psi} \right)$$

The differential equations (2) become.

$$\begin{aligned} & \left(\frac{\partial\phi}{\partial r}, \frac{1}{r} \frac{\partial\phi}{\partial\theta}, \frac{1}{r \sin\theta} \frac{\partial\phi}{\partial\psi} \right) \times (dr, rd\theta, r \sin\theta d\psi) = \bar{0} \\ \Rightarrow & \frac{dr}{\partial\phi/\partial r} = \frac{rd\theta}{\frac{1}{r} \partial\phi/\partial\theta} = \frac{r \sin\theta d\psi}{\frac{1}{r \sin\theta} \partial\phi/\partial\psi} \end{aligned} \quad (4)$$

Equations (3) and (4) are the required differential equations.

4.9. Example. Show that if the velocity potential of an irrotational fluid motion is $\phi = \frac{A}{r^2} \psi \cos \theta$, where (r, θ, ψ) are the spherical polar co-ordinates of any point, the lines of flow lie on the surface $r = k \sin^2 \theta$, k being a constant.

Solution. The differential equations for lines of flow (streamlines) are

$$\frac{dr}{\partial \phi / \partial r} = \frac{r d\theta}{\frac{1}{r} \partial \phi / \partial \theta} = \frac{r \sin \theta d\psi}{\frac{1}{r \sin \theta} \partial \phi / \partial \psi}$$

From first two members, we have

$$\begin{aligned} \frac{dr}{-\psi \frac{2A}{r^3} \cos \theta} &= \frac{r d\theta}{\frac{1}{r} \left(-\psi \frac{A}{r^2} \sin \theta \right)} \\ \Rightarrow \frac{dr}{\cos \theta} &= \frac{2r d\theta}{\sin \theta} \quad \Rightarrow \frac{dr}{r} = 2 \frac{\cos \theta}{\sin \theta} d\theta \\ \Rightarrow \log r &= 2 \log \sin \theta + \log k \quad \Rightarrow r = k \sin^2 \theta \end{aligned}$$

Hence the result.

4.10. Note. In the above example, the velocity potential, in Cartesian co-ordinates, can be written as

$$\phi = A(x^2 + y^2 + z^2)^{-3/2} z \cdot \tan^{-1} \left(\frac{y}{x} \right),$$

where

$$x = r \sin \theta \cos \psi, y = r \sin \theta \sin \psi, z = r \cos \theta$$

are spherical polar substitutions.

Also, the streamlines $r = k \sin^2 \theta$ can be written as $r^3 = k r^2 \sin^2 \theta$

$$\Rightarrow (x^2 + y^2 + z^2)^{3/2} = k (x^2 + y^2)$$

$$\text{i.e.} \quad x^2 + y^2 + z^2 = k^{2/3} (x^2 + y^2)^{2/3}$$

which are the streamlines in Cartesian co-ordinates.

4.11. Example. At the point in an incompressible fluid having spherical polar co-ordinates (r, θ, ψ) , the velocity components are $(2M\bar{r}^3 \cos\theta, M\bar{r}^3 \sin\theta, 0)$ where M is a constant. Show that velocity is of potential kind. Find the velocity potential and the equations of streamlines.

Solution. Here $d\bar{r} = dr\hat{r} + r d\theta\hat{\theta} + r \sin\theta d\psi\hat{\psi}$

$$\bar{q} = 2M\bar{r}^3 \cos\theta\hat{r} + M\bar{r}^3 \sin\theta\hat{\theta}$$

Then,

$$\begin{aligned} \text{curl } \bar{q} &= \frac{1}{r^2 \sin\theta} \begin{vmatrix} \hat{r} & r\hat{\theta} & r\sin\theta\hat{\psi} \\ \frac{\partial}{\partial r} & \frac{\partial}{\partial \theta} & \frac{\partial}{\partial \psi} \\ 2M\bar{r}^3 \cos\theta & M\bar{r}^3 \sin\theta & 0 \end{vmatrix} \\ &= \frac{1}{r^2 \sin\theta} \left[\hat{r} \cdot 0 + r\hat{\theta} \cdot 0 + r\sin\theta\hat{\psi}(-2M\bar{r}^3 \sin\theta + 2M\bar{r}^3 \sin\theta) \right] = \bar{0} \end{aligned}$$

Therefore, the flow is of potential kind.

Now, using the relation $\bar{q} = -\nabla\phi = -\left(\frac{\partial\phi}{\partial r}\hat{r} + \frac{1}{r}\frac{\partial\phi}{\partial\theta}\hat{\theta} + \frac{1}{r\sin\theta}\frac{\partial\phi}{\partial\psi}\hat{\psi}\right)$, we have

$$2M\bar{r}^3 \cos\theta\hat{r} + M\bar{r}^3 \sin\theta\hat{\theta} = \left(-\frac{\partial\phi}{\partial r}\hat{r} - \frac{1}{r}\frac{\partial\phi}{\partial\theta}\hat{\theta} - \frac{1}{r\sin\theta}\frac{\partial\phi}{\partial\psi}\hat{\psi}\right)$$

From here,

$$\frac{\partial\phi}{\partial r} = 2M\bar{r}^3 \cos\theta, -\frac{\partial\phi}{\partial\theta} = M\bar{r}^3 \sin\theta, \frac{\partial\phi}{\partial\psi} = 0$$

Therefore,

$$\begin{aligned} d\phi &= \frac{\partial\phi}{\partial r} dr + \frac{\partial\phi}{\partial\theta} d\theta + \frac{\partial\phi}{\partial\psi} d\psi \\ &= (-2M\bar{r}^3 \cos\theta)dr - (M\bar{r}^3 \sin\theta)d\theta \\ &= d(M\bar{r}^3 \cos\theta) \end{aligned}$$

Integrating, we get

$$\phi = M\bar{r}^3 \cos\theta$$

which is the required velocity potential.

The streamlines are given by

$$\frac{dr}{\frac{\partial \phi}{\partial r}} = \frac{rd\theta}{-\frac{1}{r} \frac{\partial \phi}{\partial \theta}} = \frac{r \sin \theta d\psi}{-\frac{1}{r \sin \theta} \frac{\partial \phi}{\partial \psi}}$$

$$\text{or } \frac{dr}{2Mr^3 \cos \theta} = \frac{rd\theta}{Mr^3 \sin \theta} = \frac{r \sin \theta d\psi}{0}$$

From the last term, $\psi = \text{constant}$.

From the first two terms, we get

$$\frac{dr}{r} = \frac{2 \cos \theta}{\sin \theta} d\theta = 2 \cot \theta d\theta$$

Integrating, we get

$$\log r = \log \sin^2 \theta + \text{constant}$$

$$\Rightarrow r = A \sin^2 \theta, \psi = \text{constant}$$

The equation $\psi = \text{const.}$ shows that the streamlines lie in planes which pass through the axis of symmetry $\theta = 0$.

5. Irrotational and Rotational Motion, Vortex Lines

5.1. Vorticity. If $\bar{q} = (u, v, w)$ be the velocity vector of a fluid particle, then the vector $\bar{\xi}$ defined by

$$\bar{\xi} = \text{curl } \bar{q} = \nabla \times \bar{q}$$

is called the **vortex vector** or **vorticity** and its components are (ξ_1, ξ_2, ξ_3) , given by

$$\xi_1 = \frac{\partial w}{\partial y} - \frac{\partial v}{\partial z}, \xi_2 = \frac{\partial u}{\partial z} - \frac{\partial w}{\partial x}, \xi_3 = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}$$

5.2. Vortex Motion (or Rotational Motion). The fluid motion is said to be rotational if

$$\bar{\xi} = \text{curl } \bar{q} \neq \bar{0}$$

5.3. Irrotational Motion. If $\bar{\xi} = \text{curl } \bar{q} = \bar{0}$, then the fluid motion is said to be irrotational or of potential kind and then $\bar{q} = -\nabla \phi$.

5.4. Vortexline. It is a curve in the fluid such that the tangent at any point on the curve has the direction of the vorticity vector $\bar{\xi}$.

The differential equations of vortexlines are given by $\bar{\xi} \times d\bar{r} = \bar{0}$

$$\text{i.e. } \frac{dx}{\xi_1} = \frac{dy}{\xi_2} = \frac{dz}{\xi_3}$$

where $\bar{\xi} = (\xi_1, \xi_2, \xi_3)$.

5.5. Vortex Tube. It is the locus of vortex line drawn at each point of a closed curve i.e. vortex tube is the surface formed by drawing vortex lines through each point of a closed curve in the fluid.

A vortex tube with small cross-section is called a **vortex filament**.

5.6. Flow. Let A and B be two points in the fluid.

Then $\int_B^A \bar{q} \cdot d\bar{r}$ is called the flow along any path from A to B

If motion is irrotational then $\bar{q} = -\nabla\phi$ and

$$\text{flow} = -\int_A^B \nabla\phi \cdot d\bar{r} = -\int_A^B d\phi = \phi(A) - \phi(B)$$

5.7. Circulation It is the flow round a closed curve. If C be the closed curve in a moving fluid, then circulation Γ about C is given by

$$\Gamma = \oint_C \bar{q} \cdot d\bar{r} = \int_S \hat{n} \cdot \text{curl } \bar{q} dS = \int_S \hat{n} \cdot \bar{\xi} dS$$

If the motion is irrotational, then $\bar{q} = -\nabla\phi$ and thus,

$$\Gamma = -\oint_C \nabla\phi \cdot d\bar{r} = -\oint_C d\phi = \phi(A) - \phi(A) = 0,$$

where A is any point on the curve C. This shows that for an irrotational motion, circulation is zero.

5.8. Theorem :- The necessary and sufficient condition such that the vortex lines are at right angles to the stream lines, is

$$(u, v, w) = \mu \left(\frac{\partial\phi}{\partial x}, \frac{\partial\phi}{\partial y}, \frac{\partial\phi}{\partial z} \right)$$

i.e. $\bar{q} = \mu \nabla\phi$, where μ and ϕ are functions of x, y, z and t.

Proof. Necessary condition:- We know that the differential equation

$$\begin{array}{l|l} \bar{q} \cdot d\bar{r} = 0 \text{ is integrable if } \left(\frac{\partial Q}{\partial z} - \frac{\partial R}{\partial y} \right) + \dots = 0 & \because p dx + Q dy + R dz = 0 \text{ is integrable if} \\ \bar{q} \cdot \text{curl } \bar{q} = 0 & \text{(exactness condition)} \end{array}$$

i.e. $\bar{q} \cdot \bar{\xi} = 0, \quad \bar{\xi} = \text{curl } \bar{q}$

This shows that the streamlines are at right angles to the vortex lines. Thus the streamlines and vortex lines are at right angles to each other if the differential equation $\bar{q} \cdot d\bar{r} = 0$ is integrable.

The exactness condition $\bar{q} \cdot \text{curl } \bar{q} = 0$ implies that $\bar{q} = -\nabla\phi$.

But $\text{curl } \bar{q} = \text{curl}(-\nabla\phi) = \bar{0}$. Thus the vortexlines do not exist. The equations $\bar{q} \cdot d\bar{r} = 0$ are therefore not exact.

So, there exists an integrating factor μ (function of x, y, z, t) such that

$$\mu^{-1} \bar{q} \cdot d\bar{r} = 0 \text{ is integrable.}$$

If this differential equation is integrable, then we can write

$$\mu^{-1} \bar{q} \cdot d\bar{r} = d\phi, \text{ where } \phi \text{ is a scalar function of } x, y, z, t$$

$$\Rightarrow \mu^{-1} \bar{q} \cdot d\bar{r} = \nabla\phi \cdot d\bar{r} \quad | \quad \because d\phi = \nabla\phi \cdot d\bar{r}$$

$$\Rightarrow \bar{q} = \mu \nabla\phi.$$

Sufficient condition :- Let us take $\bar{q} = \mu \nabla\phi \Rightarrow \nabla\phi = \mu^{-1} \bar{q}$

$$\text{Then, } \text{curl } \bar{q} = \text{curl } (\mu \nabla\phi)$$

$$\Rightarrow \bar{\xi} = \nabla \times (\mu \nabla\phi) = \mu(\nabla \times \nabla\phi) + \nabla \mu \times \nabla\phi = \nabla \mu \times \nabla\phi$$

Therefore,

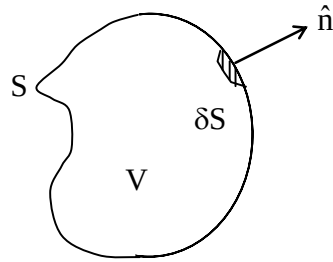
$$\begin{aligned} \bar{q} \cdot \bar{\xi} &= (\nabla \mu \times \nabla\phi) \cdot \bar{q} = \nabla \mu \cdot (\nabla\phi \times \bar{q}) \\ &= \nabla \mu \cdot (\mu^{-1} \bar{q} \times \bar{q}) = 0 \end{aligned}$$

This shows that the directions of streamlines and vortexlines are at right angles to each other.

6. Equation of Continuity

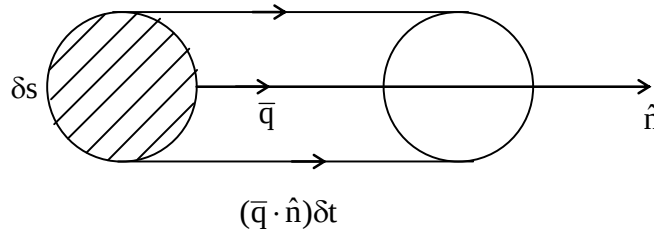
6.1. Equation of Continuity by Euler's Method (Equation of conservation of Mass): Equation of continuity is obtained by using the fact that the mass contained inside a given volume of fluid remains constant throughout the motion. Consider a region of fluid in which there is no inlets (sources) or outlets (sinks) through which the fluid can enter or leave the region. Let S be the surface enclosing volume V of this region and let $\hat{n}$ denotes the unit vector normal to an element δS of S drawn outwards.

Let $\vec{q}$ be the fluid velocity and ρ be the fluid density.



First, we consider the mass of fluid which leaves V by flowing across an element δS of S in time δt . This quantity is exactly that which is contained in a small cylinder of cross-section δS of length $(\vec{q} \cdot \hat{n}) \delta t$.

Thus, mass of the fluid is = density $\times$ Volume = $\rho (\vec{q} \cdot \hat{n}) \delta t \delta S$



Hence the rate at which fluid leaves V by flowing across the element δS is

$$\rho (\vec{q} \cdot \hat{n}) \delta S.$$

Summing over all such elements δS , we obtain the rate of flow of fluid coming out of V across the entire surface S . Hence, the rate at which mass flows out of the region V is

$$\int_S \rho (\vec{q} \cdot \hat{n}) dS = \int_S (\rho \vec{q}) \cdot \hat{n} dS$$

By Gauss divergence theorem

$$\int_S \vec{F} \cdot \hat{n} dS = \int_V \nabla \cdot \vec{F} dV.$$

$$= \int_V \text{div}(\rho \bar{q}) dV \quad (1)$$

Now, the mass M of the fluid possessed by the volume V of the fluid is

$$M = \int_V \rho dV, \text{ where } \rho = \rho(x, y, z, t) \text{ with } (x, y, z) \text{ the Cartesian}$$

co-ordinates of a general point of V , a fixed region of space. Since the space co-ordinates are independent of time t , therefore the rate of increase of mass within V is

$$\frac{dM}{dt} = \frac{d}{dt} \left(\int_V \rho dV \right) = \int_V \frac{\partial \rho}{\partial t} dV \quad | \text{ V does not change w.r.t. time } (2)$$

But the considered region is free from source or sink i.e. the mass is neither created nor destroyed, therefore the total rate of change of mass is zero and thus from (1) & (2), we get

$$\int_V \frac{\partial \rho}{\partial t} dV + \int_V \text{div}(\rho \bar{q}) dV = 0$$

$$\Rightarrow \int_V \left[\frac{\partial \rho}{\partial t} + \text{div}(\rho \bar{q}) \right] dV = 0$$

Since V is arbitrary, we conclude that at any point of the fluid which is neither a source nor a sink,

$$\frac{\partial \rho}{\partial t} + \text{div}(\rho \bar{q}) = 0$$

$$\text{i.e.} \quad \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \bar{q}) = 0 \quad (3)$$

Equation (3) is known as equation of continuity.

Corollary (1). We know that

$$\text{div}(\rho \bar{q}) = \rho \text{div} \bar{q} + \bar{q} \cdot (\text{grad} \rho)$$

Therefore, (3) takes the form

$$\frac{\partial \rho}{\partial t} + \rho(\nabla \cdot \bar{q}) + (\bar{q} \cdot \nabla) \rho = 0 \quad (4)$$

Corollary (2). We know that the differential operator $\frac{D}{Dt}$ is given by

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + (\bar{q} \cdot \nabla)$$

Therefore, from (4), we obtain the equation of continuity as $\frac{D\rho}{Dt} + \rho(\nabla \cdot \bar{q}) = 0$

$$\text{i.e.} \quad \frac{D\rho}{Dt} + \rho \operatorname{div} \bar{q} = 0 \quad (5)$$

Corollary (3). Equation (5) can be written as

$$\begin{aligned} & \frac{1}{\rho} \frac{D\rho}{Dt} + \operatorname{div} \bar{q} = 0 \\ \Rightarrow & \frac{D}{Dt}(\log \rho) + \operatorname{div} \bar{q} = 0 \end{aligned} \quad (6)$$

Corollary (4). When the motion of fluid is steady, then $\frac{\partial \rho}{\partial t} = 0$ and thus the equation of continuity (3) becomes

$$\operatorname{div}(\rho \bar{q}) = 0 \quad \text{Here } \rho \text{ is not a function of time i.e. } \rho = \rho(x, y, z) \quad (7)$$

Corollary (5). When the fluid is incompressible, then $\rho = \text{constant}$ and thus $\frac{D\rho}{Dt} = 0$.

The equation of continuity becomes

$$\operatorname{div} \bar{q} = 0 \quad (8)$$

which is same for homogeneous and incompressible fluid.

Corollary (6). If in addition to homogeneity and incompressibility, the flow is of potential kind such that $\bar{q} = -\nabla \phi$, then the equation of continuity becomes single word

$$\operatorname{div}(-\nabla \phi) = 0 \Rightarrow \nabla \cdot (\nabla \phi) = 0 \Rightarrow \nabla^2 \phi = 0 \quad (9)$$

which is known as the Laplace equation.

6.2. Equation of continuity in Cartesian co-ordinates :- Let (x, y, z) be the rectangular Cartesian co-ordinates.

$$\text{Let} \quad \bar{q} = u \hat{i} + v \hat{j} + w \hat{k} \quad (1)$$

$$\text{and } \nabla = \frac{\partial}{\partial x} \hat{i} + \frac{\partial}{\partial y} \hat{j} + \frac{\partial}{\partial z} \hat{k} \quad (2)$$

Then, the equation of continuity $\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \bar{q}) = 0$ can be written as

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) + \frac{\partial}{\partial y}(\rho v) + \frac{\partial}{\partial z}(\rho w) = 0 \quad (3)$$

$$\text{i.e.} \quad \frac{\partial \rho}{\partial t} + u \frac{\partial \rho}{\partial x} + v \frac{\partial \rho}{\partial y} + w \frac{\partial \rho}{\partial z} + \rho \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} \right) = 0 \quad (4)$$

which is the required equation of continuity in Cartesian co-ordinates.

Corollary (1). If the fluid motion is steady, then $\frac{\partial \rho}{\partial t} = 0$ and the equation (3) becomes

$$\frac{\partial}{\partial x}(\rho u) + \frac{\partial}{\partial y}(\rho v) + \frac{\partial}{\partial z}(\rho w) = 0 \quad (5)$$

Corollary (2). If the fluid is incompressible, then $\rho = \text{constant}$ and the equation of continuity is $\nabla \cdot \bar{q} = 0$

i.e.
$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0 \quad (6)$$

Corollary (3). If the fluid is incompressible and of potential kind, then equation of continuity is

$$\nabla^2 \phi = 0$$

i.e.
$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2} = 0, \text{ where } \bar{q} = -\nabla \phi.$$

6.14. Equation of Continuity by Lagrange's Method. Let initially a fluid element be at (a, b, c) at time $t = t_0$ when its volume is dV_0 and density is ρ_0 . After time t , let the same fluid element be at (x, y, z) when its volume is dV and density is ρ . Since mass of the fluid element remains invariant during its motion, we have

$$\rho_0 dV_0 = \rho dV \text{ i.e. } \rho_0 da db dc = \rho dx dy dz$$

$$\text{or} \quad \rho_0 da db dc = \rho \frac{\partial(x, y, z)}{\partial(a, b, c)} da db dc$$

$$\text{or} \quad \rho J = \rho_0 \quad (1)$$

$$\text{where } J = \frac{\partial(x, y, z)}{\partial(a, b, c)}$$

which is the required equation of continuity.

6.15. Remark. By simple property of Jacobians, we get

$$\frac{dJ}{dt} = J \nabla \cdot \bar{q}$$

$$\text{Thus (1) gives } \frac{d}{dt}(\rho J) = 0 \Rightarrow \frac{d\rho}{dt} J + \rho \frac{dJ}{dt} = 0$$

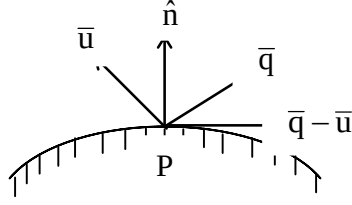
$$\Rightarrow \frac{d\rho}{dt} J + \rho J \nabla \cdot \bar{q} = 0 \Rightarrow \frac{d\rho}{dt} + \rho \nabla \cdot \bar{q} = 0 \quad \text{or} \quad \frac{D\rho}{Dt} + \rho \nabla \cdot \bar{q} = 0$$

which is the Euler's equation of continuity.

7. Boundary Surfaces

Physical conditions that should be satisfied on given boundaries of the fluid in motion, are called boundary conditions. The simplest boundary condition occurs where an ideal and incompressible fluid is in contact with rigid impermeable boundary, e.g., wall of a container or the surface of a body which is moving through the fluid.

Let P be any point on the boundary surface where the velocity of fluid is $\bar{q}$ and velocity of the boundary surface is $\bar{u}$.



The velocity at the point of contact of the boundary surface and the liquid must be tangential to the surface otherwise the fluid will break its contact with the boundary surface. Thus, if $\hat{n}$ be the unit normal to the surface at the point of contact, then

$$(\bar{q} - \bar{u}) \cdot \hat{n} = 0 \Rightarrow \bar{q} \cdot \hat{n} = \bar{u} \cdot \hat{n} \quad (1)$$

In particular, if the boundary surface is at rest, then $\bar{u} = \bar{0}$ and the condition becomes

$$\bar{q} \cdot \hat{n} = 0 \quad (2)$$

Another type of boundary condition arrives at a free surface where liquid borders a vacuum eg. the interface between liquid and air is usually regarded as free surface. For this free surface, pressure p satisfies

$$P = \Pi \quad (3)$$

where Π denotes the pressure outside the fluid i.e. the atmospheric pressure. Equation (3) is a dynamic boundary condition.

Third type of boundary condition occurs at the boundary between two immiscible ideal fluids in which the velocities are $\bar{q}_1$ & $\bar{q}_2$ and pressures are p_1 & p_2 respectively.

Now, we find the condition that a given surface satisfies to be a boundary surface.

7.1. Article. To obtain the differential equation satisfied by boundary surface of a fluid in motion

or

To find the condition that the surface.

$$F(\bar{r}, t) = F(x, y, z, t) = 0$$

may represent a boundary surface :-

If $\bar{q}$ be the velocity of fluid and $\bar{u}$ be the velocity of the boundary surface at a point P of contact, then

$$(\bar{q} - \bar{u}) \cdot \hat{n} = 0 \Rightarrow \bar{q} \cdot \hat{n} = \bar{u} \cdot \hat{n} \quad (1)$$

where $\bar{q} - \bar{u}$ is the relative velocity and $\hat{n}$ is a unit vector normal to the surface at P.

The equation of the given surface is

$$F(\bar{r}, t) = F(x, y, z, t) = 0 \quad (2)$$

We know that a unit vector normal to the surface (2) is given by

$$\hat{n} = \frac{\nabla F}{|\nabla F|}$$

Thus, from (1), we get $\bar{q} \cdot \nabla F = \bar{u} \cdot \nabla F \quad (3)$

since the boundary surface is itself in motion, therefore at time $(t + \delta t)$, it's equation is given by

$$F(\bar{r} + \delta\bar{r}, t + \delta t) = 0. \quad (4)$$

From (2) & (4), we have

$$F(\bar{r} + \delta\bar{r}, t + \delta t) - F(\bar{r}, t) = 0$$

$$\text{i.e. } F(\bar{r} + \delta\bar{r}, t + \delta t) - F(\bar{r}, t + \delta t) + F(\bar{r}, t + \delta t) - F(\bar{r}, t) = 0$$

By Taylor's series, we can have

$$(\delta\bar{r} \cdot \nabla)F(\bar{r}, t + \delta t) + \delta t \frac{\partial}{\partial t} \{F(\bar{r}, t)\} = 0$$

$$\begin{aligned} \therefore F(x + \delta x, y + \delta y, z + \delta z) &= F(x, y, z) + \delta x \frac{\partial F}{\partial x} + \delta y \frac{\partial F}{\partial y} + \delta z \frac{\partial F}{\partial z} + \dots \\ &= F(x, y, z) + \delta\bar{r} \cdot \nabla F \end{aligned}$$

$$\Rightarrow \left(\frac{\delta\bar{r}}{\delta t} \cdot \nabla \right) F(\bar{r}, t + \delta t) + \frac{\partial F}{\partial t} = 0$$

Taking limit as $\delta t \rightarrow 0$, we get

$$\left(\frac{d\bar{r}}{dt} \cdot \nabla \right) F + \frac{\partial F}{\partial t} = 0$$

$$\Rightarrow \frac{\partial F}{\partial t} + (\bar{q} \cdot \nabla)F = 0 \quad \text{i.e. } \frac{DF}{Dt} = 0 \quad (5)$$

which is the required condition for any surface F to be a boundary surface

Corollary (1) If $\bar{q} = (u, v, w)$, then the condition (5) becomes

$$\frac{\partial F}{\partial t} + u \frac{\partial F}{\partial x} + v \frac{\partial F}{\partial y} + w \frac{\partial F}{\partial z} = 0$$

In case, the surface is rigid and does not move with time, then $\frac{\partial F}{\partial t} = 0$ and the

boundary condition is $u \frac{\partial F}{\partial x} + v \frac{\partial F}{\partial y} + w \frac{\partial F}{\partial z} = 0$ i.e. $(\bar{q} \cdot \nabla)F = 0$

Corollary (2) The boundary condition

$$\frac{\partial F}{\partial t} + u \frac{\partial F}{\partial x} + v \frac{\partial F}{\partial y} + w \frac{\partial F}{\partial z} = 0$$

is a linear equation and its solution gives

$$\frac{dt}{1} = \frac{dx}{u} = \frac{dy}{v} = \frac{dz}{w} \quad \left| \frac{D}{Dt} \equiv \frac{d}{dt} \text{ in Lagrangian view} \right.$$

$$\Rightarrow \frac{dx}{dt} = u, \frac{dy}{dt} = v, \frac{dz}{dt} = w$$

which are the equations of path lines.

Hence once a particle is in contact with the surface, it never leaves the surface.

Corollary (3) From equation (5), we have

$$\bar{q} \cdot \nabla F = \frac{-\partial F}{\partial t}$$

$$\Rightarrow \bar{q} \cdot \frac{\nabla F}{|\nabla F|} = \frac{-\partial F/\partial t}{|\nabla F|}$$

$$\Rightarrow \bar{q} \cdot \hat{n} = \frac{-\partial F/\partial t}{|\nabla F|}$$

which gives the normal velocity.

Also from (1), we get

$$\bar{u} \cdot \hat{n} = \frac{-\partial F/\partial t}{|\nabla F|} \quad \left| \because \bar{q} \cdot \hat{n} = \bar{u} \cdot \hat{n} \right.$$

which gives the normal velocity of the boundary surface.

7.2. Example. Show that the ellipsoid

$$\frac{x^2}{a^2 k^2 t^{2n}} + k t^n \left[\left(\frac{y}{b} \right)^2 + \left(\frac{z}{c} \right)^2 \right] = 1$$

is a possible form of the boundary surface of a liquid.

Solution. The surface $F(x, y, z, t) = 0$ can be a possible boundary surface, if it satisfies the boundary condition.

$$\frac{DF}{Dt} = \frac{\partial F}{\partial t} + u \frac{\partial F}{\partial x} + v \frac{\partial F}{\partial y} + w \frac{\partial F}{\partial z} = 0 \quad (1)$$

where u, v, w satisfy the equation of continuity

$$\nabla \cdot \vec{q} = 0 \text{ i.e. } \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0 \quad (2)$$

$$\text{Here, } F(x, y, z, t) \equiv \frac{x^2}{a^2 k^2 t^{2n}} + k t^n \left[\left(\frac{y}{b} \right)^2 + \left(\frac{z}{c} \right)^2 \right] - 1 = 0$$

$$\begin{aligned} \text{Therefore, } \frac{\partial F}{\partial t} &= -\frac{x^2 \cdot 2n}{a^2 k^2 t^{2n+1}} + n k t^{n-1} \left[\left(\frac{y}{b} \right)^2 + \left(\frac{z}{c} \right)^2 \right] \\ \frac{\partial F}{\partial x} &= \frac{2x}{a^2 k^2 t^{2n}}, \quad \frac{\partial F}{\partial y} = \frac{2k t^n y}{b^2}, \quad \frac{\partial F}{\partial z} = \frac{2k t^n z}{c^2}. \end{aligned}$$

Thus, from (1), we get

$$\begin{aligned} &\frac{-x^2}{a^2 k^2 t^{2n+1}} + n k t^{n-1} \left[\left(\frac{y}{b} \right)^2 + \left(\frac{z}{c} \right)^2 \right] \\ &\quad + \frac{2x u}{a^2 k^2 t^{2n}} + \frac{2k t^n y v}{b^2} + \frac{2k t^n z w}{c^2} = 0 \end{aligned}$$

$$\text{or } \left(u - \frac{nx}{t} \right) \frac{2x}{a^2 k^2 t^{2n}} + \left(v + \frac{ny}{2t} \right) \frac{2k y t^n}{b^2} + \left(w + \frac{nz}{2t} \right) \frac{2k z t^n}{c^2} = 0$$

which will hold. if we take

$$u - \frac{nx}{t} = 0, \quad v + \frac{ny}{2t} = 0, \quad w + \frac{nz}{2t} = 0$$

$$\text{i.e. } u = \frac{nx}{t}, \quad v = -\frac{ny}{2t}, \quad w = -\frac{nz}{2t} \quad (3)$$

It will be a justifiable step if equation (2) is satisfied.

$$\text{i.e. } \frac{n}{t} + \frac{-n}{2t} + \frac{-n}{2t} = 0.$$

which is true.

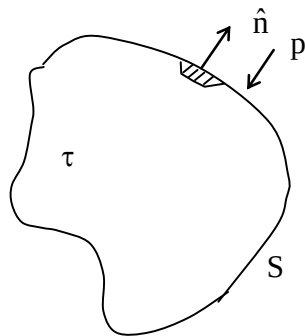
Hence the given ellipsoid is a possible form of boundary surface of a liquid.

9. Equations of Motion

9.1. Euler's Equation of Motion of an Ideal Fluid (Equation of Conservation of Momentum). To obtain Euler's dynamical equation, we shall make use of Newton's second law of motion.

Consider a region τ of fluid bounded by a closed surface S which consists of the same fluid particles at all times. Let $\bar{q}$ be the velocity and ρ be the density of the fluid.

Then $\rho \, d\tau$ is an element of mass within S and it remains constant.



The linear momentum of volume τ is

$$\bar{M} = \int_{\tau} \bar{q} \rho d\tau \quad | \text{ mass} \times \text{velocity} = \text{momentum.}$$

Rate of change of momentum is

$$\frac{d\bar{M}}{dt} = \frac{d}{dt} \int_{\tau} \bar{q} \rho d\tau = \int_{\tau} \frac{d\bar{q}}{dt} \rho d\tau \quad (1)$$

The fluid within τ is acted upon by two types of forces

The first type of forces are the surface forces which are due to the fluid exterior to τ .

Since the fluid is ideal, the surface force is simply the pressure p directed along the inward normal at all point of S .

The total surface force on S is

$$\int_S p(-\hat{n})dS = -\int_S p \hat{n}dS = \int_{\tau} \nabla p d\tau \quad (\text{By Gauss div. Theorem}) \quad (2)$$

The second type of forces are the body forces which are due to some external agent. Let $\bar{F}$ be the body force per unit mass acting on the fluid. Then $\bar{F} \rho d\tau$ is the body force on the element of mass $\rho d\tau$ and the total body force on the mass within τ is

$$\int_{\tau} \bar{F} \rho d\tau \quad (3)$$

By Newton's second law of motion, we have

Rate of change of momentum = total force

$$\Rightarrow \int_{\tau} \frac{d\bar{q}}{dt} \rho d\tau = \int_{\tau} \bar{F} \rho d\tau - \int_{\tau} \nabla p d\tau$$

$$\Rightarrow \int_{\tau} \left(\frac{d\bar{q}}{dt} \rho - \bar{F} \rho + \nabla p \right) d\tau = 0$$

Since $d\tau$ is arbitrary, we get

$$\frac{d\bar{q}}{dt} \rho - \bar{F} \rho + \nabla p = 0$$

i.e.

$$\frac{d\bar{q}}{dt} = \bar{F} - \frac{1}{\rho} \nabla p \quad (4)$$

which holds at every point of the fluid and is known as Euler's dynamical equation for an ideal fluid.

9.2. Remark. The above method for obtaining the Euler's equation of motion, is also known as **flux method**.

9.3. Other Forms of Euler's Equation of Motion. (i) We know that

$$\frac{d}{dt} \equiv \frac{D}{Dt} = \frac{\partial}{\partial t} + \bar{q} \cdot \nabla,$$

therefore equation (4) becomes.

$$\frac{\partial \bar{q}}{\partial t} + (\bar{q} \cdot \nabla) \bar{q} = \bar{F} - \frac{1}{\rho} \nabla p \quad (5)$$

$$\text{But } (\bar{q} \cdot \nabla) \bar{q} = \nabla \left(\frac{1}{2} \bar{q}^2 \right) + \bar{\xi} \times \bar{q}, \quad \bar{\xi} = \text{curl } \bar{q}$$

Therefore, Euler's equation becomes

$$\frac{\partial \bar{q}}{\partial t} + \nabla \left(\frac{1}{2} \bar{q}^2 \right) + \bar{\xi} \times \bar{q} = \bar{F} - \frac{1}{\rho} \nabla p. \quad (6)$$

Equation (6) is called Lamb's hydrodynamical equation

(ii) Cartesian Form. Let $\bar{q} = (u, v, w)$, $\bar{F} = (X, Y, Z)$ and $\nabla p = \left(\frac{\partial p}{\partial x}, \frac{\partial p}{\partial y}, \frac{\partial p}{\partial z} \right)$,

then equation (5) gives

$$\left. \begin{aligned} \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} + w \frac{\partial u}{\partial z} &= X - \frac{1}{\rho} \frac{\partial p}{\partial x} \\ \frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} + w \frac{\partial v}{\partial z} &= Y - \frac{1}{\rho} \frac{\partial p}{\partial y} \\ \frac{\partial w}{\partial t} + u \frac{\partial w}{\partial x} + v \frac{\partial w}{\partial y} + w \frac{\partial w}{\partial z} &= Z - \frac{1}{\rho} \frac{\partial p}{\partial z} \end{aligned} \right\} \quad (7)$$

Equation (7) are the required equations in Cartesian form.

9.4. Remark :- The two equations, the equation of continuity and the Euler's equation of motion, comprise the equations of motion of an ideal fluid. Thus the equations

$$\frac{\partial \rho}{\partial t} + \text{div}(\rho \bar{q}) = 0$$

$$\text{and } \frac{\partial \bar{q}}{\partial t} + (\bar{q} \cdot \nabla) \bar{q} = \bar{F} - \frac{1}{\rho} \nabla p$$

are fundamental to any theoretical study of ideal fluid flow. These equations are solved subject to the appropriate boundary and initial conditions dictated by the physical characteristics of the flow.

9.5. Lagrange's Equation of Motion. Let initially a fluid element be at (a, b, c) at time $t = t_0$ when its volume is dV_0 and density is ρ_0 . After time t , let the same fluid element be at (x, y, z) when its volume is dV and density is ρ . The equation of continuity is

$$\rho J = \rho_0 \quad (1)$$

$$\text{where } J = \frac{\partial(x, y, z)}{\partial(a, b, c)}$$

The components of acceleration are

$$\ddot{x} = \frac{\partial^2 x}{\partial t^2}, \ddot{y} = \frac{\partial^2 y}{\partial t^2}, \ddot{z} = \frac{\partial^2 z}{\partial t^2}$$

Let the body force $\bar{F}$ be conservative so that we can express it in terms of a body force potential function Ω as

$$\bar{F} = -\nabla \Omega \quad (2)$$

By Euler's equation of motion,

$$\frac{d\bar{q}}{dt} = \bar{F} - \frac{1}{\rho} \nabla p = -\nabla \Omega - \frac{1}{\rho} \nabla p \quad (3)$$

Its Cartesian equivalent is

$$\left. \begin{aligned} \frac{\partial^2 x}{\partial t^2} &= -\frac{\partial \Omega}{\partial x} - \frac{1}{\rho} \frac{\partial p}{\partial x} \\ \frac{\partial^2 y}{\partial t^2} &= -\frac{\partial \Omega}{\partial y} - \frac{1}{\rho} \frac{\partial p}{\partial y} \\ \frac{\partial^2 z}{\partial t^2} &= -\frac{\partial \Omega}{\partial z} - \frac{1}{\rho} \frac{\partial p}{\partial z} \end{aligned} \right\} \quad (4)$$

We note that a, b, c, t are the independent variables and our object is to determine x, y, z in terms of a, b, c, t and so investigate completely the motion.

To deduce equations containing only differentiations w.r.t. the independent variables a, b, c, t we multiply the equations in (4) by $\partial x/\partial a, \partial y/\partial a, \partial z/\partial a$ and add to get

$$\frac{\partial^2 x}{\partial t^2} \frac{\partial x}{\partial a} + \frac{\partial^2 y}{\partial t^2} \frac{\partial y}{\partial a} + \frac{\partial^2 z}{\partial t^2} \frac{\partial z}{\partial a} = -\frac{\partial \Omega}{\partial a} - \frac{1}{\rho} \frac{\partial p}{\partial a} \quad (5)$$

Similarly, we get

$$\frac{\partial^2 x}{\partial t^2} \frac{\partial x}{\partial b} + \frac{\partial^2 y}{\partial t^2} \frac{\partial y}{\partial b} + \frac{\partial^2 z}{\partial t^2} \frac{\partial z}{\partial b} = -\frac{\partial \Omega}{\partial b} - \frac{1}{\rho} \frac{\partial p}{\partial b} \quad (6)$$

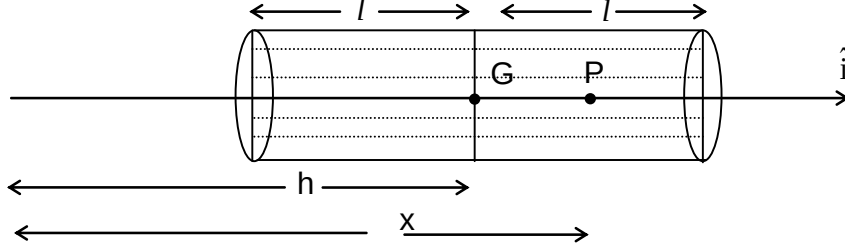
$$\frac{\partial^2 x}{\partial t^2} \frac{\partial x}{\partial c} + \frac{\partial^2 y}{\partial t^2} \frac{\partial y}{\partial c} + \frac{\partial^2 z}{\partial t^2} \frac{\partial z}{\partial c} = -\frac{\partial \Omega}{\partial c} - \frac{1}{\rho} \frac{\partial p}{\partial c} \quad (7)$$

These equations (5), (6), (7) together with equation (1) constitute Lagrange's Hydrodynamical Equations.

9.6. Example. A homogeneous incompressible liquid occupies a length $2l$ of a straight tube of uniform small bore and is acted upon by a body force which is such that the fluid is attracted to a fixed point of the tube, with a force varying as the distance from the point. Discuss the motion and determine the velocity and pressure within the liquid.

Solution. We note that the small bore of the tube permits us to ignore any variation of velocity across any cross-section of the tube and to suppose that the flow is unidirectional.

We u be the velocity along the tube and p be the pressure at a general point P at distance x from the centre of force O . Also, let h be the distance of the centre of mass G of the fluid, as shown in the figure.



Equations of motion of the fluid are :

(i) Equation of Continuity : Here, $\bar{q} = (u, 0, 0)$

Therefore, equation of continuity becomes

$$\frac{\partial u}{\partial x} = 0 \Rightarrow u = u(t) \quad (1)$$

(ii) Euler's Equation : In this case, it becomes

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + X - \frac{1}{\rho} \frac{\partial p}{\partial x} = -\mu x - \frac{1}{\rho} \frac{\partial p}{\partial x}$$

$$\Rightarrow \frac{\partial u}{\partial t} = -\mu x - \frac{1}{\rho} \frac{\partial p}{\partial x} \quad (2) \quad | \text{ using (1)}$$

where $-\mu x \hat{i}$ is the body force per unit mass, μ being a positive constant.

We observe that equation (2) can be written as

$$\frac{du}{dt} = -\mu x - \frac{1}{\rho} \frac{dp}{dx} \quad (3)$$

Integrating w.r.t. x , we get

$$x \frac{du}{dt} = -\mu \frac{x^2}{2} - \frac{p}{\rho} + C \quad (4)$$

where C is a constant and at most can be a function of t only. w.r.t. (x, y, z)

Let Π be the pressure at the free surfaces $x = h-l$ and $x = h+l$ of the liquid. Then using these boundary conditions, equation (4) becomes

$$(h-l) \frac{du}{dt} = -\frac{1}{2} \mu (h-l)^2 - \frac{\Pi}{\rho} + C$$

$$(h+l) \frac{du}{dt} = -\frac{1}{2}\mu(h-l)^2 - \frac{\Pi}{\rho} + C$$

which on subtraction give

$$\frac{du}{dt} = -\mu h \quad (5)$$

But in the fluid motion all fluid particles move with the same velocity u and u

$$= \frac{dh}{dt}$$

$\therefore$ Equation (5) becomes

$$\frac{d^2h}{dt^2} = -\mu h \quad (6)$$

Now, we solve the different equation (6), which can be written as

$$(D^2 + \mu) h = 0$$

Here auxiliary equation is

$$D^2 + \mu = 0 \Rightarrow D = \pm \sqrt{\mu} i$$

Therefore, the solution of (6) is

$$h = A \cos(\sqrt{\mu} t + \epsilon)$$

where A and ϵ are constants which can be determined from initial conditions.

To Calculate Pressure :- We have from (3) & (5)

$$-\mu x - \frac{1}{\rho} \frac{dp}{dx} = -\mu h$$

$$\Rightarrow \frac{1}{\rho} \frac{dp}{dx} = \mu(h - x)$$

Integrating w.r.t. x , we get

$$\frac{p}{\rho} = \frac{\mu(h-x)^2}{2(-1)} + D \quad (7)$$

The boundary condition $x = h - l$, $p = \Pi$ gives

$$\frac{\Pi}{\rho} = \mu \frac{l^2}{-2} + D$$

i.e. $D = \Pi/\rho + \frac{\mu l^2}{2}$

Therefore, equation (7) becomes

$$\begin{aligned} \frac{p}{\rho} &= \frac{\mu(h-x)^2}{-2} + \Pi/\rho + \frac{\mu l^2}{2} \\ &= \frac{\Pi}{\rho} - \frac{\mu}{2} [(h-x)^2 - l^2] \\ &= \frac{\Pi}{\rho} - \frac{\mu}{2} [(h-x+l)(h-x-l)] \end{aligned}$$

10. Bernoulli's Equation (Theorem)

10.1. For Steady Flow. We shall obtain a special form of Euler's dynamical equation in terms of pressure. The Euler's dynamical equation is

$$\frac{d\bar{q}}{dt} = \bar{F} - \frac{1}{\rho} \nabla p \quad (1)$$

where $\bar{q}$ is velocity, $\bar{F}$ is the body force, p and ρ are pressure and density respectively.

$\bar{F}$ be conservative so that it can be expressed in terms of a body force potential function Ω as

$$\bar{F} = -\nabla \Omega \quad (2)$$

When the flow is steady, then $\frac{\partial \bar{q}}{\partial t} = 0$ (3)

Therefore, in case of steady motion with a conservative body force equation (1), on using (2) and (3), gives

$$\nabla \left(\frac{1}{2} \bar{q}^2 \right) - \bar{q} \times \bar{\xi} = -\nabla \Omega - \frac{1}{\rho} \nabla p$$

$$\left| \begin{array}{l} \because \frac{d\bar{q}}{dt} = \frac{\partial \bar{q}}{\partial t} + (\bar{q} \cdot \nabla) \bar{q} \\ \text{or } \frac{d\bar{q}}{dt} = \frac{\partial \bar{q}}{\partial t} + \nabla \left(\frac{1}{2} \bar{q}^2 \right) - \bar{q} \times \bar{\xi} \text{ and } \frac{\partial \bar{q}}{\partial t} = 0 \end{array} \right.$$

$$\Rightarrow \nabla \left(\frac{1}{2} \bar{q}^2 + \Omega \right) + \frac{1}{\rho} \nabla p = \bar{q} \times \bar{\xi} \quad (4)$$

Further, if we suppose that the liquid is barotropic i.e. density is a function of pressure p only, then we can write

$$\frac{1}{\rho} \nabla p = \nabla \int \frac{dp}{\rho}$$

Using this in (4), we get

$$\nabla \left[\frac{1}{2} \bar{q}^2 + \Omega + \int \frac{dp}{\rho} \right] = \bar{q} \times \bar{\xi}. \quad (5)$$

Multiplying (5) scalarly by $\bar{q}$ and noting that

$$\bar{q} \cdot (\bar{q} \times \bar{\xi}) = (\bar{q} \times \bar{q}) \cdot \bar{\xi} = 0, \text{ we get}$$

$$\bar{q} \cdot \nabla \left[\frac{1}{2} \bar{q}^2 + \Omega + \int \frac{dp}{\rho} \right] = 0 \quad (6)$$

If $\hat{s}$ is a unit vector along the streamline through general point of the fluid and s measures distance along this stream line, then since $\hat{s}$ is parallel to $\bar{q}$, therefore equation (6) gives

$$\frac{\partial}{\partial s} \left[\frac{1}{2} \bar{q}^2 + \Omega + \int \frac{dp}{\rho} \right] = 0 \quad \left| \begin{array}{l} \because \hat{s} \text{ is parallel to } \bar{q} \\ \bar{q} = k \hat{s} \\ \hat{s} \cdot \nabla \equiv \frac{\partial}{\partial s} \end{array} \right.$$

Hence along any particular streamline, we have

$$\frac{1}{2} \bar{q}^2 + \Omega + \int \frac{dp}{\rho} = C \quad (7)$$

where C is constant which takes different values for different streamlines. Equation (7) is known as Bernaull's equation. This result applies to steady flow of ideal. barotropic fluids in which the body forces are conservative.

Now, if $\hat{s}$ is a unit vector taken along a vortexline, then, similarly, we get

$$\frac{1}{2}\bar{q}^2 + \Omega + \int \frac{dp}{\rho} = C \quad \text{along any particular vortexline.} \quad (\text{Here, we}$$

multiply scalarly by $\bar{\xi}$)

10.2. Remark. (i) If $\bar{q} \times \bar{\xi} = \bar{0}$ i.e. if $\bar{q}$ & $\bar{\xi}$ are parallel, then streamlines and vortex lines coincide and $\bar{q}$ is said to be **Beltrami vector**.

If $\bar{\xi} = \bar{0}$, the flow is irrotational.

For both of these flow patterns,

$$\frac{1}{2}\bar{q}^2 + \Omega + \int \frac{dp}{\rho} = C$$

where C is same at all points of the fluid.

(ii) For homogeneous incompressible fluids, ρ is constant and

$$\int \frac{dp}{\rho} = \frac{p}{\rho}.$$

The Bernoulli's equation becomes

$$\frac{p}{\rho} + \frac{1}{2}\bar{q}^2 + \Omega = C$$

so that if $\bar{q}$ is known, the pressure can be calculated.

10.3. For Unsteady Irrotational Flow. Here also, we suppose that the body forces are conservative i.e. $\bar{F} = -\nabla\Omega$

For irrotational flow, $\bar{q} = -\nabla\phi \Rightarrow \nabla \times \bar{q} = 0$

The equation of motion

$$\frac{\partial \bar{q}}{\partial t} + \nabla \left(\frac{1}{2}\bar{q}^2 \right) - \bar{q} \times (\nabla \times \bar{q}) = \bar{F} - \frac{1}{\rho} \nabla p \quad (1)$$

in the present case becomes.

$$-\nabla \left(\frac{\partial \phi}{\partial t} \right) + \nabla \left(\frac{1}{2}\bar{q}^2 \right) = -\nabla \Omega - \frac{1}{\rho} \nabla p$$

$$\Rightarrow \nabla \left(\frac{1}{2} \bar{q}^2 + \Omega + \int \frac{dp}{\rho} - \frac{\partial \phi}{\partial t} \right) = 0 \quad | \text{ Barotropic fluid.}$$

Integrating, we get

$$\frac{1}{2} \bar{q}^2 + \Omega + \int \frac{dp}{\rho} - \frac{\partial \phi}{\partial t} = f(t) \quad (2)$$

which is the required equation.

If the liquid is homogeneous, then $\int \frac{dp}{\rho} = \frac{p}{\rho}$ and the equation (2) become

$$\frac{1}{2} \bar{q}^2 + \Omega + \frac{p}{\rho} - \frac{\partial \phi}{\partial t} = f(t).$$

Further, for study case,

$$\frac{\partial \phi}{\partial t} = 0, f(t) = \text{const}$$

$$\frac{1}{2} \bar{q}^2 + \Omega + \frac{p}{\rho} = \text{const}$$

10.4. Example. A long straight pipe of length L has a slowly tapering circular cross section. It is inclined so that its axis makes an angle α to the horizontal with its smaller cross-section downwards. The radius of the pipe at its upper end is twice that of at its lower end and water is pumped at a steady rate through the pipe to emerge at atmospheric pressure. If the pumping pressure is twice the atmospheric pressure, show that the fluid leaves the pipe with a speed U given by

$$U^2 = \frac{32}{15} \left[gL \sin \alpha + \frac{\Pi}{\rho} \right],$$

where Π is atmospheric pressure

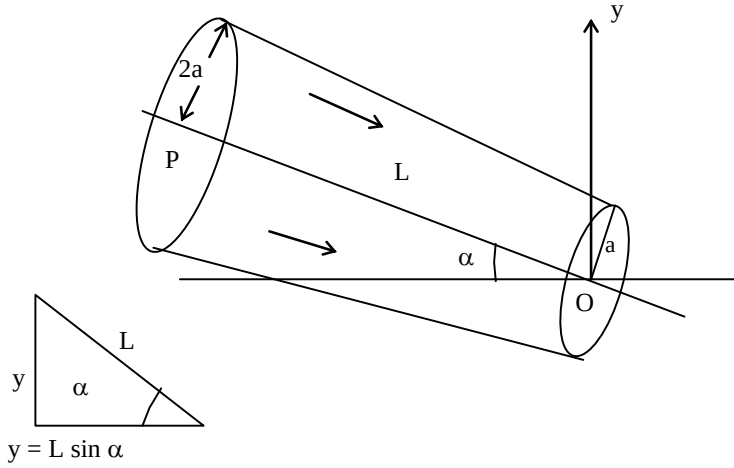
Solution. The assumption that the pipe is slowly tapering means that any variation in the velocity over any cross-section can be ignored. Let the velocity at the wider end of the pipe be V and the emerging velocity be U (velocity at the lower end). The only body force is that of gravity, so $\bar{F} = -g \hat{j}$ and consequently $\Omega = gy$

$$\left| \begin{array}{l} \because \vec{F} = -\nabla\Omega \Rightarrow -q\hat{j} = -\nabla\Omega = -\frac{\partial\Omega}{\partial x}\hat{i} - \frac{\partial\Omega}{\partial y}\hat{j} - \frac{\partial\Omega}{\partial z}\hat{k} \\ \Rightarrow -g = -\frac{\partial\Omega}{\partial y} \Rightarrow \Omega = gy \end{array} \right.$$

Bernoulli's equation, $\frac{p}{\rho} + \frac{1}{2}q^2 + \Omega = C$ | $\because$ For water ρ is const.

becomes $\frac{p}{\rho} + \frac{1}{2}q^2 + gy = C$ (1)

Applying this equation of the two ends of the pipe, we get



$$\frac{2\Pi}{\rho} + \frac{1}{2}V^2 + gL\sin\alpha = \frac{\Pi}{\rho} + \frac{1}{2}U^2 \quad (2) \quad \text{for}$$

lower end $y = 0$

Let a and $2a$ be the radii of the lower and upper ends respectively, then by the principle of conservation of mass

$$\begin{aligned} \pi(2a)^2V &= \pi a^2 U \\ \Rightarrow V &= \frac{U}{4} \end{aligned} \quad (3)$$

From (2) and (3), we obtain

$$\Pi + \frac{1}{2}\rho\left(\frac{U^2}{16}\right) + g\rho L\sin\alpha = \frac{1}{2}\rho U^2$$

$$\Rightarrow \quad \frac{1}{2}\rho\left(U^2 - \frac{U^2}{16}\right) = \Pi + g\rho L \sin \alpha.$$

$$\Rightarrow \quad \frac{15}{32}\rho U^2 = \Pi + g\rho L \sin \alpha.$$

$$\Rightarrow \quad U^2 = \frac{32}{15}\left[gL \sin \alpha + \frac{\Pi}{\rho}\right]$$

Hence the result.

Stream Function

When motion is the same in all planes parallel to xy plane (say) and there is no velocity parallel to the z-axis, i.e. when u, v are functions of x, y, t only and $w = 0$, we may regard the motion as **two-dimensional** and consider only the cases confined to the xy plane. When we speak of the flow across a curve in this plane, we shall mean the flow across unit length of a cylinder whose trace on the xy plane is the curve in question, the generators of the cylinder being parallel to the z-axis.

For a two-dimensional motion in xy-plane, $\bar{q}$ is a function of x, y, t only and the differential equation of the streamlines (lines of flow) are

$$\frac{dx}{u} = \frac{dy}{v} \text{ i.e. } vdx - udy = 0 \quad (1)$$

and the corresponding equation of continuity is

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0 \quad (2)$$

We note that equation (2) is the condition of exactness of (1), it follows that (1) must be an exact differential, $d\psi$ (say). Thus

$$vdx - udy = d\psi = \frac{\partial \psi}{\partial x} dx + \frac{\partial \psi}{\partial y} dy$$

so that
$$u = - \frac{\partial \psi}{\partial y}, \quad v = \frac{\partial \psi}{\partial x}$$

This function ψ is called the **stream function** or the **current function** or **Lagrange's stream function**.

Obviously, the streamlines are given by the solution of (1) i.e. $d\psi = 0$ i.e. $\psi = \text{constant}$. (For unsteady flow, streamlines are given by $\psi = f(t)$) Thus, the stream function is constant along a streamline.

From the above discussion, it is clear that the existence of stream function is merely a consequence of the continuity and incompressibility of the fluid. The

stream function always exists in all types of **two dimensional** motion whether rotational or irrotational. However, it should be noted again that velocity potential exists only for **irrotational motion** whether two dimensional or three dimensional.

12.1. Physical Interpretation of Stream Function :-

Let P be a point on a curve C in xy-plane. Let an element ds of the curve makes an angle θ with x-axis. The direction cosines of the normal at P are

$$(\cos(90^\circ + \theta), \cos \theta, 0)$$

i.e. $(-\sin \theta, \cos \theta, 0)$.

The flow across the curve C from right to left is

$$= \int_C \bar{q} \cdot \hat{n} ds, \text{ where } \hat{n} = -\sin \theta \hat{i} + \cos \theta \hat{j},$$

$$\bar{q} = u \hat{i} + v \hat{j}$$

$$= \int_C (-u \sin \theta + v \cos \theta) ds$$

=

$$\int_C \left(\frac{\partial \psi}{\partial y} \sin \theta + \frac{\partial \psi}{\partial x} \cos \theta \right) ds$$

$$\left| u = -\frac{\partial \psi}{\partial y}, v = \frac{\partial \psi}{\partial x} \right.$$

=

$$\int_C \left(\frac{\partial \psi}{\partial x} \frac{dx}{ds} + \frac{\partial \psi}{\partial y} \frac{dy}{ds} \right) ds$$

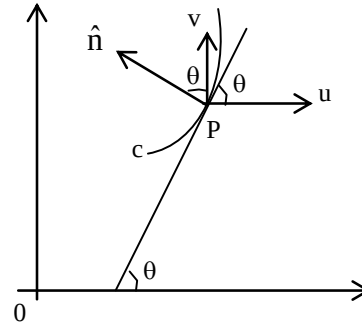
$$\left| \cos \theta = \frac{dx}{ds}, \sin \theta = \frac{dy}{ds} \right.$$

$$= \int_C \left(\frac{\partial \psi}{\partial x} dx + \frac{\partial \psi}{\partial y} dy \right)$$

$$= \int_C d\psi = (\psi_B - \psi_A)$$

where ψ_A and ψ_B are the values of ψ at the initial and final points of the curve. Thus, the difference of the values of a stream function at any two points represents the flow across that curve, joining the two points.

12.2. Corollary. If we suppose that the curve C be the streamline, then no fluid crosses its boundary, then



$$(\psi_B - \psi_A) = 0 \Rightarrow \psi_B = \psi_A$$

i.e. ψ is constant along c.

12.3. Relation Between ϕ and ψ (i.e. C-R equations) :-

We know that the velocity potential ϕ is given by

$$\bar{q} = -\nabla\phi = -\left(\frac{\partial\phi}{\partial x}, \frac{\partial\phi}{\partial y}\right)$$

$$\text{i.e. } u = -\frac{\partial\phi}{\partial x}, v = -\frac{\partial\phi}{\partial y} \quad (1)$$

Also, the stream function ψ is given by

$$u = -\frac{\partial\psi}{\partial y}, v = \frac{\partial\psi}{\partial x} \quad (2)$$

From (1) and (2), we get

$$\frac{\partial\phi}{\partial x} = \frac{\partial\psi}{\partial y} \text{ and } \frac{\partial\phi}{\partial y} = -\frac{\partial\psi}{\partial x} \quad (3)$$

Equations in (3) imply

$$\nabla^2\phi = 0 \text{ and } \nabla^2\psi = 0$$

i.e. ϕ and ψ are harmonic functions.

Again, from (3), we get

$$\begin{aligned} \nabla\phi &= \text{grad } \phi = -\bar{q} = -(u\hat{i} + v\hat{j}) \\ &= -\left(-\frac{\partial\psi}{\partial y}\hat{i} + \frac{\partial\psi}{\partial x}\hat{j}\right) \\ &= \frac{\partial\psi}{\partial y}\hat{i} - \frac{\partial\psi}{\partial x}\hat{j} \\ &= \frac{\partial\psi}{\partial y}(\hat{j} \times \hat{k}) + \frac{\partial\psi}{\partial x}(\hat{i} \times \hat{k}) \\ &= \left(\frac{\partial\psi}{\partial x}\hat{i} + \frac{\partial\psi}{\partial y}\hat{j}\right) \times \hat{k} \\ &= \nabla\psi \times \hat{k} = \text{grad } \psi \times \hat{k} \end{aligned}$$

$$\text{i.e. } \text{grad } \phi = (\text{grad } \psi) \times \hat{k} = -\hat{k} \times \text{grad } \psi$$

$$\text{i.e.} \quad \nabla \phi = \nabla \psi \times \hat{k} \quad (4)$$

Again, from (3), we note that

$$\begin{aligned} \frac{\partial \phi}{\partial x} \frac{\partial \psi}{\partial x} &= \frac{\partial \psi}{\partial y} \left(-\frac{\partial \phi}{\partial y} \right) \\ \Rightarrow \quad \frac{\partial \phi}{\partial x} \frac{\partial \psi}{\partial x} + \frac{\partial \phi}{\partial y} \frac{\partial \psi}{\partial y} &= 0 \end{aligned}$$

$$\text{i.e.} \quad \nabla \phi \cdot \nabla \psi = 0 \quad (5)$$

Thus, for irrotational incompressible two-dimensional flow (steady or unsteady), $\phi(x, y)$, $\psi(x, y)$ are harmonic functions and the family of curves $\phi = \text{constant}$ (equipotentials) and $\psi = \text{constant}$ (streamlines) intersect orthogonally.

12.4. Exercise. Show that $u = 2cxy$, $v = c(a^2 + x^2 - y^2)$ are the velocity components of a possible fluid motion. Determine the stream function and the streamlines.

12.5. Remark. We shall consider the study of two dimensional motion later on. At present we continue discussing three dimensional irrotational flow of incompressible fluids.

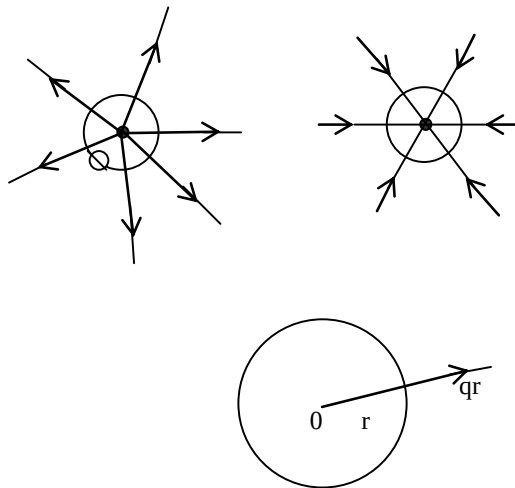
UNIT - III & IV

4. Sources, Sinks and Doublets

4.1. Source : An outward symmetrical radial flow of fluid in all directions is termed as a three dimensional source or a **point source** or a **simple source**.

Thus, a source is a point at which fluid is continuously created and distributed e.g. an expanding bubble of gas pushing away the surrounding fluid. If the volume of fluid per unit time which is emitted from a simple source at 0 is constant and equal to $4\pi m$, then m is termed as strength of the source.

4.2. Sink : A negative source is called a **sink**. At such points, the fluid is constantly moving radically inwards from all directions. Thus a simple sink of strength m is a simple source of strength $-m$.



4.3. Velocity Potential due to a Simple Source of Strength m . Let there be a source of strength m at a point 0. With 0 as the centre, we draw a sphere of radius r around 0.

The flow across the sphere per unit volume is given by

$$\int_S \bar{q} \cdot \hat{n} dS$$

In case of a source there is only the radial velocity i.e. $\bar{q}$ has only radial component q_r .

Therefore, the flow is

$$= \int_S q_r dS \quad |\bar{q} \cdot \hat{n} = q_r, \text{ since } \bar{q} \text{ and } \hat{n} \text{ have same directions i.e.}$$

radial direction.

$$= q_r (4\pi r^2).$$

Thus, we get

$$4\pi m = q_r (4\pi r^2)$$

$$\Rightarrow q_r = \frac{m}{r^2} = -\frac{\partial}{\partial r} \left(\frac{m}{r} \right) \quad (1)$$

It is observed that $\text{curl } \bar{q} = \bar{0}$ (except at $r = 0$), therefore for irrotational flow,

$$q_r = -\frac{\partial \phi}{\partial r} \quad |\bar{q} = -\nabla \phi \quad (2)$$

From (1) & (2), we find

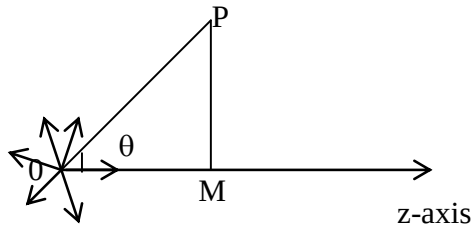
$$\phi = \frac{m}{r}$$

which is the required expression for the velocity potential for a source.

4.4. Remarks. (i) For a simple sink of strength m , the velocity potential is $\phi =$

$$-\frac{m}{r}$$

(ii) A source or sink implies the creation or annihilation of liquid at a point. Both are points at which the velocity potential (and stream function for two dimensional case) become infinite and therefore, they require special analysis.



4.5. A simple Source in Uniform Stream. Let us consider a simple source of strength m at O in a uniform stream having undisturbed velocity $U \hat{k}$, $\hat{k}$ be the unit vector along z -axis which is taken as the axis of symmetry of the flow.

We shall find the velocity potential at any point $P(z, \theta, \psi)$. From P , draw $\perp$ on OZ . Let $OP = r$, $\angle POZ = \theta$; $OM = z$

We observe that the velocity potential of the uniform stream in the absence of source is

$$\left| \begin{array}{l} \bar{q} = -\nabla\phi \Rightarrow U\hat{k} = -\frac{\partial\phi}{\partial z}\hat{k} \\ \Rightarrow \frac{\partial\phi}{\partial z} = -U \Rightarrow \phi = -Uz \end{array} \right.$$

$$\phi_1 = -Uz = -Ur \cos \theta \quad (1)$$

and the velocity potential of the simple source is

$$\phi_2 = \frac{m}{r} \quad (2)$$

Thus, the velocity potential of the combination is

$$\begin{aligned} \phi &= \phi_1 + \phi_2 = -Ur \cos \theta + \frac{m}{r} \\ &= -\left(Ur \cos \theta - \frac{m}{r} \right) \end{aligned} \quad (3)$$

From here, the velocity components at $P(r, \theta, \psi)$ are

$$\begin{aligned} q_r &= -\frac{\partial\phi}{\partial r} = U \cos \theta + \frac{m}{r^2} \\ q_\theta &= -\frac{1}{r} \frac{\partial\phi}{\partial \theta} = -U \sin \theta \\ q_\psi &= -\frac{1}{r \sin \theta} \frac{\partial\phi}{\partial \psi} = 0 \end{aligned} \quad \left| \begin{array}{l} 0 \leq \theta \leq \pi \\ 0 \leq \psi \leq 2\pi \\ \frac{\partial}{\partial \psi} \equiv 0 \end{array} \right.$$

The stagnation points ($\bar{q} = 0$) are given by $U \cos \theta + \frac{m}{r^2} = 0$, $\sin \theta = 0 \Rightarrow \theta = 0$

or π

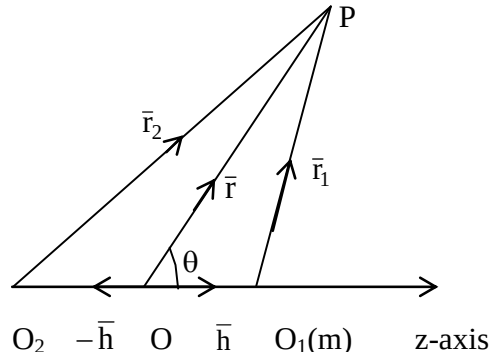
But $\theta = 0$ gives r to be imaginary $\Rightarrow \theta = \pi$ and $r = \sqrt{\frac{m}{U}}$

Thus there is only one stagnation point $\left(\sqrt{\frac{m}{U}}, \pi, 0\right)$

4.6. Doublet (Dipole). The combination of a source and a sink of equal strength, at a small distance apart, is called a doublet.

4.7. To Find the Velocity Potential of Doublet. Suppose that there is a simple source of

strength m at O_1 and a simple sink of strength m at O_2 . Origin O is taken as the mid point. of $O_1 O_2$. It is also assumed that there is no other source or sink. Let P be a fixed point within the fluid and



$$\overline{OP} = \bar{r}, \quad \overline{O_1P} = \bar{r}_1, \quad \overline{O_2P} = \bar{r}_2, \quad \angle POO_1 = \theta,$$

$$\overline{OO_1} = \bar{h}, \quad \overline{OO_2} = -\bar{h}, \quad h = |\bar{h}|$$

The velocity potential at P due to the combination of source and sink at O_1 and O_2 is

$$\begin{aligned} \phi &= \frac{m}{r_1} - \frac{m}{r_2} = \frac{mr_2 - mr_1}{r_1 r_2} \\ &= \frac{m(r_2 - r_1)}{r_1 r_2} = \frac{m(r_2^2 - r_1^2)}{r_1 r_2 (r_1 + r_2)} \\ &= \frac{m(\bar{r}_2 - \bar{r}_1) \cdot (\bar{r}_2 + \bar{r}_1)}{r_1 r_2 (r_1 + r_2)} \end{aligned}$$

But

$$\bar{r}_2 - \bar{r}_1 = 2\bar{h} \text{ and } \bar{r}_2 + \bar{r}_1 = 2\bar{r}$$

$$\begin{cases} \bar{r}_2 = \bar{h} + \bar{r} \\ \bar{r}_1 = -\bar{h} + \bar{r} \end{cases}$$

$$\begin{aligned}
 \text{Thus } \phi &= \frac{m(2\bar{h}).(2\bar{r})}{r_1 r_2 (r_1 + r_2)} = \frac{4m\bar{h}.\bar{r}}{r_1 r_2 (r_1 + r_2)} \\
 &= \frac{2\bar{\mu}.\bar{r}}{r_1 r_2 (r_1 + r_2)}, \text{ where } \bar{\mu} = 2m\bar{h}
 \end{aligned} \tag{1}$$

In equation (1), let us first keep $\bar{\mu}$ a finite constant and non-zero vector, so that $\mu = |\bar{\mu}|$ is a finite constant and non-zero scalar. Let $\bar{h} \rightarrow \bar{0}$ along $\overline{O_1 O_2}$.

Then $m \rightarrow \infty$ in such a way that $\bar{\mu}$ remains the same finite non-zero constant vector. In that case, both $r_1, r_2 \rightarrow r$ and thus under this limiting process, (1) results in

$$\phi = \frac{2\bar{\mu}.\bar{r}}{2r^3} = \frac{\mu r \cos \theta}{r^3} = \frac{\mu \cos \theta}{r^2} \tag{2}$$

The limiting source sink combination obtained at 0 when we keep the direction of $\bar{h}$ fixed but let $h \rightarrow 0$ and $m \rightarrow \infty$ with $\mu = 2mh$ remaining a finite non-zero constant, is called a three-dimensional doublet (or dipole). The scalar quantity μ is called the moment or strength of the doublet. The vector quantity $\bar{\mu} = \mu \hat{\mu}$ is called the vector moment of the doublet & $\hat{\mu}$ (unit vector from O_2 to O_1) determines the direction of the axis of the doublet from sink to source.

From (2), the velocity components are given by

$$q_r = -\frac{\partial \phi}{\partial r} = \frac{2\mu \cos \theta}{r^3}$$

$$q_\theta = -\frac{1}{r} \frac{\partial \phi}{\partial \theta} = \frac{\mu \sin \theta}{r^3}$$

$$q_\psi = 0$$

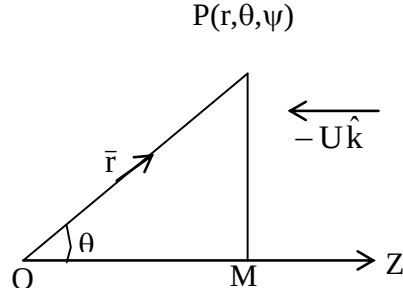
The streamlines due to the doublet are given by

$$\frac{dr}{\frac{2\mu \cos \theta}{r^3}} = \frac{r d\theta}{\frac{\mu \sin \theta}{r^3}} = \frac{r \sin \theta d\psi}{0}$$

$$\Rightarrow d\psi = 0 \Rightarrow \psi = \text{constant and } \frac{dr}{r} = 2 \cot \theta d\theta$$

$$\Rightarrow r = A \sin^2 \theta$$

4.8. Doublet in a Uniform Stream. Let there be a doublet of vector moment $\bar{\mu} = \mu \hat{k}$ at O in a uniform stream whose velocity in the absence of the doublet is $-U \hat{k}$ ($U = \text{constant}$).



Let P be a point in the fluid having spherical polar co-ordinates (r, θ, ψ) , the direction $\overline{OZ}$ of the doublets axis being the line $\theta = 0$. We shall find the resultant velocity potential due to the combination of the uniform stream and the doublet. We know that the velocity potential due to the uniform stream is

$$\phi_1 = Uz = Ur \cos \theta \quad (1)$$

and the velocity potential due to a doublet at O, is

$$\phi_2 = \frac{\mu \cos \theta}{r^2} \quad (2)$$

Thus, the resultant velocity potential at P. due to the combination, is

$$\phi = \phi_1 + \phi_2 = (Ur + \mu \bar{r}^2) \cos \theta$$

From here, the velocity component are

$$q_r = -\frac{\partial \phi}{\partial r} = -\left(U - \frac{2\mu}{r^3}\right) \cos \theta$$

$$q_\theta = -\frac{1}{r} \frac{\partial \phi}{\partial \theta} = \left(U + \frac{\mu}{r^3}\right) \sin \theta$$

$$q_\psi = \frac{1}{r \sin \theta} \frac{\partial \phi}{\partial \psi} = 0$$

Stagnation points are determined by solving.

$$\left(U - \frac{2\mu}{r^3}\right) \cos \theta = 0, \left(U + \frac{\mu}{r^3}\right) \sin \theta = 0 \quad | \quad \bar{q} = \bar{0}$$

which are satisfied when $\sin\theta = 0$ and $r = \left(\frac{2\mu}{U}\right)^{1/3}$

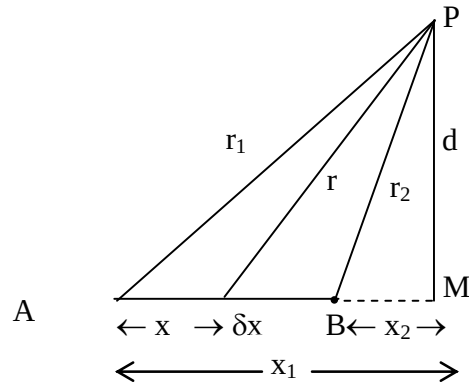
Thus, we have the two stagnation points.

$$\left(\left(\frac{2\mu}{U}\right)^{1/3}, 0\right) \text{ and } \left(\left(\frac{2\mu}{U}\right)^{1/3}, \pi\right)$$

which lie on the axis of symmetry.

If we write $r = a$ i.e. $a = \left(\frac{2\mu}{U}\right)^{1/3}$ i.e. $\mu = \frac{1}{2} U a^3$, then for the region $r \geq a$, we obtain the same velocity potential as for a uniform flow past a fixed impermeable sphere of radius a and centre 0. Thus, for $r \geq a$, the effect of the sphere is that of a doublet of strength $\mu = \frac{1}{2} U a^3$ situated at its centre, its axis pointing upstream. So the sphere can be represented by a suitably chosen singularity at its centre.

4.9. Line Distribution of Sources. Let us consider a uniform line source AB of strength m per unit length. This means that the elemental section of AB at a distance x from A and of length δx is a point source of strength $m\delta x$.



Let P be a point in the fluid at a distance r from this element, then the velocity potential at P due to the point source is $\frac{m\delta x}{r}$.

The total velocity potential at P due to the entire line distribution AB ($= 2l$) is

$$\phi = m \int_0^{2l} \frac{dx}{r} \quad (1)$$

Let $AM = x_1$, $BM = x_2$, where AM is the orthogonal projection of AP on AB . Also, let $PM = d$, $AP = r_1$, $BP = r_2$. Since $r^2 = (x_1 - x)^2 + d^2 = (x_1 - x)^2 + r_1^2 - x_1^2$, therefore from (1), we get

$$\begin{aligned}
 \phi &= m \int_0^{2l} \frac{dx}{\sqrt{(x_1 - x)^2 + (r_1^2 - x_1^2)}} \\
 &= m \left[\frac{\log \left\{ \frac{(x_1 - x) + \sqrt{(x_1 - x)^2 + (r_1^2 - x_1^2)}}{-1} \right\}}{-1} \right]_0^{2l} \quad \left| \begin{array}{l} \because \int_{\alpha}^{\beta} \frac{1}{\sqrt{x^2 + a^2}} dx \\ = \left[\log \left(x + \sqrt{x^2 + a^2} \right) \right]_{\alpha}^{\beta} \end{array} \right. \\
 &= m \left[\log \left\{ (x_1 - x) + \sqrt{(x_1 - x)^2 + (r_1^2 - x_1^2)} \right\} \right]_{2l}^0 \\
 &= m \left[\log(x_1 + r_1) - \log \left\{ x_2 + \sqrt{x_2^2 + r_1^2 - x_1^2} \right\} \right] \quad | \because x_1 - 2l = x_1 - AB = x_2 \\
 &= m \log \left(\frac{x_1 + r_1}{x_2 + r_2} \right), \text{ where } r_1^2 - x_1^2 = d^2 = r_2^2 - x_2^2.
 \end{aligned}$$

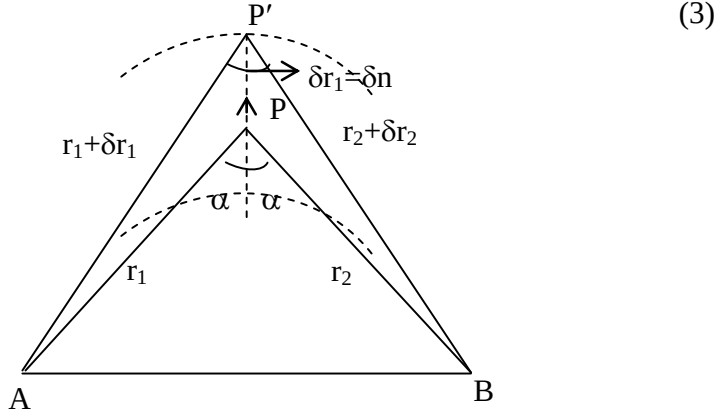
Again, the relation $r_1^2 - x_1^2 = r_2^2 - x_2^2$

$$\begin{aligned}
 \Rightarrow \quad \frac{r_1 + x_1}{r_2 + x_2} &= \frac{r_2 - x_2}{r_1 - x_1} = \frac{r_1 + r_2 + x_1 - x_2}{r_1 + r_2 + x_2 - x_1} \\
 &= \frac{r_1 + r_2 + 2l}{r_1 + r_2 - 2l}
 \end{aligned}$$

$$\begin{aligned}
 \text{Thus,} \quad \phi &= m \log \left(\frac{r_1 + r_2 + 2l}{r_1 + r_2 - 2l} \right) \\
 &= m \log \left(\frac{a + l}{a - l} \right) \quad (2)
 \end{aligned}$$

where $2a$ is the length of major axis of the ellipsoid of revolution through P having A and B as foci since for such an ellipsoid $r_1 + r_2 = \text{constant}$. It follows from here that the equipotential surfaces $\phi = \text{constant}$ are precisely the family of confocal ellipsoid $r_1 + r_2 = 2a$ obtained when a is allowed to vary.

Expression for Velocity :- The velocity at P is given by $\bar{q} = -\nabla\phi = -\left(\frac{\partial\phi}{\partial n}\right)\hat{n}$



Let P be any point on the ellipsoid specified by parameter a and P' the neighbouring point on the ellipsoid specified by parameter a + delta a, where $\overline{PP'} = \delta n \hat{n}$

$$\text{Thus } \bar{q} = -m \frac{\partial}{\partial n} \left[\log \frac{a+l}{a-l} \right] \hat{n} = -m \left[\frac{1}{a+l} - \frac{1}{a-l} \right] \frac{\partial a}{\partial n} \hat{n} = \frac{2lm}{a^2 - l^2} \frac{\partial a}{\partial n} \hat{n} \quad (4)$$

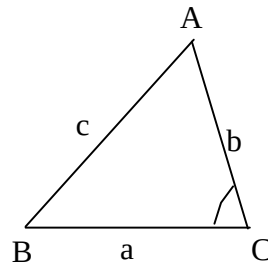
The normal at P to the a-surface bisects the angle 2α between the focal radii AP, BP.

Now,

$$(r_1 + \delta r_1)^2 = r_1^2 + (\delta n)^2 - 2r_1 \delta n \cos (180-\alpha)$$

$$= r_1^2 + (\delta n)^2 + 2r_1 \delta n \cos \alpha$$

$$\left| \begin{array}{l} \cos C = \frac{a^2 + b^2 - c^2}{2ab} \\ \Rightarrow c^2 = a^2 + b^2 - 2ab \cos C \end{array} \right.$$



$$\Rightarrow 2r_1 \delta r_1 = 2r_1 \delta n \cos \alpha + (\delta n)^2 - (\delta r_1)^2$$

$$\Rightarrow \delta r_1 = \delta n \cos \alpha \quad | \quad (\delta r_1)^2 = (\delta n)^2$$

$$\Rightarrow \frac{\partial r_1}{\partial n} = \cos \alpha$$

$$\text{Similarly, } \frac{\partial r_2}{\partial n} = \cos \alpha$$

$$\text{Since, } 2a = r_1 + r_2$$

$$\Rightarrow 2 \frac{\partial a}{\partial n} = \frac{\partial r_1}{\partial n} + \frac{\partial r_2}{\partial n} = \cos \alpha + \cos \alpha = 2 \cos \alpha$$

$$\Rightarrow \frac{\partial a}{\partial n} = \cos \alpha$$

and thus from equation (4), the velocity of fluid at P is given by

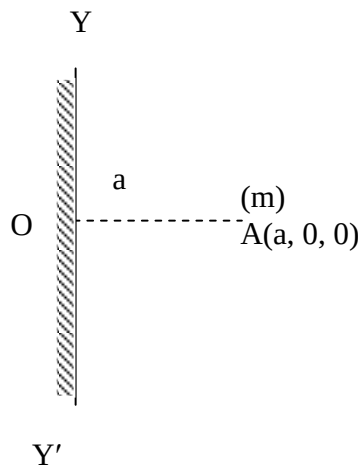
$$\bar{q} = \left[\frac{2lm \cos \alpha}{a^2 - l^2} \right] \hat{n}$$

5. Hydrodynamical Images for Three Dimensional Flows

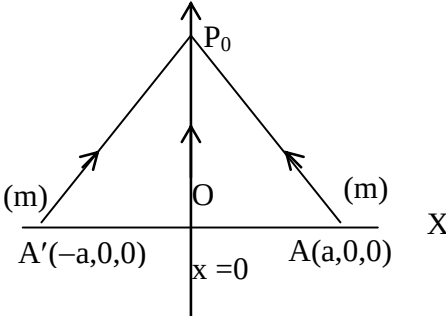
Let us consider a fluid containing a distribution of sources, sinks and doublets. If a surface S can be drawn in the fluid across which there is no flow, then any system of sources, sinks and doublets on opposite sides of this surface S may be said to be images of one another w.r.t. to the surface. Further, if the surface S be considered as a rigid boundary and the liquid removed from one side of it, the motion on the other side will remain unaltered.

5.1. Images in a Rigid Impermeable Infinite Plane. (i) Image of a source in a plane : consider a simple source of strength m situated at A(a, 0, 0) at a distance a from an infinite plane YY'.

We shall show that the appropriate image system for this is an equal source of strength m at A'(-a, 0, 0), the reflection of A in the plane.



To prove this, we consider two equal sources of strength m at A(a, 0, 0) & A'(-a, 0, 0) with no rigid boundary. Let P₀ be any point on the plane YY'. Then the fluid velocity at P₀ due to the two sources is

$$\bar{q} = \frac{m}{(AP_0)^3} \overline{AP_0} + \frac{m}{(A'P_0)^3} \overline{A'P_0} \quad \left| \quad \bar{q} = \frac{m}{r^2} \hat{r} = \frac{m}{r^3} \bar{r} \right.$$


$$\Rightarrow \bar{q} = \frac{m}{(AP_0)^3} (\overline{AP_0} + \overline{A'P_0})$$

$$= \frac{m}{(AP_0)^3} (2\overline{OP_0}) = \frac{2m}{(AP_0)^3} (\overline{OP_0}) \quad \left| \quad \begin{aligned} &\because \overline{AP_0} + \overline{A'P_0} \\ &= (\overline{AO} + \overline{OP_0}) + (\overline{A'O} + \overline{OP_0}) \\ &= 2\overline{OP_0} \end{aligned} \right.$$

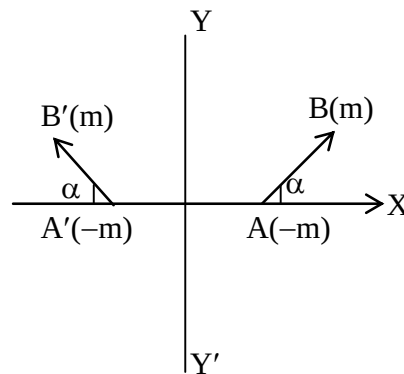
This shows that at any point P_0 of the plane YY' , the fluid flows tangentially to the plane $x = 0$ and so there is no transport of fluid across this plane.

Let ϕ denotes the velocity potential then, at all points P_0 on the plane YY' , the normal component of velocity is zero

$$\Rightarrow \frac{\partial \phi}{\partial n} = 0. \text{ Hence, the image of a source at } A \text{ in the rigid plane } YY' \text{ is}$$

a source at A' , as required.

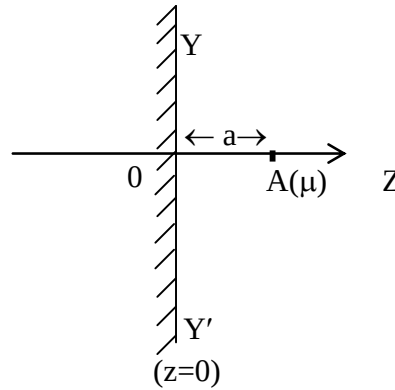
(ii) Image of Doublet in a Plane : Consider a pair of sources $-m$ at A and m at B , taken close together and on one side of the rigid plane YY' . The image system is $-m$ at A' , m at B' , where A' & B' are respectively the reflections of A and B in the plane YY' . In the limiting case, when $B \rightarrow A$ along $\overline{BA}$ in such a way as to form a doublet at A , we find that the image of



a doublet in an infinite impermeable rigid plane is a doublet of equal strength and symmetrically disposed to the other w.r.t the plane.

5.2. Example. A three dimensional doublet of strength μ whose axis is in the direction $\overline{OZ}$ is distant a from the rigid plane $z = 0$ which is the sole boundary of liquid of constant density ρ , infinite in extent. If p_∞ be the pressure at ∞ , show that the pressure on the plane is least at a distance $\frac{a\sqrt{5}}{2}$ from the doublet

Solution. Let there be a doublet of strength μ at the point A with $OA = a$ and YY' (i.e. $z = 0$) be the infinite plane. Then the image system is an equal doublet of strength μ at A' , the reflection of A in the plane $z = 0$, and the axis along $\overline{ZO}$. The line OZ is taken as the initial line $\theta = 0$ and plane $z = 0$ is $\theta = \pi/2$.



so that $P(r, \theta, \psi)$ is confined to the region $0 \leq \theta \leq \pi/2$. Let $AP = r_1$, $A'P = r_2$ and α_1, α_2 be the angles which these lines make with the axis of the doublets as shown in the figure.

Then, the velocity potential at P is

$$\phi = \frac{\mu \cos \alpha_1}{r_1^2} + \frac{\mu \cos \alpha_2}{r_2^2} \quad (1)$$

where

$$\left. \begin{aligned} r_1^2 &= r^2 + a^2 - 2ra \cos \theta \\ r_2^2 &= r^2 + a^2 + 2ra \cos \theta \end{aligned} \right\} \quad (2)$$

(By cosine formulae in ΔPOA , POA')

But $\cos \alpha_1 = \frac{AM}{r_1} = \frac{OM - OA}{r_1} = \frac{r \cos \theta - a}{r_1}$

and $\cos (180 - \alpha_2) = \frac{A'M}{r_2} = \frac{A'O + OM}{r_2} = \frac{a + r \cos \theta}{r_2}$

$\Rightarrow \cos \alpha_2 = -\frac{(a + r \cos \theta)}{r_2}$

Using these relations in (1), we get

$$\begin{aligned} \phi &= \frac{\mu}{r_1^2} \left(\frac{r \cos \theta - a}{r_1} \right) + \frac{\mu}{r_2^2} \left[\frac{-(a + r \cos \theta)}{r_2} \right] \\ &= \mu \left[\frac{r \cos \theta - a}{r_1^3} - \frac{r \cos \theta + a}{r_2^3} \right] \end{aligned} \quad (3)$$

Further from (2), we have

$$2 r_1 \frac{\partial r_1}{\partial r} = 2r - 2a \cos \theta \Rightarrow \frac{\partial r_1}{\partial r} = \frac{r - a \cos \theta}{r_1}$$

Similarly, $\frac{\partial r_2}{\partial r} = \frac{r + a \cos \theta}{r_2}$, $\frac{\partial r_1}{\partial \theta} = \frac{ra \sin \theta}{r_1}$

$$\frac{\partial r_2}{\partial \theta} = -\frac{ra \sin \theta}{r_2}.$$

Thus from (3), the velocity components are given by

$$\begin{aligned} q_r &= -\frac{\partial \phi}{\partial r} = \mu \left[\frac{\cos \theta}{r_2^3} - 3 \left(\frac{\partial r_2}{\partial r} \right) \frac{1}{r_2^4} (r \cos \theta + a) - \frac{\cos \theta}{r_1^3} + 3 \left(\frac{\partial r_1}{\partial r} \right) \frac{1}{r_1^4} (r \cos \theta - a) \right] \\ &= \mu \left[\frac{\cos \theta}{r_2^3} - 3 \frac{(r + a \cos \theta)(r \cos \theta + a)}{r_2^5} - \frac{\cos \theta}{r_1^3} + \frac{3(r - a \cos \theta)(r \cos \theta - a)}{r_1^5} \right] \\ q_\theta &= -\frac{1}{r} \frac{\partial \phi}{\partial \theta} = \frac{\mu}{r} \left[\frac{r \sin \theta}{r_1^3} + 3 \frac{(r \cos \theta - a) \left(\frac{\partial r_1}{\partial \theta} \right)}{r_1^4} - \frac{r \sin \theta}{r_2^3} - 3 \frac{(r \cos \theta + a) \left(\frac{\partial r_2}{\partial \theta} \right)}{r_2^4} \right] \end{aligned}$$

$$= \frac{\mu}{r} \left[\frac{r \sin \theta}{r_1^3} + \frac{3r \sin \theta (r \cos \theta - a)}{r_1^5} - \frac{r \sin \theta}{r_2^3} + 3 \frac{r \sin \theta (r \cos \theta + a)}{r_2^5} \right]$$

$$q_\psi = 0$$

When the point P lies on the plane YY' or $\theta = \pi/2$, we have $r_1^2 = r_2^2 = r^2 + a^2$ and so at $(r, \pi/2, \psi)$, the velocity components are

$$q_r = -6\mu ra/(r^2 + a^2)^{5/2}, q_\theta = 0, q_\psi = 0.$$

Along the streamline through this point, Bernoulli's equation is

$$\frac{p}{\rho} + \frac{1}{2} \bar{q}^2 = \text{const} = \frac{p_\infty}{\rho},$$

where $\bar{q} = 0$ at infinity.

Thus, the pressure at any point on the plane YY' is given by

$$p = p_\infty - \frac{1}{2} \rho \left[36\mu^2 a^2 r^2 / (r^2 + a^2)^5 \right]$$

i.e.
$$p(r) = p_\infty - \frac{18\rho\mu^2 a^2 r^2}{(r^2 + a^2)^5}$$

Now,

$$p'(r) = \frac{dp}{dr} = 36\rho\mu^2 a^2 r(4r^2 - a^2)/(r^2 + a^2)^6$$

which gives $p'(r) = 0$ when $r = \frac{1}{2}a$

Also

$$p'\left(\frac{a}{2} -\right) < 0, p'\left(\frac{a}{2} +\right) > 0$$

i.e. $p'(r)$ changes sign from negative to positive when r passes through $\frac{a}{2}$

$\Rightarrow p$ is minimum at $r = \frac{a}{2}$ $\theta = \pi/2$

i.e. at the point $P_0 \left(\frac{a}{2}, \pi/2, \psi \right)$

The distance P_0A is given by

$$\sqrt{\left(\frac{a}{2}\right)^2 + a^2} = \frac{\sqrt{5}}{2}a$$

Hence p is least at a distance $\frac{\sqrt{5}}{2}a$ from the doublet and the minimum value is

$$p_{\min.} = p_{\infty} - \frac{9}{2}\rho\mu^2\left(\frac{4}{5}\right)^5 \frac{1}{a^6}$$

5.3. Images in Impermeable Spherical Surfaces. We have already studied the effect of placing a solid impermeable sphere in a uniform stream of incompressible fluid, taking the case of axial symmetry. Here, we discuss the disturbance produced when a sphere is placed in more general flow.

We shall make use of Weiss's Sphere Theorem which states as follows :

“Let $\phi(r, \theta, \psi)$ be the velocity potential at a point P having spherical polar co-ordinates (r, θ, ψ) in an incompressible fluid having irrotational motion and no rigid boundaries. Also suppose that ϕ has no singularities within the region $r \leq a$. Then if a solid impermeable sphere of radius a is introduced into the flow with its centre at the origin of co-ordinates, the new velocity potential at P in the fluid is

$$\phi(r, \theta, \psi) + \frac{a}{r}\phi\left(\frac{a^2}{r}, \theta, \psi\right) - \frac{1}{a}\int_0^{a^2/r} \phi(R, \theta, \psi) dR, (r > a)$$

where r and $\frac{a^2}{r}$ are the inverse points w.r.t the sphere of radius a .”

Here, the last two terms refer to perturbation potential due to the presence of the sphere.

(i) Image of a Source in a Sphere : Suppose a source of strength m is situated at point A at a distance $f(> a)$ from the centre of the sphere of radius a .

7. Irrotational Motion in Two-dimensions

Suppose that a fluid moves in such a way that at any given instant, the flow pattern in a certain plane within the fluid is the same as that in all other parallel planes within the fluid. Then at the considered instant, the flow is said to be two-dimensional flow or plane flow. Any one of the parallel planes is then termed as plane of flow.

If we take the plane of flow as the plane $z = 0$, then at any point in the fluid having cartesian co-ordinates (x, y, z) , all physical quantities i.e. velocity, density, pressure etc, associated with the fluid are independent of z .

Thus $\bar{q} = \bar{q}(x, y, t)$ $\rho = \rho(x, y, t)$ etc

Plane flows, as described above, cannot be achieved in reality, but in certain important cases, close approximation to planarity of flow may occur.

We have already considered such flow when defining Lagrange's stream function. We consider here some special methods for treating two-dimensional irrotational motion.

7.1. Use of Cylindrical Polar Co-ordinates. For an incompressible irrotational flow of uniform density, the equation of continuity $\nabla^2\phi = 0$ for the velocity potential $\phi(r, \theta, z)$ in cylindrical polar co-ordinates (r, θ, z) is

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial \phi}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 \phi}{\partial \theta^2} + \frac{\partial^2 \phi}{\partial z^2} = 0 \quad (1)$$

If the flow is two dimensional and the co-ordinate axes are so chosen that all physical quantities associated with the fluid are independent of z , then $\phi = \phi(r, \theta)$ and (1) simplifies to

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial \phi}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 \phi}{\partial \theta^2} = 0 \quad (2)$$

Let us seek solutions of (2) by putting

$$\phi(r, \theta) = -f(r) g(\theta) \quad (3)$$

in (2) for separation of variables. Thus, we get

$$g(\theta) \frac{1}{r} \frac{d}{dr} [r f'(r)] + \frac{1}{r^2} f(r) g''(\theta) = 0$$

$$\text{i.e.} \quad \frac{r \frac{d}{dr}[rf'(r)]}{f(r)} = -\frac{g''(\theta)}{g(\theta)} \quad (4)$$

Thus, L.H.S. of (4) is a function of r only and R.H.S. is a function of θ only. As r and θ are independent variables, so each side of (4) is a constant λ (say). Thus, we have

$$\frac{r^2 f''(r) + r f'(r)}{f(r)} = \lambda; \quad \frac{g''(\theta)}{g(\theta)} = -\lambda$$

$$\text{i.e.} \quad r^2 f''(r) + r f'(r) - \lambda f(r) = 0 \quad (5)$$

$$\text{and} \quad g''(\theta) + \lambda g(\theta) = 0 \quad (6)$$

Equation (6) has periodic solutions when $\lambda > 0$. Normally the physical problem requires that $g(\theta + 2\pi) = g(\theta)$ and this is satisfied when $\lambda = n^2$ for $n = 1, 2, 3, \dots$

Thus, the basic solution of (6) are

$$g(\theta) = c_1 \cos n\theta + c_2 \sin n\theta \quad (7)$$

Now, (5) is of Euler-homogeneous type and it is reduced to a linear different equation of constant co-efficients by putting $r = e^t$ i.e. $t = \log r \Rightarrow \frac{dt}{dr} = \frac{1}{r}$

$$\text{Also,} \quad f'(r) = \frac{df}{dr} = \frac{df}{dt} \cdot \frac{dt}{dr} = \frac{1}{r} \frac{df}{dt}$$

$$\begin{aligned} \text{and} \quad f''(r) &= \frac{d^2 f}{dr^2} = \frac{d}{dr} \left(\frac{df}{dr} \right) = \frac{d}{dr} \left(\frac{1}{r} \frac{df}{dt} \right) \\ &= \frac{1}{r} \frac{d}{dr} \left(\frac{df}{dt} \right) + \frac{df}{dt} \left(-\frac{1}{r^2} \right) \\ &= \frac{1}{r} \left[\frac{d}{dt} \left(\frac{df}{dt} \right) \frac{dt}{dr} \right] - \frac{1}{r^2} \frac{df}{dt} \\ &= \frac{1}{r^2} \frac{d^2 f}{dt^2} - \frac{1}{r^2} \frac{df}{dt} \end{aligned}$$

$$\Rightarrow r^2 f''(r) = \frac{d^2 f}{dt^2} - \frac{df}{dt}$$

Therefore, equation (5) reduces to

$$\frac{d^2 f}{dt^2} - \frac{df}{dt} + \frac{df}{dt} - n^2 f = 0$$

$$\Rightarrow \frac{d^2 f}{dt^2} - n^2 f = 0$$

It's solution is

$$f = \exp(\pm nt) = e^{\pm nt} = (e^t)^{\pm n} = r^{\pm n}$$

$$\text{i.e. } f = c_3 r^n + c_4 r^{-n} \quad (8)$$

A special solution of (2) is obtained by linear superposition of the forms (7) & (8) to give

$$\begin{aligned} \phi(r, \theta) &= -f(r) g(\theta) \\ &= -(A_n r^n + B_n r^{-n}) (C_n \cos n\theta + D_n \sin n\theta) \end{aligned} \quad (9)$$

The most general solution is of the form

$$\phi(r, \theta) = - \sum_{n=1}^{\infty} (A_n r^n + B_n r^{-n}) (C_n \cos n\theta + D_n \sin n\theta) \quad (10)$$

7.2. Particular cases. (i) for $n = 0$, we have

$$f = k_1 + k_2 t = k_1 + k_2 \log r$$

$$\text{and } g = k_3 + k_4 \theta$$

so that another solution of (2) is

$$\phi(r, \theta) = -(k_1 + k_2 \log r) (k_3 + k_4 \theta)$$

(ii) for $n = 1$, we get a special solution as

$$\phi = -r \cos \theta, \phi = -r \sin \theta, \phi = -r^{-1} \cos \theta, \phi = -r^{-1} \sin \theta$$

7.3. Example. Discuss the uniform flow past an infinitely long circular cylinder.

Solution. Let P be a point with cylindrical polar co-ordinates (r, θ, z) in the flow region of an unbounded incompressible fluid of uniform density moving irrotationally with uniform velocity $-U\hat{i}$ at infinity past the fixed solid cylinder $r \leq a$

When the cylinder $r = a$ is introduced, it will produce a perturbation which is such as to satisfy Laplace equation and to become vanishingly small for large r . This suggests taking the velocity potential for $r > a$, $0 \leq \theta \leq 2\pi$ in the form

$$\phi(r, \theta) = Ur \cos\theta - Ar^{-1} \cos\theta, \quad (1)$$

where the velocity potential of the uniform stream is $Ux = Ur \cos\theta$ and due to perturbation, it is $-Ar^{-1} \cos\theta$ which $\rightarrow 0$ as $r \rightarrow \infty$ and gives rise to a velocity pattern which is symmetrical about $\theta = 0, \pi$. (the term $r^{-1} \sin\theta$ is not there since it does not give symmetric flow)

As there is no flow across $r = a$, so the boundary condition on the surface is

$$\frac{\partial\phi}{\partial r} = 0, \text{ when } r = a \quad (2)$$

Applying (2) in (1), we get $A = -Ua^2$ for all θ satisfying $0 \leq \theta \leq 2\pi$.

Thus, the velocity potential for a uniform flow past a fixed infinite cylinder is

$$\phi(r, \theta) = U \cos\theta \left(r + \frac{a^2}{r} \right), \quad r > a, \quad 0 \leq \theta \leq 2\pi \quad (3)$$

From here, the cylindrical components of velocity are ($\bar{q} = -\nabla\phi$)

$$q_r = -\frac{\partial\phi}{\partial r} = -U \cos\theta \left(1 - \frac{a^2}{r^2} \right)$$

$$q_\theta = -\frac{1}{r} \frac{\partial\phi}{\partial\theta} = \frac{1}{r} U \sin\theta \left(r + \frac{a^2}{r} \right) = U \sin\theta \left(1 + \frac{a^2}{r^2} \right)$$

$$q_z = -\frac{\partial\phi}{\partial z} = 0$$

We note that as $r \rightarrow \infty$, $q_r = -U \cos\theta$, $q_\theta = U \sin\theta$ which are consistent with the velocity at infinity $-U\hat{i}$ of the uniform stream.

8. The Complex Potential

Here, we confine our attention to irrotational plane flows of incompressible fluid of uniform density for which the velocity potential $\phi(x, y)$ and the stream function $\psi(x, y)$ exist. Here (x, y) specify two dimensional Cartesian co-ordinates in a plane of flow. Let us write

$$W = \phi(x, y) + i\psi(x, y) \quad (1)$$

We suppose that all four first-order partial derivatives of ϕ & ψ with respect to x, y exist and are continuous throughout the plane of flow. Now, the velocity $\bar{q} = (u, v)$ has components satisfying $\bar{q} = -\nabla\phi$.

$$u = -\frac{\partial\phi}{\partial x} = -\frac{\partial\psi}{\partial y}, \quad v = -\frac{\partial\phi}{\partial y} = \frac{\partial\psi}{\partial x} \quad (2)$$

Thus ϕ and ψ satisfy the C-R equations and so W must be an analytic function of $z = x + iy$

Therefore, we can write (1) as

$$W = f(z) = \phi + i\psi \quad (3)$$

The function $W = f(z)$ is called the **complex potential** of the plane flow.

8.1. Complex Velocity. We have

$$W = \phi + i\psi \text{ and } z = x + iy$$

Differentiating partially w.r.t. x , we get

$$\frac{\partial W}{\partial x} = \frac{\partial\phi}{\partial x} + i \frac{\partial\psi}{\partial x} = \frac{\partial\phi}{\partial x} - i \frac{\partial\phi}{\partial y} = -u + iv$$

$$\text{But } \frac{\partial W}{\partial x} = \frac{dW}{dz} \cdot \frac{\partial z}{\partial x} = \frac{dW}{dz} \quad \left| \because \frac{\partial z}{\partial x} = 1 \right.$$

$$\text{Thus, } \frac{dW}{dz} = -u + iv$$

$$\begin{aligned} \Rightarrow -\frac{dW}{dz} &= u - iv = q \cos\theta - iq \sin\theta \\ &= q(\cos\theta - i \sin\theta) = q e^{-i\theta} \end{aligned}$$

The combination $u - iv$ is known as **complex velocity**

$$\text{Thus, speed } q = \left| -\frac{dW}{dz} \right| = \sqrt{u^2 + v^2}$$

$$\text{and for stagnation points, } \frac{dW}{dz} = 0$$

8.2. Example. Discuss the flow for which complex potential is

$$W = z^2$$

Solution. We have

$$W = \phi + i\psi = z^2 = (x + iy)^2 = x^2 - y^2 + 2i xy$$

$$\Rightarrow \phi(x, y) = x^2 - y^2, \quad \psi(x, y) = 2xy$$

The equipotentials, $\phi = \text{constant}$, are the rectangular hyperbolae $x^2 - y^2 = \text{constant}$ having asymptotes $y = \pm x$.

The streamlines, $\psi = \text{constant}$, are the rectangular hyperbolae $xy = \text{constant}$ having the axes $x = 0$, $y = 0$ as asymptotes. Also $\frac{dW}{dz} = 2z$, therefore the only stagnation point is the origin. The two families of the hyperbolae cut orthogonally in accordance with general theory.

8.3. Complex Potential for a Uniform Stream. Let the uniform stream advance with a velocity having magnitude U and being inclined at angle α to the positive direction of the x -axis.

Then, we have $u = U \cos \alpha$, $v = U \sin \alpha$ and thus

$$-\frac{dW}{dz} = u - iv = U e^{-i\alpha}$$

The simplest form of W , ignoring the constant of integration, is

$$W = -Uz e^{-i\alpha}$$

$$\begin{aligned} \text{i.e.} \quad \phi + i\psi &= -U(x + iy)(\cos \alpha - i \sin \alpha) \\ &= -U(x \cos \alpha + y \sin \alpha) - U i (y \cos \alpha - x \sin \alpha) \end{aligned}$$

Equating real and imaginary parts, we get

$$\phi = -U(x \cos \alpha + y \sin \alpha)$$

$$\psi = -U(y \cos \alpha - x \sin \alpha)$$

Thus, the equations of equipotentials are

$$x \cos \alpha + y \sin \alpha = \text{constant} \quad (1)$$

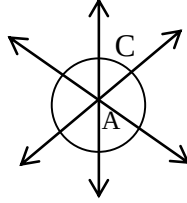
These equations represent a family of parallel streamlines. The equations of the streamlines are

$$y \cos \alpha - x \sin \alpha = \text{constant} \quad (2)$$

These equations represent another family of parallel streamlines inclined at angle α to the positive x-direction. The two family of streamlines intersect orthogonally in accordance with general theory.

8.4. Line Source and Line Sink. Line source and line sink are the two-dimensional analogues of the three-dimensional simple source and sink. Let A be any point of the considered plane of flow and C be any closed curve surrounding A. We construct a cylinder having its generators through the points of C and normal to the plane of flow. Suppose that in each plane of flow, fluid is emitted radially and symmetrically from all points on the infinite line through A normal to the plane of flow and such that the rate of emission from all such points as A is the same. Then the line through A is called a **line source**. We may take the closed curve C to be a circle having centre A and radius r.

Suppose the line source emits fluid at the rate $2\pi m\rho$ units of mass per unit length of the source per unit time, in all directions in the plane of flow (say, xy-plane). We define the strength of the line source to be m. A line source of strength $-m$ is called a line sink.



An example of a line source is a long straight hose with perforations along its length, commonly used for watering lawns for long periods of time.

8.5. Complex Potential for a Line Source. Let there be a line source of strength m per unit length at $z = 0$. Since the flow is radial, the velocity has the radial component q_r only. Then the flow across a circle of radius r is (by law of conservation of mass)

$$(2\pi r q_r)\rho = 2\pi m\rho$$

$$\Rightarrow q_r = \frac{m}{r}$$

The complex potential is obtained from the relation

$$\begin{aligned} -\frac{dW}{dz} &= u - iv = q_r \cos \theta - i q_r \sin \theta \\ &= q_r (\cos \theta - i \sin \theta) = \frac{m}{r} e^{-i\theta} \end{aligned}$$

$$\Rightarrow \frac{dW}{dz} = -\frac{m}{r} e^{-i\theta} = \frac{-m}{re^{i\theta}} = \frac{-m}{z}$$

Integrating, we get

$$W = -m \log z$$

where we have ignored the constant of integration.

We can write it as

$$\begin{aligned} \phi + i\psi &= -m \log (r e^{i\theta}) \\ &= -m \log r - i m \theta \end{aligned}$$

$$\Rightarrow \phi = -m \log r, \psi = -m\theta$$

Thus, the equipotentials and streamlines have the respective forms

$$r = \text{constant}, \theta = \text{constant}$$

$$\text{i.e. } x^2 + y^2 = \text{constant}, \tan^{-1} \frac{y}{x} = \text{constant}$$

$$\text{i.e. } x^2 + y^2 = C_1, y = C_2 x.$$

Thus the equipotentials are circles and streamlines are straight lines passing through origin.

If the line source is at $z = z_0$ instead of $z = 0$, then the complex potential is

$$W = -m \log (z - z_0)$$

For a line sink of strength m per unit length at $z = z_0$, the complex potential is

$$W = m \log (z - z_0).$$

If there are a number of line sources at $z = z_1, z_2, \dots, z_n$ of respective strengths $m_1, m_2, \dots, m_n$ per unit length, then the complex potential is

$$W = -m_1 \log(z - z_1) - m_2 \log(z - z_2) - \dots - m_n \log(z - z_n).$$

8.6. Complex Potential for a Line Doublet. The combination of a line source and a line sink of equal strength when placed close to each other gives a line doublet. Let us take a line source of strength m per unit length at $z = a e^{i\alpha}$ and a line sink of strength m per unit length at $z = 0$

Therefore, the complex potential due to the combination is

$$\begin{aligned}
W &= -m \log(z - ae^{i\alpha}) + m \log(z-0) \\
&= -m \log \frac{z - ae^{i\alpha}}{z} = -m \log \left(1 - \frac{ae^{i\alpha}}{z} \right) \\
&= m \left[\frac{ae^{i\alpha}}{z} + \frac{a^2 e^{2i\alpha}}{2z^2} + \frac{a^3 e^{3i\alpha}}{3z^3} \dots \right]
\end{aligned}$$

In the figure,

OP = a = δs where a is the distance between the source and sink.

As $a \rightarrow 0$, $m \rightarrow \infty$ so that $ma \rightarrow \mu$ and thus, we get

$$W = \frac{\mu e^{i\alpha}}{z}$$

If the line sink is situated at $z = z_0$, then the complex potential is

$$W = \frac{\mu e^{i\alpha}}{z - z_0}$$

If $\alpha = 0$, then the line source is on x-axis and thus.

$$W = \frac{\mu}{z - z_0}$$

If there are number of line doublets of strengths $\mu_1, \mu_2, \dots, \mu_n$ per unit length with line sinks at points $z_1, z_2, \dots, z_n$ and their axis being inclined at angles $\alpha_1, \alpha_2, \dots, \alpha_n$ with the positive direction of x-axis, then the complex potential is given by

$$W = \mu_1 \frac{e^{i\alpha_1}}{z - z_1} + \mu_2 \frac{e^{i\alpha_2}}{z - z_2} + \dots + \mu_n \frac{e^{i\alpha_n}}{z - z_n}$$

8.7. Example. Discuss the flow due to a uniform line doublet at origin of strength μ per unit length and its axis being along the x-axis.

Solution. We know that the complex potential for a doublet is

$$W = \frac{\mu e^{i\alpha}}{z - z_0}$$

and when the doublet is at origin having its axis along x-axis, then $\alpha = 0$, $z_0 = 0$

$$\therefore W = \frac{\mu}{z} = \frac{\mu}{x + iy} = \frac{\mu(x - iy)}{x^2 + y^2}$$

$$\Rightarrow \phi + i\psi = \frac{\mu x}{x^2 + y^2} - i \frac{\mu y}{x^2 + y^2}$$

$$\Rightarrow \phi = \frac{\mu x}{x^2 + y^2}, \quad \psi = \frac{-\mu y}{x^2 + y^2}$$

Thus the equipotentials, $\phi = \text{constant}$, are the coaxial circles

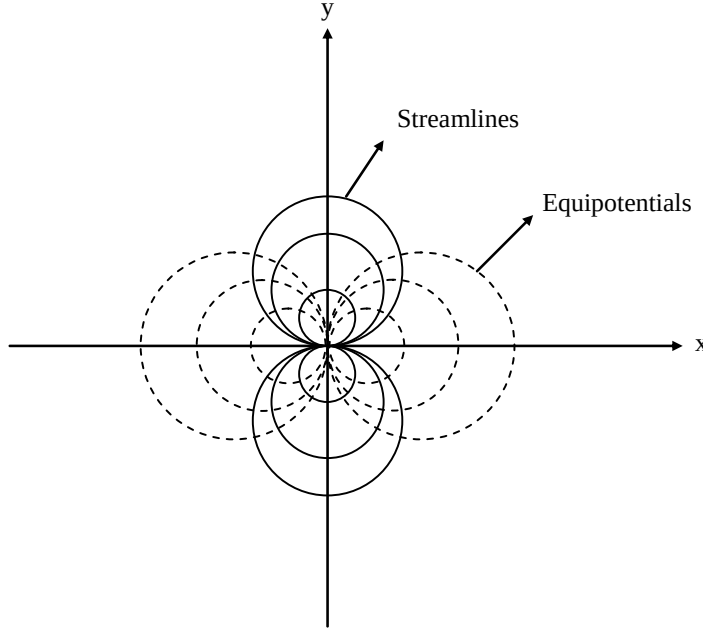
$$x^2 + y^2 = 2k_1 x \quad (1)$$

and the streamlines, $\psi = \text{constant}$, are the coaxial circles

$$x^2 + y^2 = 2k_2y \quad (2)$$

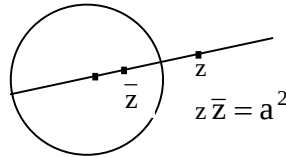
Family (1) have centres $(k_1, 0)$ and radii k_1 and family (2) have centres $(0, k_2)$ and radii k_2

The two families are orthogonal



8.8. Milne-Thomson Circle Theorem : Let $f(z)$ be the complex potential for a flow having no rigid boundaries and such that there are no singularities within the circle $|z| = a$. Then on introducing the solid circular cylinder $|z| = a$, with impermeable boundary, into the flow, the new complex potential for the fluid outside the cylinder is given by

$$W = f(z) + \bar{f}(a^2/z), \quad |z| \geq a$$



Proof. Let C be the **cross-section** of the cylinder with equation $|z| = a$.

Therefore, on the circle C , $|z| = a \Rightarrow z \bar{z} = a^2 \Rightarrow \bar{z} = a^2/z$

where $\bar{z}$ is the image of the point z w.r.t. the circle. If z is outside the circle, then $\bar{z} = a^2/z$ is inside the circle. Further, all the singularities of $f(z)$ lie outside C and the singularities of $f(a^2/z)$ and therefore those of $\bar{f}(a^2/z)$ lie inside C . Therefore $\bar{f}(a^2/z)$ introduces no singularity outside the cylinder. Thus, the functions $f(z)$ and $f(z) + \bar{f}(a^2/z)$ both have the same singularities outside C . Therefore the conditions satisfied by $f(z)$ in the absence of the cylinder are satisfied by $f(z) + \bar{f}(a^2/z)$ in the presence of the cylinder. Further, the complex potential, after insertion of the cylinder $|z| = a$, is

$$\begin{aligned}
W &= f(z) + \bar{f}(a^2/z) = f(z) + \bar{f}(\bar{z}) \\
&= f(z) + \overline{f(z)} \\
&= \text{a purely real quantity}
\end{aligned}$$

But we know that $W = \phi + i\psi$

It follows that $\psi = 0$

This proves that the circular cylinder $|z| = a$ is a streamline i.e. C is a streamline. Therefore, the new complex potential justifies the fluid motion and hence the circle theorem.

8.9. Uniform Flow Past a Fixed Infinite Circular Cylinder. We have already dealt with this problem using cylindrical polar co-ordinates. Here, we use the concept of complex potential.

The velocity potential due to an undisturbed uniform stream having velocity $-U\hat{i}$ (U is real) is $Ux = U \operatorname{Re}(z)$.

Since z is an analytic function, the corresponding complex potential is

$$f(z) = Uz$$

Thus

$$\bar{f}(\bar{z}) = \overline{f(z)} = \overline{Uz} = \bar{U}\bar{z} = U\bar{z}$$

and so

$$\bar{f}(a^2/z) = Ua^2/z.$$

With the cylinder $|z| = a$ present, by circle theorem, the complex potential, for the liquid region $|z| \geq a$, is

$$W = f(z) + \bar{f}(a^2/z)$$

$$\text{i.e. } \phi + i\psi = U \left(z + \frac{a^2}{z} \right)$$

Taking $z = re^{i\theta}$, where $r \geq a$, equating real and imaginary parts, we get

$$\phi = \operatorname{Re}(W) = U \cos\theta \left(r + \frac{a^2}{r} \right) \quad | \text{ Same expression as derived earlier}$$

10. Blasius Theorem

In a steady two dimensional irrotational flow given by the complex potential $W = f(z)$, if the pressure forces on the fixed cylindrical surface C are represented by a force (X, Y) and a couple of moment M about the origin of co-ordinates, then neglecting the external forces,

$$X - iY = \frac{i\rho}{2} \int_C \left(\frac{dW}{dz} \right)^2 dz$$

$$M = \text{Real part of} \left[-\frac{\rho}{2} \int_C z \left(\frac{dW}{dz} \right)^2 dz \right]$$

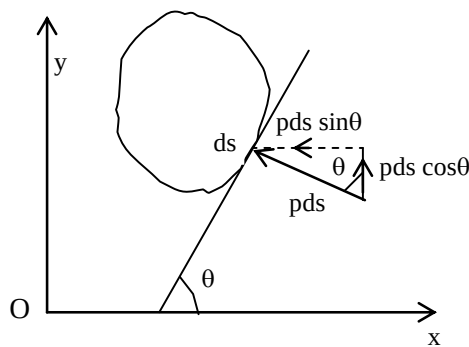
where ρ is the density of the fluid

Proof. Let ds be an element of arc at a point $P(x, y)$ and the tangent at p makes an angle θ with the x -axis. The pressure at $P(x, y)$ is pds , p is the pressure per unit length. pds acts along the inward normal to the cylindrical surface and its components along the co-ordinate axes are

$$pds \cos (90 + \theta), \quad pds \cos \theta$$

$$\text{i.e.} \quad -pds \sin \theta, \quad pds \cos \theta$$

The pressure at the element ds is



$$dF = dX + idY$$

$$= -p \sin \theta ds + ip \cos \theta ds$$

$$= ip (\cos \theta + i \sin \theta) ds$$

$$\left| \begin{array}{l} pds \sin \theta \text{ along negative } x - \text{axis} \\ \Rightarrow -pds \sin \theta \text{ along positive } x - \text{axis} \end{array} \right.$$

$$= ip \left(\frac{dx}{ds} + i \frac{dy}{ds} \right) ds \quad \left| \cos \theta = \frac{dx}{ds}, \quad \sin \theta = \frac{dy}{ds} \right.$$

$$= ip (dx + idy) = ip dz \quad (1)$$

The pressure equation, in the absence of external forces, is

$$\frac{p}{\rho} + \frac{1}{2} q^2 = \text{constant}$$

$$\text{or} \quad p = -\frac{1}{2} \rho q^2 + k \quad (2)$$

$$\text{Further} \quad \frac{dW}{dz} = -u + iv = -q \cos \theta + iq \sin \theta$$

$$= -q (\cos\theta - i \sin\theta) = -q e^{-i\theta} \quad (3)$$

$$\text{and } dz = dx + i dy = \left(\frac{dx}{ds} + i \frac{dy}{ds} \right) ds = (\cos\theta + i \sin\theta) ds = e^{i\theta} ds \quad (4)$$

The pressure on the cylinder is obtained by integrating (1). Therefore,

$$\begin{aligned} F = X + iY &= \int_C i p dz = \int_C i (k - 1/2 \rho q^2) dz \\ &= -\frac{i\rho}{2} \int_C q^2 dz \quad | \because \int_C dz = 0 \\ &= -\frac{i\rho}{2} \int_C q^2 e^{i\theta} ds \end{aligned}$$

From here ;

$$\begin{aligned} X - iY &= \frac{i\rho}{2} \int_C q^2 e^{-i\theta} ds \\ &= \frac{i\rho}{2} \int_C (q^2 e^{-2i\theta}) e^{i\theta} ds \\ &= \frac{i\rho}{2} \int_C \left(\frac{dW}{dz} \right)^2 dz \quad | \text{ using (3) \& (4)} \end{aligned}$$

The moment M is given by

$$\begin{aligned} M &= \int_C |\vec{r} \times d\vec{F}| = \int_C [(p ds \sin\theta) y + (p ds \cos\theta) x] \\ &= \int_C \left[p \left(\frac{dy}{ds} \right) y ds + p \left(\frac{dx}{ds} \right) x ds \right] \\ &= \begin{vmatrix} \vec{r} \times d\vec{F} \\ \hat{i} & \hat{j} & \hat{k} \\ x & y & 0 \\ -pd \sin\theta & pd \cos\theta & 0 \end{vmatrix} \\ &= \int_C p(x dx + y dy) \\ &= \int_C \left(k - \frac{1}{2} \rho q^2 \right) (x dx + y dy) \end{aligned}$$

$$= k \int_C d \left[\frac{1}{2} (x^2 + y^2) \right] - \frac{\rho}{2} \int_C q^2 (x dx + y dy)$$

$$= -\frac{\rho}{2} \int_C q^2 (x dx + y dy) \quad | \quad \because \text{1st integral}$$

vanishes.

$$= \frac{-\rho}{2} \int_C q^2 (x \cos \theta + y \sin \theta) ds \quad \left| \begin{array}{l} dx = \cos \theta ds \\ dy = \sin \theta ds \end{array} \right.$$

$$= \text{R.P. of} \left[\frac{-\rho}{2} \int_C q^2 (x + iy)(\cos \theta - i \sin \theta) ds \right]$$

$$= \text{R.P. of} \left[\frac{-\rho}{2} \int_C q^2 z e^{-i\theta} ds \right]$$

$$= \text{R.P. of} \left[\frac{-\rho}{2} \int_C z (q^2 e^{-2i\theta}) e^{i\theta} ds \right]$$

$$= \text{R.P. of} \left[-\frac{\rho}{2} \int_C z \left(\frac{dW}{dz} \right)^2 dz \right].$$

Hence the theorem.

UNIT – V & VI

1. Three Dimensional Irrotational Flow

1.1. Acyclic and Cyclic Irrotational Motion. An irrotational motion is called acyclic if the velocity potential ϕ is a single valued function i.e. when at every field point, a unique velocity potential exists, otherwise the irrotational motion is said to be cyclic. Clearly, only acyclic irrotational motion is possible in a simply connected region.

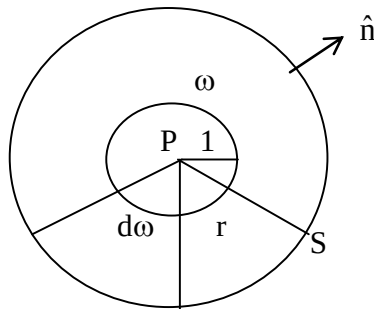
For a possible fluid motion, even if ϕ is multivalued at a particular point, the velocity at that point must be single-valued. Hence if we obtain two different values of ϕ , these values can only differ by a constant.

At present, we restrict ourselves to acyclic irrotational motion for which we prove a number of results related to ϕ .

1.2. Mean Value of Velocity Potential Over Spherical Surfaces. Theorem :

The mean value of a ϕ over any spherical surface S drawn in the fluid throughout whose interior $\nabla^2\phi = 0$, is equal to the value of ϕ at the centre of the sphere.

Proof. Let $\phi(P)$ be the value of ϕ at the centre P of a spherical surface S of radius r , wholly lying in the liquid and let $\bar{\phi}$ denotes the mean value of ϕ over S . Let us draw another concentric sphere ω of unit radius. Then a cone with vertex P which intercepts area dS from the sphere S , intercepts an area $d\omega$ from the sphere ω and we have



$$\frac{dS}{d\omega} = \frac{r^2}{1^2} \Rightarrow dS = r^2 d\omega \quad (1)$$

Now, by definition

$$\begin{aligned}
 \bar{\phi} &= \frac{\int_S \phi dS}{\int_S dS} = \frac{1}{4\pi r^2} \int_S \phi dS \\
 &= \frac{1}{4\pi r^2} \int_S \phi r^2 d\omega = \frac{1}{4\pi} \int_S \phi d\omega \\
 \Rightarrow \quad \frac{\partial \bar{\phi}}{\partial r} &= \frac{1}{4\pi} \int_S \frac{\partial \phi}{\partial r} d\omega = \frac{1}{4\pi} \int_S \frac{\partial \phi}{\partial r} \frac{dS}{r^2} \\
 &= \frac{1}{4\pi r^2} \int_S \frac{\partial \phi}{\partial r} dS \quad (2) \\
 &\quad | \because r^2 \text{ is constant on } S
 \end{aligned}$$

Since the normal $\hat{n}$ to the surface is along the radius r , therefore on S , we have

$$\frac{\partial \phi}{\partial r} = \frac{\partial \phi}{\partial n} = \nabla \phi \cdot \hat{n} \quad (3)$$

From (2) & (3), we find

$$\begin{aligned}
 \frac{\partial \bar{\phi}}{\partial r} &= \frac{1}{4\pi r^2} \int_S \nabla \phi \cdot \hat{n} dS \\
 &= \frac{1}{4\pi r^2} \int_{\tau} \text{div}(\nabla \phi) d\tau \quad | \text{ Gauss theorem} \\
 &= \frac{1}{4\pi r^2} \int_{\tau} \nabla^2 \phi d\tau = 0, \quad | \nabla^2 \phi = 0
 \end{aligned}$$

where τ is the volume enclosed by the surface S .

Thus $\frac{\partial \bar{\phi}}{\partial r} = 0 \Rightarrow \bar{\phi} = \text{constant}$.

This shows that $\bar{\phi}$ is independent of choice of r and hence mean value of ϕ is same over all spherical surfaces having the same centre P . When S shrinks to point P , then $\bar{\phi} = \phi(P)$

1.3. Corollary. The velocity potential ϕ can not have a maximum or minimum value in the interior of any region throughout which $\nabla^2 \phi = 0$.

Proof. If possible, suppose that ϕ has a maximum value $\phi(P)$ at a point P . We draw a sphere with centre P and radius ϵ , where ϵ is small. Then the mean value $\bar{\phi}$ of ϕ must be less than $\phi(P)$ i.e. $\bar{\phi} < \phi(P)$ as $\phi(P)$ is maximum. This

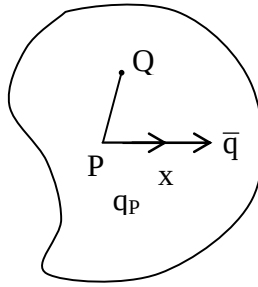
is a contradiction to the mean potential theorem in which $\bar{\phi} = \phi(P)$. Thus ϕ cannot have a maximum value. Similarly ϕ cannot have a minimum value.

1.4. Theorem. In an irrotational motion the maximum value of the fluid velocity occurs at the boundary.

Proof. Let P be any interior point of the fluid and Q be a neighbouring point also lying in the fluid. Let us take the direction of x-axis along the direction of $\bar{q}$ at P. Let q_P and q_Q denote the speed of particles at P & Q respectively.

Then
$$q_P^2 = \left(\frac{\partial \phi}{\partial x} \right)_P^2$$

and
$$q_Q^2 = \left(\frac{\partial \phi}{\partial x} \right)_Q^2 + \left(\frac{\partial \phi}{\partial y} \right)_Q^2 + \left(\frac{\partial \phi}{\partial z} \right)_Q^2$$



Since $\nabla^2 \phi = 0 \Rightarrow \frac{\partial}{\partial x} (\nabla^2 \phi) = 0 \Rightarrow \nabla^2 \left(\frac{\partial \phi}{\partial x} \right) = 0$

$\Rightarrow \frac{\partial \phi}{\partial x}$ satisfies Laplace equation. Therefore, by mean value theorem

(corollary), $\frac{\partial \phi}{\partial x}$ cannot be maximum or minimum at P. Thus, there are points such as Q in the neighbourhood of P such that

$$\left(\frac{\partial \phi}{\partial x} \right)_Q^2 > \left(\frac{\partial \phi}{\partial x} \right)_P^2 \Rightarrow q_Q^2 > q_P^2$$

$\Rightarrow q_P$ cannot be maximum in the interior of fluid and its maximum value $|\bar{q}|$, if any, must therefore occur on the boundary.

1.5. Note. $q = |\bar{q}|$ may be minimum in the interior of the fluid as $\bar{q} = \bar{0}$ at the stagnation point. i.e. q is minimum at stagnation points.

1.6. Corollary. In steady irrotational flow, the pressure has its minimum value on the boundary.

Proof. From Bernoulli's equation, we have

$$\frac{p}{\rho} + \frac{1}{2}q^2 = \text{constant} \quad (1)$$

Equation (1) shows that p is least when q^2 is greatest and by above theorem, q^2 is greatest at the boundary. Thus, the minimum value of p must occur only on the boundary.

1.7. Note. The maximum value of p occurs at the stagnation points, where $\bar{q} = \bar{0}$.

1.8. Theorem. If liquid of infinite extent is in irrotational motion and is bounded internally by one or more closed surfaces S , the mean value of ϕ over a large sphere Σ , of radius R , which encloses S , is of the form

$$\bar{\phi} = \frac{M}{R} + C$$

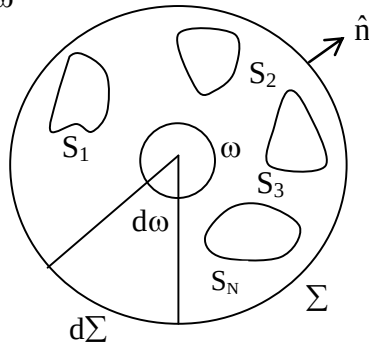
where M and C are constants, provided that the liquid is at rest at infinity.

Proof. Suppose that the volume of fluid acrossing each of internal surfaces contained within Σ , per unit time, is a finite quantity say $-4\pi M$ (i.e. $-4\pi M$ represents the flux of fluid across Σ or S). Since the fluid velocity at any point of Σ is $\frac{\partial \phi}{\partial R}$ radially outwards, the equation of continuity gives

$$\int_{\Sigma} \frac{\partial \phi}{\partial R} d\Sigma = -4\pi M \quad (1)$$

But

$$d\Sigma = R^2 d\omega$$



Therefore,

$$\begin{aligned} \frac{1}{4\pi} \int_{\Sigma} \frac{\partial \phi}{\partial R} R^2 d\omega &= -M \\ \Rightarrow \frac{1}{4\pi} \int_{\Sigma} \frac{\partial \phi}{\partial R} d\omega &= \frac{-M}{R^2} \\ \Rightarrow \frac{1}{4\pi} \frac{\partial}{\partial R} \int_{\Sigma} \phi d\omega &= \frac{-M}{R^2} \end{aligned}$$

Integrating w.r.t. R , we get

$$\frac{1}{4\pi} \int_{\Sigma} \phi d\omega = \frac{M}{R} + C$$

where C is independent of R .

$$\begin{aligned} \Rightarrow \frac{1}{4\pi} \int_{\Sigma} \phi \left(\frac{d\Sigma}{R^2} \right) &= \frac{M}{R} + C \\ \Rightarrow \frac{\int_{\Sigma} \phi d\Sigma}{4\pi R^2} &= \frac{M}{R} + C \\ \Rightarrow \bar{\phi} &= \frac{M}{R} + C \end{aligned} \tag{2}$$

To show that C is an absolute constant, we have to prove that it is independent of co-ordinates of centre of sphere Σ . Let the centre of the sphere Σ be displaced by distance δx in an arbitrary direction while keeping R constant, then from (2),

$$\frac{\partial \bar{\phi}}{\partial x} = \frac{\partial C}{\partial x} \tag{3}$$

| $\because R$ is constant

Also,
$$\frac{\partial \bar{\phi}}{\partial x} = \frac{\partial}{\partial x} \left[\frac{1}{4\pi} \int_{\Sigma} \phi d\omega \right] = \frac{1}{4\pi} \int_{\Sigma} \frac{\partial \phi}{\partial x} d\omega$$

$$= 0, \text{ since } \frac{\partial \phi}{\partial x} = 0 \text{ on } \Sigma \text{ when } R \rightarrow \infty \text{ as the liquid is at rest at}$$

infinity.

∴ From (3), we get

$$\frac{\partial C}{\partial x} = 0 \Rightarrow C \text{ is an absolute constant.}$$

Hence

$$\bar{\phi} = \frac{M}{R} + C, \text{ where } M \text{ and } C \text{ are constants.}$$

1.9. Corollary. When closed surfaces within Σ are rigid then no flow can take place across them, therefore, in that case $M = 0$ and $\bar{\phi} = C$.

This shows that mean value of ϕ over any sphere enclosing solid rigid boundaries is constant.

2. Kinetic Energy of Irrotational Flow

We shall prove that K.E. is given by

$$T = \frac{\rho}{2} \int_S \phi \frac{\partial \phi}{\partial n} dS,$$

where ϕ is the velocity potential.

We know that if τ be the finite region occupied by the fluid, then the K.E. is given by

$$\begin{aligned} T &= \frac{1}{2} \int_{\tau} \rho \bar{q}^2 d\tau = \frac{1}{2} \int_{\tau} \rho (\bar{q} \cdot \bar{q}) d\tau \\ &= \frac{1}{2} \int_{\tau} \rho (\nabla \phi \cdot \nabla \phi) d\tau \quad | \quad \bar{q} = -\nabla \phi \end{aligned}$$

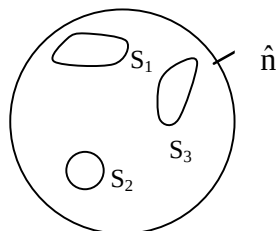
If fluid density is constant, then

$$T = \frac{\rho}{2} \int_{\tau} (\nabla \phi \cdot \nabla \phi) d\tau \quad (1)$$

$$\text{Now, } \text{div} (\phi \nabla \phi) = \nabla \cdot (\phi \nabla \phi) = \nabla \phi \cdot \nabla \phi + \phi (\nabla \cdot \nabla \phi)$$

$$= \nabla \phi \cdot \nabla \phi + \phi \nabla^2 \phi$$

$$= \nabla \phi \cdot \nabla \phi \quad | \quad \because \nabla^2 \phi = 0$$



Therefore, from (1) & (2), we get

$$\begin{aligned} T &= \frac{\rho}{2} \int_{\tau} \operatorname{div}(\phi \nabla \phi) d\tau = \frac{\rho}{2} \int_S \phi \nabla \phi \cdot \hat{n} dS \quad | \text{ By Gauss theorem} \\ &= \frac{\rho}{2} \int_S \phi (\nabla \phi \cdot \hat{n}) dS = \frac{\rho}{2} \int_S \phi \frac{\partial \phi}{\partial n} dS, \text{ where } S = S_0 + S_1 + S_2 + \dots + S_n \end{aligned}$$

denotes the sum of the outer boundary surface S_0 and the inverse boundaries $S_1, S_2, \dots, S_n$ and $\hat{n}$ is unit normal to S drawn out of the fluid on each boundary.

Also $T = -\frac{\rho}{2} \int_S \phi \frac{\partial \phi}{\partial n} dS$, where $\hat{n}$ is unit normal to S drawn inside the fluid on each boundary.

2.1. Kelvin's Minimum Energy Theorem. The kinetic energy of irrotational motion of a liquid occupying a finite simply connected region is less than that of any other motion of the liquid which is consistent with the same normal velocity of the boundary.

Proof. Let T be the K.E., $\bar{q}$ be the fluid velocity and ϕ be the velocity potential of the given irrotational motion. Let τ be the region occupied by the fluid and S be the surface of this region, then

$$\begin{aligned} T &= \frac{\rho}{2} \int_{\tau} \bar{q}^2 d\tau = \frac{\rho}{2} \int_{\tau} (-\nabla \phi)^2 d\tau \\ &= \frac{\rho}{2} \int_S \phi \frac{\partial \phi}{\partial n} dS \end{aligned} \quad (1)$$

Let T_1 be the K.E. and $\bar{q}_1$ be the velocity of any other motion of the fluid consistent with the same normal velocity of the boundary S (or consistent with the same kinetic boundary condition)

For both the motions, the continuity equation is satisfied i.e.

$$\nabla \cdot \bar{q} = 0 = \nabla \cdot \bar{q}_1 \quad (2)$$

The boundaries have the same normal velocity

i.e. $\bar{q} \cdot \hat{n} = \bar{q}_1 \cdot \hat{n}$

$$\text{i.e.} \quad (\bar{q}_1 - \bar{q}) \cdot \hat{n} = 0 \quad (3)$$

Now, let us consider

$$\begin{aligned} T_1 - T &= \frac{\rho}{2} \int_{\tau} (q_1^2 - q^2) d\tau \\ &= \frac{\rho}{2} \int_{\tau} [2\bar{q} \cdot (\bar{q}_1 - \bar{q}) + (\bar{q}_1 - \bar{q})^2] d\tau \\ &= \frac{\rho}{2} \int_{\tau} 2\bar{q} \cdot (\bar{q}_1 - \bar{q}) d\tau + \frac{\rho}{2} \int_{\tau} (\bar{q}_1 - \bar{q})^2 d\tau \\ &= -\rho \int_{\tau} \nabla \phi \cdot (\bar{q}_1 - \bar{q}) d\tau + \frac{\rho}{2} \int_{\tau} (\bar{q}_1 - \bar{q})^2 d\tau \quad (4) \end{aligned}$$

From vector calculus, we have

$$\nabla \cdot [\phi (\bar{q}_1 - \bar{q})] = \nabla \phi \cdot (\bar{q}_1 - \bar{q}) + \phi \nabla \cdot (\bar{q}_1 - \bar{q})$$

$$\text{i.e.} \quad \nabla \phi \cdot (\bar{q}_1 - \bar{q}) = \nabla \cdot [\phi (\bar{q}_1 - \bar{q})] - \phi \nabla \cdot (\bar{q}_1 - \bar{q})$$

Therefore, from (4), we find

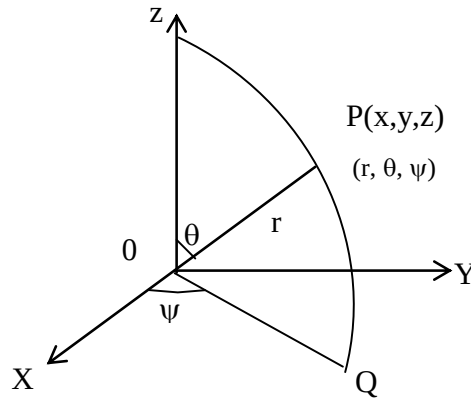
$$\begin{aligned} T_1 - T &= -\rho \int_{\tau} \nabla \cdot [\phi (\bar{q}_1 - \bar{q})] d\tau + \rho \int_{\tau} \phi \nabla \cdot (\bar{q}_1 - \bar{q}) d\tau \\ &\quad + \frac{\rho}{2} \int_{\tau} (\bar{q}_1 - \bar{q})^2 d\tau \\ &= -\rho \int_S \phi (\bar{q}_1 - \bar{q}) \cdot \hat{n} dS + \rho \int_{\tau} \phi \nabla \cdot (\bar{q}_1 - \bar{q}) d\tau \\ &\quad + \frac{\rho}{2} \int_{\tau} (\bar{q}_1 - \bar{q})^2 d\tau \quad \text{|By Gauss theorem} \\ &= \frac{\rho}{2} \int_{\tau} (\bar{q}_1 - \bar{q})^2 d\tau \quad \text{| using (2) \& (3)} \\ &> 0 \end{aligned}$$

$$\Rightarrow T_1 > T.$$

Hence the theorem.

3. Axially Symmetric Flows

A potential flow which is axially symmetric about the axis $\theta = 0, \pi$ (i.e. z-axis is taken as the axis of symmetry) has the property that at any point P, all the scalar and vector quantities associated with the flow are independent of azimuthal angle ψ such that $\frac{\partial}{\partial \psi} \equiv 0$, where (r, θ, ψ) are spherical polar co-ordinates.



The equation of continuity $\text{div } \bar{q} = 0$ for steady flow of an incompressible fluid becomes.

$$\frac{1}{r^2} \frac{\partial}{\partial r} (r^2 q_r) + \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} (\sin \theta q_\theta) = 0 \quad (1)$$

For irrotational motion $\bar{q} = -\nabla \phi$, where ϕ is velocity potential and thus

$$q_r = -\frac{\partial \phi}{\partial r}, \quad q_\theta = -\frac{1}{r} \frac{\partial \phi}{\partial \theta}$$

From equation (1), we have

$$\frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial \phi}{\partial r} \right) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial \phi}{\partial \theta} \right) = 0 \quad (2)$$

Let a solution of (2) in separable variables r, θ has the form

$$\phi = -R(r) \Theta(\theta) \quad (3)$$

Using (3) in (2), we get

$$\begin{aligned} & \frac{\partial}{\partial r} \left[r^2 \frac{\partial}{\partial r} (R\Theta) \right] + \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \left[\sin \theta \frac{\partial}{\partial \theta} (R\Theta) \right] = 0 \\ \Rightarrow & \Theta \frac{\partial}{\partial r} \left(r^2 \frac{\partial R}{\partial r} \right) + \frac{R}{\sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial \Theta}{\partial \theta} \right) = 0 \\ \Rightarrow & \frac{1}{R} \frac{d}{dr} \left(r^2 \frac{dR}{dr} \right) = - \frac{1}{\Theta \sin \theta} \frac{d}{d\theta} \left(\sin \theta \frac{d\Theta}{d\theta} \right) \end{aligned} \quad (4)$$

The L.H.S. of (4) is a function of r only while the R.H.S. is a function of θ only. The equation can therefore be satisfied if and only if either side is a constant, say $n(n+1)$ and thus we get

$$\frac{1}{R} \frac{d}{dr} \left(r^2 \frac{dR}{dr} \right) = n(n+1) \quad (5)$$

and

$$\frac{d}{d\theta} \left(\sin \theta \frac{d\Theta}{d\theta} \right) + n(n+1) \Theta \sin \theta = 0 \quad (6)$$

To solve (5), we put

$$R = r^m \Rightarrow \frac{dR}{dr} = m r^{m-1}$$

$$\text{Thus (5)} \Rightarrow \frac{1}{r^m} \frac{d}{dr} \left(r^2 m r^{m-1} \right) = n(n+1)$$

$$\Rightarrow m \frac{d}{dr} \left(r^{m+1} \right) = r^m n(n+1)$$

$$\Rightarrow m(m+1)r^m = r^m n(n+1)$$

$$\Rightarrow (m^2 + m - n^2 - n) = 0$$

$$\Rightarrow (m-n)(m+n+1) = 0$$

$$\Rightarrow m = n \text{ or } m = -(n+1)$$

Therefore, solution of (5) can be written as

$$R(r) = A_n r^n + B_n r^{-(n+1)} \quad (7)$$

To solve (6), we put

$$\cos\theta = \mu$$

$$\Rightarrow \frac{d}{d\theta} \equiv \frac{d\mu}{d\theta} \frac{d}{d\mu} \equiv -\sin\theta \frac{d}{d\mu}$$

Therefore, equation (6) becomes.

$$\begin{aligned} & -\sin\theta \frac{d}{d\mu} \left[\sin\theta (-\sin\theta) \frac{d\Theta}{d\mu} \right] + n(n+1) \Theta \sin\theta = 0 \\ \Rightarrow & \frac{d}{d\mu} \left(\sin^2\theta \frac{d\Theta}{d\mu} \right) + n(n+1) \Theta = 0 \\ \Rightarrow & \frac{d}{d\mu} \left[(1 - \cos^2\theta) \frac{d\Theta}{d\mu} \right] + n(n+1) \Theta = 0 \\ \Rightarrow & \frac{d}{d\mu} \left[(1 - \mu^2) \frac{d\Theta}{d\mu} \right] + n(n+1) \Theta = 0 \end{aligned} \quad (8)$$

Equation (8) is a Legendre's Equation and possesses a solution known as Legendre Function of the first kind $P_n(\mu)$

Therefore,

$$\Theta = P_n(\mu)$$

Hence the general solution of (3) is of the form

$$\begin{aligned} \phi(r, \theta) &= -R(r) \Theta(\theta) \\ &= -[A_n r^n + B_n r^{-(n+1)}] P_n(\cos\theta) \end{aligned} \quad (9)$$

(complete solution is the sum of all such solutions i.e. $\sum_{n=0}^{\infty} \dots\dots\dots$)

3.1. Uniform Flow. Consider the flow which corresponds to a potential given by (9) with

$$A_n = US_{1n}, B_n = 0, \quad (n = 0, 1, 2, \dots\dots) \quad | \quad S_{ij} \text{ is knonecker delta}$$

$$S_{ii} = 1 \quad S_{ij} = 0 \text{ for } i \neq j$$

Where U is a constant.

Since $P_1(\cos\theta) = \cos\theta$, equation (9) becomes

$$\phi(r, \theta) = -Ur \cos \theta \equiv -Uz \quad | \quad z = r \cos\theta$$

Thus

$$\bar{q} = -\nabla\phi = -\frac{\partial\phi}{\partial z}\hat{k} = U\hat{k}$$

which is a uniform streaming motion of the fluid with speed U along the direction of z-axis or the axis $\theta = 0$.

3.2. Sphere at Rest in a Uniform Stream. Consider an impermeable solid sphere of radius ‘a’ at rest with its centre at the pole of a system of spherical polar co-ordinates (r, θ , ψ). The sphere is immersed in an infinite homogeneous liquid with constant density ρ , which, in the absence of the sphere, would be flowing as a uniform stream with speed U along the direction $\theta = 0$.

The presence of the sphere will produce a local perturbation of the uniform streaming motion such that the disturbance diminishes with increasing distance r from centre of sphere. We say that the perturbation of the uniform stream tends to zero as $r \rightarrow \infty$.

In this problem z-axis is the axis of symmetry. Undisturbed velocity of incompressible fluid is $U\hat{k}$ i.e. $\bar{q} = U\hat{k}$

$\Rightarrow$ the velocity potential ϕ_0 due to such a uniform flow would be

$$\phi_0 = -Uz = -Ur \cos\theta$$

When the sphere is inserted, the undisturbed potential $-Ur \cos\theta$ of uniform stream has to be modified by “perturbation potential” due to the presence of the sphere. This must have the following properties.

- (i) It must satisfy Laplace equation for the case of axial symmetry
- (ii) It must tend to zero at large distances from the sphere

So, we write $\phi(r, \theta) = -U r \cos \theta + \phi_1(r, \theta) \quad (r \geq a)$

where ϕ_1 satisfies the Laplace equation together with boundary conditions

$$\frac{\partial \phi}{\partial r} = -U \cos \theta + \frac{\partial \phi_1}{\partial r} \quad \left| \frac{\partial \phi}{\partial r} = 0 \text{ i.e. velocity normal to} \right.$$

sphere is zero at $r = a$

$$\Rightarrow \frac{\partial \phi_1}{\partial r} = + U \cos \theta \quad (r = a, a \leq \theta \leq \pi)$$

and

$$|\nabla \phi_1| \rightarrow 0 \text{ as } r \rightarrow \infty.$$

Hence a suitable form of function ϕ_1 is

$$\phi_1 = -B \bar{r}^2 \cos \theta$$

So, we assume (in view of (9)) that

$$\phi(r, \theta) = -Ur \cos \theta - \frac{B}{r^2} \cos \theta \quad (1)$$

The constant B is to be determined from the fact that there is no flow normal to

the surface $r = a$ i.e. $\left(\frac{\partial \phi}{\partial r} \right)_{r=a} = 0$

$$\Rightarrow -U \cos \theta + \frac{2B}{a^3} \cos \theta = 0 \Rightarrow B = \frac{1}{2} U a^3$$

Thus (1) becomes

$$\begin{aligned} \phi(r, \theta) &= -Ur \cos \theta - \frac{U a^3}{2r^2} \cos \theta \\ &= -U \left(r + \frac{a^3}{2r^2} \right) \cos \theta \end{aligned} \quad (2)$$

Now, the uniqueness theorem II infer that the velocity potential in (2) is unique.

The velocity components at $P(r, \theta, \psi)$, ($r \geq a$), are

$$\begin{aligned} q_r &= -\frac{\partial \phi}{\partial r} = U \left(1 - \frac{a^3}{r^3} \right) \cos \theta \\ q_\theta &= -\frac{1}{r} \frac{\partial \phi}{\partial \theta} = -U \left(1 + \frac{a^3}{2r^3} \right) \sin \theta \end{aligned} \quad (3)$$

$$q_{\psi} = -\frac{1}{r \sin \theta} \frac{\partial \phi}{\partial \psi} = 0.$$

Different terms related to motion are obtained as follows.

(i) Stagnation Points : Stagnation points are those points in the flow where the velocity vanishes i.e. $\bar{q} = \bar{0}$. Thus these points are obtained by solving the equations

$$\text{and} \quad \left. \begin{aligned} U \left(1 - \frac{a^3}{r^3} \right) \cos \theta &= 0 \\ U \left(1 + \frac{a^3}{2r^3} \right) \sin \theta &= 0 \end{aligned} \right\} \quad (4)$$

which are satisfied only by $r = a$, $\sin \theta = 0$. i.e. $r = a$, $\theta = 0, \pi$. Thus the stagnation points are $(r = a, \theta = a)$ and $(r = a, \theta = \pi)$ on the sphere. These are referred to respectively as the **rear** and **forward** stagnation points.

(ii) Streamlines : The equations of streamlines

$$\frac{dr}{q_r} = \frac{rd\theta}{q_{\theta}} = \frac{r \sin \theta d\psi}{q_{\psi}}$$

for the present case, become

$$\frac{dr}{U \left(1 - \frac{a^3}{r^3} \right) \cos \theta} = \frac{rd\theta}{-U \left(1 + \frac{a^3}{2r^3} \right) \sin \theta} = \frac{r \sin \theta d\psi}{0}, \quad r \geq a$$

$$\Rightarrow d\psi = 0 \Rightarrow \psi = \text{constant.}$$

and

$$r \left(1 - \frac{a^3}{r^3} \right) \cos \theta d\theta = - \left(1 + \frac{a^3}{2r^3} \right) \sin \theta dr$$

$$\Rightarrow \frac{1}{r} \left(\frac{2r^3 + a^3}{r^3 - a^3} \right) dr = -2 \cot \theta d\theta$$

$$\Rightarrow \left(\frac{2r + a^3 r^{-2}}{r^2 - a^3 r^{-1}} \right) dr = -2 \cot \theta d\theta$$

Integrating, we get

$$\log (r^2 - a^3 r^{-1}) = -2 \log \sin \theta + \log C$$

$$\Rightarrow \log \left(\frac{r^3 - a^3}{r} \right) = -\log \sin^2 \theta + \log C$$

$$\Rightarrow \sin^2 \theta = \frac{Cr}{r^3 - a^3}, \text{ where } C \geq 0$$

For each value of C , above equation gives a streamline in the plane $\psi = \text{constant}$. The choice of $c = 0$ corresponds to the sphere and the axis of symmetry.

(iii) Pressure at Any Point : The pressure at any point of the fluid is obtained by applying Bernoulli's equation along the streamline through that point, taking the pressure at ∞ to be of constant value p_∞ . Thus, in the absence of body force, the Bernoulli's equation for homogeneous steady flow is

$$\frac{p}{\rho} + \frac{1}{2}(-\nabla\phi)^2 = C$$

At infinity, $p = p_\infty$ and $-\nabla\phi = U\hat{k}$, we get

$$C = \frac{p_\infty}{\rho} + \frac{1}{2}U^2$$

Thus

$$p = p_\infty + \frac{1}{2}\rho U^2 - \frac{1}{2}\rho(\nabla\phi)^2$$

$\Rightarrow$

$$p = p_\infty +$$

$$\frac{1}{2}\rho U^2 - \frac{1}{2}\rho \left[U^2 \left(1 - \frac{a^3}{r^3} \right)^2 \cos^2 \theta + U^2 \left(1 + \frac{a^3}{2r^3} \right)^2 \sin^2 \theta \right]$$

$$|\nabla\phi| = -\bar{q}$$

$$\Rightarrow p = p_\infty - \frac{1}{2}\rho U^2 \left[\left(1 - \frac{a^3}{r^3} \right)^2 \cos^2 \theta + \left(1 + \frac{a^3}{2r^3} \right)^2 \sin^2 \theta - 1 \right] \quad (5)$$

which gives the pressure at any point of the fluid. Of particular interest is the distribution of pressure on the boundary of the sphere. It is obtained by putting $r = a$ in (5) and thus

$$\begin{aligned} p &= p_\infty - \frac{1}{2}\rho U^2 \left[\left(1 + \frac{a^3}{2r^3} \right)^2 \sin^2 \theta - 1 \right] \\ &= p_\infty - \frac{1}{2}\rho U^2 \left(\frac{9}{4} \sin^2 \theta - 1 \right) = p_\infty + \frac{1}{8}\rho U^2 (4 - 9 \sin^2 \theta) \end{aligned}$$

$$= p_{\infty} + \frac{1}{8} \rho U^2 (9 \cos^2 \theta - 5)$$

The maximum pressure occurs at the stagnation points, where $\theta = 0$ or π . Thus

$$p_{\max} = p_{\infty} + \frac{1}{2} \rho U^2$$

($p_{\max}$ is also called stagnation pressure)

The minimum pressure occurs along the equatorial circle of the sphere where $\theta = \pi/2$

Therefore,

$$p_{\min} = p_{\infty} - \frac{5}{8} \rho U^2$$

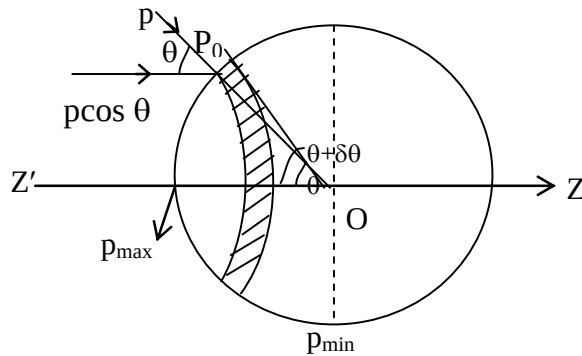
A fluid is presumed to be incapable of sustaining a negative pressure, thus

$$p_{\min} = 0 \Rightarrow U = \sqrt{\frac{8p_{\infty}}{5\rho}}$$

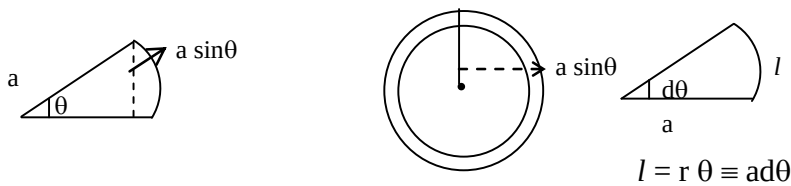
At this stage the fluid will tend to break away from the surface of the sphere and **cavitation** is said to occur. i.e. a vacuum is formed.

(iv) Thrust on the Hemisphere : Now, we find the **thrust** (force) on the hemisphere on which the liquid impinges, $r = a$, $0 \leq \theta \leq \pi/2$.

Let δS be a small element at $P_0(a, \theta, \psi)$ of the hemisphere bounded by circles at $r = a$ and at angular distances θ and $\theta + \delta\theta$ from axis of symmetry (i.e. z-axis)



The component of thrust on δS is $p \cos \theta \delta S$. Hence the total thrust on the hemisphere is along $Z'O$ and is given by



$$\begin{aligned}
 dS &= (2\pi a \sin \theta)(a d\theta) \\
 F &= \int_{\text{hemisphere}} p \cos \theta \, dS \\
 &= \int_0^{\pi/2} p \cos \theta (2\pi a \sin \theta) (a d\theta) \\
 &= \int_0^{\pi/2} (2\pi a^2) \sin \theta \cos \theta \left[p_{\infty} + \frac{1}{8} \rho U^2 (9 \cos^2 \theta - 5) \right] d\theta
 \end{aligned}$$

(using value of p at boundary)

$$= \pi a^2 \left[p_{\infty} - \frac{1}{16} \rho U^2 \right].$$

3.3. Sphere in Motion in Fluid at Rest at Infinity. Let a solid sphere of radius 'a' centred at 0 be moving with uniform velocity $-U \hat{k}$ in incompressible fluid of infinite extent, which is at rest at infinity. Z-axis is the axis of symmetry and $\hat{k}$ is unit vector in this direction. (As the sphere is moving with velocity $-U \hat{k} \Rightarrow$ the relative velocity of fluid if the sphere be considered to be at rest is $U \hat{k}$.)

The boundary value problem for ϕ is now to solve

$$\nabla^2 \phi = 0 \quad (1)$$

$$\text{such that } -\frac{\partial \phi}{\partial r} = -U \cos \theta, \quad (r = a) \quad (2)$$

and

$$|\nabla \phi| \rightarrow 0, \quad (r \rightarrow \infty) \quad (3)$$

The present case is also a problem with axial symmetry about the axis $\theta = 0, \pi$, so

$$\phi = \phi(r, \theta)$$

Also, since $P_1(\cos \theta) = \cos \theta$ | Legendre's function

and the boundary condition (2) implies that the dependence of ϕ on θ must be like $\cos \theta$, therefore ϕ has the form

$$\phi = - \left(Ar + \frac{B}{r^2} \right) P_1(\cos \theta) = - \left(Ar + \frac{B}{r^2} \right) \cos \theta$$

However, to satisfy (3), it is necessary that $A = 0$, and then from (2), we get $B = \frac{1}{2} U a^3$.

Thus the solution for ϕ is

$$\phi = \frac{-Ua^3}{2r^2} \cos\theta \quad (4)$$

From here, the velocity components are obtained to be

$$q_r = -\frac{\partial\phi}{\partial r} = \frac{Ua^3}{r^3} \cos\theta, \quad q_\theta = \frac{-1}{r} \frac{\partial\phi}{\partial\theta} = \frac{Ua^3}{2r^3} \sin\theta, \quad q_\psi = 0,$$

where (r, θ, ψ) are spherical polar co-ordinates. The various terms of particular importance related to this motion are obtained as follows.

(i) Streamlines : The differential equations for streamlines are

$$\frac{dr}{q_r} = \frac{rd\theta}{q_\theta} = \frac{r \sin\theta d\psi}{q_\psi}$$

i.e.
$$\frac{dr}{-\frac{Ua^3}{r^3} \cos\theta} = \frac{rd\theta}{-\frac{Ua^3}{2r^3} \sin\theta} = \frac{r \sin\theta d\psi}{0}$$

$\Rightarrow d\psi = 0 \Rightarrow \psi = \text{constant.}$

and

$$\frac{dr}{r} = 2 \cot\theta d\theta \Rightarrow \log r = 2 \log \sin\theta + \log C$$

$\Rightarrow r = C \sin^2\theta$

Therefore, streamline lines are given by $r = C \sin^2\theta, \psi = \text{constant}$

(ii) K.E. of the Liquid : Let S be the surface of sphere and ρ be the density of liquid, then K.E. is given by

$$T_1 = \frac{\rho}{2} \int_S \phi \frac{\partial\phi}{\partial n} dS \quad (5)$$

Where $\hat{n}$ is the outwards unit normal. But for the sphere $\hat{n}$ is along radius vector

Therefore,
$$\left(\phi \frac{\partial\phi}{\partial n} \right)_S = \left(-\phi \frac{\partial\phi}{\partial r} \right)_{r=a}$$

$$= \left(\frac{1}{2} Ua \cos\theta \right) (U \cos\theta)$$

$$= \frac{1}{2} U^2 a \cos^2 \theta$$

Therefore,

$$\begin{aligned} T_1 &= \frac{\rho}{2} \int_S \frac{1}{2} U^2 a \cos^2 \theta dS = \frac{\rho a U^2}{4} \int_0^\pi \cos^2 \theta (2\pi a \sin \theta) (a d\theta) \\ &= \frac{\pi \rho a^3 U^2}{2} \int_0^\pi \cos^2 \theta \sin \theta d\theta \\ &\quad \left| \begin{array}{l} 0 \leq \theta \leq \pi \\ 0 \leq \psi \leq 2\pi \end{array} \right. \\ &= \frac{\pi \rho a^3 U^2}{2} \left[-\frac{\cos^3 \theta}{3} \right]_0^\pi \\ &= \frac{1}{3} \pi \rho a^3 U^2 = \left(\frac{4}{3} \pi \rho a^3 \right) \left(\frac{U^2}{4} \right) \\ &= \frac{1}{4} M^1 U^2 \end{aligned} \quad (6)$$

where $M^1 = \frac{4}{3} \pi \rho a^3$ is the mass of the liquid displaced by the sphere.

Also, K.E. of the sphere moving with speed U is given by

$$T_2 = \frac{1}{2} M U^2 \quad (7)$$

where $M = \frac{4}{3} \pi \sigma a^3$ is the mass of the sphere, σ being the density of the material of the sphere.

Therefore, from (6) and (7), total K.E. T is given by

$$T = T_1 + T_2 = \frac{1}{2} \left(M + \frac{M^1}{2} \right) U^2 \quad (8)$$

The quantity $M + \frac{M^1}{2}$ is called the virtual mass of the sphere.

3.4. Accelerating Sphere Moving in a Fluid at Rest at Infinity. The solution derived above for ϕ is applicable when the sphere translates unsteadily along a straight line. In the present case, we take $U = U(t)$ and get the velocity potential as

$$\phi = \phi(r, \theta, t) = -\frac{U(t)a^3}{2r^2} \cos \theta \quad (1)$$

The instantaneous values of velocity components and K.E. at time t are given by

$$q_r = \frac{-U(t)a^3}{r^3} \cos\theta, q_\theta = \frac{-U(t)a^3}{2r^3} \sin\theta, q_\psi = 0 \quad | \text{ similar to steady case}$$

and
$$T = \frac{1}{2} \left(M + \frac{1}{2} M' \right) U^2(t) \quad (2)$$

The pressure at any point of the fluid is obtained by using Bernoulli's equation for unsteady flow of a homogeneous liquid, in the absence of body force, as

$$\frac{p}{\rho} + \frac{1}{2} \bar{U}^2 - \frac{\partial \phi}{\partial t} = f(t) \quad (3)$$

where $f(t)$ is a function of time t only.

Let p_∞ be the pressure at infinity where the fluid is at rest, then from (3), we get

$$f(t) = \frac{p_\infty}{\rho} \text{ and thus}$$

$$\frac{p}{\rho} = \frac{p_\infty}{\rho} - \frac{1}{2} \bar{U}^2 + \frac{\partial \phi}{\partial t} \quad (4)$$

To find $\frac{\partial \phi}{\partial t}$, we proceed as follows :

Since $\bar{U} = -U\hat{k} = -U(t)\hat{k}$ is the velocity of the sphere, the velocity potential given in (1) can be expressed in the form

$$\phi = \frac{1}{2} \frac{a^3 (\bar{U} \cdot \bar{r})}{r^3} \quad (5)$$

since $\bar{r}$ is the position vector of a fixed point P of the fluid relative to the moving centre O of the sphere, it follows that

$$\bar{U} = \frac{\partial}{\partial t}(-\bar{r}) \quad (6)$$

Also, since $r^2 = \bar{r} \cdot \bar{r} \Rightarrow r \frac{\partial r}{\partial t} = \bar{r} \cdot \frac{\partial \bar{r}}{\partial t} = -\bar{r} \cdot \bar{U} \quad | \text{ using (6)}$

$$\begin{aligned} &= (-\bar{r}) \cdot (-U\hat{k}) \\ &= rU (\hat{r} \cdot \hat{k}) \quad | \bar{r} = r \hat{r} \\ &= rU \cos\theta \end{aligned}$$

$$\Rightarrow \frac{\partial r}{\partial t} = U \cos \theta \quad (7)$$

Differentiating (5) w.r.t. time t and using (6) & (7), we get

$$\begin{aligned} \frac{\partial \phi}{\partial t} &= \frac{1}{2} a^3 \left[-\frac{U^2}{r^3} - \frac{\cos \theta}{r^2} \frac{\partial U}{\partial t} + \frac{3U^2}{r^3} \cos^2 \theta \right] \\ &= -\frac{a^3}{2} \left[\frac{U \cos \theta}{r^2} + \frac{U^2}{r^3} - \frac{3U^2 \cos^2 \theta}{r^3} \right] \quad \left| \dot{U} = \frac{\partial U}{\partial t} \right. \end{aligned}$$

Also,

$$\begin{aligned} \bar{U}^2 &= q_r^2 + q_\theta^2 = \frac{U^2 a^6}{r^6} \cos^2 \theta + \frac{U^2 a^6}{4r^6} \sin^2 \theta \\ &= \frac{U^2 a^6}{r^6} \left(\cos^2 \theta + \frac{1}{4} \sin^2 \theta \right) \end{aligned}$$

The pressure at any point of the fluid can be obtained from equation (4).

In particular, at a point on the sphere $r = a$

$$\frac{\partial \phi}{\partial t} = \frac{-1}{2} [\dot{U} a \cos \theta + U^2 - 3U^2 \cos^2 \theta]$$

and
$$\bar{U} = \frac{U^2}{4} (4 \cos^2 \theta + \sin^2 \theta)$$

and the corresponding pressure is given by

$$\frac{p}{\rho} = \frac{p_\infty}{\rho} - \frac{1}{2} \dot{U} a \cos \theta + \frac{1}{8} U^2 (9 \cos^2 \theta - 5) \quad (8)$$

The force (thrust) acting on the sphere is given by

$$\begin{aligned} \bar{F} &= \int_0^\pi p \cos \theta (2\pi a \sin \theta) (a d\theta) \hat{k} \\ &= 2\pi a^2 \hat{k} \int_0^\pi \left[p_\infty - \frac{1}{2} \rho \dot{U} a \cos \theta + \frac{1}{8} \rho U^2 (9 \cos^2 \theta - 5) \right] \cos \theta \sin \theta d\theta \\ &= \frac{2}{3} \pi \rho a^3 \dot{U} \hat{k} = \frac{1}{2} \left(\frac{4}{3} \pi a^3 \rho \right) \dot{U} \hat{k} = \frac{1}{2} M' \dot{U} \hat{k} \end{aligned}$$

where $M' = \frac{4}{3} \pi a^3 \rho$ is mass of the liquid displaced. This shows that the force acts in the direction opposing the sphere's motion.

3.5. Equation of Motion of the Sphere. Let R be the external force per unit mass in the direction of motion of the sphere. Let us use the result that the rate of doing work is equal to the rate of increase in K.E.

Thus
$$RU = \frac{d\tau}{dt} = \frac{1}{2} \frac{d}{dt} \left[\left(M + \frac{M'}{2} \right) U^2(t) \right]$$

| From (2)

$$= \left(M + \frac{M'}{2} \right) U \frac{dU}{dt}$$

$$\Rightarrow M \frac{dU}{dt} = R - \frac{1}{2} M' \frac{dU}{dt} \quad (9)$$

If the liquid is not there, then $M' = 0$ and the equation of motion of the sphere is

$$M \frac{dU}{dt} = R \quad (10)$$

Comparing equation (9) & (10), we note that the presence of the liquid offers a resistance of the amount $\frac{1}{2} M' \frac{dU}{dt}$ to the motion of the sphere

Let R' be the external force per unit mass on the sphere when there is no liquid, then

MR = external force on the sphere in the presence of the liquid.

$$= MR' - M'R' = (M - M') R'$$

Since, $M = \frac{4\pi\sigma a^3}{3}, M' = \frac{4\pi\rho a^3}{3}$

$$\therefore R = \left(\frac{\sigma - \rho}{\sigma} \right) R' \quad (11)$$

From equations (9) & (11), we find

$$M \frac{dU}{dt} = \left(\frac{\sigma - \rho}{\sigma} \right) R' - \frac{1}{2} M' \frac{dU}{dt}$$

or
$$\left(M + \frac{M'}{2} \right) \frac{dU}{dt} = \left(\frac{\sigma - \rho}{\sigma} \right) R' = \left(\frac{M - M'}{M} \right) R'$$

UNIT – VII

1. Vortex Motion

So far we have confined our attention to the cases involving irrotational motion only. But the most general displacement of a fluid involves rotation such that the rotational vector (vortex vector or vorticity) $\vec{\xi} = \text{curl } \vec{q} \neq \vec{0}$. Here we consider the theory of rotational or vortex motion. First of all we revisit some elementary definitions.

Lines drawn in the fluid so as at every point to coincide with the instantaneous axis of rotation of the corresponding fluid element are called **vortex lines**. Portions of the fluid bounded by vortex lines drawn through every point of an infinity small closed curve are called **vortex filaments** or simply **vortices** and the boundary of a vortex filament is called a **vortex tube**.

If C is a closed curve, then **circulation** about C is given by

$$\Gamma = \oint_C \vec{q} \cdot d\vec{r} = \int_S \hat{n} \cdot \text{curl } \vec{q} \, dS = \int_S \hat{n} \cdot \vec{\xi} \, dS = \int_S \vec{\xi} \cdot d\vec{S}$$

The quantity $|\hat{n} \cdot \vec{\xi}| \delta S$ is called the strength of the vortex tube. A vortex tube with a unit strength is called a unit vortex tube.

We shall observe some important results for vortex motion which are consequences of the following theorem due to Lord Kelvin.

1.1. Kelvin's Circulation Theorem (Consistency of circulation). The circulation around a closed contour C moving with the inviscid (non-viscous) fluid is constant for all times provided that the external forces (body forces) are conservative and the density is a function of pressure only.

Proof. The circulation round a closed curve C of fluid particles is defined by

$$\Gamma = \oint_C \vec{q} \cdot d\vec{r},$$

where $\vec{q}$ is the velocity and $\vec{r}$ is the position vector of a fluid particle at any time t.

Time derivative of Γ following the motion of fluid is

$$\frac{d\Gamma}{dt} = \frac{d}{dt} \oint_C \vec{q} \cdot d\vec{r} = \oint_C \frac{d}{dt} (\vec{q} \cdot d\vec{r})$$

$$\begin{aligned}
 &= \oint_C \left[\frac{d\bar{q}}{dt} \cdot d\bar{r} + \bar{q} \cdot \frac{d}{dt}(d\bar{r}) \right] \\
 &= \oint_C \left[\frac{d\bar{q}}{dt} \cdot d\bar{r} + \bar{q} \cdot d\bar{q} \right] \quad (1) \quad \left| \because \frac{d}{dt}(d\bar{r}) = d\left(\frac{d\bar{r}}{dt}\right) = d\bar{q} \right.
 \end{aligned}$$

Since the system of forces is conservative; therefore $\bar{F} = -\nabla\Omega$, where Ω is a potential function Euler's equation of motion is

$$\frac{d\bar{q}}{dt} = \bar{F} - \frac{1}{\rho} \nabla p = -\nabla\Omega - \frac{1}{\rho} \nabla p \quad (2)$$

Multiplying each term of (2) scalarly by $d\bar{r}$, we get

$$d\bar{r} \cdot \frac{d\bar{q}}{dt} = -d\bar{r} \cdot \nabla\Omega - \frac{1}{\rho} d\bar{r} \cdot \nabla p$$

$$\text{i.e.} \quad \frac{d\bar{q}}{dt} \cdot d\bar{r} = -d\Omega - \frac{dp}{\rho} \quad (3) \quad \left| \because d\bar{r} \cdot \nabla \equiv d \right.$$

Thus from (1), we get

$$\begin{aligned}
 \frac{d\Gamma}{dt} &= \oint_C \left(-d\Omega - \frac{dp}{\rho} + \bar{q} \cdot d\bar{q} \right) \\
 &= \oint_C \left[d\left(\frac{1}{2} \bar{q}^2 - \Omega \right) - \frac{1}{\rho} dp \right] \\
 &= \oint_C d\left(\frac{1}{2} \bar{q}^2 - \Omega \right) - \oint_C \frac{1}{\rho} dp \\
 &= \left[\frac{1}{2} \bar{q}^2 - \Omega \right]_A^A - \oint_C \frac{dp}{\rho} \\
 &= 0 - \oint_C \frac{dp}{\rho} \quad (4)
 \end{aligned}$$

where A is any point on the closed contour C. Now, if density is a function of pressure only, then the integral $\oint_C \frac{dp}{\rho}$ vanishes and hence we get

$$\frac{d\Gamma}{dt} = 0 \Rightarrow \Gamma = \text{constant for all time}$$

Corollary (1). In a closed circuit C of fluid particles moving under the same conditions as in the theorem,

$$\int_S \text{curl } \bar{q} \cdot d\bar{S} = \int_S \bar{\xi} \cdot d\bar{S} = \text{constant} \quad (5)$$

where S is any open surface whose rim is C. To establish (5), we note that, by Stock's theorem,

$$\int_S \text{curl } \bar{q} \cdot d\bar{S} = \oint_C \bar{q} \cdot d\bar{r} = \Gamma = \text{constant}$$

This shows that the product of the cross-section and angular velocity at any point on a vortex filament is constant all along the vortex filament and for all times.

Corollary (2). Under the conditions of the theorem, vortex lines move with the fluid.

Proof. Let C be any closed curve drawn on the surface of a vortex tube. Let S be the portion of the vortex tube rimmed by C. By definition vortex lines lie on S. Thus

$$0 = \int_S \text{curl } \bar{q} \cdot d\bar{S} = \oint_C \bar{q} \cdot d\bar{r} \quad | \because \text{on surface circulation is zero}$$

Let C be a material curve and S be a material surface, then

$$\frac{d}{dt} \int_S (\hat{n} \cdot \text{curl } \bar{q}) dS = \int_S \frac{D}{Dt} (\hat{n} \cdot \text{curl } \bar{q}) dS = 0$$

Thus $\hat{n} \cdot \text{curl } \bar{q}$ remains zero, so that S remains a surface composed of vortex lines. Consequently vortex lines and tubes move with the fluid i.e. vortex filaments are composed of the same fluid particles. This explains why smoke rings maintain their forms for long periods of time.

Corollary (3). Under the conditions of the theorem, if the flow is irrotational in a material region of the fluid at some particular time (e.g. $t = 0$ or $t = t_0$), the flow is always irrotational in that material region thereafter.

i.e. If the motion of an ideal fluid is once irrotational it remains irrotational for ever afterwards provided the external forces are conservative and density ρ is a function of pressure p only.

Proof. Suppose that at some instant ($t = t_0$), the fluid on the material surface S is irrotational

$$\text{Then, } \bar{\xi} = 0 \quad (1)$$

for all points of S .

Let C be the boundary of surface S , then

$$\Gamma = \oint_C \bar{q} \cdot d\bar{r} = \int_S (\hat{n} \cdot \text{curl } \bar{q}) dS = \int_S (\hat{n} \cdot \bar{\xi}) dS = 0 \quad | \text{ using (1)}$$

But by Kelvin's circulation theorem, Γ is constant for all times. Hence circulation Γ is zero for all subsequent times. At any later time,

$$\int_S \hat{n} \cdot \bar{\xi} dS = 0$$

If we now take S to be non-zero infinitesimal element, say ΔS , then

$\hat{n} \cdot \bar{\xi} \Delta S = 0 \Rightarrow \bar{\xi} = 0$ at all points of S for all times and the motion is irrotational permanently. This proves the permanency of irrotational motion.

1.2. Remarks (i) The above three corollaries are properties of vortex filaments.

(ii) The Kelvin's theorem is true whether the motion be rotational or irrotational. In case of irrotational motion, $\bar{\xi} = \bar{0}$ and thus $\Gamma = 0$

(iii) From the results of the theorem, we conclude that vortex filaments must either form closed curves or have their ends on the bounding surface of the fluid. A vortex in an ideal fluid is therefore permanent.

1.3. Vorticity Equation. Euler's equation of motion for an ideal fluid under the action of a conservative body force with potential Ω per unit mass is

$$\frac{D\bar{q}}{Dt} = \frac{\partial \bar{q}}{\partial t} + \nabla \left(\frac{1}{2} \bar{q}^2 \right) - \bar{q} \times \bar{\xi} = -\nabla \Omega - \frac{1}{\rho} \nabla p \quad (1)$$

where the vorticity $\bar{\xi} = \text{curl } \bar{q} = \nabla \times \bar{q}$. If the fluid has constant density, then taking curl of equation (1), we get

$$\nabla \times \frac{\partial \bar{q}}{\partial t} + \nabla \times \left[\nabla \left(\frac{1}{2} \bar{q}^2 \right) \right] - \nabla \times (\bar{q} \times \bar{\xi}) = \nabla \times \left(-\nabla \Omega - \frac{1}{\rho} \nabla p \right)$$

$$\Rightarrow \nabla \times \frac{\partial \bar{q}}{\partial t} - \nabla \times (\bar{q} \times \bar{\xi}) = 0$$

$$\Rightarrow \frac{\partial}{\partial t} - (\nabla \times \bar{q}) - \nabla \times (\bar{q} \times \bar{\xi}) = 0$$

$$\begin{aligned} \Rightarrow \frac{\partial \bar{\xi}}{\partial t} &= \nabla \times (\bar{q} \times \bar{\xi}) \\ &= (\bar{\xi} \cdot \nabla) \bar{q} - (\bar{q} \cdot \nabla) \bar{\xi} \end{aligned}$$

$$\Rightarrow \frac{\partial \bar{\xi}}{\partial t} + (\bar{q} \cdot \nabla) \bar{\xi} = (\bar{\xi} \cdot \nabla) \bar{q}$$

$$\text{i.e.} \quad \frac{D\bar{\xi}}{Dt} = (\bar{\xi} \cdot \nabla) \bar{q} \quad (2)$$

which is the required vorticity equation.

Equation (2) is called Helmholtz's vorticity equation. For two-dimensional motion, the vorticity vector $\bar{\xi}$ is perpendicular to the velocity vector $\bar{q}$ and the R.H.S. of (2) is identically zero. Thus, for two dimensional motion of an ideal fluid, vorticity is constant.

In the case, when body force is not conservative, equation (2) becomes

$$\frac{D\bar{\xi}}{Dt} = (\bar{\xi} \cdot \nabla) \bar{q} + \text{curl } \bar{F}$$

where $\bar{F}$ is body force per unit mass.

1.4. Example. A motion of in viscid incompressible fluid of uniform density is symmetrical about the axis $r = 0$ where (r, θ, z) are cylindrical polar co-ordinates. The cylindrical polar resolutives of velocity are $[q_r(r, z), 0, q_z(r, z)]$. Show that if a fluid particle has vorticity of magnitude ξ_0 when $r = r_0$, its vorticity when at general distance r from the axis of symmetry has magnitude $\xi = (\xi_0/r_0)r$, if any body forces acting are conservative.

Solution. The vorticity vector $\bar{\xi}$ satisfies the vorticity equation

$$\frac{D\bar{\xi}}{Dt} = (\bar{\xi} \cdot \nabla) \bar{q} \quad (1)$$

Now,

$$\begin{aligned} \bar{\xi} = \text{curl } \bar{q} &= \frac{1}{r} \begin{vmatrix} \hat{r} & r\hat{\theta} & \hat{z} \\ \frac{\partial}{\partial r} & \frac{\partial}{\partial \theta} & \frac{\partial}{\partial z} \\ q_r(r, z) & 0 & q_z(r, z) \end{vmatrix} \\ &= \frac{1}{r} \left[\hat{r} \frac{\partial}{\partial \theta} \{q_z(r, z)\} + r\hat{\theta} \left\{ \frac{\partial}{\partial z} q_r(r, z) - \frac{\partial}{\partial r} q_z(r, z) \right\} + \hat{z}(0) \right] \\ &= \left[\frac{\partial q_r}{\partial z} - \frac{\partial q_z}{\partial r} \right] \hat{\theta} . \end{aligned} \quad (2)$$

Therefore,

$$\begin{aligned} (\bar{\xi} \cdot \nabla) &= \bar{\xi} \cdot \left(\hat{r} \frac{\partial}{\partial r} + \hat{\theta} \frac{1}{r} \frac{\partial}{\partial \theta} + \hat{z} \frac{\partial}{\partial z} \right) \\ &= \hat{\theta} \left(\frac{\partial q_r}{\partial z} - \frac{\partial q_z}{\partial r} \right) \cdot \left(\hat{r} \frac{\partial}{\partial r} + \hat{\theta} \frac{1}{r} \frac{\partial}{\partial \theta} + \hat{z} \frac{\partial}{\partial z} \right) \\ &= \frac{1}{r} \left(\frac{\partial q_r}{\partial z} - \frac{\partial q_z}{\partial r} \right) \frac{\partial}{\partial \theta} \quad \left| \hat{\theta} \hat{r} = 0 = \hat{\theta} \hat{z} \right. \end{aligned}$$

Thus

$$\begin{aligned} (\bar{\xi} \cdot \nabla) \bar{q} &= \frac{1}{r} \left(\frac{\partial q_r}{\partial z} - \frac{\partial q_z}{\partial r} \right) \frac{\partial}{\partial \theta} (q_r \hat{r} + q_z \hat{z}) \\ &= \frac{q_r}{r} \left[\frac{\partial q_r}{\partial z} - \frac{\partial q_z}{\partial r} \right] \frac{\partial \hat{r}}{\partial \theta} \end{aligned}$$

$$= \frac{q_r}{r} \left(\frac{\partial q_r}{\partial z} - \frac{\partial q_z}{\partial r} \right) \hat{\theta} \quad | \quad \frac{\partial \hat{r}}{\partial \theta} = \hat{\theta}, \quad \frac{\partial \hat{z}}{\partial \theta} = 0 \quad (3)$$

$$\text{Hence } (\bar{\xi} \cdot \nabla) \bar{q} = \frac{q_r}{r} \bar{\xi} \quad | \text{ using (2)} \quad (4)$$

∴ From (1) & (4), we get

$$\frac{D\bar{\xi}}{Dt} = \frac{q_r}{r} \bar{\xi} \quad (5)$$

Now, $q_r = \bar{q} \cdot \hat{r}$, so equation (5) becomes

$$r \frac{D\bar{\xi}}{Dt} = \bar{q} \cdot \hat{r} \bar{\xi} \quad (6)$$

$$\text{Since } \bar{r}^2 = r^2 \Rightarrow \bar{r} \cdot \frac{D\bar{r}}{Dt} = r \frac{Dr}{Dt}$$

$$\Rightarrow \frac{\bar{r}}{r} \cdot \frac{D\bar{r}}{Dt} = \frac{Dr}{Dt} \Rightarrow \hat{r} \cdot \frac{D\bar{r}}{Dt} = \frac{Dr}{Dt} \Rightarrow \hat{r} \cdot \bar{q} = \frac{Dr}{Dt}$$

Using this in (6), we get

$$r \frac{D\bar{\xi}}{Dt} = \frac{Dr}{Dt} \bar{\xi}$$

$$\Rightarrow r \frac{D\bar{\xi}}{Dt} - \bar{\xi} \frac{Dr}{Dt} = 0 \quad \Rightarrow \frac{r \frac{D\bar{\xi}}{Dt} - \bar{\xi} \frac{Dr}{Dt}}{r^2} = 0$$

$$\Rightarrow \frac{D}{Dt} \left(\frac{\bar{\xi}}{r} \right) = 0 \Rightarrow \frac{\bar{\xi}}{r} = \text{const} = \frac{\xi_0}{r_0}$$

$$\Rightarrow \frac{r^2}{\xi} = \left(\frac{\xi_0}{r_0} \right) r$$

Hence the result.

2. Vorticity in Two-dimensions

For an incompressible fluid in the xy-plane, we have

$$\bar{q} = (u, v, 0), \quad \nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, 0 \right)$$

Therefore,
$$\bar{\xi} = \nabla \times \bar{q} = (0, 0, \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y})$$

$$= \hat{k} \left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \right)$$

which shows that in two-dimensional flow, the vorticity vector is perpendicular to the plane of flow.

Also,
$$\xi = |\bar{\xi}| = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}$$

Thus
$$\bar{\xi} = \hat{k} \xi$$

Now, for this case, the Helmholtz's vorticity equation

$$\frac{d\bar{\xi}}{dt} = (\bar{\xi} \cdot \nabla) \bar{q} \text{ gives}$$

$$\frac{d\bar{\xi}}{dt} = 0 \Rightarrow \bar{\xi} = \text{constant}$$

i.e. $\xi = \text{constant}$.

which shows that in the two-dimensional motion of an incompressible fluid, the vorticity of any particle remains constant.

Here, we may regard ξ as a vortex strength per unit area.

Also, in terms of stream function, we have

$$u = -\frac{\partial \psi}{\partial y}, \quad v = \frac{\partial \psi}{\partial x}$$

Therefore,
$$\bar{\xi} = \hat{k} \left(\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} \right) = \hat{k} \nabla^2 \psi$$

i.e.
$$\xi = \nabla^2 \psi$$

This gives vorticity in terms of the stream function.

2.1. Circular Vortex. The section of a cylindrical vortex tube whose cross-section is a circle of radius a , by the plane of motion is a circle and the liquid inside such a tube is said to form a circular vortex.

If ω is the angular velocity and πa^2 the cross-sectional area of the vortex tube, then circulation

$$\begin{aligned}\Gamma &= \oint_C \bar{q} \cdot d\bar{r} = \int_S \text{curl } \bar{q} \cdot \hat{n} dS = \int_S \text{curl } \bar{q} \cdot d\bar{S} \\ &= \omega \int_S dS = \omega \pi a^2 = k(\text{say})\end{aligned}$$

This product of the cross-section and angular velocity at any point of the vortex tube is constant along the vortex and is known as the strength of the circular vortex.

2.2. Rectilinear or Columnar Vortex Filament. The strength k of circular vortex is given by $k = \omega \pi a^2$. If we let $a \rightarrow 0$ and $\omega \rightarrow \infty$ such that the product $\omega \pi a^2$ remains constant, we get a rectilinear vortex filament and represent it by a point in the plane of motion. Such vortex filament may be regarded as straight gravitating rod of fluid lying perpendicular to the plane of flow. It is also termed as a uniform line vortex. The strength of a vortex filament is positive when the circulation round it is anticlockwise and negative when clockwise.

2.3. Different Types of Vortices. We may divide vortices into the following four types

- (i) **Forced vortex** in which the fluid rotates as a rigid body with constant angular velocity.
- (ii) **Free cylindrical vortex** for which the fluid moves along streamlines which are concentric circles in horizontal planes and there is no variation of total energy with radius.
- (iii) **Free spiral vortex** which is a combination the free cylindrical vortex and a source (radial flow)
- (iv) **Compound vortex** in which the fluid rotates as a forced vortex at the centre and as a free vortex outside.

2.4. Complex Potential for Circulation about a Circular Cylinder (Circular vortex). In case of a doubly connected region, the possibility of cyclic motion does exist and as such we proceed to explain it presently in the case of circle.

If the circulation in a closed circuit is $2\pi k$, then k is called the strength of the circulation.

Let us consider the complex potential

$$W = \phi + i\psi = ik \log z \quad (1)$$

On the circular cylinder $|z| = a$, $z = a e^{i\theta}$

Thus, $W = ik \log (a e^{i\theta}) = ik (\log a + i\theta)$

$$\text{i.e.} \quad \phi + i\psi = -k\theta + ik \log a$$

$$\Rightarrow \quad \phi = -k\theta, \psi = k \log a = \text{constant.}$$

This shows that the circular cylinder is a streamline and thus equation (1) gives the required complex potential for circulation about a circular cylinder.

When the fluid moves once round the cylinder in the positive sense, θ increases by 2π and then

$$\begin{aligned} \phi_1 &= -k(\theta + 2\pi) = -k\theta - 2\pi k \\ &= \phi - 2\pi k \end{aligned}$$

Therefore, circulation $= 2\pi k = \phi - \phi_1$

$=$ decrease in ϕ moving once round the circuit.

Hence there is a circulation of amount $2\pi k$ about the cylinder.

$$\text{Also,} \quad \frac{dw}{dz} = \frac{ik}{z}$$

$$\Rightarrow \quad q = \left| -\frac{dW}{dz} \right| = \frac{k}{r}$$

$$\text{i.e. } k = rq$$

Therefore, $k = q$ when $r = 1$

Thus k is the speed at unit distance from the origin.

2.5. Complex Potential for Rectilinear Vortex (Line Vortex). Let us consider a cylindrical vortex tube whose cross-section is a circle of radius a ; surrounded by infinite mass of liquid. We assume that vorticity over the area of the circle is constant and is zero outside the circle.

Let ψ be the stream function, then

$$\bar{\xi} = \nabla^2 \psi \hat{k}$$

$$\begin{aligned} \text{i.e.} \quad \xi &= \nabla^2 \psi = \frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} \\ &= \frac{\partial^2 \psi}{\partial r^2} + \frac{1}{r} \frac{\partial \psi}{\partial r} + \frac{1}{r^2} \frac{\partial^2 \psi}{\partial \theta^2} \end{aligned}$$

Since there is a symmetry about the origin ψ is a function of r only and so $\frac{\partial^2 \psi}{\partial \theta^2} = 0$.

$$\begin{aligned} \therefore \quad \xi &= \frac{1}{r} \frac{d}{dr} \left(r \frac{d\psi}{dr} \right), \text{ for } r < a \\ &= 0, \text{ for } r > a \end{aligned}$$

$$\begin{aligned} \text{i.e.} \quad \frac{d}{dr} \left(r \frac{d\psi}{dr} \right) &= r\xi, \text{ for } r < a \\ &= 0, \text{ for } r > a \end{aligned}$$

Integrating, we find

$$\begin{aligned} r \frac{d\psi}{dr} &= \xi \frac{r^2}{2} + A, \text{ for } r < a \\ &= B, \text{ for } r > a \end{aligned}$$

We are interested in the fluid motion outside the cylinder $|z| = a$. Therefore, integrating the second of the above result, we get

$$\psi = B \log r + C, \text{ for } r > a.$$

The constant C may be chosen to be zero. Further, for $r > a$, the vorticity is zero and the fluid motion is irrotational, therefore velocity potential ϕ exists and is related to ψ as

$$-\frac{1}{r} \frac{\partial \phi}{\partial \theta} = \frac{\partial \psi}{\partial r} = \frac{B}{r}$$

$$\Rightarrow \quad \phi = -B\theta + D$$

$$\Rightarrow \quad \phi = -B\theta, \text{ neglecting } D$$

Let k be the circulation while moving once round the cylinder, then

$$\begin{aligned} k &= \text{decrease in value of } \phi \text{ on describing the circuit once} \\ &= -B [\theta - (\theta + 2\pi)] = 2\pi B \end{aligned}$$

$$\Rightarrow \quad B = k/2\pi = K(\text{say})$$

Thus, $\phi = -K\theta$ and $\psi = K \log r$

Hence $W = \phi + i\psi = -k\theta + iK \log r$

$$\begin{aligned} &= iK (\log r + i\theta) \\ &= iK \log z = i \frac{k}{2\pi} \log z. \end{aligned}$$

If the rectilinear vortex is situated at the point $z = z_0$, then by shifting the origin, we get

$$W = iK \log (z - z_0)$$

If there are vortices of strengths $K_1, K_2, \dots, K_n$ situated at $z_1, z_2, \dots, z_n$ respectively, then the complex potential is

$$W = iK_1 \log(z - z_1) + iK_2 \log(z - z_2) + \dots + iK_n \log(z - z_n).$$

2.6. Remarks (i) By a vortex, we mean a rectilinear vortex or line vortex.

(ii) $K = k/2\pi$, where K is the strength of a vortex and k that of circulation

2.7. Complex Potential for a Spiral Vortex. The combination of a source and a vortex is called a spiral vortex or a **vortex source**.

Let us consider a source of strength m and a vortex of strength K both at the origin. Then the complex potential is

$$\begin{aligned} W &= -m \log z + iK \log z \\ &= (-m + iK) \log z = (-m + iK) \log (re^{i\theta}) \\ &= (-m + iK) (\log r + i\theta) \end{aligned}$$

$$\Rightarrow \quad \phi + i\psi = -m \log r - K\theta + i(-m\theta + K \log r)$$

Therefore, $\phi = -(m \log r + K\theta), \psi = -m\theta + K \log r$

If we go once round the origin, then ϕ decrease by $2\pi K$ and ψ be $2\pi m$.

2.8. Example. Find the complex potential for the motion due to a system consisting of a coincident line-source of strength m per unit length and line-vortex of strength K per unit length in the presence of a circular cylinder of radius a , whose axis is parallel to and at a distance b ($> a$) from the line of the source and vortex. Show that the cylinder is attracted by a force of magnitude

$$2\pi\rho a^2 (m^2 + K^2)/ b(b^2-a^2)$$

per unit length.

Solution. We suppose the line-source and line-vortex to be at the origin, then the complex potential is

$$W = -m \log z + iK \log z = (iK - m) \log z \quad (1)$$

When the circular cylinder $|z - b| = a$ ($b > a$) is inserted, the complex potential, by circle theorem, becomes

$$W = (iK - m) \log z + (-iK - m) \log \left(\frac{a^2}{z - b} + b \right) \quad (2)$$

where

$$|z - b| = a \Rightarrow (\bar{z} - b)(z - b) = a^2$$

$$\Rightarrow \bar{z} - b = \frac{a^2}{b} \Rightarrow \bar{z} = \frac{a^2}{z - b} + b$$

By Blasius theorem, force on the cylinder C is given by

$$X - iY = -\pi\rho \left[\text{sum of residues of } \left(\frac{dW}{dz} \right)^2 \text{ within } C \right] \quad (3)$$

$$\text{Now } \left(\frac{dW}{dz} \right)^2 = \left\{ \frac{iK - m}{z} - \frac{iK + m}{a^2 + b(z - b)} \frac{a^2}{z - b} \right\}^2 \quad (4)$$

The only singularities of $\left(\frac{dW}{dz}\right)^2$ within C are at $z = b$ and $z = b - \frac{a^2}{b}$ since $z = 0$ is not inside C.

Now,

$$\text{residue } (z = b) = -2 \left(\frac{K^2 + m^2}{b} \right) \quad \left| \begin{array}{l} \text{Only product term of (4)} \\ \\ \text{will contribute} \end{array} \right.$$

and

$$\text{residue } (z = b - \frac{a^2}{b}) = \frac{2(K^2 + m^2)b}{(b^2 - a^2)}.$$

Therefore, from (3), we get

$$\begin{aligned} X - iY &= -2\pi\rho (K^2 + m^2) \left[\frac{b}{b^2 - a^2} - \frac{1}{b} \right] \\ &= -2\pi\rho a^2 (K^2 + m^2) / b(b^2 - a^2) \end{aligned}$$

Thus

$$Y = 0, \quad X = -2\pi\rho a^2 (K^2 + m^2) / b(b^2 - a^2).$$

The negative sign implies that the cylinder is attracted towards the origin where the spiral vortex is situated.

2.9. Complex Potential for a Vortex Doublet. Two equal and opposite vortices placed at small distance apart, form a vortex doublet.

Let us consider a vortex of strength K at $z = ae^{i\alpha}$ and another vortex of strength $-K$ at $z = 0$, then the complex potential is

$$\begin{aligned} W &= iK \log (z - ae^{i\alpha}) - iK \log z \\ &= iK \log \left(\frac{z - ae^{i\alpha}}{z} \right) = iK \log \left(1 - \frac{ae^{i\alpha}}{z} \right) \\ &= -iK \left(\frac{ae^{i\alpha}}{z} + \frac{a^2 e^{2i\alpha}}{2z^2} + \dots \right) \end{aligned}$$

As $a \rightarrow 0$, $K \rightarrow \infty$, then $Ka \rightarrow \mu$ and we obtain

$$W = \frac{-i\mu e^{i\alpha}}{z} = \frac{\mu e^{i(\alpha-\pi/2)}}{z}$$

This is the required complex potential for a vortex doublet at the origin.

Also, we note that the complex potential for a doublet at the origin is $\frac{\mu e^{i\alpha}}{z}$.

Thus, it follows that the complex potential of a vortex doublet is the same as that for a doublet with its axes rotated through a right angle.

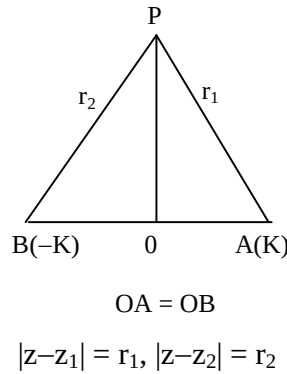
2.10. Image of Vortex in a Plane. Let us consider two line vortices of strengths K and $-K$ per unit length at $A(z = z_1)$ and $B(z = z_2)$ respectively. The complex potential due to these line vortices is

$$W = \phi + i\psi = iK \log(z - z_1) - iK \log(z - z_2)$$

$$\Rightarrow \psi = K \log \left| \frac{z - z_1}{z - z_2} \right| = K \log \frac{r_1}{r_2}$$

If $r_1 = r_2$, then $\psi = K \log 1 = 0$

Thus the plane boundary OP is a streamline so that there is no flow across OP . Hence the line vortex at B with strength $-K$ per unit length is the image of the line vortex at A with strength K per unit length so that A and B are at equal distances from OP .



2.11. Remark. In case of two dimensions (as for sources, sinks and doublets), a vortex means a line vortex and strength means strength per unit length.

2.12. Image of a Vortex in a Circular Cylinder (or in a circle). Let a vortex of strength k be present at $z = d$, then the complex potential is $iK \log(z - d)$. When the cylinder $|z| = a$ is introduced into the fluid, the complex potential, by circle theorem, becomes

$$W = iK \log(z - d) - iK \log\left(\frac{a^2}{z} - d\right)$$

$$\text{i.e.} \quad \phi + i\psi = iK \log(z - d) - iK \log\left(z - \frac{a^2}{d}\right) + iK \log z + \text{constant}$$

(1)

$$\begin{aligned}
 &= iK \left[\log \{(x-d)^2 + y^2\}^{1/2} + i \tan^{-1} \left(\frac{y}{x-d} \right) \right] \\
 &- iK \left[\log \left\{ \left(x - \frac{a^2}{d} \right)^2 + y^2 \right\}^{1/2} + i \tan^{-1} \left(\frac{y}{x - \frac{a^2}{d}} \right) \right] \\
 &+ iK \left[\log \sqrt{x^2 + y^2} + i \tan^{-1} \frac{y}{x} \right]
 \end{aligned}$$

where we have ignored the constant term

$$\Rightarrow \quad \psi = K \log \left[\frac{a \{(a \cos \theta - d)^2 + a^2 \sin^2 \theta\}^{1/2}}{\left\{ \left(a \cos \theta - \frac{a^2}{d} \right)^2 + a^2 \sin^2 \theta \right\}^{1/2}} \right]$$

$$\begin{aligned}
 |z| = a &\Rightarrow z = ae^{i\theta} \\
 \Rightarrow x &= a \cos \theta, \\
 y &= a \sin \theta
 \end{aligned}$$

$$= K \log d = \text{constant.}$$

This shows that the cylinder is a streamline. Thus (1) represents the complex potential of the fluid motion. From (1), we observe that the image of a vortex of strength K at $z = d$ is a vortex of strength $-K$ at the inverse point $z = a^2/d$ together with a vortex of strength K at $z = 0$ i.e. centre of the circle.

2.13. Circulation about a Circular Cylinder in a Uniform Stream. Let a liquid be in motion with a velocity $-U$ along the x -axis. The complex potential due to the stream is Uz . If the circular cylinder of radius a is introduced inside the liquid, then the complex potential, by circle theorem, becomes $Uz + U \frac{a^2}{z}$.

Let there be a circulation k about the cylinder. The complex potential due to circulation is $ik \log z$. Thus the complex potential of the whole system is

$$W = Uz + U \frac{a^2}{z} + ik \log z. \quad (1)$$

$$\Rightarrow -\bar{q} = \frac{dW}{dz} = U - U \frac{a^2}{z^2} + \frac{ik}{z}$$

At the stagnation points, $\bar{q} = 0$ i.e. $q = 0$

$$\Rightarrow U - \frac{Ua^2}{z^2} + \frac{ik}{z} = 0$$

$$\Rightarrow Uz^2 + ikz - Ua^2 = 0$$

$$\Rightarrow z = -\frac{ik}{2U} \pm a \sqrt{1 - \frac{k^2}{4a^2U^2}}$$

Since a and U are constants, therefore the flow potential term depends very much on the magnitude of k . We shall consider three cases.

Case I. When $k < 2aU$ i.e. $\frac{k^2}{4a^2U^2} < 1$, we put

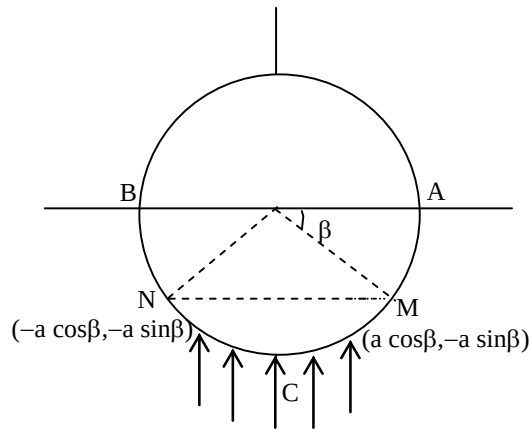
$$\frac{k^2}{4a^2U^2} = \sin^2 \beta \text{ and then}$$

$$z = -ia \sin \beta \pm a \cos \beta$$

Thus the stagnation points are $(a \cos \beta, -a \sin \beta)$ and $(-a \cos \beta, -a \sin \beta)$

Further $|z| = a |\pm \cos \beta - i \sin \beta| = a$

$\therefore$ The stagnation points lie on the boundary of the cylinder. They lie on the line MN below the diameter AB as shown in the fig. The velocity increases above MN and decreases below MN .



Further, from Bernoulli's equation,

$$\frac{p}{\rho} + \frac{1}{2}q^2 = \text{constant}$$

we observe that the pressure decreases above MN and increases below MN.

Thus, there is an increase of pressure by the liquid due to circulation. If there is no circulation, then $k = 0 \Rightarrow \sin\beta = 0$

$$\Rightarrow \beta = 0, \pi, z = \pm a$$

Therefore, MN coincides with AB and thus the stagnation points are at A and B. Therefore we conclude that the circulation brings the stagnation points downwards and put an upward thrust on the cylinder.

Case II. When $k = 2aU$ i.e. $\frac{k^2}{4a^2U^2} = 1$, then $\sin\beta = 1$

$$\Rightarrow \beta = \pi/2, z = -ia \Rightarrow |z| = a$$

and thus the stagnation points coincide at C, the bottom of the cylinder.

Case III. When $k > 2aU$ i.e. $\frac{k^2}{4a^2U^2} > 1$, then we put $\frac{k^2}{4a^2U^2} = \cosh^2\beta$ so that

$$z = a(-i \cosh \beta \pm \sinh \beta)$$

$$= -ia e^\beta, -ia e^{-\beta}$$

$\therefore$ The stagnation points lie on y-axis.

$$\text{Further } |(-iae^\beta)(-iae^{-\beta})| = a^2$$

this shows that the stagnation points are inverse points w.r.t. the circular boundary of the cylinder. One of these points lie inside and other is outside the cylinder. The point which is inside the cylinder does not belong to the motion.

$$\begin{cases} |z_1| = |-iae^\beta| = ae^\beta, \text{ outside the circle} \\ |z_2| = |-iae^{-\beta}| = ae^{-\beta}, \text{ inside the circle} \end{cases}$$

since $ae^{-\beta} < ae^\beta$.

We know that at the **stagnation points (critical points)**, there are two branches of the streamlines which are at right angles to each other. Thus the liquid inside the loop formed at the stagnation points will not be carried by the stream but will circulate round the cylinder

Pressure (Force) on the circular cylinder :- From (1), we have

$$\frac{dW}{dz} = U - \frac{Ua^2}{z^2} + \frac{ik}{z}$$

Therefore, by Blasius theorem,

$$X - iy = \frac{i\rho}{2} \int_C \left(\frac{dW}{dz} \right)^2 dz.$$

$$= -\pi\rho \left(\text{sum of the residues of } \left(\frac{dW}{dz} \right)^2 \text{ within the circle } |z| = a \right)$$

$$\left| \text{By Cauchy's Residue theorem as } \left(\frac{dW}{dz} \right)^2 \text{ is a meromorphic function} \right.$$

where X, Y are components of the pressure of the liquid and ρ is the density of the liquid

$$\text{Now, } \left(\frac{dW}{dz} \right)^2 = U^2 \left(1 - \frac{a^2}{z^2} \right)^2 + \frac{2ikU}{z} \left(1 - \frac{a^2}{z^2} \right) - \frac{k^2}{z^2}$$

The only pole inside the cylinder $|z| = a$ is $z = 0$ i.e. a simple pole. The residue at $z = 0$ is $2ikU$

$$\text{Therefore, } X - iy = -\pi\rho(2ikU)$$

$$\Rightarrow X = 0, Y = 2\pi k\rho U$$

This represents an upward thrust on the cylinder due to circulation. The lifting tendency ($k \neq 0$) is called the **Magnus effect**. The moment M is obtained to be **zero**, since residue is zero in that case.

2.14. Exercise. Show that the complex potential

$$W = U \left(z + \frac{a^2}{z} \right) + ik \log z \text{ represents a possible flow past a}$$

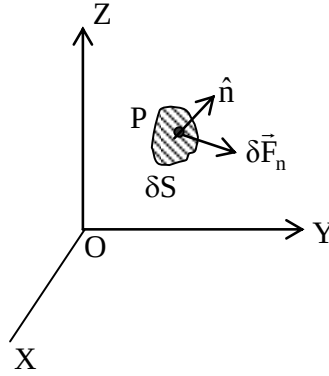
circular cylinder. Sketch the streamlines, find the stagnation points and calculate the force on the cylinder.

2.15. Example. Verify that $W = iK \log \left(\frac{z - ia}{z + ia} \right)$, K and a both real, is the complex potential of a steady flow of liquid about a circular cylinder, the plane

UNIT – VII & IX

1. Stress Components in a Real Fluid

Let δS be a small rigid plane area inserted at a point P in a viscous fluid. Cartesian co-ordinates (x, y, z) are referred to a set of fixed axes OX, OY, OZ. Suppose that $\delta \vec{F}_n$ is the force exerted by the moving fluid on one side of δS , the unit vector $\hat{n}$ being taken to specify the normal at P to δS on this side. We know that in the case of an inviscid fluid, $\delta \vec{F}_n$ is aligned with $\hat{n}$. For a viscous fluid, however, frictional forces are called into play between the fluid and the surface so that $\delta \vec{F}_n$ will also have a component tangential to δS . We suppose the Cartesian components of $\delta \vec{F}_n$ to be $(\delta F_{nx}, \delta F_{ny}, \delta F_{nz})$ so that



$$\delta \vec{F}_n = \delta F_{nx} \hat{i} + \delta F_{ny} \hat{j} + \delta F_{nz} \hat{k}.$$

Then the components of stress parallel to the axes are defined to be σ_{nx} , σ_{ny} , σ_{nz} , where

$$\sigma_{nx} = \lim_{\delta S \rightarrow 0} \frac{\delta F_{nx}}{\delta S} = \frac{dF_{nx}}{dS},$$

$$\sigma_{ny} = \lim_{\delta S \rightarrow 0} \frac{\delta F_{ny}}{\delta S} = \frac{dF_{ny}}{dS},$$

$$\sigma_{nz} = \lim_{\delta S \rightarrow 0} \frac{\delta F_{nz}}{\delta S} = \frac{dF_{nz}}{dS}.$$

In the components σ_{nx} , σ_{ny} , σ_{nz} , the first suffix n denotes the direction of the normal to the elemental plane δS whereas the second suffix x or y or z denotes the direction in which the component is measured.

If we identify $\hat{n}$ in turn with the unit vectors $\hat{i}, \hat{j}, \hat{k}$ in $(\overline{OX}), (\overline{OY}), (\overline{OZ})$, which is achieved by suitably re-orientating δS , we obtain the following three sets of stress components

$$\sigma_{xx}, \sigma_{xy}, \sigma_{xz};$$

$$\sigma_{yx}, \sigma_{yy}, \sigma_{yz};$$

$$\sigma_{zx}, \sigma_{zy}, \sigma_{zz}.$$

The diagonal elements $\sigma_{xx}, \sigma_{yy}, \sigma_{zz}$ of this array are called normal or direct stresses. The remaining six elements are called shearing stresses. For an inviscid fluid, we have

$$\sigma_{xx} = \sigma_{yy} = \sigma_{zz} = -p$$

$$\sigma_{xy} = \sigma_{xz} = \sigma_{yx} = \sigma_{yz} = \sigma_{zx} = \sigma_{zy} = 0$$

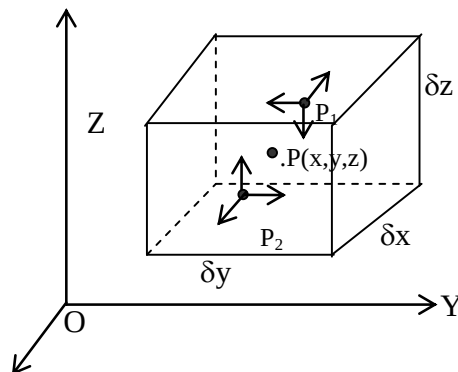
Here, we consider the normal stresses as positive when they are tensile and negative when they are compressive, so that p is the hydrostatic pressure. The matrix

$$\begin{bmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{bmatrix} \quad (1)$$

is called the stress matrix. If its components are known, we can calculate the total forces on any area at any chosen point. The quantities σ_{ij} ($i, j = x, y, z$) are called the components of the stress tensor whose matrix is of the form (1). Further we observe that σ_{ij} is a tensor of order two.

2. Relation Between Rectangular (Cartesian) Components of Stress

Let us consider the motion of a small rectangular parallelopiped of viscous fluid, its centre being $P(x, y, z)$ and its edges of lengths $\delta x, \delta y, \delta z$, parallel to fixed Cartesian axes, as shown in the figure.



X

Let ρ be the density of the fluid. The mass $\rho \delta x \delta y \delta z$ of the fluid element remains constant and the element is presumed to move along with the fluid. In the figure, the points P_1 and P_2 have been taken on the centre of the faces so that they have co-ordinates $\left(x - \frac{\delta x}{2}, y, z\right)$ and $\left(x + \frac{\delta x}{2}, y, z\right)$ respectively.

At $P(x, y, z)$, the force components parallel to $\overline{OX}, \overline{OY}, \overline{OZ}$ on the surface area $\delta y \delta z$ through P and having $\hat{i}$ as unit normal, are

$$(\sigma_{xx} \delta y \delta z, \sigma_{xy} \delta y \delta z, \sigma_{xz} \delta y \delta z)$$

At $P_2\left(x + \frac{\delta x}{2}, y, z\right)$, since $\hat{i}$ is the unit normal measured outwards from the fluid, the corresponding force components across the parallel plane of area $\delta y \delta z$, are

$$\left[\left\{ \sigma_{xx} + \frac{\delta x}{2} \left(\frac{\partial \sigma_{xx}}{\partial x} \right) \right\} \delta y \delta z, \left\{ \sigma_{xy} + \frac{\delta x}{2} \left(\frac{\partial \sigma_{xy}}{\partial x} \right) \right\} \delta y \delta z, \left\{ \sigma_{xz} + \frac{\delta x}{2} \left(\frac{\partial \sigma_{xz}}{\partial x} \right) \right\} \delta y \delta z \right].$$

For the parallel plane through $P_1\left(x - \frac{\delta x}{2}, y, z\right)$, since $-\hat{i}$ is the unit normal drawn outwards from the fluid element, the corresponding components are

$$\left[- \left\{ \sigma_{xx} - \frac{\delta x}{2} \left(\frac{\partial \sigma_{xx}}{\partial x} \right) \right\} \delta y \delta z, - \left\{ \sigma_{xy} - \frac{\delta x}{2} \left(\frac{\partial \sigma_{xy}}{\partial x} \right) \right\} \delta y \delta z, - \left\{ \sigma_{xz} - \frac{\delta x}{2} \left(\frac{\partial \sigma_{xz}}{\partial x} \right) \right\} \delta y \delta z \right]$$

The forces on the parallel planes through P_1 and P_2 are equivalent to a single force at P with components

$$\left[\frac{\partial \sigma_{xx}}{\partial x}, \frac{\partial \sigma_{xy}}{\partial x}, \frac{\partial \sigma_{xz}}{\partial x} \right] \delta x \delta y \delta z$$

together with couples whose moments (upto third order terms) are

$$\begin{cases} -\sigma_{xz} \delta x \delta y \delta z \text{ about } Oy, \\ \sigma_{xy} \delta x \delta y \delta z \text{ about } Oz. \end{cases}$$

Similarly, the pair of faces perpendicular to the y axis give a force at P having components

$$\left[\frac{\partial \sigma_{yx}}{\partial y}, \frac{\partial \sigma_{yy}}{\partial y}, \frac{\partial \sigma_{yz}}{\partial y} \right] \delta x \delta y \delta z$$

together with couples of moments

$$\begin{cases} -\sigma_{yx} \delta x \delta y \delta z \text{ about Oz,} \\ \sigma_{yz} \delta x \delta y \delta z \text{ about Ox.} \end{cases}$$

The pair of faces perpendicular to the z-axis give a force at P having components

$$\left[\frac{\partial \sigma_{zx}}{\partial z}, \frac{\partial \sigma_{zy}}{\partial z}, \frac{\partial \sigma_{zz}}{\partial z} \right] \delta x \delta y \delta z$$

together with couples of moments

$$\begin{cases} -\sigma_{zy} \delta x \delta y \delta z \text{ about Ox,} \\ \sigma_{zx} \delta x \delta y \delta z \text{ about Oy.} \end{cases}$$

Combining the surface forces of all six faces of the parallelopiped, we observe that they reduce to a single force at P having components

$$\left[\left(\frac{\partial \sigma_{xx}}{\partial x} + \frac{\partial \sigma_{yx}}{\partial y} + \frac{\partial \sigma_{zx}}{\partial z} \right), \left(\frac{\partial \sigma_{xy}}{\partial x} + \frac{\partial \sigma_{yy}}{\partial y} + \frac{\partial \sigma_{zy}}{\partial z} \right), \left(\frac{\partial \sigma_{xz}}{\partial x} + \frac{\partial \sigma_{yz}}{\partial y} + \frac{\partial \sigma_{zz}}{\partial z} \right) \right] \delta x \delta y \delta z,$$

together with a vector couple having Cartesian components

$$[(\sigma_{yz} - \sigma_{zy}), (\sigma_{zx} - \sigma_{xz}), (\sigma_{xy} - \sigma_{yx})] \delta x \delta y \delta z.$$

Now, suppose the external body forces acting at P are $[X, Y, Z]$ per unit mass, so that the total body force on the element has components $[X, Y, Z] \rho \delta x \delta y \delta z$. Let us take moments about $\hat{i}$ –direction through P. Then, we have

Total moment of forces = Moment of inertia about axis $\times$ Angular acceleration

i.e. $(\sigma_{yz} - \sigma_{zy}) \delta x \delta y \delta z + \text{terms of 4}^{\text{th}} \text{ order in } \delta x, \delta y, \delta z = \text{terms of 5}^{\text{th}} \text{ order in } \delta x, \delta y, \delta z.$

Thus, to the third order of smallness in δx , δy , δz , we obtain

$$(\sigma_{yz} - \sigma_{zy}) \delta x \delta y \delta z = 0$$

Hence, as the considered fluid element becomes vanishingly small, we obtain

$$\sigma_{yz} = \sigma_{zy}.$$

Similarly, we get

$$\sigma_{zx} = \sigma_{xz}, \quad \sigma_{xy} = \sigma_{yx}$$

Thus, the stress matrix is diagonally symmetric and contains only six unknowns. In other words, we have proved that

$$\sigma_{ij} = \sigma_{ji}, \quad (i, j = x, y, z)$$

i.e. σ_{ij} is symmetric.

In fact, σ_{ij} is a symmetric second order Cartesian tensor.

2.1. Transnational Motion of Fluid Element. Considering the surface forces and body forces, we note (from the previous article) that the total force component in the $\hat{i}$ -direction, acting on the fluid element at point $P(x, y, z)$, is

$$\left(\frac{\partial \sigma_{xx}}{\partial x} + \frac{\partial \sigma_{yx}}{\partial y} + \frac{\partial \sigma_{zx}}{\partial z} \right) \delta x \delta y \delta z + X \rho \delta x \delta y \delta z \quad (1)$$

where (X, Y, Z) is the body force per unit mass and ρ being the density of the viscous fluid. As the mass $\rho \delta x \delta y \delta z$ is considered constant, if $\bar{q} = (u, v, w)$ be the velocity of point P at time t , then the equation of motion in the $\hat{i}$ -direction is

$$\left(\frac{\partial \sigma_{xx}}{\partial x} + \frac{\partial \sigma_{yx}}{\partial y} + \frac{\partial \sigma_{zx}}{\partial z} \right) \delta x \delta y \delta z + \rho X \delta x \delta y \delta z = (\rho \delta x \delta y \delta z) \frac{du}{dt}$$

or
$$\frac{\partial \sigma_{xx}}{\partial x} + \frac{\partial \sigma_{yx}}{\partial y} + \frac{\partial \sigma_{zx}}{\partial z} + \rho X = \rho \frac{du}{dt} \quad (2)$$

If $u = u(x, y, z, t)$, then

$$\frac{du}{dt} = \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} + w \frac{\partial u}{\partial z} \quad \text{where} \quad \frac{d}{dt} \equiv \frac{\partial}{\partial t} + \bar{q} \cdot \nabla$$

Thus, (2) becomes

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} + w \frac{\partial u}{\partial z} = X + \frac{1}{\rho} \left(\frac{\partial \sigma_{xx}}{\partial x} + \frac{\partial \sigma_{yx}}{\partial y} + \frac{\partial \sigma_{zx}}{\partial z} \right) \quad (3)$$

Similarly the equations of motion in $\hat{j}$ and $\hat{k}$ directions are

$$\frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} + w \frac{\partial v}{\partial z} = Y + \frac{1}{\rho} \left(\frac{\partial \sigma_{xy}}{\partial x} + \frac{\partial \sigma_{yy}}{\partial y} + \frac{\partial \sigma_{zy}}{\partial z} \right) \quad (4)$$

$$\frac{\partial w}{\partial t} + u \frac{\partial w}{\partial x} + v \frac{\partial w}{\partial y} + w \frac{\partial w}{\partial z} = Z + \frac{1}{\rho} \left(\frac{\partial \sigma_{xz}}{\partial x} + \frac{\partial \sigma_{yz}}{\partial y} + \frac{\partial \sigma_{zz}}{\partial z} \right) \quad (5)$$

Equations (3), (4), (5) provide the equations of motion of the fluid element at $P(x, y, z)$.

In tensor form, if the co-ordinates are x_i , the velocity components u_i , the body force components X_i , where $i = 1, 2, 3$, the equations of motion can be expressed as

$$\frac{\partial u_i}{\partial t} + u_j \frac{\partial u_i}{\partial x_j} = X_i + \frac{1}{\rho} \sigma_{ji,j} \quad (i, j = 1, 2, 3).$$

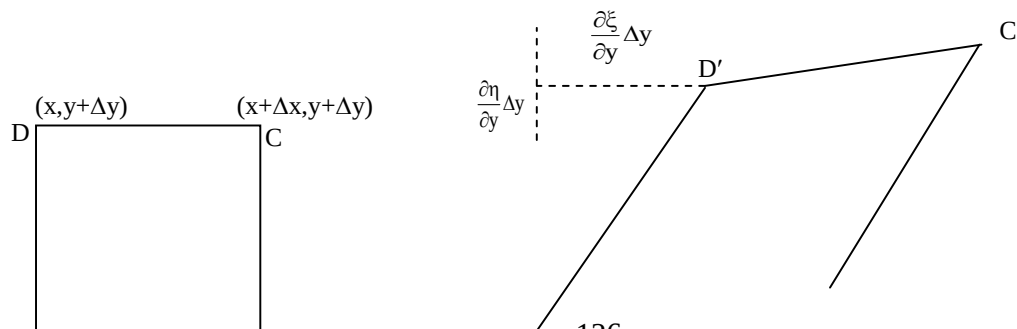
3. Nature of Strains (Rates of Strain)

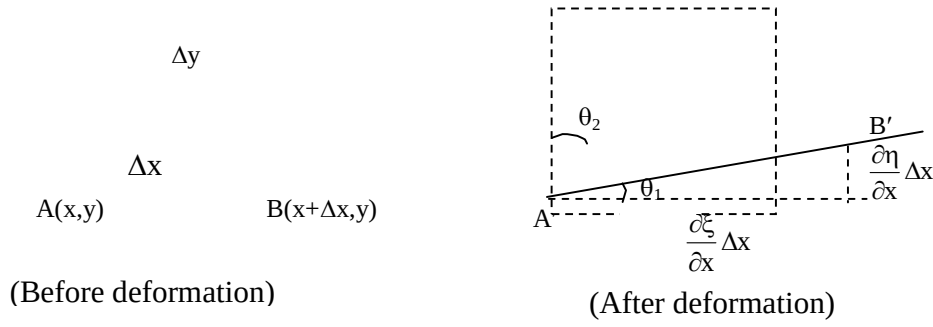
The change in the relative position of the parts of the body under some force, is termed as **deformation**. By Hooke's law, the stress is proportional to strain in case of elastic bodies, while in case of non-elastic bodies the stress is proportional to the **rate of strain**.

Strain is of two kinds, the normal and the shearing. The ratio of change in length to the original length of a line element is called **normal (or direct) strain**. The shearing strain measures the change in angle between two line elements from the natural state to some standard state. We shall consider two dimensional case and then extend it to three dimensions. Let us consider a rectangular element ABCD of an elastic solid with co-ordinates of A as (x, y) and length of sides as Δx and Δy in the natural state.

Let the point A. be defined to a point $A'(x + \xi, y + \eta)$ then

$$B(x + \Delta x, y) \text{ goes to } B'(x + \xi + \Delta x + \frac{\partial \xi}{\partial x} \Delta x, y + \eta + \frac{\partial \eta}{\partial x} \Delta x)$$





The point $D(x, y + \Delta y)$ goes to the point

$$D'(x + \xi + \frac{\partial \xi}{\partial y} \Delta y, y + \eta + \Delta y + \frac{\partial \eta}{\partial y} \Delta y).$$

Therefore, projected lengths of $A'B'$ along x and y axes are $\Delta x + \frac{\partial \xi}{\partial x} \Delta x$ and

$$\frac{\partial \eta}{\partial x} \Delta x$$

Thus,

$$(A'B')^2 = \left(\Delta x + \frac{\partial \xi}{\partial x} \Delta x \right)^2 + \left(\frac{\partial \eta}{\partial x} \Delta x \right)^2 \quad (1)$$

The normal strain along x-axis is defined by

$$\epsilon_{xx} = \frac{A'B' - AB}{AB}$$

$$\Rightarrow A'B' = (1 + \epsilon_{xx}) AB = (1 + \epsilon_{xx}) \Delta x \quad | \quad AB = \Delta x \quad (2)$$

From (1) & (2), we have

$$(1 + \epsilon_{xx})^2 (\Delta x)^2 = (\Delta x)^2 \left[\left(1 + \frac{\partial \xi}{\partial x} \right)^2 + \left(\frac{\partial \eta}{\partial x} \right)^2 \right]$$

$$\Rightarrow (1 + \epsilon_{xx})^2 = \left(1 + \frac{\partial \xi}{\partial x} \right)^2 + \left(\frac{\partial \eta}{\partial x} \right)^2$$

From here, to the first order terms only, we get

$$\epsilon_{xx} = \frac{\partial \xi}{\partial x}.$$

Similarly, the normal strain along the y-axis is

$$\epsilon_{yy} = \frac{\partial \eta}{\partial y}$$

The shearing strain γ_{xy} at the point A is the change in the angle between the sides AB and AD. The right angle $\angle DAB$ between AB and AD is diminished by $\gamma_{xy} = \theta_1 + \theta_2 = \tan\theta_1 + \tan\theta_2$, θ_1 & θ_2 being small.

$$\begin{aligned} \text{i.e. } \gamma_{xy} &= \frac{\frac{\partial \eta}{\partial x} \Delta x}{\left(1 + \frac{\partial \xi}{\partial x}\right) \Delta x} + \frac{\frac{\partial \xi}{\partial y} \Delta y}{\left(1 + \frac{\partial \eta}{\partial y}\right) \Delta y} \\ &= \frac{\partial \eta}{\partial x} \left(1 + \frac{\partial \xi}{\partial x}\right)^{-1} + \frac{\partial \xi}{\partial y} \left(1 + \frac{\partial \eta}{\partial y}\right)^{-1} \\ \epsilon_{xy} &= \frac{1}{2}(\gamma_{xy}) = \frac{1}{2} \left(\frac{\partial \eta}{\partial x} + \frac{\partial \xi}{\partial y} \right), \quad \text{upto first order.} \end{aligned}$$

We observe that the strains have the nature of change in displacement in a given unit length in a given direction. Hence strain is a tensor of order two.

In the case of fluids, there is no resistance to deformation but only to the time rate of deformation. Hence in fluid dynamics the rate of change of strain with time i.e. rate of strain is to be used in place of strain in elasticity. Thus, for viscous fluids, replacing strains by rates of strain, the corresponding results are obtained to be

$$\begin{aligned} \epsilon_{xx} &= \frac{\partial}{\partial t} \left(\frac{\partial \xi}{\partial x} \right) = \frac{\partial}{\partial x} \left(\frac{\partial \xi}{\partial t} \right) = \frac{\partial}{\partial x} (u) = \frac{\partial u}{\partial x} \\ \epsilon_{yy} &= \frac{\partial v}{\partial y}, v = \frac{\partial \eta}{\partial t} \\ \epsilon_{xy} &= \frac{1}{2} \frac{\partial}{\partial t} \left(\frac{\partial \eta}{\partial x} + \frac{\partial \xi}{\partial y} \right) = \frac{1}{2} \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \right) \end{aligned}$$

In case of three dimensions, these become

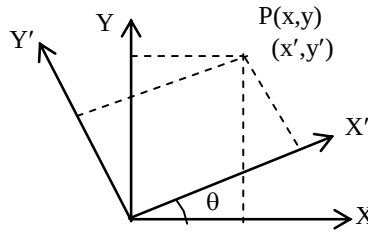
$$\left. \begin{aligned} \epsilon_{xx} &= \frac{\partial u}{\partial x}, \epsilon_{yy} = \frac{\partial v}{\partial y}, \epsilon_{zz} = \frac{\partial w}{\partial z} \\ \epsilon_{xy} &= \frac{1}{2}(\gamma_{xy}) = \frac{1}{2}\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right) \\ \epsilon_{yz} &= \frac{1}{2}(\gamma_{yz}) = \frac{1}{2}\left(\frac{\partial w}{\partial y} + \frac{\partial v}{\partial z}\right) \\ \epsilon_{zx} &= \frac{1}{2}(\gamma_{zx}) = \frac{1}{2}\left(\frac{\partial w}{\partial x} + \frac{\partial u}{\partial z}\right) \end{aligned} \right\} \quad (A)$$

where u, v, w are the velocity components of the viscous fluid along x, y, z axis respectively.

The six quantities $\epsilon_{xx}, \epsilon_{yy}, \epsilon_{zz}, \gamma_{xy}, \gamma_{yz}, \gamma_{zx}$ in (A) are called components of the rates of strain or **gradients of velocity**

3.1. Transformation of Rates of Strain.

We shall obtain the rates of strain in term of the new co-ordinates x', y' , changing from x, y to x', y' . Let us obtain the new axes by rotating the original axes through angle θ and let $l = \cos\theta, m = \sin\theta$



Then $x' = lx + my, y' = -mx + ly$

$\Rightarrow x = lx' - my', y = mx' + ly'$

Further, $\frac{\partial}{\partial t}(x') = \frac{\partial}{\partial t}(lx + my)$

$\Rightarrow u' = lu + mv$

and $v' = -mu + lv$

Also, $(OP)^2 = x^2 + y^2 = x'^2 + y'^2 \quad | \because \text{they are still perpendicular}$

Now, $\epsilon'_{xx} = \frac{\partial u'}{\partial x'} = \left(\frac{\partial u'}{\partial x}\right)\frac{\partial x}{\partial x'} + \left(\frac{\partial u'}{\partial y}\right)\frac{\partial y}{\partial x'}$

or $\epsilon'_{xx} = \left(l\frac{\partial u}{\partial x} + m\frac{\partial v}{\partial x}\right)l + \left(l\frac{\partial u}{\partial y} + m\frac{\partial v}{\partial y}\right)m$

$$= l^2\frac{\partial u}{\partial x} + m^2\frac{\partial v}{\partial y} + lm\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right)$$

$$= l^2 \epsilon_{xx} + m^2 \epsilon_{yy} + lm \gamma_{xy}$$

Similarly $\epsilon'_{yy} = \frac{\partial v'}{\partial y'} = m^2 \epsilon_{xx} + l^2 \epsilon_{yy} - lm \gamma_{xy}$

$$\gamma'_{xy} = \frac{\partial u'}{\partial y'} + \frac{\partial v'}{\partial x'} = 2lm (\epsilon_{yy} - \epsilon_{xx}) + (l^2 - m^2) \gamma_{xy}.$$

which are the rates of strain of the new system in terms of rates of strain in the original system. If we put back $l = \cos\theta$, $m = \sin\theta$, then

$$\left. \begin{aligned} \epsilon'_{xx} &= \frac{\epsilon_{xx} + \epsilon_{yy}}{2} + \frac{\epsilon_{xx} - \epsilon_{yy}}{2} \cos 2\theta + \frac{\gamma_{xy}}{2} \sin 2\theta \\ \epsilon'_{yy} &= \frac{\epsilon_{xx} + \epsilon_{yy}}{2} - \frac{\epsilon_{xx} - \epsilon_{yy}}{2} \cos 2\theta - \frac{\gamma_{xy}}{2} \sin 2\theta \\ \epsilon'_{xy} &= \frac{1}{2} (\gamma'_{xy}) = -\frac{\epsilon_{xx} - \epsilon_{yy}}{2} \sin 2\theta + \frac{\gamma_{xy}}{2} \cos 2\theta \end{aligned} \right\} \quad (B)$$

These equations give the transformation formulae for the rates of strain.

We observe that the rate of strain is also a tensor of order two, there must exist at least two invariants of the rate of strain to the choice of co-ordinate systems. These can be obtained as follows.

$$\begin{aligned} \epsilon'_{xx} + \epsilon'_{yy} &= (l^2 + m^2) (\epsilon_{xx} + \epsilon_{yy}) \\ &= \epsilon_{xx} + \epsilon_{yy} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = \text{div } \bar{q}, \quad \bar{q} = (u, v) \end{aligned} \quad (1)$$

$$\begin{aligned} \epsilon'_{xx} \epsilon'_{yy} - \frac{(\gamma'_{xy})^2}{4} &= (l^2 \epsilon_{xx} + m^2 \epsilon_{yy} + lm \gamma_{xy}) (m^2 \epsilon_{xx} + l^2 \epsilon_{yy} - lm \gamma_{xy}) \\ &\quad - \frac{1}{4} [2lm (\epsilon_{yy} - \epsilon_{xx}) + (l^2 - m^2) \gamma_{xy}]^2 \\ &= (l^4 + 2l^2 m^2 + m^4) \epsilon_{xx} \epsilon_{yy} - \frac{\gamma_{xy}^2}{4} (l^4 + 2l^2 m^2 + m^4) \\ &= \epsilon_{xx} \epsilon_{yy} - \frac{\gamma_{xy}^2}{4} \end{aligned} \quad (2)$$

Equation (1) shows that the divergence of the velocity vector at a given point is independent of the orientation of the co-ordinate axes. Equation (2) is related to the dissipation function. i.e. loss of energy due to viscosity.

Let us now consider the general case of the rates of strain in three dimensions. The direction cosines between x, y, z and x', y', z' are related as follows.

	x	y	z
x'	l_1	m_1	n_1
y'	l_2	m_2	n_2
z'	l_3	m_3	n_3

The relations between co-ordinates in the two systems are

$$x' = l_1x + m_1y + n_1z$$

$$y' = l_2x + m_2y + n_2z$$

$$z' = l_3x + m_3y + n_3z$$

and

$$x = l_1x' + l_2y' + l_3z'$$

$$y = m_1x' + m_2y' + m_3z'$$

$$z = n_1x' + n_2y' + n_3z'$$

From here, we get

$$u' = l_1u + m_1v + n_1w$$

$$v' = l_2u + m_2v + n_2w$$

$$w' = l_3u + m_3v + n_3w$$

We shall use these relations to find out the rates of strain w. r. t. the new co-ordinates x', y', z' .

Let us work out

$$\begin{aligned}
 \epsilon'_{xx} &= \frac{\partial u'}{\partial x'} = \frac{\partial u'}{\partial x} \frac{\partial x}{\partial x'} + \frac{\partial u'}{\partial y} \frac{\partial y}{\partial x'} + \frac{\partial u'}{\partial z} \frac{\partial z}{\partial x'} \\
 &= \left(l_1 \frac{\partial u}{\partial x} + m_1 \frac{\partial v}{\partial x} + n_1 \frac{\partial w}{\partial x} \right) l_1 \\
 &\quad + \left(l_1 \frac{\partial u}{\partial y} + m_1 \frac{\partial v}{\partial y} + n_1 \frac{\partial w}{\partial y} \right) m_1 \\
 &\quad + \left(l_1 \frac{\partial u}{\partial z} + m_1 \frac{\partial v}{\partial z} + n_1 \frac{\partial w}{\partial z} \right) n_1 \\
 &= l_1^2 \epsilon_{xx} + m_1^2 \epsilon_{yy} + n_1^2 \epsilon_{zz} + l_1 m_1 \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \right) \\
 &\quad + m_1 n_1 \left(\frac{\partial v}{\partial z} + \frac{\partial w}{\partial y} \right) + n_1 l_1 \left(\frac{\partial w}{\partial x} + \frac{\partial u}{\partial z} \right) \\
 &= l_1^2 \epsilon_{xx} + m_1^2 \epsilon_{yy} + n_1^2 \epsilon_{zz} + l_1 m_1 \gamma_{xy} + m_1 n_1 \gamma_{yz} + n_1 l_1 \gamma_{zx}
 \end{aligned}$$

Similarly, we have

$$\begin{aligned}
 \epsilon'_{yy} &= \frac{\partial v'}{\partial y'} = l_2^2 \epsilon_{xx} + m_2^2 \epsilon_{yy} + n_2^2 \epsilon_{zz} + l_2 m_2 \gamma_{xy} + m_2 n_2 \gamma_{yz} + n_2 l_2 \gamma_{zx} \\
 \epsilon'_{zz} &= \frac{\partial w'}{\partial z'} = l_3^2 \epsilon_{xx} + m_3^2 \epsilon_{yy} + n_3^2 \epsilon_{zz} + l_3 m_3 \gamma_{xy} + m_3 n_3 \gamma_{yz} + n_3 l_3 \gamma_{zx} \\
 \gamma'_{xy} &= \frac{\partial v'}{\partial x'} + \frac{\partial u'}{\partial y'} = \frac{\partial v'}{\partial x} \frac{\partial x}{\partial x'} + \frac{\partial v'}{\partial y} \frac{\partial y}{\partial x'} + \frac{\partial v'}{\partial z} \frac{\partial z}{\partial x'} + \frac{\partial u'}{\partial x} \frac{\partial x}{\partial y'} + \frac{\partial u'}{\partial y} \frac{\partial y}{\partial y'} + \frac{\partial u'}{\partial z} \frac{\partial z}{\partial y'} \\
 &= 2l_1 l_2 \epsilon_{xx} + 2m_1 m_2 \epsilon_{yy} + 2n_1 n_2 \epsilon_{zz} \\
 &\quad + (l_1 m_2 + m_1 l_2) \gamma_{xy} + (m_1 n_2 + n_1 m_2) \gamma_{yz} + (n_1 l_2 + l_1 n_2) \gamma_{zx} \\
 \gamma'_{yz} &= \frac{\partial w'}{\partial y'} + \frac{\partial v'}{\partial z'} = 2l_2 l_3 \epsilon_{xx} + 2m_2 m_3 \epsilon_{yy} + 2n_2 n_3 \epsilon_{zz} \\
 &\quad + (l_2 m_3 + m_2 l_3) \gamma_{xy} + (m_2 n_3 + n_2 m_3) \gamma_{yz} + (n_2 l_3 + l_2 n_3) \gamma_{zx}
 \end{aligned}$$

$$\begin{aligned}\gamma'_{zx} &= \frac{\partial u'}{\partial z'} + \frac{\partial w'}{\partial x'} = 2l_3l_1 \epsilon_{xx} + 2m_3m_1 \epsilon_{yy} + 2n_3n_1 \epsilon_{zz} \\ &\quad + (l_3m_1 + m_3l_1) \gamma_{xy} + (m_3n_1 + n_3m_1) \gamma_{yz} + (n_3l_1 + l_3n_1) \gamma_{zx}\end{aligned}$$

From here, we find

$$\begin{aligned}\epsilon'_{xx} + \epsilon'_{yy} + \epsilon'_{zz} &= (l_1^2 + l_2^2 + l_3^2) \epsilon_{xx} + (m_1^2 + m_2^2 + m_3^2) \epsilon_{yy} \\ &\quad + (n_1^2 + n_2^2 + n_3^2) \epsilon_{zz} + (l_1m_1 + l_2m_2 + l_3m_3) \gamma_{xy} \\ &\quad + (m_1n_1 + m_2n_2 + m_3n_3) \gamma_{yz} + (n_1l_1 + n_2l_2 + n_3l_3) \gamma_{zx} \\ &= \epsilon_{xx} + \epsilon_{yy} + \epsilon_{zz}\end{aligned}$$

where we have used the orthogonality relations

$$l_1^2 + l_2^2 + l_3^2 = 1 \text{ etc}$$

and $l_1m_1 + l_2m_2 + l_3m_3 = 0$ etc.

Thus we conclude that

$$\begin{aligned}\epsilon'_{xx} + \epsilon'_{yy} + \epsilon'_{zz} &= \epsilon_{xx} + \epsilon_{yy} + \epsilon_{zz} \\ &= \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = \text{div } \bar{q}\end{aligned}$$

is invariant.

Similarly,

$$\begin{aligned}\epsilon'_{xx} \epsilon'_{yy} + \epsilon'_{yy} \epsilon'_{zz} + \epsilon'_{zz} \epsilon'_{xx} &- \frac{1}{4} [(\gamma'_{xy})^2 + (\gamma'_{yz})^2 + (\gamma'_{zx})^2] \\ &= \epsilon_{xx} \epsilon_{yy} + \epsilon_{yy} \epsilon_{zz} + \epsilon_{zz} \epsilon_{xx} - \frac{1}{4} [(\gamma_{xy})^2 + (\gamma_{yz})^2 + (\gamma_{zx})^2]\end{aligned}$$

is also invariant.

3.2. Remark. The stress tensor σ_{ij} and the rates of strain ϵ_{ij} follow the same rules of transformation. Thus, the three equations in (B) can also be written for stress components so that we get the relations between the original and the new stress components as

$$\left. \begin{aligned} \sigma'_{xx} &= \frac{\sigma_{xx} + \sigma_{yy}}{2} + \frac{\sigma_{xx} - \sigma_{yy}}{2} \cos 2\theta + \sigma_{xy} \sin 2\theta \\ \sigma'_{yy} &= \frac{\sigma_{xx} + \sigma_{yy}}{2} - \frac{\sigma_{xx} - \sigma_{yy}}{2} \cos 2\theta - \sigma_{xy} \sin 2\theta \\ \sigma'_{xy} &= -\frac{\sigma_{xx} - \sigma_{yy}}{2} \sin 2\theta + \sigma_{xy} \cos 2\theta \end{aligned} \right\} \quad (C)$$

4. Relations Between the Stress and Gradients of Velocity

(Equivalence of Hooke's Law in Case of Viscous Fluids)

In elasticity, generalized Hooke's law gives a relation between the stress and the strain components.

For viscous fluid, the following assumptions are to be made to find the relations between the stress and the rate of strain.

- (i) The stress components may be expressed as linear functions of rates of strain components.
- (ii) The relations between stress and rates of strain are invariant w.r.t rotation and reflection of co-ordinate axes (symmetry).
- (iii) The stress components reduce to the hydrostatic pressure when all the gradients of velocity are zero.

i.e. $\sigma_{xx} = -p = \sigma_{yy} = \sigma_{zz}, \epsilon_{xx} = \frac{\partial u}{\partial x} = 0 = \epsilon_{yy} = \epsilon_{zz}.$

First we consider two dimensional case and then we extend it to three dimensions.

Under the assumption (i), we can write

$$\begin{aligned} \sigma_{xx} &= A_1 \epsilon_{xx} + B_1 \epsilon_{yy} + C_1 \gamma_{xy} + D_1 \\ \sigma_{yy} &= A_2 \epsilon_{xx} + B_2 \epsilon_{yy} + C_2 \gamma_{xy} + D_2 \\ \sigma_{xy} &= A_3 \epsilon_{xx} + B_3 \epsilon_{yy} + C_3 \gamma_{xy} + D_3 \end{aligned} \quad (1)$$

where A's, B's, C's and D's are constants to be determined.

From the assumption (ii), we have

$$\begin{aligned} \sigma'_{xx} &= A_1 \epsilon'_{xx} + B_1 \epsilon'_{yy} + C_1 \gamma'_{xy} + D_1 \\ \sigma'_{yy} &= A_2 \epsilon'_{xx} + B_2 \epsilon'_{yy} + C_2 \gamma'_{xy} + D_2 \end{aligned} \quad (2)$$

$$\sigma'_{xy} = A_3 \epsilon'_{xx} + B_3 \epsilon'_{yy} + C_3 \gamma'_{xy} + D_3$$

But the relations between the original and the new stress components are (from equation (C))

$$\left. \begin{aligned} \sigma'_{xx} &= \frac{\sigma_{xx} + \sigma_{yy}}{2} + \frac{\sigma_{xx} - \sigma_{yy}}{2} \cos 2\theta + \sigma_{xy} \sin 2\theta \\ \sigma'_{yy} &= \frac{\sigma_{xx} + \sigma_{yy}}{2} - \frac{\sigma_{xx} - \sigma_{yy}}{2} \cos 2\theta - \sigma_{xy} \sin 2\theta \\ \sigma'_{xy} &= -\frac{\sigma_{xx} - \sigma_{yy}}{2} \sin 2\theta + \sigma_{xy} \cos 2\theta \end{aligned} \right\} \quad (3)$$

Using the equation (1) in 1st of (3), we get

$$\begin{aligned} \sigma'_{xx} &= \frac{1}{2} (A_1 + A_2) \epsilon_{xx} + \frac{1}{2} (B_1 + B_2) \epsilon_{yy} + \frac{1}{2} (C_1 + C_2) \gamma_{xy} \\ &+ \frac{1}{2} (D_1 + D_2) + \frac{1}{2} (A_1 - A_2) \epsilon_{xx} \cos 2\theta \\ &+ \frac{1}{2} (B_1 - B_2) \epsilon_{yy} \cos 2\theta + \frac{1}{2} (C_1 - C_2) \gamma_{xy} \cos 2\theta \\ &+ \frac{1}{2} (D_1 - D_2) \cos 2\theta + (A_3 \epsilon_{xx} + B_3 \epsilon_{yy} + C_3 \gamma_{xy} + D_3) \sin 2\theta \end{aligned} \quad (4)$$

Also, the relations between the original and the new rates of strain are

$$\left. \begin{aligned} \epsilon'_{xx} &= \frac{\epsilon_{xx} + \epsilon_{yy}}{2} + \frac{\epsilon_{xx} - \epsilon_{yy}}{2} \cos 2\theta + \frac{\gamma_{xy}}{2} \sin 2\theta \\ \epsilon'_{yy} &= \frac{\epsilon_{xx} + \epsilon_{yy}}{2} - \frac{\epsilon_{xx} - \epsilon_{yy}}{2} \cos 2\theta - \frac{\gamma_{xy}}{2} \sin 2\theta \\ \gamma'_{xy} &= -\frac{\epsilon_{xx} - \epsilon_{yy}}{2} \sin 2\theta + \frac{\gamma_{xy}}{2} \cos 2\theta \end{aligned} \right\} \quad (5)$$

Using equation (5) in 1st of equations (2), we get

$$\begin{aligned} \sigma'_{xx} &= \frac{A_1}{2} (\epsilon_{xx} + \epsilon_{yy}) + \frac{A_1}{2} (\epsilon_{xx} - \epsilon_{yy}) \cos 2\theta + \frac{A_1}{2} \gamma_{xy} \sin 2\theta \\ &+ \frac{B_1}{2} (\epsilon_{xx} + \epsilon_{yy}) - \frac{B_1}{2} (\epsilon_{xx} - \epsilon_{yy}) \cos 2\theta - \frac{B_1}{2} \gamma_{xy} \sin 2\theta \end{aligned}$$

$$-C_1(\epsilon_{xx} - \epsilon_{yy}) \sin 2\theta + C_1 \gamma_{xy} \cos 2\theta + D_1 \quad (6)$$

Comparing co-efficients in (4) & (6), we get

$$\begin{aligned} \frac{A_1}{2} (1+\cos 2\theta) + \frac{A_2}{2} (1-\cos 2\theta) + A_3 \sin 2\theta \\ = \frac{A_1}{2} (1 + \cos 2\theta) + \frac{B_1}{2} (1-\cos 2\theta) - C_1 \sin 2\theta \quad | \epsilon_{xx} \end{aligned}$$

$$\begin{aligned} \frac{B_1}{2} (1+\cos 2\theta) + \frac{B_2}{2} (1-\cos 2\theta) + B_3 \sin 2\theta \\ = \frac{A_1}{2} (1 - \cos 2\theta) + \frac{B_1}{2} (1+\cos 2\theta) + C_1 \sin 2\theta \quad | \epsilon_{yy} \end{aligned}$$

$$\begin{aligned} \frac{C_1}{2} (1+\cos 2\theta) + \frac{C_2}{2} (1-\cos 2\theta) + C_3 \sin 2\theta \\ = \frac{A_1}{2} \sin 2\theta - \frac{B_1}{2} \sin 2\theta + C_1 \cos 2\theta \quad | \gamma_{xy} \end{aligned}$$

$$\frac{D_1}{2} (1+\cos 2\theta) + \frac{D_2}{2} (1-\cos 2\theta) + D_3 \sin 2\theta = D_1$$

From these equations, we get

$$A_2 = B_1 = B(\text{say}), B_2 = A_1 = A(\text{say})$$

$$C_2 = A_3 = -C_1 = -B_3 = -C(\text{say})$$

$$C_3 = \frac{A_1 - B_1}{2} = \frac{A - B}{2}, D_1 = D_2 = D(\text{say}), D_3 = 0$$

The stress components in terms of the rates of strain are now obtained to be

$$\left. \begin{aligned} \sigma_{xx} &= A\epsilon_{xx} + B\epsilon_{yy} + C\gamma_{xy} + D \\ \sigma_{yy} &= B\epsilon_{xx} + A\epsilon_{yy} - C\gamma_{xy} + D \\ \sigma_{xy} &= -C(\epsilon_{xx} - \epsilon_{yy}) + \frac{A - B}{2}\gamma_{xy} \end{aligned} \right] \quad (7)$$

To find A, B, C and D, we make use of the assumption that there is symmetry of the fluid about the co-ordinate axes.

Let us take the symmetry w.r.t. the y-axis. If (x_1, y_1) are the new co-ordinates of the point with co-ordinates (x, y) , then

$$x_1 = -x, y_1 = y$$

i.e. $u_1 = -u, v_1 = v$

The rates of strain w.r.t. (x_1, y_1) co-ordinates are

$$\begin{aligned} \epsilon_{x_1 x_1} &= \frac{\partial u_1}{\partial x_1} = \frac{-\partial u}{\partial x_1} = -\frac{\partial u}{\partial x} \frac{\partial x}{\partial x_1} - \frac{\partial u}{\partial y} \frac{\partial y}{\partial x_1} \\ &= \frac{\partial u}{\partial x} = \epsilon_{xx} \end{aligned} \quad \left| \because \frac{\partial x}{\partial x_1} = -1, \frac{\partial y}{\partial x_1} = 0 \right.$$

Similarly,

$$\epsilon_{y_1 y_1} = \epsilon_{yy}, \quad \gamma_{x_1 y_1} = -\gamma_{xy}$$

and

$$\sigma_{x_1 x_1} = \sigma_{xx}, \sigma_{y_1 y_1} = \sigma_{yy}, \sigma_{x_1 y_1} = -\sigma_{xy}$$

Using these in (7), we get

$$\left. \begin{aligned} \sigma_{x_1 x_1} &= A\epsilon_{x_1 x_1} + B\epsilon_{y_1 y_1} - C\gamma_{x_1 y_1} + D \\ \sigma_{y_1 y_1} &= B\epsilon_{x_1 x_1} + A\epsilon_{y_1 y_1} + C\gamma_{x_1 y_1} + D \\ \sigma_{x_1 y_1} &= C(\epsilon_{x_1 x_1} - \epsilon_{y_1 y_1}) + \frac{A-B}{2}\gamma_{x_1 y_1} \end{aligned} \right\} \quad (8)$$

The relations (7) are invariant where there is a symmetry w.r.t. any co-ordinate transformation and so

$$\left. \begin{aligned} \sigma_{x_1 x_1} &= A\epsilon_{x_1 x_1} + B\epsilon_{y_1 y_1} + C\gamma_{x_1 y_1} + D \\ \sigma_{y_1 y_1} &= B\epsilon_{x_1 x_1} + A\epsilon_{y_1 y_1} - C\gamma_{x_1 y_1} + D \\ \sigma_{x_1 y_1} &= -C(\epsilon_{x_1 x_1} - \epsilon_{y_1 y_1}) + \frac{A-B}{2}\gamma_{x_1 y_1} \end{aligned} \right\} \quad (9)$$

Comparing (8) & (9), we find $C = 0$. According to the assumption (iii), we have

$$\sigma_{xx} = \sigma_{yy} = -p, \quad \epsilon_{xx} = \epsilon_{yy} = 0$$

Thus from (7), we find $D = -p$, since $C = 0$.

The last equation in (7) becomes

$\sigma_{xy} = \frac{A-B}{2} \gamma_{xy} = \mu \gamma_{xy}$, where $\mu = \frac{A-B}{2}$ is called the **co efficient of viscosity**.

The relations in (7) are now,

$$\begin{aligned}\sigma_{xx} &= A \epsilon_{xx} + B \epsilon_{yy} - p = (A-B) \epsilon_{xx} + B (\epsilon_{xx} + \epsilon_{yy}) - p \\ &= 2\mu \epsilon_{xx} + B \nabla \cdot \bar{q} - p \\ \bar{q} &= (u, v) \\ \epsilon_{xx} + \epsilon_{yy} &= \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = \nabla \cdot \bar{q} \\ \sigma_{yy} &= 2\mu \epsilon_{yy} + B \nabla \cdot \bar{q} - p. \\ \sigma_{xy} &= \mu \gamma_{xy} = 2\mu \epsilon_{xy}\end{aligned}$$

These are the required relations between the stress components and the rates of strain in two dimensions.

For three dimensional case, we can write.

$$\left. \begin{aligned}\sigma_{xx} &= 2\mu \epsilon_{xx} + B \nabla \cdot \bar{q} - p = 2\mu \frac{\partial u}{\partial x} + \lambda \nabla \cdot \bar{q} - p \\ \sigma_{yy} &= 2\mu \epsilon_{yy} + B \nabla \cdot \bar{q} - p = 2\mu \frac{\partial v}{\partial y} + \lambda \nabla \cdot \bar{q} - p \\ \sigma_{zz} &= 2\mu \epsilon_{zz} + B \nabla \cdot \bar{q} - p = 2\mu \frac{\partial w}{\partial z} + \lambda \nabla \cdot \bar{q} - p\end{aligned}\right\} \quad (10)$$

$$\left. \begin{aligned}\sigma_{xy} &= \mu \gamma_{xy} = \mu \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \right), \\ \sigma_{yz} &= \mu \gamma_{yz} = \mu \left(\frac{\partial w}{\partial y} + \frac{\partial v}{\partial z} \right) \\ \sigma_{zx} &= \mu \gamma_{zx} = \mu \left(\frac{\partial w}{\partial x} + \frac{\partial u}{\partial z} \right)\end{aligned}\right\} \quad (11)$$

where $B \equiv \lambda$.

$$\begin{aligned}
 \text{Also, } \sigma_{xx} + \sigma_{yy} + \sigma_{zz} &= 2\mu(\epsilon_{xx} + \epsilon_{yy} + \epsilon_{zz}) + 3\lambda \nabla \cdot \bar{q} - 3p \\
 &= 2\mu \nabla \cdot \bar{q} + 3\lambda \nabla \cdot \bar{q} - 3p \\
 &= (2\mu + 3\lambda) \nabla \cdot \bar{q} - 3p
 \end{aligned}$$

For incompressible fluid $\nabla \cdot \bar{q} = 0$.

$$\Rightarrow \sigma_{xx} + \sigma_{yy} + \sigma_{zz} = -3p$$

$$\text{i.e. } \frac{\sigma_{xx} + \sigma_{yy} + \sigma_{zz}}{3} = -p$$

This shows that the mean normal stress is equal to the hydrostatic pressure (i.e. constant)

4.1. Remarks : (i) For compressible fluids, $B \equiv \lambda = -\frac{2\mu}{3}$

(ii) Equations (10) and (11) may be combined in tensor form. Thus, if x_i denote the Cartesian co-ordinates, u_i the velocity components ($i = 1, 2, 3$), then (10) & (11) may be collectively written as

$$\sigma_{ij} = (\lambda\theta - p) S_{ij} + \mu(u_{i,j} + u_{j,i}), \quad (i, j = 1, 2, 3)$$

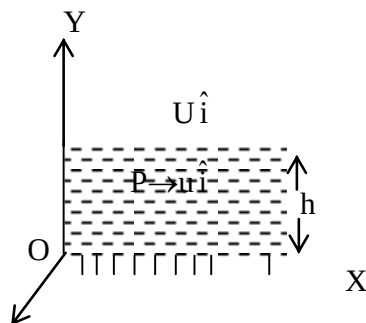
where $\theta = \text{div } \bar{q} = u_{j,i}$,

$$p = -\frac{1}{3} \sigma_{i,i}, \quad \theta = 0 \text{ for incompressible flow,}$$

$$\lambda = -\frac{2}{3} \mu \text{ for compressible flow.}$$

(iv) For viscous fluids, stress is linearly proportional to rate of strain. This law is known as Newton's law of viscosity and such fluids are known as **Newtonian fluids**.

4.2. The Co-efficient of Viscosity and Laminar Flow :





The figure shows two parallel planes $y = 0$, $y = h$, a small distance h apart, the space between them being occupied by a thin film of viscous fluid. The plane $y = 0$ is held fixed and the upper plane is given a constant velocity $U \hat{i}$. If U is not very large, the layers of liquid in contact with $y = 0$ are at rest and those in contact with $y = h$ are moving with velocity $U \hat{i}$ i.e. there is no slip between fluid and either surface. A velocity gradient is set up in the fluid between the planes. At some point $P(x, y, z)$ in between the planes, the fluid velocity will be $U \hat{i}$, where $0 < u < U$ and u is independent of x and z . Thus, when y is fixed, u is fixed i.e. fluid moves in layers parallel to two planes. Such flow is termed as **Laminar flow**. Due to viscosity of the fluid there is friction between these layers. Experimental work shows that the shearing stress on the moving plane is proportional to U/h when h is sufficiently small. Thus, we write this stress in the form

$$\sigma_{yx} = \mu \lim_{h \rightarrow 0} \frac{U}{h} = \mu \frac{du}{dy}$$

where μ is the co-efficient of viscosity. In aerodynamics, a more important quality is the Kinematic co-efficient of viscosity ν defined by

$$\nu = \mu/\rho.$$

For most fluids μ depends on the pressure and temperature. For gases, according to the Kinetic theory, μ is independent of the pressure but decreases with the temperature.

5. Navier-Stoke's Equations of Motion (Conservation of Linear Momentum)

Let us consider a mass of volume τ enclosed by the surface S in motion at time t . Let $d\tau$ be an element of volume, then the mass of this element is $\rho d\tau$, ρ being the density of the viscous fluid.

Let the element moves with the velocity $\bar{q}$. The inertial force on the element is

$$\rho d\tau \left(\frac{d\bar{q}}{dt} \right) \quad | \quad \bar{F} = m\bar{a}$$

The resultant of inertial forces (or the rate of change of linear momentum) is

$$\bar{F}_I = \iiint \rho \frac{d\bar{q}}{dt} d\tau \quad (1)$$

Let $\bar{X}$ be the body force per unit mass, then the resultant of body force is

$$\bar{F}_B = \iiint \rho \bar{X} d\tau \quad (2)$$

The surface force on an element $d\bar{A}$ of the surface is given by the vector

$$\begin{aligned} \bar{f} &= f_x \hat{i}_x + f_y \hat{i}_y + f_z \hat{i}_z \\ &= (\bar{P}_x \cdot d\bar{A}) \hat{i}_x + (\bar{P}_y \cdot d\bar{A}) \hat{i}_y + (\bar{P}_z \cdot d\bar{A}) \hat{i}_z \end{aligned} \quad (3)$$

where $\hat{i}_x, \hat{i}_y, \hat{i}_z$ are unit vectors, $d\bar{A}$ is the vectorial area of the element and $\bar{P}_x, \bar{P}_y, \bar{P}_z$ are components of stress vector, given by

$$\left. \begin{aligned} \bar{P}_x &= \sigma_{xx} \hat{i}_x + \sigma_{xy} \hat{i}_y + \sigma_{xz} \hat{i}_z \\ \bar{P}_y &= \sigma_{yx} \hat{i}_x + \sigma_{yy} \hat{i}_y + \sigma_{yz} \hat{i}_z \\ \bar{P}_z &= \sigma_{zx} \hat{i}_x + \sigma_{zy} \hat{i}_y + \sigma_{zz} \hat{i}_z \end{aligned} \right\} | T_i^x = \tau_{ij} x_j \quad (4)$$

The resultant of the surface forces is given by

$$\bar{F}_S = \hat{i}_x \iint \bar{P}_x \cdot d\bar{A} + \hat{i}_y \iint \bar{P}_y \cdot d\bar{A} + \hat{i}_z \iint \bar{P}_z \cdot d\bar{A} \quad (5)$$

Using Gauss divergence theorem this can be written as

$$\bar{F}_S = \hat{i}_x \iiint \nabla \cdot \bar{P}_x d\tau + \hat{i}_y \iiint \nabla \cdot \bar{P}_y d\tau + \hat{i}_z \iiint \nabla \cdot \bar{P}_z d\tau \quad (6) \quad | \because d\bar{A} = \hat{n} dS$$

Let us use the law of conservation of momentum. By this law, the time rate of change of linear momentum is equal to the total force on the fluid mass. Equating the resultant of body and surface forces with that of inertial forces, we obtain.

$$\iiint \rho \frac{d\bar{q}}{dt} d\tau = \iiint \rho \bar{X} d\tau + \hat{i}_x \iiint \nabla \cdot \bar{P}_x d\tau + \hat{i}_y \iiint \nabla \cdot \bar{P}_y d\tau + \hat{i}_z \iiint \nabla \cdot \bar{P}_z d\tau \quad (7)$$

Since $d\tau$ is an arbitrary volume element, so we have

$$\rho \frac{d\bar{q}}{dt} = \rho \bar{X} + \nabla \cdot \bar{P}_x \hat{i}_x + \nabla \cdot \bar{P}_y \hat{i}_y + \nabla \cdot \bar{P}_z \hat{i}_z \quad (8)$$

This is the required equation of motion in vector form using the values of $\bar{P}_x, \bar{P}_y, \bar{P}_z$, we get

$$\begin{aligned}\nabla \cdot \bar{P}_x &= \frac{\partial \sigma_{xx}}{\partial x} + \frac{\partial \sigma_{xy}}{\partial y} + \frac{\partial \sigma_{xz}}{\partial z} \\ \nabla \cdot \bar{P}_y &= \frac{\partial \sigma_{yx}}{\partial x} + \frac{\partial \sigma_{yy}}{\partial y} + \frac{\partial \sigma_{yz}}{\partial z} \\ \nabla \cdot \bar{P}_z &= \frac{\partial \sigma_{zx}}{\partial x} + \frac{\partial \sigma_{zy}}{\partial y} + \frac{\partial \sigma_{zz}}{\partial z}\end{aligned}$$

and let $\bar{q} = (u, v, w)$, $\bar{X} = (X_x, X_y, X_z)$ then the equations of motion can be put as

$$\left. \begin{aligned}\rho \frac{du}{dt} &= \rho X_x + \frac{\partial \sigma_{xx}}{\partial x} + \frac{\partial \sigma_{xy}}{\partial y} + \frac{\partial \sigma_{xz}}{\partial z} \\ \rho \frac{dv}{dt} &= \rho X_y + \frac{\partial \sigma_{yx}}{\partial x} + \frac{\partial \sigma_{yy}}{\partial y} + \frac{\partial \sigma_{yz}}{\partial z} \\ \rho \frac{dw}{dt} &= \rho X_z + \frac{\partial \sigma_{zx}}{\partial x} + \frac{\partial \sigma_{zy}}{\partial y} + \frac{\partial \sigma_{zz}}{\partial z}\end{aligned}\right\} \quad (9)$$

These are the equations of motion in terms of the stress components. (We have also drawn these equations previously)

Also, we know that

$$\frac{d}{dt} \equiv \frac{\partial}{\partial t} + \bar{q} \cdot \nabla \equiv \frac{\partial}{\partial t} + u \frac{\partial}{\partial x} + v \frac{\partial}{\partial y} + w \frac{\partial}{\partial z}$$

and the relations between stress and rates of strain are

$$\begin{aligned}\sigma_{xx} &= 2\mu \frac{\partial u}{\partial x} + \lambda \nabla \cdot \bar{q} - p \\ \sigma_{yy} &= 2\mu \frac{\partial v}{\partial y} + \lambda \nabla \cdot \bar{q} - p \\ \sigma_{zz} &= 2\mu \frac{\partial w}{\partial z} + \lambda \nabla \cdot \bar{q} - p \\ \sigma_{xy} &= \mu \gamma_{xy} = \mu \left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \right)\end{aligned}$$

$$\sigma_{yz} = \mu \left(\frac{\partial w}{\partial y} + \frac{\partial v}{\partial z} \right),$$

$$\sigma_{zx} = \mu \left(\frac{\partial w}{\partial x} + \frac{\partial u}{\partial z} \right)$$

Using these in (9), we get

$$\left. \begin{aligned} \rho \frac{du}{dt} &= \rho X_x - \frac{\partial p}{\partial x} + \frac{\partial}{\partial x} \left[\mu \left(2 \frac{\partial u}{\partial x} - \frac{2}{3} \nabla \cdot \bar{q} \right) \right] + \frac{\partial}{\partial y} \left[\mu \left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \right) \right] + \frac{\partial}{\partial z} \left[\mu \left(\frac{\partial u}{\partial z} + \frac{\partial w}{\partial x} \right) \right] \\ \rho \frac{dv}{dt} &= \rho X_y - \frac{\partial p}{\partial y} + \frac{\partial}{\partial y} \left[\mu \left(2 \frac{\partial v}{\partial y} - \frac{2}{3} \nabla \cdot \bar{q} \right) \right] + \frac{\partial}{\partial z} \left[\mu \left(\frac{\partial v}{\partial z} + \frac{\partial w}{\partial y} \right) \right] + \frac{\partial}{\partial x} \left[\mu \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \right) \right] \\ \rho \frac{dw}{dt} &= \rho X_z - \frac{\partial p}{\partial z} + \frac{\partial}{\partial z} \left[\mu \left(2 \frac{\partial w}{\partial z} - \frac{2}{3} \nabla \cdot \bar{q} \right) \right] + \frac{\partial}{\partial x} \left[\mu \left(\frac{\partial w}{\partial x} + \frac{\partial u}{\partial z} \right) \right] + \frac{\partial}{\partial y} \left[\mu \left(\frac{\partial w}{\partial y} + \frac{\partial v}{\partial z} \right) \right] \end{aligned} \right\} \quad (10)$$

where $\lambda = -\frac{2\mu}{3}$ compressible fluids.

The equation in (10) are called Navier-Stoke's equations for a viscous **compressible** fluid.

5.1. Deductions (i) If μ = co-efficient of viscosity = constant, then Navier-Stoke's equations (10) become

$$\rho \frac{du}{dt} = \rho X_x - \frac{\partial p}{\partial x} + \frac{1}{3} \mu \frac{\partial}{\partial x} \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} \right) + \mu \nabla^2 u$$

$$\rho \frac{dv}{dt} = \rho X_y - \frac{\partial p}{\partial y} + \frac{1}{3} \mu \frac{\partial}{\partial y} \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} \right) + \mu \nabla^2 v$$

$$\rho \frac{dw}{dt} = \rho X_z - \frac{\partial p}{\partial z} + \frac{1}{3} \mu \frac{\partial}{\partial z} \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} \right) + \mu \nabla^2 w$$

which can be expression in vector form as

$$\rho \frac{d\bar{q}}{dt} = \rho \left[\frac{\partial \bar{q}}{\partial t} + (\bar{q} \cdot \nabla) \bar{q} \right] = \rho \bar{X} - \nabla p + \mu \nabla^2 \bar{q} + \frac{\mu}{3} \nabla (\nabla \cdot \bar{q})$$

(ii) For incompressible fluid, ρ = constant,

$$\mu = \text{constant}, \quad \nabla \cdot \bar{\mathbf{q}} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0.$$

Thus the equations become

$$\frac{d\bar{\mathbf{q}}}{dt} = \frac{\partial \bar{\mathbf{q}}}{\partial t} + (\bar{\mathbf{q}} \cdot \nabla) \bar{\mathbf{q}} = \bar{\mathbf{X}} - \frac{1}{\rho} \nabla p + \frac{\mu}{\rho} \nabla^2 \bar{\mathbf{q}}$$

i.e.
$$\frac{d\bar{\mathbf{q}}}{dt} = \bar{\mathbf{X}} - \frac{\nabla p}{\rho} + \nu \nabla^2 \bar{\mathbf{q}}$$

where $\nu = \mu/\rho$ is called the Kinematic co-efficient of viscosity.

For steady motion with no body forces, we have

$$(\bar{\mathbf{q}} \cdot \nabla) \bar{\mathbf{q}} = \frac{-\nabla p}{\rho} + \frac{\mu}{\rho} \nabla^2 \bar{\mathbf{q}} \quad \left| \frac{\partial \bar{\mathbf{q}}}{\partial t} = 0, \bar{\mathbf{X}} = 0 \right.$$

(iii) If there is no shear at all i.e $\mu = 0$, then

$$\frac{d\bar{\mathbf{q}}}{dt} = \frac{\partial \bar{\mathbf{q}}}{\partial t} + (\bar{\mathbf{q}} \cdot \nabla) \bar{\mathbf{q}} = \bar{\mathbf{X}} - \frac{\nabla p}{\rho}$$

These are Euler's dynamical equations for an incompressible non-viscous fluid.

UNIT X

Steady Flow Between Parallel Planes

For a viscous incompressible fluid in steady flow, the Navier Stoke's equation with negligible body forces, are

$$\frac{d\bar{q}}{dt} = \frac{-\nabla p}{\rho} + \frac{\mu}{\rho} \nabla^2 \bar{q} = \frac{-\nabla p}{\rho} + \nu \nabla^2 \bar{q}, \quad \nu = \frac{\mu}{\rho}$$

In Cartesian co-ordinates; these are

$$u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} + w \frac{\partial u}{\partial z} = -\frac{1}{\rho} \frac{\partial p}{\partial x} + \nu \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right)$$

$$\begin{aligned}
 & \left| \because \text{For steady case, } \frac{du}{dt} = \frac{\partial u}{\partial t} + (\bar{q} \cdot \nabla)u = (\bar{q} \cdot \nabla)u, \frac{\partial}{\partial t} = 0 \right. \\
 & = \left(u \frac{\partial}{\partial x} + v \frac{\partial}{\partial y} + w \frac{\partial}{\partial z} \right) u \\
 & = u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} + w \frac{\partial u}{\partial z} \\
 & u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} + w \frac{\partial v}{\partial z} = -\frac{1}{\rho} \frac{\partial p}{\partial y} + \nu \left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial z^2} \right) \\
 & u \frac{\partial w}{\partial x} + v \frac{\partial w}{\partial y} + w \frac{\partial w}{\partial z} = -\frac{1}{\rho} \frac{\partial p}{\partial z} + \nu \left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} + \frac{\partial^2 w}{\partial z^2} \right) \quad (1)
 \end{aligned}$$

The equation of continuity for incompressible flow is

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0 \quad (2) \quad | \because \nabla \cdot \bar{q} = 0$$

The equations (1) are non-linear 2nd order partial differential equations and there is no known general method for solving them. However, we shall find some exact solutions of the Navier-Stoke's equations in some special cases. This is one of those cases.

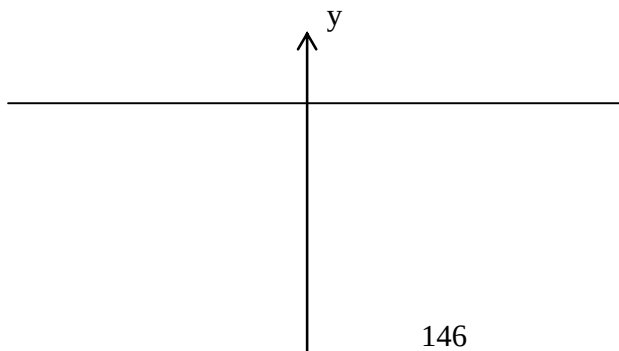
Let us consider a two dimensional steady laminar flow of a viscous incompressible fluid between two parallel straight plates. Let x-axis be the direction of flow, y-axis be perpendicular to it and z-axis be parallel to the width of the plates and let h be the distance between the plates.

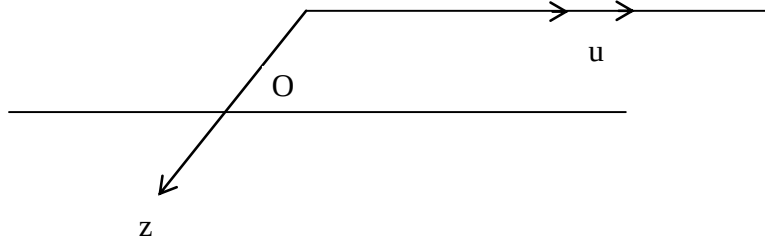
We have the conditions

$$v = 0, w = 0 \text{ and } \frac{\partial}{\partial z} \equiv 0 \quad (3)$$

From the continuity equation (2), we have

$$\frac{\partial u}{\partial x} = 0 \Rightarrow u = u(y) \quad (4)$$





The second equation of equations (1) gives

$$-\frac{\partial p}{\partial y} = 0 \Rightarrow p = p(x) \quad (5)$$

The 3rd equation of equations (1) is identically satisfied and the 1st equation gives

$$0 = -\frac{1}{\rho} \frac{dp}{dx} + \nu \frac{d^2 u}{dy^2} \Rightarrow \frac{dp}{dx} = \mu \frac{d^2 u}{dy^2} \quad \left| \because \frac{\mu}{\rho} = \nu \right. \quad (6)$$

Since u is a function of y only, so $\frac{dp}{dx}$ is either a function of y or a constant.

But from (5), p is a function of x alone.

Hence $\frac{dp}{dx}$ is constant. i.e. pressure gradient is constant.

Integrating equation (6) w.r.t y twice, we get the general solution to be

$$u = \frac{1}{\mu} \frac{dp}{dx} \frac{y^2}{2} + Ay + B \quad (7)$$

where A and B are constants to be determined from the boundary conditions.

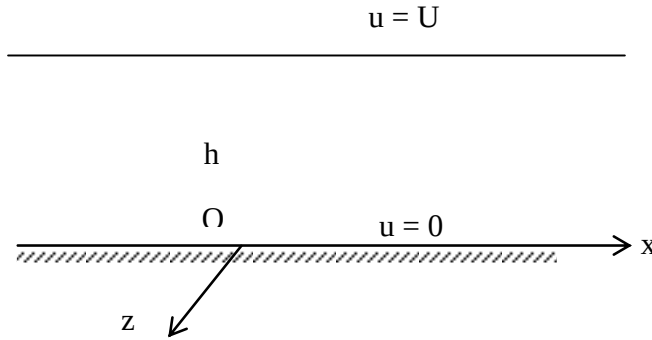
Now we take the following particular cases

6.1. Couette's Flow : It is the flow between two parallel planes (flat plates) one of which is at rest and other moving with velocity U parallel to the fixed plate. Here, the constants A and B in (7) are determined from the conditions

$$u = 0, y = 0 \quad \left. \vphantom{\frac{dp}{dx}} \right\}$$

and $u = U, y = h \quad (8)$





Using these conditions, we get

$$B = 0, \quad U = \frac{1}{\mu} \left(\frac{dp}{dx} \right) \frac{h^2}{2} + Ah$$

$$\Rightarrow \quad A = \frac{U}{h} - \frac{h}{2\mu} \left(\frac{dp}{dx} \right), \quad B = 0 \quad (9)$$

Therefore, the solution (7) becomes

$$u = \frac{1}{\mu} \left(\frac{dp}{dx} \right) \frac{y^2}{2} + y \left[\frac{U}{h} - \frac{h}{2\mu} \left(\frac{dp}{dx} \right) \right] \quad (10)$$

$$= \frac{y^2 - hy}{2\mu} \left(\frac{dp}{dx} \right) + \frac{Uy}{h} \quad (*)$$

$$= \frac{U}{h} y - \frac{h^2}{2\mu} \frac{dp}{dx} \frac{y}{h} \left(1 - \frac{y}{h} \right) \quad (11)$$

We note that equation (10) represents a parabolic curve.

This equation is known as the equation of Couette's flow. Thus the velocity profile for Couette's flow is parabolic. The flow Q per unit breadth is given by

$$\begin{aligned} Q &= \int_0^h u \, dy = \int_0^h \left[\frac{1}{\mu} \frac{dp}{dx} \frac{y^2}{2} + y \left(\frac{U}{h} - \frac{h}{2\mu} \frac{dp}{dx} \right) \right] dy \\ &= \frac{hU}{2} - \frac{h^3}{12\mu} \frac{dp}{dx} \quad (**) \end{aligned}$$

$$= \frac{hU}{2} + \frac{h^3}{12\mu} P, \quad P = -\frac{dp}{dx} \quad (12)$$

In non-dimensional form (11) can be written as

$$\frac{u}{U} = \frac{y}{h} + \alpha \frac{y}{h} \left(1 - \frac{y}{h} \right) \quad (13)$$

where $\alpha = \frac{h^2}{2\mu U} \left(-\frac{dp}{dx} \right)$ (14)

α is the non-dimensional pressure gradient. If $\alpha > 0$, the pressure is decreasing in the direction of flow and the velocity is positive between the plates. If $\alpha < 0$, the equation (13) can be put as

$$\frac{u}{U} = \frac{y}{h} (1 + \alpha) - \frac{\alpha y^2}{h^2} \quad (15)$$

The pressure is increasing in the direction of flow and the reverse flow begins when $\alpha < -1$

| $\because y$ is small. i.e.

y^2 is neglected

If $\alpha = 0$ (i.e. $\frac{dp}{dx} = 0$), then the particular case is known as simple Couette's flow and the velocity is given by

$$\frac{u}{U} = \frac{y}{h}$$

which gives $u = 0$ where $y = 0$ i.e. on the stationary plane.

(i) Average and Extreme Values of Velocity : The average velocity of a Couette's flow between two parallel straight plates is given by

$$u_0 = \frac{1}{h} \int_0^h u \, dy \quad (16) \quad | \because u = u(y)$$

Using the value of u from (13), we get

$$\begin{aligned} u_0 &= \frac{1}{h} \int_0^h \left[\frac{Uy}{h} + U\alpha \frac{y}{h} \left(1 - \frac{y}{h} \right) \right] dy \\ &= \frac{Uh^2}{2h^2} + U\alpha \left(\frac{h^2}{2h^2} - \frac{h^3}{3h^3} \right) \end{aligned}$$

$$= \frac{U}{2} + \frac{U\alpha}{6} = \left(\frac{1}{2} + \frac{\alpha}{6}\right)U \quad (17)$$

$$= \frac{U}{2} - \frac{\mu^2}{12\mu} \frac{dp}{dx} = \frac{U}{2} + \frac{h^2}{12\mu} P, \quad P = -\frac{dp}{dx} \quad (18)$$

In the case of a simple Couette's flow, the velocity increases from zero on the stationary plate to U on the moving plate such that the average velocity is $\frac{U}{2}$.

When the non-dimensional pressure gradient is $\alpha = -3$, then from (17), we get $u_0 = 0$. This means that there is no flow because the pressure gradient is balanced by the viscous force.

For maximum & minimum values of u , we have

$$\begin{aligned} \frac{du}{dy} = 0 &\Rightarrow \frac{U}{h} + U\alpha\left(\frac{1}{h} - \frac{2y}{h^2}\right) = 0 \\ &\Rightarrow y = \left(\frac{1+\alpha}{2\alpha}\right)h \end{aligned} \quad (19)$$

From here, $\frac{y}{h} = 1$, when $\alpha = 1$

and $\frac{y}{h} = 0$, when $\alpha = -1$

So, from (13), we get

$$\begin{aligned} u &= \left[\frac{1+\alpha}{2\alpha} + \alpha \left(\frac{1+\alpha}{2\alpha} \right) \left(1 - \frac{1+\alpha}{2\alpha} \right) \right] U \\ &= \frac{(1+\alpha)^2}{4\alpha} U \end{aligned}$$

and thus u is maximum for $\alpha \geq 1$ and minimum for $\alpha \leq -1$.

(ii) Shearing Stress : The shearing stress (drag per unit area) in a Couette's flow is given by

$$\sigma_{yx} = \mu \frac{du}{dy} = \mu \frac{U}{h} + \frac{\mu\alpha U}{h} \left(1 - \frac{2y}{h} \right) \quad (20)$$

$$= \frac{\mu U}{h}, \text{ for a simple Couette's flow } (\alpha = 0).$$

When $y = \frac{h}{2}$, then the second term in (20) vanishes. Thus the shearing stress is independent of α on the line midway between the flow. The shearing stress at the stationary plane is positive for $\alpha > -1$ and negative for $\alpha < -1$.
| $y = 0$ at stationary plate

The velocity gradient at the stationary plate is zero for $\alpha = -1$ and the shearing stress is zero for $\alpha = -1$.

Thus $\sigma_{yx} \gtrless 0$ when $\alpha \gtrless -1$.

Further, drag per unit area on the lower and the upper plates are obtained from (20) by putting $y = 0$ and $y = h$, as

$$\frac{\mu U}{h} + \frac{\mu \alpha U}{h} \text{ and } \frac{\mu U}{h} - \frac{\mu \alpha U}{h}$$

combining the two results, drag per unit area on the two plates is

$$\frac{\mu U}{h} \pm \frac{\mu \alpha U}{h} \text{ i.e. } \frac{\mu U}{h} \mp \frac{h}{2} \frac{dp}{dx} \quad (***)$$

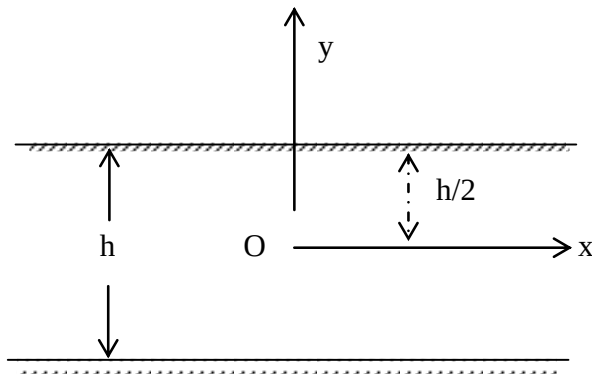
i.e. $\frac{\mu U}{h} \pm \frac{Ph}{2}, P = -\frac{dp}{dx}$

6.2. Plane Poiseuille Flow : A flow between two parallel stationary plates is said to be a plane Poiseuille Flow.

The origin is taken on the line midway between the plates which are placed at a distance h and x -axis is along this line.

The conditions to be used in this problem are

$$u = 0, \text{ when } y = \pm \frac{h}{2} \quad (21)$$



Using these conditions in (7), we get

$$A = 0, B = \frac{1}{\mu} \left(-\frac{dp}{dx} \right) \frac{h^2}{8}$$

and thus the solution (7) is

$$u = \frac{1}{\mu} \left(\frac{dp}{dx} \right) \left(\frac{y^2}{2} - \frac{h^2}{8} \right) \quad (22)$$

This represents a parabola and thus the laminar flow in a Plane Poiseuille Flow is parabolic.

(i) Average and Maximum Velocity : For extreme values of u , we have $\frac{du}{dy} = 0$ and thus from (22), we get

$$\frac{1}{\mu} \left(\frac{dp}{dx} \right) y = 0 \Rightarrow y = 0$$

Therefore ,
$$U_{\max} = \frac{h^2}{8\mu} \left(-\frac{dp}{dx} \right) \quad (23)$$

The average velocity in the plane Poiseuille flow is defined by

$$u_0 = \frac{1}{h} \int_{-h/2}^{h/2} u \, dy$$

Using the value of u from (22), we get

$$\begin{aligned} u_0 &= \frac{1}{h} \int_{-h/2}^{h/2} \frac{-h^2}{8\mu} \frac{dp}{dx} \left(1 - \frac{4y^2}{h^2} \right) dy \\ &= \frac{2}{3} \left(\frac{-h^2}{8\mu} \frac{dp}{dx} \right) = \frac{2}{3} U_{\max} \end{aligned} \quad (24)$$

From (23) & (24), decrease in the pressure is given by

$$\frac{dp}{dx} = -\frac{8\mu}{h^2} u_{\max.} = -\frac{8\mu}{h^2} \frac{3}{2} u_0 = -\frac{12\mu}{h^2} u_0 \quad (25)$$

This further shows that $\frac{dp}{dx}$ is a negative constant.

(ii) Shearing Stress : The shearing stress at a plate (lower plate) for a plane Poiseuille Flow is

$$\begin{aligned} (\sigma_{yx})_{y=-\frac{h}{2}} &= \left(\mu \frac{du}{dy} \right)_{y=-\frac{h}{2}} = -\mu \frac{1}{\mu} \cdot \frac{dp}{dx} \cdot \frac{h}{2} \\ &= -\frac{h}{2} \frac{dp}{dx} \\ &= \frac{4\mu}{h} u_{\max.} \end{aligned} \quad (26)$$

The local frictional (skin) co-efficient C_f is defined by

$$\begin{aligned} C_f &= \frac{(\sigma_{yx})_{-h/2}}{\rho u_{0/2}^2} = \frac{4\mu}{h} u_{\max.} \bigg/ \frac{\rho u_0^2}{2} \\ &= \frac{4\mu}{\rho h} \left(\frac{3}{2} \frac{u_0}{u_0^2/2} \right) = \frac{12\mu}{h u_0} = \frac{12}{Re} \end{aligned}$$

Where $Re = \frac{u_0 h}{\nu}$ is the Reynolds number of the flow based on the average velocity and the channel height.

Elective Paper

Applied Stream

MATA 2.4

Block - II

Marks : 40 (SSE : 30; IA : 10)

Stochastic Processes

Syllabus

• Unit 11 •

Review of Probability: Random variables, conditional probability and independence, bivariate and multi-variate distributions.

• Unit 12 •

Probability generating functions, characteristic functions, convergence concepts.

• Unit 13 •

Conditional Expectation: Conditioning on an event, conditioning on a discrete random variable, conditioning on an arbitrary random variable, conditioning on a sigma-field.

• Unit 14 •

The Random Walk: unrestricted random walk, types of stochastic processes, gambler's ruin problem, generalisation of the random walk model.

• Unit 15 •

Markov Chains: Definitions, Chapman-Kolmogorov equation, Equilibrium distributions, Classification of states, Long-time behaviour. Stationary distribution. Branching process.

• Unit 16 •

Stochastic process in continuous time: Poisson process and Brownian motion.

UNIT - 11, 12, 13

Probability.

The axioms of probability are that there is a “sample space” Ω (or S) containing all possible outcomes and a function P that assigns to subsets of Ω (called “events”) a number in $[0, 1]$ such that

- (i) $P(\Omega) = 1$ and
- (ii) if $E_1, E_2, \dots$ are disjoint events, then

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

[Actually, sometimes only certain subsets of Ω can be given a probability, but that will not concern us. Part of the development of measure theory elucidates this issue.]

The axioms imply the particularly useful consequences:

- (i) $P(\emptyset) = 0$
- (ii) If $E \subseteq F$, then $P(E) \leq P(F)$
- (iii) $P(E^c) = 1 - P(E)$
- (iv) (subadditivity) For any events E_n , we have

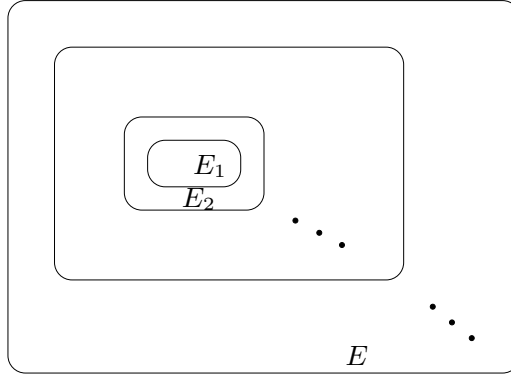
$$P\left(\bigcup_{n=1}^{\infty} E_n\right) \leq \sum_{n=1}^{\infty} P(E_n).$$

...

- (v) If $P(E_n) = 0$, then $P(\bigcup E_n) = 0$. If $P(E_n) = 1$, then $P(\bigcap E_n) = 1$

PROPOSITION 1.1.1. If $E_n \uparrow E$ or $E_n \downarrow E$, then $P(E_n) \rightarrow P(E)$.

Proof. In the first case that $E_n \uparrow E$, write E as a disjoint union $\bigcup (E_{n+1} \setminus E_n)$. (See the figure.) ... When $E_n \downarrow E$, use $E_n^c \uparrow E^c$



If you flip a sequence of coins and the n th coin has chance $1/n^2$ of landing H, will you get an infinite number of heads? What if the chance is $1/n$? To answer these questions, we prove the Borel-Cantelli Lemmas.

PROPOSITION 1.1.2 (FIRST BOREL-CANTELLI LEMMA). *If $\sum_n P(E_n) < \infty$, then $P(E_n \text{ i.o.}) = 0$.*

Proof. We have

$$P(E_n \text{ i.o.}) = P\left(\bigcap_{n \geq 1} \bigcup_{k \geq n} E_k\right) = \lim_n P\left(\bigcup_{k \geq n} E_k\right) \leq \liminf_n \sum_{k \geq n} P(E_k) = 0. \quad \blacksquare$$

...

PROPOSITION 1.1.3 (SECOND BOREL-CANTELLI LEMMA). *If $\sum P(E_n) = \infty$ and $\{E_n\}$ are independent, then $P(E_n \text{ i.o.}) = 1$.*

Proof. We have

$$\begin{aligned} P(E_n \text{ i.o.}) &= \lim_n P\left(\bigcup_{k \geq n} E_k\right) = \lim_n \left[1 - P\left(\bigcap_{k \geq n} E_k^c\right)\right] \\ &= \lim_n \left[1 - \prod_{k \geq n} (1 - P(E_k))\right] \geq \limsup_n \left[1 - \prod_{k \geq n} e^{-P(E_k)}\right] \quad \text{since } 1 - x \leq e^{-x} \\ &= \limsup_n \left[1 - e^{-\sum_{k \geq n} P(E_k)}\right] = 1. \end{aligned}$$

§1.2. Random Variables.

A **(real-valued) random variable** is a function $X: \Omega$ (or S) $\rightarrow \mathbb{R}$. Its **(cumulative) distribution function** (c.d.f.) $F = F_X$ is $F(x) := P[X \leq x]$. Often $\overline{F}(x) := 1 - F(x) = P[X > x]$ is useful. We use the notation $X \sim F$, especially when F has a name, like $\text{Bin}(n, p)$ or $\text{Unif}[0, 1]$. When two random variables X and Y have the same c.d.f., we write $X \stackrel{D}{=} Y$. In case we have a collection of identically distributed random variables X_i , we often write X for a random variable with the same distribution as all of the X_i .

If the range of X is countable, we call X **discrete**; if its values are isolated, then F_X is a step function. If no value has positive probability, X is **continuous**; this is the same as saying that F_X is a continuous function. The random variable X could be neither discrete nor continuous. For example, if we flip a coin and get H, then set $X := 0$; but if we get T, then choose $X \sim \text{Unif}[0, 1]$. What is F_X ? ... If $\exists f: \mathbb{R} \rightarrow \mathbb{R}$ such that $\forall x \ F(x) = \int_{-\infty}^x f(s) ds$, ... then X is **absolutely continuous** [called “continuous” in the book] and f is its **probability density function**. In this case, $f(x) = F'(x)$ and $P[X \in B] = \int_B f(x) dx$. Almost always, we will use the case that B is an interval. Defn for more general B

For two random variables X and Y , their **joint distribution function** is $F_{X,Y}(x, y) := P[X \leq x, Y \leq y]$. If $F_{X,Y}(x, y) = \int_{-\infty}^x \int_{-\infty}^y f_{X,Y}(s, t) dt ds$, then $f_{X,Y}$ is called the **joint density** of X and Y .

Note that $F_X(x) = F_{X,Y}(x, \infty)$. We have that X and Y are independent $\iff \forall x, y \ F_{X,Y}(x, y) = F_X(x)F_Y(y) \iff \forall A, B \ P[X \in A, Y \in B] = P[X \in A]P[Y \in B]$.

§1.3. Expected Value.

The **expectation** of X is defined by $E(X) := \int_{-\infty}^{\infty} x dF_X(x)$. However, we shall only say what this means in two cases: If X is absolutely continuous, then $E(X) = \int_{-\infty}^{\infty} xf(x) dx$. If X is discrete, then $E(X) = \sum_x xP[X = x]$. Give idea of Stieltjes integral with $\int h(x) dF_X(x)$ to explain notation. ... It can be shown that

$$E(h(X)) = \int_{-\infty}^{\infty} h(x) dF_X(x)$$

so that, in particular,

$$P[X \in A] = \int_A dF_X(x).$$

It can also be shown that $E(\sum_{i=1}^n X_i) = \sum_{i=1}^n E(X_i)$. This linearity is actually a key property of expectation; the proper definition of expectation using measure theory makes

the proof of linearity easy. Since $E[Z] \geq 0$ when $Z \geq 0$, it follows that $E[Y] \geq E[X]$ when $Y \geq X$. Another particular case of the previous formula for $E(h(X))$ uses $h(x) := \int_{-\infty}^{\infty} g(x, y) dy$, which gives

$$\begin{aligned} E\left[\int_{-\infty}^{\infty} g(X, y) dy\right] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) dy dF_X(x) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) dF_X(x) dy \\ &= \int_{-\infty}^{\infty} E[g(X, y)] dy. \end{aligned}$$

In other words, we can interchange expectation and integral. Define the **variance** of X as

$$\text{Var}(X) := E[(X - E(X))^2] = E(X^2) - E(X)^2$$

and the **covariance** of X and Y as

$$\text{Cov}(X, Y) := E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y).$$

... Recall that $\text{Cov}(X, Y) = 0$ if (but not only if) X and Y are independent. We have

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) + \sum_{i \neq j} \text{Cov}(X_i, X_j).$$

... In particular, if X_i are independent, then $\text{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \text{Var}(X_i)$.

If X and Y have a joint density, then it can be shown that

$$E[h(X, Y)] = \iint h(x, y) f_{X,Y}(x, y) dx dy.$$

§1.4. Moment Generating, Characteristic Functions, and Laplace Transforms.

We will occasionally need the **moment generating function**

$$E[e^{tX}] = \sum_{n \geq 0} \frac{E[X^n]}{n!} t^n.$$

(Although we will not pay close attention to when this equality holds, it does in all situations we will encounter. For example, if the left-hand side is finite for some $t_+ \geq 0$ and some $t_- \leq 0$, then equality holds for all $t \in [t_-, t_+]$.)

EXAMPLE: $\text{Exp}(\lambda)$. Recall that $X \sim \text{Exp}(\lambda)$ (λ is called the **parameter** or **rate**) if it has probability density function

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases}$$

...

Note that

$$E[X^n] = \left(\frac{d}{dt} \right)^n E[e^{tX}] \Big|_{t=0}.$$

...

Thus, for $X \sim \text{Exp}(\lambda)$, we have $E[X] = 1/\lambda$ and $\text{Var}(X) = 1/\lambda^2$

In the case of the exponential distribution, one could also derive the moments from the fact that $\int_0^\infty x^n e^{-x} dx = n!$ (which follows by induction and integration by parts). Conversely, one can derive this integral by differentiating the moment generating function for $\text{Exp}(1)$.

§1.5. Conditional Expectation.

Use an example for all the following, such as X the number of the first die and Y the sum of two dice. Suppose that X and Y are discrete. Then

$$P[X = x \mid Y = y] = \frac{P[X = x, Y = y]}{P[Y = y]}.$$

We can regard this as a function of x or of y . As a function of x , it gives the distribution function of a random variable since $\sum_x P[X = x \mid Y = y] = 1$. It is called the **conditional distribution of X given $Y = y$** . It has, thus, an expectation,

$$\sum_x x P[X = x \mid Y = y] =: E[X \mid Y = y].$$

...

If we regard $P[X = x \mid Y = y]$ as a function of y , then we may pre-compose it with Y to get a random variable denoted $P[X = x \mid Y]$ We can also pre-compose the function $y \mapsto E[X \mid Y = y]$ with Y to get a random variable denoted $E[X \mid Y]$ We have

$$(1.5.1) \quad E[X] = E[E[X \mid Y]].$$

Proof: Write it out. ...

As a special case, we get that

$$P[X = x] = E[P[X = x \mid Y]].$$

...

These ideas extend to all random variables. For the case that X and Y are jointly absolutely continuous, the density of X given $Y = y$ is $x \mapsto f_{X,Y}(x, y)/f_Y(y)$ (by Exercise 3, this is a probability density function). Think of this as follows. Note that

$$f_X(x) dx = P[X \in (x, x + dx)],$$

so that

$$\begin{aligned} \frac{f_{X,Y}(x, y)}{f_Y(y)} &= \frac{P[X \in (x, x + dx), Y \in (y, y + dy)]/(dx dy)}{P[Y \in (y, y + dy)]/dy} \\ &= P[X \in (x, x + dx) \mid Y \in (y, y + dy)]/dx \\ &= P[X \in (x, x + dx) \mid Y = y]/dx. \end{aligned}$$

Equation (1.5.1) holds too:

$$\begin{aligned}
E\left[\underbrace{E[X | Y]}_{\text{function of } Y}\right] &= \int_{-\infty}^{\infty} E[X | Y = y] dF_Y(y) = \int_{-\infty}^{\infty} E[X | Y = y] f_Y(y) dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x \frac{f_{X,Y}(x, y)}{f_Y(y)} dx f_Y(y) dy \\
&= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy dx = \int_{-\infty}^{\infty} x f_X(x) dx \quad [\text{by Exercise 3}] \\
&= E(X).
\end{aligned}$$

We also want to define $E[X | Y]$ when only one of X or Y is absolutely continuous. First, if X is absolutely continuous and A is an event of positive probability, then (it can be shown using measure theory that) the function

$$x \mapsto P[X \leq x | A]$$

is the c.d.f. of an absolutely continuous random variable; its expectation is denoted $E[X | A]$. This allows us to define $P[X \leq x | Y = y]$ and $E[X | Y = y]$ when X is absolutely continuous and Y is discrete. Second, if A is an event and Y is an absolutely continuous random variable, we define

$$P[A | Y = y] := \lim_{\epsilon \rightarrow 0} \frac{P(A \cap \{Y \in (y - \epsilon, y + \epsilon)\})}{P[Y \in (y - \epsilon, y + \epsilon)]}$$

when the limit exists. When X is discrete and Y is absolutely continuous, this allows us to define

$$E[X | Y = y] := \sum_x x P[X = x | Y = y].$$

Once we have defined $E[X | Y = y]$ in either of these cases, we define $E[X | Y]$ as before by pre-composing with Y .

In all cases, it can be shown that (1.5.1) still holds. In particular,

$$P(A) = E[\mathbf{1}_A] = E[E[\mathbf{1}_A | Y]] = E[P(A | Y)].$$

... Also, it follows from (1.5.1) that

$$E[h(Y)E[X | Y]] = E[E[h(Y)X | Y]] = E[h(Y)X].$$

We can condition on several random variables, too:

$$E[E[X | Y_1, Y_2, \dots]] = E[X].$$

Two random variables X and Y are independent iff the conditional distribution of X given $Y = y$ is equal to the unconditional distribution of X (for all y).

We now give some applications of conditioning.

EXAMPLE PM 3.15 (ANALYZING THE QUICK-SORT ALGORITHM). Given distinct numbers $x_1, \dots, x_n$, the goal is to place them in increasing order, that is, to **sort** them, as quickly as possible. The quick-sort algorithm works as illustrated in an example: Suppose that the original list is 10, 5, 8, 2, 1, 4, 7. Choose one at random, say, 4. Compare 4 to the others: $\{2, 1\}$, 4, $\{10, 5, 8, 7\}$. Now apply the same procedure to the set < 4 and the set > 4 :

$\rightarrow 1, 2, 4, \{10, 5, 8, 7\} \rightarrow$ choose at random from 2nd set, say 7:

$\rightarrow 1, 2, 4, 5, 7, \{10, 8\} \rightarrow 1, 2, 4, 5, 7, 8, 10.$

The number of comparisons here was $6 + 1 + 3 + 1 = 11$. This is a random “divide and conquer” algorithm. How well does it do? The slowest would be if we always pick the smallest or the largest one; then every pair must be compared and it takes $\sim n^2/2$ comparisons. The fastest possible would be if every time, the median were chosen; then the number of comparisons would be

$$\sim n + \frac{n}{2} \times 2 + \frac{n}{4} \times 4 + \dots \quad (\sim \log_2 n \text{ terms}) = n \log_2 n.$$

It turns out that this is quite close to M_n , the expected number of comparisons in quick-sort.

To calculate M_n , condition on the rank of the initial value selected:

$$\begin{aligned} M_n &= \sum_{j=1}^n E[\text{number of comparisons} \mid \text{initial value is } j\text{th smallest}] P[\text{initial value is } j\text{th smallest}] \\ &= \sum_{j=1}^n (n-1 + M_{j-1} + M_{n-j}) \cdot \frac{1}{n} = n-1 + \frac{2}{n} \sum_{k=1}^{n-1} M_k. \end{aligned}$$

Thus

$$nM_n = n(n-1) + 2 \sum_{k=1}^{n-1} M_k.$$

Substitute $n-1$ for n and subtract:

$$nM_n - (n-1)M_{n-1} = 2(n-1) + 2M_{n-1},$$

whence

$$nM_n = (n+1)M_{n-1} + 2(n-1),$$

which is

$$\frac{M_n}{n+1} = \frac{M_{n-1}}{n} + \frac{2(n-1)}{n(n+1)}.$$

Iterating gives

$$\begin{aligned}\frac{M_n}{n+1} &= 2 \sum_{n \geq k \geq 1} \frac{k-1}{k(k+1)} = 2 \sum_{k=1}^n \left[\frac{2}{k+1} - \frac{1}{k} \right] \\ &\sim 2(2 \log n - \log n) = 2 \log n,\end{aligned}$$

whence

$$M_n \sim 2n \log n.$$

Note that $2 > (\log 2)^{-1}$.

EXAMPLE 1.5(E) (THE BALLOT THEOREM). In an election, A receives n votes and B receives m votes, $n > m$. If all orderings of the $n + m$ votes are equally likely, then

$$P[A \text{ always ahead of } B] = \frac{n-m}{n+m}.$$

Proof. Let $P_{n,m}$ be the desired probability. Then

$$\begin{aligned}P_{n,m} &= P[A \text{ always ahead} \mid A \text{ gets last vote}]P[A \text{ gets last vote}] \\ &\quad + P[A \text{ always ahead} \mid B \text{ gets last vote}]P[B \text{ gets last vote}] \\ &= P_{n-1,m} \cdot \frac{n}{n+m} + P_{n,m-1} \cdot \frac{m}{n+m}.\end{aligned}$$

Here, we make the convention that $P_{n-1,m} := 0$ if $n = m + 1$; note that this fits our formula nicely, so we needn't consider that case separately when we claim that our formula fits the equation. **Note why we conditioned on the last vote, rather than the first.** Now use induction on $n + m$.

EXAMPLE 1.5(A) (THE SUM OF A RANDOM NUMBER OF RANDOM VARIABLES). Let X_i be i.i.d. ($i \geq 1$) and N be an independent random variable with values in $\mathbb{N} := \{0, 1, 2, \dots\}$. Let $Y := \sum_{i=1}^N X_i$.

Examples:

- **Queueing:** $N :=$ the number of customers arriving in a specific time period, $X_i :=$ the service time required by the i th customer. Then $\sum_1^N X_i =$ the total service time required by customers arriving in that time period.

- **Risk Theory:** $N :=$ the number of claims arriving at an insurance company in a given week, $X_i :=$ the amount of the i th claim. Then $\sum_1^N X_i =$ the total liability for that week.

- **Population Model:** $N :=$ the number of plants of a given species in a certain area, $X_i :=$ the number of seeds produced by the i th plant.

To compute moments of Y , we compute the moment generating function:

$$E[e^{tY}] = E[E[e^{tY} \mid N]].$$

Now

$$\begin{aligned} E[e^{tY} \mid N = n] &= E[e^{t \sum_1^N X_i} \mid N = n] = E[e^{t \sum_1^n X_i} \mid N = n] \\ &= E[e^{t \sum_1^n X_i}] \text{ by independence} \\ &= E[e^{tX}]^n \text{ by independence,} \end{aligned}$$

where $X \stackrel{\mathcal{D}}{=} X_i$. Therefore $E[e^{tY}] = E[E[e^{tX}]^N]$,

$$\frac{d}{dt} E[e^{tY}] = E[Y e^{tY}] = E[N E[e^{tX}]^{N-1} E[X e^{tX}]],$$

$$\begin{aligned} \frac{d^2}{dt^2} E[e^{tY}] &= E[Y^2 e^{tY}] \\ &= E[N(N-1) E[e^{tX}]^{N-2} E[X e^{tX}]^2] + E[N E[e^{tX}]^{N-1} E[X^2 e^{tX}]], \end{aligned}$$

so

$$\begin{aligned} E[Y] &= E[N E[X]] = E[N] E[X], \\ E[Y^2] &= E[N(N-1) E[X]^2 + E[N] E[X^2]], \end{aligned}$$

and

$$\begin{aligned} \text{Var}(Y) &= E[N] E[X^2] + \{E[N^2] - E[N] - E[N]^2\} E[X]^2 \\ &= E[N] \text{Var}(X) + \text{Var}(N) E[X]^2. \end{aligned}$$

Given an event A of positive probability and a random variable X with finite expectation, we have defined $E[X | A]$ as $E[Y]$, where Y has the distribution of X given A . There is another definition of conditional probability used in measure-theory-based courses, which we will occasionally find useful:

$$E[X | A] = E[X \mathbf{1}_A] / P(A).$$

To see that this is the same, we may assume that $X \geq 0$ by decomposing $X = X^+ - X^-$. (This would give a corresponding decomposition $Y = Y^+ - Y^-$.) Then we can use the tail formula, Exercise 4 (p. 27, 1.1 in the book), as follows:

$$\begin{aligned} E[Y] &= \int_0^\infty P[Y > y] dy = \int_0^\infty P[X > y | A] dy = \int_0^\infty P[A, X > y] / P(A) dy \\ &= \frac{1}{P(A)} \int_0^\infty E[\mathbf{1}_A \mathbf{1}_{\{X > y\}}] dy = \frac{1}{P(A)} E\left[\int_0^\infty \mathbf{1}_A \mathbf{1}_{\{X > y\}} dy\right] \\ &= \frac{1}{P(A)} E\left[\mathbf{1}_A \int_0^\infty \mathbf{1}_{\{X > y\}} dy\right] = \frac{1}{P(A)} E[\mathbf{1}_A X]. \end{aligned}$$

§1.6. The Exponential Distribution, Lack of Memory, and Hazard Rate Functions.

If $X \sim \text{Exp}(\lambda)$, then $\overline{F}_X(x) = e^{-\lambda x}$. Such a random variable is **memoryless**:

$$P(X > s + t | X > t) = P(X > s) \quad \text{for all } s, t \geq 0. \quad (\text{N1})$$

EXAMPLE: A post office has 2 clerks. The customer service time of each clerk is $\text{Exp}(\lambda)$. Neither clerk is busy. One customer arrives at a random time and, while that customer is still being served, another customer arrives at a random time and begins service with the other clerk. What is the chance that the first customer finishes first?

Solution. The answer is $1/2$ by the (strong) memoryless property and symmetry. Note that we assume the arrival times and service times are mutually independent.

The **strong memoryless property** says that if $X \sim \text{Exp}(\lambda)$ and $Y \geq 0$ is independent of X , then for all $s \geq 0$, we have $P[X > s + Y | X > Y] = P[X > s]$. The book does not mention that the strong memoryless property needs to be established. Prove this by writing the conditional probability as a quotient; calculate both numerator and denominator by conditioning on Y . E.g., $P[X > s + Y, X > Y] = P[X > s + Y] = E[P[X > s + Y | Y]]$ and $P[X > s + Y | Y = y] = P[X > s + y | Y = y] = e^{-\lambda(s+y)}$ by independence,

whence $P[X > s + Y] = E[e^{-\lambda(s+Y)}]$. Likewise, $P[X > Y] = E[e^{-\lambda Y}]$. Thus, the quotient is $e^{-\lambda s}$.

For this problem, let A_1 and A_2 be the arrival times and S_1, S_2 be the service times. We want to show symmetry, i.e., that for all $t > 0$, we have

$$P[A_1 + S_1 > A_2 + t \mid A_1 < A_2 < A_1 + S_1] = P[S_2 > t \mid A_1 < A_2 < A_1 + S_1] = e^{-\lambda t}$$

and that $A_1 + S_1 - A_2$ is independent of S_2 given $A_1 < A_2 < A_1 + S_1$, which follows from the assumed mutual independence. ... We have

$$\begin{aligned} P[A_1 + S_1 > A_2 + t \mid A_1 < A_2 < A_1 + S_1] &= P[S_1 > A_2 - A_1 + t \mid 0 < A_2 - A_1 < S_1] \\ &= e^{-\lambda t} \end{aligned}$$

by the strong memoryless property, where we use the random variables S_1 and $A_2 - A_1$ with respect to the probability measure where $A_2 - A_1$ is conditioned to be positive. ...

Alternatively, we can formulate an even stronger memoryless property: if $X \sim \text{Exp}(\lambda)$ and $Y, Z \geq 0$ with X, Y and Z being mutually independent, then $P[X > Z + Y \mid X > Y] = P[X > Z]$. We prove this by re-using some of what we proved above, namely, $P[X > Y] = E[e^{-\lambda Y}]$, $P[X > Z] = E[e^{-\lambda Z}]$ and $P[X > Z + Y, X > Y] = P[X > Z + Y] = E[e^{-\lambda(Z+Y)}] = E[e^{-\lambda Z}]E[e^{-\lambda Y}]$

We can use this stronger memoryless property to give another solution:

$$\begin{aligned} P[A_1 + S_1 > A_2 + S_2 \mid A_1 < A_2 < A_1 + S_1] &= P[S_1 > S_2 + (A_2 - A_1) \mid 0 < A_2 - A_1 < S_1] \\ &= P[S_1 > S_2] = 1/2. \end{aligned}$$

EXAMPLE 1.6(A). A post office has 2 clerks. The customer service time of each clerk is $\text{Exp}(\lambda)$. You enter and are first in line, with both clerks already serving customers. What is the chance that both customers currently being served will be finished before you are?

Solution. Answer: $1/2$ by the strong memoryless property and symmetry, measuring time from when the first customer leaves. By the previous example, we may assume that both previous customers began service when you entered. Let their service times be S_1 and S_2 , while yours is X . We want $P[X + \min\{S_1, S_2\} > \max\{S_1, S_2\}]$. Calculate this by conditioning whether $S_1 < S_2$ or not and apply the strong memoryless property measuring time from $\min\{S_1, S_2\}$. Again, establish symmetry.

Note that (N1) is the same as $\overline{F}_X(s+t) = \overline{F}_X(s)\overline{F}_X(t)$. Thus, $\log \overline{F}_X$ satisfies the functional equation

$$g(x+y) = g(x) + g(y) \quad (x, y \geq 0).$$

Also, $\log \overline{F}_X$ is right continuous (i.e., continuous *from* the right.)... To show that the exponential random variables are the *only* memoryless random variables, we show that this equation has only the linear solutions $g(x) = cx$ provided g is right continuous. (This result will be useful later, too.) Here are the steps:

- (1) Let $c := g(1)$. Then $g(x) = cx$ for $x \in \mathbb{Q}^+$
- (2) If $g \in C_r(\mathbb{R}^+)$, we're done. (Here, C_r is the space of right-continuous functions, i.e., functions that are continuous *from* the right.) ...
- (3) If g is assumed only to be right continuous at 0, then actually $g \in C_r(\mathbb{R}^+)$ since $g(x_0+h) - g(x_0) = g(h)$

For later use, we note that if g is bounded in some interval $[0, \delta]$ ($\delta > 0$), say, by M , then $|g(x)| = |g(nx)|/n \leq M/n$ for $0 \leq x \leq \delta/n$, whence g is right continuous at 0.

Exponential as a limit of geometric, which is the discrete memoryless random variable: $n^{-1}\text{Geom}(\lambda/n) \Rightarrow \text{Exp}(\lambda)$

In general, if X has a probability density function, the **failure** or **hazard rate** function $\lambda_X(t)$ is $\lambda_X(t) := f_X(t)/\overline{F}_X(t)$. Thus $P[X \in (t, t+dt) \mid X > t] \approx \lambda_X(t) dt$, which explains the name.

§1.8. Some Limit Theorems.

WLLN. If X_i are i.i.d. with mean $\mu \in (-\infty, \infty)$, then for all $\epsilon > 0$, we have

$$\lim_{n \rightarrow \infty} P\left(\frac{1}{n} \sum_{i=1}^n X_i \in (\mu - \epsilon, \mu + \epsilon)\right) = 1.$$

SLLN. If X_i are i.i.d. with mean $\mu \in [-\infty, \infty]$, then

$$P\left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i = \mu\right) = 1.$$

CLT. If X_i are i.i.d. with finite mean μ and finite variance σ^2 , then $\forall a \in \mathbb{R}$

$$\lim_{n \rightarrow \infty} P\left(\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^n (X_i - \mu) \leq a\right) = \int_{-\infty}^a \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

I.e.,

$$\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^n (X_i - \mu) \Rightarrow N(0, 1).$$

(Let the random variables X_n have c.d.f. F_n and Y have c.d.f. F . We write that $X_n \Rightarrow Y$, $X_n \Rightarrow F$, or $F_n \Rightarrow F$ if $F_n(a) \rightarrow F(a)$ at every $a \in \mathbb{R}$ where $F(a)$ is continuous. This is called **convergence in distribution**, **convergence in law**, or **weak convergence**. The last name is because this kind of convergence follows from a.s. convergence; that is, if $P[X_n \rightarrow Y] = 1$, then $X_n \Rightarrow Y$. The WLLN is about weak convergence to a constant random variable, while the SLLN is about a.s. convergence. If X_n and Y are integer valued, then $X_n \Rightarrow Y$ iff $P[X_n = k] \rightarrow P[Y = k]$ for all $k \in \mathbb{Z}$.)

We will use the following generalization only once:

CLT OF LINDBERBERG. *Let X_i be independent, $F_i :=$ the c.d.f. of X_i . Suppose that $E(X_i) = 0$, $\text{Var}(X_i) = \sigma_i^2 < \infty$,*

$$s_n^2 := \sum_{i=1}^n \sigma_i^2,$$

and

$$\forall t > 0 \quad \lim_{n \rightarrow \infty} \frac{1}{s_n^2} \sum_{i=1}^n \int_{|x| \geq ts_n} x^2 dF_i(x) = 0.$$

Then

$$\frac{1}{s_n} \sum_{i=1}^n X_i \Rightarrow N(0, 1).$$

Note that this *is* a generalization. ...

POISSON CONVERGENCE (THE LAW OF RARE EVENTS). For any $\lambda > 0$, we have

$$\text{Bin}(n, \lambda/n) \Rightarrow \text{Pois}(\lambda)$$

as $n \rightarrow \infty$. (See Exercise 7 (Exercise 1.3, p. 46).) More generally, suppose that $\forall n$ $X_{n,i}$ ($1 \leq i \leq n$) are independent random variables with values in $\mathbb{N}$ such that

$$p_{n,i} := P[X_{n,i} = 1]$$

and

$$\varepsilon_{n,i} := P[X_{n,i} \geq 2]$$

satisfy

$$\sum_{i=1}^n p_{n,i} \rightarrow \lambda \in [0, \infty],$$

$$\max_{1 \leq i \leq n} p_{n,i} \rightarrow 0,$$

and

$$\sum_{i=1}^n \varepsilon_{n,i} \rightarrow 0.$$

Then

$$\sum_{i=1}^n X_{n,i} \Rightarrow \text{Pois}(\lambda).$$

Note: There are two special interpretations: $\text{Pois}(0)$ means the distribution of the random variable that is identically 0; and $\text{Pois}(\infty)$ means the distribution of the random variable that is identically ∞ . In the latter case, to say that random variables X_n converge weakly to ∞ means that for all $t < \infty$, we have $F_{X_n}(t) \rightarrow 0$ as $n \rightarrow \infty$.

THE MONOTONE CONVERGENCE THEOREM (MCT). If $X_n \rightarrow X$ a.s. and $0 \leq X_n \leq X$ a.s., then $E[X_n] \rightarrow E[X]$.

THE LEBESGUE DOMINATED CONVERGENCE THEOREM (LDCT). If $X_n \rightarrow X$ a.s., $|X_n| \leq Y$, and $E[Y] < \infty$, then $E[X_n] \rightarrow E[X]$.

THE BOUNDED CONVERGENCE THEOREM (BCT). The LDCT for Y a constant.

Definition of independent increments and stationary increments for a stochastic process. We will be dealing with two stochastic processes that have independent and stationary increments: ones that jump (Poisson processes) and ones that don't (Brownian motion). ...

UNIT 14

Random Walks

Duality in Random Walks.

If $\langle X_i \rangle$ are i.i.d. and $S_n := \sum_{i=1}^n X_i$ is the corresponding random walk, then a useful observation is the “duality” property that $\langle X_1, X_2, \dots, X_n \rangle \stackrel{\mathcal{D}}{=} \langle X_n, X_{n-1}, \dots, X_1 \rangle$. We give two applications. Let

$$N := \min\{n; S_n > 0\},$$

$$M := \text{number of new minima of } \langle S_n \rangle$$

$$= |\{n; \forall k \in [0, n) \quad S_n \leq S_k\}|,$$

$$R_n := \text{number of distinct values of } \langle S_k; 0 \leq k \leq n \rangle$$

$$= |\{S_0, S_1, \dots, S_n\}|$$

$$= |\{k \in [0, n]; \forall j \in [0, k) \quad S_k \neq S_j\}|$$

$$= |\{k \in [0, n]; \forall j \in (k, n] \quad S_k \neq S_j\}|.$$

R_n is called (the size of) the **range** of $\langle S_k; 0 \leq k \leq n \rangle$. If $E[X]$ exists and is positive, then by the SLLN, $S_n \rightarrow \infty$ a.s., whence $N < \infty$ a.s. and $M < \infty$ a.s. Always $R_n \rightarrow \infty$ a.s. (except if $X = 0$ a.s.).

PROPOSITION 7.1.1. *If $\mu := E[X] > 0$, then $E[N] = E[M] < \infty$ and $E[S_N] = \mu E[N]$.*

PROPOSITION 7.1.2. *Without any assumption on $E[X]$,*

$$\lim_{n \rightarrow \infty} \frac{E[R_n]}{n} = P[\text{no return to } 0] = P[\forall n > 0 \quad S_n \neq 0].$$

Hence $\lim E[R_n]/n = 0 \iff$ the random walk is recurrent.

Proof of Proposition 7.1.1. We have

$$E[N] = \sum_{n=0}^{\infty} P[N > n] = \sum_{n \geq 0} P[\forall k \leq n \quad S_k \leq 0].$$

By duality, this ...

$$= \sum_{n \geq 0} P[\forall k \leq n \ S_n - S_{n-k} \leq 0] = \sum_{n \geq 0} P[S_n \text{ is a new minimum}] = E[M].$$

mes at which minima occur form a renewal sequence with interarrival times allowed to be infinite. Indeed, there are only a finite number of renewals since $\mu > 0$; their number is a geometric random variable, so $E[M] < \infty$. Since this means that $E[N] < \infty$ also, we may apply Wald's equation (not its extension) to conclude when $\mu < \infty$. If $\mu = \infty$, then $E[S_N] \geq E[S_N \mathbf{1}_{\{N=1\}}] = E[X_1 \mathbf{1}_{\{X_1 > 0\}}] \geq E[X_1] = \infty$, so the result holds still. ■

Proof of Proposition 7.1.2. We have

$$\begin{aligned} E[R_n] &= \sum_{k=0}^n P[\forall j \in (k, n] \ S_k \neq S_j] \\ \dots \\ &= \sum_{k=0}^n P[\forall i \in (0, n-k] \ S_0 \neq S_i] \\ &= \sum_{k=0}^n P[\forall i \in (0, k] \ S_i \neq 0]. \end{aligned}$$

Since the summands $\rightarrow P[\text{no return to } 0]$, the result follows. ■

REMARK. For the proof of Proposition 7.1.2, we used stationarity and reversed counting, rather than duality (which the book uses). Thus, the result is more general: it applies to stationary $\langle X_i \rangle$. We also don't need the values of X to lie in $\mathbb{R}$. It is also true that $R_n/n \rightarrow P[\text{no return}]$ a.s. (Kesten, Spitzer, and Whitman; use the Kingman subadditive ergodic theorem to prove this).

EXAMPLE 7.1(A). Suppose that $X = \pm 1$ with probability $\frac{1}{2} \pm \alpha$. Then $P[\text{no return}] = 1 - f_{00} = 2|\alpha|$ (see the calculation of the Green function or the gambler's ruin problem).
...

▷ Read Proposition 7.1.3 in the book.

Unit 15

Markov Chains

We now go beyond having real-valued random variables. In this chapter, we consider stochastic processes indexed by $\mathbb{N}$ (or $\mathbb{Z}^+$) and which can take values in a finite or countable set called the **state space**. For simplicity, the states will often be labelled $0, 1, 2, \dots$, but there may be no numerical significance to the labels.

What replaces independence of increments is the **Markovian property** that the future and the past are independent given the present: given $n, r, i_0, i_1, \dots, i_{n+r}$ with $P[X_n = i_n] > 0$, consider the past event $A := [X_0 = i_0, X_1 = i_1, \dots, X_{n-1} = i_{n-1}]$ and the future event $B := [X_{n+1} = i_{n+1}, X_{n+2} = i_{n+2}, \dots, X_{n+r} = i_{n+r}]$. Then

$$P(A, B \mid X_n = i_n) = P(A \mid X_n = i_n)P(B \mid X_n = i_n). \quad (\text{N1})$$

These events A and B are very basic events. If we sum over all possibilities, then we see that the same equation holds for all pairs of events A and B where A depends only on X_j for $j < n$ and B depends only on X_k for $k > n$

The Markovian property is equivalent to the following property: $\forall n \ \forall i_0, \dots, i_{n+1}$ with $P[X_0 = i_0, \dots, X_n = i_n] > 0$,

$$P[X_{n+1} = i_{n+1} \mid X_0 = i_0, \dots, X_n = i_n] = P[X_{n+1} = i_{n+1} \mid X_n = i_n]. \quad (\text{N2})$$

↓ To see this, suppose first that the Markov property (N1) holds; let A be as in the definition and let $B := [X_{n+1} = i_{n+1}]$. Then the left-hand side of (N2) is equal to

$$P(B \mid A, X_n = i_n) = \frac{P(A, B \mid X_n = i_n)}{P(A \mid X_n = i_n)} = P(B \mid X_n = i_n)$$

by (N1), which shows (N2). Conversely, suppose that (N2) holds. The calculation we just did shows that (N1) holds for A as in the definition and for B of the special form $B := [X_{n+1} = i_{n+1}]$. Summing over possibilities shows that it also holds for every A that depends only on times before the present, n . We can also reformulate (N1) as

$$P(B \mid A, X_n = i_n) = P(B \mid X_n = i_n). \quad (*)$$

Now suppose that B is general, as in the definition. Let A depend only on the past, X_j for $j < n$. Then we have

$$\begin{aligned} P(B \mid A, X_n = i_n) &= \prod_{j=1}^r P(X_{n+j} = i_{n+j} \mid A, X_n = i_n, X_{n+1} = i_{n+1}, \dots, X_{n+j-1} = i_{n+j-1}) \\ &= \prod_{j=1}^r P(X_{n+j} = i_{n+j} \mid X_{n+j-1} = i_{n+j-1}) \end{aligned}$$

by (*). Since this does not depend on A , it follows that (N1) holds.

The right-hand side of (N2) is known as a **transition probability**. The analogue of stationary increments is that this doesn't depend on n , only on i_n and i_{n+1} ; that is,

$$P[X_{n+1} = j \mid X_n = i] =: p_{ij}$$

does not depend on n . In this case, the process is called a **(homogeneous) Markov chain**. From the transition probabilities and the initial distribution $p_i := P[X_0 = i]$, we can calculate all probabilities:

$$P[X_0 = i_0, X_1 = i_1, \dots, X_n = i_n] = p_{i_0} p_{i_0 i_1} p_{i_1 i_2} \dots p_{i_{n-1} i_n}.$$

...

It's really better to say that a Markov chain is a collection of probability measures P_i , representing the chain when it starts in state i , with the property that

$$\begin{aligned} P_{i_0}[X_1 = i_1, \dots, X_{n-1} = i_{n-1}, X_n = i_n, X_{n+1} = i_{n+1}, X_{n+2} = i_{n+2}, \dots, X_{n+r} = i_{n+r}] \\ = P_{i_0}[X_1 = i_1, \dots, X_{n-1} = i_{n-1}, X_n = i_n] P_{i_n}[X_1 = i_{n+1}, \dots, X_r = i_{n+r}] \end{aligned}$$

for all n, r , and $i_0, \dots, i_{n+r}$. This avoids problems of p_i possibly being 0 for some (even most) i . Again, by summing over all possibilities, we see that for all pairs of events A and B where A depends only on X_m for $m < n$ and B depends only on X_k for $k > n$, we have

$$P_i[A, X_n = j, B] = P_i[A, X_n = j] P_j(\tau_n B).$$

Here, we define τ_n of an event that depends only on X_k for $k > n$ by subtracting n from all the indices of the random variables; i.e.,

$$\tau_n[X_{n+1} = i_{n+1}, X_{n+2} = i_{n+2}, \dots, X_{n+r} = i_{n+r}] := [X_1 = i_{n+1}, X_2 = i_{n+2}, \dots, X_r = i_{n+r}]$$

and similarly for unions of such events.

EXAMPLE (I.I.D. TRIALS). If X_n are i.i.d., then $\langle X_n \rangle$ is a Markov chain.

EXAMPLE 4.1(C) (SUMS OF I.I.D. $\mathbb{Z}$ -VALUED RANDOM VARIABLES). Here, the state space is $\mathbb{Z}$

EXAMPLE 4.1(A) (THE M/G/1 QUEUE). Let $X_n :=$ the number of customers in the system when the n th customer leaves the system, and $X_0 := 0$. The memoryless property of the arrival stream shows that this is a Markov chain. ... Now

$$X_{n+1} = \begin{cases} X_n - 1 + Y_{n+1} & \text{if } X_n \geq 1, \\ Y_{n+1} & \text{if } X_n = 0, \end{cases}$$

where $Y_{n+1} :=$ the number of arrivals during the period of service of the $(n+1)$ st customer. (Note that those customers who arrive during a free period are not counted by any of the Y_n . Only customers who have to wait in queue are counted by some Y_n .) Thus, Y_n are i.i.d. and for $j \in \mathbb{N}$, with λ denoting the rate of the arrivals, we have

$$\begin{aligned} P[Y = j] &= E \left[P[Y = j \mid \underbrace{\text{service time}}_{Z \sim G}] \right] \\ &= E \left[e^{-\lambda Z} (\lambda Z)^j / j! \right] = \int_0^\infty e^{-\lambda x} \frac{(\lambda x)^j}{j!} dG(x). \end{aligned}$$

Thus,

$$\begin{aligned} p_i &= \begin{cases} 1 & \text{if } i = 0, \\ 0 & \text{otherwise,} \end{cases} \\ p_{0j} &= \int_0^\infty e^{-\lambda x} \frac{(\lambda x)^j}{j!} dG(x) \quad (j \geq 0), \\ p_{i,i-1+j} &= p_{0j} \quad (i \geq 1, j \geq 0), \\ p_{i,j} &= 0 \quad (i \geq 2, j \leq i-2). \end{aligned}$$

▷ Read pp. 163--165 in the book.

Note: In Example 4.1(d), Ross says that the case of summands $X_i = \pm 1$ is “simple random walk”. Usually, this is called “nearest-neighbor random walk” and the term “simple” is reserved for the case when $P[X_i = 1] = \frac{1}{2}$. We will *not* use Ross’s terminology.

EXAMPLE 4.4(A) (THE GAMBLER’S RUIN PROBLEM). A gambler needs $\$N$ but has only $\$i$ ($1 \leq i \leq N-1$). He plays games that give him chance p of winning $\$1$ and $q := 1-p$ of losing $\$1$ each time. When his fortune is either $\$N$ or 0, he stops. What is his chance of success?

Solution. The gambler's fortune is a Markov chain on $\{0, 1, \dots, N\}$. Let α_i be the probability of success. Then $\alpha_0 = 0$, $\alpha_N = 1$, and

$$\text{for } 1 \leq i \leq N-1 \quad \alpha_i = p\alpha_{i+1} + q\alpha_{i-1},$$

which gives

$$\alpha_{i+1} - \alpha_i = \frac{q}{p}(\alpha_i - \alpha_{i-1}).$$

Therefore

$$\alpha_{i+1} - \alpha_i = \left(\frac{q}{p}\right)^i (\alpha_1 - \alpha_0) = \left(\frac{q}{p}\right)^i \alpha_1.$$

To determine α_1 , add these up:

$$1 = \alpha_N = \sum_{i=0}^{N-1} (\alpha_{i+1} - \alpha_i) = \sum_{i=0}^{N-1} \left(\frac{q}{p}\right)^i \alpha_1,$$

so

$$\alpha_1 = 1 / \sum_{i=0}^{N-1} \left(\frac{q}{p}\right)^i.$$

By adding only the equations for $\alpha_1 - \alpha_0, \alpha_2 - \alpha_1, \dots, \alpha_i - \alpha_{i-1}$, we get

$$\alpha_i = \sum_{j=0}^{i-1} \left(\frac{q}{p}\right)^j / \sum_{j=0}^{N-1} \left(\frac{q}{p}\right)^j.$$

For example, when $p = \frac{1}{2}$, we have $\alpha_i = i/N$.

For an application to statistics, see Exercise 4.30, p. 224 (with answer in the back).

§4.2. Chapman-Kolmogorov Equations and Classification of States.

Let $p_{ij}^{(n)}$ be the n -step transition probabilities, i.e.,

$$p_{ij}^{(n)} := P_i[X_n = j].$$

Note that ...

$$p_{ij}^{(n+m)} = \sum_k P_i[X_{n+m} = j, X_n = k] = \sum_k P_i[X_n = k] P_k[X_m = j] = \sum_k p_{ik}^{(n)} p_{kj}^{(m)}.$$

Notice that this is matrix multiplication: If $P^{(n)} := (p_{ij}^{(n)})$, then the above equation is $P^{(n+m)} = P^{(n)} P^{(m)}$, whence $P^{(n)} = P^n$, where $P := (p_{ij})$.

We say that state j is **accessible** from state i if $\exists n \geq 0$ $p_{ij}^{(n)} > 0$. If i and j are accessible from each other, we say they **communicate** and write $i \leftrightarrow j$. It is not hard to check that $\leftrightarrow$ is an equivalence relation. ... If there is only one equivalence class, the Markov chain is called **irreducible**. The **period** of state i is the g.c.d. of $\{n \geq 0; p_{ii}^{(n)} > 0\}$, written $d(i)$. If $d(i) = 1$, then state i is called **aperiodic**.

PROPOSITION 4.2.2. If $i \leftrightarrow j$, then $d(i) = d(j)$.

Proof. It suffices to show that $d(j) \mid d(i)$. [The symbol $k \mid n$ here stands for “divides” and means that $n/k \in \mathbb{Z}$.] Let $p_{ii}^{(s)} > 0$ and choose m, n such that $p_{ij}^{(m)} > 0$ and $p_{ji}^{(n)} > 0$. First, we have $p_{jj}^{(n+m)} \geq p_{ji}^{(n)} p_{ij}^{(m)} > 0$, so $d(j) \mid (n+m)$. Second, we have

$$p_{jj}^{(n+s+m)} \geq p_{ji}^{(n)} p_{ii}^{(s)} p_{ij}^{(m)} > 0,$$

so $d(j) \mid (n+s+m)$. Therefore, $d(j) \mid s$, so $d(j) \mid d(i)$. ■

Let $f_{ij}^{(n)}$ be the probability that the first* transition into j is at time n when the chain starts in state i : $f_{ij}^{(0)} := 0$ and for $n \geq 1$,

$$f_{ij}^{(n)} := P_i[X_n = j \text{ and } \forall k \in [1, n-1] X_k \neq j].$$

Then $f_{ij} := \sum_{n=1}^{\infty} f_{ij}^{(n)}$ is the probability of *ever* making a transition into state j when the chain starts in i . We call j **recurrent** if $f_{jj} = 1$ and **transient** otherwise. The function

$$G(i, j) := \sum_{n=0}^{\infty} p_{ij}^{(n)},$$

the expected number of visits to j for the chain started at i , ... is called the **Green function** of the Markov chain.

* f stands for “first”

PROPOSITION 4.2.3. *State j is transient iff $G(j, j) < \infty$. If state j is transient, then a.s. the number of visits to j starting from j is finite, while if state j is recurrent, then a.s. the number of visits to j starting from j is infinite.*

Proof. Each visit to j is followed (at some time) by another visit to j with probability f_{jj} . Hence the number of visits is geometric with mean $(1 - f_{jj})^{-1}$ But we already know that the mean number of visits to j when the chain starts in j is $G(j, j)$. Therefore $(1 - f_{jj})^{-1} = G(j, j)$, so that $f_{jj} < 1$ iff $G(j, j) < \infty$ iff the geometric distribution of the number of visits to j has finite mean and hence is finite a.s. ... ■

COROLLARY. *If a Markov chain has only finitely many states, then some state is recurrent.*

...

COROLLARY 4.2.4. *If $i \leftrightarrow j$ and i is recurrent, then j is recurrent.*

Proof. Fix m and n such that $p_{ij}^{(m)} > 0$ and $p_{ji}^{(n)} > 0$. Then

$$\forall s \geq 0 \quad p_{jj}^{(n+s+m)} \geq p_{ji}^{(n)} p_{ii}^{(s)} p_{ij}^{(m)},$$

whence $\sum_s p_{jj}^{(n+s+m)} \geq p_{ji}^{(n)} p_{ij}^{(m)} G(i, i) = \infty$. ■

PROPOSITION. *If i is recurrent and j is accessible from i , then $f_{ij} = 1$ and $i \leftrightarrow j$.*

Proof. Let $X_0 = i$ and fix n such that $p_{ij}^{(n)} > 0$. Let $A_0 := [X_n = j]$ and let $T_1 := \min\{k \geq n; X_k = i\}$. Let $A_1 := [X_{T_1+n} = j]$ and $T_2 := \min\{k \geq T_1 + n; X_k = i\}$. In general, set $A_r := [X_{T_r+n} = j]$ and $T_{r+1} := \min\{k \geq T_r + n; X_k = i\}$. Then $\langle A_r \rangle$ are independent and each have probability $p_{ij}^{(n)}$, so one of them occurs. ... Thus, $f_{ij} = 1$; since $f_{ii} = 1$, it follows that i is accessible from j ■

Actually, we are using something stronger than the Markov property here and in the proof of Proposition 4.2.3, namely, a special case of what is called the **strong Markov property**. It always holds for discrete-time Markov chains, and usually, but not always, for continuous-time ones. It says the following. Given a Markov chain $\langle X_n \rangle$, call a random variable N with values in $\mathbb{N} \cup \{\infty\}$ a **stopping time** if for all n , the event $[N = n]$ (or, if you prefer, its indicator $\mathbf{1}_{[N=n]}$) depends only on $X_0, X_1, \dots, X_n$, written $[N = n] \in \sigma(X_0, X_1, \dots, X_n)$; in other words, there are functions $\phi_n: \mathbb{N}^{n+1} \rightarrow \{0, 1\}$ such that $N = n$ iff $\phi_n(X_0, X_1, \dots, X_n) = 1$. Write $\psi(i, B) := P_i[(X_0, X_1, \dots) \in B]$. The strong Markov property says that if N is a stopping time and B is an event, then

$$P\left[(X_N, X_{N+1}, \dots) \in B \mid X_0, X_1, \dots, X_N\right] = \psi(X_N, B)$$

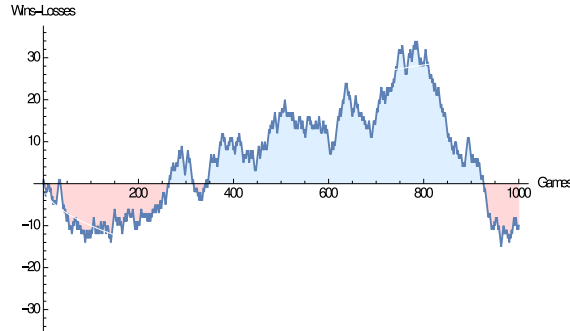
on the event that $N < \infty$. In the cases we are using, X_N is a fixed state, so this is easier to interpret. Note that the conditioning on $X_0, X_1, \dots, X_N$ implicitly includes conditioning on N . The proof of the strong Markov property is not hard: Given $i_0, i_1, \dots$ and n with $\phi_n(i_0, i_1, \dots, i_n) = 1$, the event $[N = n]$ is implied by the event $[X_0 = i_0, X_1 = i_1, \dots, X_n = i_n]$, whence

$$\begin{aligned} P\left[(X_N, X_{N+1}, \dots) \in B \mid N = n, X_0 = i_0, X_1 = i_1, \dots, X_n = i_n\right] \\ &= P\left[(X_n, X_{n+1}, \dots) \in B \mid X_0 = i_0, X_1 = i_1, \dots, X_n = i_n\right] \\ &= P\left[(X_n, X_{n+1}, \dots) \in B \mid X_n = i_n\right] \\ &= \psi(i_n, B) \end{aligned}$$

by the Markov property and homogeneity.

An irreducible Markov chain is called **transient** or **recurrent** according as its states are.

In particular, the Markov chain of Example 1.9(a) is recurrent and a.s. every bead is visited.



EXAMPLE 4.2(A). Consider the Markov chain on $\mathbb{Z}$ such that, for a given p and all i , $p_{i,i+1} = p$ and $p_{i,i-1} = 1 - p$. If $p \neq \frac{1}{2}$, then the SLLN shows that the chain is transient. We show that for $p = 1/2$, it is recurrent. Note that it has period 2. Now

$$p_{00}^{(2n)} = \binom{2n}{n} \frac{1}{2^{2n}}.$$

... Stirling's approximation $n! \sim \sqrt{2\pi n}(n/e)^n$ yields

$$p_{00}^{(2n)} \sim \frac{1}{\sqrt{\pi n}}, \tag{N3}$$

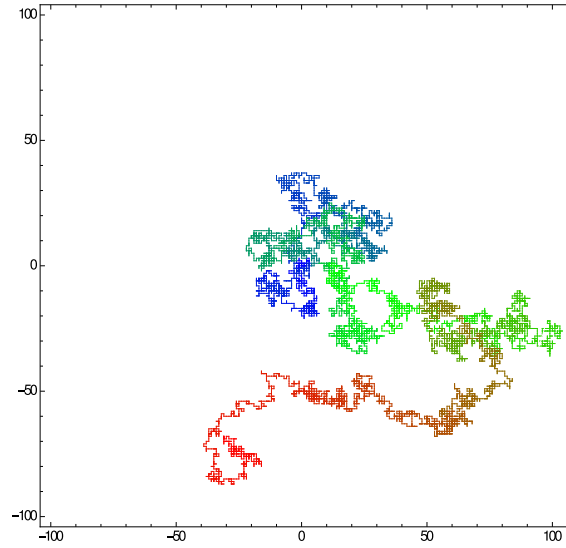
whence $G(0, 0) = \infty$.

Second proof. Let $q := 1 - p$. Then $p_{00}^{(2n)} = \binom{2n}{n} p^n q^n$. Now

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

... Therefore $G(0, 0) = \sum_{n \geq 0} \binom{-1/2}{n} (-4pq)^n = (1 - 4pq)^{-1/2} = |1 - 2p|^{-1}$. Thus, ... $G(0, 0) = \infty$ iff $p = 1/2$. Also, we see that $f_{00} = 2(p \wedge q)$

Third proof when $p = 1/2$. Let $a := f_{10}$. By symmetry, we have $a = f_{-1,0}$, whence $f_{00} = a$ We also have $a = (1 + f_{20})/2$... and $f_{20} = f_{21}f_{10} = a^2$, ... whence $(a - 1)^2 = 0$, so $a = 1$.



10,000 steps of 2D random walk.

One of the most famous theorems in probability extends this to higher dimensions:

PÓLYA'S THEOREM. *Simple random walk on the lattice $\mathbb{Z}^d$ is transient iff $d \geq 3$.*

This follows from:

PROPOSITION. *For simple random walk on $\mathbb{Z}^d$,*

$$p_{00}^{(2n)} \sim 2 \left(\frac{d}{4\pi n} \right)^{d/2}$$

as $n \rightarrow \infty$.

Idea of proof: Let $N_i(n)$ be the number of steps among the first n in direction i . By the WLLN, $N_i(2n) \sim 2n/d$ and $P[\forall i \ N_i(2n) \text{ is even}] \rightarrow 2^{-(d-1)}$. Given the values of $N_i(2n)$, the d coordinates of X_{2n} are independent, so the result follows from (N3). ...

§4.3. Limit Theorems.

Let μ_{ii} denote the expected number of transitions needed to return to i starting from i . Let $N_j(n)$ be the number of visits to j by time n . These visits form a delayed renewal process if j is recurrent. Actually, even when j is transient, the visits form a delayed renewal process, albeit one with only a finite number of renewals a.s. Thus, with the convention that $a/\infty := 0$ for any finite a , our results on renewal theory give:

THEOREM 4.3.1. *If $i \leftrightarrow j$, then*

- (i) *when the Markov chain starts from state i , we have $\lim_{n \rightarrow \infty} \frac{N_j(n)}{n} = \frac{1}{\mu_{jj}}$ a.s.;*
- (ii) $\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n p_{ij}^{(k)} = \frac{1}{\mu_{jj}};$
- (iii) *when j is aperiodic, we have $\lim_{n \rightarrow \infty} p_{ij}^{(n)} = \frac{1}{\mu_{jj}};$*
- (iv) $\lim_{n \rightarrow \infty} p_{jj}^{(nd)} = \frac{d}{\mu_{jj}},$ *where $d := d(j)$ is the period of j .*

The results we are using are: (i) Proposition 3.3.1; (ii) The Elementary Renewal Theorem; (iii) and (iv) Blackwell's Renewal Theorem. However, if j is transient, then we use the fact that the expected number of visits to j is finite, even when starting in i , whence $\sum_{k=1}^{\infty} p_{ij}^{(k)} < \infty$.

We call a recurrent state i **positive recurrent** if $\mu_{ii} < \infty$ and **null recurrent** otherwise.

PROPOSITION. *Every finite-state Markov chain has a state that is positive recurrent.*

...

PROPOSITION 4.3.2. *If $i \leftrightarrow j$ and i is null recurrent, then so is j .*

Proof. Let k and ℓ be such that $p_{ij}^{(k)} > 0$ and $p_{ji}^{(\ell)} > 0$. Let $d = d(i) = d(j)$. Since i is null recurrent, Theorem 4.3.1(iv) tells us that

$$0 = \lim_{n \rightarrow \infty} p_{ii}^{(nd+k+\ell)} \geq \limsup_{n \rightarrow \infty} p_{ij}^{(k)} p_{jj}^{(nd)} p_{ji}^{(\ell)} = p_{ij}^{(k)} p_{ji}^{(\ell)} \cdot \frac{d}{\mu_{jj}},$$

whence $\mu_{jj} = \infty$. ■

As was the case for renewal processes, stationary Markov chains arise from limiting probabilities used as initial distributions. Recall that $\langle X_n \rangle$ is stationary if $\forall k \geq 0$ $\langle X_n, X_{n+1}, \dots, X_{n+k} \rangle$ has a joint distribution that is the same for each n . The homogeneous Markov property shows that this follows for all k if it holds for $k = 0$ and $n \in \{0, 1\}$:

$$\begin{aligned} \forall j \quad p_j &= P[X_0 = j] = P[X_1 = j] = \sum_{i=0}^{\infty} P[X_0 = i, X_1 = j] \\ &= \sum_{i=0}^{\infty} p_i p_{ij}. \end{aligned}$$

... We call an initial distribution $\langle p_j \rangle$ **stationary** if this holds.

An irreducible Markov chain is called **positive recurrent** or **null recurrent** according as its states are.

THEOREM 4.3.3. *Consider an irreducible aperiodic Markov chain and write*

$$\pi_j := \lim_{n \rightarrow \infty} p_{ij}^{(n)} = \frac{1}{\mu_{jj}}.$$

The following are equivalent:

- (i) *the chain is positive recurrent;*
- (ii) $\exists$ *a stationary probability distribution;*
- (iii) $\langle \pi_j \rangle$ *is the unique stationary probability distribution.*

In this case, if $x_i \geq 0$, $c := \sum_j x_j > 0$, and $\forall j \quad x_j = \sum_i x_i p_{ij}$, then $c < \infty$ and $\forall i \quad x_i = c\pi_i$.

LEMMA (FATOU'S LEMMA FOR SERIES). *If $\alpha_n(j) \geq 0$, $\lim_{n \rightarrow \infty} \alpha_n(j) = \alpha(j)$, and $\lim_{n \rightarrow \infty} \sum_j \alpha_n(j) = \alpha$, then $\alpha \geq \sum_j \alpha(j)$.*

Proof. $\forall J \quad \sum_{j \leq J} \alpha(j) = \lim_n \sum_{j \leq J} \alpha_n(j) \leq \lim_n \sum_{j=0}^{\infty} \alpha_n(j) = \alpha$, so $\sum_{j=0}^{\infty} \alpha(j) \leq \alpha$. ■

LEMMA (LDCT FOR SERIES). *If $|\alpha_n(j)| \leq \beta(j)$, $\sum_{j=0}^{\infty} \beta(j) < \infty$, and $\lim_{n \rightarrow \infty} \alpha_n(j) = \alpha(j)$, then*

$$\lim_{n \rightarrow \infty} \sum_{j=0}^{\infty} \alpha_n(j) = \sum_{j=0}^{\infty} \alpha(j).$$

Proof. We have

$$\left| \sum_{j=0}^{\infty} \alpha_n(j) - \sum_{j=0}^{\infty} \alpha(j) \right| \leq \sum_{j=0}^{\infty} \left| \alpha_n(j) - \alpha(j) \right| \leq \sum_{j=0}^J \left| \alpha_n(j) - \alpha(j) \right| + \sum_{j>J} 2\beta(j).$$

Now let $n \rightarrow \infty$. Then let $J \rightarrow \infty$. ■

Proof of Theorem 4.3.3. (i) $\Rightarrow$ (ii): Since $p_{ij}^{(n+1)} = \sum_k p_{ik}^{(n)} p_{kj}$, Fatou's Lemma gives $\pi_j \geq \sum_k \pi_k p_{kj}$. Since $\sum_j p_{ij}^{(n)} = 1$, Fatou's Lemma also gives $\sum_{j=0}^{\infty} \pi_j \leq 1$. Thus $\sum_j \pi_j \geq \sum_j \sum_k \pi_k p_{kj} = \sum_k \pi_k \sum_j p_{kj} = \sum_k \pi_k$, whence $\forall j \pi_j = \sum_k \pi_k p_{kj}$ Therefore, $p_i := \pi_i / \sum \pi_j$ form a stationary probability distribution, where the denominator is positive since each $\pi_j > 0$.

(ii) $\Rightarrow$ (iii): If $\langle p_i \rangle$ is any stationary probability distribution, then $\forall n, j \ p_j = \sum_{i=0}^{\infty} p_i p_{ij}^{(n)}$, ... whence by the LDCT, $p_j = \sum_{i=0}^{\infty} p_i \pi_j = \pi_j$. That is, $\langle \pi_j \rangle$ is a stationary probability distribution and is unique.

(iii) $\Rightarrow$ (i) since some $\pi_j > 0$.

Finally, in the case of positive recurrence and $x_j = \sum_i x_i p_{ij}$, we have $\forall n \ x_j = \sum_i x_i p_{ij}^{(n)}$ (as $[x_j]$ is a left eigenvector for P), ... so $x_j \geq \sum_i x_i \pi_j = c \pi_j$. This means that $c < \infty$, whence the LDCT gives $x_j = c \pi_j$. ■

In the irreducible positive recurrent periodic case, the unique stationary probability distribution is still $\pi_j := 1/\mu_{jj} = \lim_{n \rightarrow \infty} p_{jj}^{(nd)}/d$. See Exercise 4.17, p. 221 (solution in the back of the book).

EXAMPLE (SIMPLE RANDOM WALK ON $\mathbb{Z}$). This is null recurrent. If not, then every solution to $\forall j \ x_j = \sum_i x_i p_{ij}$ satisfies $\sum_j x_j < \infty$. However, $x_j \equiv 1$ is a solution. ...

§4.5. Branching Processes.

An individual has k children with probability p_k , where $\sum_{k=0}^{\infty} p_k = 1$. The children reproduce independently according to the same offspring distribution. Let Z_n be the size of the n th generation. Clearly Z_n is a Markov chain, called a **Galton-Watson branching process**. The initial state Z_0 is usually assumed to be 1.

This was introduced in order to study British family names. It is also of interest in biology, chain reactions, electron multipliers, and analysis of various probabilistic processes. Many, many variations have been (and continue to be) studied.

Let L be a random variable with $P[L = k] = p_k$ and let $\langle L_i^{(n)} ; n, i \geq 1 \rangle$ be independent copies of L , so that $Z_{n+1} = \sum_{i=1}^{Z_n} L_i^{(n+1)}$. The **probability generating function** (p.g.f.) of L is

$$f(s) := E[s^L] = \sum_{k \geq 0} p_k s^k \quad (0 \leq s \leq 1).$$

PROPOSITION. The p.g.f. of Z_n is $f^{(n)} := \underbrace{f \circ \dots \circ f}_{n \text{ times}}$.

Proof. $E[s^{Z_n}] = E\left[E\left[s^{\sum_{i=1}^{Z_{n-1}} L_i^{(n)}} \mid Z_{n-1}\right]\right] = E\left[\prod_{i=1}^{Z_{n-1}} E[s^{L_i^{(n)}}]\right] = E[f(s)^{Z_{n-1}}]$. Apply this n times. ■

Write $q := P[Z_n \rightarrow 0] = P[\exists n \ Z_n = 0]$.

COROLLARY. $q = \lim_{n \rightarrow \infty} f^{(n)}(0)$.

Proof. $q = \lim_n P[Z_n = 0] = \lim_n f^{(n)}(0)$. ■

Looking at a graph of f , which is increasing and convex, we see that ...

PROPOSITION. *Suppose $p_1 \neq 1$. We have $q = 1 \iff f'(1) \leq 1$. Also, q is the smallest root of $f(s) = s$ in $[0, 1]$ — the only other possible root being 1.*

Note that $f'(1) = E[L] =: m$, the mean number of offspring per individual.

Unit 16

Continuous-Time Markov Chains

§5.2. Continuous-Time Markov Chains.

This section consists of definitions.

A stochastic process with time being an interval in $\mathbb{R}$ is called **Markov** if the future and past are independent given the present: $\forall t \{X(s); s > t\}$ and $\{X(s); s < t\}$ are independent given $X(t)$. If the number of states is countable, the process is called a **chain**. We then identify the states with $\mathbb{N}$. We will deal only with Markov chains that do not have instantaneous states and are right continuous, i.e., with probability 1 $\forall i \in \mathbb{N} \forall t X(t) = i \Rightarrow \exists \varepsilon > 0 \forall s \in (0, \varepsilon) X(t+s) = i$. We also assume homogeneous transition probabilities $p_{ij}(s) := P[X(t+s) = j \mid X(t) = i]$.

By the Markov property, the time spent in each state is a memoryless random variable, hence is an exponential random variable; call the rate ν_i when in state i . Let the probability distribution of the next state visited be P_{ij} ; the next state is independent of the time spent $\downarrow$ in i by the Markov property again. Indeed, fix a state j . Let A_t be the event that the next state after t is j and B_t be the event that $X(t) = i$. For a fixed $t \geq 0$, let τ be the maximal length of an interval containing t during which the Markov chain is in state i . We claim that A_t and τ are independent given B_t , i.e., $P[A_t, \tau > s \mid B_t] = P(A_t \mid B_t)P[\tau > s \mid B_t]$ for all $s > 0$. To see this, let $T := \sup\{u < t; X(u) \neq i\}$. Then

$$P[A_t, \tau > s \mid B_t] = E[P[A_t, \tau > s \mid B_t, T]]. \quad (*)$$

Furthermore, $P[A_t, \tau > s \mid B_t, T = u] = P[A_t, \forall v \in [t, u+s] X(v) = i \mid B_t]$ by the Markov property. This in turn equals $P[\forall v \in [t, u+s] X(v) = i \mid B_t]P[A_{u+s} \mid X(u+s) = i] = P[\tau > s \mid B_t, T = u]P[A_{u+s} \mid B_{u+s}]$ by the Markov property. By homogeneity, $P(A_{u+s} \mid B_{u+s}) = P(A_t \mid B_t)$. Substituting this in $(*)$ gives the desired result.

This gives a constructive view of a continuous time Markov chain: use a timer at each state; when it rings, move according to a discrete-time Markov chain. However, there is a difficulty: what if we make an infinite number of transitions in a finite time period? Example: $P_{i,i+1} = 1$, $\nu_i = i^2$. If τ_i is the time spent in i , then

$$E[\tau_i] = \frac{1}{i^2}, \quad \text{so} \quad E\left[\sum \tau_i\right] < \infty,$$

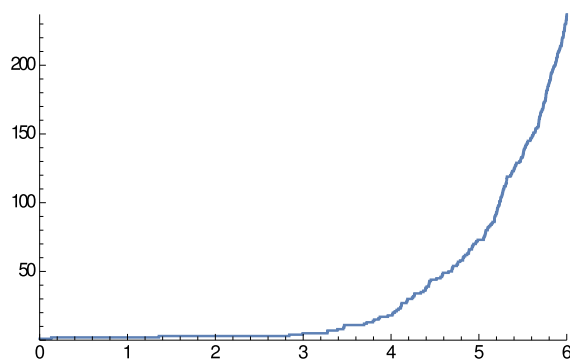
so $\sum \tau_i < \infty$ a.s. **The paradoxes of Lincoln's penny.** We will not treat such chains, only **regular** ones, i.e., ones defined on $[0, \infty)$ that with probability 1 make only a finite number of transitions in $[0, N)$ for every $N < \infty$.

Another construction of continuous-time Markov chains is as follows. Let $q_{ij} := \nu_i P_{ij}$ for $i \neq j$. This is the **transition rate** from i to j . We could have at state i timers ... with rates q_{ij} ; the first to ring determines the next state: from Problem 1.34, p. 53, the probability of being the first to ring is proportional to the rate. ...

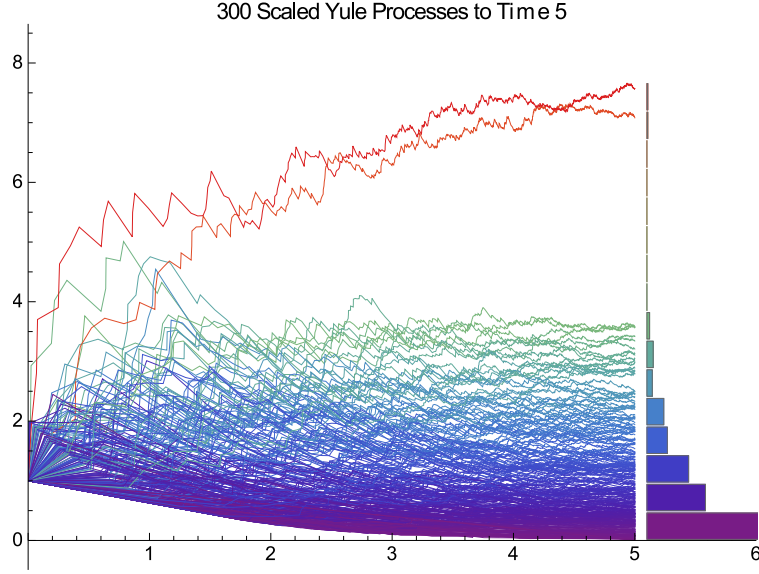
§5.3. Birth and Death Processes.

In case $P_{ij} = 0$ for $|i - j| > 1$, the chain is called a **birth and death process**. We think of the state as representing the size of a population. Let the **birth rates** be $\lambda_i := q_{i,i+1}$ and the **death rates** be $\mu_i := q_{i,i-1}$.

If there are no deaths, the process is called a **pure birth process**. The Poisson process, $\lambda_n = \lambda$ for $n \geq 0$, is such a process. Another is the **Yule process**, where $\lambda_n = n\lambda$ for $n \geq 0$.



A sample of a Yule process.



A sample of 300 scaled Yule processes, dividing by e^t .

... The Yule process is regular: as shown on p. 235 of the book, if τ_i denotes the time spent in state i , then for every $k \geq 1$ and $t \in (0, \infty)$, we have (if the chain starts in state 1)

$$P\left[\sum_{i=1}^k \tau_i \leq t\right] = (1 - e^{-\lambda t})^k,$$

whence $P\left[\sum_{i=1}^{\infty} \tau_i \leq t\right] = 0$

Another way to see this result, which says that $P[X(t) > k] = (1 - e^{-\lambda t})^k$, or that $X(t) \sim \text{Geom}(e^{-\lambda t})$, is the following. The time $\tau_i \sim \text{Exp}(\lambda i)$, which is also the distribution of the minimum of i independent $\text{Exp}(\lambda)$ random variables. Thus, $\sum_{i=1}^k \tau_i$ has the same distribution as $\max_{1 \leq i \leq k} Z_i$, where $Z_i \sim \text{Exp}(\lambda)$ are independent: τ_k has the same distribution as $\min_{1 \leq i \leq k} Z_i$, then τ_{k-1} the same as the minimum of the time from the minimum of Z_i to the next smallest, etc. But the cdf of $\max_{1 \leq i \leq k} Z_i$ is easy to calculate.

EXAMPLE 5.3(A).

- (i) Let $X(t)$ be the number of people in the system of an M/M/s queue, where arrivals have rate λ and service has rate μ . Then $\lambda_n = \lambda$ for $n \geq 0$, $\mu_n = n\mu$ for $1 \leq n \leq s$, and $\mu_n = s\mu$ for $n > s$.
- (ii) A linear growth process with immigration assumes that each individual in the population gives birth at exponential rate λ and dies at rate μ , while there is also immigration at rate θ . Thus $\lambda_n = n\lambda + \theta$ for $n \geq 0$ and $\mu_n = n\mu$ for $n \geq 1$. This can be shown

to be regular by a more general method than the one we used for a Yule process. Namely, let τ_i denote the i th sojourn time, i.e., the time spent between the transitions number i and $i + 1$. We wish to show that $\sum_i \tau_i = \infty$ a.s. If some state is visited i.o., then this sum contains a sum of infinitely many i.i.d. random variables, whence the sum is infinite. Otherwise, it contains the sum of the first sojourn times at each state. Let $\tau'_i \sim \text{Exp}(\nu_i)$. To see that $\sum_i \tau'_i = \infty$ a.s., note that $M_n := E\left[\sum_{i=1}^n \tau'_i\right] \rightarrow \infty$ as $n \rightarrow \infty$. Then show that $P\left[M_n^{-1} \sum_{i=1}^n \tau'_i \rightarrow 1\right] = 1$ by calculating the variance of $\sum_{i=1}^n \tau'_i \dots$

§5.4. The Kolmogorov Differential Equations.

A pure birth process is the easiest to analyze, since it can always be reduced to a sum of independent (though not necessarily identically distributed) exponential random variables. ... For other processes, we require new tools.

Recall that $p_{ij}(t) = P[X(s+t) = j \mid X(s) = i]$. There are two sets of differential equations that these functions satisfy, obtained by conditioning on intermediate states.

THEOREM 5.4.3 (KOLMOGOROV'S BACKWARD EQUATIONS). $\forall i, j, t$

$$p'_{ij}(t) = \sum_{k \neq i} q_{ik} p_{kj}(t) - \nu_i p_{ij}(t).$$

This can be written with matrices as

$$P'(t) = QP(t),$$

where $P(t) := (p_{ij}(t))_{i,j}$, $Q := (q_{ij})_{i,j}$, and

$q_{ii} := -\nu_i =$ the negative of the rate of transition out of i .

Proof. First we write an integral equation for $p_{ij}(t)$. Either the chain has jumped by time t or not; if it has, then its first jump is to some state $k \neq i$, from which it eventually reaches state j . Let τ be the time of the first jump. Thus, we have

$$\begin{aligned} p_{ij}(t) &= E_i[P_i[X(t) = j \mid \tau]] = E_i[\delta_{ij} \mathbf{1}_{\{\tau > t\}} + P_i[X(t) = j \mid \tau] \mathbf{1}_{\{\tau \leq t\}}] \\ &= \delta_{ij} e^{-\nu_i t} + E_i\left[\sum_{k \neq i} P_{ik} p_{kj}(t - \tau) \mathbf{1}_{\{\tau \leq t\}}\right] \\ &= \delta_{ij} e^{-\nu_i t} + \int_0^t \sum_{k \neq i} P_{ik} p_{kj}(t - s) \nu_i e^{-\nu_i s} ds \\ &= \delta_{ij} e^{-\nu_i t} + \int_0^t \sum_{k \neq i} q_{ik} p_{kj}(u) e^{-\nu_i(t-u)} du \\ &= e^{-\nu_i t} H_{ij}(t), \end{aligned}$$

where

$$H_{ij}(t) := \delta_{ij} + \int_0^t \sum_{k \neq i} q_{ik} p_{kj}(u) e^{\nu_i u} du.$$

Observe that the integrand is bounded by $\nu_i e^{\nu_i t}$ Therefore $H_{ij}(\cdot)$ is continuous, and thus so is $p_{ij}(\cdot)$. But this means that the integrand is continuous by the LDCT for series, ... whence $H_{ij}(\cdot)$ and $p_{ij}(\cdot)$ are differentiable and we can apply the Fundamental Theorem of Calculus to derive

$$p'_{ij}(t) = -\nu_i p_{ij}(t) + e^{-\nu_i t} \sum_{k \neq i} q_{ik} p_{kj}(t) e^{\nu_i t},$$

as desired. ■

If we use $t := 0$, then we get that $p'_{ij}(0) = q_{ij}$.

The name “backward equations” arises because we conditioned all the way back to the time of the first jump. The forward equations come from a more natural conditioning, yet are more difficult to establish—indeed, they do not always hold. The forward equations arise as follows. Since

$$p_{ij}(t+h) = \sum_k p_{ik}(t) p_{kj}(h) \quad (h > 0),$$

we have

$$\frac{p_{ij}(t+h) - p_{ij}(t)}{h} = \sum_{k \neq j} p_{ik}(t) \frac{p_{kj}(h)}{h} - p_{ij}(t) \frac{1 - p_{jj}(h)}{h}.$$

If we could interchange $\lim_{h \rightarrow 0+}$ with $\sum_{k \neq j}$, e.g., if there are only finitely many states, then we would get

$$p'_{ij}(t) = \sum_{k \neq j} p_{ik}(t) q_{kj} - p_{ij}(t) \nu_j,$$

or

$$P'(t) = P(t)Q$$

in matrix notation. (Note that the continuity of $p_{ik}(\cdot)$ allows a similar argument for the left-hand derivative.)

EXAMPLE 5.4(A) (THE TWO-STATE CHAIN). Let $q_{01} = \lambda$ and $q_{10} = \mu$. Then $\nu_0 = \lambda$, $\nu_1 = \mu$, and

$$Q = \begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix}.$$

Since the matrices are finite, the solution to $P'(t) = QP(t)$ is

$$P(t) = e^{Qt} \quad (\text{the multiplicative constant} = 1 \text{ since } P(0) = I).$$

This is intuitive from the following calculation:

$$P(t) = P(t/n)^n = (I + [P(t/n) - I])^n = (I + Qt/n + o(1/n))^n = e^{Qt}.$$

Exponentiation is calculated via diagonalization: If $Q = ADA^{-1}$, then

$$e^{Qt} = Ae^{Dt}A^{-1}.$$

Here, it is easily calculated that the eigenvalues of Q are 0 and $-(\lambda + \mu)$, with corresponding eigenvectors $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} \lambda \\ -\mu \end{pmatrix}$. Thus, we use

$$D := \begin{pmatrix} 0 & 0 \\ 0 & -(\lambda + \mu) \end{pmatrix}, \quad A := \begin{pmatrix} 1 & \lambda \\ 1 & -\mu \end{pmatrix}, \quad A^{-1} = \frac{1}{\lambda + \mu} \begin{pmatrix} \mu & \lambda \\ 1 & -1 \end{pmatrix},$$

so

$$\begin{aligned} e^{Qt} &= \frac{1}{\lambda + \mu} \begin{pmatrix} 1 & \lambda \\ 1 & -\mu \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & e^{-(\lambda + \mu)t} \end{pmatrix} \begin{pmatrix} \mu & \lambda \\ 1 & -1 \end{pmatrix} \\ &= \frac{1}{\lambda + \mu} \begin{pmatrix} \mu + \lambda e^{-(\lambda + \mu)t} & \lambda - \lambda e^{-(\lambda + \mu)t} \\ \mu - \mu e^{-(\lambda + \mu)t} & \lambda + \mu e^{-(\lambda + \mu)t} \end{pmatrix}. \end{aligned}$$

E.g., $p_{00}(t) = (\mu + \lambda e^{-(\lambda + \mu)t})/(\lambda + \mu)$.

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

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

Self Learning Materials (SLMs) have been published by the Directorate of Open and Distance Learning, University of Kalyani, Kalyani-741235, West Bengal and all the copyright reserved for University of Kalyani. No part of this work should be reproduced in any form without permission in writing from the appropriate authority of the University of Kalyani.

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

Director
Directorate of Open and Distance Learning
University of Kalyani

COURSE PREPARATION TEAM

1. Dr. Avijit Sarkar
Assistant Professor,
Department of Mathematics
University of Kalyani
2. Mr. Biswajit Mallick
Assistant Professor (Cont.)
Department of Mathematics
DODL, University of Kalyani
3. Ms. Audrija Choudhury
Assistant Professor (Cont.)
Department of Mathematics
DODL, University of Kalyani

June, 2019

Directorate of Open and Distance Learning, University of Kalyani
Published by the Directorate of Open and Distance Learning,
University of Kalyani, Kalyani-741235, West Bengal and Printed by
Printtech, 15A, Ambika Mukherjee Road, Kolkata-700056

All right reserved. No part of this work should be reproduced in any form without the permission in writing from the Directorate of Open and Distance Learning, University of Kalyani.

POST-GRADUATE DEGREE PROGRAMME (CBCS)
IN
MATHEMATICS

SEMESTER-II
SELF LEARNING MATERIAL

PAPER–MATP 2.4
(Pure Stream)

Differential Geometry - II
Topology - II



DIRECTORATE OF OPEN AND DISTANCE LEARNING
UNIVERSITY OF KALYANI
Kalyani, Nadia
West Bengal, India

Contents

Sl. No	Blocks	Units	Page
1	Differential Geometry – II	I	2-6
		II	7-11
		III	12-17
		IV	18-21
		V	22-28
		VI	29-32
		VII	33-34
2	Topology - II	IX	37-42
		X	43-48
		XI	49-55
		XII	56-61
		XIII	62-68
		XIV	69-74
		XV	75-80
		XVI	81-88

Elective Paper

MATP 2.4

Block - I

Marks : 50 (SEE : 40; IA : 10)

• Differential Geometry - II

Syllabus

Unit	Topic
1	Curves in the plane and space, surfaces in three-dimension, Smooth surface
2	Tangents and derivatives, normal and orientability, Examples of surfaces.
3	The first fundamental form, Length of curves on surfaces
4	Isometries of surfaces, Conformal mapping of surfaces
5	Curvature of surfaces, The second fundamental form, The Gauss and Weingarten map
6	Normal and geodesic curvatures, Parallel transport and covariant derivative.
7	Gaussian, mean and principal curvatures, Gauss Theorema Egregium, Minimal surface
8	The Gauss Bonnet Theorem. Abstract differentiable manifolds and examples, Tangent Spaces

Unit I

Course Structure

- Curves in the plane and space
 - Surfaces in three-dimension
 - Smooth surface
-

1 Curves

What are curves? As we have read so far, the line $y - 2x = 1$ is a curve (even though its not curved). A curve can be said to be a generalisation of a straight line whose curvature is not zero. Say a circle $x^2 + y^2 = 1$, or the parabola, $y = x^2$. All these curves can be described by means of their Cartesian equation $f(x, y) = c$, where, f is a function of x and y and c is a constant. In this perspective, a curve may be considered as the set of points, namely

$$\mathcal{C} = \{(x, y) \in \mathbb{R}^2 | f(x, y) = c\}$$

These are all curves in the $\mathbb{R}^2$. In $\mathbb{R}^3$, for example, the x -axis is the straight line given by $y = 0, z = 0$. More generally, a curve in $\mathbb{R}^3$ is defined as a pair of equations

$$f_1(x, y, z) = c_1, f_2(x, y, z) = c_2$$

There is another way to define curves which is more useful in many situations, which is viewed as the path traced out by a moving point in space.

Definition 1.1. A parametrized curve is a path in the xy -plane traced out by the point $(x(t), y(t))$ as the parameter t ranges over an interval I .

$$C = \{(x(t), y(t)) : t \in I\}$$

For example, the graph of the function $y = f(x), x \in I$ is a curve C parametrised by the equations

$$x(t) = t, y(t) = f(t), t \in I$$

The above equations give a parametrisation of the curve $y = f(x)$. Also, a given curve may have more than one parametrisation. For example, the curve $y = 2x, x \in [1, 3]$ can be parametrised as

$$x(t) = t, y(t) = 2t, t \in [1, 3]$$

It can also be parametrised as

$$x(t) = t + 1, y(t) = 2t + 2, t \in [0, 2]$$

Example 1.2. Consider the parabola $x = 1 - y^2$, $-1 \leq y \leq 1$. Then we can think of two parametrisations of it as

1. When we set $y(t) = t$ and $x(t) = 1 - t^2$, $t \in [-1, 1]$. Changing the domain to all real t gives us the whole parabola.
2. When we set $y(t) = \cos t$ and $x(t) = 1 - \cos^2 t$, $t \in [0, \pi]$. Changing the domain to all real t does not give us any more of the parabola.

Example 1.3. Now, consider the circle $x^2 + y^2 = 1$. If we take $x(t) = t$ and $y(t) = \sqrt{1 - t^2}$, then this will only represent the upper half of the circle since $\sqrt{1 - t^2} \geq 0$ always. Similarly, we could not have taken $y(t) = -\sqrt{1 - t^2}$ since that would have traced the lower half of the circle. So, let us take the parametrisation as $x(t) = \cos t$ and $y(t) = \sin t$, where $t \in [0, 2\pi]$. Then we see that the curve $(x(t), y(t))$ traces the whole of the circle as t traverses $[0, 2\pi]$.

There are other several examples which we will omit now since we have already studied those in DG I. We will straightaway move on to surfaces in three dimensions.

1.1 Surface in Three Dimensions

Just like a curve is the basic building block for figures in a plane, a surface is the basic building block for figures in space. A surface is essentially a curve with depth. Curves and surfaces are analogous in many ways. If you think of a curve as being the trace of the motion of a point in a plane, a surface is like the trace of the motion of a curve in space. Surfaces are continuous, meaning that given two points on a surface, you can start from one and reach the other without leaving that surface. Just like a curve is still one-dimensional, a surface, although it exists in three dimensions, is still two-dimensional. For example, when you build a curve by tracing the motion of a point, that curve, although it spans both length and width, has no width of its own. The curve doesn't have area, it only has length, one dimension. Similarly, a surface can span more than one plane, but it still does not have depth of its own. It only has two dimensions, length and width. We will work mostly with the simplest surface, a plane. Before formally defining a surface, we will see certain useful definitions.

Definition 1.4. A subset U of $\mathbb{R}^n$ is called Open if, whenever $\mathbf{a}$ is a point in U , there exists a positive ϵ such that

$$\mathbf{a} \in U \text{ and } \|\mathbf{u} - \mathbf{a}\| < \epsilon \text{ implies } \mathbf{u} \in U$$

The whole of $\mathbb{R}^n$ is an open set. Also, the open ball

$$\mathcal{D}_r(\mathbf{a}) = \{\mathbf{u} \in \mathbb{R}^n : \|\mathbf{u} - \mathbf{a}\| < r\},$$

with centre at $\mathbf{a}$ and radius $r > 0$. However, the set

$$\overline{\mathcal{D}}_r(\mathbf{a}) = \{\mathbf{u} \in \mathbb{R}^n : \|\mathbf{u} - \mathbf{a}\| \leq r\}$$

is not open.

Definition 1.5. If X and Y are subsets of $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively, a map $f : X \rightarrow Y$ is said to be continuous at a point $\mathbf{a} \in X$ if given any $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\mathbf{u} \in X \text{ and } \|\mathbf{u} - \mathbf{a}\| < \delta \text{ implies } \|f(\mathbf{u}) - f(\mathbf{a})\| < \epsilon$$

If f is continuous at every point of X , then it is said to be continuous on X . Composites of continuous maps are continuous.

Continuous maps can also be equivalently defined as follows:

f is continuous if and only if, for any open set $V \in \mathbb{R}^n$, there is an open set U in $\mathbb{R}^m$ such that $U \cap X = \{x \in X | f(x) \in V\}$.

Let us now formally define a surface in $\mathbb{R}^3$.

Definition 1.6. A subset S of $\mathbb{R}^3$ is a surface if, for every $\mathbf{p} \in S$, there is an open set U in $\mathbb{R}^2$ and an open set W in $\mathbb{R}^3$ containing $\mathbf{p}$ such that $S \cap W$ is homeomorphic to U . A subset of a surface S of the form $S \cap W$, where W is an open subset of $\mathbb{R}^3$, is called an open subset of S . A homeomorphism $\sigma : U \rightarrow S \cap W$ as in this definition is called a surface patch or parametrisation of the open subset $S \cap W$ of S . A collection of such surface patches whose images cover the whole of S is called an atlas of S .

A surface in three dimensions is often presented by a defining equation; one can also describe it by parameters, but one requires two of them: varying one parameter can only produce a one dimensional figure, some sort of curve.

Example 1.7. Every plane in $\mathbb{R}^3$ is a surface with an atlas consisting of a single surface patch. In fact, let $\mathbf{a}$ be a point on the plane, and let $\mathbf{p}$ and $\mathbf{q}$ be two unit vectors that are parallel to the plane and perpendicular to each other. If $\mathbf{v}$ is any point of the plane, $\mathbf{v} - \mathbf{a}$ is parallel to the plane, and so

$$\mathbf{v} - \mathbf{a} = u\mathbf{p} + v\mathbf{q}$$

for some scalars u and v . Thus, the desired surface patch is

$$\sigma(u, v) = \mathbf{a} + u\mathbf{p} + v\mathbf{q}$$

and its inverse map is

$$\sigma^{-1}(\mathbf{v}) = ((\mathbf{v} - \mathbf{a}) \cdot \mathbf{p}, (\mathbf{v} - \mathbf{a}) \cdot \mathbf{q}).$$

These formulae make it clear that σ and σ^{-1} are continuous, and hence σ is a homeomorphism.

Example 1.8. A circular cylinder is the set of points of $\mathbb{R}^3$ that are at a fixed distance (the radius of the cylinder) from a fixed straight line (its axis). For example, the circular cylinder of radius 1 and axis the z -axis, which we shall call the unit cylinder, is

$$S = \{(x, y, z) \in \mathbb{R}^3 | x^2 + y^2 = 1\}$$

It can be parametrised as

$$\sigma(u, v) = (\cos u, \sin u, v)$$

Clearly, $\sigma(u, v) \in S$ for all $(u, v) \in \mathbb{R}^2$, and every point of S is of this form. Moreover σ is continuous. But it is not injective, and so is not a homeomorphism because $\sigma(u, v) = \sigma(u + 2\pi, v)$ for all (u, v) . To get an injective map we can restrict u to lie in an interval of length $\leq 2\pi$, say $0 \leq u < 2\pi$. However, although the restriction $\sigma|_V$ to σ where

$$V = \{(u, v) \in \mathbb{R}^2 | 0 \leq u < 2\pi\}$$

is injective, V is not an open subset of $\mathbb{R}^2$ and so $\sigma|_V$ is not a surface patch. The largest open subset of $\mathbb{R}^2$ contained in V is

$$U = \{(u, v) \in \mathbb{R}^2 | 0 < u < 2\pi\}$$

and the restriction $\sigma|_U$ to U is a surface patch. However, $\sigma|_U$ does not cover the whole of S , but only the open subset obtained by removing the line $x = 1, y = 0$ from S .

To get an atlas for S , we therefore need at least one more surface patch. We can take $\sigma|_{\tilde{U}}$, where

$$\tilde{U} = \{(u, v) \in \mathbb{R}^2 | -\pi < u < \pi\}$$

which covers the open subset of S obtained by removing the line $x = -1, y = 0$. Every point of S is in the image of at least one of the surface patches $\sigma|_U, \sigma|_{\tilde{U}}$. So, $\{\sigma|_U, \sigma|_{\tilde{U}}\}$ is an atlas for S , and S is a surface.

1.2 Smooth Surfaces

In Differential Geometry we use calculus to analyse surfaces (and other geometric objects). We must be able to make sense of the statement that a function on a surface is differentiable, for example. For this, we have to consider surfaces with some extra structure.

First, if U is an open subset of $\mathbb{R}^m$, we say that a map $f : U \rightarrow \mathbb{R}^n$ is smooth if each of the n components of f , have continuous partial derivatives of all orders. For example, if $m = 2$ and $n = 3$, and

$$f(u, v) = (f_1(u, v), f_2(u, v), f_3(u, v)),$$

then

$$\frac{\partial f}{\partial u} = \left(\frac{\partial f_1}{\partial u}, \frac{\partial f_2}{\partial u}, \frac{\partial f_3}{\partial u} \right), \quad \frac{\partial f}{\partial v} = \left(\frac{\partial f_1}{\partial v}, \frac{\partial f_2}{\partial v}, \frac{\partial f_3}{\partial v} \right)$$

and similarly for higher derivatives. We often use the following abbreviations:

$$\frac{\partial f}{\partial u} = f_u, \quad \frac{\partial f}{\partial v} = f_v,$$

$$\frac{\partial^2 f}{\partial u^2} = f_{uu}, \quad \frac{\partial^2 f}{\partial u \partial v} = f_{uv}, \quad \frac{\partial^2 f}{\partial v \partial u} = f_{vu}, \quad \frac{\partial^2 f}{\partial v^2} = f_{vv},$$

and so on. From advanced calculus we know that $f_{uv} = f_{vu}$, if f is smooth.

It now makes sense to say that a surface patch $\sigma : U \rightarrow \mathbb{R}^3$ is smooth. We have further definitions.

Definition 1.9. A surface patch $\sigma : U \rightarrow \mathbb{R}^3$ is called regular if it is smooth and the vectors σ_u and σ_v are linearly independent at all points $(u, v) \in U$. Equivalently, σ should be smooth and the vector product $\sigma_u \times \sigma_v$ should be non-zero at every point of U .

Definition 1.10. If S is a surface, an allowable surface patch for S is a regular surface patch $\sigma : U \rightarrow \mathbb{R}^3$ such that σ is a homeomorphism from U to an open subset of S . A **smooth surface** is a surface S such that, for any point $\mathbf{p} \in S$, there is an allowable surface patch σ such that $\mathbf{p} \in \sigma(U)$. A collection $\mathcal{A}$ of allowable surface patches for a surface S such that every point of S is in the image of at least one patch in $\mathcal{A}$ is called an atlas for the smooth surface S .

Example 1.11. The plane we have dealt with earlier is a smooth surface, since

$$\sigma(u, v) = \mathbf{a} + u\mathbf{p} + v\mathbf{q}$$

is clearly smooth and $\sigma_u = \mathbf{p}$ and $\sigma_v = \mathbf{q}$ are linearly independent because $\mathbf{p}$ and $\mathbf{q}$ were chosen to be perpendicular unit vectors.

Example 1.12. The unit cylinder we have previously dealt with is also a smooth surface since

$$\sigma(u, v) = (\cos u, \sin u, v)$$

is clearly smooth and

$$\sigma_u = (-\sin u, \cos u, 0), \sigma_v = (0, 0, 1)$$

are obviously linearly independent for all (u, v) , so $\sigma|_U$ and $\sigma|_{\tilde{U}}$ are regular surface patches.

We have the following important result in this connection:

Theorem 1.13. Let U and $\tilde{U}$ be open subsets of $\mathbb{R}^2$ and let $\sigma : U \rightarrow \mathbb{R}^3$ be a regular surface patch. Let $\Phi : \tilde{U} \rightarrow U$ be a bijective smooth map with smooth inverse map $\Phi^{-1} : U \rightarrow \tilde{U}$. Then, $\tilde{\sigma} = \sigma \circ \Phi : \tilde{U} \rightarrow \mathbb{R}^3$ is a regular surface patch.

Proof. The patch $\tilde{\sigma}$ is smooth because any composite of smooth maps is smooth. As for regularity, let $(u, v) = \Phi(\tilde{u}, \tilde{v})$. By the chain rule,

$$\tilde{\sigma}_{\tilde{u}} = \frac{\partial u}{\partial \tilde{u}} \sigma_u + \frac{\partial v}{\partial \tilde{u}} \sigma_v, \quad \tilde{\sigma}_{\tilde{v}} = \frac{\partial u}{\partial \tilde{v}} \sigma_u + \frac{\partial v}{\partial \tilde{v}} \sigma_v,$$

so,

$$\tilde{\sigma}_{\tilde{u}} \times \tilde{\sigma}_{\tilde{v}} = \left(\frac{\partial u}{\partial \tilde{u}} \frac{\partial v}{\partial \tilde{v}} - \frac{\partial u}{\partial \tilde{v}} \frac{\partial v}{\partial \tilde{u}} \right) \sigma_u \times \sigma_v \quad (1.1)$$

The scalar on the right-hand side of this equation is the determinant of the Jacobian matrix

$$J(\Phi) = \begin{bmatrix} \frac{\partial u}{\partial \tilde{u}} & \frac{\partial v}{\partial \tilde{u}} \\ \frac{\partial u}{\partial \tilde{v}} & \frac{\partial v}{\partial \tilde{v}} \end{bmatrix}$$

of Φ . We recall from calculus that, if ψ and $\tilde{\psi}$ are two smooth maps between open sets of $\mathbb{R}^2$,

$$J(\tilde{\psi} \circ \psi) = J(\tilde{\psi})J(\psi)$$

Taking $\psi = \Phi$ and $\tilde{\psi} = \Phi^{-1}$, we see that $J(\Phi^{-1}) = J(\Phi)^{-1}$. In particular, $J(\Phi)$ is invertible, so its determinant is non-zero and equation (1.1) shows that $\tilde{\sigma}$ is regular. $\square$

If regular surface patches σ and $\tilde{\sigma}$ related as in this theorem, then $\tilde{\sigma}$ is said to be a reparametrisation of σ , and that Φ is a reparametrisation map.

Unit II

Course Structure

- Tangents and derivatives
 - Normal and orientability
-

2 Tangents and Derivatives

Definition 2.1. A tangent vector to a surface S at a point $\mathbf{p} \in S$ is the tangent vector at $\mathbf{p}$ of a curve in S passing through $\mathbf{p}$. The tangent space $T_{\mathbf{p}}S$ of S at $\mathbf{p}$ is the set of all tangent vectors to S at $\mathbf{p}$.

To understand the tangent space $T_{\mathbf{p}}S$, choose a surface patch $\sigma : U \rightarrow \mathbb{R}^3$ of S such that $\mathbf{p}$ is the image of σ , say $\sigma(u_0, v_0) = \mathbf{p}$. If a curve γ lies in S and passes through $\mathbf{p}$ when $t = t_0$, say, there are functions $u(t)$ and $v(t)$ such that

$$\gamma(t) = \sigma(u(t), v(t)) \quad (2.1)$$

for all values of t close to t_0 , and $u(t_0) = u_0$, $v(t_0) = v_0$. The functions u and v are necessarily smooth; conversely, it is obvious that if $t \mapsto (u(t), v(t))$ is smooth, then equation (2.1) defines a curve lying in S .

Theorem 2.2. Let $\sigma : U \rightarrow \mathbb{R}^3$ be a patch of a surface S containing a point $\mathbf{p} \in S$, and let (u, v) be coordinates in U . The tangent space to S at $\mathbf{p}$ is the vector subspace of $\mathbb{R}^3$ spanned by the vectors σ_u and σ_v (the derivatives are evaluated at the point $(u_0, v_0) \in U$ such that $\sigma(u_0, v_0) = \mathbf{p}$).

Proof. Let γ be a smooth curve in S , say

$$\gamma(t) = \sigma(u(t), v(t)).$$

Denoting d/dt be a dot, we have, by the chain rule,

$$\dot{\gamma} = \sigma_u \dot{u} + \sigma_v \dot{v}.$$

Thus, $\dot{\gamma}$ is a linear combination of σ_u and σ_v .

Conversely, any vector in the vector subspace of $\mathbb{R}^3$, spanned by σ_u and σ_v is of the form $\lambda\sigma_u + \mu\sigma_v$ for some scalars λ and μ . Define

$$\gamma(t) = \sigma(u_0 + \lambda t, v_0 + \mu t).$$

Then, γ is a smooth curve in S and at $t = 0$, that is, at the point $\mathbf{p} \in S$, we have

$$\dot{\gamma} = \lambda\sigma_u + \mu\sigma_v$$

This shows that every vector in the span of σ_u and σ_v is the tangent vector at $\mathbf{p}$ of some curve in S . □

Since σ is assumed to be regular, σ_u and σ_v are linearly independent so the tangent space is two-dimensional, and will be called the tangent plane from now on.

As a first application of the tangent plane to a smooth surface, we shall explain what is meant by the derivative of a smooth map between surfaces. Suppose then that $f : S \rightarrow \tilde{S}$ is such a map. The derivative of f at a point $\mathbf{p} \in S$ should measure how the point $f(\mathbf{p}) \in \tilde{S}$ changes when $\mathbf{p}$ moves to a nearby point, say $\mathbf{q}$ of S . If the points $\mathbf{p}$ and $\mathbf{q}$ are very close together, the straight line through them should be nearly tangent to S at $\mathbf{p}$. So we should expect that the derivative of f at $\mathbf{p}$ associates to any tangent vector to S at $\mathbf{p}$ a tangent vector to $\tilde{S}$ at $f(\mathbf{p})$, in other words, the derivative of f at $\mathbf{p}$ should be a map $D_{\mathbf{p}}f : T_{\mathbf{p}}S \rightarrow T_{f(\mathbf{p})}\tilde{S}$.

To give a precise definition of $D_{\mathbf{p}}f$ we let $\mathbf{w} \in T_{\mathbf{p}}S$ be a tangent vector to S at $\mathbf{p}$. By definition, $\mathbf{w}$ is the tangent vector at $\mathbf{p}$ of a curve γ in S passing through $\mathbf{p}$, say $\mathbf{w} = \dot{\gamma}(t_0)$. Then, $\tilde{\gamma} = f \circ \gamma$ is a curve in $\tilde{S}$ passing through $f(\mathbf{p})$ when $t = t_0$, so $\tilde{\mathbf{w}} = \dot{\tilde{\gamma}}(t_0) \in T_{f(\mathbf{p})}\tilde{S}$.

Definition 2.3. With the above notation, the derivative $D_{\mathbf{p}}f$ of f at the point $\mathbf{p} \in S$ is the map $D_{\mathbf{p}}f : T_{\mathbf{p}}S \rightarrow T_{f(\mathbf{p})}\tilde{S}$ such that $D_{\mathbf{p}}f(\mathbf{w}) = \tilde{\mathbf{w}}$ for any tangent vector $\mathbf{w} \in T_{\mathbf{p}}S$.

The first thing we must do now is to show that this definition makes sense, that is, that $D_{\mathbf{p}}f(\mathbf{w})$ depends only on f , $\mathbf{p}$, and $\mathbf{w}$: there are (infinitely) many curves γ with the correct tangent vector $\mathbf{w}$ at $\mathbf{p}$ and a priori $D_{\mathbf{p}}f(\mathbf{w})$ could depend on which curve is chosen.

Let $\sigma : U \rightarrow \mathbb{R}^3$ be a surface patch of S containing $\mathbf{p}$, say $\mathbf{p} = \sigma(u_0, v_0)$, and let α, β be the smooth functions on U such that

$$f(\sigma(u, v)) = \tilde{\sigma}(\alpha(u, v), \beta(u, v)).$$

Let $\mathbf{w} = \lambda\sigma_u + \mu\sigma_v$ be the tangent vector at $\mathbf{p}$ of a curve $\gamma(t) = \sigma(u(t), v(t))$, where u and v are smooth functions such that $\dot{u}(t_0) = \lambda, \dot{v}(t_0) = \mu$. Since the corresponding curve on $\tilde{S}$ is $\tilde{\gamma}(t) = \tilde{\sigma}(\tilde{u}(t), \tilde{v}(t))$, where $\tilde{u}(t) = \alpha(u(t), v(t))$ and $\tilde{v}(t) = \beta(u(t), v(t))$, we have

$$\begin{aligned} D_{\mathbf{p}}f(\mathbf{w}) &= \dot{\tilde{u}}\tilde{\sigma}_{\tilde{u}} + \dot{\tilde{v}}\tilde{\sigma}_{\tilde{v}} \\ &= (\dot{u}\alpha_u + \dot{v}\alpha_v)\tilde{\sigma}_{\tilde{u}} + (\dot{u}\beta_u + \dot{v}\beta_v)\tilde{\sigma}_{\tilde{v}}, \end{aligned}$$

the derivative of u and v being evaluated at t_0 . Thus,

$$D_{\mathbf{p}}f(\mathbf{w}) = (\lambda\alpha_u + \mu\alpha_v)\tilde{\sigma}_{\tilde{u}} + (\lambda\beta_u + \mu\beta_v)\tilde{\sigma}_{\tilde{v}}. \quad (2.2)$$

The RHS depends only on $\mathbf{p}$, f , λ and μ , that is, on $\mathbf{p}$, f and $\mathbf{w}$ as desired. The equation (2.2) also establishes the following theorem

Theorem 2.4. *If $f : S \rightarrow \tilde{S}$ is a smooth map between surfaces and $\mathbf{p} \in S$, the derivative $D_{\mathbf{p}}f : T_{\mathbf{p}}S \rightarrow T_{f(\mathbf{p})}\tilde{S}$ is a linear map.*

Theorem 2.5. *1. If S is a surface and $\mathbf{p} \in S$, the derivative at $\mathbf{p}$ of the identity map $S \rightarrow S$ is the identity map $T_{\mathbf{p}}S \rightarrow T_{\mathbf{p}}S$.*

2. If S_1, S_2, S_3 are surfaces and $f_1 : S_1 \rightarrow S_2$ and $f_2 : S_2 \rightarrow S_3$ are smooth maps, then for all $\mathbf{p} \in S_1$,

$$D_{\mathbf{p}}(f_2 \circ f_1) = D_{f_1(\mathbf{p})}f_2 \circ D_{\mathbf{p}}f_1.$$

3. If $f : S_1 \rightarrow S_2$ is a diffeomorphism, then for all $\mathbf{p} \in S_1$, the linear map $D_{\mathbf{p}}f : T_{\mathbf{p}}S_1 \rightarrow T_{f(\mathbf{p})}S_2$ is invertible.

Proof. Part 1 is obvious. For 2, let $\mathbf{w} \in T_{\mathbf{p}}S_1$ be the tangent vector at $\mathbf{p}$ of a curve γ_1 on S_1 . Then $\gamma_2 = f_1 \circ \gamma_1$ is a curve on S_2 with tangent vector $D_{\mathbf{p}}f_1(\mathbf{w})$ at $f_1(\mathbf{p})$, so $\gamma_3 = f_2 \circ \gamma_2 = (f_2 \circ f_1) \circ \gamma_1$ is a curve on S_3 with tangent vector $D_{f_1(\mathbf{p})}f_2(D_{\mathbf{p}}f_1(\mathbf{w}))$ at $f_2(f_1(\mathbf{p}))$. But the tangent vector of γ_3 at $\mathbf{p}$ is also $D_{\mathbf{p}}(f_2 \circ f_1)(\mathbf{w})$.

Finally, for 3, let $g : S_3 \rightarrow S_1$ be the inverse map of f , so that $g \circ f$ and $f \circ g$ are the identity maps $S_1 \rightarrow S_1$ and $S_2 \rightarrow S_2$, respectively. Parts 1 and 2 show that $D_{f(\mathbf{p})}g$ is the inverse of the linear map $D_{\mathbf{p}}f$. $\square$

2.1 Normals and Orientability

Since the tangent plane $T_{\mathbf{p}}S$ of a surface S at point $\mathbf{p} \in S$ passes through the origin of $\mathbb{R}^3$, it is completely determined by giving a unit vector perpendicular to it, called a unit normal to S at $\mathbf{p}$. There are, of course, two such vectors, but by a previous theorem, choosing a surface patch $\sigma : U \rightarrow \mathbb{R}^3$ containing $\mathbf{p}$ leads to a definite choice, namely

$$N_{\sigma} = \frac{\sigma_u \times \sigma_v}{\|\sigma_u \times \sigma_v\|}$$

(with the derivatives evaluated at the point of U corresponding to $\mathbf{p}$), for this is clearly a unit vector perpendicular to every linear combination of σ_u and σ_v . This is called the standard unit normal of the surface patch σ at $\mathbf{p}$. Unlike the tangent plane, however, N_{σ} is not quite independent of the choice of patch σ containing $\mathbf{p}$. In fact, if $\tilde{\sigma} : \tilde{U} \rightarrow \mathbb{R}^3$ is another surface patch in the atlas of S containing $\mathbf{p}$ that

$$\tilde{\sigma}_{\tilde{u}} \times \tilde{\sigma}_{\tilde{v}} = \det(J(\Phi))\sigma_u \times \sigma_v$$

where $J(\Phi)$ is the Jacobian matrix of the transition map Φ from σ to $\tilde{\sigma}$. So the standard unit normal of $\tilde{\sigma}$ is

$$\begin{aligned} N_{\tilde{\sigma}} &= \frac{\tilde{\sigma}_{\tilde{u}} \times \tilde{\sigma}_{\tilde{v}}}{\|\tilde{\sigma}_{\tilde{u}} \times \tilde{\sigma}_{\tilde{v}}\|} \\ &= \pm \frac{\sigma_u \times \sigma_v}{\|\sigma_u \times \sigma_v\|} \\ &= \pm N_{\sigma}, \end{aligned}$$

where the sign is that of the determinant of $J(\Phi)$. This leads to the following definition.

Definition 2.6. A surface S is said to be orientable if there exists an atlas $\mathcal{A}$ for S with the property that, if Φ is the transition map between any two surface patches in $\mathcal{A}$, then $\det(J(\Phi)) > 0$, where Φ is defined.

The preceding discussion gives the following theorem.

Theorem 2.7. Let S be an orientable surface equipped with an atlas $\mathcal{A}$ as in the above definition. Then, there is a smooth choice of unit normal at any point of S ; take the standard unit normal of any surface patch in $\mathcal{A}$.

An oriented surface is a surface S together with a smooth choice of unit normal N at each point, i.e., a smooth map $N : S \rightarrow \mathbb{R}^3$ (meaning that each of the three components of N is a smooth function $S \rightarrow \mathbb{R}$) such that, for all $\mathbf{p} \in S$, $N(\mathbf{p})$ is a unit vector perpendicular to $T_{\mathbf{p}}S$. Any oriented surface is orientable. To see this, start with the maximal atlas of S and retain a patch $\sigma(u, v)$ if $\sigma_u \times \sigma_v$ is a positive multiple of N at all points in the image of σ , otherwise discard it. The patches that remain form an atlas $\mathcal{A}$ satisfying the condition in the previous definition.

Example 2.8. The Möbius band is the surface obtained by rotating a straight line segment l around its midpoint $\mathbf{p}$ at the same time as $\mathbf{p}$ moves around a circle $\mathcal{C}$, in such a way as $\mathbf{p}$ moves once around $\mathcal{C}$, l makes a half-turn about $\mathbf{p}$. If we take $\mathcal{C}$ to be the circle $x^2 + y^2 = 1$ in the xy -plane, and l to be a segment of length 1 that is initially parallel to the z -axis with its midpoint $\mathbf{p}$ at $(1, 0, 0)$, and then after $\mathbf{p}$ has rotated by an angle θ around the z -axis, l should have rotated by $\theta/2$ around $\mathbf{p}$ in the plane containing $\mathbf{p}$ and the z -axis. The point of l initially at $(1, 0, t)$ is then at the point

$$\sigma(t, \theta) = \left(\left(1 - t \sin \frac{\theta}{2}\right) \cos \theta, \left(1 - t \sin \frac{\theta}{2}\right) \sin \theta, t \cos \frac{\theta}{2} \right).$$

We take the domain of definition of σ to be

$$U = \{(t, \theta) \in \mathbb{R}^2 \mid -1/2 < t < 1/2, 0 < \theta < 2\pi\}$$

We can define a second patch $\tilde{\sigma}$ by the same formula as σ but with domain of definition

$$\tilde{U} = \{(t, \theta) \in \mathbb{R}^2 \mid -1/2 < t < 1/2, -\pi < \theta < \pi\}$$

It can be checked that these two patches form an atlas for the Möbius band consisting of regular surface patches, making the Möbius band into a smooth surface S .

We compute the standard unit normal N_σ at points on the median circle (where $t = 0$). At such points, we have

$$\sigma_t = \left(-\sin \frac{\theta}{2} \cos \theta, -\sin \frac{\theta}{2} \sin \theta, \cos \frac{\theta}{2} \right), \quad \sigma_\theta = (\sin \theta, \cos \theta, 0),$$

So,

$$\sigma_t \times \sigma_\theta = \left(-\cos \theta \cos \frac{\theta}{2}, -\sin \theta \cos \frac{\theta}{2}, -\sin \frac{\theta}{2} \right).$$

This is a unit vector, so it is equal to N_σ .

If the Möbius band was orientable, there would be a well-defined unit normal N defined at every point of S and varying smoothly over S . At a point $\sigma(0, \theta)$ on the median circle, we would have

$$N = \lambda(\theta)N_\sigma,$$

where $\lambda : (0, 2\pi) \rightarrow \mathbb{R}$ is smooth and $\lambda(\theta) = \pm 1$ for all θ . It follows that either $\lambda(\theta) = +1$ for all $\theta \in (0, 2\pi)$, or $\lambda(\theta) = -1$ for all $\theta \in (0, 2\pi)$. Replacing N by $-N$ if necessary, we can assume that $\lambda = 1$. At the point $\sigma(0, 0) = \sigma(0, 2\pi)$, we would have (since N is smooth)

$$N = \lim_{\theta \downarrow 0} N_\sigma = (-1, 0, 0) \text{ and also } N = \lim_{\theta \uparrow 2\pi} N_\sigma = (1, 0, 0).$$

This contradiction shows that the Möbius band is not orientable.

If a surface S is oriented, it is possible to give a sign to the angle between two tangent vectors at a point of S .

Let $\mathbf{p} \in S$ and let N be the chosen unit normal at $\mathbf{p}$. A rotation in the tangent plane $T_{\mathbf{p}}S$ is said to be in the positive sense, or anticlockwise, if rotation in this sense of a right-handed screw held perpendicular to $T_{\mathbf{p}}S$ would cause it to advance in the direction of N . Put another way, the choice of N enables us to distinguish the two sides of $T_{\mathbf{p}}S$: the positive side is the half-space into which N points. Then, if we view $T_{\mathbf{p}}S$ from a point on the positive side, a positive rotation in $T_{\mathbf{p}}S$ would be seen as anticlockwise in the usual sense.

If $\mathbf{v}$ and $\mathbf{w}$ are non-zero vectors in $T_{\mathbf{p}}S$, the oriented angle (which we shall sometimes just call the angle) between $\mathbf{v}$ and $\mathbf{w}$ is the angle through which $\mathbf{v}$ must be rotated in the positive sense in order for the resulting vector to be a positive scalar multiple of $\mathbf{w}$. We shall denote this angle by $\hat{\mathbf{v}\mathbf{w}}$. Note that

$$\hat{\mathbf{w}\mathbf{v}} = -\hat{\mathbf{v}\mathbf{w}},$$

and that the sign of $\hat{\mathbf{v}\mathbf{w}}$ will change if we change the choice of unit normal to $T_{\mathbf{p}}S$. Note also that $\hat{\mathbf{v}\mathbf{w}}$ is determined only up to the addition of an integer multiple of 2π .

2.2 Examples of Surfaces

2.2.1 Level Surfaces

Surfaces are often given to us as level surfaces

$$\{(x, y, z) \in \mathbb{R}^3 \mid f(x, y, z) = 0\}$$

where f is a smooth function. The following result gives general conditions under which a level surface is a smooth surface.

Theorem 2.9. *Let S be a subset of $\mathbb{R}^3$ with the following property: for each point $\mathbf{p} \in S$, there is an open subset W of $\mathbb{R}^3$ containing $\mathbf{p}$ and a smooth function $f : W \rightarrow \mathbb{R}$ such that*

1. $S \cap W = \{(x, y, z) \in W \mid f(x, y, z) = 0\}$
2. $\nabla f = (f_x, f_y, f_z)$ of f does not vanish at $\mathbf{p}$.

Unit III

Course Structure

- The first Fundamental form
 - Length of curves on surfaces
 - Isometries of surfaces
 - Conformal mapping of surfaces
-

Introduction

In our Euclidean plane we use the quadratic form $ds^2 = dx^2 + dy^2$ for the purpose of measurement. This formula is valid in $\mathbb{R}^2$ with rectangular cartesian coordinate system. The expression $ds = \sqrt{dx^2 + dy^2}$ is called Euclidean metric. Observe that $ds^2 = dx^2 + dy^2$ may be expressed as

$$ds^2 = \begin{bmatrix} dx & dy \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} dx \\ dy \end{bmatrix}$$

The matrix in the above equation is called the Euclidean matrix.

But, if our coordinate system is not rectangular, then the metric is not Euclidean. On a surface we can't define rectangular cartesian coordinate system. The coordinate system on surface is a type of curvilinear coordinate system.

In this chapter we shall deduce a quadratic form on a surface which is known as the First Fundamental form and is used to measure lengths and angles on surfaces.

2.3 Lengths of curves on surfaces

If $\gamma(t) = \sigma(u(t), v(t))$ is a curve in a surface patch σ , its arc length starting at a point $\gamma(t_0)$ is given by

$$s = \int_{t_0}^t \|\dot{\gamma}(u)\| du$$

By chain rule,

$$\dot{\gamma}^2 = \sigma_u \dot{u} + \sigma_v \dot{v}$$

Hence

$$\begin{aligned} \|\dot{\gamma}\|^2 &= (\sigma_u \dot{u} + \sigma_v \dot{v})(\sigma_u \dot{u} + \sigma_v \dot{v}) \\ &= (\sigma_u \sigma_u) \dot{u}^2 + 2(\sigma_u \sigma_v) \dot{u} \dot{v} + (\sigma_v \sigma_v) \dot{v}^2 \\ &= E \dot{u}^2 + 2F \dot{u} \dot{v} + G \dot{v}^2 \end{aligned}$$

where $E = \|\sigma_u\|^2$, $F = \sigma_u \sigma_v$, $G = \|\sigma_v\|^2$. So

$$s = \int_{t_0}^t (E\dot{u}^2 + 2F\dot{u}\dot{v} + G\dot{v}^2)^{1/2} dt$$

Comparing the above equation with

$$s = \int_{t_0}^t (\sqrt{ds})^2$$

We have,

$$ds^2 = Edu^2 + 2Fdudv + Gdv^2$$

This is called the First Fundamental Quadratic form on surface.

2.4 First Fundamental form on Sphere

For the sphere in latitude longitude coordinates

$$\sigma(\theta, \phi) = (\cos \theta \cos \phi, \cos \theta \sin \phi, \sin \theta),$$

We have

$$\begin{aligned}\sigma_\theta &= (-\sin \theta \cos \phi, -\sin \theta \sin \phi, \cos \theta) \\ \sigma_\phi &= (-\cos \theta \sin \phi, \cos \theta \cos \phi, 0)\end{aligned}$$

Hence, $E = \|\sigma_\theta\|^2 = 1$, $F = \sigma_\theta \sigma_\phi = 0$, $G = \|\sigma_\phi\|^2 = \cos^2 \theta$. So the First Fundamental form is $d\theta^2 + \cos^2 \theta d\phi^2$.

Exercise 2.10. 1. Find the First Fundamental form on cylinder.

2. Calculate the First Fundamental form on $\sigma(u, v) = (u - v, u + v, u^2 + v^2)$.

3. Show that applying rigid motion to a surface does not change its First Fundamental form.

Definition 2.11. If S_1 and S_2 are surfaces, a diffeomorphism $f : S_1 \rightarrow S_2$ is called an isometry if it takes curves in S_1 to curves of same length in S_2 . If an isometry $f : S_1 \rightarrow S_2$ exists, then we say that S_1 and S_2 are isomorphic.

Theorem 2.12. A diffeomorphism $f : S_1 \rightarrow S_2$ is an isometry if and only if, for any surface patch σ_1 of S_1 , the patches σ_1 and $f \circ \sigma_1$ of S_1 and S_2 respectively, have the same first fundamental form.

Proof. Suppose the length of any curve can be computed as the sum of the lengths of curves each lying in a single surface patch, we can assume that S_1 and S_2 are covered by single surface patches. Moreover, since f is a diffeomorphism, we can assume that these patches are of the form $\sigma_1 : U \rightarrow \mathbb{R}^3$ (for S_1) and $f \circ \sigma_1 = \sigma_2$ (for S_2). We have to show that f is an isometry if and only if σ_1 and σ_2 have the same first fundamental form.

Suppose first that σ_1 and σ_2 have same first fundamental form. If $t \rightarrow (u(t), v(t))$ is any curve in U , say $\gamma_1(t) = \sigma_1(u(t), v(t))$ and $\gamma_2(t) = \sigma_2(u(t), v(t))$ are the corresponding curves in S_1 and S_2 , then f takes γ_1 to γ_2 , since

$$\begin{aligned} f_1(\gamma_1(t)) &= f(\sigma_1(u(t)), \sigma_1(v(t))) \\ &= \sigma_2(u(t), v(t)) \\ &= \gamma_2(t) \end{aligned}$$

It is clear that γ_1 and γ_2 have the same length, since both lengths are found by integrating the expression $(E\dot{u}^2 + 2F\dot{u}\dot{v} + G\dot{v}^2)^{1/2}$, where $Edu^2 + 2Fdu dv + Gdv^2$ is the common first fundamental form of S_1 and S_2 .

Conversely suppose that f is an isometry. If $f \mapsto (u(t), v(t))$ is any curve in U defined for $t \rightarrow (\alpha, \beta)$, say, the curve $\gamma_1(t) = \sigma_1(u(t), v(t))$ and $\gamma_2(t) = (\sigma_2(u(t)), \sigma_2(v(t)))$ have the same length. Hence

$$\int_{t_0}^{t_1} (E_1\dot{u}^2 + 2F_1\dot{u}\dot{v} + G_1\dot{v}^2)^{1/2} dt = \int_0^{t_0} (E_2\dot{u}^2 + 2F_2\dot{u}\dot{v} + G_2\dot{v}^2)^{1/2} dt$$

for all $t_0, t_1 \in (\alpha, \beta)$, where E_1, F_1 and G_1 are the coefficients of the first fundamental form of σ_1 and E_2, F_2 and G_2 are those of σ_2 . This implies that the two integrands are the same and hence that

$$E_1\dot{u}^2 + 2F_1\dot{u}\dot{v} + G_1\dot{v}^2 = E_2\dot{u}^2 + 2F_2\dot{u}\dot{v} + G_2\dot{v}^2$$

Fix $t_0 \in (\alpha, \beta)$ and let $u_0 = u(t_0), v_0 = v(t_0)$. We now apply the above equation for the following three choices of the curve $t \rightarrow (u(t), v(t))$ in U :

- (i) $u = u_0 + t - t_0; v = v_0 \implies E_1 = E_2$
- (ii) $v = v_0 + t - t_0; u = u_0 \implies G_1 = G_2$
- (iii) $u = u_0 + t - t_0; v = v_0 + t - t_0$. This gives $E_1 + 2F_1 + G_1 = E_2 + 2F_2 + G_2$ and hence $F_1 = F_2$.

□

2.5 Conformal Mappings

A mapping from a surface σ_1 to another surface σ_2 is called conformal if angle between two intersecting curves is preserved under the mapping.

2.5.1 Angle between two curves on a surface

Suppose that two curves γ and $\tilde{\gamma}$ on a surface S intersect at a point $\mathbf{p}$ that lies in a surface patch σ on S . Then $\gamma(t) = \sigma(u(t), v(t))$ and $\tilde{\gamma}(t) = \sigma(\tilde{u}(t), \tilde{v}(t))$ for some smooth functions $u, v, \tilde{u}, \tilde{v}$ and for some parameter values t_0 and $\tilde{t}_0$, we have $\sigma(u(t_0), v(t_0)) = P = \sigma(\tilde{u}(t_0), \tilde{v}(t_0))$.

The angle θ between the curves γ and $\tilde{\gamma}$ is given by

$$\cos \theta = \frac{\dot{\gamma} \cdot \dot{\tilde{\gamma}}}{\|\dot{\gamma}\| \cdot \|\dot{\tilde{\gamma}}\|}$$

By chain rule

$$\begin{aligned}\dot{\gamma} &= \sigma_u \dot{u} + \sigma_v \dot{v} \\ \dot{\tilde{\gamma}} &= \sigma_u \dot{\tilde{u}} + \sigma_v \dot{\tilde{v}}\end{aligned}$$

So

$$\begin{aligned}\dot{\gamma} \cdot \dot{\tilde{\gamma}} &= (\sigma_u \dot{u} + \sigma_v \dot{v})(\sigma_u \dot{\tilde{u}} + \sigma_v \dot{\tilde{v}}) \\ &= (\sigma_u \sigma_u) \dot{u} \dot{\tilde{u}} + (\sigma_u \sigma_v)(\dot{u} \dot{\tilde{v}} + \dot{\tilde{u}} \dot{v}) + (\sigma_v \sigma_v) \dot{v} \dot{\tilde{v}} \\ &= E \dot{u} \dot{\tilde{u}} + F(\dot{u} \dot{\tilde{v}} + \dot{\tilde{u}} \dot{v}) + G \dot{v} \dot{\tilde{v}}\end{aligned}$$

Similarly, $||\dot{\gamma}|| = \sqrt{\dot{\gamma} \cdot \dot{\gamma}}$ and $||\dot{\tilde{\gamma}}|| = \sqrt{\dot{\tilde{\gamma}} \cdot \dot{\tilde{\gamma}}}$ can be evaluated and finally

$$\cos \theta = \frac{E \dot{u} \dot{\tilde{u}} + F(\dot{u} \dot{\tilde{v}} + \dot{\tilde{u}} \dot{v}) + G \dot{v} \dot{\tilde{v}}}{(E \dot{u}^2 + 2F \dot{u} \dot{v} + G \dot{v}^2)^{1/2} (E \dot{\tilde{u}}^2 + 2F \dot{\tilde{u}} \dot{\tilde{v}} + G \dot{\tilde{v}}^2)^{1/2}}$$

Theorem 2.13. *The parameter curves on a surface patch $\sigma(u, v)$ can be parametrized by $\gamma(t) = \sigma(a, t)$, $\tilde{\gamma} = \sigma(t, b)$ receptively, where a is the constant value of u and b the constant value of v in the two cases. Thus $u(t) = a$, $v(t) = t$, $\tilde{u}(t) = t$, $\tilde{v}(t) = b$, $\dot{u} = 0$, $\dot{v} = 1$, $\dot{\tilde{u}} = 1$, $\dot{\tilde{v}} = 0$. These parameter curves intersect at the point $\sigma(a, b)$ of the surface and their angle of intersection θ is given by*

$$\cos \theta = \frac{F}{\sqrt{EG}}$$

In particular, the parameter curves are orthogonal if and only if $F = 0$.

Definition 2.14. If S_1 and S_2 are surfaces, a diffeomorphism $f : S_1 \rightarrow S_2$ is said to be conformal if, whenever f takes two intersecting curves γ_1 and $\tilde{\gamma}_1$ on S_1 to curves γ_2 and $\tilde{\gamma}_2$ on S_2 , the angle of intersection of γ_1 and $\tilde{\gamma}_1$ is equal to the angle of intersection of γ_2 and $\tilde{\gamma}_2$.

Theorem 2.15. *A diffeomorphism $f : S_1 \rightarrow S_2$ is conformal if and only if, for any surface patch σ_1 on S_1 , the first fundamental forms of σ_1 and $f \circ \sigma_1$ are proportional.*

Proof. Assume that S_1 and S_2 are covered by the single surface patches $\sigma_1 : U \rightarrow \mathbb{R}^3$ and $\sigma_2 = f \circ \sigma_1$, respectively. Suppose that their first fundamental forms $E_1 du^2 + 2F_1 dudv + G_1 dv^2$ and $E_2 du^2 + 2F_2 dudv + G_2 dv^2$ are proportional, say

$$E_1 du^2 + 2F_1 dudv + G_1 dv^2 = \lambda(E_2 du^2 + 2F_2 dudv + G_2 dv^2)$$

for some smooth function $\lambda(u, v)$, where (u, v) are coordinates on U . Note that $\lambda > 0$ everywhere, since (for example) E_1 and E_2 are both greater than 0. If $\gamma(t) = \sigma_1(u(t), v(t))$ and $\tilde{\gamma}(t) = \sigma_1(\tilde{u}(t), \tilde{v}(t))$ are curves in S_1 , then f takes γ and $\tilde{\gamma}$ to the curves $\sigma_2(u(t), v(t))$ and $\sigma_2(\tilde{u}(t), \tilde{v}(t))$ in S_2 , respectively. Now,

$$\begin{aligned}\cos \theta &= \frac{E_2 \dot{u} \dot{\tilde{u}} + F_2(\dot{u} \dot{\tilde{v}} + \dot{\tilde{u}} \dot{v}) + G_2 \dot{v} \dot{\tilde{v}}}{(E_2 \dot{u}^2 + 2F_2 \dot{u} \dot{v} + G_2 \dot{v}^2)^{1/2} (E_2 \dot{\tilde{u}}^2 + 2F_2 \dot{\tilde{u}} \dot{\tilde{v}} + G_2 \dot{\tilde{v}}^2)^{1/2}} \\ &= \frac{\lambda E_1 \dot{u} \dot{\tilde{u}} + \lambda F_1(\dot{u} \dot{\tilde{v}} + \dot{\tilde{u}} \dot{v}) + \lambda G_1 \dot{v} \dot{\tilde{v}}}{(\lambda E_1 \dot{u}^2 + 2\lambda F_1 \dot{u} \dot{v} + \lambda G_1 \dot{v}^2)^{1/2} (\lambda E_1 \dot{\tilde{u}}^2 + 2\lambda F_1 \dot{\tilde{u}} \dot{\tilde{v}} + \lambda G_1 \dot{\tilde{v}}^2)^{1/2}} \\ &= \frac{E_1 \dot{u} \dot{\tilde{u}} + F_1(\dot{u} \dot{\tilde{v}} + \dot{\tilde{u}} \dot{v}) + G_1 \dot{v} \dot{\tilde{v}}}{(E_1 \dot{u}^2 + 2F_1 \dot{u} \dot{v} + G_1 \dot{v}^2)^{1/2} (E_1 \dot{\tilde{u}}^2 + 2F_1 \dot{\tilde{u}} \dot{\tilde{v}} + G_1 \dot{\tilde{v}}^2)^{1/2}}\end{aligned}$$

Hence f is conformal.

For the converse, we must show that if

$$\frac{E_1 \dot{u} \ddot{u} + F_1(\dot{u} \ddot{v} + \ddot{u} \dot{v}) + G_1 \dot{v} \ddot{v}}{(E_1 \dot{u}^2 + 2F_1 \dot{u} \dot{v} + G_1 \dot{v}^2)^{1/2} (E_1 \ddot{u}^2 + 2F_1 \ddot{u} \ddot{v} + G_1 \ddot{v}^2)^{1/2}} = \frac{E_2 \dot{u} \ddot{u} + F_2(\dot{u} \ddot{v} + \ddot{u} \dot{v}) + G_2 \dot{v} \ddot{v}}{(E_2 \dot{u}^2 + 2F_2 \dot{u} \dot{v} + G_2 \dot{v}^2)^{1/2} (E_2 \ddot{u}^2 + 2F_2 \ddot{u} \ddot{v} + G_2 \ddot{v}^2)^{1/2}} \quad (2.3)$$

for all pairs of intersecting curves

$$\begin{aligned} \gamma(t) &= \sigma_1(u(t), v(t)) \\ \text{and } \tilde{\gamma}(t) &= \sigma_1(\tilde{u}(t), \tilde{v}(t)) \end{aligned}$$

in S_1 , then the first fundamental form of σ_1 and σ_2 are proportional. Fix $(a, b) \in U$ and consider the curves

$$\begin{aligned} \gamma(t) &= \sigma_1(a + t, b) \\ \tilde{\gamma}(t) &= \sigma_1(a + t \cos \phi, b + t \sin \phi) \end{aligned}$$

where ϕ is a constant, for which

$$\dot{u} = 1, \dot{v} = 0, \ddot{u} = \cos \phi, \ddot{v} = \sin \phi$$

Using these in (2.3), we get

$$\frac{E_1 \cos \phi + F_1 \sin \phi}{\sqrt{E_1(E_1 \cos^2 \phi + 2F_1 \sin \phi \cos \phi + G_1 \sin^2 \phi)}} = \frac{E_2 \cos \phi + F_2 \sin \phi}{\sqrt{E_2(E_2 \cos^2 \phi + 2F_2 \sin \phi \cos \phi + G_2 \sin^2 \phi)}} \quad (2.4)$$

Squaring both sides of equation (2.4) and writing

$$(E_1 \cos \phi + F_1 \sin \phi)^2 = E_1(E_1 \cos^2 \phi + 2F_1 \sin \phi \cos \phi + G_1 \sin^2 \phi) - (E_1 G_1 - F_1^2) \sin^2 \phi,$$

we get

$$(E_1 G_1 - F_1^2) E_2 ((E_2 \cos^2 \phi + 2F_2 \sin \phi \cos \phi + G_2 \sin^2 \phi) = (E_2 G_2 - F_2^2) E_1 (E_1 \cos^2 \phi + 2F_1 \sin \phi \cos \phi + G_1 \sin^2 \phi)$$

or setting

$$\lambda = \frac{(E_2 G_2 - F_2^2) E_1}{(E_1 G_1 - F_1^2) E_2},$$

we get

$$(E_2 - \lambda E_1) \cos^2 \phi + (2F_2 - \lambda F_1) \sin \phi \cos \phi + (G_2 - \lambda G_1) \sin^2 \phi = 0$$

Taking $\phi = 0$ and $\phi = \pi/2$ gives $E_2 = \lambda E_1$, $G_2 = \lambda G_1$. Then by the last equation, $F_2 = \lambda F_1$. $\square$

Exercise 2.16. 1. Calculate the first fundamental form of the surface

$$\sigma(u, v) = (\sinh u \sinh v, \sinh u \cosh v, \sinh u).$$

2. Show that applying a rigid motion to a surface does not change its first fundamental form.
3. Show that every isometry is a conformal map. Give an example of a conformal map that is not an isometry.

Summary

In this unit, we have determined the expression of first fundamental form of surfaces. We have calculated first fundamental form of sphere. We have also discussed about conformal maps.

Unit IV

Course Structure

- Introduction
 - Curvature of curves on surfaces
 - Second Fundamental form
-

Introduction

To introduce the notion of curvature on surfaces, we start by finding a new interpretation of the curvature of a plane curve. Suppose that γ is a unit speed curve in $\mathbb{R}^2$. As the parameter t of γ changes to $t + \Delta t$, the curve moves away from its tangent line at $\gamma(t)$ by a distance $(\gamma(t + \Delta t) - \gamma(t)) \cdot \eta$, where η is the principal normal to γ at $\gamma(t)$. By Taylor's theorem,

$$\gamma(t + \Delta t) = \gamma(t) + \dot{\gamma}(t)\Delta t + \frac{1}{2}\ddot{\gamma}(t)(\Delta t)^2 + \text{remainder},$$

where $\text{remainder}/(\Delta t)^2$ tends to 0 as $\Delta t \rightarrow 0$. Now, η is perpendicular to the unit tangent vector $t = \dot{\gamma}$ and $\ddot{\gamma} = \dot{t} = \kappa\eta$, where κ is the curvature of γ . Hence, $\ddot{\gamma} \cdot \eta = \kappa$ and the derivative of γ forms its tangent line is

$$(\dot{\gamma}(t)\Delta t + \frac{1}{2}\ddot{\gamma}(t)(\Delta t)^2 + \cdots) \cdot \eta = \frac{1}{2}\kappa(\Delta t)^2 + \text{remainder} \quad (2.5)$$

Now let σ be a surface patch in $\mathbb{R}^3$ with standard unit normal N (surface normal). As the parameters (u, v) of σ changes to $(u + \Delta u, v + \Delta v)$, the surface moves away from its tangent plane at $\sigma(u, v)$ by a distance $(\sigma(u + \Delta u, v + \Delta v) - \sigma(u, v)) \cdot N$. By the two variable form of Taylor's theorem,

$$\sigma(u + \Delta u, v + \Delta v) - \sigma(u, v) = \sigma_u \Delta u + \sigma_v \Delta v + \frac{1}{2}(\sigma_{uu}(\Delta u)^2 + 2\sigma_{uv}\Delta u\Delta v + \sigma_{vv}(\Delta v)^2) + \text{remainder},$$

where $(\text{remainder})/[(\Delta u)^2 + (\Delta v)^2]$ tends to zero as $(\Delta u)^2 + (\Delta v)^2 \rightarrow 0$. Now, σ_u and σ_v are tangents to the surface, hence perpendicular to N , so the derivation of σ from its tangent plane is

$$\frac{1}{2}(\mathcal{L}(\Delta u)^2 + 2\mathcal{M}\Delta u\Delta v + \mathcal{N}(\Delta v)^2) + \text{remainder} \quad (2.6)$$

where, $\mathcal{L} = \sigma_{uu} \cdot N$, $\mathcal{M} = \sigma_{uv} \cdot N$, $\mathcal{N} = \sigma_{vv} \cdot N$. Comparing equation (2.6) with equation (2.5), we see that the expression is the analogue for the surface of the curvature term $\kappa(\Delta t)^2$ in the case of a curve. One calls the expression

$$\mathcal{L}du^2 + 2\mathcal{M}dudv + \mathcal{N}dv^2 \quad (2.7)$$

is called the second fundamental form of σ . As in the case of the first fundamental form we regard the expression (2.7) simply as a convenient way of keeping track of the three functions $\mathcal{L}$, $\mathcal{M}$ and $\mathcal{N}$. We shall soon see that a knowledge of these functions (together with that of first fundamental form) will enable us to compute the curvature of any curve on the surface.

Note 2.17. $\mathcal{L} = \sigma_{uu} \cdot N$, $\mathcal{M} = \sigma_{uv} \cdot N$, $\mathcal{N} = \sigma_{vv} \cdot N$, where

$$N = \frac{\sigma_u \times \sigma_v}{\|\sigma_u \times \sigma_v\|}.$$

Exercise 2.18. Calculate the second fundamental form of unit sphere and right circular cylinder.

2.6 Normal and geodesic curvature of a curve on a surface

If $\gamma(t) = \sigma(u(t), v(t))$ is a unit speed curve in a surface patch γ , then $\dot{\gamma}$ is a unit vector, and is by definition, a tangent vector to σ . Hence $\dot{\gamma}$ is perpendicular to the standard unit normal N of σ , so $\dot{\gamma}$, N and $N \times \dot{\gamma}$ are mutually perpendicular unit vectors. Again, since γ is unit speed, $\ddot{\gamma}$ is perpendicular to $\dot{\gamma}$ and hence is a linear combination of N and $N \times \dot{\gamma}$. So,

$$\ddot{\gamma} = \kappa_n N + \kappa_g N \times \dot{\gamma} \quad (2.8)$$

The scalars κ_n and κ_g are called the normal curvature and the geodesic curvature of γ respectively. Since N and $N \times \dot{\gamma}$ are perpendicular unit vectors, equation (2.8) implies $\kappa_n = \ddot{\gamma} \cdot N$, $\kappa_g = \ddot{\gamma} \cdot (N \times \dot{\gamma})$ and,

$$\|\ddot{\gamma}\|^2 = \kappa_n^2 + \kappa_g^2 \quad (2.9)$$

Hence the curvature $\kappa = \|\ddot{\gamma}\|$ of γ is given by

$$\kappa^2 = \kappa_n^2 + \kappa_g^2$$

Moreover, if η is principal normal of γ , so that $\ddot{\gamma} = \kappa\eta$, we have

$$\kappa_\eta = \kappa\eta \cdot N = \kappa \cos \psi$$

where ψ is the angle between η and N . Then from equation (2.9)

$$\kappa_g = \pm \kappa \sin \psi$$

It is clear from their definition that κ_η and κ_g either stay the same or both change sign when σ is reparametrized.

When $\psi = 0$, the curve is called a normal section. Then $\kappa_\eta = \pm \kappa$, $\kappa_g = 0$.

Theorem 2.19. If $\gamma(t) = \sigma(u(t), v(t))$ is a unit speed curve on a surface patch σ , its normal curvature is given by

$$\kappa_\eta = \mathcal{L}\dot{u}^2 + 2\mathcal{M}\dot{u}\dot{v} + \mathcal{N}\dot{v}^2,$$

where $\mathcal{L}\dot{u}^2 + 2\mathcal{M}\dot{u}\dot{v} + \mathcal{N}\dot{v}^2$ is the second fundamental form of σ .

Proof. We have, with N denoting the standard unit normal of σ ,

$$\begin{aligned}
\kappa_\eta &= N \cdot \ddot{\gamma} \\
&= N \cdot \frac{d}{dt}(\dot{\gamma}) \\
&= N \cdot \frac{d}{dt}(\sigma_u \dot{u} + \sigma_v \dot{v}) \\
&= N \cdot (\sigma_u \ddot{u} + \sigma_v \ddot{v} + (\sigma_{uu} \dot{u} + \sigma_{uv} \dot{v}) \dot{u} + (\sigma_{uv} \dot{u} + \sigma_{vv} \dot{v}) \dot{v}) \\
&= \mathcal{L} \dot{u}^2 + 2\mathcal{M} \dot{u} \dot{v} + \mathcal{N} \dot{v}^2
\end{aligned}$$

□

Theorem 2.20. (Meusnier's Theorem) *Let P be a point on a surface S and let V be a unit vector to S at P . Let π_θ be the plane containing the line through P parallel to V and making an angle θ with the tangent plane to S at P . Suppose that π_θ intersects S in a curve with curvature κ_θ . Then $\kappa_\theta \sin \theta$ is independent of θ .*

Proof. Assume that γ_θ is a unit speed parametrisation of the curve of intersection of π_θ and S . Then at P , $\dot{\gamma}_\theta = \pm v$, so $\ddot{\gamma}_\theta$ is perpendicular to v and is parallel to π_θ . Thus if ψ is the angle between η and N , then $\psi = \pi/2 - \theta$. Again, we know

$$\kappa_\eta = \kappa \cos \psi$$

Hence,

$$\kappa_\eta = \kappa \sin \theta = \kappa_\theta \sin \theta$$

(since, here $\kappa = \kappa_\theta$ by notation).

□

2.7 Matrix Representation of Normal Curvature

To analyse κ further, it is useful to use matrix notation. If $Edu^2 + 2Fdu dv + Gdv^2$ and $\mathcal{L}du^2 + 2\mathcal{M}du dv + \mathcal{N}dv^2$ are the first and second fundamental forms of a surface σ , we introduce the following symmetric 2×2 matrices

$$F_1 = \begin{bmatrix} E & F \\ F & G \end{bmatrix}, \quad F_{II} = \begin{bmatrix} \mathcal{L} & \mathcal{M} \\ \mathcal{M} & \mathcal{N} \end{bmatrix}$$

Let

$$t_1 = \xi_1 \sigma_u + \eta_1 \sigma_v, \quad t_2 = \xi_2 \sigma_u + \eta_2 \sigma_v$$

be two tangent vectors at some point of σ . Then

$$\begin{aligned}
t_1 \cdot t_2 &= (\xi_1 \sigma_u + \eta_1 \sigma_v) \cdot (\xi_2 \sigma_u + \eta_2 \sigma_v) \\
&= E \xi_1 \xi_2 + F(\xi_1 \eta_2 + \xi_2 \eta_1) + G \eta_1 \eta_2 \\
&= \begin{bmatrix} \xi_1 & \eta_1 \end{bmatrix} \begin{bmatrix} E & F \\ F & G \end{bmatrix} \begin{bmatrix} \xi_2 \\ \eta_2 \end{bmatrix}
\end{aligned}$$

Thus writing

$$T_1 = \begin{bmatrix} \xi_1 \\ \eta_1 \end{bmatrix}, \quad T_2 = \begin{bmatrix} \xi_2 \\ \eta_2 \end{bmatrix} \tag{2.10}$$

we get $t_1.t_2 = T_1^t F_1 T_2$. On the other hand, tangent vector

$$\dot{\gamma} = \dot{u}\sigma_u + \dot{v}\sigma_v$$

and if

$$T = \begin{bmatrix} \dot{u} \\ \dot{v} \end{bmatrix}$$

then by using theorem 2.19, we see that

$$\kappa_\eta = T^t F_{II} T$$

Exercise 2.21. Compute the normal curvature of the circle $\gamma(t) = (\cos t, \sin t, 1)$ on the elliptic paraboloid $\sigma(u, v) = (u, v, u^2 + v^2)$.

Unit V

Course Structure

- Introduction
 - Geodesic curvatures
 - Gaussian and Mean Curvatures
-

3 Introduction

In the present unit, we shall study curvature of surfaces using curvature of curves defined on surfaces.

In the previous unit, we have seen the normal curvature of a curve defined on a surface is given by

$$\kappa_\eta = \mathcal{L}(\dot{u})^2 + 2\mathcal{M}\dot{u}\dot{v} + \mathcal{N}(\dot{v})^2$$

So,

$$\begin{aligned}\kappa_\eta &= \mathcal{L} \left(\frac{du}{ds} \right)^2 + 2\mathcal{M} \frac{du}{ds} \frac{dv}{ds} + \mathcal{N} \left(\frac{dv}{ds} \right)^2 \\ &= \frac{\mathcal{L}(du)^2 + 2\mathcal{M}dudv + \mathcal{N}(dv)^2}{ds^2} \\ &= \frac{\mathcal{L}(du)^2 + 2\mathcal{M}dudv + \mathcal{N}(dv)^2}{E(du)^2 + 2Fdudv + G(dv)^2}\end{aligned}$$

Hence

$$\mathcal{L}(du)^2 + 2\mathcal{M}dudv + \mathcal{N}(dv)^2 = \kappa_\eta(E(du)^2 + 2Fdudv + G(dv)^2) \quad (3.1)$$

From equation (3.1), after some calculations, we can show that κ_η has maximum or minimum value if

$$|F_{II} - \kappa_\eta F_I| = 0$$

The maximum or minimum values of κ_η are called principal curvatures of the surface (not curve).

Definition 3.1. The principal curvatures of a surface patch are the roots of the equation

$$\begin{vmatrix} \mathcal{L} - \kappa E & \mathcal{M} - \kappa F \\ \mathcal{M} - \kappa F & \mathcal{N} - \kappa G \end{vmatrix} = 0$$

If κ_1 and κ_2 are the roots, then $\kappa_1\kappa_2$ is called Gaussian curvature of the surface and $\frac{\kappa_1 + \kappa_2}{2}$ is called mean curvature of the surface.

Theorem 3.2. Let κ_1 and κ_2 be the principal curvatures at a point P of a surface patch σ . Then

1. κ_1 and κ_2 are real numbers
2. If $\kappa_1 = \kappa_2 = \kappa$, say, then $F_{II} = \kappa F_I$ and hence every tangent vector to σ at P is principal vector
3. If $\kappa_1 \neq \kappa_2$, then any two non-zero principal vectors d_1 and d_2 corresponding to κ_1 and κ_2 , respectively, are perpendicular.

(For case 2, P is called an umbilic).

Proof. For 1, let t_1 and t_2 be any two perpendicular unit tangent vectors to the surface at P (not yet known to be principal vectors). Define ξ_i, η_i, T_i for $i = 1, 2$ as done in the previous unit. Let

$$A = \begin{bmatrix} \xi_1 & \xi_2 \\ \eta_1 & \eta_2 \end{bmatrix}$$

By multiplying out the matrices, it is easy to check that

$$\begin{aligned} A^t F_I A &= \begin{bmatrix} T_1^t F_I T_1 & T_1^t F_I T_2 \\ T_2^t F_I T_1 & T_2^t F_I T_2 \end{bmatrix} \\ &= \begin{bmatrix} t_1 \cdot t_1 & t_1 \cdot t_2 \\ t_2 \cdot t_1 & t_2 \cdot t_2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{aligned}$$

Since t_1 and t_2 are perpendicular unit vectors.

Let $G_{II} = A^t F_{II} A$. Then G_{II} is still (real and) symmetric because

$$\begin{aligned} G_{II}^t &= A^t F_{II} (A^t)^t \\ &= A^t F_{II} A \\ &= G_{II} \end{aligned}$$

From the theory of Linear algebra, we can say, there is an orthogonal matrix B such that

$$B^t G_{II} B = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

for some real numbers λ_1 and λ_2 . Let $C = AB$. Then

$$\begin{aligned} C^t F_I C &= B^t (A^t F_I A) B \\ &= B^t B \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \end{aligned} \tag{3.2}$$

because B is orthogonal, and

$$\begin{aligned} C^t F_{II} C &= B^t (A^t F_{II} A) B \\ &= B^t G_{II} B \\ &= \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \end{aligned} \tag{3.3}$$

Now, C is invertible. So,

$$\det(F_{II} - \kappa F_I) = 0 \text{ if and only if } \det(C^t(F_{II} - \kappa F_I)C) = 0,$$

hence,

$$\begin{aligned} \det(F_{II} - \kappa F_I) &= 0 \text{ if and only if} \\ \det\left(\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} - \kappa \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) &= 0 \end{aligned}$$

Hence the principal curvatures are the roots of

$$\begin{vmatrix} \lambda_1 - \kappa & 0 \\ 0 & \lambda_2 - \kappa \end{vmatrix} = 0,$$

For 2., suppose that the principal curvatures are equal to κ_1 , say. Then $\lambda_1 = \lambda_2 = \kappa$, and equations (3.2) and (3.3) give

$$C^t F_I C = I, \quad C^t F_{II} C = \kappa I$$

Hence

$$\begin{aligned} C^t(F_{II} - \kappa F_I)C &= 0 \\ \implies F_{II} - \kappa F_I &= 0 \\ \implies F_{II} &= \kappa F_I \end{aligned}$$

since C and C^t are invertible.

Obviously, if D is any 2×1 column matrix,

$$(F_{II} - \kappa F_I)D = 0$$

It follows that every tangent vector to σ at P is a principal vector.

Finally, for 3, let

$$\begin{aligned} d_i &= \alpha_i \sigma_u + \beta_i \sigma_v \\ D_i &= \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix}, \end{aligned}$$

for $i = 1, 2$. Then by $t_1, t_2 = T_1^t F_I T_2$, $d_1 d_2 = D_1^t F_I D_2$. From the definition of principal vector

$$F_{II} D_1 = \kappa_1 F_I D_1, \quad F_{II} D_2 = \kappa_2 F_I D_2$$

Hence,

$$\begin{aligned} D_2^t F_{II} D_1 &= \kappa_1 (d_1 d_2) \\ D_2^t F_{II} D_2 &= \kappa_2 (d_1 d_2) \end{aligned} \tag{3.4}$$

But since $D_1^t F_{II} D_2$ is a 1×1 matrix, it is equal to its transpose

$$\begin{aligned} D_1^t F_{II} D_2 &= (D_1^t F_{II} D_2)^t \\ &= D_2^t F_{II}^t D_1 \\ &= D_2^t F_{II} D_1, \end{aligned}$$

the last equality comes from the fact that F_{II} is symmetric. Hence equation (3.4) gives

$$\kappa_1(d_1 d_2) = \kappa_2(d_1 d_2)$$

So, $\kappa_1 \neq \kappa_2$ implies $d_1 d_2 = 0$, that is, d_1 and d_2 are perpendicular. □

Example 3.3. Find the Gaussian curvature of the right circular cylinder.

We consider the circular cylinder of radius 1 and axis z -axis, parametrized in the usual way

$$\sigma(u, v) = (\cos v, \sin v, u).$$

Here,

$$E = \sigma_u \sigma_u = 1, \quad F = \sigma_u \sigma_v = 0, \quad G = \sigma_v \sigma_v = 1$$

Also

$$\mathcal{L} = 0, \quad \mathcal{M} = 0, \quad \mathcal{N} = 1.$$

So, the principal curvatures are the roots of

$$\begin{vmatrix} 0 - \kappa & 0 \\ 0 & 1 - \kappa \end{vmatrix} = 0$$

Hence

$$\kappa(\kappa - 1) = 0, \text{ or, } \kappa = 0, \text{ or } 1$$

Hence the principal curvatures are 0 and 1. Thus,

$$\text{Gaussian curvature} = 0 \times 1 = 0$$

$$\text{Mean Curvature} = \frac{0 + 1}{2} = \frac{1}{2}$$

Exercise 3.4. Find the Gaussian curvature of the unit sphere.

Note 3.5. In the following, we shall study Euler's theorem which will relate the curvature of a surface and curvature of a curve defined on a surface.

Theorem 3.6. (Euler's Theorem) Let γ be a curve on a surface patch σ , and let κ_1 and κ_2 be the principal curvatures of σ , with non-zero principal non-zero vectors t_1 and t_2 . Then the normal curvature of γ is

$$\kappa_\eta = \kappa_1 \cos^2 \theta + \kappa_2 \sin^2 \theta$$

where θ is the angle between $\dot{\gamma}$ and t_1 .

Proof. We can assure that γ is a unit speed. Let t be the tangent vector of γ , and let

$$t = \xi\sigma_u + \eta\sigma_v,$$

$$T = \begin{bmatrix} \xi \\ \eta \end{bmatrix}$$

Suppose that

$$\kappa_1 = \kappa_2 = \kappa, \text{ say.}$$

Now,

$$\begin{aligned} \kappa_\eta &= T^t F_{II} t \\ &= \kappa T^t F_I T \\ &= \kappa t \cdot t \\ &= \kappa \end{aligned}$$

This agrees with the formula in the statement since

$$\begin{aligned} \kappa_1 \cos^2 \theta + \kappa_2 \sin^2 \theta &= \kappa(\cos^2 \theta + \sin^2 \theta) \\ &= \kappa \end{aligned}$$

Now, assume that $\kappa_1 \neq \kappa_2$, so that t_1 and t_2 are perpendicular. We might as well assume that t_1 and t_2 are unit vectors.

$$\begin{aligned} t_i &= \xi_i \sigma_u + \eta_i \sigma_v \\ T_i &= \begin{bmatrix} \xi_i \\ \eta_i \end{bmatrix} \end{aligned}$$

for $i = 1, 2$. Now

$$\dot{\gamma} = \cos \theta t_1 + \sin \theta t_2$$

So,

$$\cos \theta (\xi_1 \sigma_u + \eta_1 \sigma_v) + \sin \theta (\xi_2 \sigma_u + \eta_2 \sigma_v) = \xi \sigma_u + \eta \sigma_v$$

Hence

$$\begin{aligned} \xi_1 \cos \theta + \xi_2 \sin \theta &= \xi \\ \eta_1 \cos \theta + \eta_2 \sin \theta &= \eta \end{aligned}$$

Thus,

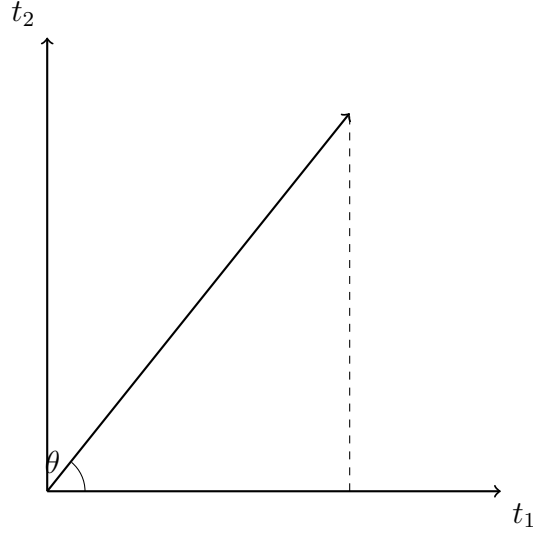
$$\begin{bmatrix} \xi \\ \eta \end{bmatrix} = \cos \theta \begin{bmatrix} \xi_1 \\ \eta_1 \end{bmatrix} + \sin \theta \begin{bmatrix} \xi_2 \\ \eta_2 \end{bmatrix}$$

Therefore,

$$T = \cos \theta T_1 + \sin \theta T_2.$$

Hence

$$\begin{aligned} \kappa_\eta &= T^t F_{II} T \\ &= (\cos \theta T_1^t + \sin \theta T_2^t) F_{II} (\cos \theta T_1 + \sin \theta T_2) \\ &= \cos^2 \theta T_1^t F_{II} T_1 + \cos \theta \sin \theta (T_1^t F_{II} T_2 + T_2^t F_{II} T_1) + \sin^2 \theta T_2^t F_{II} T_2 \end{aligned}$$



Now,

$$\begin{aligned} T_i^t F_{II} T_j &= \kappa_i T_i^t F_{II} T_j = \kappa_i, \text{ if } i = j \\ &= 0, \text{ otherwise} \end{aligned}$$

Hence

$$\kappa_\eta = \kappa_1 \cos^2 \theta + \kappa_2 \sin^2 \theta$$

□

Theorem 3.7. Let $\sigma(u, v)$ be a surface patch with first and second fundamental forms $Edu^2 + 2Fdu dv + Gdv^2$ and $\mathcal{L}du^2 + 2\mathcal{M}du dv + \mathcal{N}dv^2$, respectively. Then

1. $\kappa = \frac{\mathcal{L}\mathcal{N} - \mathcal{M}^2}{EG - F^2}$
2. $H = \frac{\mathcal{L}G - 2\mathcal{M}F + \mathcal{N}E}{2(EG - F^2)}$
3. The principal curvatures are

$$H \pm \sqrt{H^2 - \kappa}.$$

Proof. The principal curvatures are the roots of

$$\begin{vmatrix} \mathcal{L} - \kappa E & \mathcal{M} - \kappa F \\ \mathcal{M} - \kappa F & \mathcal{N} - \kappa G \end{vmatrix} = 0$$

Hence

$$(EG - F^2)\kappa^2 - (\mathcal{L}G - 2\mathcal{M}F + \mathcal{N}E)\kappa + \mathcal{L}\mathcal{N} - \mathcal{M}^2 = 0, \quad (3.5)$$

The above equation is quadratic in κ . Hence

$$\kappa_1 + \kappa_2 = \frac{\mathcal{L}G - 2\mathcal{M}F + \mathcal{N}E}{(EG - F^2)}$$

Hence,

$$H = \frac{\kappa_1 + \kappa_2}{2} = \frac{\mathcal{L}G - 2\mathcal{M}F + \mathcal{N}E}{2(EG - F^2)} \quad (3.6)$$

and

$$K = \kappa_1\kappa_2 = \frac{\mathcal{L}\mathcal{N} - \mathcal{M}^2}{EG - F^2} \quad (3.7)$$

K is the Gaussian curvature. Then, from equations (3.5), (3.6) and (3.7),

$$\kappa^2 - 2H\kappa + K = 0.$$

Hence

$$H \pm \sqrt{H^2 - K}$$

are the principal curvatures. □

Definition 3.8. A surface is called flat if its Gaussian curvature is zero.

Summary

In this unit we have learnt how to calculate the Gaussian and mean curvatures of a surface.

Unit VI

Course Structure

- Introduction
 - Geodesics on surfaces
 - Differential equation for geodesics on surfaces
-

4 Introduction

In simple language a geodesic means the line of shortest distance between two points. In our Euclidean plane, a geodesic is a straight line. In this unit, we shall see that on a sphere, geodesics are great circles. Here, we will first deduce the differential equation of geodesics on a surface.

4.1 Geodesics on Surface

Definition 4.1. A curve γ on a surface S is called a geodesic if $\ddot{\gamma}(t)$ is zero or perpendicular to the surface at the point $\gamma(t)$, that is, parallel to its unit normal, for all values of the parameter t .

Theorem 4.2. *Any geodesic has constant speed.*

Proof. Let $\gamma(t)$ be a geodesic on a surface S . Then denoting $\frac{d}{dt}$ by a dot, we have

$$\frac{d}{dt} \|\dot{\gamma}\|^2 = \frac{d}{dt} (\dot{\gamma} \cdot \dot{\gamma}) = 2\ddot{\gamma} \cdot \dot{\gamma}$$

Since γ is a geodesic, $\ddot{\gamma}$ is perpendicular to the tangent plane and is therefore perpendicular to the tangent vector $\dot{\gamma}$. So, $\ddot{\gamma} \cdot \dot{\gamma} = 0$ and the last equation shows that $\|\dot{\gamma}\|$ is constant. $\square$

Theorem 4.3. *A curve on a surface is a geodesic if and only if its geodesic curvature is zero.*

Proof. It is sufficient to consider a unit speed curve γ contained in a patch σ of the surface. Let N be the standard unit normal of σ , so that

$$\kappa g = \ddot{\gamma} \cdot (N \times \dot{\gamma}) \quad (4.1)$$

If $\ddot{\gamma}$ is parallel to N , it is obviously perpendicular to $N \times \dot{\gamma}$, so by (4.1),

$$\kappa g = 0.$$

Conversely, suppose that $\kappa g = 0$. Then $\ddot{\gamma}$ is perpendicular to $N \times \dot{\gamma}$. But then, since $\dot{\gamma}$, N and $N \times \dot{\gamma}$ are perpendicular unit vectors in $\mathbb{R}^3$, and since $\ddot{\gamma}$ is perpendicular to $\dot{\gamma}$, it follows that $\ddot{\gamma}$ is perpendicular to N . $\square$

Exercise 4.4. 1. Prove that for normal section, $\kappa g = 0$.

2. All great circles on a sphere are geodesics.

4.2 Geodesic Equations

A curve γ on a surface S is a geodesic if and only if for any part $\gamma(t) = \sigma(u(t), v(t))$ of γ contained in a surface patch σ of S , the following two equations are satisfied

$$\begin{aligned}\frac{d}{dt}(E\dot{u} + F\dot{v}) &= \frac{1}{2}(E_u\dot{u}^2 + 2F_u\dot{u}\dot{v} + G_u\dot{v}^2) \\ \frac{d}{dt}(F\dot{u} + G\dot{v}) &= \frac{1}{2}(E_v\dot{u}^2 + 2F_v\dot{u}\dot{v} + G_v\dot{v}^2)\end{aligned}$$

where $Edu^2 + 2Fdudv + Gdv^2$ is the first fundamental form of σ .

Proof. Since $\{\sigma_u, \sigma_v\}$ is a basis of the tangent plane of σ , γ is a geodesic if and only if $\ddot{\gamma}$ is perpendicular to σ_u and σ_v . Since

$$\dot{\gamma} = \sigma_u\dot{u} + \sigma_v\dot{v},$$

this is equivalent to

$$\left(\frac{d}{dt}(\sigma_u\dot{u} + \sigma_v\dot{v})\right) \cdot \sigma_u = 0 \quad (4.2)$$

$$\left(\frac{d}{dt}(\sigma_u\dot{u} + \sigma_v\dot{v})\right) \cdot \sigma_v = 0 \quad (4.3)$$

we show that these two equations are equivalent to the two geodesic equations. The left hand side of equation (4.2) is equal to

$$\begin{aligned}\frac{d}{dt}(\sigma_u\dot{u} + \sigma_v\dot{v}) \cdot \sigma_u - (\sigma_u\dot{u} + \sigma_v\dot{v}) \cdot \frac{d}{dt}\sigma_u &= \frac{d}{dt}(E\dot{u} + F\dot{v}) - (\sigma_u\dot{u} + \sigma_v\dot{v})(\sigma_{uu}\dot{u} + \sigma_{uv}\dot{v}) \\ &= \frac{d}{dt}(E\dot{u} + F\dot{v}) \\ &\quad - (\dot{u}^2(\sigma_u\sigma_{uu})\dot{u}\dot{v}(\sigma_u\sigma_{uv} + \sigma_v\sigma_{uu}) + \dot{v}^2(\sigma_v\sigma_{uv}))(4.4)\end{aligned}$$

Now,

$$\begin{aligned}E_u &= (\sigma_u \cdot \sigma_u)_u \\ &= \sigma_{uu}\sigma_u + \sigma_u\sigma_{uu} \\ &= 2\sigma_u\sigma_{uu}\end{aligned}$$

So,

$$\sigma_u\sigma_{uu} = \frac{1}{2}E_u$$

Similarly,

$$\sigma_v\sigma_{uv} = \frac{1}{2}G_u$$

Finally,

$$\sigma_v\sigma_{uv} + \sigma_v\sigma_{uu} = (\sigma_u \cdot \sigma_u)_u = F_u$$

Substituting these values in equation (4.4), gives

$$\left(\frac{d}{dt}(\sigma_u\dot{u} + \sigma_v\dot{v})\right) \cdot \sigma_u = \frac{d}{dt}(E\dot{u} + F\dot{v}) - \frac{1}{2}(E_u\dot{u}^2 + 2F_u\dot{u}\dot{v} + G_u\dot{v}^2).$$

This shows that the equation (4.2) is equivalent to the first geodesic equation. Similarly the other. $\square$

Example 4.5. Determine the geodesics on S^2 by solving the geodesic equations.

For the usual parametrisation by latitude and longitude coordinate

$$\sigma(u, v) = (\cos \theta \cos \phi, \cos \theta \sin \phi, \sin \theta),$$

we know that the first fundamental form of S^2 is

$$d\theta^2 + \cos^2 \theta d\phi^2$$

We might as well restrict ourselves to unit speed curves

$$\gamma(t) = \sigma(\theta(t), \phi(t)),$$

so that

$$\dot{\theta}^2 + \dot{\phi}^2 \cos^2 \theta = 1,$$

and if γ is a geodesic the second geodesic equation gives

$$\frac{d}{dt}(\dot{\phi} \cos^2 \theta) = 0,$$

so that,

$$\dot{\phi} \cos^2 \theta = \Omega,$$

where, Ω is a constant. If $\Omega = 0$, then $\dot{\phi} = 0$, so ϕ is constant and γ is part of a meridian. We assume that $\Omega \neq 0$ from now on.

The unit speed condition gives

$$\begin{aligned} \dot{\theta} &= \sqrt{1 - \dot{\phi}^2 \cos^2 \theta} \\ \dot{\theta}^2 &= 1 - \frac{\Omega^2}{\cos^2 \theta}, \end{aligned}$$

so along the geodesics we have

$$\left(\frac{d\phi}{d\theta}\right)^2 = \frac{\dot{\phi}^2}{\dot{\theta}^2} = \frac{1}{\cos^2 \theta \left(\frac{\cos^2 \theta}{\Omega^2} - 1\right)}$$

and hence

$$\pm(\phi - \phi_0) = \int \frac{d\theta}{\cos \theta \sqrt{\frac{\cos^2 \theta}{\Omega^2} - 1}},$$

where, ϕ_0 is a constant. The integral can be evaluated by making the substitution $u = \tan \theta$. This gives

$$\pm(\phi - \phi_0) = \int \frac{du}{\sqrt{\frac{1}{\Omega^2} - 1 - u^2}} \arcsin \left(\frac{u}{\sqrt{\frac{1}{\Omega^2} - 1}} \right),$$

and hence

$$\tan \theta = \pm \sqrt{\frac{1}{\frac{1}{\Omega^2} - 1}} \sin(\phi - \phi_0).$$

This implies that the coordinates $x = \cos \theta \cos \phi$, $y = \cos \theta \sin \phi$, $z = \sin \theta$ of $\gamma(t)$ satisfy $z = ax + by$.

Hence the geodesics are the intersection of the sphere and planes passing through the centre. So geodesics are the great circles.

Summary

In this unit, we have studied properties of geodesics and have deduced differential equations of geodesics and have seen examples for the same.

Unit VII

Course Structure

- Introduction
 - Introduction to manifolds
 - Manifolds as a generalization of surface
-

5 Introduction

In the definition of surface, we have seen that for each point $\mathbf{p} \in S$, where, S is a surface, there is an open neighbourhood of $\mathbf{p}$ which is homeomorphic to an open subset of $\mathbb{R}^2$. Also, we have taken the surface S as a subset of $\mathbb{R}^3$.

If instead of taking S as a subset of $\mathbb{R}^3$, we take S as a topological space and for if every point $\mathbf{p} \in S$, there is a neighbourhood which is homeomorphic to an open set of $\mathbb{R}^n$, then S will be called a topological manifold or simply a manifold.

Like the concept of smooth surface, there is a concept of smooth manifold or differentiable manifold. In the following, we give definition of smooth manifold.

5.1 Smooth Manifold

Let M be a topological space. Let for each $\mathbf{p} \in M$, there is a neighbourhood of $\mathbf{p}$ which is homeomorphic to an open subset of $\mathbb{R}^n$.

Let $\mathbf{p}$ has a neighbourhood U which is homeomorphic to an open subset A of $\mathbb{R}^n$ by the homeomorphism ϕ .

Let $\mathbf{p}$ has another neighbourhood V which is homeomorphic to another open subset B of $\mathbb{R}^n$ by the homeomorphism ψ . See that $\phi^{-1}(U \cap V)$ is a subset of A and $\psi^{-1}(U \cap V)$ is a subset of B .

The topological space M will be called a smooth manifold or differentiable manifold if the map

$$\psi^{-1} \circ \phi : \phi^{-1}(U \cap V) \rightarrow \psi^{-1}(U \cap V)$$

or,

$$\phi^{-1} \circ \psi : \psi^{-1}(U \cap V) \rightarrow \phi^{-1}(U \cap V)$$

is differentiable.

- Exercise 5.1.**
1. Show that a circle is a smooth manifold.
 2. Show that a sphere is a smooth manifold.
 3. Show that the set of all $n \times n$ matrices is a smooth manifold.
 4. Show that the set of all polynomials with real coefficient and degree less than three is a smooth manifold.
 5. Show that an open interval of $\mathbb{R}$ is a smooth manifold.
 6. Show that $\mathbb{R}^n$ is a smooth manifold.

Summary

Here, we have defined smooth manifold.

References

- [1] *Elementary Differential Geometry*, Andrew Pressley
- [2] *Differential Geometry of Curves and Surfaces*, M. P do Carmo
- [3] *A course in Differential Geometry and Lie Groups*, S. Kumaresan
- [4] *Differential Geometry of Manifolds*, U.C De and A. A Shaikh
- [5] *Multivariable Calculus and Geometry*, S. Dineen
- [6] *Elementary Differential Geometry*, C. Bär

Elective Paper

MATP 2.4

Block - II

Marks : 50 (SEE : 40; IA : 10)

● Topology - II

Syllabus

Unit	Topic
9	Connectedness : Examples, various characterizations and basic properties. Connectedness on the real line.
10	Components and quasi components. Path connectedness and path components.
11	Compactness : Characterizations and basic properties of compactness, Lebesgue, lemma. Sequential compactness
12	BW Compactness and countable compactness. Local compactness and Baire Category Theorem
13	Identification spaces: Constructing a Mobius strip, identification topology, Orbit spaces.
14	Some Matrix Lie Groups : Some elementary properties of topological groups.
15	$GL(n, \mathbb{R})$ as a topological group and its subgroups.
16	Fundamental groups, calculation of fundamental group of S .

Unit IX

Course Structure

- Introduction
 - Objectives
 - Connectedness : Examples, various characterizations and basic properties
 - Connectedness on the real line
-

6 Introduction

The intermediate value theorem is a very important property of the continuous functions in the real line. It says that if a function $f : [a, b] \rightarrow \mathbb{R}$ is continuous, and if r is a real number between $f(a)$ and $f(b)$, then there exists an element $c \in [a, b]$ such that $f(c) = r$. Can you see the geometric interpretation of this theorem? What if we define a continuous function on the interval $[1, 2] \cup [3, 4]$, then is the theorem true for this domain also? You will notice that its not. The intermediate value theorem is true for $[a, b]$ due to a particular property of the interval which is rightly termed as "connectedness" of the interval. As the name suggests, the interval is not "broken" or "separated" in some sense. In this unit, we will learn the concept of connectedness and its properties.

Objectives

After reading this unit, you will be able to

- define connected sets and differentiate between connected and separated sets.
- learn various equivalent definitions for connected spaces.
- learn the characteristics of connected sets.
- know the structure of connected sets in the real line.
- learn examples of connected and separated sets.

6.1 Connected Spaces

Roughly speaking, a topological space is said to be connected if it does not allow itself to be partitioned into two disjoint proper open subsets of itself. In an arbitrary topological space, we have only the open sets at our disposal. So we first define separated set as follows:

Definition 6.1. Let (X, τ) be a topological space. Then, X is said to be disconnected if there exists two open sets U and V , such that

1. $U \cap V = \phi$.
2. $U \cup V = X$.

Then the pair (U, V) is said to form a disconnection of X .

Definition 6.2. A topological space X is said to be connected if it does not admit of a disconnection.

Does the closed interval $[a, b]$ admit of a disconnection?

Connectedness is obviously a topological property since its defined in terms of open sets entirely. In other terms, we can say that if X is connected, then any space homeomorphic to X is also connected. In fact, if X and Y are homeomorphic, then there exists a homeomorphism $f : X \rightarrow Y$. If X is not connected, then it does not have a disconnection. If possible, let Y is disconnected. Then there exists disjoint open sets, say U and V . Since f is continuous, so $f^{-1}(U)$ and $f^{-1}(V)$ are open sets in X that forms a disconnection for X (verify!). Which is a contradiction. Hence Y is also connected.

Consider a subset A of a topological space X , which is both open and closed. We know that in any topological space X , the empty set $\emptyset$ and X are always both open and closed. Does there exist any other subset in X with the property? Consider the following example:

Example 6.3. The set $Y = [0, 1] \cup [3, 4]$ is a subspace of the real space with the usual topology. Then, the sets $[0, 1] = Y \cap (-1/2, 3/2)$ is open in Y . Similarly, $[3, 4] = Y \cap (3/2, 9/2)$ is open in Y . Also, see that $[0, 1] = Y \setminus [3, 4]$ and $[3, 4] = Y \setminus [0, 1]$. Hence, both are closed also. Thus the subspace Y has two proper subsets which are both open and closed.

As you can probably figure out, the subspace Y is not connected (the sets $[0, 1]$ and $[3, 4]$ forms a disconnection for Y). In fact, for connected space, we won't find any other proper subset which is both open and closed. We have the following equivalent definition for connectedness of a space X .

Theorem 6.4. A space X is said to be connected if and only if X and $\emptyset$ are the only subsets of X which are both open and closed.

Proof. Let X be connected and if possible, let A be a subset of X which is both open and closed. Since A is closed, $X \setminus A$ is open. Also, A and $X \setminus A$ are disjoint open sets whose union is X . Thus A and $X \setminus A$ forms a disconnection of X , a contradiction. Thus no proper subset of X can be both open and closed.

Conversely, suppose X is disconnected. Let U and V forms a disconnection of X . Then $X = U \cup V$. Which means, $U = X \setminus V$ and $V = X \setminus U$ implying that U and V are both open and closed. Hence the theorem. $\square$

Example 6.5. 1. Let X be a two-point set with the indiscrete topology. Then X has no separation. In fact, any set with the indiscrete topology is always connected. But what happens if we take the discrete topology?

2. Consider $\mathbb{R}$ with co-finite topology. Then $\mathbb{R}$ is connected. If possible, let A and B be a separation of $\mathbb{R}$. Then $A \cap B = \emptyset$ and $\mathbb{R} \setminus A$ and $\mathbb{R} \setminus B$ are both finite sets. So, $(\mathbb{R} \setminus A) \cup (\mathbb{R} \setminus B)$ is a finite set. That is, $\mathbb{R} \setminus (A \cap B)$ is a finite set, that is, $\mathbb{R}$ is a finite set, since $A \cap B = \emptyset$. Thus, $\mathbb{R}$ is connected with the co-finite topology.

3. $\mathbb{R}$ with the usual topology is also connected. (can you prove it?)

Definition 6.6. Let A and B be two subsets of a topological space (X, τ) . Then A and B are said to be separated if $\overline{A} \cap B = \emptyset$ and $A \cap \overline{B} = \emptyset$ simultaneously hold.

Definition 6.7. Let Y be a subspace of a topological space (X, τ) . Then Y is said to be connected if it is connected with respect to the topology induced by τ , that is, if there does not exist disjoint open sets A and B in Y such that $Y = A \cup B$.

Is a subspace of a connected space always connected? Let's consider the following example:

Example 6.8. The subspace of all rationals is disconnected with the usual topology. Let $x, y \in \mathbb{Q}$ such that $x < y$. Then, there exists an irrational number a such that $x < a < y$. Then we can write

$$\mathbb{Q} = \{(-\infty, a) \cap \mathbb{Q}\} \cup \{(a, \infty) \cap \mathbb{Q}\}$$

The sets $\{(-\infty, a) \cap \mathbb{Q}\}$ and $\{(a, \infty) \cap \mathbb{Q}\}$ are non-empty since $x \in \{(-\infty, a) \cap \mathbb{Q}\}$ and $y \in \{(a, \infty) \cap \mathbb{Q}\}$ and open in the subspace topology. Thus $\mathbb{Q}$ is disconnected. Can you give other examples of disconnected subspaces of a connected topological space?

Note that, by choosing different $a \in \mathbb{Q}$, we can get a different disconnection for $\mathbb{Q}$. In fact, in this way, we can get infinite number of disconnections for $\mathbb{Q}$. We could similarly show that the set of irrationals is also disconnected. Also we see that the union of rationals and irrationals give us $\mathbb{R}$. Thus, we can conclude that two the union of two disconnected sets may be connected. Some additional properties are required for the union to be disconnected too. We have the following theorem in this direction:

Theorem 6.9. If A and B are two separated non-empty sets of (X, τ) , then $A \cup B$ is disconnected.

Proof. Let A and B are two separated non-empty sets of (X, τ) . Then we have $\overline{A} \cap B = \emptyset$ and $A \cap \overline{B} = \emptyset$. Put $G = X \setminus \overline{B}$ and $H = X \setminus \overline{A}$. Then G and H are disjoint non-empty sets in X . Then we have

$$\begin{aligned} (A \cup B) \cap G &= (A \cap G) \cup (B \cap G) \\ &= A \cup \emptyset \\ &= A \end{aligned}$$

Similarly, we have

$$(A \cup B) \cap H = B$$

Thus,

$$A \cup B = ((A \cup B) \cap G) \cup ((A \cup B) \cap H)$$

Hence, $(A \cup B) \cap G$ and $(A \cup B) \cap H$ clearly forms a disconnection for $A \cup B$. Hence the result. $\square$

Theorem 6.10. *A subset Y of (X, τ) is disconnected if and only if Y is the union of two non-empty separated sets.*

Proof. The sufficient condition immediately follows from the previous theorem.

For the necessary part, let Y be disconnected. Then Y can be partitioned as

$$Y = (Y \cap A) \cup (Y \cap B)$$

where A and B are open sets in X whose intersections with Y are non-empty. We will check for the separatedness of $Y \cap A$ and $Y \cap B$. If possible, let a be a limit point of $Y \cap A$ and $a \in Y \cap B$. Since B is an open set containing a , so

$$B \cap \{(Y \cap A) \setminus \{a\}\} \neq \emptyset$$

which is a contradiction since

$$B \cap (Y \cap A) = (Y \cap A) \cap (Y \cap B) = \emptyset.$$

Hence $Y \cap A$ and $Y \cap B$ are separated sets. □

Example 6.11. 1. We told that the set $Y = [0, 1] \cup [3, 4]$ is disconnected. But in light of the previous theorem, we can see that $[0, 1]$ and $[3, 4]$ forms a disconnection for Y in which neither of them contains a limit point of the other.

2. The set $Y = [-1, 0) \cup (0, 1]$ of the real line is disconnected since $[-1, 0)$ and $(0, 1]$ are separated sets whose union is Y (neither of them contains a limit point of the other).
3. The set $Y = [0, 2]$ of the real line is connected. The sets $[0, 1)$ and $[1, 2]$ are of course disjoint whose union gives Y . But $[1, 2]$ is not open in the subspace Y . So this does not form a disconnection for Y . In fact, by the previous theorem, we can say that since $[0, 1)$ and $[1, 2]$ are not separated, so it does not form a disconnection for Y .

Now we will see certain results for connectedness in series.

Lemma 6.12. *If the sets C and D form a disconnection for X and Y is a connected subspace of X , then Y is either in C or D .*

Proof. Since C and D are disjoint open sets in X , the sets $C \cap Y$ and $D \cap Y$ are open in Y . These two sets are disjoint and their union is Y . If they were both non-empty, they would constitute a separation for Y . Hence, one of them should be empty. Hence Y entirely lies either in C or D . □

Theorem 6.13. *The union of a collection of connected subspaces of X that have a point in common is connected.*

Proof. Let $\{A_\alpha\}$ be a collection of connected subspaces of a space X and let $\cap_\alpha A_\alpha = p$. We prove that the space $Y = \cup_\alpha A_\alpha$ is connected. If possible, let C and D form a separation for Y . Then p lies in either C or D . Without any loss of generality, let us assume that $p \in C$. For each α , A_α must lie entirely on C or D since it is connected and it can't be in D since $p \in C$. Since all $A_\alpha \in C$, so $\cup_\alpha A_\alpha \subset C$ and hence $Y \cap D = \emptyset$. We arrive at a contradiction. Hence Y is connected. □

Theorem 6.14. *Let A be a connected subspace of X and if $A \subset B \subset \overline{A}$, then B is connected.*

Proof. Let A be connected and let $A \subset B \subset \overline{A}$. Suppose that $B = C \cup D$ be a separation for B . Then, A must lie entirely in either C or D . Suppose $A \subset C$. Then $\overline{A} \subset \overline{C}$. Since $\overline{C}$ and D are disjoint, B can't intersect D . This contradicts the fact that D is a non-empty subset of B . $\square$

Corollary 6.15. *Closure of a connected set is connected in a Topological space*

Theorem 6.16. *The continuous image of a connected space is connected.*

Proof. Let $f : X \rightarrow Y$ be a continuous function and let X be connected. We show that the space $Z = f(X)$ is connected. If not, then there exists a separation $Z = C \cup D$ of Z . Since f is continuous, so $f^{-1}(C)$ and $f^{-1}(D)$ are both disjoint open sets in X , whose union is X . Then they form a disconnection for X , a contradiction. Hence $Z = f(X)$ has to be connected. $\square$

Let us consider a topological space consisting of two members $\{a, b\}$ with the discrete topology, $a, b \in \mathbb{R}$. Then the discrete two-point space is disconnected since $\{a, b\} = \{a\} \cup \{b\}$ is clearly a disconnection for the space. One can characterize a connected space as follows.

Theorem 6.17. *A space X is disconnected if and only if there is a continuous function $f : X \rightarrow \{a, b\}$, which is onto.*

Proof. Left as an exercise. $\square$

Theorem 6.18. *A topological space (X, τ) is connected if and only if given any two distinct points in X , there is a connected subspace of X containing both.*

Proof. Let (X, τ) be connected. Given any two distinct points in X , then X itself is the required subspace containing both.

Conversely, let the given condition holds. If possible, let X is disconnected. Then there exist a disconnection C, D for X such that $X = C \cup D$. Let $c \in C$ and $d \in D$. Since C and D are disjoint, so $c \neq d$. By the hypothesis, there exists a connected subspace of X , say G that contains both c and d . Clearly, $G \subset C$ or $G \subset D$. Let $G \subset C$. Then we get $d \in (C \cap D) = \emptyset$, which is absurd. Hence, X has to be connected. $\square$

6.2 Connected Sets on the Real line

We will now deal with the connected sets in the real line with the usual topology. With all the examples that we have so far seen, you must have got an idea about the connected sets on the real line. Yes. They are intervals: (a, b) , $[a, b]$, $[a, b)$, $(a, b]$, (a, ∞) , $[a, \infty)$, $(-\infty, b)$, or $(-\infty, b]$, whatever may be the form. The following theorem guarantees the fact.

Theorem 6.19. *A subset of $\mathbb{R}$ is connected if and only if it is an interval.*

Proof. Suppose E is a connected subset of $\mathbb{R}$ without being an interval. Then we find a pair of distinct members $a, b \in E$ such that $[a, b] \not\subset E$. Thus, there is a member u such that $a < u < b$ and $u \notin E$. Then we write $E = ((-\infty, u) \cap E) \cup ((u, \infty) \cap E)$ and that is a disconnection for E , a

contradiction. Hence E has to be an interval.

Conversely suppose that I be an interval of the reals. If possible, let I has a disconnection

$$I = A \cup B,$$

where, A and B are a pair of non-empty disjoint closed sets in I . Take $x \in A$ and $z \in B$. Since $A \cap B = \emptyset$, $x \neq z$, and without any loss of generality, assume that $x < z$. Because I is an interval we have the closed interval $[x, z] \subset I$. Thus

$$[x, z] \subset A \cup B$$

Put $y = \sup([x, z] \cap A)$. Then $x \leq y \leq z$, so $y \in I$. Since A is closed in I , we have,

$$y \in A \tag{6.1}$$

Therefore $y \neq z$ and we have $y < z$. By the property of supremum, for large natural numbers n , all numbers $y + \frac{1}{n}$ belong to B and since B is closed, taking limit as n tends to infinity, we have

$$\lim_{n \rightarrow \infty} \left(y + \frac{1}{n} \right) = y \in B$$

But this contradicts (6.1) because $A \cap B = \emptyset$. Hence proved. $\square$

We can say from the theorem that $\mathbb{R}$ is connected.

Exercise 6.20. 1. Let τ and σ be two topologies on X . If $\sigma \supset \tau$, what does connectedness of one topology imply about the connectedness of the other?

2. Let $\{A_n\}$ be a sequence of connected subspaces of X , such that $A_n \cap A_{n+1} \neq \emptyset$ for all n . Show that $\bigcup_n A_n$ is connected.

3. Let $\{A_\alpha\}$ be a collection of connected subspaces of X and let A be a connected subspace X . If $A \cap A_\alpha \neq \emptyset$ for all α , then show that $A \cup (\bigcup_\alpha A_\alpha)$ is connected.

Unit X

Course Structure

- Introduction
 - Objectives
 - Components and quasi components
 - Path connectedness and path components
-

Introduction

Given any arbitrary space X , there is a natural way to break it into pieces as we shall see in this unit. Those pieces are called components. We will study various components of a topological space and further characterisations of connectedness of a topological space. Let us start with the idea of components of a topological space.

7 Components

Definition 7.1. Given a topological space X , define an equivalence relation on X as follows

$x \sim y$ if and only if there is a connected subspace containing both of them.

The equivalence classes are called the components of X .

The proof that ' $\sim$ ' is an equivalence relation has been left as an exercise.

Theorem 7.2. *The components of X are connected disjoint subspaces of X whose union is X ; such that every non-empty connected subspace of X is contained in exactly one of them.*

The statement of the theorem says that the components of X partition X into disjoint sets. The proof is as follows

Proof. Since the components are equivalence classes, they are disjoint and their union is X . For the second part, let A be a connected subspace of X . If possible, let A intersect two of the components, say C and D at points, say c and d respectively. Then by the definition of $\sim$, we can say that $c \sim d$ and A is the connected subspace containing both. This can't be true unless $C = D$.

To show that a component C is connected, choose a point $x_0 \in C$. For each point $x \in C$, we know that $x_0 \sim x$, so there is a connected subspace A_x of X containing x_0 and x . By the result just proved, $A_x \subset C$. Hence,

$$C = \bigcup_{x \in C} A_x.$$

Since the subspace C is the union of connected subspaces containing a common point x_0 , their union is connected. $\square$

Theorem 7.3. *A component is a maximal connected subspace of X , that is, it is not contained in a connected subspace of X .*

Proof. If possible let a component C be properly contained in a connected subspace D of X . Then there exists a $d \in D$ such that $d \notin C$. Since X is partitioned into components, so d is contained in some other component, say E of X . Now, let $c \in C$. Since $C \subset D$, so $c \in D$. Thus, $c \sim d$ and D is the connected subspace of X containing both c and d . Also, D intersects both the components C and E which is a contradiction by the previous theorem. Hence C can't be contained in any connected subspace of X . $\square$

Theorem 7.4. *A connected set in X that is both open and closed is a component of X .*

Proof. Let E be a connected set in X which is both open and closed. Now, since each connected set is contained in a component of X , so $E \subset C$ for some component C of X . Suppose E is a proper subset of C . Then we can write

$$C = (C \cap E) \cup (C \cap (X \setminus E)).$$

Since E is both open and closed, so the above decomposition yields a disconnection for C , a contradiction. Hence, E can't be a proper subset of C and so $E = C$. $\square$

Theorem 7.5. *Every component of X is a closed set.*

Proof. Let C be a component of X without being closed. So $\overline{C}$ strictly contains C . But, we know that the closure of connected set is connected and thus it contradicts the maximality of C . Hence C is closed. $\square$

Definition 7.6. A topological space (X, τ) is said to be totally disconnected if for every pair of disjoint points $x, y \in X$, X has a disconnection $X = G \cup H$ with $x \in G$ and $y \in H$.

Does the above definition remind you of some similar definition that you have learnt earlier?

Of course a totally disconnected space is always disconnected.

Example 7.7. The real numbers with the upper limit topology is totally disconnected.

We know that the upper limit topology for $\mathbb{R}$ is generated by the left-open intervals like $(a, b]$ for $a, b \in \mathbb{R}$ with $a < b$. Let $x, y \in \mathbb{R}$ with $x < y$. Then we can write

$$\mathbb{R} = (-\infty, x] \cup (x, \infty)$$

where the sets in the r.h.s of the above equation are clearly disjoint open sets in the upper limit topology containing x and y respectively. So, $\mathbb{R}$ is totally disconnected.

Theorem 7.8. *The components for a totally disconnected space are its singletons.*

Proof. Let X be totally disconnected and C be a component of X . We show that C can't have more than one point. Let $x, y \in C$ with $x \neq y$. Since X is totally disconnected, it has a disconnection as

$$X = G \cup H$$

where G and H are disjoint non-empty open sets such that $x \in G$ and $y \in H$. We write

$$\begin{aligned} C &= C \cap X \\ &= C \cap (G \cup H) \\ &= (C \cap G) \cup (C \cap H) \end{aligned}$$

each of which are disjoint open sets in C thus yielding a disconnection for C , a contradiction. Hence the theorem. $\square$

Definition 7.9. A topological space X is said to be locally connected at $x \in X$, if every neighbourhood of x contains an open connected neighbourhood of x . And X is said to be locally connected if it is locally connected at each of its point.

It is interesting to note that neither local-connectedness implies connectedness nor conversely.

Example 7.10. If $X = (0, 1) \cup (2, 3)$ be a space with usual topology, then it is locally connected without being connected. Take a real number a with $1 < a < 2$. We write

$$X = ((-\infty, a) \cap X) \cup ((a, \infty) \cap X),$$

and this forms a disconnection for X . Also let $u \in X$ and let $0 < u < 1$, and given a neighbourhood N_u of u in X , we can find an open interval like $(u - \delta, u + \delta)$, $\delta > 0$ such that $(u - \delta, u + \delta) \subset N_u$. Since open intervals of reals with the usual topology are connected, so N_u contains an open neighbourhood of u , and hence X is locally connected at u . Since u is arbitrary, so X is locally connected at each point of $(0, 1)$. By similar argument, we can show that X is locally connected at each point of $(2, 3)$. Hence X is locally-connected without being connected.

Can you think of a space which is connected without being locally-connected?

Theorem 7.11. *A topological space X is locally-connected if and only if the components of each open subspace of X are open in X .*

Proof. Let X be locally-connected and let Y be an open subspace of X . Suppose C is a component of Y . Take $x \in C$. Since X is locally-connected at x , there is an open connected set U in X such that $x \in U \subset Y$. Now, $x \in C \cap U$, where U and C are connected and hence $C \cup U$ is connected and $C \cup U \subset Y$. Since C is a component, by maximality of C , we have $C \cup U = C$ or $U \subset C$. Hence, $x \in U \subset C$. Since x is arbitrary, so we conclude that C is open.

Conversely let the given condition holds. Let $x \in X$ and let N_x be an open neighbourhood of x in X . Take C as a component such that $x \in C \subset N_x$. By hypothesis, C is open. This shows that there is an open connected neighbourhood C of x such that $C \subset N_x$. Thus X is locally-connected at x . Since x is arbitrary, so X is locally-connected. $\square$

7.1 Path Connectedness

Let us first define path in a topological space.

Definition 7.12. Let X be a topological space and let $x, y \in X$. A path in X from x to y is a continuous map $\gamma : [a, b] \rightarrow X$ satisfying $\gamma(a) = x$ and $\gamma(b) = y$. Here, $a, b \in \mathbb{R}$ and $a < b$.

Definition 7.13. We define another equivalence relation on X as follows:

$$x \sim y \text{ if and only if there is a path between the two.}$$

Hence the equivalence classes for ' $\sim$ ' partitions X , and they are called the path-components of X .

The proof of the equivalence of the relation is easy and has been left as an exercise.

Definition 7.14. A topological space X is path-connected if for any $x, y \in X$, there exists a path from x to y .

Theorem 7.15. Let X be a path-connected space. Then X is connected.

Proof. Let X be path-connected. We will use paths in X to show that if X is not connected then $[0, 1]$ is not connected, which of course is a contradiction, so X has to be connected.

Suppose X is not connected, so we can write

$$X = U \cup V,$$

where U and V are non-empty disjoint open sets of X . Let $x \in U$ and $y \in V$. Then, there is a path $\gamma : [0, 1] \rightarrow X$ such that $\gamma(0) = x$ and $\gamma(1) = y$. The partition of X into U and V leads via this path to a partition of $[0, 1]$.

$$[0, 1] = A \cup B,$$

where, $A = \gamma^{-1}(U)$ and $B = \gamma^{-1}(V)$. Note that $0 \in A$ and $1 \in B$, so A and B are non-empty subsets of $[0, 1]$. Obviously A and B are disjoint, since no point in $[0, 1]$ can have a common γ value in both U and V . Since γ is continuous and U and V are both open in X , A and B are open in $[0, 1]$. Thus we get a disconnection for $[0, 1]$, a contradiction. Hence, X has to be connected. $\square$

The converse does not hold in general.

Example 7.16. The space

$$A = \{(x, y) \in \mathbb{R}^2 \mid x > 0, y = \sin \frac{1}{x}\} \subset \mathbb{R}^2$$

is path-connected and hence connected. So, its closure

$$\overline{A} = A \cup (\{0\} \times [-1, 1])$$

is connected. But $\overline{A}$ is not path-connected.

Though the converse is sometimes true in special cases. We have the following theorem in this direction.

Theorem 7.17. *If a non-empty open subset of $\mathbb{R}^n$ is connected, then it is path-connected.*

We will omit the proof of the theorem.

Theorem 7.18. *The path components of a topological space X are path connected disjoint subspaces of X whose union is X , such that every non-empty path-connected subspace intersects only one of them.*

Proof. Since the path components are equivalence classes of the equivalence relation for paths, so each of the components are disjoint subspaces whose union is X . Let A be a non-empty path-connected subspace of X . If possible, let A intersects two path-components P and S of X , say at points p and s respectively. Then, since A is path-connected, and s and p are in A , so there exists a path between s and p . But by the definition of path-components, this is impossible unless $P = S$. Thus A can intersect only one of the path-components.

Let P be a path-component of X . Then by the definition of path components, for any two distinct points of P , there exists a path between the two. Then it is obviously path-connected. $\square$

We saw in case of components that they are closed. But they need not be open. But in case of path-components, we can say nothing about this. They need be neither open nor closed.

Example 7.19. In the subspace of rationals $\mathbb{Q}$ of $\mathbb{R}$, each component consists of a single point and hence is not open in $\mathbb{Q}$.

Definition 7.20. A space X is said to be locally path-connected at a point $x \in X$ if for every neighbourhood U of x , there exists a path-connected neighbourhood V of x contained in U . If X is locally path-connected at each of its points, then it is called locally path-connected.

Theorem 7.21. *A space X is locally path-connected if and only if for every open set U of X , each path component of U is open in X .*

Proof. The proof is similar as that in the case of local connectedness and has been left as an exercise. $\square$

Theorem 7.22. *If X is a topological space, each path component of X lies in a component of X .*

Proof. Let C be a component of X and let $x \in C$. Let P be the path component of X containing x . Since P is connected, so by maximality of C , $P \subset C$. Hence proved. $\square$

Theorem 7.23. *If X is locally path-connected, then the components and path-components are the same.*

Proof. Let C be a component and P be a path-component as in the previous theorem. Then by the previous theorem, we have

$$P \subset C.$$

If possible, let $P \neq C$. Let Q denotes the union of all path-components of X that are different from P and intersect C ; each of them necessarily lie in C , so that

$$C = P \cup Q.$$

Since X is locally path-connected, each path-component of X is open in X . Hence, P and Q (which is the union of path-components) are open in X , so they constitute a disconnection of C , a contradiction. Thus P has to be equal to C . Hence the theorem. $\square$

7.2 Quasicomponents

Definition 7.24. Let X be a space. We define a relation ' $\sim$ ' as follows: $x \sim y$ if and only if there is no separation

$$X = A \cup B,$$

where A and B are disjoint open sets in X such that $x \in A$ and $y \in B$. Then it is an equivalence relation on X and the equivalence classes are called quasicomponents of X .

Theorem 7.25. In a space X , each component lies in a quasicomponent of X .

Proof. If X is connected then the result is trivially true. Let X be not connected and let C be a component of X . Let $c \in C$ and also let Q be the quasicomponent of X containing c . We show that $C \subset Q$. If not, then there exists at least a pair of elements $x, y \in C$ for which there exists a disconnection

$$X = A \cup B$$

such that $x \in A$ and $y \in B$. Then, we will get

$$\begin{aligned} C &= (C \cap X) \\ &= C \cap (A \cup B) \\ &= (C \cap A) \cup (C \cap B) \end{aligned}$$

which yields a disconnection for C , a contradiction. Hence our assumption is wrong and C has to be in Q . $\square$

- Exercise 7.26.**
1. Examine if the real number space with the lower limit topology is connected.
 2. Show that the set of irrational numbers with the usual topology is disconnected.
 3. Show that a totally disconnected space is T_2 .
 4. Show that continuous image of a locally connected space may not be so.
 5. Show that any ball in $\mathbb{R}^2$ with usual topology is connected.

Unit XI

Course Structure

- Introduction
 - Objectives
 - Compactness
 - Sequential Compactness
-

Introduction

As we have seen, the closed interval $[a, b]$ has certain crucial properties which has many applications such as the maximum value theorems. But for a long time it was not clear how to formulate it in any arbitrary topological space. Mathematicians formulated it in terms of open coverings of the space. This formulation is what is called compactness. We do have preliminary ideas about compactness when we learned real analysis and metric spaces. We are very familiar with the Heine-Borel theorem which says that any closed and bounded set is compact in the real line $\mathbb{R}$. In fact, the theorem is true for any $\mathbb{R}^n$. By virtue of this, it can be said that $[a, b]$ is compact in $\mathbb{R}$. Let's look at it in a more general way, that is, for any arbitrary topological space and let's find out whether we can connect with whatever we have learned earlier.

Objectives

After reading this unit, you will be able to

- learn compactness in a more general setting
- relate the definitions with the earlier ones that you have learnt
- learn various characterizations and properties of compact spaces
- see certain examples of compact sets in familiar spaces
- learn about sequential compactness
- learn about various applications of compact spaces

8 Compact Spaces

Before starting with the definitions of compact spaces, let's go with the tradition of defining open covers, and subcovers first.

Definition 8.1. A collection $\mathcal{A}$ of subsets of a space X is said to cover X if X is equal to the union of the elements of $\mathcal{A}$. And $\mathcal{A}$ is called open cover if each of the elements of $\mathcal{A}$ are open sets of X .

Definition 8.2. A finite subfamily of $\mathcal{A}$ that also covers X is called a finite subcover for X .

Example 8.3. For example, the family of open intervals $\mathcal{A} = \{(-n, n) | n \in \mathbb{N}\}$, forms an open cover for $\mathbb{R}$. Also the subfamily $\mathcal{A}_e = \{(-2n, 2n) | n \in \mathbb{N}\}$ forms a subcover for $\mathbb{R}$.

We now define compactness of a space X .

Definition 8.4. A topological space (X, τ) is said to be compact if each open cover of X has a finite subcover.

Example 8.5. Let X be an infinite set with the co-finite topology. Then X is compact. Let

$$\mathcal{A} = \{U_\alpha | \alpha \in I\}$$

be an open cover for X . Take any $U_\alpha \in \mathcal{A}$. Then, by the definition of co-finite topology, $X \setminus U_\alpha$ is a finite set, say

$$X \setminus U_\alpha = \{a_1, a_2, \dots, a_n\}$$

Since $\mathcal{A}$ covers X so for each $i = 1(1)n$, there exists a $U_{\alpha_i} \in \mathcal{A}$ such that $a_i \in U_{\alpha_i}$. So clearly, the finite subcollection $\{U_\alpha, U_{\alpha_1}, U_{\alpha_2}, \dots, U_{\alpha_n}\}$ covers X . Hence X with the co-finite topology is compact.

Definition 8.6. A subset A of a topological space X is said to be compact if it is compact as a subspace of X with the topology induced by the topology on X .

Example 8.7. 1. The real line with the usual topology is not compact since the open cover

$$\mathcal{A} = \{(n, n+2) | n \in \mathbb{Z}\}$$

does not have any finite subcover.

2. The interval $(0, 1]$ is not compact. The open covering

$$\mathcal{A} = \{(1/n, 1] | n \in \mathbb{Z}_+\}$$

contains no finite subcover for $(0, 1]$.

3. The closed interval $[0, 1]$ is compact and it is by the well known Heine Borel theorem for $\mathbb{R}$.

Lemma 8.8. Let A be a subspace of X . Then A is compact if and only if every open covering of A by the sets open in X contains a finite subcollection covering A .

Proof. Let A be compact and let $\mathcal{A} = \{A_\alpha | \alpha \in I\}$ be an open covering for X by open sets in X . Then the family

$$\{A_\alpha \cap A | \alpha \in I\}$$

is an open covering for X by open sets in A . Then there exists a finite subcollection $\{A_{\alpha_1} \cap A, A_{\alpha_2} \cap A, \dots, A_{\alpha_n} \cap A\}$ which covers A . Thus, the subcollection $\{A_{\alpha_1}, A_{\alpha_2}, \dots, A_{\alpha_n}\}$ covers A by means of open sets in A .

Conversely, suppose that the given condition holds. We prove that A is compact. Let $\{B_\alpha | \alpha \in J\}$ be an open cover for A by open sets in A . For each α , choose a set A_α open in X such that

$$B_\alpha = A \cap A_\alpha$$

Then the collection $\{A_\alpha | \alpha \in J\}$ is an open cover for A by open sets in X . By hypothesis, there exists a finite subcollection of $\{A_\alpha | \alpha \in J\}$, say $\{A_{\alpha_1}, A_{\alpha_2}, \dots, A_{\alpha_n}\}$ that covers A . Thus the subcollection $\{B_{\alpha_1}, B_{\alpha_2}, \dots, B_{\alpha_n}\}$ covers A . Thus, A is compact. $\square$

The usefulness of the above lemma lies in the fact that whenever we are dealing with the compactness of a subspace, then it becomes redundant to differentiate between open cover by the open sets in X or in A .

Theorem 8.9. *Every closed subspace of a compact space is compact.*

Proof. Let C be a closed subspace of a compact topological space X . Let $\mathcal{A}$ be an open cover of C by open sets in X . Then the family of open sets in $\mathcal{A}$ along with the open set $X \setminus C$ forms an open cover of X . Since X is compact, so the said family has a finite subcover say

$$\{G_1, G_2, \dots, G_n\} \cup (X \setminus C)$$

of X . Thus, the subfamily

$$\{G_1, G_2, \dots, G_n\}$$

clearly covers C . Thus, C is compact. $\square$

Theorem 8.10. *Every compact subspace of a Hausdorff space is closed.*

Proof. Let Y be a compact subspace of a Hausdorff space X . We will prove that $X \setminus Y$ is open. Let $a \in X \setminus Y$. For each point $y \in Y$, we will get disjoint open neighbourhoods U_y and V_y of the points a and y respectively. The collection $\{V_y | y \in Y\}$ is a cover of Y by open sets in X , and hence we will get a finite subfamily of it, say $\{V_{y_1}, V_{y_2}, \dots, V_{y_n}\}$ that cover Y . The open set

$$V = V_{y_1} \cup V_{y_2} \cup \dots \cup V_{y_n}$$

contains Y and it is disjoint from the open set

$$U = U_{y_1} \cap U_{y_2} \cap \dots \cap U_{y_n}$$

formed by taking the corresponding intersection of the open neighbourhoods of a . For if $z \in V$, then $z \in V_{y_i}$ for each i and hence $z \notin U_{y_i}$ for each i and hence $z \notin U$. Thus U is the required neighbourhood of a that is disjoint from Y , that is $U \subset X \setminus Y$. Hence $X \setminus Y$ is open and thus, Y is closed. $\square$

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

Lemma 8.11. *If Y is a compact subspace of a Hausdorff space X and $a \notin Y$. Then there exist disjoint open sets U and V of X containing a and Y respectively.*

Theorem 8.12. *The image of a compact space under a continuous map is compact.*

Proof. Let $f : X \rightarrow Y$ be continuous and X be compact. Let $\mathcal{A}$ be an open covering of the set $f(X)$ by sets open in Y . Then the collection

$$\{f^{-1}(A) | A \in \mathcal{A}\}$$

is a collection of open sets covering X . Thus, due to the compactness of X , there is a finite subfamily, say

$$\{f^{-1}(A_i) | i = 1(1)n\}$$

that covers X . Then the subfamily $\{A_1, A_2, \dots, A_n\}$ that cover $f(X)$. $\square$

An important application of the above theorem can be stated as the following theorem.

Theorem 8.13. *Let $f : X \rightarrow Y$ be a bijective continuous map where X is compact and Y is Hausdorff. Then f is a homeomorphism.*

Proof. We will prove that the images of a closed sets of X under f are closed in Y . If A is closed in X , then its compact. Thus, by the previous theorem, $f(A)$ is compact. Since Y is Hausdorff, so $f(A)$ is closed in Y . $\square$

We are now heading to prove another equivalent definition of compactness. First we will define the Finite Intersection Property (FIP) of a collection of subsets of a space X .

Definition 8.14. A collection $\mathcal{C}$ of subsets of X is said to have FIP, if the intersection of a finite subfamily of $\mathcal{C}$ has non-empty intersection.

The following theorem gives us an equivalent definition of compactness in terms of FIP.

Theorem 8.15. *A topological space X is compact if and only if for every collection $\mathcal{C}$ of closed sets in X having FIP, the intersection $\bigcap_{C \in \mathcal{C}} C$ is non-empty.*

Proof. Given a collection $\mathcal{A}$ of subsets of X , let

$$\mathcal{C} = \{X - A | A \in \mathcal{A}\}$$

be the collection of their complements. Then the following statements hold:

- $\mathcal{A}$ is a collection of open sets if and only if $\mathcal{C}$ is a collection of closed sets.
- The collection $\mathcal{A}$ covers X if and only if the intersection $\bigcap_{C \in \mathcal{C}} C$ of all the elements of $\mathcal{C}$ is empty.
- The finite subcollection $\{A_1, A_2, \dots, A_n\}$ of $\mathcal{A}$ covers X if and only if the intersection of the corresponding elements $C_i = X - A_i$ of $\mathcal{C}$ is empty.

The first statement is trivial, while the second and third follow from DeMorgan's law:

$$X - (\cup_{\alpha \in J} A_{\alpha}) = \cap_{\alpha \in J} (X - A_{\alpha}).$$

The proof of the theorem now proceeds in two easy steps: taking the contrapositive (of the theorem), and then the complement(of the steps).

The statement that X is compact is equivalent to saying: "Given any collection $\mathcal{A}$ of open subsets of X , if $\mathcal{A}$ covers X , then some finite subcollection of $\mathcal{A}$ covers X ."

This statement is equivalent to its contrapositive, which is the following: "Given any collection $\mathcal{A}$ of open sets, if no finite subcollection of $\mathcal{A}$ covers X , then $\mathcal{A}$ does not cover X ." Letting $\mathcal{C}$ be, as earlier, the collection $\{X - A | A \in \mathcal{A}\}$ and applying to all the three above bullets, we see that this statement is in turn equivalent to the following:

"Given any collection $\mathcal{C}$ of closed sets, if every finite intersection of elements of $\mathcal{C}$ is nonempty, then the intersection of all the elements of $\mathcal{C}$ is nonempty." This is just the condition of our theorem. $\square$

A special case of this theorem occurs when we have a nested sequence $C_1 \supset C_2 \supset \dots \supset C_n \supset C_{n+1} \supset \dots$ of closed sets in a compact space X . If each of the sets C_n is nonempty, then the collection $\mathcal{C} = \{C_n\}_{n \in \mathbb{Z}_+}$ automatically has the finite intersection

$$\cap_{n \in \mathbb{Z}_+} C_n$$

is nonempty.

We shall use the closed set criterion for the compactness in the next section to prove the uncountability of the set of real numbers.

8.1 Lebesgue Lemma

The concept of Lebesgue number is new and highly useful for an open covering of a metric space. First lets recapitulate some preliminary definitions.

Definition 8.16. Let (X, d) be a metric space and let A be a non-empty subset of X . Then for each $x \in X$, the distance of x from A is defined as

$$d(x, A) = \inf\{d(x, a) | a \in A\}$$

It is easy to show that the function $d(x, A)$ is continuous.

Also recall the diameter of a set A in X , which is defined as

$$\text{diam} A = \sup\{d(a_1, a_2) | a_1, a_2 \in A\}$$

Then a set A is called bounded if $\text{diam}(A)$ is finite.

We will now state a lemma which is known as the Lebesgue number lemma.

Lemma 8.17. *Let $\mathcal{A}$ be an open covering of the metric space (X, d) . If X is compact, there is a $\delta > 0$ such that for each subset of X having diameter less than δ , there exists an element of $\mathcal{A}$ containing it.*

This number δ is called the Lebesgue number for the covering $\mathcal{A}$.

Proof. Let $\mathcal{A}$ be an open cover of X . If X is itself an element of $\mathcal{A}$, then any positive number is a Lebesgue number for $\mathcal{A}$. So, we assume that X is not an element of $\mathcal{A}$. Choose a finite subcollection $\{A_1, A_2, \dots, A_n\}$ of $\mathcal{A}$ that covers X . For each i , set

$$C_i = X \setminus A_i,$$

and define $f : X \rightarrow \mathbb{R}$ by letting

$$f(x) = \frac{d(x, C_1) + d(x, C_2) + \dots + d(x, C_n)}{n}$$

We show that $f(x) > 0$ for all x . Given $x \in X$, choose i so that $x \in A_i$. Then choose ϵ so that the ϵ -neighbourhood of x lies in A_i . Then $d(x, C_i) \geq \epsilon$, so that $f(x) \geq \frac{\epsilon}{n}$.

Since f is continuous, it has a minimum value δ . We show that this δ is our required Lebesgue number. Let B be a subset of X of diameter less than δ . Choose a point a of B . Then B lies in the δ -neighbourhood of a . Now

$$\delta \leq f(a) \leq d(a, C_m),$$

where $d(a, C_m)$ is the largest of the numbers $d(a, C_i)$. Then the δ -neighbourhood of a is contained in the element $A_m = X \setminus C_m$ of the cover $\mathcal{A}$. $\square$

8.2 Limit Point Compactness

Definition 8.18. A space X is said to be limit point compact if every infinite subset of X has a limit point.

Theorem 8.19. *Compactness implies limit point compactness.*

Proof. Let X be a compact set and let A be a subset of X . Suppose A has no limit point. Then we show that A is finite.

Since A has no limit point, so A is closed. Further, for each $a \in A$, we can choose a neighbourhood U_a of a that intersects A at the point a alone. The space X is covered by the open set $X \setminus A$ along with the open sets U_a . Since X is compact, it can be covered by finitely many of these sets. Since $X \setminus A$ does not intersect A , and each U_a contains only one point of A , the set A must be finite. Hence contrapositively, we have proved the theorem. $\square$

The converse of the theorem is not necessarily true.

Example 8.20. Let Y consists of two points endowed with the indiscrete topology. The space $X = \mathbb{Z} \times Y$ is limit point compact, for every non-empty subset of X has a limit point. But it is not compact since for the covering of X by open sets of the form

$$U_n = \{n\} \times Y$$

there is no finite subcover for X .

We will now move to define the sequential compactness of a space. But before that, let us see the convergence of sequences in a topological space.

Definition 8.21. 1. Let X be a topological space and let $\{x_n\}$ be a sequence of points in X . Then $\{x_n\}$ is said to converge to a point $a \in X$, if for every open set U containing a , we can find $m \in \mathbb{N}$ such that $x_n \in U$ for all $n \geq m$.

2. If

$$n_1 < n_2 < \cdots < n_k < \cdots$$

is an increasing sequence of natural numbers, then the sequence $\{x_{n_k}\}$ is called a subsequence of the sequence $\{x_n\}$.

Definition 8.22. A topological space is said to be sequentially compact if every sequence of points of X has a convergent subsequence.

Example 8.23. 1. Let A be a finite subset of a space X . Let $\{x_n\}$ be a sequence of points in A . Since A is finite, then at least one element of the sequence, say x_k has to appear infinite number of times in the sequence. Thus, if we construct a subsequence as $\{x_k, x_k, \dots, x_k, \dots\}$, then it surely converges to the point x_k of A . Thus, A is sequentially compact.

2. The open interval $A = (0, 1)$ with the usual topology is not sequentially compact. Since the sequence of points $\{1/2, 1/3, 1/4, \dots, 1/n, \dots\}$ clearly does not have any subsequence that converges to a point of A .

Now, for any arbitrary metric space, the concept of sequential compactness and compactness are the same. But this is not true for any arbitrary topological space. In fact, for a topological space, neither implies the other.

Exercise 8.24. 1. Show that a finite union of compact subspaces of X is compact.

2. Show that if $f : X \rightarrow Y$ is a continuous map where X is compact and Y is Hausdorff, then f is a closed mapping.
3. Show that every subspace of $\mathbb{R}$ with the co-finite topology is compact.
4. Show that sequential compactness may not necessarily imply compactness.
5. Show that compactness may not necessarily imply sequential compactness.

Unit XII

Course Structure

- Introduction
 - Objectives
 - BW compactness and Countable compactness
 - Local compactness
 - Baire Category Theorem
-

Introduction

In this section we will learn still further types of compactness and their relationship with compactness and sequential compactness. In the previous unit, we saw that sequential compactness and compactness are equivalent in any arbitrary metric space but we can't conclude anything in case of any arbitrary topological space. In this unit also we will see the relationship between the definitions in a metric space as well as topological space.

Objectives

After reading this unit, you will be able to

- define countable compactness
- define BW compactness
- define locally compact sets
- relate them with compactness and sequential compactness
- understand the explicit relationships between these definitions in topological spaces
- relate them in case of metric spaces
- know various properties relating to them
- show various examples for each type
- state the Baire Category theorem

9 Countable Compactness

Definition 9.1. A topological space X is called countably compact if every countable open cover of X has a finite subcover. A family of sets $\mathcal{A}$ is called a countable open cover of X if

- it covers X
- it has countable number of elements
- each element of $\mathcal{A}$ is open

The definition can also be given as the following theorem.

Theorem 9.2. *A topological space X is countably compact if and only if each sequence has an accumulation point.*

From the definition, it is obvious that any compact space is always countably compact. Also, any sequentially compact space is also countably compact. But the converse is not true in both the cases.

Example 9.3. Consider the natural number space $\mathbb{N}$ with the topology generated by

$$\{1, 2\}, \{3, 4\}, \{5, 6\}, \dots$$

Let A be a non-empty subset of $\mathbb{N}$ and let $m \in A$. If a is odd, then $a + 1$ is a limit point of A and if a is even, then $a - 1$ is a limit point of A . In either case, A has an accumulation point. So, $\mathbb{N}$ is countably compact. But it is not compact since the open cover

$$\{1, 2\}, \{3, 4\}, \{5, 6\}, \dots$$

clearly has no finite subcover. Also, it is not sequentially compact since the sequence $\{1, 2, 3, \dots\}$ contains no convergent subsequence.

Theorem 9.4. *A countably compact space is limit point compact.*

Proof. Without any loss of generality, assume that A be a countably infinite subset of a countably compact topological space X without any limit points. Then, A is closed in X . So for each $a \in A$, the open set U_a containing a intersects A only at a . Let $V_a = U_a \cup (X \setminus A)$. Then V_a is an open set. Still $V_a \cap A = \{a\}$. Also, $\{V_a | a \in A\}$ is a countable cover of X . Thus it has a finite subcover say $\{V_{a_1}, V_{a_2}, \dots, V_{a_n}\}$ of X . But then

$$\begin{aligned} A &= \left(\bigcup_{i=1}^n V_{a_i} \right) \cap A \\ &= \bigcup_{i=1}^n (V_{a_i} \cap A) \\ &= \{a_1, a_2, \dots, a_n\} \end{aligned}$$

which shows that A is finite, a contradiction. Hence the result. □

Theorem 9.5. *A space that is T_1 and limit point compact is countably compact.*

Proof. Suppose that X is not countably compact. So there is a countable open cover $\{U_n | n \in \mathbb{N}\}$ without a finite subcover. For each i , pick $x_i \in X \setminus (U_1 \cup U_2 \cup \dots \cup U_i)$. This is possible since $\{U_n\}$ does not have a finite subcover for X . Let

$$A = \{x_1, x_2, \dots\}$$

Suppose now that x is in X . Then $x \in U_N$ for some $N \in \mathbb{N}$ (since $\{U_n\}$ covers X). This U_N can only contain x_i for $i < N$ by definition of x_i 's. So, U_N is a neighbourhood of x that contains only finite number of points of A . Since X is T_1 , this implies that x can't be a limit point of A . Since $x \in X$ was arbitrary, so no point of X can be a limit point of A , which is a contradiction since X is limit point compact. Hence, X has to be countably compact. $\square$

We have another equivalent definition for countable compactness given in the form of the following theorem.

Theorem 9.6. *A topological space is countably compact if and only if every nested sequence $C_1 \supset C_2 \supset \dots$ of closed non-empty sets of X has a non-empty intersection.*

Proof. Left as an exercise. $\square$

We will now look at a new type of topological space called a Bolzano Weierstrass space which we define below.

Definition 9.7. A topological space X is called Bolzano-Weierstrass Space (often called BW-space) if every infinite subset of X has a limit point.

Example 9.8. The real line with the usual topology is not a BW-space. Since the subset $\mathbb{Z}$ contains no limit point in $\mathbb{R}$.

Theorem 9.9. *If a space X is compact, then it is BW.*

Proof. Let X be a compact space. Then every open cover of X has a finite subcover. If possible, let X is not a BW-space. Then there exists an infinite subset A of X that does not have a limit point, that is, $A' = \emptyset$. So A is closed. So, A is compact in X . Further, since A has no limit point, for each $a \in A$, there exists an open neighbourhood U_a of a , such that

$$A \cap U_a = \{a\}$$

So, $\{U_a | a \in A\}$ is an open cover for A . Since A is compact, there is a finite subcover, say $\{U_{a_1}, U_{a_2}, \dots, U_{a_n}\}$ which covers A . But $\bigcup_{i=1}^n U_{a_i}$ contains only n points, which implies that A is a finite set, a contradiction. Hence X has to be a BW-space. $\square$

Does this proof have any resemblance to that we did for countably compact spaces? So it certainly must have something to do with countably compact spaces. Lets look at the following theorem.

Theorem 9.10. *If X is Hausdorff, then X is BW if and only if it is countably compact.*

9.1 Local Compactness

Definition 9.11. A space X is said to be locally compact if each point in X has a compact neighbourhood.

If X is itself compact, then it is locally compact and in that case X itself acts as a compact neighbourhood of each of its points. If the topology is discrete and X is infinite, then X is locally compact without being compact (verify!).

Example 9.12. The real number space $\mathbb{R}$ with the usual topology is locally compact without being compact. Since, for $x \in \mathbb{R}$, the closed interval $[x - \delta, x + \delta]$, $\delta > 0$ is a neighbourhood of x which is compact.

Theorem 9.13. Let X be a Locally compact T_2 -space

Proof. Let E be a closed set of X which is locally compact. We will show that E with the subspace topology induced by the topology on X , is locally compact. Let $x \in E$. Since X is locally compact, so we can find a compact neighbourhood N of x in X . Put $M = N \cap E$. So, M is a neighbourhood of x in E and as M is a closed set in a compact space N , so M is a compact subset of N , that is, M is a compact neighbourhood of x in E . Thus, E is locally compact. $\square$

Compact Hausdorff spaces are one of the "good" spaces to work with since they have many useful properties. If a given space is not so, then the next best thing that one can hope for is that whether it is a subspace of one such space, that is to say, whether it is homeomorphic to the subspace of such a space. If so, then under what conditions is it possible to define such a homeomorphism. This section mainly deals with the problem of embedding a given space into a Compact Topological space. We start with the following theorem.

Theorem 9.14. A space X is locally compact Hausdorff if and only if there exists a space Y such that

1. X is a subspace of Y .
2. The set $Y \setminus X$ consists of a single point.
3. Y is a compact Hausdorff space.

If Y and Y' are two spaces satisfying the above conditions, then there is a homeomorphism of Y and Y' that equals the identity map on X .

If X itself happens to be compact, then the space Y is not so interesting since it is formed by adjoining a single isolated point to X . However, if X is not compact, then the point of $Y \setminus X$ is a limit point of X and $\overline{X} = Y$. If Y is a compact Hausdorff space and X is a proper subspace of Y whose closure equals Y , then Y is said to be a *compactification* of X and if $Y \setminus X$ is a singleton set, then Y is called a *one-point compactification* of X .

Theorem 9.15. Let X be a Hausdorff space. Then it is locally compact if and only if given $x \in X$ and a neighbourhood U of x , there is a neighbourhood V of x such that $\overline{V}$ is compact and $\overline{V} \subset U$.

Proof. Suppose X is locally compact and let $x \in X$ and U be a neighbourhood of x . Take the one-point compactification Y of X and let $C = Y \setminus U$. Then C is closed in Y and hence compact subspace of Y . Then, by a previous lemma, we can find disjoint open sets V and W containing x and C respectively. Then the closure $\overline{V}$ of V in Y is compact and moreover, $\overline{V}$ is disjoint from C , so that $\overline{V} \subset U$. Hence the result. $\square$

9.2 Baire Spaces

Definition 9.16. We have learnt about the interior of a set A as the union of all the open sets contained in A . If A has an empty interior, then it contains no open set of X other than the emptyset. Remember that

$$\overline{(X \setminus A)} = X \setminus \text{Int}A$$

So, equivalently, A has empty interior if every point of A is a limit point of the complement of A , that is, if the complement of A is dense in X .

Definition 9.17. A space X is said to be a Baire space if the following condition holds: Given any countable collection $\{A_n\}$ of closed sets of X each of which has an empty interior in X , their union $\bigcup A_n$ also has an empty interior in X .

Example 9.18. The space $\mathbb{Q}$ of rationals is not a Baire space since each one point set in $\mathbb{Q}$ is closed and has an empty interior in $\mathbb{Q}$ and $\mathbb{Q}$ is the countable union of these one-point sets. Whereas, the set $\mathbb{Z}_+$ forms a Baire space. Every subset of $\mathbb{Z}_+$ is open and hence have non-empty interior. Thus, $\mathbb{Z}_+$ satisfies the Baire condition trivially.

Theorem 9.19. (Baire Category Theorem:) *If X is a compact Hausdorff space, or a complete metric space, then X is a Baire space.*

Proof. Given a countable collection $\{A_n\}$ of closed sets of X having empty interiors, we will show that $\bigcup A_n$ also has an empty interior in X . So, given the non-empty open set U_0 of X , we must find a point x of U_0 that does not lie in any of the sets A_n .

Consider the first set A_1 . By hypothesis, A_1 does not contain U_0 . Hence, we may choose a point y of U_0 that is not in A_1 . Regularity of X , along with the fact that A_1 is closed, enables us to choose a neighbourhood U_1 of y such that

$$U_1 \cap A_1 = \emptyset, \overline{U_1} \subset U_0.$$

If X is a metric, then we choose U_1 small enough that its diameter is less than 1.

In general, given the non-empty open set U_{n-1} , we choose a point of U_{n-1} that is not in the closed set A_n , and then we choose U_n to be a neighbourhood of this point such that

$$\overline{U_n} \cap A_n = \emptyset, \overline{U_n} \subset U_{n-1}$$

such that $\text{diam}U_n < 1/n$ in the metric case. We assert that the intersection $\bigcap \overline{U_n}$ is non-empty. From this fact, our theorem will follow. For, if $x \in \bigcap \overline{U_n}$, then x is in U_0 since $\overline{U_1} \subset U_0$. And for each n , the point x is not in A_n because $\overline{U_n}$ is disjoint from A_n .

If X is a compact Hausdorff space, then we consider the nested sequence

$$\overline{U_1} \supset \overline{U_2} \supset \cdots$$

of non-empty subsets of X . The collection $\{\overline{U_n}\}$ has the finite intersection property. Since X is compact, $\bigcap \overline{U_n}$ must be non-empty.

If on the other hand, X is a complete metric, then we use the Cantor's Intersection theorem to complete the proof. $\square$

Exercise 9.20. 1. Show that a sequentially compact space is countably compact.

2. Show that an infinite set with the discrete topology is compact without being locally compact.

3. Show that any open subspace of a Baire space is also a Baire space.

4. Show that a space X is countably compact if and only if every nested sequence

$$C_1 \supset C_2 \supset \dots \supset C_n \dots$$

of closed non-empty sets of X has a non-empty intersection.

5. Show that a space X is limit point compact if and only if each sequence in X has a limit point.

6. Let X be limit point compact space and let A be a closed subset of X . Is A limit point compact?

Unit XIII

Course Structure

- Constructing a Möbius strip
 - Identification topology
 - Orbit spaces
-

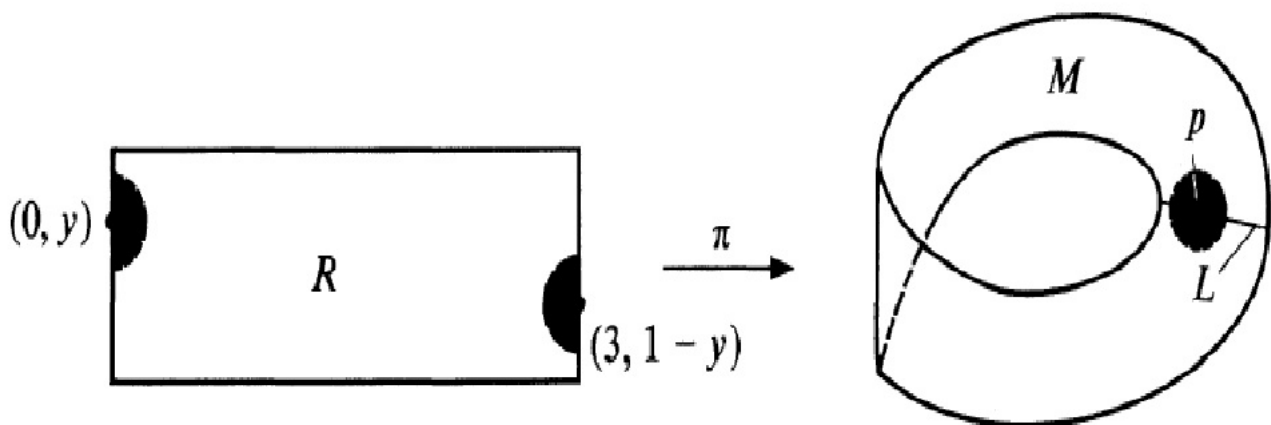
10 Constructing a Möbius Strip

To construct a Möbius strip, one takes a rectangle and identifies a pair of opposite edges with a half twist. Our first job is to translate this process into precise mathematical language. For the rectangle take the subspace R of $\mathbb{R}^2$ consisting of the points (x, y) for which $0 \leq x \leq 3$ and $0 \leq y \leq 1$. To describe the identification of the vertical edges of R with a half twist, we partition R into disjoint non-empty sub sets in such a way that two points lie in the same subset if and only if we wish them to be identified. If we now take these sub sets as the points of our Möbius strip, then we have made the required identifications. The appropriate partition of R consists of:

1. sets consisting of a pair of points of the form $(0, y), (3, 1 - y)$, where $0 \leq y \leq 1$.
2. sets consisting of a single point (x, y) , where $0 < x < 3, 0 \leq y \leq 1$.

So far we have defined a set which we shall call M , its points being the subsets of the above partition of R . There is a natural function π from R onto M that sends each point of R to the subset of the partition in which it lies. The identification topology on M is defined to be the largest topology for which π is continuous. That is to say, a subset O is defined to be open in the identification topology on M if and only if $\pi^{-1}(O)$ is open in the rectangle R .

A glance at the figure above shows the sort of open sets we obtain. We represent the points of M in the usual way as a sub set of $\mathbb{R}^3$, and we label with the letter L the image under π of the two vertical edges of R . If we use R_* to denote R minus its vertical edges, then the restriction of π to R_* is one-one and is a homeomorphism of R_* with $M \setminus L$. Therefore, we know all about the neighbourhoods of points of $M \setminus L$; they are simply the images under π of neighbourhoods of the points of R_* . If p lies on the line L , then $\pi^{-1}(p)$ consists of two distinct points, situated on the vertical edges of R , of the form $(0, y), (3, 1 - y)$. The union of two open half-discs in R , centres $(0, y), (3, 1 - y)$ and of equal radius, maps via π to an open neighbourhood of p in the identification topology on M . Notice that if we take a single half-disc, its image in M is not a neighbourhood of p and is not open, so π is not an open mapping. The points of L are in no sense special in the Möbius strip; they have the same sort of neighbourhoods in the identification topology as all the other points of M . In fact,



it is easy to check that the identification topology coincides with that induced from $\mathbb{R}$ on our set M .

For convenience we have illustrated M in $\mathbb{R}^3$. However, we emphasize that the definition of the Möbius strip as an 'identification space' given in this section is entirely abstract, and in no way relies on a particular representation of the strip as a set of points in euclidean space.

10.1 The Identification Topology

Let X be a topological space and let $\mathcal{P}$ be a family of disjoint non-empty subsets of X such that $\cup \mathcal{P} = X$. Such a family is usually called a partition of X . We form a new space Y , called an identification space, as follows:

The points of Y are the members of $\mathcal{P}$ and, if $\pi : X \rightarrow Y$ sends each point of X to the subset of $\mathcal{P}$ containing it, the topology of Y is the largest for which π is continuous. Hence, a subset of Y is open if and only if $\pi^{-1}(O)$ is open in X . This topology is called the identification topology on Y . We think of Y as a space obtained from X by identifying each of the subsets of $\mathcal{P}$ to a single point.

Our construction of the Möbius strip was a special case of this procedure. We will now prove one or two general results on identification spaces. We begin with a theorem that is useful when checking the continuity of a function which has an identification space as domain.

Theorem 10.1. *Let Y be an identification space defined as above and let Z be an arbitrary topological space. A function $f : Y \rightarrow Z$ is continuous if and only if the composition $f\pi : X \rightarrow Z$ is continuous.*

Proof. Let U be an open subset of Z . Then $f^{-1}(U)$ is open in Y if and only if $\pi^{-1}(f^{-1}(U))$ is open in X , that is, if and only if $(f\pi)^{-1}(U)$ is open in X . $\square$

Let $f : X \rightarrow Y$ be an onto map and suppose that the topology on Y is the largest for which f is continuous. Then we call f an identification map, the reason for our terminology being as follows. Any function $f : X \rightarrow Y$ gives rise to a partition of X whose members are the subsets $\{f^{-1}(y)\}$,

where $y \in Y$. Let Y_* denote the identification space associated with this partition, and $\pi : X \rightarrow Y_*$ the usual map.

Theorem 10.2. *If f is an identification map, then*

1. *the spaces Y and Y_* are homeomorphic;*
2. *a function $g : Y \rightarrow Z$ is continuous if and only if the composition $gf : X \rightarrow Z$ is continuous.*

Proof. The proof of 2 is exactly that of the previous theorem since Y has the largest topology for which f is continuous. The points of Y_* are the sets $\{f^{-1}(y)\}$, where $y \in Y$. Define $h : Y_* \rightarrow Y$ by $h(\{f^{-1}(y)\}) = y$. Then h is a bijection and satisfies $h\pi = f$, $h^{-1}f = \pi$. By the previous theorem, h is continuous, and h^{-1} is continuous by 2. Hence h is a homeomorphism. $\square$

Theorem 10.3. *Let $f : X \rightarrow Y$ be an onto map. If f maps open sets of X to open sets of Y , or closed sets to closed sets, then f is an identification map.*

Proof. Suppose f maps open sets to open sets. Let U be a subset of Y for which $f^{-1}(U)$ is open in X . Since f is onto, we have $f(f^{-1}(U)) = U$, and therefore U must be open in the given topology on Y . So this topology is the largest for which f is continuous, and f is an identification map. The proof for closed maps is similar. $\square$

Corollary 10.4. *Let $f : X \rightarrow Y$ be an onto map. If X is compact and Y is Hausdorff, then f is an identification map.*

Proof. A closed subset of the compact space X is compact and its image under the continuous function f is therefore a compact subset of Y . But a compact subset of a Hausdorff space is closed. Therefore f takes closed sets to closed sets, and we can apply the previous theorem. $\square$

We shall use the previous theorem and corollary in order to compare different descriptions of the same topological space. We begin with two methods of constructing a torus.

The Torus. Take X to be the unit square $[0, 1] \times [0, 1]$ in $\mathbb{R}^2$, with the subspace topology, and partition X into the following subsets:

1. the set $\{(0, 0), (1, 0), (0, 1), (1, 1)\}$ of four corner points;
2. sets consisting of pairs of points $(x, 0), (x, 1)$, where $0 < x < 1$;
3. sets consisting of pairs of points $(0, y), (1, y)$, where $0 < y < 1$;
4. sets consisting of a single point (x, y) , where $0 < x < 1$ and $0 < y < 1$;

The resulting identification space is the torus. An equally common description is to say that the torus is the product $S^1 \times S^1$ of two circles. Thinking of the points of S^1 as complex numbers, we can define a map $f : [0, 1] \times [0, 1] \rightarrow S^1 \times S^1$ by $f(x, y) = (e^{2\pi ix}, e^{2\pi iy})$. The partition of $[0, 1] \times [0, 1]$ which contains the inverse images under f of the points of $S^1 \times S^1$ is exactly that given earlier. By the previous corollary, f is an identification map and hence our two descriptions of the torus are homeomorphic.

Let us recollect the gluing lemma.

Theorem 10.5. *If X and Y are closed in $X \cup Y$, and if both f and g are continuous on X and Y respectively, with a common co-domain Z such that $f(x) = g(x)$ when $x \in X \cap Y$, then the map $h : X \cup Y \rightarrow Z$, defined as*

$$\begin{aligned} h(x) &= f(x), \quad x \in X \\ &= g(x), \quad x \in Y \end{aligned}$$

is continuous.

The glueing lemma remains true if we ask that X and Y are both open in $X \cup Y$. The lemma is also true for open sets. But, it fails if no restrictions are placed on the sets X and Y .

As we shall see, the glueing lemma can be explained in terms of identification maps and interpreted as a special case of the previous theorem. In order to do this, we introduce the disjoint union $X + Y$ of the spaces X, Y and the function $j : X + Y \rightarrow X \cup Y$ which when restricted to either X or Y is just the inclusion in $X \cup Y$. The function is important for our purpose because

1. it is continuous;
2. the composition $hj : X + Y \rightarrow Z$ is continuous if and only if f and g are continuous;

By combining parts 1 and 2 of theorem 10.2, we have the following theorem.

Theorem 10.6. *If j is an identification map, and if both $f : X \rightarrow Z$ and $g : Y \rightarrow Z$ are continuous, then $h(= f \cup g) : X \cup Y \rightarrow Z$ is continuous.*

The glueing lemma is a special case of this result, since if both X and Y are closed in $X \cup Y$, then j sends closed sets to closed sets and is an identification map by theorem 10.3.

If j is an identification map, then then we can think of $X \cup Y$ as an identification space formed from the disjoint union $X + Y$ by identifying certain points of X with points of Y . In this case, we often say that $X \cup Y$ has the identification topology. The open(closed) sets of $X \cup Y$ are those sets A for which $A \cap X$ and $A \cap Y$ are open (closed) sets of X and Y respectively.

The above theorem generalizes to the case of an arbitrary union. Let $X_\alpha, \alpha \in A$ be a family of subsets of a topological space and give each X_α , and the union $\cup X_\alpha$, the induced topology. Let Z be a space and suppose we are given maps $f_\alpha : X_\alpha \rightarrow Z$, one for each $\alpha \in A$, such that if $\alpha, \beta \in A$,

$$f_\alpha|_{X_\alpha \cap X_\beta} = f_\beta|_{X_\alpha \cap X_\beta}$$

Define a function $F : \cup X_\alpha \rightarrow Z$ by gluing together the f_α , that is,

$$F(x) = f_\alpha(x), \quad x \in X_\alpha.$$

Let $\boxplus X_\alpha$ denote the disjoint union of the spaces X_α , and let $j : \boxplus X_\alpha \rightarrow \cup X_\alpha$ be the function which when restricted to each X_α is the inclusion in $\cup X_\alpha$.

Theorem 10.7. *If j is an identification map, and if each f_α is continuous, then F is continuous.*

Proof. Observe that $Fj : \boxplus X_\alpha \rightarrow Z$ is continuous if and only if each f_α is continuous and apply part 2 of theorem 10.2. $\square$

Attaching maps. As a final example of an identification space we formalize the notion of attaching one space to another by means of a continuous function.

Let X, Y be spaces, let A be a subspace of Y , and let $f : A \rightarrow X$ be a continuous function. Our aim is to attach Y to X using f and to form a new space which we shall denote by $X \cup_f Y$. We begin with the disjoint union $X + Y$ and define a partition so that two points lie in the same subset if and only if they are identified under f . precisely, the subsets of the partition are

1. pairs of points $\{a, f(a)\}$ where $a \in A$;
2. individual points of $Y \setminus A$;
3. the individual points of $X \setminus \text{image}(f)$.

The identification space associated with this partition is $X \cup_f Y$. The map f is called the attaching map.

One final comment: if Y is an identification space formed from X , then Y is the image of X under a continuous function and therefore inherits properties such as compactness, connectedness, and path-connectedness from X . However, X may be Hausdorff and yet Y not satisfy the Hausdorff axiom. As an example, take X to be the real line with the usual topology, and partition X so that real numbers r and s lie in the same elements of the partition if and only if $r - s$ is rational. Then the corresponding identification is an indiscrete space.

10.2 Orbit Spaces

We will now consider spaces which have, in addition to their topology, the structure of a group. A good example is the circle.

Definition 10.8. A topological group G is both a Hausdorff space and a group, the two structures being compatible in the sense that the group multiplication $m : G \times G \rightarrow G$, and the function $i : G \rightarrow G$ which sends each group element to its inverse, are continuous.

Example 10.9. 1. The real line, the group structure being addition of real numbers.

2. The circle S^1 , thought of as the set of complex numbers of unit modulus. Then the two functions

$$S^1 \times S^1 \rightarrow S^1, (e^{i\theta}, e^{i\phi}) \mapsto e^{i(\theta+\phi)}$$

and

$$S^1 \rightarrow S^1, e^{i\theta} \mapsto e^{-i\theta}$$

are continuous.

3. Any abstract group with the discrete topology.
4. The Euclidean n -space.

5. The group of invertible $n \times n$ matrices with real entries. The group structure is matrix multiplication. For the topology we identify each $n \times n$ matrix $A = (a_{ij})$ with the corresponding point

$$(a_{11}, a_{12}, \dots, a_{1n}, a_{21}, \dots, a_{2n}, \dots, a_{n1}, \dots, a_{nn})$$

of $\mathbb{R}^{n^2}$ and take the subspace topology. This topological group is called the general linear group, denoted by $GL(n, \mathbb{R})$.

6. The orthogonal group $O(n)$ consisting of $n \times n$ orthogonal matrices with real entries. $O(n)$ has both its topology and its group structure induced from $GL(n, \mathbb{R})$.

The terms 'isomorphism' and 'subgroup' for topological groups require a few words of explanation. In each case we need to take into consideration both the topological and the algebraic structures. So an isomorphism between two topological groups is a homeomorphism which is also a group isomorphism. In the same spirit, a subset of a topological group is called a subgroup if it is algebraically a subgroup and in addition has the subspace topology. Hence, the integers $\mathbb{Z}$ with the discrete topology form a subspace of the real line $\mathbb{R}$.

Definition 10.10. A topological group G is said to act as a group of homeomorphisms on a space X if each group element induces a homeomorphism of the space in such a way that;

1. $hg(x) = h(g(x))$ for all $g, h \in G$ and all $x \in X$.
2. $e(x) = x$ for all $x \in X$, where e is the identity element of G .
3. the function $G \times X \rightarrow X$ defined by $(g, x) \mapsto g(x)$ is continuous.

If $x \in X$, then for each $g \in G$, the corresponding homeomorphism either fixes x or maps it to a new point $g(x)$. The subset of X consisting of all such images $g(x)$ as g varies over G is called the orbit of x and written as $O(x)$. If two orbits intersect then they must coincide: the relation defined by $x \sim y$ if and only if $x = g(y)$ for some $g \in G$ is an equivalence relation on X whose equivalence classes are precisely the orbits of the given action. So the orbits define a partition of X . The corresponding identification space is called the orbit space written as X/G . In constructing X/G , we 'divide' by G in the sense that we identify two points of X if and only if they differ by one of the homeomorphisms $x \mapsto g(x)$.

Example 10.11. The orbit of the real number x consists of all the points $x + n$ where $n \in \mathbb{Z}$. Therefore in forming $\mathbb{R}/\mathbb{Z}$, we identify two points of $\mathbb{R}$ if and only if they differ by an integer and we obtain the circle as orbit space.

Example 10.12. Taking the product of our first example with itself in the natural way gives an action of $\mathbb{Z} \times \mathbb{Z}$ on the plane. An ordered pair of integers $(m, n) \in \mathbb{Z} \times \mathbb{Z}$ sends the point $(x, y) \in \mathbb{R}^2$ to $(x + m, y + n)$. The orbit space is the product of two circles, in other words the torus. It may help to think of this action geometrically. Divide the plane into squares of unit side by drawing in all horizontal and vertical lines through the points with integer coordinates. The homeomorphisms of our group action preserve this pattern of squares, and any single square contains points from each orbit and therefore maps onto the torus under the identification map

$$\mathbb{R}^2 \xrightarrow{\pi} \mathbb{R}^2/\mathbb{Z} \times \mathbb{Z} = T$$

Each square has its sides identified by π in the usual way in order to form T .

Theorem 10.13. *Let G act on X and suppose that both G and X/G are connected. Then X is connected.*

Proof. Suppose X is the union of two disjoint non-empty open subsets U and V . Since the identification map $\pi : X \rightarrow X/G$ always takes open sets to open sets, and since X/G is connected, $\pi(U)$ and $\pi(V)$ can't be disjoint. Now, if $x \in \pi(U) \cap \pi(V)$, then both $U \cap O(x)$ and $V \cap O(x)$ are non-empty. These two sets decompose the orbit $O(x)$ as a disjoint union of two non-empty open sets. But $O(x)$ is the image of G under the continuous function $f : G \rightarrow X$ defined by $f(g) = g(x)$. $O(x)$ is therefore connected, and we have established the required contradiction. $\square$

Unit XIV

Course Structure

- Introduction
 - Some Elementary properties of Topological groups
-

Introduction

This unit can be thought of as a continuation of the previous unit. It deals with primarily the topological groups along with certain properties of topological groups.

11 Elementary Properties of Topological Groups

Let G be a topological group and x an element of G . The function $L_x : G \rightarrow G$ defined by $L_x(g) = xg$ is called the left translation by the element x . It is clearly one-one and onto, and it is continuous because it is the composition

$$G \rightarrow G \times G \xrightarrow{m} G$$

as $g \mapsto (x, g) \mapsto xg$. The inverse of L_x is L_x^{-1} and therefore L_x is a homeomorphism. Similarly the right translation $R_x : G \rightarrow G$ given by $R_x(g) = gx$ is also a homeomorphism. Thus, if we have $U \subset G$ and $x \in G$, then

$$U \text{ open} \iff tU \text{ open} \iff Ut \text{ open} \iff U^{-1} \text{ open}.$$

These translations show that a topological group has a certain 'homogeneity' as a topological space. For if x and y are any two points of a topological group G there is a homeomorphism of G that maps x to y , namely the translation $L_{yx^{-1}}$. Hence G exhibits the same topological structure locally near each point.

Theorem 11.1. *Let G be a topological group and let K denote the connected component of G which contains the identity element. Then K is a closed normal subgroup of G .*

Proof. Components are always closed as we have seen previously. For any $x \in K$, the set $Kx^{-1} = R_x^{-1}(K)$ is connected (since R_x^{-1} is a homeomorphism) and contains $e = xx^{-1}$. Since K is the maximal connected subset of G containing e , we must have $Kx^{-1} \subseteq K$. Hence $KK^{-1} = K$, and K is a subgroup of G . Normality follows in a similar manner. For any $g \in G$ the set $gKg^{-1} = R_g^{-1}L_g(K)$ is connected and contains e . Hence $gKg^{-1} \subseteq K$. $\square$

Theorem 11.2. *In a connected topological group any neighbourhood of the identity element is a set of generators for the whole group.*

Proof. Let G be a connected topological group and let V be a neighbourhood of e in G . Let $H = \langle V \rangle$ be the subgroup of G generated by the elements of V . If $h \in H$ then the whole neighbourhood $hV = L_h(V)$ of h lies in H , so H is open. We claim that the complement of H is also open. For if $g \in G \setminus H$, consider the set gV . If $gV \cap H$ is non-empty, say $x \in gV \cap H$, then $x = gv$ for some $v \in V$. This gives $g = xv^{-1}$, which implies the contradiction $g \in H$ since both x and v^{-1} lie in H . Hence the neighbourhood $L_g(V) = gV$ of g lies in $G \setminus H$ and we see that $G \setminus H$ is an open set. Now, G is connected and so cannot be partitioned into two disjoint non-empty open sets. Since H is non-empty, we must have $G \setminus H = \emptyset$, that is, $G = H$. $\square$

If G is a group and S and T are subsets of G , we let ST and S^{-1} denote

$$ST = \{st : s \in S, t \in T\} \text{ and } S^{-1} = \{s^{-1} : s \in S\}.$$

The subset S is called symmetric if $S^{-1} = S$. Also, we denote e as the identity element of G .

Theorem 11.3. *Let G be a topological group. Every neighbourhood U of e contains an open symmetric neighbourhood V of e such that $VV \subset U$.*

Proof. Let $\text{Int}U$ be the interior of U . Consider the multiplication map

$$\mu : \text{Int}U \times \text{Int}U \rightarrow G.$$

Since μ is continuous, then $\mu^{-1}(\text{Int}U)$ is open and contains (e, e) . So, there are open sets $V_1, V_2 \in U$ such that $(e, e) \in V_1 \times V_2$, and $V_1V_2 \in U$. Let $V_3 = V_1 \cap V_2$, then $V_3V_3 \subset U$ and V_3 is an open neighbourhood of e . Finally, let $V = V_3 \cap V_3^{-1}$, which is open, contains e , is symmetric, and satisfies $VV \subset U$. $\square$

Theorem 11.4. *If G is a topological group, then every open subgroup of G is also closed.*

Proof. Let H be an open subgroup of G . Then any coset xH is also open. So,

$$Y = \bigcup_{x \in G \setminus H} xH$$

is also open. From elementary group theory,

$$H = G \setminus Y,$$

and so H is closed. $\square$

Theorem 11.5. *If G is a topological group, and if K_1 and K_2 are compact subsets of G , then K_1K_2 is compact.*

Proof. The set $K_1 \times K_2$ is compact in $G \times G$, and the multiplication mapping is continuous. Since the continuous image of a compact set is compact, so K_1K_2 is compact. $\square$

Recall the closure of a set A in a topological space X . We have studied that a necessary and sufficient condition for $x \in \overline{A}$ is that, for every open neighbourhood U of x , the set $U \cap A \neq \emptyset$. It can also be seen as follows:

Theorem 11.6. *If $x \notin \overline{A}$, then there is a closed set F which contains A , but $x \notin F$.*

Proof. Left as an exercise. □

Theorem 11.7. *If G is a topological group, and H is a subgroup of G , then $\overline{H}$ is a subgroup of G .*

Proof. Let $g, h \in \overline{H}$, and let U be an open neighbourhood of the product gh . Let

$$\mu : G \times G \rightarrow G$$

be the multiplication map, which is continuous. So, $\mu^{-1}(U)$ is open in $G \times G$, and contains (g, h) . So, there are open neighbourhoods V_1 of g and V_2 of h such that $V_1 \times V_2 \subset \mu^{-1}(U)$. Since $g, h \in \overline{H}$, then there are points $x \in V_1 \cap \overline{H}$ and $y \in V_2 \cap \overline{H}$. Since $x, y \in H$, we have $xy \in H$ and since $(x, y) \in \mu^{-1}(U)$, then $xy \in U$. Thus, $xy \in U \cap H$, and since U was an arbitrary open neighbourhood of gh , then we have $gh \in \overline{H}$.

Now, let

$$i : G \rightarrow G$$

denote the inverse map, and let W be an open neighbourhood of h^{-1} . Then, $i^{-1}(W) = W^{-1}$ is open and contains h . So there is a point $z \in H \cap W^{-1}$. Then $z^{-1} \in H \cap W$ and as before, this implies that $h^{-1} \in \overline{H}$. Thus, $\overline{H}$ forms a subgroup of G . □

In the course of proving the above theorem, we have shown that the closure of a symmetric neighbourhood of e is again symmetric.

Lemma 11.8. *Let G be a topological group, F a closed subset of G , and K a compact subset of G , such that $F \cap K = \emptyset$. Then there is an open neighbourhood V of e such that $F \cap VK = \emptyset$ (and an open neighbourhood V' of e such that $F \cap KV' = \emptyset$).*

Proof. Let $x \in K$, so $x \in G \setminus F$ and $G \setminus F$ is open. So, $(G \setminus F)x^{-1}$ is an open neighbourhood of e . By theorem 11.3, there is an open neighbourhood W_x of e such that $W_x W_x \subset (G \setminus F)x^{-1}$. Now,

$$K \subset \bigcup_{x \in K} W_x x,$$

and K is compact, so there exists a finite number of points $x_1, x_2, \dots, x_n \in K$, such that

$$K \subset \bigcup_{i=1}^n W_i x_i$$

where $W_i = W_{x_i}$. Now let

$$V = \bigcap_{i=1}^n W_i.$$

For any $x \in K$, $x \in W_i x_i$, for some i . Now we have

$$Vx \subset W_i x \subset W_i W_i x_i \subset G \setminus F.$$

In other words, $F \cap Vx = \emptyset$. Since this is true for any $x \in K$, we have,

$$F \cap VK = \emptyset.$$

□

From theorem 11.3, the neighbourhood V in above lemma can be taken to be symmetric.

Theorem 11.9. *Let G be a topological group, K a compact subset of G , and F a closed subset of G . Then FK and KF are closed subsets of G .*

Proof. If $FK = G$, then the result is trivial. So, let $y \in G \setminus FK$. This means that $F \cap yK^{-1} = \emptyset$. Since K is compact, yK^{-1} is compact. So, by the previous lemma, there is an open neighbourhood V of e such that $F \cap VyK^{-1} = \emptyset$, or $FK \cap Vy = \emptyset$. Since V_y is an open neighbourhood of y contained in $G \setminus FK$, we have FK is closed. Similarly, we can show that KF is closed. □

11.1 Separation properties and functions

You are quite familiar with the definitions of T_1 and T_2 spaces by now. We will investigate in this section, the relationship between these separation axioms and a topological group. If G is a topological group, then if G is T_1 , then by homogeneity, $\{e\}$ is a closed set in G and conversely.

We have seen that, if a space X is T_2 , then it is T_1 but the converse is not true in general. But, we see that the converse is true in case of topological groups.

Theorem 11.10. *Let G be a T_1 topological group. Then G is Hausdorff.*

Proof. Let g, h be distinct elements of G . By the T_1 axiom, let U be an open set containing e , such that $gh^{-1} \notin U$. By theorem 11.3, let V be an open symmetric neighbourhood containing e , such that $VV \subset U$. Now, Vg is open and contains g , and Vh is open containing h . We must have $Vg \cap Vh = \emptyset$, otherwise there are $v_1, v_2 \in V$ such that

$$v_1g = v_2h,$$

which would mean

$$gh^{-1} = v_2v_1^{-1} \in VV^{-1} = VV \subset U,$$

while gh^{-1} was chosen to be an element not in U . Thus, G is Hausdorff. □

You might remember the regular and completely regular spaces as well as Tychonoff space. We learnt that a completely regular space is always regular. We will now see that any topological group which is T_1 is also completely regular, and thus, regular.

Theorem 11.11. *Let G be a topological group, and let e_G denote the identity element in G and let F be a closed subset of G such that $e_G \notin F$. Then there is a continuous function $f : G \rightarrow [0, 1]$ such that*

$$f(e_G) = 1, \text{ and } f(y) = 0, y \in F.$$

Theorem 11.12. *If G is a topological group which is T_1 , then G is completely regular and thus, regular.*

Proof. Let $x \in G$ and let F be a closed subset of G such that $x \notin F$. Then $x^{-1}F$ is a closed subset of G not containing e_G , and thus from the previous theorem, there is a continuous function $f : G \rightarrow [0, 1]$ such that $f(e_G) = 0$ and $f(y) = 1$ for $y \in x^{-1}F$. Now the function $h(g) = f(x^{-1}g)$ is the desired continuous function, and since G is also T_1 , G is completely regular, and hence is also regular. $\square$

11.2 Connectedness

We have certain important properties of a topological group concerning connectedness. Let us study them in some details.

Theorem 11.13. *The connected component of the identity in a topological group is a subgroup.*

Proof. Let G_0 be the connected component of G containing the identity and $h, k \in G_0$ be arbitrary. The set $h^{-1}G_0$ is connected and contains the identity and so $G_0 \cup h^{-1}G_0$ is also connected. Since G_0 is a component, we have $G_0 \cup h^{-1}G_0 = G_0$ which implies that $h^{-1}G_0 \subset G_0$. In particular, $h^{-1}k$ belongs to G_0 from which we can conclude that G_0 is a subgroup. $\square$

Theorem 11.14. *Suppose that G is a topological group and K is a subgroup and the coset space G/K is given the quotient topology. Then*

1. *If K and G/K are connected, then G is connected.*
2. *If K and G/K are compact, then G is compact.*

Proof. If G is connected then so is G/K since the quotient map $\eta : G \rightarrow G/K$ is a continuous surjection. To prove the converse, suppose that K and G/K are connected and $f : G \rightarrow \{0, 1\}$ be an arbitrary continuous map. We have to show that f is constant. The restriction of f to K must be constant and since each coset gK is connected, f must be constant on gK as well taking value $f(g)$. Thus, we have a well-defined map $f' : G/K \rightarrow \{0, 1\}$ such that $f' \circ \eta = f$. By the fundamental property of quotient spaces, it follows that f' is continuous and so must be constant since G/K is connected. $\square$

Theorem 11.15. *If G is a connected topological group and H is a subgroup which contains a neighbourhood of the identity then $H = G$. In particular, an open subgroup of G equals G .*

Proof. Let U be the open neighbourhood of the identity that is contained in H and $h \in H$ be arbitrary. Since multiplication by h is a homeomorphism, the set $Uh = \{uh \mid u \in U\}$ is also open and also contained in H . Hence the set

$$L = \bigcup_{h \in H} Uh$$

is open and contained in H . Since U contains the identity, $H \subset L$ and we conclude that H is open. Our job will be over if we can show that H is closed as well. Let $x \in \overline{H}$ be arbitrary. Since the neighbourhood Ux of x contains a point $y \in H$, there exists $u \in U$ such that $y = ux$ which, in view of the fact that $U \subset H$, implies $x \in H$. Hence $\overline{H} = H$. $\square$

Theorem 11.16. *Suppose G is a connected topological group and H is a discrete normal subgroup of G then H is contained in the center of G .*

Proof. Since H is discrete, the identity element is not a limit point of H and so there is a neighbourhood U of the identity such that $U \cap H = \{e\}$. We may assume U has the property that if $u_1, u_2 \in U$, then the product $u_1^{-1}u_2 \in U$. This follows from the continuity of the group operation and a detailed verification is left as an exercise. It is easy to see that if h_1 and h_2 are two distinct elements of H , then

$$Uh_1 \cap Uh_2 = \emptyset.$$

Fix $h \in H$ and consider now the set K given by

$$K = \{g \in G \mid gh = hg\}.$$

We shall show that the subgroup K contains a neighbourhood of the identity. Pick a neighbourhood V of the identity such that $V = V^{-1}$ and that $(hVh^{-1}V) \cap H = \{e\}$. Then for any $g \in V$, we have on the one hand

$$hgh^{-1}g^{-1} \in hVh^{-1}V$$

and on the other hand, $hgh^{-1}g^{-1} \in H$ since H is normal. Hence

$$hgh^{-1}g^{-1} \in (hVh^{-1}V) \cap H = \{e\}$$

which shows that g belongs to K and K contains a neighbourhood of the unit element. We may now invoke the previous theorem. $\square$

Unit XV

Course Structure

- Introduction
 - Objectives
 - $GL(n, \mathbb{R})$ as a topological group
 - Its subgroups
-

Introduction

Though we have already read a bit about $GL(n, \mathbb{R})$ previously, we will read about it in more details in this unit.

12 The Group $GL(n, \mathbb{R})$

We will first show that $GL(n, \mathbb{R})$ actually forms a topological group.

Theorem 12.1. *The matrix group $GL(n, \mathbb{R})$ is a topological group.*

Proof. Let M denote the set of all $n \times n$ matrices which have real entries, and let $A = (a_{ij})$ represent a typical element of M . We can identify M with euclidean space of dimension n^2 by associating $A = (a_{ij})$ with the point $(a_{11}, a_{12}, \dots, a_{1n}, a_{21}, \dots, a_{2n}, a_{n1}, \dots, a_{nn})$. The identification gives us a topology on M and we claim that, with respect to this topology, matrix multiplication $m : M \times M \rightarrow M$ is continuous. To see this, we need only examine the well-known formula for the entries of a product matrix: If $A = (a_{ij})$ and $B = (b_{ij})$ then the ij th entry in the product $m(A, B)$ is $\sum_{k=1}^n a_{ik}b_{kj}$. Now, M has the topology of the product space

$$\mathbb{R} \times \mathbb{R} \times \dots \times \mathbb{R} \text{ (} n^2 \text{ copies),}$$

and for each i, j satisfying $1 \leq i, j \leq n$ we have a projection

$$\pi_{ij} : M \rightarrow \mathbb{R}$$

which sends a given matrix A to its ij th entry. By the property of continuous functions, we can say that m is continuous if and only if all the composite functions

$$M \times M \xrightarrow{m} M \xrightarrow{\pi_{ij}} \mathbb{R}$$

are continuous. But, $\pi_{ij}m(A, B) = \sum_{k=1}^n a_{ik}b_{kj}$, a polynomial in the entries of A and B . Hence $\pi_{ij}m$ is continuous.

The elements of $GL(n, \mathbb{R})$ are invertible matrices in M . If we give $GL(n, \mathbb{R})$ the subspace topology from M , then, by the above, matrix multiplication

$$GL(n, \mathbb{R}) \times GL(n, \mathbb{R}) \rightarrow GL(n, \mathbb{R})$$

is continuous. It remains to prove that the inverse function

$$i : GL(n, \mathbb{R}) \rightarrow GL(n, \mathbb{R})$$

is also continuous. We use the same technique:

$$i : GL(n, \mathbb{R}) \rightarrow GL(n, \mathbb{R}) \subseteq \mathbb{R} \times \mathbb{R} \times \cdots \times \mathbb{R}$$

is continuous if and only if all of the composite functions

$$GL(n, \mathbb{R}) \xrightarrow{i} GL(n, \mathbb{R}) \xrightarrow{\pi_{jk}} \mathbb{R}, \quad 1 \leq j, k \leq n$$

are continuous. Now the composition of π_{jk} with i sends a matrix A to the jk th element of A^{-1} , that is, to $(1/\det A)(kj$ th cofactor of $A)$. It should be clear that the determinant of A and the cofactors of A are polynomials in the entries of A . Since $\det A$ does not vanish on $GL(n, \mathbb{R})$, our composition $\pi_{jk}i$ is continuous. This completes the proof that $GL(n, \mathbb{R})$ is a topological group. $\square$

We note in passing that $GL(n, \mathbb{R})$ is the inverse image of the nonzero real numbers under the determinant function

$$\det : M \rightarrow \mathbb{R}.$$

So, $GL(n, \mathbb{R})$ is not compact (it is an open subset of M), and is not connected (the matrices with positive and negative determinants partition $GL(n, \mathbb{R})$ into two disjoint non-empty open sets). Let us do it in some details.

Theorem 12.2. *$GL(n, \mathbb{R})$ is open and unbounded.*

Proof. The complement of $GL(n, \mathbb{R})$ in M is

$$\{A \in M \mid \det A = 0\}.$$

Since the determinant is continuous and $\{0\}$ is closed in $\mathbb{R}$, so the above set is closed in M , and hence its complement, that is, $GL(n, \mathbb{R})$ is open in M . Also, $kI_n \in GL(n, \mathbb{R})$ for all $k > 0$. So, $GL(n, \mathbb{R})$ is unbounded. Thus, proved. $\square$

Theorem 12.3. *$GL(n, \mathbb{R})$ is connected.*

Proof. Note that,

$$\det : GL(n, \mathbb{R}) \rightarrow \mathbb{R} \setminus \{0\}$$

is a surjective continuous map and $\mathbb{R} \setminus \{0\}$ is not connected and since we know that a continuous map sends a connected set into a connected set, so $GL(n, \mathbb{R})$ can't be connected. $\square$

In fact, as we have said earlier, we can partition $GL(n, \mathbb{R})$ into two disjoint open sets, namely,

$$\begin{aligned} GL_+(n, \mathbb{R}) &= \{A \in GL(n, \mathbb{R}) : \det A > 0\} \\ GL_-(n, \mathbb{R}) &= \{A \in GL(n, \mathbb{R}) : \det A < 0\} \end{aligned}$$

Since the determinant is a continuous function, so the above sets are open. Hence, there are two (path) components of $GL(n, \mathbb{R})$.

12.1 Subgroups of $GL(n, \mathbb{R})$

We have already seen that the set of all orthogonal matrices $O(n)$ forms a subgroup of $GL(n, \mathbb{R})$. In this section, we will study more about the subgroups of $GL(n, \mathbb{R})$. Let us list some of the subgroups of $GL(n, \mathbb{R})$ as follows:

1. $SL(n, \mathbb{R}) = \{A \in GL(n, \mathbb{R}) : \det A = 1\}$. This group is called the special linear group. Note that, since the determinant function is continuous, so $\det^{-1}(\{1\}) = SL(n, \mathbb{R})$ is a closed set in $GL(n, \mathbb{R})$.
2. $O(n) = \{A \in GL(n, \mathbb{R}) : AA^T = I_n = A^T A\}$. As we have already seen, this is the orthogonal group.
3. $SO(n) = \{A \in O(n) : \det A = 1\}$. This is called the special orthogonal group.

From the elementary properties of matrices, we can say that each of the above groups actually form a subgroup of $GL(n, \mathbb{R})$ and has been left as an exercise. If instead of $\mathbb{R}$, we take the field of complex, that is, if we consider the set $GL(n, \mathbb{C})$ then we have some additional subgroups, which are listed below:

1. $U(n) = \{U \in GL(n, \mathbb{C}) : UU^* = I_n = U^*U\}$, where, U^* denotes the conjugate transpose of U . This is called the unitary group.
2. $SU(n) = \{U \in U(n) : \det(U) = 1\}$. This particular group is called the special unitary group.

Let us now see certain properties of these subgroups.

Theorem 12.4. *The groups $O(n)$ and $SO(n)$ are compact.*

Proof. Write any matrix $A \in O(n)$ as

$$A = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

where, each v_i is a row matrix. Then from the identity $AA^T = I_n$, we get,

$$v_i v_i^T = 1, \quad 1 \leq i \leq n.$$

This implies that A is inside the unit ball of $\mathbb{R}^{n^2}$. Hence $O(n)$ is a bounded subset of the Euclidean space $\mathbb{R}^{n^2}$. Let $\{A_k\}$ be any sequence in $O(n)$ and let $A_k \rightarrow A$ in M . Taking limit as $k \rightarrow \infty$ in the relation

$$A_k A_k^T = A_k^T A_k = I_n,$$

by continuity of multiplication, we get

$$AA^T = A^T A = I_n$$

proving that $A \in O(n)$. Thus, $O(n)$ is closed too. Hence by Heine-Borel theorem, $O(n)$ is compact. If, in addition, each of the matrices A_k above have the determinant 1, then by the continuity of the determinant, we also see that $SO(n)$ is closed in $O(n)$. And since we know that a closed subset of a compact space is compact, so $SO(n)$ is compact. $\square$

Before going into other details, let us see the following theorem.

Theorem 12.5. *Let $P \in M$ be a positive semidefinite matrix. Then P is symmetric, $\det P \geq 0$ and there exists a unique positive definite matrix $P^{1/2} \in M$ such that*

$$(P^{1/2})^2 = P$$

and $P^{1/2}$ is invertible if and only if P is so.

Theorem 12.6. *Let $A \in SL(n, \mathbb{R})$. Then there exists a rotation matrix $R \in SO(n)$ and a real, symmetric and positive semidefinite matrix $P \in SL(n, \mathbb{R})$ such that $A = RP$.*

Proof. Let $P = (A^T A)^{1/2}$ and this forces R to be defined as

$$R = AP^{-1}.$$

Clearly, P is real, symmetric and positive semidefinite matrix. Further,

$$\begin{aligned} RR^T &= AP^{-1}P^{-1}A^T \\ &= AP^{-2}A^T \\ &= A(A^T A)^{-1}A^T \\ &= I_n \end{aligned}$$

and similarly, $R^T R = I_n$ implying that R is orthogonal. In particular,

$$1 = \det(R^T R) = \det(R^T) \det R = (\det R)^2$$

which implies that $\det R = 1$ or -1 . Now,

$$1 = \det A = \det R \det P$$

and so by the previous theorem, and the fact that P is invertible, we have $\det P > 0$. Hence, we must have $\det R = 1$, that is, $R \in SO(n)$. $\square$

Theorem 12.7. *Any matrix in $SO(n)$ is orthogonally similar to a block diagonal form $A_1 \boxplus A_2 \boxplus \cdots \boxplus A_r$, where, each A_r is $[1]$ or a rotation matrix of the type*

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

for some $\theta \in \mathbb{R}$.

Theorem 12.8. *$O(n)$ is not connected, whereas $SO(n)$ is path-connected.*

Proof. Let $M \in O(n)$. Then, by the above theorem, $\det M \in \{+1, -1\}$. Let

$$O_{\pm}(n) = \{M \in O(n) : \det M = \pm 1\} = GL_{\pm}(n, \mathbb{R}) \cap O(n).$$

Then, $O_+(n)$, which is the same as the subgroup $SO(n)$ and $O_-(n)$ are open in the subspace topology and they form a disconnection of $O(n)$, implying that $O(n)$ is not connected.

Next we show that any matrix in $SO(n)$ is joined to I_n by a path. Since $SO(1) = \{[1]\}$, let us assume that $n \geq 2$. Let $R \in SO(n)$. Then by a previous theorem, there exists an orthogonal matrix $M \in O(n)$ such that

$$MRM^T = A_1 \boxplus A_2 \boxplus \cdots \boxplus A_r$$

where each A_i is $[1]$ or a 2×2 rotation matrix of the type

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

for some $\theta \in \mathbb{R}$. Without any loss of generality, assume that, for some $k \leq r$, for $1 \leq i \leq k$,

$$A_i = \begin{bmatrix} \cos \theta_i & \sin \theta_i \\ -\sin \theta_i & \cos \theta_i \end{bmatrix}$$

for some $\theta_i \in \mathbb{R}$ and that $A_i = [1]$ for $k \leq i \leq r$.

We can now look for an appropriate path. For each $1 \leq i \leq k$, consider the map

$$\phi_i : [0, 1] \rightarrow SO(2)$$

as

$$\phi_i(t) = \begin{bmatrix} \cos(t\theta_i) & \sin(t\theta_i) \\ -\sin(t\theta_i) & \cos(t\theta_i) \end{bmatrix}.$$

Thus, each ϕ_i is a path in $SO(2)$ with end points I_2 and

$$\begin{bmatrix} \cos \theta_i & \sin \theta_i \\ -\sin \theta_i & \cos \theta_i \end{bmatrix}.$$

Hence, the map $\phi : [0, 1] \rightarrow SO(n)$ given by

$$\phi(t) = M^T(\phi_1(t) \boxplus \phi_2(t) \boxplus \cdots \boxplus \phi_k(t) \boxplus I_{n-2k})M$$

is a path in $SO(n)$ with end points $\phi(0) = I_n$ and $\phi(1) = R$, thereby establishing that $SO(n)$ is path-connected. $\square$

We thus get the following corollary.

Corollary 12.9. $O(n)$ has precisely two path-components, namely $O_+(n)$ and $O_-(n)$.

Proof. By the previous theorem, we get that

$$O_+(n) = SO(n)$$

is path-connected. Now, let $A, B \in O_-(n)$ and fix a $C \in O_-(n)$. Then

$$AC, BC \in O_+(n)$$

and therefore, there exists a path ϕ in $O_+(n)$ joining AC and BC . Consider the map

$$\tilde{\phi} : [0, 1] \rightarrow O_-(n)$$

given by

$$\tilde{\phi}(t) = \phi(t)C^{-1}.$$

This $\tilde{\phi}$ is a path in $O_-(n)$ joining A and B .

We also know that $O(n)$ is a disjoint union of $O_+(n)$ and $O_-(n)$, so these are the only two path-components of $O(n)$. $\square$

Note that $SL(1, \mathbb{R}) = \{[1]\}$ is clearly path-connected and compact. However, in higher dimensions, this is not the case.

Corollary 12.10. *$SL(n, \mathbb{R})$ is closed, path-connected and is not compact for $n \geq 2$.*

Unit XVI

Course Structure

- Introduction
 - Objectives
 - Fundamental Groups
 - Preliminary Properties
-

Introduction

One of the main problems of topological spaces is to determine whether two spaces are homeomorphic or not. There is no method in general to solve this problem but we have some techniques to apply in particular cases. To find homeomorphism is to find bijective continuous maps, and constructing continuous maps is a problem for which there are techniques of solving.

However showing two space not homeomorphic is quite easy. This can be done with the help of topological properties. If a space has a topological property which is not satisfied by the other, then we can clearly say that they are not homeomorphic. But this technique also has very limited application so far. So we must introduce new properties and techniques. Hence comes the idea of fundamental groups which is a generalisation of all the ideas we have done so far. Two spaces that are homeomorphic, have their fundamental groups isomorphic. This unit deals in the fundamental groups and their basic properties.

Objectives

After reading this unit, you will be able to

- define homotopy of paths
- define the various properties related to homotopy and path homotopies
- define fundamental group of a topological space
- learn the properties of the fundamental groups

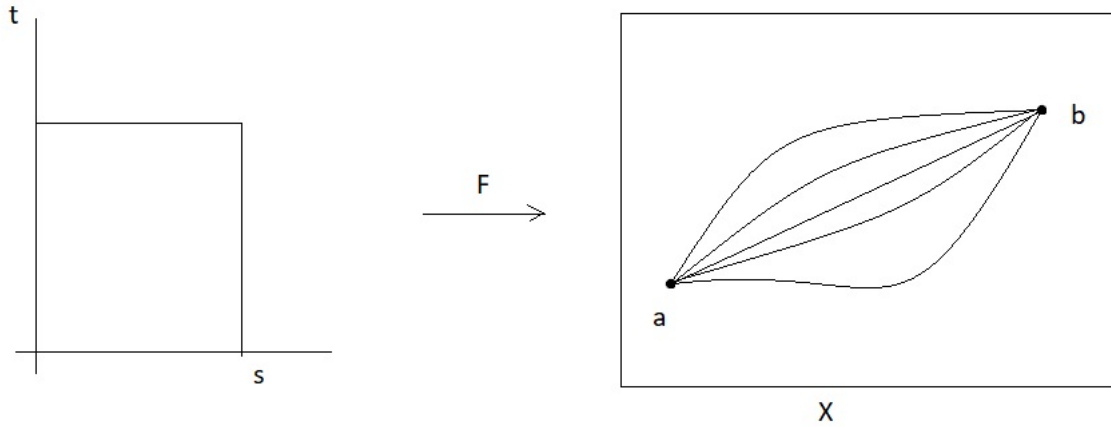


Figure 1: Path-Homotopy

13 Fundamental Group

Before defining the fundamental group of a space X , we shall consider paths on X and an equivalence relation called path homotopy between them.

Definition 13.1. If f and f' are continuous maps of the space X into the space Y , we say that f is homotopic to f' if there is a continuous map

$$F : X \times I \rightarrow Y$$

such that

$$F(x, 0) = f(x) \text{ and } F(x, 1) = f'(x)$$

for each x . Here, $I = [0, 1]$. The map F is called a homotopy between f and f' . If f is homotopic to f' then we write $f \simeq f'$. If $f \simeq f'$ and f' is a constant function, we say that f is nulhomotopic.

Now we consider a special case in which f is a path in X . Recall that if $f : [0, 1] \rightarrow X$ is a continuous map such that $f(0) = x_0$ and $f(1) = x_1$, we say that x_0 is the initial point and x_1 , the final point, of the path f . We will now define a stronger relation, called path-homotopy between two paths.

Definition 13.2. Two paths f and f' mapping the interval $I = [0, 1]$ to X are said to be path-homotopic if they have the same initial point a and same final point b , where $a, b \in X$ and if there exists a continuous map $F : I \times I \rightarrow X$ such that

$$F(s, 0) = f(s) \text{ and } F(s, 1) = f'(s)$$

$$F(0, t) = a \text{ and } F(1, t) = b$$

for each $s \in I$ and each $t \in I$. We call F a path-homotopy between f and f' . If f and f' are path-homotopic, then we write $f \simeq_p f'$. (See Figure 1)

The first condition says that F is a homotopy between f and f' and the second condition says that for each t , the path f_t defined by the equation $f_t(s) = F(s, t)$ is a path from a to b . Said differently, the first condition says that F represents a continuous way of deforming the path f to f' and the second condition ensures that the end points remain fixed in this deformation.

Theorem 13.3. *The relations $\simeq$ and $\simeq_p$ are equivalence relations.*

If f is a path, then we shall denote its path-homotopy class by $[f]$.

Proof. Given f , it is trivial that $f \simeq f$; the map $F(x, t) = f(x)$ is the required homotopy. If f is a path, then F is a path-homotopy.

Given $f \simeq f'$, we show that $f' \simeq f$. Let F be a homotopy between f and f' . Then $G(x, t) = F(x, 1 - t)$ is the homotopy between f' and f . If F is a path homotopy, then G is so.

Suppose $f \simeq f'$ and $f' \simeq f''$. Let F be a homotopy between f and f' and let F' be a homotopy between f' and f'' . Define $G : X \times I \rightarrow Y$ by the equation

$$\begin{aligned} G(x, t) &= F(x, 2t), \quad t \in \left[0, \frac{1}{2}\right] \\ &= F'(x, 2t - 1), \quad t \in \left[\frac{1}{2}, 1\right] \end{aligned}$$

The map G is well-defined, since if $t = \frac{1}{2}$, we have

$$F(x, 2t) = f'(x) = F'(x, 2t - 1).$$

Because G is continuous on the two closed subsets $X \times [0, \frac{1}{2}]$ and $X \times [\frac{1}{2}, 1]$ of $X \times I$, it is continuous on all of $X \times I$, by the gluing lemma. Thus, G is the required homotopy between f and f'' .

If F and F' are path-homotopies, then G is so. We only need to show that the initial and final points are fixed. Let f and f' be path-homotopic, then all the previous conditions remain the same except in this case, we have their initial and final points the same, say they are a and b respectively. Similarly, if f' and f'' are path-homotopic, then f'' also has the initial and final points as a and b respectively. We have,

$$\begin{aligned} G(0, t) &= F(0, 2t) = a, \quad t \in \left[0, \frac{1}{2}\right] \\ &= F'(0, 2t - 1) = a, \quad t \in \left[\frac{1}{2}, 1\right] \end{aligned}$$

So, we are getting $G(0, t) = a$. Similarly, we can show that $G(1, t) = b$, for $t \in I$. Hence, G is a path-homotopy. $\square$

Example 13.4. Let C be a convex sub set of a euclidean space and let $f, g : X \rightarrow C$ maps, where X is an arbitrary topological space. For each point $x \in X$, the straight line joining $f(x)$ and $g(x)$

lies in C , and we can define a homotopy from f to g simply by sliding f along these straight lines. To be precise, define $F : X \times I \times C$ by

$$F(x, t) = (1 - t)f(x) + tg(x).$$

Notice that if f and g happen to agree on a subset A of X then this homotopy is a homotopy relative to A . The homotopy F is called a straight-line homotopy.

We now introduce some algebra into this geometric definition.

Definition 13.5. If f is a path in X from a to b , and if g is a path in X from b to c , we define the product $f * g$ of f and g to be the path h given by the equations

$$\begin{aligned} h(s) &= f(2s), \quad s \in \left[0, \frac{1}{2}\right], \\ &= g(2s - 1), \quad s \in \left[\frac{1}{2}, 1\right]. \end{aligned}$$

The function h is well-defined and continuous, by the pasting lemma; it is a path in X from a to c . We think of h as the path whose first half is the path f and second half is the path g .

The product operation on paths induces a well-defined operation on path-homotopy classes, defined by the equation

$$[f] * [g] = [f * g].$$

To verify this fact, let F be a path-homotopy between f and f' and let G be a path homotopy between g and g' . Define

$$\begin{aligned} H(s, t) &= F(2s, t), \quad s \in \left[0, \frac{1}{2}\right], \\ &= G(2s - 1, t), \quad s \in \left[\frac{1}{2}, 1\right]. \end{aligned}$$

Because $F(1, t) = b = G(0, t)$ for all t , the map H is well-defined, it is continuous by the pasting lemma. It can be checked that H is the required path homotopy between $f * g$ and $f' * g'$.

The operation $*$ on path-homotopy classes turns out to satisfy properties that look very much like the axioms for a group. They are called the groupoid properties of $*$. One difference from the properties of a group is that $[f] * [g]$ is not defined for every pair of classes, but only for those pairs $[f], [g]$ for which $f(1) = g(0)$.

Theorem 13.6. *The operation $*$ has the following properties*

1. *If $[f] * ([g] * [h])$ is defined, so is $([f] * [g]) * [h]$, and they are equal.(associativity)*
2. *Given $x \in X$, let e_x denote the constant path $e_x : I \rightarrow X$ carrying all of I to the point x . If f is a path in X from a to b , then*

$$[f] * [e_b] = [f] \text{ and } [e_a] * [f] = [f].$$

(right and left inverses)

3. Given the path f in X from a to b , let $\bar{f}$ be the path defined by $\bar{f}(s) = f(1 - s)$. It is called the reverse of f . Then

$$[f] * [\bar{f}] = [e_a] \text{ and } [\bar{f}] * [f] = [e_b].$$

(inverse)

The set of path-homotopy classes of paths in a space X does not form a group under the operation $*$ because the product of two path-homotopy classes is not always defined. But if we pick out a point a of X serving as a base point and restrict ourselves to those beginning and ending at a , then those homotopy classes form a group under $*$, which is called the fundamental group.

Before starting with the definitions, we assume that the reader is well accustomed to the idea of group homomorphism, cosets, normal subgroups, factor or quotient groups, etc. Let us start with the definition of fundamental group relative to a base point a .

Definition 13.7. Let X be a space and let a be a point of X . A path that begins and ends at a is called a loop based at a . The set of path homotopy classes of loops based at a , with the operation $*$, is called the **fundamental group** of X relative to the base point a . It is denoted by $\pi_1(X, a)$.

It follows from the previous theorem that the operation $*$, when restricted to this set, satisfies the axioms for groups. Given two loops f and g based at a , the product $f * g$ is always defined and is a loop based at a . The other properties are immediate.

Sometimes this group is called the first homotopy group of X . From the term, it seems that there is also a second-homotopy group. There are in fact groups $\pi_n(X, a)$ for all $n \in \mathbb{N}$. But we shall omit them here and concentrate on the first homotopy group only.

Example 13.8. Let $\mathbb{R}^n$ denote the euclidean n -space. Then $\pi_1(\mathbb{R}^n, a)$ is the trivial group (the group consisting of the identity alone). For if f is a loop in $\mathbb{R}^n$ based at a , the straight-line homotopy is a path-homotopy between f and the constant path at a . More generally, if X is any convex subset of $\mathbb{R}^n$, then $\pi_1(X, a)$ is the trivial group. In particular, the open ball B^n in $\mathbb{R}^n$,

$$B^n = \{\mathbf{x} \mid x_1^2 + x_2^2 + \cdots + x_n^2 \leq 1\},$$

has a trivial fundamental group.

A valid question is that how much the fundamental group depends upon the base point. First let us see the following definition.

Definition 13.9. Let α be a path in X from a to b . We define a map

$$\hat{\alpha} : \pi_1(X, a) \rightarrow \pi_1(X, b)$$

by the equation

$$\hat{\alpha}([f]) = [\bar{\alpha}] * [f] * [\alpha].$$

The map $\hat{\alpha}$ is well-defined because the operation $*$ is well-defined. If f is a loop based at a , then $\bar{\alpha} * (f * \alpha)$ is a loop based at b . Hence $\hat{\alpha}$ maps $\pi_1(X, a)$ into $\pi_1(X, b)$ as desired. Note that, it only depends on the path-homotopy class of α .

Theorem 13.10. *The map $\hat{\alpha}$ is a group isomorphism.*

Proof. To show that $\hat{\alpha}$ is a homomorphism, we compute

$$\begin{aligned}\hat{\alpha}([f]) * \hat{\alpha}([g]) &= ([\bar{\alpha}] * [f] * [\alpha]) * ([\bar{\alpha}] * [g] * [\alpha]) \\ &= [\bar{\alpha}] * [f] * [g] * [\alpha] \\ &= \hat{\alpha}([f] * [g]).\end{aligned}$$

To show that $\hat{\alpha}$ is an isomorphism, we show that if β denotes the path $\bar{\alpha}$, which is the reverse of α , then $\hat{\beta}$ is an inverse for $\hat{\alpha}$. We compute for each element $[h]$ of $\pi_1(X, b)$,

$$\begin{aligned}\hat{\beta}([h]) &= [\bar{\beta}] * [h] * [\beta] = [\alpha] * [f] * [\bar{\alpha}], \\ \hat{\alpha}(\hat{\beta}([h])) &= [\bar{\alpha}] * ([\alpha] * [h] * [\bar{\alpha}]) * [\alpha] = [h].\end{aligned}$$

A similar computation shows that

$$\hat{\beta}(\hat{\alpha}([f])) = [f]$$

for each $[f] \in \pi_1(X, a)$. □

Corollary 13.11. *If X is path-connected and a and b are two points of X , then $\pi_1(X, a)$ and $\pi_1(X, b)$ are isomorphic.*

Let X be a space and P be a path-component of it containing a . Then its easy to see that

$$\pi_1(X, a) = \pi_1(P, a),$$

since all the loops and homotopies in X are that are based at a must lie in P . Thus, $\pi_1(X, a)$ depends only on the path component of X containing a .

Also, if X is path-connected, all the groups $\pi_1(X, a)$ are isomorphic, so its tempting to try to identify all these groups with one another and to speak simply of the fundamental group of X , without reference to the base point. The difficulty in this approach is that there is no natural way to identify $\pi_1(X, a)$ with $\pi_1(X, b)$. Different paths α and β from a to b may give rise to different isomorphisms between these groups. For this reason, omitting the base point may lead to error.

We now move on to define the simple-connectedness of a space.

Definition 13.12. A space X is simply-connected if it is a path-connected space and if $\pi_1(X, a)$ is the trivial one-element group for some $a \in X$, and hence for every $a \in X$. We often express the fact that $\pi_1(X, a)$ is the trivial group by writing

$$\pi_1(X, a) = 0.$$

Lemma 13.13. *In a simply connected space X , any two paths having the same initial and final points are path-homotopic.*

Proof. Let α and β be two paths from a to b . Then $\alpha * \bar{\beta}$ is defined and is a loop on X based at a . Since X is simply connected, this loop is path-homotopic to the constant loop at a . Then

$$[\alpha * \bar{\beta}] * [\beta] = [e_a] * [\beta]$$

from which it follows that

$$[\alpha] = [\beta].$$

□

By now, it might be clear that fundamental groups are a topological invariant. But in order to prove it mathematically, one has to introduce the notion of the homomorphism induced by a continuous map.

Suppose that $h : X \rightarrow Y$ is a continuous map that carries the point a of X to the point a' of Y . We often denote this fact by writing

$$h : (X, a) \rightarrow (Y, a').$$

If f is a loop in X based at a , then the composite $h \circ f : I \rightarrow Y$ is a loop in Y based at a' . The correspondence $f \rightarrow h \circ f$ thus gives rise to a map carrying $\pi_1(X, a)$ into $\pi_1(Y, a')$. We define it formally as follows.

Definition 13.14. Let $h : (X, a) \rightarrow (Y, a')$ be a continuous map. Define

$$h_* : \pi_1(X, a) \rightarrow \pi_1(Y, a')$$

by the equation

$$h_*([f]) = [h \circ f].$$

The map h_* is called the homomorphism induced by h , relative to the base point a .

The map h_* is well-defined, for if F is a path-homotopy between the paths f and f' , then $h \circ F$ is a path-homotopy between the paths $h \circ f$ and $h \circ f'$. The fact that h_* is a homomorphism follows from the equation

$$(h \circ f) * (h \circ g) = h \circ (f * g).$$

The homomorphism h_* depends not only on the map $h : X \rightarrow Y$ but also on the choice of the base point a (a' is determined by h). So some notational difficulty will arise if we consider several different base points for X .

The induced homomorphism has two properties that are crucial in the applications. They are called the functional properties and are given in the following theorem.

Theorem 13.15. If $h : (X, a) \rightarrow (Y, a')$ and $k : (Y, a') \rightarrow (Z, a'')$ are continuous, then $(k \circ h)_* = k_* \circ h_*$. If $i : (X, a) \rightarrow (X, a)$ is the identity map, then i_* is the identity homomorphism.

Proof. The proof is trivial. By definition,

$$(k \circ h)_*([f]) = [(k \circ h) \circ f],$$

$$\begin{aligned} (k_* \circ h_*)([f]) &= k_*(h_*([f])) \\ &= k_*([h \circ f]) \\ &= [k \circ (h \circ f)]. \end{aligned}$$

Similarly,

$$i_*([f]) = [i \circ f] = [f].$$

Hence the theorem. □

Corollary 13.16. *If $h : (X, a) \rightarrow (Y, a')$ is a homeomorphism of X with Y , then h_* is an isomorphism of $\pi_1(X, a)$ with $\pi_1(Y, a')$.*

Proof. Let $k : (Y, a') \rightarrow (X, a)$ be the inverse of h . Then

$$k_* \circ h_* = (k \circ h)_* = i_*,$$

where i is the identity map of (X, a) and

$$h_* \circ k_* = (h \circ k)_* = j_*,$$

where j is the identity map of (Y, a') . Since i_* and j_* are the identity homomorphisms of the groups $\pi_1(X, a)$ and $\pi_1(Y, a')$ respectively, k_* is the inverse of h_* . □

References:

- [1] *Introduction to Toplogy and Modern Analysis*, G.F Simmons
- [2] *Topology*, James. R. Munkres
- [3] *Introduction to Topological Manifolds*, John. M. Lee
- [4] *Theory and Problems of General Topology*, Seymour Lipschutz