

2
Edition

Core Concepts in **Mathematics: Calculus**

Diana Dias

CORE CONCEPTS IN MATHEMATICS:
CALCULUS

Diana Dias

2nd Edition



www.bibliotex.com

CORE CONCEPTS IN MATHEMATICS: CALCULUS

2nd Edition

Diana Dias



www.bibliotex.com

email: info@bibliotex.com

e-book Edition 2024

ISBN: 978-1-98469-080-7 (e-book)

This book contains information obtained from highly regarded resources. Reprinted material sources are indicated. Copyright for individual articles remains with the authors as indicated and published under Creative Commons License. A Wide variety of references are listed. Reasonable efforts have been made to publish reliable data and views articulated in the chapters are those of the individual contributors, and not necessarily those of the editors or publishers. Editors or publishers are not responsible for the accuracy of the information in the published chapters or consequences of their use. The publisher assumes no responsibility for any damage or grievance to the persons or property arising out of the use of any materials, instructions, methods or thoughts in the book. The editors and the publisher have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission has not been obtained. If any copyright holder has not been acknowledged, please write to us so we may rectify.

Notice: Registered trademark of products or corporate names are used only for explanation and identification without intent of infringement.

© 2024 3G E-learning LLC

In Collaboration with 3G E-Learning LLC. Originally Published in printed book format by 3G E-Learning LLC with ISBN 978-1-98468-863-7

EDITORIAL BOARD



Aleksandar Mratinković was born on May 5, 1988. in Arandjelovac, Serbia. He has graduated on Economic high school (2007), The College of Tourism in Belgrade (2013), and also has a master degree of Psychology (Faculty of Philosophy, University of Novi Sad). He has been engaged in different fields of psychology (Developmental Psychology, Clinical Psychology, Educational Psychology and Industrial Psychology) and has published several scientific works



Dan Piestun (PhD) is currently a startup entrepreneur in Israel working on the interface of Agriculture and Biomedical Sciences and was formerly president-CEO of the National Institute of Agricultural Research (INIA) in Uruguay. Dan is a widely published scientist who has received many honours during his career including being a two-time recipient of the Amit Golda Meir Prize from the Hebrew University of Jerusalem, his areas of expertise includes stem cell molecular biology, plant and animal genetics and bioinformatics. Dan's passion for applied science and technological solutions did not stop him from pursuing a deep connection to the farmer, his family and nature. Among some of his interest and practices counts enjoying working as a beekeeper and onboard fishing.



Felecia Killings is the Founder and CEO of LiyahAmore Publishing, a publishing company committed to providing technical and educational services and products to Christian Authors. She operates as the Senior Editor and Writer, the Senior Writing Coach, the Content Marketing Specialist, Editor-in-Chief to the company's quarterly magazine, the Executive and Host of an international virtual network, and the Executive Director of the company's online school for Authors. She is a former high-school English instructor and professional development professor. She possesses a Master of Arts degree in Education and a Bachelor's degree in English and African American studies.



Dr. Sandra El Hajj, Ph.D. in Health Sciences from Nova Southeastern University, Florida, USA is a health professional specialized in Preventive and Global Health. With her 12 years of education obtained from one of the most prominent universities in Beirut, in addition to two leading universities in the State of Florida (USA), Dr. Sandra made sure to incorporate interdisciplinary and multicultural approaches in her work. Her long years of studies helped her create her own miniature world of knowledge linking together the healthcare field with Medical Research, Statistics, Food Technology, Environmental & Occupational Health, Preventive Health and most noteworthy her precious last degree of Global Health. Till today, she is the first and only doctor specialized in Global Health in the Middle East area.



Hazem Shawky Fouda has a PhD. in Agriculture Sciences, obtained his PhD. From the Faculty of Agriculture, Alexandria University in 2008, He is working in Cotton Arbitration & Testing General Organization (CATGO).



Fozia Parveen has a Dphil in Sustainable Water Engineering from the University of Oxford. Prior to this she has received MS in Environmental Sciences from National University of Science and Technology (NUST), Islamabad Pakistan and BS in Environmental Sciences from Fatima Jinnah Women University (FJWU), Rawalpindi.



Igor Kronic 2003-2007 in the School of Economics. After graduating in 2007, he went on to study at The College of Tourism, at the University of Belgrade where he got his bachelor degree in 2010. He was active as a third-year student representative in the student parliament. Then he went on the Faculty of science, at the University of Novi Sad where he successfully defended his master's thesis in 2013. The crown of his study was the work titled "Opportunities for development of cultural tourism in Cacak". Later on, he became part of a multinational company where he got promoted to a deputy director of logistic. Nowadays he is a consultant and writer of academic subjects in the field of tourism.



Dr. Jovan Pehcevski obtained his PhD in Computer Science from RMIT University in Melbourne, Australia in 2007. His research interests include big data, business intelligence and predictive analytics, data and information science, information retrieval, XML, web services and service-oriented architectures, and relational and NoSQL database systems. He has published over 30 journal and conference papers and he also serves as a journal and conference reviewer. He is currently working as a Dean and Associate Professor at European University in Skopje, Macedonia.



Dr. Tanjina Nur finished her PhD in Civil and Environmental Engineering in 2014 from University of Technology Sydney (UTS). Now she is working as Post-Doctoral Researcher in the Centre for Technology in Water and Wastewater (CTWW) and published about eight International journal papers with 80 citations. Her research interest is wastewater treatment technology using adsorption process.



Stephen obtained his PhD from the University of North Carolina at Charlotte in 2013 where his graduate research focused on cancer immunology and the tumor microenvironment. He received postdoctoral training in regenerative and translational medicine, specifically gastrointestinal tissue engineering, at the Wake Forest Institute of Regenerative Medicine. Currently, Stephen is an instructor for anatomy and physiology and biology at Forsyth Technical Community College.



Michelle holds a Masters of Business Administration from the University of Phoenix, with a concentration in Human Resources Management. She is a professional author and has had numerous articles published in the Henry County Times and has written and revised several employee handbooks for various YMCA organizations throughout the United States.

Table of Contents

Preface

ix

1 Infinite Sequences and Series 1

Introduction.....	2
Core Concept: Real And Complex Numbers	4
Core Concept: Convergence/Divergence.....	11
Core Concept: Power Series	23
Summary	35
Multiple Choice Questions	36
Review Questions	37
References.....	38

2 Curve Sketching 39

Introduction.....	40
Core Concept: Maxima and Minima	40
Core Concept: Polar Coordinate System	46
Summary	60
Multiple Choice Questions	61
Review Questions	62
References.....	63

3 Indeterminate Forms and Improper Integral 65

Introduction.....	66
Core Concept: Indeterminate Forms	68
Core Concept: Improper Integral	72
Summary	83
Multiple Choice Questions	84

Review Questions	85
References.....	86

4 **Differential Calculus of Several Variables** **87**

Introduction.....	88
Core Concept: Differential Calculus	88
Core Concept: Differentiation of Implicit Functions.....	102
Summary	117
Multiple Choice Questions	118
Review Questions	119
References.....	120

5 **Situational Problems Involving Related Rates** **121**

Introduction.....	122
Core Concept: Higher-order Derivatives of Functions.....	122
Core Concept: Chain Rule of Differentiation	129
Summary	139
Multiple Choice Questions	140
Review Questions	141
References.....	142

6 **Circular Functions** **143**

Introduction.....	144
Core Concept: Theory of Circular Functions.....	145
Core Concept: Domain and Range of a Function	157
Summary	167
Multiple Choice Questions	168
Review Questions	169
References.....	170

7 Conic Sections and Systems 171

Introduction.....	172
Core Concept: Conic Sections	173
Core Concept: Systems of Nonlinear Equations	191
Summary	207
Multiple Choice Questions	208
Review Questions	209
References.....	210

8 Line Integrals 211

Introduction.....	212
Core Concept: Line Integral Definition	212
Core Concept: Line Integrals in Green's Theorem	236
Core Concept: Line Integral in Stokes's Theorem.....	241
Core Concept: Line Integrals in Vector Fields.....	243
Summary	247
Multiple Choice Questions	248
References.....	249

Index 251

Preface

Calculus is a field of study that focuses on understanding how things change. It provides a frame-work for modeling systems that experience change and allows us to make predictions based on these models. The development of calculus, particularly its applications in physics and engineering, has had a profound impact on the advancement of modern science. In fact, it played a pivotal role in the industrial revolution and subsequent scientific breakthroughs. No other branch of mathematics boasts such a rich history and cultural significance as calculus. From its foundational principles rooted in geometry, such as Pythagorean Theorem as part of vector calculus, to its application in measuring volumes inspired by Archimedes, calculus has been instrumental in addressing questions about our place in the universe. By analyzing velocities and forces, it enabled advancements in astronomy and aided our exploration of celestial bodies like planets, galaxies, and phenomena like black holes. The quest to understand our role within the vast universe itself has propelled us to embrace calculus. Moreover, calculus holds vital importance for our survival as it equips us with tools to develop climate models or predict population growth. It helps us comprehend disease spread dynamics and enables conflict resolution mechanisms alongside tackling economic crises effectively. Throughout history, humanity's reliance on calculus has only grown stronger as we continue pushing past boundaries and seeking greater insight into natural phenomena - making this field an indispensable pillar upon which our collective progress rests.

This book, *Core Concepts in Calculus 2nd edition* is organized into Eight Chapters. The book provides a succinct account of the essentials of calculus, presented in an engaging style that is both readable and mathematically precise. All chapters are revised and restructured including newly added chapters.

Chapter One is taking a look at sequences and (infinite) series. Actually, this chapter will deal almost exclusively with series. In experimental science and engineering, as well as in everyday life, we deal with integers, or at most rational numbers. In this chapter, you will understand the core concepts of real and complex numbers, convergence/divergence, and power series.

Chapter Two focuses on curve sketching, which includes techniques that can be used to produce a rough idea of overall shape of a plane curve given its equation without computing the large numbers of points required for a detailed plot.

Chapter Three presents insight into indeterminate forms and improper integral. An indeterminate form is an expression involving two functions whose limit cannot be determined solely from the limits of the individual functions. These forms are common

in calculus; indeed, the limit definition of the derivative is the limit of an indeterminate form.

In Chapter Four, you will understand the core concepts of differential calculus and differentiation of implicit functions.

A focus on Cartesian product and cardinality is given in Chapter Five. In this chapter, you will understand the core concepts of higher-order derivatives of functions and chain rule of differentiation.

Chapter Six provides a focus on circular functions which may be thought of as a way to extend these extremely useful mathematical relationships. Theory of circular functions and domain and range of a function are covered in this chapter.

Chapter Seven immerses itself into the technical aspects of conic sections and systems, shedding light on their functional implications in distinct fields. Constructing a comprehensive overview commences with understanding ellipses and hyperbola configurations crucial to plot data points.

Chapter Eight focuses on Line Integrals. This transition elaborates on contours' ponderous necessity while calculating path integrals in scalar fields significantly. Such knowledge expounded here paves the way for comprehending complex calculus concepts necessary for an array of professions like engineering and physics.

INFINITE SEQUENCES AND SERIES



A sequence is an ordered list of objects (or events). Like a set, it contains members (also called elements, or terms). The number of ordered elements (possibly infinite) is called the length of the sequence. A series is, informally speaking, the sum of the terms of a sequence. Finite sequences and series have defined first and last terms, whereas infinite sequences and series continue indefinitely.

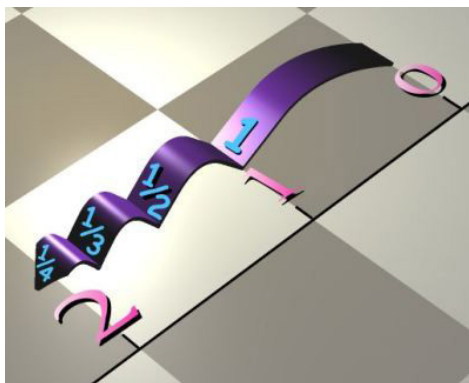
After studying this chapter, you will understand the core concepts of:

- Real and complex numbers
- Convergence/divergence
- Power series

INTRODUCTION

An infinite sequence (sometimes just called a sequence) is a function with a domain of all positive integers. In beginning calculus, the range of an infinite sequence is usually the set of real numbers, although it's also possible for the range to include complex numbers. A series is, roughly speaking, the operation of adding infinitely many quantities, one after the other, to a given starting quantity. The study of series is a major part of calculus and its generalization, mathematical analysis.

An infinite sequence of real numbers is an ordered unending list of real numbers. E.g.: 1, 2, 3, 4,... We will be taking a look at sequences and (infinite) series. Actually, this chapter will deal almost exclusively with series. In experimental science and engineering, as well as in everyday life, we deal with integers, or at most rational numbers. Yet in theoretical analysis, we use real and complex numbers, as well as far more abstract mathematical constructs, fully expecting that this analysis will eventually provide useful models of natural phenomena. Hence we proceed through the construction of the real and complex numbers starting from the positive integers.



We start with the positive integers and zero, and introduce negative integers to allow subtraction of integers. Then we introduce rational numbers to permit division by integers. From arithmetic we proceed to analysis, which begins with the concept of convergence of infinite sequences of (rational) numbers, as defined here by the Cauchy criterion. Then we define irrational numbers as limits of convergent (Cauchy) sequences of rational numbers. In order to solve algebraic equations in general, we must introduce complex numbers and the representation of complex numbers as points in the complex plane. The fundamental theorem of algebra states that every polynomial has at least one root in the complex plane, from which it follows that every polynomial of degree n has exactly n roots in the complex plane when these roots are suitably counted.

Once we understand convergence of infinite sequences, we can deal with infinite series of the form

$$\sum_{n=1}^{\infty} x_n$$

and the closely related infinite products of the form

$$\prod_{n=1}^{\infty} x_n$$

Infinite series are central to the study of solutions, both exact and approximate, to the differential equations. Many functions that arise in calculus are defined only through infinite series, and it is important to understand the convergence properties of these series, both for theoretical analysis and for approximate evaluation of the functions. We review some of the standard tests (comparison test, ratio test, root test, integral test) for convergence of infinite series, and give some illustrative examples. We note that absolute convergence of an infinite series is necessary and sufficient to allow the terms of a series to be rearranged arbitrarily without changing the sum of the series.

Infinite sequences of functions have more subtle convergence properties. In addition to point wise convergence of the sequence of values of the functions taken at a single point, there is a concept of uniform convergence on an interval of the real axis, or in a region of the complex plane. Uniform convergence guarantees that properties such as continuity and differentiability of the functions in the sequence are shared by the limit function. There is also a concept of weak convergence, defined in terms of the sequences of numbers generated by integrating each function of the sequence over a region with functions from a class of smooth functions (test functions). For example, the Dirac δ -function and its derivatives are defined in terms of weakly convergent sequences of well-behaved functions. It is a short step from sequences of functions to consider infinite series of functions, especially power series of the form

$$\sum_{n=0}^{\infty} a_n z^n$$

in which the a_n are real or complex numbers and z is a complex variable. These series are central to the theory of functions of a complex variable. We show that a power series converges absolutely and uniformly inside a circle in the complex plane (the circle of convergence), with convergence on the circle of convergence an issue that must be decided separately for each particular series.

Even divergent series can be useful. We show some examples that illustrate the idea of a semi convergent, or asymptotic, series. These can be used to determine the asymptotic behavior and approximate asymptotic values of a function, even though the series is actually divergent. We give a general description of the properties of such series, and explain Laplace's method for finding an asymptotic expansion of a function defined by an integral representation (Laplace integral) of the form

$$I(z) = \int_0^a f(t) e^{zh(t)} dt$$

Beyond the sequences and series generated by the mathematical functions that occur in solutions to differential equations, there are sequences generated by dynamical systems themselves through the equations of motion of the system. These sequences can be viewed as iterated maps of the coordinate space of the system into itself; they arise in classical mechanics, for example, as successive intersections of a particle orbit with a fixed plane. They also arise naturally in population dynamics as a sequence of population counts at periodic intervals.

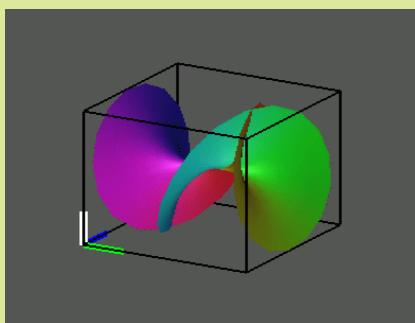
The asymptotic behavior of these sequences exhibits new phenomena beyond the simple convergence or divergence familiar from before studies. In particular, there are sequences that converge, not to a single limit, but to a periodic limit cycle, or that diverge in such a way that the points in the sequence are dense in some region in a coordinate space. An elementary prototype of such a sequence is the logistic map defined by

$$T_\lambda : x \rightarrow x_\lambda = \lambda x(1 - x)$$

This map generates a sequence of points $\{x_n\}$ with

$$x_{n+1} = \lambda x_n(1 - x_n)$$

($0 < \lambda < 4$) starting from a generic point x_0 in the interval $0 < x_0 < 1$. The behavior of this sequence as a function of the parameter λ as λ increases from 0 to 4 provides a simple illustration of the phenomena of period doubling and transition to chaos that have been an important focus of research in the past 30 years or so.



CORE CONCEPT: REAL AND COMPLEX NUMBERS

Real numbers are numbers that can be found on the number line. This includes both the rational and irrational numbers. A complex number is a number that can be expressed in the form $a + bi$, where a and b are real numbers, and i is a solution of the equation $x^2 = -1$, which is called an imaginary

number because there is no real number that satisfies this equation.

In mathematics, a real number is a value of a continuous quantity that can represent a distance along a line. The numbers we deal with in the real world (ignoring any units that go along with them, such as dollars, inches, degrees, etc.) are usually real numbers. A real number is any number that can be placed on a number line that extends to **infinity** in both the positive and negative directions.

A complex number is any number that includes i . Thus, $3i$, $2 + 5.4i$, and $-\pi i$ are all complex numbers. (In fact, the real numbers are a subset of the complex numbers-any real number r can be written as $r + 0i$, which is a complex representation.) Complex numbers are an important part of algebra, and they do have relevance to such things as solutions to polynomial equations.

Infinity is a concept describing something without any bound or larger than any natural number.

Arithmetic

The construction of the real and complex number systems starting from the positive integers illustrates several of the structures studied extensively by mathematicians. The positive integers have the property that we can add, or we can multiply, two of them together and get a third. Each of these operations is commutative:

$$x \circ y = y \circ x \quad (1)$$

and associative:

$$x \circ (y \circ z) = (x \circ y) \circ z \quad (2)$$

(here $\circ$ denotes either addition or multiplication), but only for multiplication is there an identity element e , with the property that

$$e \circ x = x = x \circ e \quad (3)$$

Of course the identity element for addition is the number zero, but zero is not a positive integer. Properties (2) and (3) are enough to characterize the positive integers as a semigroup under multiplication, denoted by Z_* or, with the inclusion of zero, a semigroup under addition, denoted by Z_+ . Neither addition nor multiplication has an inverse defined within the positive integers. In order to define an inverse for addition, it is necessary to include zero and the negative integers. Zero is defined as the identity for addition, so that

$$x + 0 = x = 0 + x \quad (4)$$

and the negative integer $-x$ is defined as the inverse of x under addition,

$$x + (-x) = 0 = (-x) + x \quad (5)$$

With the inclusion of the negative integers, the equation

$$p + x = q \quad (6)$$

has a unique integer solution $x (\equiv q - p)$ for every pair of integers p, q . Properties (2)–(5) characterize the integers as a group Z under addition, with 0 as an identity element. The fact that addition is commutative makes Z a commutative, or Abelian, group. The combined operations of addition with zero as identity, and multiplication satisfying Equations. (2) and (3) with 1 as identity, characterize Z as a ring, a commutative ring since multiplication is also commutative. To proceed further, we need an inverse for multiplication, which leads to the introduction of fractions of the form p/q (with integers p, q). One important property of fractions is that they can always be reduced to a form in which the integers p, q have no common factors. Numbers of this form are rational. With both addition and multiplication having well-defined inverses (except for division by zero, which is undefined), and the distributive law

$$a * (x + y) = a * x + a * c = y \quad (7)$$

satisfied, the rational numbers form a field, denoted by Q .

Algebraic Equations

The rational numbers are adequate for the usual operations of arithmetic, but to solve algebraic (polynomial) equations, or to carry out the limiting operations of calculus, we need more. For example, the quadratic equation

$$x^2 - 2 = 0 \quad (8)$$

has no rational solution, yet it makes sense to enlarge the rational number system to include the roots of this equation. The real algebraic numbers are introduced as the real roots of polynomials of any degree with integer coefficients. The algebraic numbers also form a field.

Exercise. Show that the roots of a polynomial with rational coefficients can be expressed as roots of a polynomial with integer coefficients.

Complex numbers are introduced in order to solve algebraic equations that would otherwise have no real roots. For example, the equation

$$x^2 + 1 = 0 \quad (9)$$

has no real solutions; it is “solved” by introducing the imaginary unit $i \equiv \sqrt{-1}$ so that the roots are given by $x = \pm i$. Complex numbers are then introduced as ordered pairs $(x, y) \sim x + iy$, of real numbers; x, y can be restricted to be rational (algebraic) to define the complex rational (algebraic) numbers.

Complex numbers can be represented as points (x, y) in a plane (the complex plane) in a natural way, and the magnitude of the complex number $x + iy$ is defined by

$$|x + iy| \equiv \sqrt{x^2 + y^2} \quad (10)$$

In view of the identity

$$e^{i\theta} = \cos \theta + i \sin \theta \quad (11)$$

we can also write

$$x + iy = re^{i\theta} \quad (12)$$

with $r = |x + iy|$ and $\tan \theta = y/x$. These relations have an obvious interpretation in terms of the polar coordinates of the point (x, y) . We also define

$$\arg z \equiv \theta \quad (13)$$

for $z \neq 0$. The angle $\arg z$ is the phase of z . Evidently it can only be defined as mod 2π ; adding any integer multiple of 2π to $\arg z$ does not change the complex number z , since

$$e^{2\pi i} = 1 \quad (14)$$

Equation (14) is one of the most remarkable equations of mathematics.

Infinite Sequences; Irrational Numbers

To complete the construction of the real and complex numbers, we need to look at some elementary properties of sequences, starting with the formal definitions:

Definition 1. A sequence of numbers (real or complex) is an ordered set of numbers in one-to-one correspondence with the positive integers; write $\{z_n\} \equiv \{z_1, z_2, \dots\}$.

Definition 2. The sequence $\{z_n\}$ is bounded if there is some positive number M such that $|z_n| < M$ for all positive integers n .

Definition 3. The sequence $\{x_n\}$ of real numbers is increasing (decreasing) if $x_{n+1} > x_n$ ($x_{n+1} < x_n$) for every n . The sequence is non-decreasing (non-increasing) if $x_{n+1} \geq x_n$ ($x_{n+1} \leq x_n$) for every n . A sequence belonging to one of these classes is monotone (or monotonic).

Remark. The preceding definition is restricted to real numbers because it is only for real numbers that we can define a “natural” ordering that is compatible with the standard measure of the distance between the numbers.

Definition 4. The sequence $\{z_n\}$ is a Cauchy sequence if for every $\varepsilon > 0$ there is a positive integer N such that $|z_p - z_q| < \varepsilon$ whenever $p, q > N$.

Definition 5. The sequence $\{z_n\}$ is convergent to the limit z (write $\{z_n\} \rightarrow z$) if for every $\varepsilon > 0$ there is a positive integer N such that $|z_n - z| < \varepsilon$ whenever $n > N$.

There is no guarantee that a Cauchy sequence of rational numbers converges to a rational, or even algebraic, limit. For example, the sequence $\{x_n\}$ defined by

$$x_n \equiv \left(1 + \frac{1}{n}\right)^n \quad (15)$$

converges to the limit $e = 2.71828 \dots$, the base of natural logarithms. It is true, though nontrivial to prove, that e is not an algebraic number. A real number that is not algebraic is transcendental. Another famous transcendental number is π , which is related to e through Eq. (14).

If we want to insure that every Cauchy sequence of rational numbers converges to a limit, we must include the irrational numbers, which can be defined as limits of Cauchy sequences of rational numbers. As examples of such sequences, imagine the infinite, nonterminating, non-periodic decimal expansions of transcendental numbers such as e or π , or algebraic numbers such as $\sqrt{2}$. Countless computer cycles have been used in calculating the digits in these expansions.

The set of real numbers, denoted by R , can now be defined as the set containing rational numbers together with the limits of Cauchy sequences of rational numbers. The set of complex numbers, denoted by C , is then introduced as the set of all ordered pairs $(x, y) \sim x+iy$ of real numbers. Once we know that every Cauchy sequence of real (or rational) numbers converges to a real number, it is a simple exercise to show that every Cauchy sequence of complex numbers converges to a complex number.

Sets of Real and Complex Numbers

We also need some elementary definitions and results about sets of real and complex numbers that are generalized later to other structures.

Definition 6. For real numbers, we can define an open interval:

$$(a, b) \equiv \{x \mid a < x < b\}$$

or a closed interval:

$$[a, b] \equiv \{x \mid a \leq x \leq b\}$$

as well as semiopen (or semiclosed) intervals:

$$(a, b] \equiv \{x \mid a < x \leq b\} \quad \text{and} \quad [a, b) \equiv \{x \mid a \leq x < b\}$$

A neighborhood of the real number x_0 is any open interval containing x_0 . An ε -neighborhood of x_0 is the set of all points x such that

$$|x - x_0| < \varepsilon \tag{16}$$

This concept has an obvious extension to complex numbers: An (ε) -neighborhood of the complex number z_0 , denoted by $N_\varepsilon(z_0)$, is the set of all points z such that

$$0 < |z - z_0| < \varepsilon \tag{17}$$

Note that for complex numbers, we exclude the point z_0 from the neighborhood $N_\varepsilon(z_0)$.

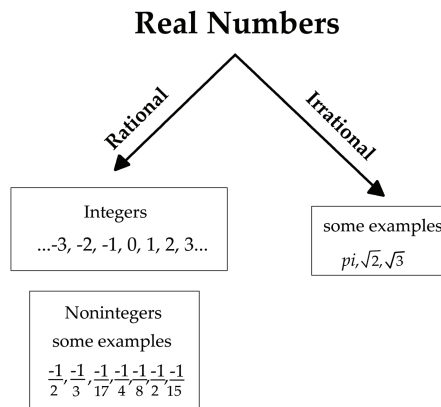
Definition 7. The set S of real or complex numbers is open if for every x in S , there is a neighborhood of x lying entirely in S . S is closed if its complement is open. S is bounded if there is some positive M such that x

Definition 8. x is a limit point of the set S if every neighborhood of x contains at least one point of S .

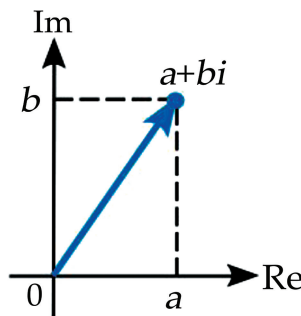
While x itself need not be a member of the set S , this definition implies that every neighborhood of x in fact contains an infinite number of points of S . An alternative definition of a closed set can be given in terms of limit points, and one of the important results of analysis is that every bounded infinite set contains at least one limit point.

Difference between Real Number and Complex Number

A real number is a number that can take any value on the number line. A real number can be any of the rational and irrational numbers. Complex numbers are the numbers that exist in the form of $a+ib$, where a and b denotes real numbers and i denotes an imaginary part.

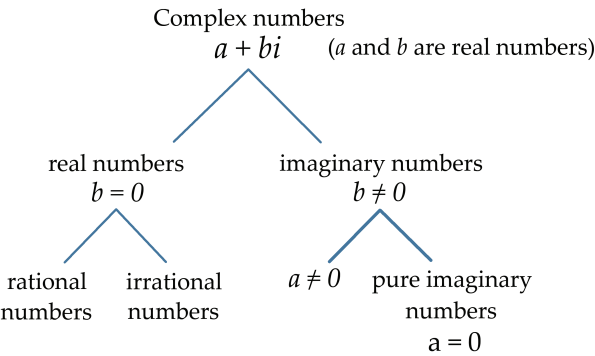


It is important to understand the concept of number line to learn about real numbers. Number line can be expressed as an actual geometric line where a point is chosen to be the origin. Points that fall in the right side of origin are considered positive numbers, whereas numbers lying in the left side of origin are considered to be negative. A real number refers to any number that can be found on this number line. Therefore, they consist of whole (0,1,3,9,26), rational ($6/9$, 78.98) and irrational numbers (square root of 3, π). Infinity does not fall in the category of real numbers. Square root of -1 is also not a real number, and therefore it is referred to as an imaginary number.



Complex Numbers are considered to be an extension of the real number system. They have been designed in order to solve the problems that cannot be solved using real numbers. Complex numbers actually combine real and imaginary number ($a+ib$), where a and b denotes real numbers, whereas i denotes an imaginary number. By

definition, imaginary numbers are those numbers which when squared give a negative result. This statement would not make out a lot of logic as when we calculate the square of a positive number, we get a positive result. Similarly, when a negative number is squared it also provides a positive number. Therefore, imaginary name is given to such numbers. However, unit imaginary number is considered to be the square root of -1



Comparison between Real Number and Complex Number

The comparison between real number and complex number are:

	Real Number	Complex Number
Definition	A real number is a number that can take any value on the number line. They can be any of the rational and irrational numbers.	Complex Numbers are considered to be an extension of the real number system. Complex numbers actually combine real and imaginary number (a+ib), where a and b denotes real numbers, whereas i denotes an imaginary number.
Number line	Can be plotted on the number line.	Cannot be plotted on the number line.
Includes	This includes (but is not limited to) positives and negatives, integers and rational numbers, square roots, cube roots , π (pi), etc.	This includes numbers like $3 - 2i$ or $5 + \sqrt{6}i$, as they can be written as the sum or difference of a real number and an imaginary number.
Symbol	R	C
Members	All real numbers are also complex numbers with zero for the imaginary part.	All imaginary numbers are also complex numbers with zero for the real part.

Determine whether the geometric series is convergent or divergent

$$\sum_{n=1}^{\infty} \frac{(-3)^{n-1}}{4^n}$$

Reminders: $\sum_{n=1}^{\infty} ar^{n-1} \Rightarrow \boxed{a} + ar + ar^2 + ar^3 + \dots$ convergent if $|r| < 1$
 divergent if $|r| \geq 1$

$\left\{ \frac{1}{4} + \frac{-3}{16} + \frac{9}{64} + \frac{-27}{256} + \dots \right\}$ $|-\frac{3}{4}| = \frac{3}{4} < 1$
 convergent

$\boxed{\frac{1}{4}} \left[\boxed{\frac{3}{4}} + \frac{9}{16} - \frac{27}{64} + \dots \right]$ $\frac{1}{4} + \frac{1}{4} \Rightarrow \frac{1}{2} \Rightarrow \frac{1}{2} \left(\frac{4}{7} \right) \Rightarrow \boxed{\frac{1}{7}}$

CORE CONCEPT: CONVERGENCE/DIVER- GENCE

A series is said to be convergent if it approaches some limit. In mathematics, a series is the sum of the terms of an infinite sequence of numbers.

Given an infinite sequence $(a_1, a_2, a_3, \dots)$, the n th partial sum S_n is the sum of the first n terms of the sequence, that is,

$$S_n = \sum_{k=1}^n a_k$$

A series is convergent if the sequence of its partial sums $\{S_1, S_2, S_3, \dots\}$ tends to a limit; that means that the partial sums become closer and closer to a given number when the number of their terms increases.

Convergence, in mathematics, property (exhibited by certain infinite series and **functions**) of approaching a limit more and more closely as an argument (variable) of the function increases or decreases or as the number of terms of the series increases. For example, the function $y = 1/x$ converges to zero as x increases. Although no finite value of x will cause the value of y to actually become zero, the limiting value of y is zero because y can be made as small as desired by choosing x large enough. The line $y = 0$ (the x -axis) is called an asymptote of the function.

Similarly, for any value of x between (but not including) -1 and $+1$, the series $1 + x + x^2 + \dots + x^n$ converges toward the limit $1/(1 - x)$ as n , the number of terms, increases. The interval $-1 < x < 1$ is called the range of convergence of the series; for values of x outside this range, the series is said to diverge.

Convergence of Infinite Products

There is a simple convergence test for infinite products that you think deserves to be better known.

Theorem. Let a_n be a sequence of positive numbers. Then the infinite product

Function is originally the idealization of how a varying quantity depends on another quantity.

$$\prod_{n=1}^{\infty} (1 + a_n)$$

converges if and only if the series

$$\sum_{n=1}^{\infty} a_n$$

converges.

Proof: Taking the logarithm of the product gives the series

$$\sum_{n=1}^{\infty} \ln(1 + a_n)$$

whose convergence is equivalent to the convergence of the product. But observe that

$$\sum_{n=1}^{\infty} \lim_{x \rightarrow 0} \frac{\ln(1+x)}{x} = 1$$

If we assume that $a_n \rightarrow 0$, this gives us that

$$\lim_{n \rightarrow \infty} \frac{\ln(1 + a_n)}{a_n} = 1$$

and the theorem follows by the limit comparison test. Q.E.D.

Using this theorem, everything you know about infinite series translates directly to the world of infinite products. For example, the product

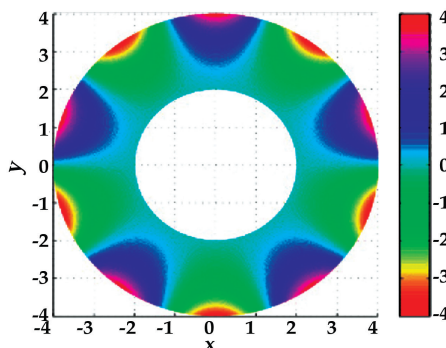
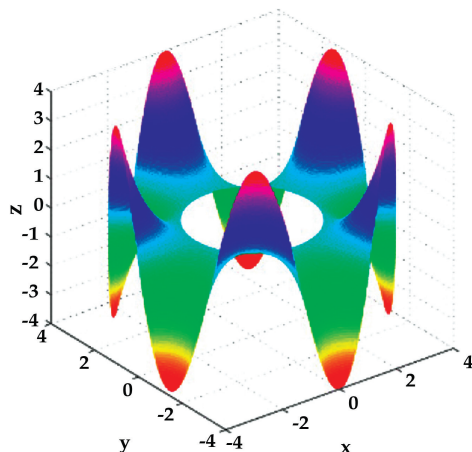
$$\prod_{n=1}^{\infty} \left(1 + \frac{1}{n^p} \right)$$

converges if and only if $p > 1$.

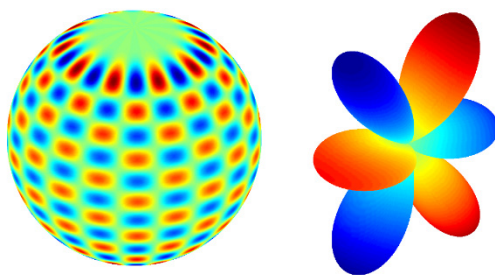
Before we learned this theorem, we had imagined that there must be an entire theory of convergence for infinite products, as complex and interesting as the theory of series from calculus, but completely unknown to me. Instead, it turns out that no one ever talks about the convergence of infinite products because there is basically nothing new to say.

The Harmonic Series

Harmonic function, mathematical function of two variables having the property that its value at any point is equal to the average of its values along any circle around that point, provided the function is defined within the circle. An infinite number of points are involved in this average, so that it must be found by means of an integral, which represents an infinite sum. In physical situations, harmonic functions describe those conditions of equilibrium such as the temperature or electrical charge distribution over a region in which the value at each point remains constant.



Harmonic functions can also be defined as functions that satisfy Laplace's equation, a condition that can be shown to be equivalent to the first definition. The surface defined by a harmonic function has zero convexity, and these functions thus have the important property that they have no maximum or minimum values inside the region in which they are defined. Harmonic functions are also analytic, which means that they possess all derivatives (are perfectly "smooth") and can be represented as polynomials with an infinite number of terms, called power series. Spherical harmonic functions arise when the spherical coordinate system is used. (In this system, a point in space is located by three coordinates, one representing the distance from the origin and two others representing the angles of elevation and azimuth, as in astronomy.) Spherical harmonic functions are commonly used to describe three-dimensional fields, such as gravitational, magnetic, and electrical fields, and those arising from certain types of fluid motion.



Another reason you like this theorem is that it gives a nice proof that the harmonic series diverges. According to the theorem, the behavior of the harmonic series is the same as the behavior of the following product:

$$(1+1)\left(1+\frac{1}{2}\right)\left(1+\frac{1}{3}\right)\left(1+\frac{1}{4}\right)\dots$$

But this is just

$$\frac{2}{1} \times \frac{3}{2} \times \frac{4}{3} \times \frac{5}{4} \times \dots$$

Sequence is an enumerated collection of objects in which repetitions are allowed. Like a set, it contains members. The number of elements is called the length of the sequence.

This clearly diverges, for the partial products are the **sequence** of positive integers.

Using the *n*th-Term Test for Divergence

The *n*th-term test for divergence is a very important test, as it enables you to identify lots of series as divergent. Fortunately, it is also very easy to use. To show you why this test works, the following sequence meets the necessary condition — that is, a sequence that does not approach 0:

$$\{a_n\} = \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \dots, \frac{n}{n+1}$$

Notice that the limit of the sequence is 1 rather than 0. So here's the related series:

$$\sum_{n=1}^{\infty} \frac{n}{n+1} = \frac{1}{2} + \frac{2}{3} + \frac{3}{4} + \dots$$

Because this series is the sum of an infinite number of terms that are very close to 1, it naturally produces an infinite sum, so it is divergent.

For example, the harmonic sequence

$$1, \frac{1}{2}, \frac{1}{3}, \dots$$

approaches 0, but the harmonic series

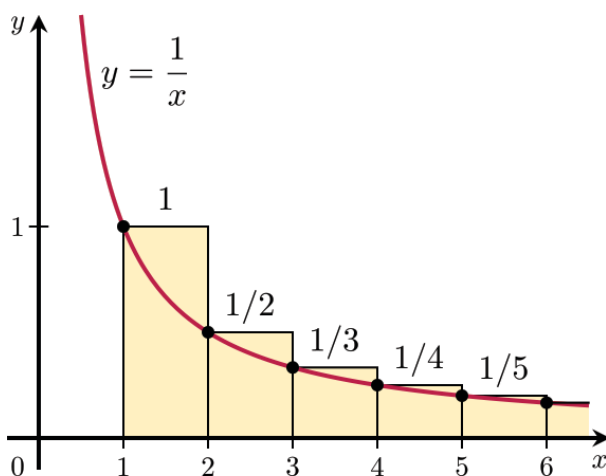
$$1 + \frac{1}{2} + \frac{1}{3} + \dots$$

is divergent.

Integral Test for Convergence

In mathematics, the integral test for convergence is a method used to test infinite series of non-negative terms for convergence. It was developed by Colin Maclaurin and Augustin-Louis Cauchy and is sometimes known as the Maclaurin–Cauchy test.

It is generally quite difficult, often impossible, to determine the value of a series exactly. In many cases it is possible at least to determine whether or not the series converges, and so we will spend most of our time on this problem. If all of the terms in a series are non-negative, then clearly the sequence of partial sums s_n is non-decreasing.



This means that if we can show that the sequence of partial sums is bounded, the series must converge. We know that if the series converges, the terms approach zero, but this does not mean that $a_n \geq a_{n+1}$ for every n . Many useful and interesting series do have this property, however, and they are among the easiest to understand. Let's look at an example.

Example

Show that $\sum_{n=1}^{\infty} \frac{1}{n^2}$ converges.

The terms $1/n^2$ are positive and decreasing, and since $\lim_{x \rightarrow \infty} 1/x^2 = 0$, the terms $1/n^2$ approach zero. We seek an upper bound for all the partial sums, that is, we want to find a number N so that $s_n \leq N$ for every n . The upper bound is provided courtesy of integration, and is inherent in the below figure.

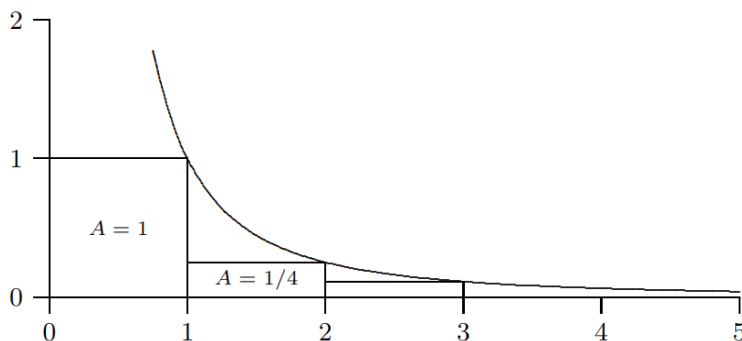


Figure 1: Graph of $y = 1/x^2$ with rectangles.

The figure shows the graph of $y = 1/x^2$ together with some rectangles that lie completely below the curve and that all have base length one. Because the heights of the rectangles are determined by the height of the curve, the areas of the rectangles are $1/1^2$, $1/2^2$, $1/3^2$, and so on—in other words, exactly the terms of the series. The partial sum s_n is simply the sum of the areas of the first n rectangles. Because the rectangles all lie between the curve and the x-axis, any sum of rectangle areas is less than the

corresponding area under the curve, and so of course any sum of rectangle areas is less than the area under the entire curve, that is, all the way to infinity. There is a bit of trouble at the left end, where there is an asymptote, but we can work around that easily. Here it is:

$$s_n = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \cdots + \frac{1}{n^2} < 1 + \int_1^n \frac{1}{x^2} dx < 1 + \int_1^\infty \frac{1}{x^2} dx = 1 + 1 = 2,$$

recalling that we computed this improper integral in section 9.7. Since the sequence of partial sums s_n is increasing and bounded above by 2, we know that $\lim_{n \rightarrow \infty} s_n = L < 2$, and so the series converges to some number less than 2. In fact, it is possible, though difficult, to show that $L = \pi^2/6 \approx 1.6$. We already know that $\sum 1/n$ diverges. What goes wrong if we try to apply this technique to it? Here's the calculation:

$$s_n = \frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} < 1 + \int_1^n \frac{1}{x} dx < 1 + \int_1^\infty \frac{1}{x} dx = 1 + \infty.$$

The problem is that the improper integral does not converge. Note well that this does not prove that $\sum 1/n$ diverges, just that this particular **calculation** fails to prove that it converges. A slight modification, however, allows us to prove in a second way that $\sum 1/n$ diverges.

Calculation is a deliberate process that transforms one or more inputs into one or more results, with variable change.

Theorem. A p -series with $p > 0$ converges if and only if $p > 1$.

Proof. We use the integral test; we have already done $p = 1$, so assume that $p \neq 1$.

$$\int_1^\infty \frac{1}{x^p} dx = \lim_{D \rightarrow \infty} \left. \frac{x^{1-p}}{1-p} \right|_1^D = \lim_{D \rightarrow \infty} \frac{D^{1-p}}{1-p} - \frac{1}{1-p}.$$

If $p > 1$ then $1 - p < 0$ and $\lim_{D \rightarrow \infty} D^{1-p} = 0$, so the integral converges. If $0 < p < 1$ then $1 - p > 0$ and $\lim_{D \rightarrow \infty} D^{1-p} = \infty$, so the integral diverges.

Example. Approximate $\sum 1/n^2$ to two decimal places.

Referring to above figure, if we approximate the sum by $\sum_{n=1}^N 1/n^2$, the error we make is the total area of the remaining rectangles, all of which lie under the curve $1/x^2$ from $x = N$ out to infinity. So we know the true value of the series is larger than the approximation, and no bigger than the approximation plus the area under the curve from N to infinity.

Roughly, then, we need to find N so that

$$\int_N^\infty \frac{1}{x^2} dx < 1/100.$$

We can compute the integral:

$$\int_N^\infty \frac{1}{x^2} dx = \frac{1}{N},$$

so $N = 100$ is a good starting point. Adding up the first 100 terms gives approximately 1.634983900, and that plus $1/100$ is 1.644983900, so approximating the series by the value halfway between these will be at most $1/200 = 0.005$ in error. The midpoint is 1.639983900, but while this is correct to ± 0.005 , we cannot tell if the correct two-decimal approximation is 1.63 or 1.64. We need to make N big enough to reduce the guaranteed error, perhaps to around 0.004 to be safe, so we would need $1/N \approx 0.008$, or $N = 125$. Now the sum of the first 125 terms is approximately 1.636965982, and that plus 0.008 is 1.644965982 and the point halfway between them is 1.640965982. The true value is then 1.640965982 ± 0.004 , and all numbers in this range round to 1.64, so 1.64 is correct to two decimal places. We have mentioned that the true value of this series can be shown to be $\pi^2/6 \approx 1.644934068$ which rounds down to 1.64 (just barely) and is indeed below the upper bound of 1.644965982, again just barely. Frequently approximations will be even better than the “guaranteed” accuracy, but not always, as this example demonstrates.

Alternating Series

Next we consider series with both positive and negative terms, but in a regular pattern: they alternate, as in the alternating harmonic series for example:

$$\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} = \frac{1}{1} + \frac{-1}{2} + \frac{1}{3} + \frac{-1}{4} + \cdots = \frac{1}{1} - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots.$$

In this series the sizes of the terms decrease, that is, $|a_n|$ forms a decreasing sequence, but this is not required in an alternating series. As with positive term series, however, when the terms do have decreasing sizes it is easier to analyze the series, much easier, in fact, than positive term series. Consider pictorially what is going on in the alternating harmonic series, shown in figure 2. Because the sizes of the terms a_n are decreasing, the partial sums s_1, s_3, s_5 , and so on, form a decreasing sequence that is bounded below by s_2 , so this sequence must converge. Likewise, the partial sums s_2, s_4, s_6 , and so on, form an increasing sequence that is bounded above by s_1 , so this sequence also converges. Since all the even numbered partial sums are less than all the odd numbered ones, and since the “jumps” (that is, the a_i terms) are getting smaller and smaller, the two sequences must converge to the same value, meaning the entire sequence of partial sums $s_1, s_2, s_3, \dots$ converges as well.

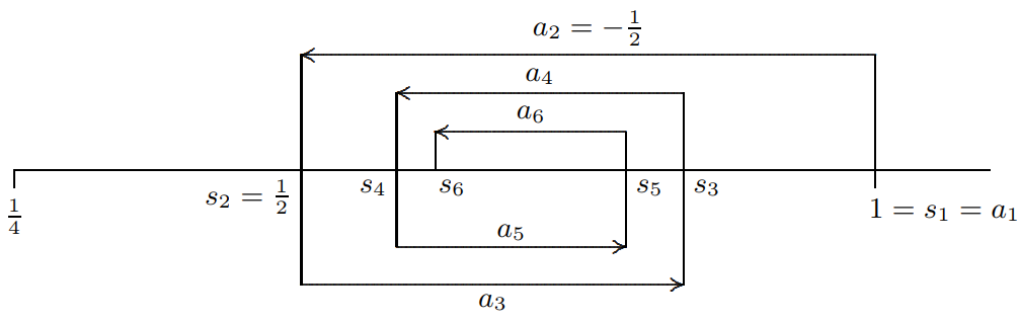


Figure 2: The alternating harmonic series.

There's nothing special about the alternating harmonic series—the same argument works for any alternating sequence with decreasing size terms. The alternating series test is worth calling a theorem.

Theorem. Suppose that $\{a_n\}_{n=1}^{\infty}$ is a non-increasing sequence of positive numbers and $\lim_{n \rightarrow \infty} a_n = 0$. Then the alternating series $\sum_{n=1}^{\infty} (-1)^{n-1} a_n$ converges.

Proof. The odd numbered partial sums, s_1, s_3, s_5 , and so on, form a non-increasing sequence, because $s_{2k+3} = s_{2k+1} - a_{2k+2} + a_{2k+3} \leq s_{2k+1}$, since $a_{2k+2} \geq a_{2k+3}$. This sequence is bounded below by s_2 , so it must converge, say $\lim_{k \rightarrow \infty} s_{2k+1} = L$. Likewise, the partial sums s_2, s_4, s_6 , and so on, form a non-decreasing sequence that is bounded above by s_1 , so this sequence also converges, say $\lim_{k \rightarrow \infty} s_{2k} = M$. Since $\lim_{n \rightarrow \infty} a_n = 0$ and $s_{2k+1} = s_{2k} + a_{2k+1}$,

$$L = \lim_{k \rightarrow \infty} s_{2k+1} = \lim_{k \rightarrow \infty} (s_{2k} + a_{2k+1}) = \lim_{k \rightarrow \infty} s_{2k} + \lim_{k \rightarrow \infty} a_{2k+1} = M + 0 = M,$$

so $L = M$, the two sequences of partial sums converge to the same limit, and this means the entire sequence of partial sums also converges to L .

Another useful fact is implicit in this discussion. Suppose that

$$L = \sum_{n=1}^{\infty} (-1)^{n-1} a_n$$

and that we approximate L by a finite part of this sum, say

$$L \approx \sum_{n=1}^N (-1)^{n-1} a_n.$$

Because the terms are decreasing in size, we know that the true value of L must be between this approximation and the next one, that is, between

$$\sum_{n=1}^N (-1)^{n-1} a_n \quad \text{and} \quad \sum_{n=1}^{N+1} (-1)^{n-1} a_n$$

Depending on whether N is odd or even, the second will be larger or smaller than the first.

Comparison Tests

As we begin to compile a list of convergent and divergent series, new ones can sometimes be analyzed by comparing them to ones that we already understand.

Comparison test can mean:

- Limit comparison test, a method of testing for the convergence of an infinite series.
- Direct comparison test, a way of deducing the convergence or **divergence** of an infinite series or an improper integral.

Divergence is a vector operator that produces a scalar field, giving the quantity of a vector field's source at each point.

Example

Does $\sum_{n=2}^{\infty} \frac{1}{n^2 \ln n}$ converge?

The obvious first approach, based on what we know, is the integral test. Unfortunately, we cannot compute the required anti-derivative. But looking at the series, it would appear that it must converge, because the terms we are adding are smaller than the terms of a p -series, that is,

$$\frac{1}{n^2 \ln n} < \frac{1}{n^2},$$

when $n \geq 3$. Since adding up the terms $1/n^2$ does not get “too big”, the new series “should” also converge. Let’s make this more precise.

The series $\sum_{n=2}^{\infty} \frac{1}{n^2 \ln n}$ converges if and only if $\sum_{n=3}^{\infty} \frac{1}{n^2 \ln n}$ converges—all we have done is dropped the initial term. We know that $\sum_{n=3}^{\infty} \frac{1}{n^2}$ converges. Looking at two typical partial sums:

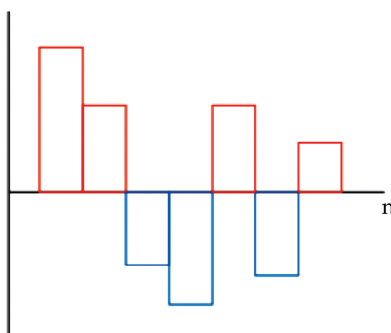
$$s_n = \frac{1}{3^2 \ln 3} + \frac{1}{4^2 \ln 4} + \frac{1}{5^2 \ln 5} + \cdots + \frac{1}{n^2 \ln n} < \frac{1}{3^2} + \frac{1}{4^2} + \frac{1}{5^2} + \cdots + \frac{1}{n^2} = t_n.$$

Since the p -series converges, say to L , and since the terms are positive, $t_n < L$. Since the terms of the new series are positive, the s_n form an increasing sequence and $s_n < t_n < L$ for all n . Hence the sequence $\{s_n\}$ is bounded and so converges. Sometimes, even when the integral test applies, comparison to

a known series is easier, so it is generally a good idea to think about doing a comparison before doing the integral test.

Absolute Convergence

In mathematics, an infinite series of numbers is said to converge absolutely (or to be absolutely convergent) if the sum of the absolute values of the summands is finite.



Roughly speaking there are two ways for a series to converge: As in the case of $\sum 1/n^2$, the individual terms get small very quickly, so that the sum of all of them stays finite, or, as in the case of $\sum (-1)^{n-1}/n$, the terms do not get small fast enough ($\sum 1/n$ diverges), but a mixture of positive and negative terms provides enough cancellation to keep the sum finite. You might guess from what we have seen that if the terms get small fast enough to do the job, then whether or not some terms are negative and some positive the series converges.

Theorem.

If $\sum_{n=0}^{\infty} |a_n|$ converges, then $\sum_{n=0}^{\infty} a_n$ converges.

Proof. Note that $0 \leq a_n + |a_n| \leq 2|a_n|$ so by the comparison test $\sum_{n=0}^{\infty} (a_n + |a_n|)$ converges.

Now

$$\sum_{n=0}^{\infty} (a_n + |a_n|) - \sum_{n=0}^{\infty} |a_n| = \sum_{n=0}^{\infty} a_n + |a_n| - |a_n| = \sum_{n=0}^{\infty} a_n$$

So given a series $\sum a_n$ with both positive and negative terms, you should first ask whether $\sum |a_n|$ converges. This may be an easier question to answer, because we have tests that apply specifically to series with non-negative terms. If $\sum |a_n|$ converges then you know that $\sum a_n$ converges as well. If $\sum |a_n|$ diverges then it still may be true that $\sum a_n$ converges—you will have to do more work to decide the question. Another way to think of this result is: it is (potentially) easier for $\sum a_n$ to converge than for $\sum |a_n|$ to converge, because the latter series cannot take advantage of cancellation.

If $\sum |a_n|$ converges we say that $\sum a_n$ converges absolutely; to say that $\sum a_n$ converges absolutely is to say that any cancellation that happens to come along is not really needed, as the terms already get small so fast that convergence is guaranteed by that

alone. If $\sum a_n$ converges but $\sum |a_n|$ does not, we say that $\sum a_n$ converges conditionally. For example $\sum_{n=1}^{\infty} (-1)^{n-1} \frac{1}{n^2}$ converges absolutely, while $\sum_{n=1}^{\infty} (-1)^{n-1} \frac{1}{n}$ converges conditionally.

The Ratio and Root Tests

Does the series $\sum_{n=0}^{\infty} \frac{n^5}{5^n}$ converge? It is possible, but a bit unpleasant, to approach this with the integral test or the comparison test, but there is an easier way. Consider what happens as we move from one term to the next in this series:

$$\cdots + \frac{n^5}{5^n} + \frac{(n+1)^5}{5^{n+1}} + \cdots$$

The denominator goes up by a factor of 5, $5^{n+1} = 5 \cdot 5^n$, but the numerator goes up by much less: $(n+1)^5 = n^5 + 5n^4 + 10n^3 + 10n^2 + 5n + 1$, which is much less than $5n^5$ when n is large, because $5n^4$ is much less than n^5 . So we might guess that in the long run it begins to look as if each term is $1/5$ of the above term. We have seen series that behave like this:

$$\sum_{n=0}^{\infty} \frac{1}{5^n} = \frac{5}{4},$$

a geometric series. So we might try comparing the given series to some variation of this geometric series. This is possible, but a bit messy. We can in effect do the same thing, but bypass most of the unpleasant work. The key is to notice that

$$\lim_{n \rightarrow \infty} \frac{a_{n+1}}{a_n} = \lim_{n \rightarrow \infty} \frac{(n+1)^5}{5^{n+1}} \frac{5^n}{n^5} = \lim_{n \rightarrow \infty} \frac{(n+1)^5}{n^5} \frac{1}{5} = 1 \cdot \frac{1}{5} = \frac{1}{5}.$$

This is really just what we noticed above, done a bit more officially: in the long run, each term is one fifth of the before term. Now pick some number between $1/5$ and 1 , say $1/2$. Because

$$\lim_{n \rightarrow \infty} \frac{a_{n+1}}{a_n} = \frac{1}{5},$$

then when n is big enough, say $n \geq N$ for some N ,

$$\frac{a_{n+1}}{a_n} < \frac{1}{2} \quad \text{and} \quad a_{n+1} < \frac{a_n}{2}.$$

So

$a_{N+1} < a_N/2$, $a_{N+2} < a_{N+1}/2 < a_N/4$, $a_{N+3} < a_{N+2}/2 < a_{N+1}/4 < a_N/8$ and so on.

The general form is $a_{N+k} < a_N/2^k$. So if we look at the series

$$\sum_{k=0}^{\infty} a_{N+k} = a_N + a_{N+1} + a_{N+2} + a_{N+3} + \cdots + a_{N+k} + \cdots,$$

its terms are less than or equal to the terms of the sequence

$$a_N + \frac{a_N}{2} + \frac{a_N}{4} + \frac{a_N}{8} + \cdots + \frac{a_N}{2^k} + \cdots = \sum_{k=0}^{\infty} \frac{a_N}{2^k} = 2a_N.$$

So by the comparison test, $\sum_{k=0}^{\infty} a_{N+k}$ converges, and this means that $\sum_{n=0}^{\infty} a_n$ converges, since we have just added the fixed number $a_0 + a_1 + \cdots + a_{N-1}$.

Under what circumstances could we do this? What was crucial was that the limit of a_{n+1}/a_n , say L , was less than 1 so that we could pick a value r so that $L < r < 1$. The fact that $L < r$ ($1/5 < 1/2$ in our example) means that we can compare the series $\sum a^n$ to $\sum r^n$, and the fact that $r < 1$ guarantees that $\sum r^n$ converges. That's really all that is required to make the argument work. We also made use of the fact that the terms of the series were positive; in general we simply consider the absolute values of the terms and we end up testing for absolute convergence.

Theorem. The Ratio Test

Suppose that $\lim_{n \rightarrow \infty} |a_{n+1}/a_n| = L$. If $L < 1$ the series $\sum a_n$ converges absolutely, if $L > 1$ the series diverges, and if $L = 1$ this test gives no information.

Proof. The example above essentially proves the first part of this, if we simply replace $1/5$ by L and $1/2$ by r . Suppose that $L > 1$, and pick r so that $1 < r < L$. Then for $n \geq N$, for some N ,

$$\frac{|a_{n+1}|}{|a_n|} > r \quad \text{and} \quad |a_{n+1}| > r|a_n|.$$

This implies that $|a_{N+k}| > r^k |a_N|$, but since $r > 1$ this means that $\lim_{k \rightarrow \infty} |a_{N+k}| \neq 0$, which means also that $\lim_{n \rightarrow \infty} a_n \neq 0$. By the divergence test, the series diverges. To see that we get no information when $L = 1$, we need to exhibit two series with $L = 1$, one that converges and one that diverges. It is easy to see that $\sum 1/n^2$ and $\sum 1/n$ do the job.

Theorem. The Root Test

Suppose that $\lim_{n \rightarrow \infty} |a_n|^{1/n} = L$. If $L < 1$ the series $\sum a_n$ converges absolutely, if $L > 1$ the series diverges, and if $L = 1$ this test gives no information.

The proof of the root test is actually easier than that of the ratio test, and is a good exercise.

Example

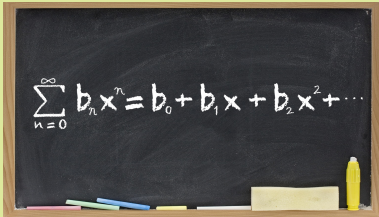
Analyze $\sum_{n=0}^{\infty} \frac{5^n}{n^n}$.

The ratio test turns out to be a bit difficult on this series (try it). Using the root test:

$$\lim_{n \rightarrow \infty} \left(\frac{5^n}{n^n} \right)^{1/n} = \lim_{n \rightarrow \infty} \frac{(5^n)^{1/n}}{(n^n)^{1/n}} = \lim_{n \rightarrow \infty} \frac{5}{n} = 0.$$

Since $0 < 1$, the series converges.

The root test is frequently useful when n appears as an exponent in the general term of the series.



CORE CONCEPT: POWER SERIES

A power series (in one variable) is an infinite series of the form $f(x) = \sum_{n=0}^{\infty} a_n (s-c)^n$, where a_n is the coefficient of the n th term and x varies around c .

Power series, in mathematics, an infinite series that can be thought of as a polynomial with an infinite number of terms, such as $1 + x + x^2 + x^3 + \dots$. Usually, a given power series will converge (that is, approach a finite sum) for all values of x within a certain interval around zero—in particular, whenever the absolute value of x is less than some positive number r , known as the radius of convergence. Outside of this interval the series diverges (is infinite), while the series may converge or diverge when $x = \pm r$. The radius of convergence can often be determined by a version of the **ratio test** for power series: given a general power series $a_0 + a_1 x + a_2 x^2 + \dots$, in which the coefficients are known, the radius of convergence is equal to the limit of the ratio of successive coefficients.

Symbolically, the series will converge for all values of x such that

$$|x| < \lim_{n \rightarrow \infty} |a_n / a_{n+1}| = r$$

For instance, the infinite series $1 + x + x^2 + x^3 + \dots$ has a radius of convergence of 1 (all the coefficients are 1)—that is, it converges for all $-1 < x < 1$ —and within that interval the infinite series is equal to $1/(1-x)$. Applying the ratio test to the series $1 + x/1! + x^2/2! + x^3/3! + \dots$ (in which the factorial notation $n!$ means the product of the counting numbers from 1 to n) gives a radius of convergence of

Ratio test is a test for the convergence of a series. Where each term is a real or complex number and a_n is nonzero when n is large.

$$\lim_{n \rightarrow \infty} |(1/n!) / (1/(n+1)!)|$$

$$\lim_{n \rightarrow \infty} |n+1| = \infty$$

so that the series converges for any value of x .

Most functions can be represented by a power series in some interval. Although a series may converge for all values of x , the convergence may be so slow for some values that using it to approximate a function will require calculating too many terms to make it useful. Instead of powers of x , sometimes a much faster convergence occurs for powers of $(x - c)$, where c is some value near the desired value of x . Power series have also been used for calculating constants such as π and the natural logarithm base e and for solving differential equations.

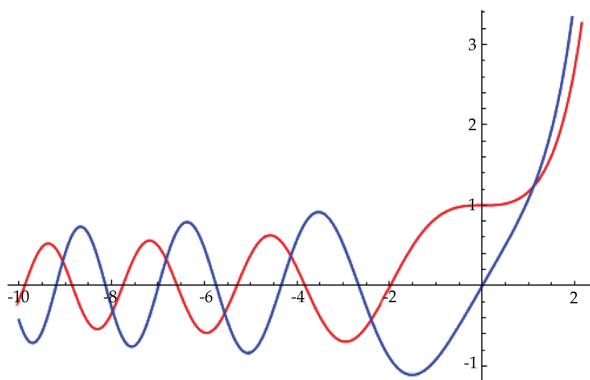
Recall that we were able to analyze all geometric series “simultaneously” to discover that

$$\sum_{n=0}^{\infty} kx^n = \frac{k}{1-x},$$

if $|x| < 1$, and that the series diverges when $|x| \geq 1$. At the time, we thought of x as an unspecified constant, but we could just as well think of it as a variable, in which case the series

$$\sum_{n=0}^{\infty} kx^n$$

is a function, namely, the function $k/(1-x)$, as long as $|x| < 1$. While $k/(1-x)$ is a reasonably easy function to deal with, the more complicated $\sum kx^n$ does have its attractions: it appears to be an infinite version of one of the simplest function types—a polynomial. This leads naturally to the questions: Do other functions have representations as series? Is there an advantage to viewing them in this way?



The geometric series has a special feature that makes it unlike a typical polynomial—the coefficients of the powers of x are the same, namely k . We will need to allow more general coefficients if we are to get anything other than the geometric series.

Definition. A power series has the form

$$\sum_{n=0}^{\infty} a_n x^n,$$

with the understanding that a_n may depend on n but not on x .

Example

$\sum_{n=1}^{\infty} \frac{x^n}{n}$ is a power series. We can investigate convergence using the ratio test:

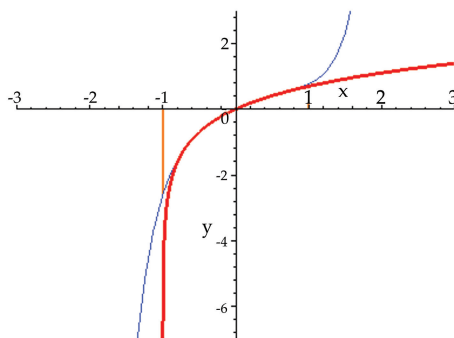
$$\lim_{n \rightarrow \infty} \frac{|x|^{n+1}}{n+1} \frac{n}{|x|^n} = \lim_{n \rightarrow \infty} |x| \frac{n}{n+1} = |x|.$$

Thus when $|x| < 1$ the series converges and when $|x| > 1$ it diverges, leaving only two values in doubt. When $x = 1$ the series is the harmonic series and diverges; when $x = -1$ it is the alternating harmonic series (actually the negative of the usual alternating

harmonic series) and converges. Thus, we may think of $\sum_{n=1}^{\infty} \frac{x^n}{n}$ as a function from the interval $[-1, 1)$ to the real numbers. A bit of thought reveals that the ratio test applied to a power series will always have the same nice form. In general, we will compute

$$\lim_{n \rightarrow \infty} \frac{|a_{n+1}| |x|^{n+1}}{|a_n| |x|^n} = \lim_{n \rightarrow \infty} |x| \frac{|a_{n+1}|}{|a_n|} = |x| \lim_{n \rightarrow \infty} \frac{|a_{n+1}|}{|a_n|} = L|x|,$$

assuming that $\lim |a_{n+1}|/|a_n|$ exists. Then the series converges if $L|x| < 1$, that is, if $|x| < 1/L$, and diverges if $|x| > 1/L$. Only the two values $x = \pm 1/L$ require further investigation. Thus the series will definitely define a function on the interval $(-1/L, 1/L)$, and perhaps will extend to one or both endpoints as well. Two special cases deserve mention: if $L = 0$ the limit is 0 no matter what value x takes, so the series converges for all x and the function is defined for all real numbers. If $L = \infty$, then no matter what value x takes the limit is infinite and the series converges only when $x = 0$. The value $1/L$ is called the radius of convergence of the series, and the interval on which the series converges is the interval of convergence.



Consider again the geometric series,

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}.$$

Whatever benefits there might be in using the series form of this function are only available to us when x is between -1 and 1 . Frequently we can address this shortcoming by modifying the power series slightly. Consider this series:

$$\sum_{n=0}^{\infty} \frac{(x+2)^n}{3^n} = \sum_{n=0}^{\infty} \left(\frac{x+2}{3} \right)^n = \frac{1}{1 - \frac{x+2}{3}} = \frac{3}{1-x},$$

because this is just a geometric series with x replaced by $(x+2)/3$. Multiplying both sides by $1/3$ gives

$$\sum_{n=0}^{\infty} \frac{(x+2)^n}{3^{n+1}} = \frac{1}{1-x},$$

the same function as before. For what values of x does this series converge? Since it is a geometric series, we know that it converges when

$$|x+2|/3 < 1$$

$$|x+2| < 3$$

$$-3 < x+2 < 3$$

$$-5 < x < 1.$$

So we have a series representation for $1/(1-x)$ that works on a larger interval than before, at the expense of a somewhat more complicated series. The endpoints of the interval of convergence now are -5 and 1 , but note that they can be more compactly described as -2 ± 3 . We say that 3 is the **radius of convergence**, and we now say that the series is centered at -2 .

Radius of convergence of a power series is the radius of the largest disk in which the series converges.

Calculus with Power Series

Now we know that some functions can be expressed as power series, which look like infinite polynomials. Since calculus, that is, computation of derivatives and anti-derivatives, is easy for polynomials, the obvious question is whether the same is true for infinite series. The answer is yes:

Theorem. Suppose the power series $f(x) = \sum_{n=0}^{\infty} a_n(x-a)^n$ has radius of convergence R . Then

$$f'(x) = \sum_{n=0}^{\infty} n a_n (x-a)^{n-1},$$

$$\int f(x) dx = C + \sum_{n=0}^{\infty} \frac{a_n}{n+1} (x-a)^{n+1},$$

and these two series have radius of convergence R as well.

Example. Starting with the geometric series:

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n$$

$$\int \frac{1}{1-x} dx = -\ln|1-x| = \sum_{n=0}^{\infty} \frac{1}{n+1} x^{n+1}$$

$$\ln|1-x| = \sum_{n=0}^{\infty} -\frac{1}{n+1} x^{n+1}$$

when $|x| < 1$. The series does not converge when $x = 1$ but does converge when $x = -1$ or $1-x = 2$. The interval of convergence is $[-1, 1)$, or $0 < 1-x \leq 2$, so we can use the series to represent $\ln(x)$ when $0 < x \leq 2$. For example

$$\ln(3/2) = \ln(1 - -1/2) = \sum_{n=0}^{\infty} (-1)^n \frac{1}{n+1} \frac{1}{2^{n+1}}$$

and so

$$\ln(3/2) \approx \frac{1}{2} - \frac{1}{8} + \frac{1}{24} - \frac{1}{64} + \frac{1}{160} - \frac{1}{384} + \frac{1}{896} = \frac{909}{2240} \approx 0.406.$$

Because this is an alternating series with decreasing terms, we know that the true value is between $909/2240$ and $909/2240 - 1/2048 = 29053/71680 \approx .4053$, so correct to two decimal places the value is 0.41. What about $\ln(9/4)$? Since $9/4$ is larger than 2 we cannot use the series directly, but

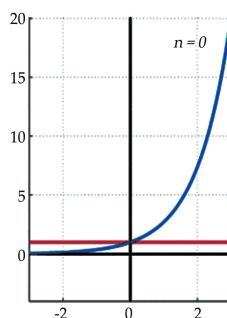
$$\ln(9/4) = \ln((3/2)^2) = 2 \ln(3/2) \approx 0.82,$$

so in fact we get a lot more from this one calculation than first meets the eye. To estimate the true value accurately we actually need to be a bit more careful. When we multiply by two we know that the true value is between 0.8106 and 0.812, so rounded to two decimal places the true value is 0.81.

Taylor Series

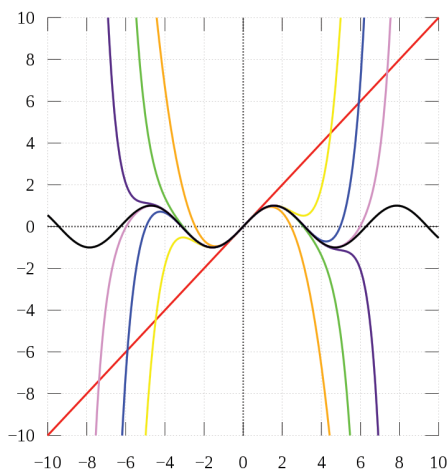
In mathematics, a Taylor series is a representation of a function as an infinite sum of terms that are calculated from the values of the function's derivatives at a single point. The concept of a Taylor series was formulated by the Scottish mathematician James Gregory and formally introduced by the English mathematician Brook Taylor in 1715. If the Taylor series is centered at zero, then that series is also called a Maclaurin series,

named after the Scottish mathematician Colin Maclaurin, who made extensive use of this special case of Taylor series in the 18th century.



A function can be approximated by using a finite number of terms of its Taylor series. Taylor's theorem gives quantitative estimates on the error introduced by the use of such an approximation. The polynomial formed by taking some initial terms of the Taylor series is called a Taylor polynomial. The Taylor series of a function is the limit of that function's Taylor polynomials as the degree increases, provided that the limit exists. A function may not be equal to its Taylor series, even if its Taylor series converges at every point. A function that is equal to its Taylor series in an open interval (or a disc in the complex plane) is known as an analytic function in that interval.

Geometric series
is a series with
a constant ratio
between successive
terms.



We have seen that some functions can be represented as series, which may give valuable information about the function. So far, we have seen only those examples that result from manipulation of our one fundamental example, the **geometric series**. We would like to start with a given function and produce a series to represent it, if possible.

$$f(x) = \sum_{n=0}^{\infty} a_n x^n$$

Suppose that on some interval of convergence. Then we know that we can compute derivatives of f by taking derivatives of the terms of the series. Let's look at the first few in general:

$$f'(x) = \sum_{n=1}^{\infty} n a_n x^{n-1} = a_1 + 2a_2x + 3a_3x^2 + 4a_4x^3 + \dots$$

$$f''(x) = \sum_{n=2}^{\infty} n(n-1)a_n x^{n-2} = 2a_2 + 3 \cdot 2a_3x + 4 \cdot 3a_4x^2 + \dots$$

$$f'''(x) = \sum_{n=3}^{\infty} n(n-1)(n-2)a_n x^{n-3} = 3 \cdot 2a_3 + 4 \cdot 3 \cdot 2a_4x + \dots$$

By examining these it is not hard to discern the general pattern. The k th derivative must be

$$\begin{aligned} f^{(k)}(x) &= \sum_{n=k}^{\infty} n(n-1)(n-2) \cdots (n-k+1) a_n x^{n-k} \\ &= k(k-1)(k-2) \cdots (2)(1)a_k + (k+1)(k) \cdots (2)a_{k+1}x + \\ &\quad + (k+2)(k+1) \cdots (3)a_{k+2}x^2 + \dots \end{aligned}$$

We can shrink this quite a bit by using factorial notation:

$$f^{(k)}(x) = \sum_{n=k}^{\infty} \frac{n!}{(n-k)!} a_n x^{n-k} = k!a_k + (k+1)!a_{k+1}x + \frac{(k+2)!}{2!}a_{k+2}x^2 + \dots$$

Now substitute $x = 0$:

$$f^{(k)}(0) = k!a_k + \sum_{n=k+1}^{\infty} \frac{n!}{(n-k)!} a_n 0^{n-k} = k!a_k,$$

and solve for a_k :

$$a_k = \frac{f^{(k)}(0)}{k!}.$$

Note the special case, obtained from the series for f itself, that gives $f(0) = a_0$.

So if a function f can be represented by a series, we know just what series it is. Given a function f , the series

$$\sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} x^n$$

is called the Maclaurin series for f .

Example. Find the Maclaurin series for $f(x) = 1/(1-x)$. We need to compute the derivatives of f (and hope to spot a pattern).

$$\begin{aligned}
 f(x) &= (1-x)^{-1} \\
 f'(x) &= (1-x)^{-2} \\
 f''(x) &= 2(1-x)^{-3} \\
 f'''(x) &= 6(1-x)^{-4} \\
 f^{(4)}(x) &= 4!(1-x)^{-5} \\
 &\vdots \\
 f^{(n)}(x) &= n!(1-x)^{-n-1}
 \end{aligned}$$

So

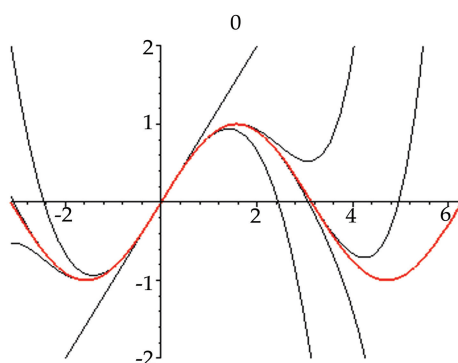
$$\frac{f^{(n)}(0)}{n!} = \frac{n!(1-0)^{-n-1}}{n!} = 1$$

and the Maclaurin series is

$$\sum_{n=0}^{\infty} 1 \cdot x^n = \sum_{n=0}^{\infty} x^n,$$

the geometric series.

A warning is in order here. Given a function f we may be able to compute the Maclaurin series, but that does not mean we have found a series representation for f . We still need to know where the series converges, and if, where it converges, it converges to $f(x)$. While for most commonly encountered functions the Maclaurin series does indeed converge to f on some interval, this is not true of all functions, so care is required.



As a practical matter, if we are interested in using a series to approximate a function, we will need some finite number of terms of the series. Even for functions with messy derivatives we can compute these using computer software like Sage. If we want to know the whole series, that is, a typical term in the series, we need a function whose derivatives fall into a pattern that we can discern. A few of the most important functions are fortunately very easy.

Example. Find the Maclaurin series for $\sin x$

The derivatives are quite easy: $f'(x) = \cos x$, $f''(x) = -\sin x$, $f'''(x) = -\cos x$, $f^{(4)}(x) = \sin x$, and then the pattern repeats. We want to know the derivatives at zero: 1, 0, -1, 0, 1, 0, -1, 0, . . . , and so the Maclaurin series is

$$x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!}.$$

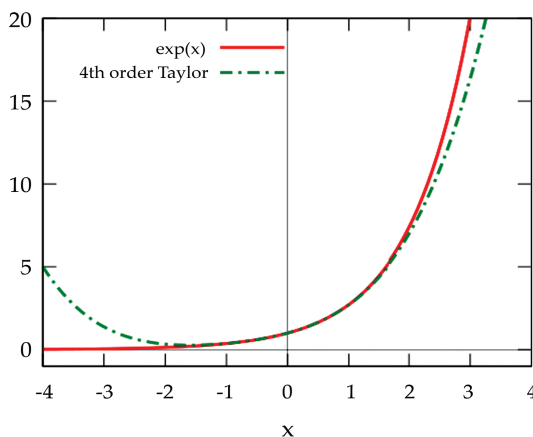
We should always determine the radius of convergence:

$$\lim_{n \rightarrow \infty} \frac{|x|^{2n+3}}{(2n+3)!} \frac{(2n+1)!}{|x|^{2n+1}} = \lim_{n \rightarrow \infty} \frac{|x|^2}{(2n+3)(2n+2)} = 0,$$

so the series converges for every x . Since it turns out that this series does indeed converge to $\sin x$ everywhere, we have a series representation for $\sin x$ for every x . Here is an interactive plot of the sine and some of its series approximations. Sometimes the formula for the n th derivative of a function f is difficult to discover, but a combination of a known Maclaurin series and some algebraic manipulation leads easily to the Maclaurin series for f .

Taylor's Theorem

In calculus, Taylor's theorem gives an approximation of a k -times differentiable function around a given point by a k -th order Taylor polynomial. For analytic functions the Taylor polynomials at a given point are finite order truncations of its Taylor series, which completely determines the function in some neighborhood of the point. The exact content of "Taylor's theorem" is not universally agreed upon. Indeed, there are several versions of it applicable in different situations, and some of them contain explicit estimates on the approximation error of the function by its Taylor polynomial.



One of the most important uses of infinite series is the potential for using an initial portion of the series for f to approximate f . We have seen, for example, that when we add up the first n terms of an alternating series with decreasing terms that

the difference between this and the true value is at most the size of the next term. A similar result is true of many Taylor series.

Theorem. Suppose that f is defined on some open interval I around a and suppose $f^{(N+1)}(x)$ exists on this interval. Then for each $x \neq a$ in I there is a value z between x and a so that

$$f(x) = \sum_{n=0}^N \frac{f^{(n)}(a)}{n!} (x-a)^n + \frac{f^{(N+1)}(z)}{(N+1)!} (x-a)^{N+1}.$$

Proof. The proof requires some cleverness to set up, but then the details are quite elementary. We want to define a function $F(t)$. Start with the equation

$$F(t) = \sum_{n=0}^N \frac{f^{(n)}(t)}{n!} (x-t)^n + B(x-t)^{N+1}.$$

Here we have replaced a by t in the first $N+1$ terms of the Taylor series, and added a carefully chosen term on the end, with B to be determined. Note that we are temporarily keeping x fixed, so the only variable in this equation is t , and we will be interested only in t between a and x . Now substitute $t = a$:

$$F(a) = \sum_{n=0}^N \frac{f^{(n)}(a)}{n!} (x-a)^n + B(x-a)^{N+1}.$$

Set this equal to $f(x)$:

$$f(x) = \sum_{n=0}^N \frac{f^{(n)}(a)}{n!} (x-a)^n + B(x-a)^{N+1}.$$

Since $x \neq a$, we can solve this for B , which is a “constant”—it depends on x and a but those are temporarily fixed. Now we have defined a function $F(t)$ with the property that $F(a) = f(x)$. Consider also $F(x)$: all terms with a positive power of $(x-t)$ become zero when we substitute x for t , so we are left with $F(x) = f(0)(x)/0! = f(x)$. So $F(t)$ is a function with the same value on the endpoints of the interval $[a, x]$. By Rolle’s Theorem, we know that there is a value $z \in (a, x)$ such that $F'(z) = 0$. Let’s look at $F'(t)$. Each term in $F(t)$, except the first term and the extra term involving B , is a product, so to take the derivative we use the product rule on each of these terms. It will help to write out the first few terms of the definition:

$$\begin{aligned} F(t) = f(t) + \frac{f^{(1)}(t)}{1!}(x-t)^1 + \frac{f^{(2)}(t)}{2!}(x-t)^2 + \frac{f^{(3)}(t)}{3!}(x-t)^3 + \cdots \\ + \frac{f^{(N)}(t)}{N!}(x-t)^N + B(x-t)^{N+1}. \end{aligned}$$

Now take the derivative:

$$\begin{aligned}
F'(t) = f'(t) &+ \left(\frac{f^{(1)}(t)}{1!} (x-t)^0 (-1) + \frac{f^{(2)}(t)}{1!} (x-t)^1 \right) \\
&+ \left(\frac{f^{(2)}(t)}{1!} (x-t)^1 (-1) + \frac{f^{(3)}(t)}{2!} (x-t)^2 \right) \\
&+ \left(\frac{f^{(3)}(t)}{2!} (x-t)^2 (-1) + \frac{f^{(4)}(t)}{3!} (x-t)^3 \right) + \dots + \\
&+ \left(\frac{f^{(N)}(t)}{(N-1)!} (x-t)^{N-1} (-1) + \frac{f^{(N+1)}(t)}{N!} (x-t)^N \right) \\
&+ B(N+1)(x-t)^N (-1).
\end{aligned}$$

Now most of the terms in this expression cancel out, leaving just

$$F'(t) = \frac{f^{(N+1)}(t)}{N!} (x-t)^N + B(N+1)(x-t)^N (-1).$$

At some z , $F'(z) = 0$ so

$$\begin{aligned}
0 &= \frac{f^{(N+1)}(z)}{N!} (x-z)^N + B(N+1)(x-z)^N (-1) \\
B(N+1)(x-z)^N &= \frac{f^{(N+1)}(z)}{N!} (x-z)^N \\
B &= \frac{f^{(N+1)}(z)}{(N+1)!}.
\end{aligned}$$

Now we can write

$$F(t) = \sum_{n=0}^N \frac{f^{(n)}(t)}{n!} (x-t)^n + \frac{f^{(N+1)}(z)}{(N+1)!} (x-t)^{N+1}.$$

Recalling that $F(a) = f(x)$ we get

$$f(x) = \sum_{n=0}^N \frac{f^{(n)}(a)}{n!} (x-a)^n + \frac{f^{(N+1)}(z)}{(N+1)!} (x-a)^{N+1},$$

which is what we wanted to show. It may not be immediately obvious that this is particularly useful.

CONCEPT: POWER SERIES

In calculus, a power series is a representation of a function as an infinite sum of terms, where each term is a constant multiplied by a variable raised to a non-negative integer exponent. Power series are a fundamental concept in analysis and have numerous applications in mathematics, physics, engineering, and other fields.

A power series is typically expressed in the following form:

$$f(x) = c_0 + c_1x + c_2x^2 + c_3x^3 + \dots = \sum c_n x^n$$

Here:

- $f(x)$ is the function being represented.
- x is the variable.
- $c_0, c_1, c_2, \dots$ are constants called the coefficients of the power series.
- n is a non-negative integer exponent.

Power series are often used to approximate more complex functions with simpler ones, especially when those functions are difficult to work with directly. The idea is that by taking a finite number of terms from the series, you can approximate the behavior of the function within a certain interval around a specific point.

The convergence of a power series is crucial. A power series may converge to the function it represents for certain values of x , and diverge (not give accurate approximations) for others. The interval of convergence is the range of values of x for which the series converges to the actual function. The radius of convergence determines how far from the center of the series (often denoted as $x = 0$) the series remains convergent.

The most well-known example of a power series is the Maclaurin series, which is a specific type of power series centered at $x = 0$. The Maclaurin series of a function is its power series expansion around that point.

For example, the Maclaurin series of the exponential function e^x is:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

Power series are extensively used in calculus to express functions like trigonometric functions, exponential functions, logarithmic functions, and more complex functions as approximations based on simpler polynomial terms. They also play a significant role in solving differential equations and analyzing the behavior of functions around specific points.

SUMMARY

- An infinite sequence (sometimes just called a sequence) is a function with a domain of all positive integers. In beginning calculus, the range of an infinite sequence is usually the set of real numbers, although it's also possible for the range to include complex numbers.
- Infinite sequences of functions have more subtle convergence properties. In addition to point wise convergence of the sequence of values of the functions taken at a single point, there is a concept of uniform convergence on an interval of the real axis, or in a region of the complex plane.
- A real number is a value of a continuous quantity that can represent a distance along a line. The numbers we deal with in the real world (ignoring any units that go along with them, such as dollars, inches, degrees, etc.) are usually real numbers.
- Convergence, in mathematics, property (exhibited by certain infinite series and functions) of approaching a limit more and more closely as an argument (variable) of the function increases or decreases or as the number of terms of the series increases
- Power series, in mathematics, an infinite series that can be thought of as a polynomial with an infinite number of terms, such as $1 + x + x^2 + x^3 + \dots$. Usually, a given power series will converge (that is, approach a finite sum) for all values of x within a certain interval around zero—in particular, whenever the absolute value of x is less than some positive number r , known as the radius of convergence.
- A Taylor series is a representation of a function as an infinite sum of terms that are calculated from the values of the function's derivatives at a single point.

MULTIPLE CHOICE QUESTIONS

1. A series can also be denoted by symbol _____
 - a. πa_n
 - b. $\sum a_n$
 - c. Φa_n
 - d. θa_n
2. **1+2+3+4 or 10 is a series?**
 - a. 1+2+3+4 only
 - b. 10 only
 - c. 1+2+3+4 and 10
 - d. Neither 1+2+3+4 nor 10
3. **Find sum of series 2+3+5+7.**
 - a. 5
 - b. 10
 - c. 17
 - d. infinite
4. **Find the sum of first 5 terms of series 2+4+6+.....**
 - a. 14
 - b. 16
 - c. 20
 - d. 30
5. **The sequence is same as series.**
 - a. True
 - b. False

Answers

1. (b) 2. (a) 3. (c) 4. (d) 5. (a)

REVIEW QUESTIONS

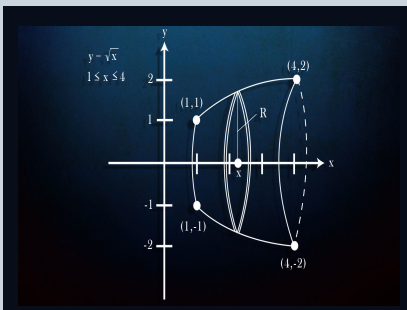
1. Explain the difference between real number and complex number.
2. Find a sequence a_n of real numbers such that $\sum a_n$ converges but $\prod(1+a_n)$ diverges.
3. Find an N so that $\sum_{n=1}^{\infty} \frac{1}{n^4}$ is between $\sum_{n=1}^N \frac{1}{n^4}$ and $\sum_{n=1}^N \frac{1}{n^4} + 0.005$.
4. Compute $\lim_{n \rightarrow \infty} |a_{n+1}/a_n|$ for the series $\sum 1/n^2$.
5. Compute $\lim_{n \rightarrow \infty} |a_{n+1}/a_n|$ for the series $\sum 1/n$.
6. Find a power series representation for $1/(1-x)^3$. What is the radius of convergence?
7. Find the first four terms of the Maclaurin series for $\tan x$ (up to and including the x^3 term).
8. Use a combination of Maclaurin series and algebraic manipulation to find a series centered at zero for $x \cos(x^2)$.
9. Drive Taylor's theorem.

REFERENCES

1. D'Angelo, J. P. and West, D. B. *Mathematical Thinking: Problem-Solving and Proofs*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 2000.
2. Dawikins, Paul. "Series and Sequences". Paul's Online Math Notes/Calc II (notes). Archived from the original on 30 November 2012. Retrieved 18 December 2012.
3. Falcon, Sergio (2003). "Fibonacci's multiplicative sequence". *International Journal of Mathematical Education in Science and Technology*. 34 (2): 310–315.
4. Gasper, G., Rahman, M. (2004). *Basic hypergeometric series*. Cambridge University Press.
5. Higham, N. J. (2008). *Functions of matrices: theory and computation*. Society for Industrial and Applied Mathematics.
6. Higham, N. J. (2009). The scaling and squaring method for the matrix exponential revisited. *SIAM review*, 51(4), 747-764.
7. Johansson, F. (2016). Computing hypergeometric functions rigorously. arXiv preprint arXiv:1606.06977.
8. Narici, Lawrence; Beckenstein, Edward (2011). *Topological Vector Spaces*. Pure and applied mathematics (Second ed.). Boca Raton, FL: CRC Press.

CHAPTER 2

CURVE SKETCHING



Curve sketching (or curve tracing) includes techniques that can be used to produce a rough idea of overall shape of a plane curve given its equation without computing the large numbers of points required for a detailed plot. It is an application of the theory of curves to find their main features.

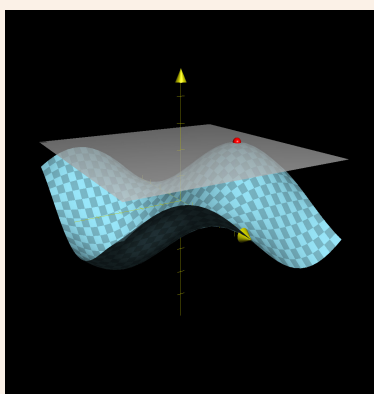
After studying this chapter, you will understand the core concepts of:

- Maxima and Minima
- Polar Coordinate System

INTRODUCTION

Curve sketching is the process of using algebra and calculus to find the features of a function and then plotting those features in order to draw the function on a coordinate plane. Sketching the curve is useful in calculus because it allows students to visualize functions and their features.

Curve sketching is another practical application of differential calculus. Whether we are interested in a function as a purely mathematical object or in connection with some application to the real world, it is often useful to know what the graph of the function looks like. We can obtain a good picture of the graph using certain crucial information provided by derivatives of the function and certain limits.



CORE CONCEPT: MAXIMA AND MINIMA

The maxima and minima (the respective plurals of maximum and minimum) of a function, known collectively as extrema (the plural of extremum), are the largest and smallest value of the function, either within a given range (the local or relative extrema) or on the entire domain of a function (the global or absolute extrema).

A local maximum point on a function is a point (x, y) on the graph of the function whose **y coordinate** is larger than all other y coordinates on the graph at points “close to” (x, y) . More precisely, $(x, f(x))$ is a local maximum if there is an interval (a, b) with $a < x < b$ and $f(x) \geq f(z)$ for every z in (a, b) . Similarly, (x, y) is a local minimum point if it has locally the smallest y coordinate. Again being more precise: $(x, f(x))$ is a local minimum if there is an interval (a, b) with $a < x < b$ and $f(x) \leq f(z)$ for every z in (a, b) . A local extremum is either a local minimum or a local maximum.

Local maximum and minimum points are quite distinctive on the graph of a function, and are therefore useful in understanding the shape of the graph. In many applied problems we want to find the largest or smallest value that a function achieves (for example, we might want to find the minimum cost at which some task can be performed) and so identifying maximum and minimum points will be useful for applied problems as well. Some examples of local maximum and minimum points are shown in figure 1.

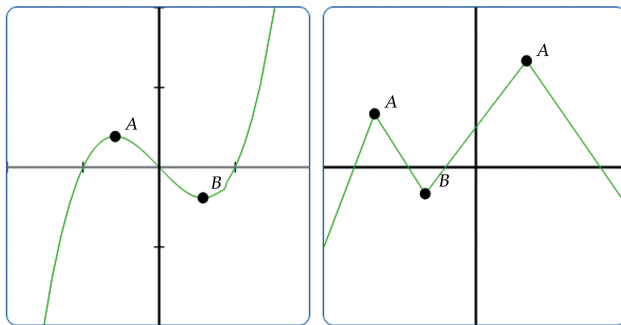


Figure 1: Some local maximum points (A) and minimum points (B).

If $(x, f(x))$ is a point where $f(x)$ reaches a local maximum or minimum, and if the derivative of f exists at x , then the graph has a tangent line and the tangent line must be horizontal. This is important enough to state as a theorem, though we will not prove it.

Theorem: (Fermat's Theorem) If $f(x)$ has a local extremum at $x=a$ and f is differentiable at a , then $f'(a)=0$.

Thus, the only points at which a function can have a local maximum or minimum are points at which the derivative is zero, as in the left hand graph in figure 1, or the derivative is undefined, as in the right hand graph. Any value of x for which $f'(x)$ is zero or undefined is called a critical value for f . When looking for local maximum and minimum points, you are likely to make two sorts of mistakes: You may forget that a maximum or minimum can occur where the derivative does not exist, and so forget to check whether the derivative exists everywhere. You might also assume that any place that the derivative is zero is a local maximum or minimum point, but this is not true. A portion of the graph of $f(x)=x^3$ is shown in figure 2. The derivative of f is $f'(x)=3x^2$, and $f'(0)=0$, but there is neither a maximum nor minimum at $(0,0)$.

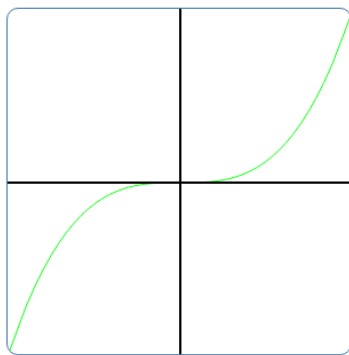


Figure 2: No maximum or minimum even though the derivative is zero.

Coordinates

are distances or angles, represented by numbers that uniquely identify points on surfaces of two dimensions (2D) or in space of three dimensions (3D).

Since the derivative is zero or undefined at both local maximum and local minimum points, we need a way to determine which, if either, actually occurs. The most elementary approach, but one that is often tedious or difficult, is to test directly whether the y coordinates “near” the potential maximum or minimum are above or below the y coordinate at the point of interest. There are too many points “near” the point to test, but a little thought shows we need only test two provided we know that f is continuous (recall that this means that the graph of f has no jumps or gaps).

Suppose, for example, that we have identified three points at which f' is zero or nonexistent: (x_1, y_1) , (x_2, y_2) , (x_3, y_3) , and $x_1 < x_2 < x_3$ (see figure 3). Suppose that we compute the value of $f(a)$ for $x_1 < a < x_2$, and that $f(a) < f(x_2)$. What can we say about the graph between a and x_2 ? Could there be a point $(b, f(b))$, $a < b < x_2$ with $f(b) > f(x_2)$? No: if there were, the graph would go up from $(a, f(a))$ to $(b, f(b))$ then down to $(x_2, f(x_2))$ and somewhere in between would have a local maximum point. But at that local maximum point the derivative of f would be zero or nonexistent, yet we already know that the derivative is zero or nonexistent only at x_1 , x_2 , and x_3 . The upshot is that one computation tells us that $(x_2, f(x_2))$ has the largest y coordinate of any point on the graph near x_2 and to the left of x_2 . We can perform the same test on the right. If we find that on both sides of x_2 the values are smaller, then there must be a local maximum at $(x_2, f(x_2))$; if we find that on both sides of x_2 the values are larger, then there must be a local minimum at $(x_2, f(x_2))$; if we find one of each, then there is neither a local maximum or minimum at x_2 .

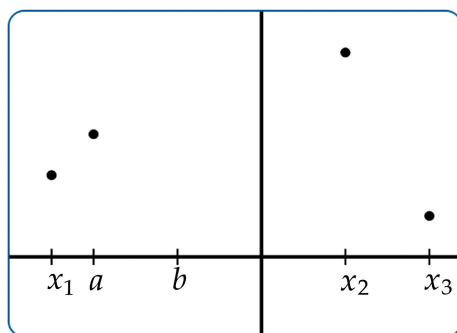


Figure 3: Testing for a maximum or minimum.

It is not always easy to compute the value of a function at a particular point. The task is made easier by the availability of calculators and computers, but they have their own drawbacks—they do not always allow us to distinguish between values that are very close together. Nevertheless, because this method is conceptually simple and sometimes easy to perform, you should always consider it.

The First Derivative Test

We can instead use information about the derivative $f'(x)$ to decide; since we have already had to compute the derivative to find the critical values, there is often relatively little extra work involved in this method.

How can the derivative tell us whether there is a maximum, minimum, or neither at a point? Suppose that $f'(a)=0$. If there is a local maximum when $x=a$, the function must be lower near $x=a$ than it is right at $x=a$. If the derivative exists near $x=a$, this means $f'(x)>0$ when x is near a and $x<a$, because the function must “slope up” just to the left of a . Similarly, $f'(x)<0$ when x is near a and $x>a$, because f slopes down from the local maximum as we move to the right. Using the same reasoning, if there is a local minimum at $x=a$, the derivative of f must be negative just to the left of a and positive just to the right. If the derivative exists near a but does not change from positive to negative or negative to positive, that is, it is positive on both sides or negative on both sides, then there is neither a maximum nor minimum when $x=a$. See the first graph in figure 1 and the graph in figure 2 for examples.

Example

Find all local maximum and minimum points for $f(x)=\sin x+\cos x$ using the first derivative test. The derivative is $f'(x)=\cos x-\sin x$ and from example 3 the critical values we need to consider are $\pi/4$ and $5\pi/4$.

The graphs of $\sin x$ and $\cos x$ are shown in figure 4. Just to the left of $\pi/4$ the cosine is larger than the sine, so $f'(x)$ is positive; just to the right the cosine is smaller than the sine, so $f'(x)$ is negative. This means there is a local maximum at $\pi/4$. Just to the left of $5\pi/4$ the cosine is smaller than the sine, and to the right the cosine is larger than the sine. This means that the derivative $f'(x)$ is negative to the left and positive to the right, so f has a local minimum at $5\pi/4$.

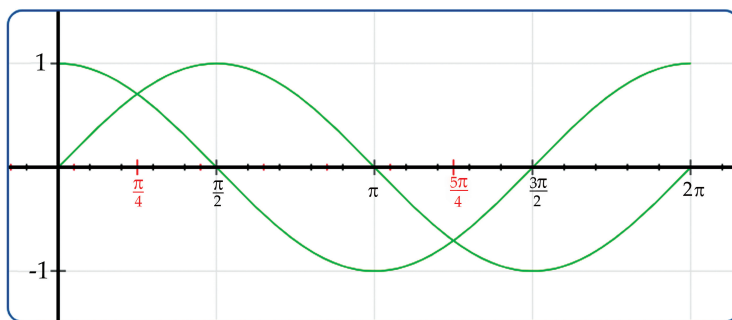


Figure 4: The sine and cosine.

The Second Derivative Test

The basis of the first derivative test is that if the derivative changes from positive to negative at a point at which the derivative is zero then there is a local maximum at the point, and similarly for a local minimum. If f' changes from positive to negative it is decreasing; this means that the derivative of f' , f'' , might be negative, and if in fact f'' is negative then f' is definitely decreasing, so there is a local maximum at the point in question. Note well that f' might change from positive to negative while f'' is zero, in which case f'' gives us no information about the critical value. Similarly, if f' changes

from negative to positive there is a local minimum at the point, and f' is increasing. If $f'' > 0$ at the point, this tells us that f' is increasing, and so there is a local minimum.

Concavity and Inflection Points

We know that the sign of the derivative tells us whether a function is increasing or decreasing; for example, when $f'(x) > 0$, $f(x)$ is increasing. The sign of the second derivative $f''(x)$ tells us whether f' is increasing or decreasing; we have seen that if f' is zero and increasing at a point then there is a local minimum at the point, and if f' is zero and decreasing at a point then there is a local maximum at the point. Thus, we extracted information about f from information about f'' .

We can get information from the sign of f'' even when f' is not zero. Suppose that $f''(a) > 0$. This means that near $x=a$, f' is increasing. If $f'(a) > 0$, this means that f slopes up and is getting steeper; if $f'(a) < 0$, this means that f slopes down and is getting less steep. The two situations are shown in figure 5. A curve that is shaped like this is called concave up.

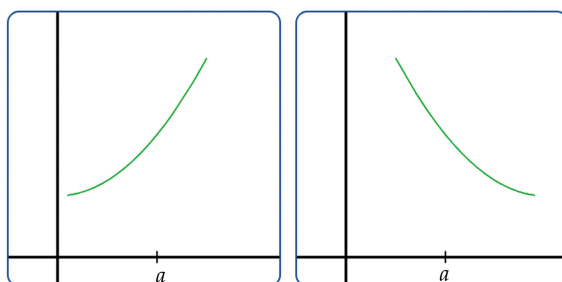


Figure 5: $f''(a) > 0$: $f'(a)$ positive and increasing, $f'(a)$ negative and increasing.

Now suppose that $f''(a) < 0$. This means that near $x=a$, f' is decreasing. If $f'(a) > 0$, this means that f slopes up and is getting less steep; if $f'(a) < 0$, this means that f slopes down and is getting steeper. The two situations are shown in figure 6. A curve that is shaped like this is called concave down.

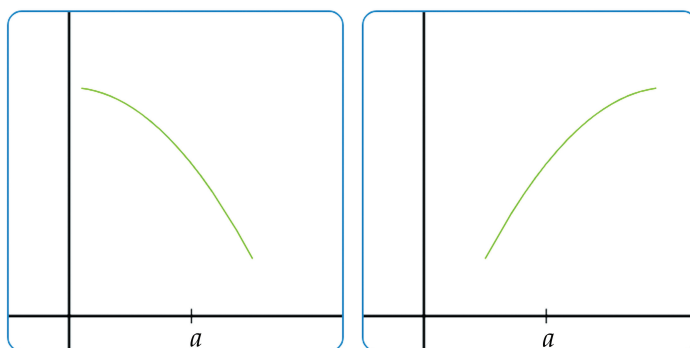


Figure 6: $f''(a) < 0$: $f'(a)$ positive and decreasing, $f'(a)$ negative and decreasing.

If we are trying to understand the shape of the graph of a function, knowing where it is concave up and concave down helps us to get a more accurate picture.

Of particular interest are points at which the concavity changes from up to down or down to up; such points are called inflection points. If the concavity changes from up to down at $x=a$, f'' changes from positive to the left of a to negative to the right of a , and usually $f''(a)=0$. We can identify such points by first finding where $f''(x)$ is zero and then checking to see whether $f''(x)$ does in fact go from positive to negative or negative to positive at these points. Note that it is possible that $f''(a)=0$ but the concavity is the same on both sides; $f(x)=x^4$ at $x=0$ is an example.

Asymptotes and Other Things to Look For

A vertical asymptote is a place where the function becomes infinite, typically because the formula for the function has a denominator that becomes zero. For example, the reciprocal function $f(x)=1/x$ has a vertical asymptote at $x=0$, and the function $\tan x$ has a vertical asymptote at $x=\pi/2$ (and also at $x=-\pi/2$, $x=3\pi/2$, etc.). Whenever the formula for a function contains a denominator it is worth looking for a vertical asymptote by checking to see if the denominator can ever be zero, and then checking the limit at such points. Note that there is not always a vertical asymptote where the denominator is zero: $f(x)=(\sin x)/x$ has a zero denominator at $x=0$, but since $\lim_{x \rightarrow 0} (\sin x)/x=1$ there is no asymptote there.

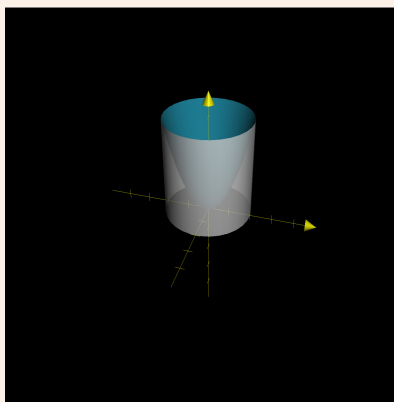
A horizontal asymptote is a horizontal line to which $f(x)$ gets closer and closer as x approaches ∞ (or as x approaches $-\infty$). For example, the reciprocal function has the x -axis for a horizontal asymptote. Horizontal asymptotes can be identified by computing the limits $\lim_{x \rightarrow \infty} f(x)$ and $\lim_{x \rightarrow -\infty} f(x)$. Since $\lim_{x \rightarrow \infty} 1/x = \lim_{x \rightarrow -\infty} 1/x = 0$, the line $y=0$ (that is, the x -axis) is a horizontal asymptote in both directions.

Some functions have asymptotes that are neither horizontal nor vertical, but some other line. Such asymptotes are somewhat more difficult to identify and we will ignore them.

If the domain of the function does not extend out to infinity, we should also ask what happens as x approaches the boundary of the domain. For example, the function $y=f(x)=1/\sqrt{r^2-x^2}$ has domain $-r < x < r$, and y becomes infinite as x approaches either r or $-r$. In this case we might also identify this behavior because when $x=\pm r$ the denominator of the function is zero.

If there are any points where the derivative fails to exist (a cusp or corner), then we should take special note of what the function does at such a point.

Finally, it is worthwhile to notice any symmetry. A function $f(x)$ that has the same value for $-x$ as for x , i.e., $f(-x)=f(x)$, is called an “even function.” Its graph is symmetric with respect to the y -axis. Some examples of even functions are: x^n when n is an even number, $\cos x$, and $\sin^2 x$. On the other hand, a function that satisfies the property $f(-x)=-f(x)$ is called an “odd function.” Its graph is symmetric with respect to the origin. Some examples of odd functions are: x^n when n is an odd number, $\sin x$, and $\tan x$. Of course, most functions are neither even nor odd, and do not have any particular symmetry.



CORE CONCEPT: POLAR COORDINATE SYSTEM

The polar coordinate system is a two-dimensional coordinate system in which points are given by an angle and a distance from a central point known as the pole (equivalent to the origin in the more familiar Cartesian coordinate system). The polar coordinate system is used in many fields, including mathematics, physics, engineering, navigation and robotics. It is especially useful in situations where

the relationship between two points is most easily expressed in terms of angles and distance; in the Cartesian coordinate system, such a relationship can only be found through trigonometric formulae.

Rectangular coordinate system is also known as the Cartesian coordinate system after Rene Descartes, who popularized its use in analytic geometry.

The polar coordinate system is a two-dimensional coordinate system in which each point on a plane is determined by a distance from a reference point and an angle from a reference direction.

When we think about plotting points in the plane, we usually think of **rectangular coordinates** (x, y) in the Cartesian coordinate plane. However, there are other ways of writing a coordinate pair and other types of grid systems. Polar coordinates are points labeled (r, θ) and plotted on a polar grid. The polar grid is represented as a series of concentric circles radiating out from the pole, or the origin of the coordinate plane.

The reference point (analogous to the origin of a Cartesian system) is called the pole, and the ray from the pole in the reference direction is the polar axis. The distance from the pole is called the radial coordinate or radius, and the angle is called the angular coordinate, polar angle, or azimuth. The radial coordinate is often denoted by r or ρ , and the angular coordinate by ϕ , θ , or t .

Angles in polar notation are generally expressed in either degrees or radians (2π rad being equal to 360°). Degrees are traditionally used in navigation, surveying, and many applied disciplines, while radians are more common in mathematics and mathematical physics. In many contexts, a positive angular coordinate means that the angle ϕ is measured counterclockwise from the axis. In mathematical literature, the polar axis is often drawn horizontal and pointing to the right.

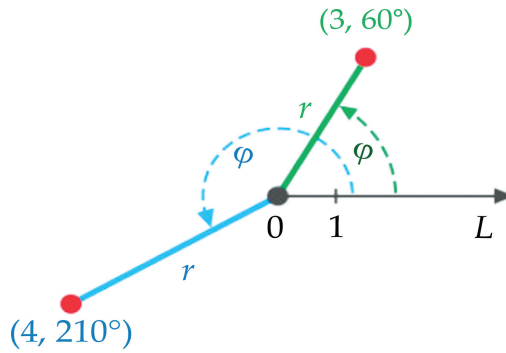


Figure 7: Examples of Polar Coordinates: Points in the polar coordinate system with pole O and polar axis L . In green, the point with radial coordinate 3 and angular coordinate 60° or $(3, 60^\circ)$. In blue, the point $(4, 210^\circ)$.

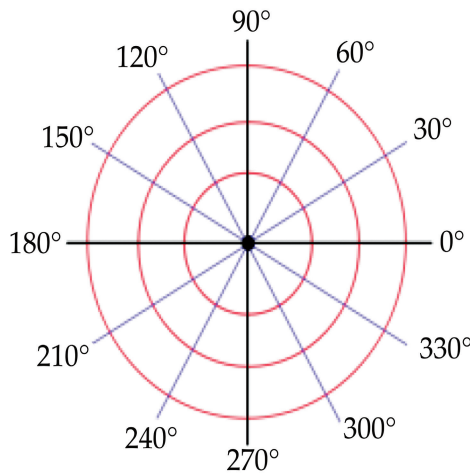


Figure 8: Polar Graph Paper: A polar grid with several angles labeled in degrees.

Plotting Points Using Polar Coordinates

The polar grid is scaled as the unit circle with the positive x -axis now viewed as the polar axis and the origin as the pole. The first coordinate is the radius or length of the directed line segment from the pole. The angle θ , measured in radians, indicates the direction of r . We move counterclockwise from the polar axis by an angle of θ , and measure a directed line segment the length of r in the direction of θ . Even though we measure θ first and then r , the polar point is written with the r -coordinate first. For

example, to plot the point $(2, \frac{\pi}{4})$, we would move $\frac{\pi}{4}$ units in the counterclockwise direction and then a length of 2 from the pole. This point is plotted on the grid in Figure 9.

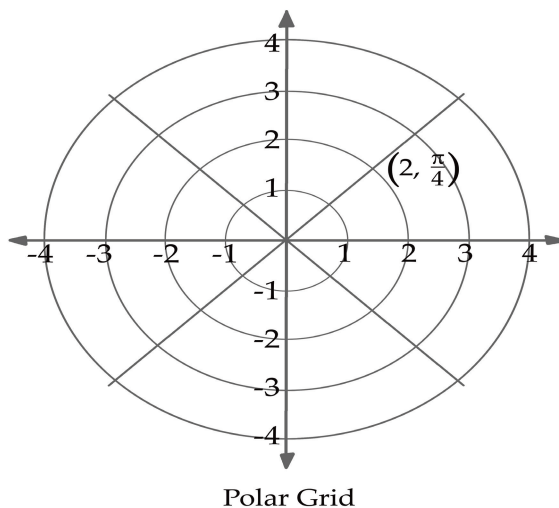


Figure 9: Plotting a point on a Polar Grid: Plot of the point $(2, \frac{\pi}{4})$, by moving $\frac{\pi}{4}$ units in the counterclockwise direction and then a length of 2 from the pole.

Uniqueness of Polar Coordinates

Adding any number of full turns (360° or 2π radians) to the angular coordinate does not change the corresponding direction. Also, a negative radial coordinate is best interpreted as the corresponding positive distance measured in the opposite direction. Therefore, the same point can be expressed with an infinite number of different polar coordinates $(r, \phi \pm n \cdot 360^\circ)$ or $(-r, \phi \pm (2n+1) \cdot 180^\circ)$, where n is any integer. Moreover, the pole itself can be expressed as $(0, \phi)$ for any angle ϕ .

Converting Between Polar and Cartesian Coordinates

Polar and Cartesian coordinates can be interconverted using the **Pythagorean Theorem** and trigonometry.

Polar Coordinates to Rectangular (Cartesian) Coordinates

When given a set of polar coordinates, we may need to convert them to rectangular coordinates. To do so, we can recall the relationships that exist among the variables x, y, r , and θ , from the definitions of $\cos \theta$ and $\sin \theta$. Solving for the variables x and y yields the following formulas:

$$\cos \theta = \frac{x}{r} \quad \Rightarrow \quad x = r \cos \theta$$

$$\sin \theta = \frac{y}{r} \quad \Rightarrow \quad y = r \sin \theta$$

An easy way to remember the equations above is to think of $\cos \theta$ as the adjacent side over the hypotenuse and $\sin \theta$ as the opposite side over the hypotenuse. Dropping a perpendicular from the point in the plane to the x -axis forms a right triangle, as illustrated in Figure 10.

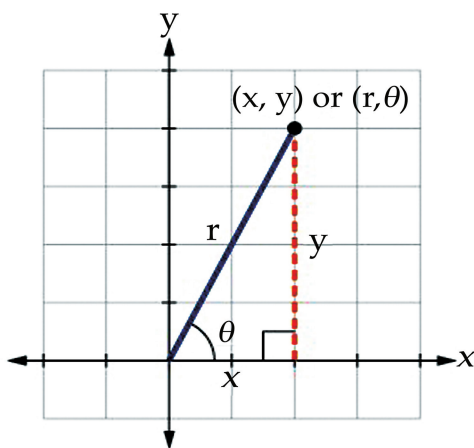


Figure 10: Trigonometry Right Triangle: A right triangle with rectangular (Cartesian) coordinates and equivalent polar coordinates.

To convert polar coordinates (r, θ) to rectangular coordinates (x, y) follow these steps:

- 1) Write $\cos \theta = \frac{x}{r} \Rightarrow x = r \cos \theta$ and $\sin \theta = \frac{y}{r} \Rightarrow y = r \sin \theta$.
- 2) Evaluate $\cos \theta$ and $\sin \theta$.
- 3) Multiply $\cos \theta$ by r to find the x -coordinate of the rectangular form.
- 4) Multiply $\sin \theta$ by r to find the y -coordinate of the rectangular form.

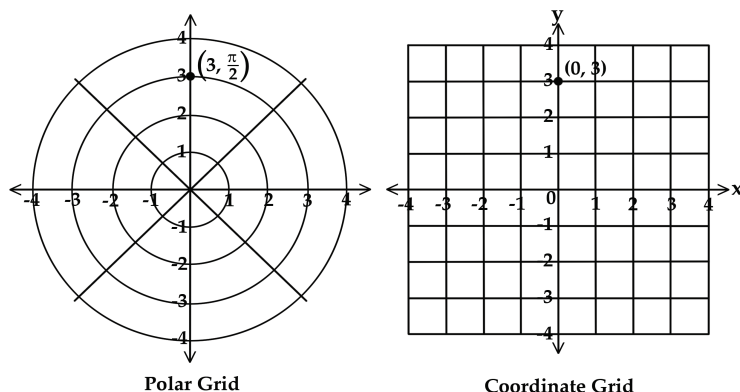
Example: Write the polar coordinates $(3, \frac{\pi}{2})$ as rectangular coordinates.

$$\begin{aligned} x &= r \cos \theta \\ &= 3 \cos \frac{\pi}{2} \\ &= 0 \end{aligned}$$

$$\begin{aligned} y &= r \sin \theta \\ &= 3 \sin \frac{\pi}{2} \\ &= 3 \end{aligned}$$

The rectangular coordinates are $(0, 3)$.

Polar and Coordinate Grid of Equivalent Points: The rectangular coordinate $(0, 3)$ is the same as the polar coordinate $(3, \frac{\pi}{2})$ as plotted on the two grids below.



Rectangular (Cartesian) Coordinates to Polar Coordinates

To convert rectangular coordinates to polar coordinates, we will use two other familiar relationships. With this conversion, however, we need to be aware that a set of rectangular coordinates will yield more than one polar point.

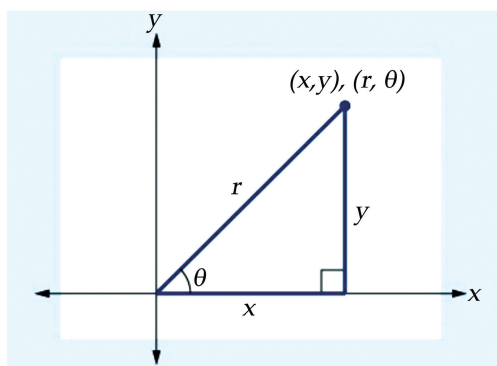
Converting from rectangular coordinates to polar coordinates requires the use of one or more of the relationships illustrated below. Recall:

$$\cos \theta = \frac{x}{r} \Rightarrow x = r \cos \theta$$

$$\sin \theta = \frac{y}{r} \Rightarrow y = r \sin \theta$$

$$r^2 = x^2 + y^2$$

$$\tan \theta = \frac{y}{x}$$



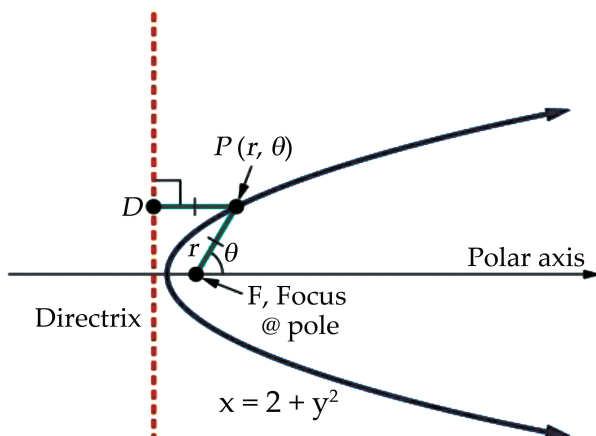
Trigonometry Right Triangle: A right triangle with rectangular (Cartesian) coordinates and equivalent polar coordinates.

Conics in Polar Coordinates

Polar coordinates allow conic sections to be expressed in an elegant way.

Defining a Conic

We learned how a parabola is defined by the focus (a fixed point) and the directrix (a fixed line).



Parts of a Parabola: Consider the parabola $x=2+y^2$. Any conic may be determined by three characteristics: a single focus, a fixed line called the directrix, and the ratio of the distances of each to a point on the graph.

We can define any conic in the polar coordinate system in terms of a fixed point, the focus $P(r, \theta)$ at the pole, and a line, the directrix, which is perpendicular to the polar axis.

For a conic with eccentricity e ,

1. If $0 \leq e < 1$, the conic is an ellipse.
2. If $e = 1$, the conic is a parabola.
3. If $e > 1$, the conic is an hyperbola .

With this definition, we may now define a conic in terms of the directrix: $x = \pm p$, the eccentricity e , and the angle θ . Thus, each conic may be written as a polar equation in terms of r and θ .

For a conic with a focus at the origin, if the directrix is $x = \pm p$, where p is a positive real number, and the eccentricity is a positive real number e , the conic has a polar equation:

$$r = \frac{e \cdot p}{1 \pm e \cdot \cos \theta}$$

For a conic with a focus at the origin, if the directrix is $y = \pm p$, where p is a positive real number, and the eccentricity is a positive real number e , the conic has a polar equation:

$$r = \frac{e \cdot p}{1 \pm e \cdot \sin \theta}$$

Polar equation is any equation that describes a relation between r and θ , where r represents the distance from the pole (origin) to a point on a curve, and θ represents the counterclockwise angle made by a point on a curve, the pole, and the positive x -axis.

Other Curves in Polar Coordinates

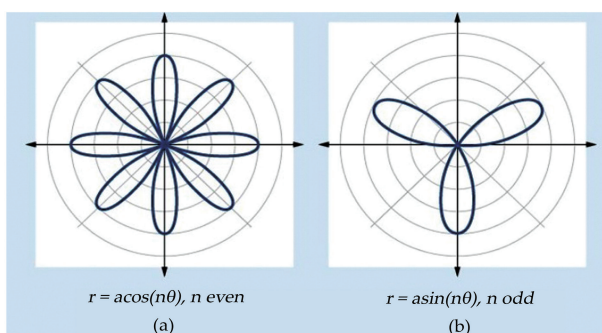
Some curves have a simple expression in polar coordinates, whereas they would be very complex to represent in Cartesian coordinates. To graph in the rectangular coordinate system we construct a table of x and y values. To graph in the polar coordinate system we construct a table of r and θ values. We enter values of θ into a polar equation and calculate r . However, using the properties of symmetry and finding key values of θ and r means fewer calculations will be needed.

Investigating Rose Curves

Polar equations can be used to generate unique graphs. The following type of polar equation produces a petal-like shape called a rose curve. Although the graphs look complex, a simple **polar equation** generates the pattern. The formulas that generate the graph of a rose curve are given by:

$$r = a \cdot \cos(n\theta) \text{ and } r = a \cdot \sin(n\theta) \text{ where } a \neq 0$$

If n is even, the curve has $2n$ petals. If n is odd, the curve has n petals.



Rose Curves: Complex graphs generated by the simple polar formulas that generate rose curves: $r = a \cos n\theta$ and $r = a \sin n\theta$ where $a \neq 0$.

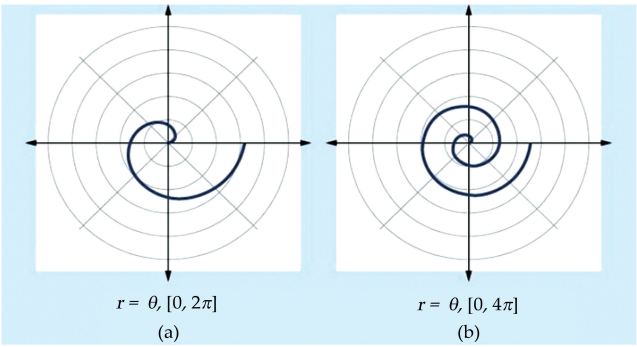
Investigating the Archimedes' Spiral

Archimedes' spiral is named for its discoverer, the Greek mathematician Archimedes, who is credited with numerous discoveries in the fields of geometry and mechanics.

The formula that generates the graph of the Archimedes' spiral is given by:

$$R = a + b\theta \text{ for } \theta \geq 0$$

As θ increases, r increases at a constant rate in an ever-widening, never-ending, spiraling path.



Archimedes' Spiral: The formula that generates the graph of a spiral is $r = \theta$ for $\theta \geq 0$.

Example

Sketch the graph of $r = \theta$ over $[0, 2\pi]$

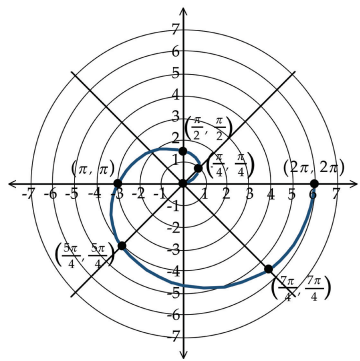
Solution

As r is equal to θ , the plot of the Archimedes' spiral begins at the pole at the point $(0, 0)$. While the graph hints of symmetry, there is no formal symmetry with regard to passing the symmetry tests. Further, there is no maximum value, unless the domain is restricted.

Create a table such as the one below.

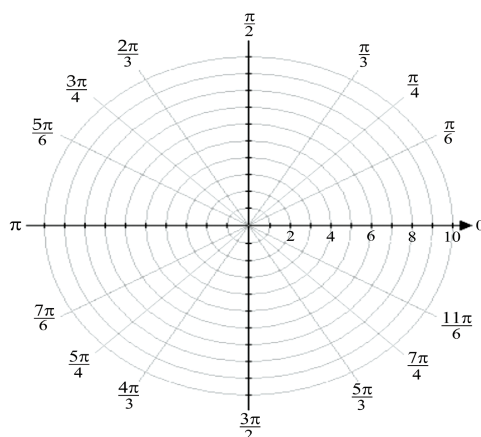
θ	$\frac{\pi}{4}$	$\frac{\pi}{2}$	π	$\frac{3\pi}{2}$	$\frac{7\pi}{4}$	2π
r	0.785	1.57	3.14	4.71	5.50	6.28

Notice that the r -values are just the decimal form of the angle measured in radians. We can see them on a graph in Figure.



Graphs of Polar Equations

The polar coordinate system is very similar to that of the rectangular coordinate system. In a polar coordinate grid, there will be a series of circles extending out from the pole (or origin in a rectangular coordinate grid) and five different lines passing through the pole to represent the angles at which the exact values are known for the trigonometric functions.



The **trigonometric functions** (also called circular functions, angle functions or goniometric functions) are functions of an angle. They relate the angles of a triangle to the lengths of its sides.

Graphing a polar equation is accomplished in pretty much the same manner as rectangular equations are graphed. They can be graphed by point-plotting, using the **trigonometric functions** period, and using the equation's symmetry (if any). When graphing rectangular equations by point-plotting you would pick values for x and then evaluate the equation to determine its corresponding y value. For a polar equation, you would pick angle measurements for θ and then evaluate the equation to determine its corresponding r value.

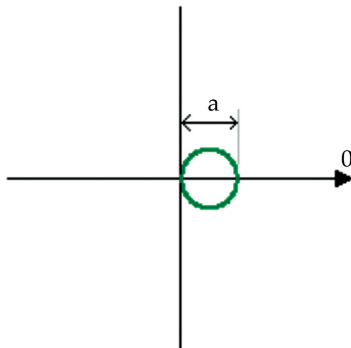
Symmetry tests for polar coordinates

1. Replace θ with $-\theta$. If an equivalent equation results, the graph is symmetric with respect to the polar axis.
2. Replace θ with $-\theta$ and r with $-r$. If an equivalent equation results, the graph is symmetric with respect to $\theta = \frac{\pi}{2}$.
3. Replace r with $-r$. If an equivalent equation results, the graph is symmetric with respect to the pole.

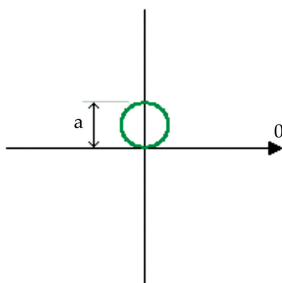
When you started to graph functions (in rectangular form) you started by learning the basic shapes of certain functions such as lines, parabolas, circles, square roots, and absolute value functions just to name a few. Polar equations also have some general types of equations. Learning to recognize the formulas of these equations will help in sketching the graphs.

Circles in Polar Form

1. $r = a \cos \theta$ is a circle where "a" is the diameter of the circle that has its left-most edge at the pole.



2. $r = a \sin \theta$ is a circle where “a” is the diameter of the circle that has its bottom-most edge at the pole.



Limaçons (Snails)

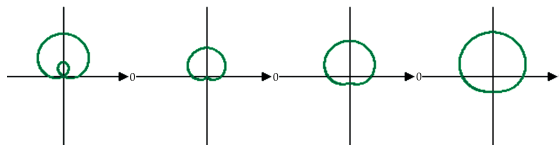
1. $r = a \pm b \sin \theta$, where $a > 0$ and $b > 0$
2. $r = a \pm b \cos \theta$, where $a > 0$ and $b > 0$

The limaçons containing sine will be above the horizontal axis if the sign between a and b is plus or below the horizontal axis if the sign is minus. If the limaçon contains the function cosine then the graph will be either to the right of the vertical axis if the sign is plus or to the left if the sign is minus.

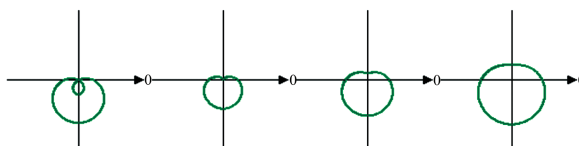
The ratio of $\frac{a}{b}$ will determine the exact shape of the limaçon

$$\frac{a}{b} < 1 \qquad \frac{a}{b} = 1 \qquad 1 < \frac{a}{b} < 2 \qquad \frac{a}{b} \geq 2$$

$$r = a + b \sin \theta$$



$$r = a - b \sin \theta$$



The graphs of limaçons with cosine would have similar shapes but along the horizontal axis.

Rose Curves

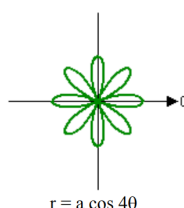
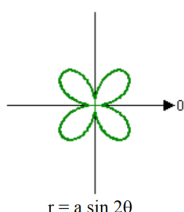
A rose curve is a graph that is produced from a polar equation in the form of:

$r = a \sin n\theta$ or $r = a \cos n\theta$, where $a \neq 0$ and n is an integer > 1

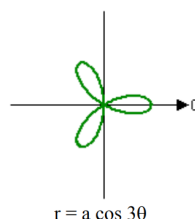
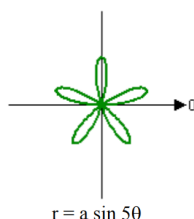
It is possible for a polar equation to fail a test and still exhibit that type of symmetry when you finish graphing the function over a full period.

They are called rose curves because the loops that are formed resemble petals. The number of petals that are present will depend on the value of n . The value of a will determine the length of the petals.

If n is an even integer, then the rose will have $2n$ petals.



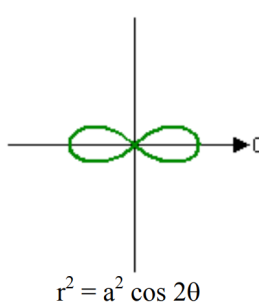
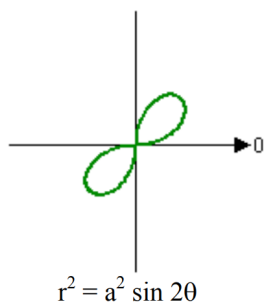
If n is an odd integer, then the rose will have n petals.



Lemniscates

The last type of polar equation that we will cover here is the lemniscates, which has the shape of a figure 8 or a propeller. Lemniscates have the general polar equation of:

$r^2 = a^2 \sin 2\theta$ or $r^2 = a^2 \cos 2\theta$, where $a \neq 0$



A lemniscate containing the sine function will be symmetric to the pole while the lemniscate containing the cosine function will be symmetric to the polar axis, to $\theta = \frac{\pi}{2}$, and the pole.

Example

Sketch the graph of $r^2 = 4\cos 2\theta$.

Solution

The equation exhibits symmetry with respect to the line $\theta = \frac{\pi}{2}$, the polar axis, and the pole.

Let's find the zeros. It should be routine by now, but we will approach this equation a little differently by making the substitution $u = 2\theta$.

$$0 = 4 \cos 2\theta$$

$$0 = 4 \cos u$$

$$0 = \cos u$$

$$\cos^{-1} 0 = \frac{\pi}{2}$$

$$u = \frac{\pi}{2} \quad \text{Substitute } 2\theta \text{ back in for } u.$$

$$2\theta = \frac{\pi}{2}$$

$$\theta = \frac{\pi}{4}$$

So, the point $\left(0, \frac{\pi}{4}\right)$ is a zero of the equation.

Now let's find the maximum value. Since the maximum of $\cos u = 1$ when $u = 0$, the maximum $\cos 2\theta = 1$ when $2\theta = 0$. Thus,

$$r^2 = 4 \cos(0)$$

$$r^2 = 4(1) = 4$$

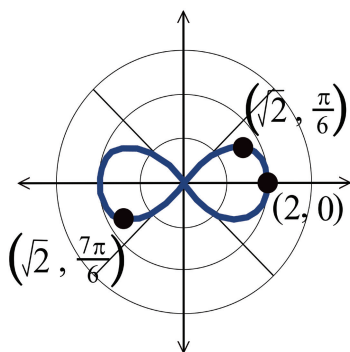
$$r = \pm\sqrt{4} = 2$$

We have a maximum at $(2, 0)$. Since this graph is symmetric with respect to the pole, the line $\theta = \frac{\pi}{2}$, and the polar axis, we only need to plot points in the first quadrant.

Make a table similar to the table below.

θ	0	$\frac{\pi}{6}$	$\frac{\pi}{4}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$
r	2	$\sqrt{2}$	0	$\sqrt{2}$	0

Plot the points on the graph, such as the one shown in Figure.



CONCEPT: POLAR COORDINATE SYSTEM

The polar coordinate system is a mathematical framework used to describe the positions of points in a plane using distance and angle relative to a reference point and axis. It provides an alternative way to represent points compared to the standard Cartesian coordinate system, which uses x and y coordinates.

In the polar coordinate system, a point is uniquely identified by two values:

- **Radius (r):** The distance from the origin (reference point) to the point in question.
- **Angle (θ):** The angle formed between the reference axis (usually the positive x -axis) and the line connecting the origin and the point. The angle is measured in a counter-clockwise direction.

To specify a point in polar coordinates, you write it as (r, θ) , where ' r ' is the radius and ' θ ' is the angle. The reference point (origin) is typically denoted as O , and the reference axis (usually the positive x -axis) is denoted as the polar axis.

Converting from Polar to Cartesian Coordinates: Given a point in polar coordinates (r, θ) , you can convert it to Cartesian coordinates (x, y) using the following relationships:

$$x = r * \cos(\theta) \quad y = r * \sin(\theta)$$

Converting from Cartesian to Polar Coordinates: Given a point in Cartesian coordinates (x, y) , you can convert it to polar coordinates (r, θ) using the following relationships:

$$r = \sqrt{x^2 + y^2} \quad \theta = \text{atan2}(y, x)$$

Here, $\text{atan2}(y, x)$ is the arctangent function that considers the signs of both ' x ' and ' y ' to correctly determine the angle in the appropriate quadrant.

Polar coordinates are particularly useful for describing points in a circular or radial manner. They find applications in various fields, including physics, engineering, and geometry. For instance, polar coordinates are used to describe circular motion, wave propagation, and antenna radiation patterns, among other phenomena.

Keep in mind that while polar coordinates can be more intuitive in certain situations, they might not be as straightforward for describing certain complex shapes and relationships as Cartesian coordinates. It's often a matter of choosing the coordinate system that suits the problem at hand.

SUMMARY

- Curve sketching is the process of using algebra and calculus to find the features of a function and then plotting those features in order to draw the function on a coordinate plane.
- Curve sketching is another practical application of differential calculus. Whether we are interested in a function as a purely mathematical object or in connection with some application to the real world, it is often useful to know what the graph of the function looks like.
- The basis of the first derivative test is that if the derivative changes from positive to negative at a point at which the derivative is zero then there is a local maximum at the point, and similarly for a local minimum.
- A vertical asymptote is a place where the function becomes infinite, typically because the formula for the function has a denominator that becomes zero.
- The polar coordinate system is a two-dimensional coordinate system in which each point on a plane is determined by a distance from a reference point and an angle from a reference direction.
- The polar grid is scaled as the unit circle with the positive x -axis now viewed as the polar axis and the origin as the pole. The first coordinate is the radius or length of the directed line segment from the pole.
- To graph in the rectangular coordinate system we construct a table of x and y values. To graph in the polar coordinate system we construct a table of r and θ values.

MULTIPLE CHOICE QUESTIONS

1. **Curve sketching is another practical application of differential calculus.**
 - a. True
 - b. False
2. **In many contexts, a positive angular coordinate means that the angle ϕ is measured counterclockwise from the axis.**
 - a. True
 - b. False
3. **Polar equation is any equation that describes a relation between r and θ , where r represents the distance from the pole (origin) to a point on a curve, and θ represents the counterclockwise angle made by a point on a curve, the pole, and the positive x -axis.**
 - a. True
 - b. False
4. **The trigonometric functions (also called circular functions, angle functions or goniometric functions) are functions of an angle.**
 - a. True
 - b. False
5. **It is possible for a polar equation to fail a test and still exhibit that type of symmetry when you finish graphing the function over a full period.**
 - a. True
 - b. False

Answers

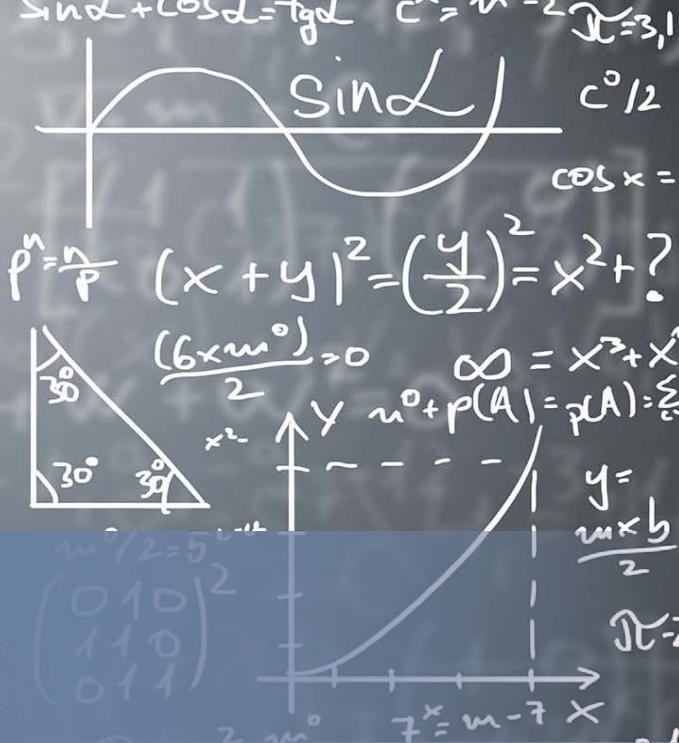
1. (a) 2. (a) 3. (a) 4. (a) 5. (a)

REVIEW QUESTIONS

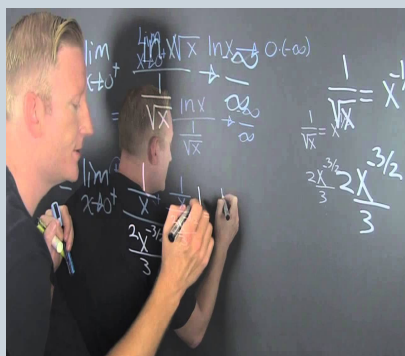
1. Explain why the function $f(x)=1/x$ has no local maxima or minima.
2. Let $f(x)=ax^2+bx+c$ with $a \neq 0$. Show that f has exactly one critical point. Give conditions on a and b which guarantee that the critical point will be a maximum. It is possible to see this without using calculus at all; explain.
3. Let $r > 0$. Find the local maxima and minima of the function $f(x) = \sqrt{r^2 - x^2}$ on its domain $[-r, r]$.
4. Identify the intervals on which the graph of the function $f(x)=x^4-4x^3+10$ is of one of these four shapes: concave up and increasing; concave up and decreasing; concave down and increasing; concave down and decreasing.
5. Convert each of the following into an equation in the given coordinate system.
6. Convert $2x - 5x^3 = 1 + xy$ into polar coordinates.
7. Convert $r = -8\cos\theta$ into Cartesian coordinates
8. Graph the polar equation $r = 1 - 2 \cos \theta$.
9. Graph the polar equation $r = 3 \cos 2\theta$.
10. Describe the concavity of $y=x^3+bx^2+cx+d$. You will need to consider different cases, depending on the values of the coefficients.
11. Let $f(x)=1/(x^2-a^2)$, where $a \geq 0$. Find any vertical and horizontal asymptotes and the intervals upon which the given function is concave up and increasing; concave up and decreasing; concave down and increasing; concave down and decreasing. Discuss how the value of a affects these features.
12. Graph $r = 5 - 5\sin \theta$, $r = 7 - 6 \cos \theta$, and $r = 2 + 4 \cos \theta$.

REFERENCES

1. Calculus I: The Derivative and Its Applicationsbooks.google.com, books, Patrick Clark · 2023
2. Calculus Manual 1: Function Definitions for Studentsbooks.google.com, books, E. E. Engstrom 2003
3. Calculus Workbook for Dummies with Online Practicebooks.google.com, books, Mark Ryan 2018
4. Fast Start Differential Calculusbooks.google.com books, Daniel Ashlock 2019, This book reviews the algebraic prerequisites of calculus, including solving equations, lines, quadratics, functions, logarithms, and trig functions.



INDETERMINATE FORMS AND IMPROPER INTEGRAL



An indeterminate form is an expression involving two functions whose limit cannot be determined solely from the limits of the individual functions. These forms are common in calculus; indeed, the limit definition of the derivative is the limit of an indeterminate form.

An integral is an improper integral if either the interval of integration is not finite (improper integral of type 1) or if the function to integrate is not continuous (not bounded) in the interval of integration (improper integral of type 2).

After studying this chapter, you will understand the core concepts of:

- Indeterminate Forms
- Improper Integral

INTRODUCTION

In calculus and other branches of mathematical analysis, when the limit of the sum, difference, product, quotient or power of two functions is taken, it may often be possible to simply add, subtract, multiply, divide or exponentiate the corresponding limits of these two functions respectively. However, there are occasions where it is unclear what the sum, difference, product or power of these two limits ought to be. For example, it is unclear what the following expressions ought to evaluate to:

$$\frac{0}{0}, \frac{\infty}{\infty}, 0 \times \infty, \infty - \infty, 0^0, 1^\infty, \text{ and } \infty^0.$$

These seven expressions are known as indeterminate forms. More specifically, such expressions are obtained by naively applying the algebraic limit theorem to evaluate the limit of the corresponding arithmetic operation of two functions, yet there are examples of pairs of functions that after being operated on converge to 0, converge to another finite value, diverge to infinity or just diverge. This inability to decide what the limit ought to be explains why these forms are regarded as indeterminate.

An improper integral is an extension of the notion of a definite integral to cases that violate the usual assumptions for that kind of integral. In the context of Riemann integrals (or, equivalently, Darboux integrals), this typically involves unboundedness, either of the set over which the integral is taken or of the integrand (the function being integrated), or both. It may also involve bounded but not closed sets or bounded but not continuous functions. While an improper integral is typically written symbolically just like a standard definite integral, it actually represents a limit of a definite integral or a sum of such limits; thus improper integrals are said to converge or diverge. If a regular definite integral (which may retronymically be called a proper integral) is worked out as if it is improper, the same answer will result.

The term “indeterminate” is sometimes used as a synonym for unknown or variable.

A mathematical expression can also be said to be indeterminate if it is not definitively or precisely determined. Certain forms of limits are said to be indeterminate when merely knowing the limiting behavior of individual parts of the expression is not sufficient to actually determine the overall limit. *For example*, a limit of the form

$0/0$, i.e., $\lim_{x \rightarrow 0} \frac{f(x)}{g(x)}$ where $\lim_{x \rightarrow 0} f(x) = \lim_{x \rightarrow 0} g(x) = 0$, is indeterminate since the value of the overall limit actually depends on the limiting behavior of the combination of the two

functions (e.g., $\lim_{x \rightarrow 0} \frac{x}{x} = 1$, while $\lim_{x \rightarrow 0} \frac{x^2}{x} = 0$).

There are seven indeterminate forms involving 0, 1, and ∞ :

$$\frac{0}{0}, 0 \cdot \infty, \frac{\infty}{\infty}, \infty - \infty, 0^0, \infty^0, 1^\infty$$

Note, however, that there is a certain ambiguity in this enumeration in the sense

that symbolic expressions of the form $\frac{\infty}{\infty}$ can perhaps be written as $\infty\infty^{-1}$, etc.

If complex infinity $\tilde{\infty}$ is allowed as well, then six additional indeterminate forms result:

$$0 \cdot \tilde{\infty}, \frac{\tilde{\infty}}{\tilde{\infty}}, \frac{\infty}{\tilde{\infty}}, \frac{\tilde{\infty}}{\infty}, \tilde{\infty} - \tilde{\infty}, 1^{\tilde{\infty}}.$$

Improper integral is the limit of a definite integral as an endpoint of the interval(s) of integration approaches either a specified real number, $\infty, -\infty$, or in some instances as both endpoints approach limits. Such an integral is often written symbolically just like a standard definite integral, in some cases with *infinity* as a limit of integration.

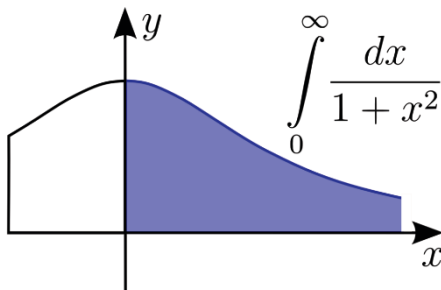
Specifically, an improper integral is a limit of the form:

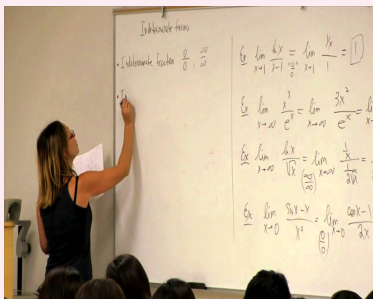
$$\lim_{b \rightarrow \infty} \int_a^b f(x) dx, \quad \lim_{a \rightarrow -\infty} \int_a^b f(x) dx,$$

or

$$\lim_{c \rightarrow b^-} \int_a^c f(x) dx, \quad \lim_{c \rightarrow a^+} \int_c^b f(x) dx,$$

in which one takes a limit in one or the other (or sometimes both) endpoints.





CORE CONCEPT: INDETERMINATE FORMS

The term “indeterminate” is sometimes used as a synonym for unknown or variable. A mathematical expression which is not definitively or precisely determined may be considered as indeterminate form.

The indeterminate forms involving 0, 1 and ∞ are

$$\frac{0}{0}, \frac{\infty}{\infty}, 0 \times \infty, \infty - \infty, \infty^0, 0^0, 1^\infty$$

An indeterminate form takes place if by knowing the limits of the functions involved we cannot determine what is the overall limit. We have to do a further analysis to solve this kind of situations.

Example

If $f(x) = x$ and $g(x) = \frac{1}{x} + 1$ then we know that $\lim_{x \rightarrow +\infty} f(x) = +\infty$ and $\lim_{x \rightarrow +\infty} g(x) = 1$ but we cannot know beforehand the result of the limit $\lim_{x \rightarrow +\infty} g(x)^{f(x)} = 1^{+\infty}$.

The main indeterminate forms are: $(+\infty) - (+\infty), 0 \cdot (\pm\infty), \frac{0}{0}, (+\infty)^0, 1^{\pm\infty}, 0^{\pm\infty}, \frac{\pm\infty}{\pm\infty}$ where all the values that appear are limits of functions.

$$\text{If } f(x) \rightarrow \infty \Rightarrow \begin{cases} \lim_{x \rightarrow +\infty} 1^{f(x)} = \lim_{x \rightarrow +\infty} 1 = 1 \\ \lim_{x \rightarrow +\infty} \frac{0}{f(x)} = \lim_{x \rightarrow +\infty} 0 = 0 \\ \lim_{x \rightarrow +\infty} f(x) \cdot 0 = \lim_{x \rightarrow +\infty} 0 = 0 \\ \lim_{x \rightarrow +\infty} \frac{0}{\frac{1}{f(x)}} = \lim_{x \rightarrow +\infty} 0 = 0 \end{cases}$$

Indeterminate Form 0/0

Let us suppose that $\lim_{x \rightarrow +\infty} f(x) = 0$ and $\lim_{x \rightarrow +\infty} g(x) = 0$, then we have that $\lim_{x \rightarrow +\infty} \frac{f(x)}{g(x)} = \frac{0}{0}$ and thus there is an indeterminate form.

In this case we have to ask ourselves what function tends more rapidly toward zero: if it is $f(x)$, then the limit will be zero, and if it is $g(x)$ then the limit will be **infinity**.

$$\lim_{x \rightarrow +\infty} \frac{f(x)}{g(x)} = \lim_{x \rightarrow +\infty} \frac{\frac{1}{\frac{g(x)}{f(x)}}}{\frac{1}{f(x)}} = \frac{\pm\infty}{\pm\infty}$$

Note that $\lim_{x \rightarrow +\infty} \frac{f(x)}{g(x)} = \frac{0}{0}$ implies that

so we have changed the indeterminate form zero over zero by the one we already know, infinity over infinity.

Infinity is a concept describing something without any bound or larger than any natural number.

Example

$$1. \quad \lim_{x \rightarrow +\infty} \frac{\frac{x}{x^2-1}}{\frac{2+x}{x^2}} = \frac{0}{0} \Rightarrow \lim_{x \rightarrow +\infty} \frac{\frac{x}{x^2-1}}{\frac{2+x}{x^2}} = \lim_{x \rightarrow +\infty} \frac{x \cdot x^2}{(x^2-1) \cdot (2+x)} = \lim_{x \rightarrow +\infty} \frac{x^3}{x^3} = 1$$

$$\frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a \cdot d}{b \cdot c}$$

since d

$$2. \quad \lim_{x \rightarrow +\infty} \frac{\frac{2+x}{\ln x}}{\frac{\ln x}{x}} = \lim_{x \rightarrow +\infty} \frac{x \cdot (2+x)}{\ln x^2} = \lim_{x \rightarrow +\infty} \frac{x^2}{\ln x^2} = +\infty$$

Indeterminate Form Infinity/Infinity (∞/∞)

Let us suppose that $\lim_{x \rightarrow +\infty} f(X) = \pm\infty$ and $\lim_{x \rightarrow +\infty} g(x) = \pm\infty$, then

we have that $\lim_{x \rightarrow +\infty} \frac{f(x)}{g(x)} = \frac{\pm\infty}{\pm\infty}$, so that we have an indeterminate form.

To know the value of the limit we will have to look at the functional form of every function and find the term of highest order. Then, according to the position of this element of highest order we will determine whether the limit is infinite, zero, or in case of being in both sides of the **fraction**, the value to which the limit is converging.

Let us remember how the limits of the main functions are ordered:

$$\log, x \ll x^n \ll x^m \ll a^x \ll b^x \ll x^x \text{ where } n < m \text{ and } b > a > 0$$

Example

$$1. \text{ If } b > a > 0 \text{ then } \lim_{x \rightarrow +\infty} \frac{x^k + \log_r a^x}{\pm^\pm} = \lim_{x \rightarrow +\infty} \frac{a^x}{\pm^\pm} = 0$$

$$2. \quad \lim_{x \rightarrow +\infty} \frac{x^2 + 3^4 x^3 - x^4}{2x - \ln x + x^2} = \lim_{x \rightarrow +\infty} \frac{-x^4}{x^2} = -\infty$$

Fraction is a number that represents a part of a whole. It consists of a numerator and a denominator. The numerator represents the number of equal parts of a whole, while the denominator is the total number of parts that make up said whole.

$$3. \lim_{x \rightarrow +\infty} \frac{2e^x - \log_2 x}{x+1-3e^x} = \lim_{x \rightarrow +\infty} \frac{2e^x}{3e^x} = \frac{2}{3}$$

$$4. \lim_{x \rightarrow +\infty} \frac{(1-x) \cdot 2^x}{(1+x^2) \cdot 2^x - 1} = \lim_{x \rightarrow +\infty} \frac{-x \cdot 2^x}{x^2 \cdot 2^x} = \lim_{x \rightarrow +\infty} \frac{-1}{x} = 0$$

Indeterminate Form 0 x Infinity (0 x ∞)

We will suppose that $\lim_{x \rightarrow +\infty} f(x) = 0$ and $\lim_{x \rightarrow +\infty} g(x) = \pm\infty$, then we will have that $\lim_{x \rightarrow +\infty} f(x) \cdot g(x) = 0 \cdot \pm(\infty)$.

In other words, we are wondering what function goes more rapidly to its limit, $f(x)$ to zero or $g(x)$ to infinity.

To solve this type of indeterminate form we will do a simple step:

$$\lim_{x \rightarrow +\infty} f(x) \cdot g(x) = \lim_{x \rightarrow +\infty} \frac{1}{\frac{1}{f(x)}} \cdot g(x) = \lim_{x \rightarrow +\infty} \frac{g(x)}{\frac{1}{f(x)}} = \frac{\pm\infty}{\pm\infty}$$

and we will solve the limit.

Example

$$1. \lim_{x \rightarrow +\infty} \frac{2x}{x^3-1} \cdot \ln x = 0 \cdot (+\infty) \Rightarrow \lim_{x \rightarrow +\infty} \frac{2x}{x^3-1} \cdot \ln x = \lim_{x \rightarrow +\infty} \frac{2x \ln x}{x^3-1} = \lim_{x \rightarrow +\infty} \frac{2x \cdot \ln x}{x^3} = \lim_{x \rightarrow +\infty} \frac{2 \cdot \ln x}{x^2} = 0$$

$$2. \lim_{x \rightarrow +\infty} x^{-x} \cdot 2^x = \lim_{x \rightarrow +\infty} \frac{2x}{x^x} = 0$$

$$3. \lim_{x \rightarrow +\infty} \frac{\ln x}{x+1} \cdot \frac{-x^2-1}{\ln x-1} = \lim_{x \rightarrow +\infty} \frac{\ln x \cdot (-x^2-1)}{(x+1) \cdot (\ln x-1)} = \lim_{x \rightarrow +\infty} -x = -\infty$$

Indeterminate Form Infinity minus Infinity (∞ - ∞)

Let us suppose that $\lim_{x \rightarrow +\infty} f(x) = \pm\infty$ and $\lim_{x \rightarrow +\infty} g(x) = \pm\infty$, then we have that $\lim_{x \rightarrow +\infty} f(x) - g(x) = (\pm\infty) - (\pm\infty)$

and thus, we have an indeterminate form.

To solve for this limit we have three options:

- When one sees the limit straight away: Sometimes it can be seen straight away that one of the two infinities is greater than the other:

Example

$$1. \lim_{x \rightarrow +\infty} x - 2^x = -\infty$$

$$2. \quad \lim_{x \rightarrow +\infty} \left(\frac{x-2}{\sqrt{x-1}} - \ln(2x-1) \right) = +\infty$$

- When it is possible simplify the expression: Another option is to do the subtraction and to obtain only one expression where probably we will have an indeterminate form of the type infinity divided by infinity, which we can already solve.

Example

$$\lim_{x \rightarrow +\infty} \left(\frac{2x^2-5}{x+3} - 2x \right) = \lim_{x \rightarrow +\infty} \frac{2x^2-5-2x \cdot (x+3)}{x+3} = \lim_{x \rightarrow +\infty} \frac{-11x}{x+3} = -11$$

- When there are square roots involved: If we have $f(x)$ or $g(x)$ with a square root is not easy to find the limit.

An option is to add expressions in the following way:

$$f(x) - g(x) = (f(x) - g(x)) \cdot \frac{f(x)^2 + g(x)^2}{f(x)^2 + g(x)^2}$$

in order to obtain an expression which we can:

Example

$$\begin{aligned} \lim_{x \rightarrow +\infty} \sqrt{x^2-2} - x &= \lim_{x \rightarrow +\infty} \frac{(\sqrt{x^2-x} - x) \cdot (\sqrt{x^2-x} + x)}{\sqrt{x^2-x} + x} \\ &= \lim_{x \rightarrow +\infty} \frac{\sqrt{x^2-x} - x^2}{\sqrt{x^2-x} + x} = \lim_{x \rightarrow +\infty} \frac{-x}{\sqrt{x^2-x} + x} \\ &= \lim_{x \rightarrow +\infty} \frac{-x}{\frac{2}{x^2+x}} = \lim_{x \rightarrow +\infty} \frac{-x}{2x} = \frac{-1}{2} \end{aligned}$$

Indeterminate Form 1 Raised To Infinity (1^∞)

Let us suppose that $\lim_{x \rightarrow +\infty} f(x) = 1$ and $\lim_{x \rightarrow +\infty} g(x) = \pm\infty$, then we have that

$$\lim_{x \rightarrow +\infty} f(x)^{g(x)} = 1^{\pm\infty}$$

and we have again an indeterminate form.

The basic problem of this indeterminate form is to know from where $f(x)$ tends to one (right or left) and what function reaches its limit more rapidly.

To solve this limit we will use the following two formulae, depending on which it is more convenient. These convert the indeterminate form to one that we can solve.

The two formulae are the following: If $\lim_{x \rightarrow +\infty} f(x) = 1$ and $\lim_{x \rightarrow +\infty} g(x) = \pm\infty$ then,

$$1. \lim_{x \rightarrow +\infty} f(x)^{g(x)} = e^{\left(\lim_{x \rightarrow +\infty} (f(x)-1) \cdot g(x) \right)}$$

$$2. \lim_{x \rightarrow +\infty} f(x)^{g(x)} = e^{\left(\lim_{x \rightarrow +\infty} g(x) \cdot \ln f(x) \right)}$$

Example

$$\lim_{x \rightarrow +\infty} \left(\frac{1}{1+x^2} \right)^2 x = e^{\lim_{x \rightarrow +\infty} \left(\frac{1}{1+x^2} - 1 \right) \cdot 2x}$$

$$= e^{\lim_{x \rightarrow +\infty} \left(\frac{1}{1+x^2} - \frac{1+x^2}{1+x^2} \right) \cdot 2x} = e^{\lim_{x \rightarrow +\infty} \frac{-x^2}{1+x^2} \cdot 2x}$$

$$1. = e^{\lim_{x \rightarrow +\infty} \frac{-2x^3}{1+x^2}} = e^{-\infty} = 0$$

$$\lim_{x \rightarrow +\infty} \left(\frac{1}{1+x^2} \right)^2 x = e^{\lim_{x \rightarrow +\infty} 2x \cdot \left(\frac{1}{1+x^2} \right)}$$

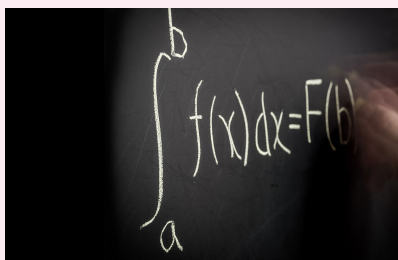
$$2. = e^{\lim_{x \rightarrow +\infty} -2 \cdot \ln(1+x^2)} = e^{-\infty} = 0$$

$$\lim_{x \rightarrow +\infty} \left(1 - \frac{1}{x} \right) = \lim_{x \rightarrow +\infty} \left(1 - \frac{1}{x} \right)$$

$$\lim_{x \rightarrow +\infty} \left(\frac{1}{x} \right) = \frac{1}{\infty} = 0$$

3.

$$4. \lim_{x \rightarrow +\infty} \left(1 + \frac{1}{2^x} \right)^x = e^{\lim_{x \rightarrow +\infty} \left(1 + \frac{1}{2^x} - 1 \right) \cdot x} = e^{\lim_{x \rightarrow +\infty} \frac{x}{2^x}} = e^0 = 1$$

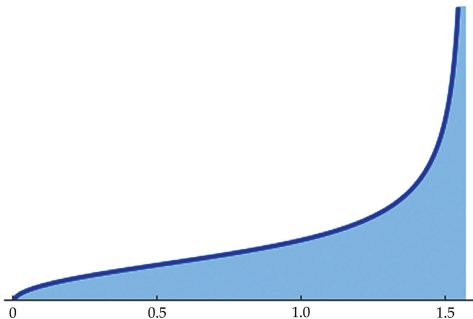


CORE CONCEPT: IMPROPER INTEGRAL

An improper integral is a definite integral that has either or both limits infinite or an integrand that approaches infinity at one or more points in the range of integration. Improper integrals cannot be computed using a normal Riemann integral.

An improper integral is a type of definite integral in which the integrand is undefined at one or both of the endpoints. Strictly speaking, it is the limit of the definite integral as the interval approaches its desired size.

Improper integrals may be evaluated by finding a limit of the indefinite integral of the integrand. However, such a value is meaningful only if the improper integral converges in the first place.



Improper integrals appear frequently in the study of probability distributions, asymptotic behavior, and calculus in general. In order to evaluate them properly, it is crucial to understand precisely what is meant by these integrals.

Definition of Improper Integral

One type of improper integral is an integral where one of the endpoints is (approaching) infinity. For instance,

$$\int_0^{\infty} \frac{1}{1+x^2} dx = \lim_{b \rightarrow \infty} \int_0^b \frac{1}{1+x^2} dx.$$

Another type of improper integral is an integral where both of the endpoints are (approaching) infinity. For instance,

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \int_{-\infty}^0 e^{-x^2} dx + \int_0^{\infty} e^{-x^2} dx$$

is an integral used in Gaussian distributions.

A third type of improper integral is an integral for which an **asymptote** appears at one (or both) of the endpoints. For instance,

$$\int_0^1 \frac{1}{\sqrt{x}} dx = \lim_{a \rightarrow 0^+} \int_a^1 \frac{1}{\sqrt{x}} dx.$$

A fourth type of improper integral is one that is not properly defined: an integral whose integrand is not defined at all the values within the interval of integration. For instance,

$$\int_0^2 \frac{1}{\sqrt[3]{(x-1)^2}} dx$$

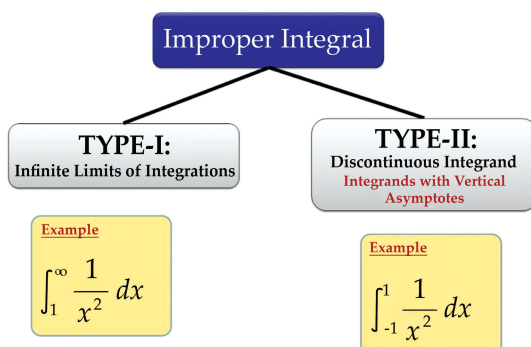
Asymptote is a line such that the distance between the curve and the line approaches zero as one or both of the x or y coordinates tends to infinity.

$$f(x) = \frac{1}{\sqrt[3]{(x-1)^2}}$$

is undefined because the function $f(x) = \frac{1}{\sqrt[3]{(x-1)^2}}$ is not defined throughout the interval of integration, even though its anti-derivative $F(x) = 3\sqrt[3]{x-1}$ is defined throughout the interval of integration. A concept known as the Cauchy principal value permits a reinterpretation of the integral as

$$\int_0^2 \frac{1}{\sqrt[3]{(x-1)^2}} dx = \lim_{b \rightarrow 1^-} \int_0^b \frac{1}{\sqrt[3]{(x-1)^2}} dx + \lim_{a \rightarrow 1^+} \int_a^2 \frac{1}{\sqrt[3]{(x-1)^2}} dx$$

in order to assign a value to determine a value for it, provided each of the integrals converges. In general, the Cauchy principal value splits the interval of integration of an improper integral into closed intervals, on whose interiors the integrand is defined.



Improper Integrals of Type- I

When we defined definite integral $\int_a^b f(x) dx$ we assumed that interval $[a, b]$ is finite and that function f does not have infinite discontinuities. Let us extend integral to the case where interval $[a, b]$ is infinite.

Consider the infinite region S that lies below the curve $y = f(x)$, above the x -axis and to the right of the line $x = a$.

You might think that, since S is infinite in extent, its area must be infinite, but let us take a closer look. The area of the part of S that lies to the

$$A(t) = \int_a^t f(x) dx = F(t) - F(a)$$

left of the line $x=t$ is

where F is antiderivative of f .

Although the curves $y = \frac{1}{x}$ and $y = \frac{1}{x^2}$ look very similar for $x > 0$, the region under $y = \frac{1}{x^2}$ to the right of $x = 1$ has finite area whereas the corresponding region under $y = \frac{1}{x}$ has infinite area. Note that both $\frac{1}{x}$ and $\frac{1}{x^2}$ approach 0 as $x \rightarrow \infty$ but $\frac{1}{x^2}$ approaches faster than $\frac{1}{x}$. The values of $\frac{1}{x}$ do not decrease fast enough for its integral to have a finite value.

For example, if $f(x) = \frac{1}{x^2}$ and $a = 1$ then $A(t) = \int_1^t \frac{1}{x^2} dx = 1 - \frac{1}{t}$.

Observe that in this case $\lim_{t \rightarrow \infty} A(t) = \lim_{t \rightarrow \infty} \left(1 - \frac{1}{t}\right) = 0$.

So, we define the integral of f (not necessarily a positive function) over an infinite interval as the limit of integrals over finite intervals.

Definition of an Improper Integral of Type 1

1. If $\int_a^t f(x) dx$ exists for every number $t \geq a$, then $\int_a^\infty f(x) dx = \lim_{t \rightarrow \infty} \int_a^t f(x) dx$ provided this limit exists (as a finite number).

2. If $\int_t^b f(x) dx$ exists for every number $t \leq b$, then $\int_{-\infty}^b f(x) dx = \lim_{t \rightarrow -\infty} \int_t^b f(x) dx$ provided this limit exists (as a finite number).

3. The improper integrals $\int_a^\infty f(x) dx$ and $\int_{-\infty}^b f(x) dx$ are called convergent if the corresponding limit exists and divergent if the limit does not exist.

4. If both $\int_a^\infty f(x) dx$ and $\int_{-\infty}^a f(x) dx$ are convergent then $\int_{-\infty}^\infty f(x) dx = \int_{-\infty}^a f(x) dx + \int_a^\infty f(x) dx$. Note, that any real number can be used. Once again only when BOTH integrals are convergent, $\int_{-\infty}^\infty f(x) dx$ is also an integral. Otherwise, it is divergent.

Any of the improper integrals in Definition can be interpreted as an area provided that f is a positive function.

Example

For which p is integral $\int_1^\infty \frac{1}{x^p} dx$ convergent?

Let us handle two cases: $p = 1$ and $p \neq 1$.

If $p = 1$ then $\int_1^\infty \frac{1}{x} dx = \lim_{t \rightarrow \infty} \int_1^t \frac{1}{x} dx = \lim_{t \rightarrow \infty} (\ln|x|)_1^t = \lim_{t \rightarrow \infty} (\ln(t) - \ln(1)) = \infty$

Thus, when $p = 1$ integral is divergent.

If $p \neq 1$ then

$$\int_1^{\infty} \frac{1}{x^p} dx = \lim_{t \rightarrow \infty} \int_1^t \frac{1}{x^p} dx = \lim_{t \rightarrow \infty} \left(\frac{1}{-p+1} x^{-p+1} \Big|_1^t \right) = \frac{1}{1-p} \lim_{t \rightarrow \infty} \left(\frac{1}{t^{p-1}} - 1 \right).$$

If $p > 1$ then $p-1 > 0$ and $\frac{1}{t^{p-1}} \rightarrow 0$ as $t \rightarrow \infty$. Therefore, $\int_1^{\infty} \frac{1}{x^p} dx = \frac{1}{p-1}$ when $p > 1$.

If $p < 1$ then $p-1 < 0$ and $\frac{1}{t^{p-1}} \rightarrow \infty$ as $t \rightarrow \infty$. Therefore, integral is divergent.

Therefore, $\int_1^{\infty} \frac{1}{x^p} dx$ is convergent for $p > 1$.

Example

Evaluate $\int_{-\infty}^{-1} e^{2x} dx$.

By definition

$$\int_{-\infty}^{-1} e^{2x} dx = \lim_{t \rightarrow -\infty} \int_t^{-1} e^{2x} dx = \lim_{t \rightarrow -\infty} \left(\frac{1}{2} e^{2x} \Big|_t^{-1} \right) = \lim_{t \rightarrow -\infty} \left(\frac{1}{2} e^{-2} - \frac{1}{2} e^{2t} \right) = \frac{1}{2} e^{-2} - 0 = \frac{1}{2e^2}$$

Example

Calculate if possible $\int_{-\infty}^{\infty} \frac{1}{r^2+4} dr$.

It is convenient to take $a = 0$ in definition.

$$\int_{-\infty}^{\infty} \frac{1}{r^2+4} dr = \int_{-\infty}^0 \frac{1}{r^2+4} dr + \int_0^{\infty} \frac{1}{r^2+4} dr.$$

Now

$$\begin{aligned} \int_{-\infty}^0 \frac{1}{r^2+4} dr &= \lim_{t \rightarrow -\infty} \int_t^0 \frac{1}{r^2+4} dr = \lim_{t \rightarrow -\infty} \left(\frac{1}{2} \tan^{-1} \left(\frac{r}{2} \right) \Big|_t^0 \right) = \lim_{t \rightarrow -\infty} \left(\frac{1}{2} \tan^{-1} \left(\frac{0}{2} \right) - \frac{1}{2} \tan^{-1} \left(\frac{t}{2} \right) \right) \\ &= \lim_{t \rightarrow -\infty} \left(0 - \frac{1}{2} \tan^{-1} \left(\frac{t}{2} \right) \right) = 0 - \left(\frac{1}{2} \cdot \left(-\frac{\pi}{2} \right) \right) = \frac{\pi}{4}. \end{aligned}$$

Similarly

$$\begin{aligned} \int_0^{\infty} \frac{1}{r^2+4} dr &= \lim_{t \rightarrow \infty} \int_0^t \frac{1}{r^2+4} dr = \lim_{t \rightarrow \infty} \left(\frac{1}{2} \tan^{-1} \left(\frac{r}{2} \right) \Big|_0^t \right) = \lim_{t \rightarrow \infty} \left(\frac{1}{2} \tan^{-1} \left(\frac{t}{2} \right) - \frac{1}{2} \tan^{-1} \left(\frac{0}{2} \right) \right) \\ &= \lim_{t \rightarrow \infty} \left(\frac{1}{2} \tan^{-1} \left(\frac{t}{2} \right) - 0 \right) = \frac{1}{2} \cdot \frac{\pi}{2} - 0 = \frac{\pi}{4}. \end{aligned}$$

Since both of these integrals are convergent, the given integral is convergent and

$$\int_{-\infty}^{\infty} \frac{1}{r^2 + 4} dr = \frac{\pi}{4} + \frac{\pi}{4} = \frac{\pi}{2}.$$

Since $\frac{1}{x^2 + 4} > 0$, the given improper integral can be interpreted as the area of the

infinite region that lies under the curve $y = \frac{1}{x^2 + 4}$ and above the x-axis.

Example

Find $\int_{-\infty}^{\infty} \frac{1}{x} dx$.

Here we will choose $a = 1$.

$$\int_{-\infty}^{\infty} \frac{1}{x} dx = \int_{-\infty}^1 \frac{1}{x} dx + \int_1^{\infty} \frac{1}{x} dx.$$

It was shown in example above that $\int_1^{\infty} \frac{1}{x} dx$ is divergent, so we do not need to

calculate second integral. Since at least one integral is divergent then $\int_{-\infty}^{\infty} \frac{1}{x} dx$ is also divergent.

Improper Integrals of Type – II

Suppose that f is a positive continuous function defined on a finite interval $[a, b)$ but has a vertical asymptote at b . Let S be the unbounded region under the graph of f and above the x-axis between a and b . The area of the part of S between a and t is

$$A(t) = \int_a^t f(x) dx.$$

If it happens that $A(t)$ approaches a definite number as $t \rightarrow b^-$, then we say that

the area of the region S is A and we write $A = \lim_{t \rightarrow b^-} \int_a^t f(x) dx$. We use this equation to define an improper integral of Type 2 even when f is not a positive function, no matter what type of discontinuity f has at b .

Definition of an Improper Integral of Type 2

1. If f is continuous on $[a, b)$ and is discontinuous at b , then

$$\int_a^b f(x) dx = \lim_{t \rightarrow b^-} \int_a^t f(x) dx \quad \text{if this limit exists (as a finite number).}$$

2. If f is continuous on $(a, b]$ and is discontinuous at a , then

$$\int_a^b f(x) dx = \lim_{t \rightarrow a^+} \int_t^b f(x) dx \quad \text{if this limit exists (as a finite number).}$$

3. The improper integral $\int_a^b f(x) dx$ is called convergent if the corresponding limit exists and divergent if the limit does not exist.

4. If f has a discontinuity at c , where $a < c < b$, and both $\int_a^c f(x) dx$ and $\int_c^b f(x) dx$ are convergent, then

$$\int_a^b f(x) dx = \int_a^c f(x) dx + \int_c^b f(x) dx.$$

Only when BOTH integrals are convergent, convergent is $\int_a^b f(x) dx$. Otherwise it is divergent.

Example

Find $\int_3^7 \frac{1}{\sqrt{x-3}} dx$.

Integral is improper because $f(x) = \frac{1}{\sqrt{x-3}}$ has infinite discontinuity at $x = 3$.

Thus,

$$\begin{aligned} \int_3^7 \frac{1}{\sqrt{x-3}} dx &= \lim_{t \rightarrow 3^+} \int_t^7 \frac{1}{\sqrt{x-3}} dx = \lim_{t \rightarrow 3^+} \left(2\sqrt{x-3} \Big|_t^7 \right) = 2 \lim_{t \rightarrow 3^+} (\sqrt{7-3} - \sqrt{t-3}) \\ &= 2(\sqrt{7-3} - \sqrt{t-3}) = 4. \end{aligned}$$

Example

Find $\int_0^{\frac{\pi}{2}} \tan(x) dx$ if possible.

Integral is divergent because $\lim_{t \rightarrow \left(\frac{\pi}{2}\right)^-} \tan(x) = \infty$.

$$\begin{aligned}\int_0^{\frac{\pi}{2}} \tan(x) dx &= \lim_{t \rightarrow \left(\frac{\pi}{2}\right)^-} \int_0^t \tan(x) dx = \lim_{t \rightarrow \left(\frac{\pi}{2}\right)^-} \left(-\ln |\cos(x)| \Big|_0^t \right) \\ &= \lim_{t \rightarrow \left(\frac{\pi}{2}\right)^-} \left(-\ln |\cos(t)| + \ln |\cos(0)| \right) = \lim_{t \rightarrow \left(\frac{\pi}{2}\right)^-} \left(-\ln |\cos(t)| \right) = \infty.\end{aligned}$$

Therefore,

Thus, integral is divergent.

Example

Evaluate $\int_0^3 \frac{1}{x-1} dx$.

This integral is improper because $x=1$ is a vertical asymptote. Thus,

$$\int_0^3 \frac{1}{x-1} dx = \int_0^1 \frac{1}{x-1} dx + \int_1^3 \frac{1}{x-1} dx.$$

$$\int_0^1 \frac{1}{x-1} dx = \lim_{t \rightarrow 1^-} \int_0^t \frac{1}{x-1} dx = \lim_{t \rightarrow 1^-} \left(\ln|x-1| \Big|_0^t \right) = \lim_{t \rightarrow 1^-} \left(\ln|t-1| - \ln|0-1| \right) = \lim_{t \rightarrow 1^-} \ln|t-1| = -\infty.$$

Therefore, $\int_0^1 \frac{1}{x-1} dx$ is divergent. This implies that $\int_0^3 \frac{1}{x-1} dx$ is divergent.

Example

Find if possible $\int_0^{\infty} \frac{1}{x^3} dx$.

Note that $f(x) = \frac{1}{x^3}$ has discontinuity at $x=0$ and also interval is infinite. So, this integral belongs to both types. To handle it, we split it into 2 integrals. We can split it up anywhere, but pick a value that will be convenient for evaluation purposes.

$$\int_0^{\infty} \frac{1}{x^3} dx = \int_0^1 \frac{1}{x^3} dx + \int_1^{\infty} \frac{1}{x^3} dx.$$

Handle first integral:

$$\int_0^1 \frac{1}{x^3} dx = \lim_{t \rightarrow 0^+} \int_t^1 \frac{1}{x^3} dx = \lim_{t \rightarrow 0^+} \left(-\frac{1}{2} \frac{1}{x^2} \Big|_t^1 \right) = -\frac{1}{2} \lim_{t \rightarrow 0^+} \left(\frac{1}{1^2} - \frac{1}{t^2} \right) = \frac{1}{2} \lim_{t \rightarrow 0^+} \left(\frac{1}{t^2} - 1 \right) = \infty.$$

So, integral is divergent and so the whole integral is also divergent.

Comparison Test for Improper Integrals

Sometimes it is impossible to find the exact value of an improper integral and yet it is

important to know whether it is convergent or divergent. In such cases the following test is useful. Although we state it for Type 1 integrals, a similar theorem is true for Type 2 integrals.

Comparison Test

Suppose that f and g are continuous functions with $f(x) \geq g(x) \geq 0$.

1. If $\int_a^\infty f(x) dx$ is convergent, then $\int_a^\infty g(x) dx$ is also convergent.
2. If $\int_a^\infty g(x) dx$ is divergent, then $\int_a^\infty f(x) dx$ is also divergent.

In terms of area the Comparison Test makes a lot of sense. If $f(x)$ is larger than $g(x)$ then the area under $f(x)$ must also be larger than the area under $g(x)$.

So, if the area under the larger function is finite ($\int_a^\infty f(x) dx$ converges) then the area under the smaller function must also be finite ($\int_a^\infty g(x) dx$ converges). Similarly, if the area under the smaller function is infinite ($\int_a^\infty g(x) dx$ diverges) then the area under the larger function must also be infinite ($\int_a^\infty f(x) dx$ diverges).

You should correctly understand this test. Divergence of $\int_a^\infty f(x) dx$ does not imply divergence of $\int_a^\infty g(x) dx$ (it may or may not diverge). Convergence of $\int_a^\infty g(x) dx$ does not imply convergence of $\int_a^\infty f(x) dx$ (it may or may not converge).

Example

Determine whether $\int_0^\infty e^{-x^2} dx$ is convergent.

We cannot evaluate the integral directly because the antiderivative of e^{-x^2} is not an elementary function. However, note that $x^2 \geq x$ for $x \geq 1$. This means that $-x^2 \leq -x$ and $e^{-x^2} \leq e^{-x}$ for $x \geq 1$.

However, we have integral from 0 to ∞ instead of 1 to ∞ .

An improper integral converges if the limit exists, that is, if the limit equals a finite number. Otherwise, an improper integral is said to diverge.

$$\int_0^{\infty} e^{-x^2} dx = \int_0^1 e^{-x^2} dx + \int_1^{\infty} e^{-x^2} dx.$$

So, we split integral as follows:

First integral on the right-hand side is just an ordinary definite integral and, thus, has finite value.

For the second integral with use comparison test ($e^{-x^2} \leq e^{-x}$ for $x \geq 1$).

$$\int_1^{\infty} e^{-x} dx = \lim_{t \rightarrow \infty} \int_1^t e^{-x} dx = \lim_{t \rightarrow \infty} (-e^{-x} \big|_1^t) = \lim_{t \rightarrow \infty} (-e^{-t} + e^{-1}) = \frac{1}{e}.$$

Thus, taking $f(x) = e^{-x}$ and $g(x) = e^{-x^2}$ in the Comparison Test, we see that $\int_1^{\infty} e^{-x^2} dx$ is convergent. It follows that is $\int_0^{\infty} e^{-x^2} dx$ is convergent.

Example

Determine if the integral $\int_1^{\infty} \frac{1}{x+e^x} dx$ is convergent or divergent.

We can write that $\frac{1}{x+e^x} < \frac{1}{x}$ but we cannot say nothing about integral because $\int_1^{\infty} \frac{1}{x} dx$ is divergent.

So, let us try to drop out x : $\frac{1}{x+e^x} < \frac{1}{e^x} = e^{-x}$ for $x \geq 1$.

$\int_1^{\infty} e^{-x} dx$ is convergent, so $\int_1^{\infty} \frac{1}{x+e^x} dx$ is also convergent by Comparison Test.

Example

Determine if the integral $\int_1^{\infty} \frac{1+e^x}{x} dx$ convergent or divergent.

Integral is divergent by Comparison Test because $\frac{1+e^x}{x} > \frac{1}{x}$ and $\int_1^{\infty} \frac{1}{x} dx$ is divergent.

Example

Determine if the integral $\int_1^{\infty} \frac{\cos(\sqrt{x})}{x^2} dx$ convergent or divergent.

Since $\cos(\sqrt{x}) \leq 1$ then integral is convergent by Comparison Test because $\frac{\cos(\sqrt{x})}{x^2} < \frac{1}{x^2}$ and $\int_1^{\infty} \frac{1}{x^2} dx$ is convergent.

CONCEPT: IMPROPER INTEGRAL

An improper integral is a type of integral that involves integrating a function over an interval where the function might not be defined at certain points or the interval extends to infinity. Improper integrals are important in calculus and analysis because they allow us to extend the concept of integration beyond the limits of a finite interval.

There are two main types of improper integrals:

- **Type 1: Infinite Interval** An improper integral of Type 1 occurs when one or both of the limits of integration are infinite. Mathematically, it's expressed as:

$$\int [a, \infty) f(x) \, dx \text{ or } \int (-\infty, b] f(x) \, dx \text{ or } \int (-\infty, \infty) f(x) \, dx$$

Here, 'a' and 'b' are real numbers, and $f(x)$ is the integrand.

To evaluate these integrals, you often take a limit as a or b approaches infinity or negative infinity, depending on the situation.

- **Type 2: Discontinuities** An improper integral of Type 2 occurs when the function has a discontinuity within the interval of integration. Mathematically, it's expressed as:

$$\int [a, b] f(x) \, dx$$

where ' $f(x)$ ' has a discontinuity at some point within $[a, b]$.

To evaluate these integrals, you break down the interval into subintervals around the point(s) of discontinuity and then take a limit as the size of these subintervals approaches zero.

When evaluating improper integrals, you need to be cautious about the convergence or divergence of the integral. An improper integral is said to converge if the limit of its value exists and is finite. It's said to diverge if the limit does not exist or if it approaches infinity.

Some improper integrals converge, some diverge, and others might be conditionally convergent (convergent under certain conditions). Determining the convergence or divergence often involves applying techniques from calculus, such as limit comparison test, comparison test, ratio test, and more.

Improper integrals are used in various fields of mathematics and science, including physics, engineering, probability theory, and more. They provide a way to calculate areas, volumes, probabilities, and other quantities that might not fit within the framework of a standard, well-behaved integral.

SUMMARY

- In calculus and other branches of mathematical analysis, when the limit of the sum, difference, product, quotient or power of two functions is taken, it may often be possible to simply add, subtract, multiply, divide or exponentiate the corresponding limits of these two functions respectively.
- An improper integral is an extension of the notion of a definite integral to cases that violate the usual assumptions for that kind of integral.
- An indeterminate form takes place if by knowing the limits of the functions involved we cannot determine what is the overall limit.
- An improper integral is a type of definite integral in which the integrand is undefined at one or both of the endpoints. Strictly speaking, it is the limit of the definite integral as the interval approaches its desired size.
- Sometimes it is impossible to find the exact value of an improper integral and yet it is important to know whether it is convergent or divergent.

MULTIPLE CHOICE QUESTIONS

1. Let $f \circ^n (f(x))$ denote the composition of $f(x)$ with itself n number of times then the value of $\lim_{n \rightarrow \infty} f \circ^n (\sin(x)) =$
 - a. -1
 - b. 2
 - c. ∞
 - d. 0
2. If differentiation of any function is zero at any point and constant at other points then it means?
 - a. Function is parallel to x-axis at that point
 - b. Function is parallel to y-axis at that point
 - c. Function is constant
 - d. Function is discontinuous at that point
3. The term “indeterminate” is sometimes used as a synonym for unknown or variable.
 - a. True
 - b. False
4. Infinity is a concept describing something without any bound or larger than any natural number.
 - a. True
 - b. False
5. Fraction is a number that represents a part of a whole. It consists of a numerator and a denominator.
 - a. True
 - b. False

Answers

1. (d) 2. (a) 3. (a) 4. (a) 5. (a)

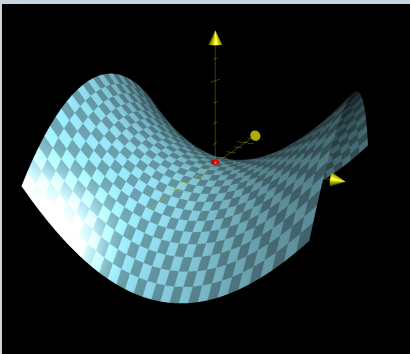
REVIEW QUESTIONS

1. What do you understand by indeterminate form? Explain.
2. Describe the indeterminate form $0/0$.
3. Discuss about the indeterminate form infinity/infinity (∞/∞).
4. Explain the indeterminate form $0 \times \text{infinity}$ ($0 \times \infty$).
5. Discuss the indeterminate form infinity minus infinity ($\infty - \infty$).
6. What is improper integral? Define.
7. Discuss about the improper integrals of type- I.
8. Explain about the improper integrals of type – II.

REFERENCES

1. Autar Kaw, Egwu Kalu (2008), Numerical Methods with Applications (1st ed.), autarkaw.com
2. Cooper, Jeffery (2005), Working analysis, Gulf Professional
3. Ghorpade, Sudhir; Limaye, Balmohan (2010), A course in multivariable calculus and analysis, Springer
4. Ghorpade, Sudhir; Limaye, Balmohan (2010), A course in multivariable calculus and analysis, Springer
5. Lipkin, Leonard J. (2003). “On the Indeterminate Form $0/0$ ”. The College Mathematics Journal. Mathematical Association of America. 34 (1): 55–56.

DIFFERENTIAL CALCULUS OF SEVERAL VARIABLES



Differential calculus is a subfield of calculus concerned with the study of the rates at which quantities change. It is one of the two traditional divisions of calculus, the other being integral calculus, the study of the area beneath a curve.

After studying this chapter, you will understand the core concepts of:

- Differential Calculus
- Differentiation of Implicit Functions

INTRODUCTION

Differential calculus, Branch of mathematical analysis, devised by Isaac Newton and G.W. Leibniz, and concerned with the problem of finding the rate of change of a function with respect to the variable on which it depends. Thus it involves calculating derivatives and using them to solve problems involving non-constant rates of change. Typical applications include finding maximum and minimum values of functions in order to solve practical problems in optimization. Differential calculus is the variation of a function with respect to changes in independent variables, especially the use of differentiation to calculate rates of change of a function and the maximum and minimum values of a function. Differential calculus has applications such as calculating compound interest, organic growth, and slopes of curves, and studying the acceleration of moving bodies. Compare calculus of variations, integral calculus, and differential calculus is a subfield of calculus concerned with the study of the rates at which quantities change. It is one of the two traditional divisions of calculus, the other being integral calculus. The primary objects of study in differential calculus are the derivative of a function, related notions such as the differential, and their applications. The derivative of a function at a chosen input value describes the rate of change of the function near that input value. The process of finding a derivative is called differentiation. Geometrically, the derivative at a point is the slope of the tangent line to the graph of the function at that point, provided that the derivative exists and is defined at that point. For a real-valued function of a single real variable, the derivative of a function at a point generally determines the best linear approximation to the function at that point. Differential calculus is the study of instantaneous rates of change; these rates of change are known as derivatives.



CORE CONCEPT: DIFFERENTIAL CALCULUS

Differential calculus deals with the study of the rates at which quantities change. It is one of the two principal areas of calculus (integration being the other).

Differential calculus describes the methods by which, given a function, you can find its associated rate of change function, called the “derivative.” The function must describe a constantly changing system, such as the temperature variation over the course of the day or the velocity of a planet around a star over the course of one rotation. The derivative of those functions would give you the rate that the temperature changed and

the acceleration of the planet, respectively. Calculus approaches the paths of objects in motion as curves, or functions, and then determines the value of these functions to calculate their rate of change, area, or volume. In the 18th century, *Sir Isaac Newton* and *Gottfried Leibniz* simultaneously, yet separately, described calculus to help solve problems in physics. The two divisions of calculus, differential and integral, can solve problems like the **velocity** of a moving object at a certain moment in time, or the surface area of a complex object like a lampshade.

Velocity is a physical vector quantity; both magnitude and direction are needed to define it, also it is the measurement of the rate at which an object changes its position.

Differential calculus relies on the fundamental principle that you can always use approximations of increasing accuracy to find the exact answer. For instance, you can approximate a curve by a series of straight lines: the shorter the lines, the closer they are to resembling a curve. You can also approximate a spherical solid by a series of cubes that get smaller and smaller with each iteration that fits inside the sphere. Using calculus, you can determine that the approximations tend toward the precise end result, called the limit, until you have accurately described and reproduced the curve, surface, or solid.

Limits and Continuity

When two variables are functionally related, it is reasonable to ask what the behavior of one (dependent) variable is as the other (independent) variable exhibits the specific behavior of “approaching” a particular number. When the dependent variable also approaches a particular number, we call that number a limit. The concept of limit is very powerful, but the Greeks were limited in their advances because modern algebraic notation had not yet been invented. Limit is the fundamental concept of calculus. In fact, a simplistic definition of calculus would be a study of limits. It is in this theme that we begin our calculus adventure.

The concept of continuity is an important first step in the analysis leading to differential and integral calculus. It is also an important analytical tool in its own right, with significant practical applications. Fortunately the main theorems are intuitive, though their proofs can be technically challenging. Nonetheless we prove most of the continuity theorems we state, while the remaining theorems we discuss and apply without proof, since they are intuitive and useful but their proofs require background material from at least a junior level real analysis or topology. Related to continuity is the concept of limit, which is vital for calculus since it puts calculus on the same

rigorous footing as other mathematical disciplines such as algebra and geometry. In our more modern times it has further conceptual appeal, as it is often only possible to approximate the solution to some problem, even though our method of approximation may be arbitrarily close to the actual solution if given enough computing resources. However, the real value of limits lies in its use in confronting an interesting phenomenon.

This is the fact that in mathematics we at times find our analysis (algebraic, geometric or otherwise) breaking down at exactly the value of some variable where we would like to compute something; that is, we are allowed to let that variable “approach” the desired value as closely as we would like but it cannot equal that value according to our classical, pre-calculus mathematics. The relatively modern mathematical tool called “limits” can often break through the analytic barrier at that value, in turn opening us to the extensive and spectacularly useful field we call calculus. Many examples of continuity and limits in action seem straightforward enough, but without a sufficiently deep understanding it is all too easy for students to fall victim to common errors. For this reason we introduce the rather technical definition of continuity here, and develop a method to prove continuity in cases which may seem obvious. Some powerful continuity theorems follow, as do applications. We then employ limits for cases where continuity is “broken,” and in numerous other contexts to make calculus possible, and as an analytical tool in its own right.

Limits and continuity sufficiently to avoid numerous common mistakes requires a care and depth of thought. After continuity, we use a “forms” approach to limits, those forms themselves being ultimately intuitive but nonetheless requiring students to study them carefully and extensively to achieve satisfactory proficiency.

Limit

A sequence is said to approach x_0 or converge to x_0 from below if it converges to x_0 and all its member are less than x_0 .

We say that the limit of f from below at x_0 is h , written as $\lim_{x \rightarrow x_0^-} f(x) = h$ if for any sequence $\{x_j\}$ that approaches x_0 from below, $\{f(x_j)\}$ approaches h .

Similarly the limit of f from above at x_0 is h , written as $\lim_{x \rightarrow x_0^+} f(x) = h$, if for any sequence $\{x_j\}$ that approaches x_0 from above, $\{f(x_j)\}$ approaches h .

We write $\lim_{x \rightarrow x_0} f(x) = h$ when the limits from above and below are both h .

With these definitions, we can prove that if

$$\lim_{x \rightarrow x_0} f(x) = a \text{ and } \lim_{x \rightarrow x_0} g(x) = b$$

Then

$$\lim_{x \rightarrow x_0} [f(x) + g(x)] = a + b \quad (1)$$

$$\lim_{x \rightarrow x_0} [f(x) \cdot g(x)] = ab \quad (2)$$

Proof of (1) Statement

Consider a sequence $\{x_j\}$ that converges to x , so that $\{f(x_j)\}$ converges to a and $\{g(x_j)\}$ converges to b . Given a criterion q , $q > 0$, choose n such that all terms x_j for $j > n$ have

$$|f(x_j) - a| < \frac{q}{2}, \text{ and } |g(x_j) - b| < \frac{q}{2}$$

Then we get $|f(x_j) + g(x_j) - a - b| < q$, and we can conclude that the limit of $f + g$ is $a + b$.

Proof of (2) Statement

As in the last argument, choose n so that all x_j for $j > n$ obey

$$|f(x_j) - a| < \frac{1}{|b| + 1} \frac{q}{2} \text{ and } |g(x_j) - b| < \frac{1}{|a| + \frac{1}{|b| + 1} \frac{q}{2}} \frac{q}{2} \quad (3)$$

We then have

$$|f(x_j)g(x_j) - ab| = |(f(x_j)a)b + (g(x_j) - b)f(x_j)| < \frac{q}{2} + \frac{q}{2} = q \quad (4)$$

which means that fg converges to ab .

Thus the limit of sum is the sum of the limits of the terms summed; and the limit of a product is the product of the limits of its factors, (when they exist).

The limit of a function $f(x)$ as x approaches c is the value that $f(x)$ approaches as x approaches the number c . mathematically this is written as:

$$\lim_{x \rightarrow c} f(x) \quad (5)$$

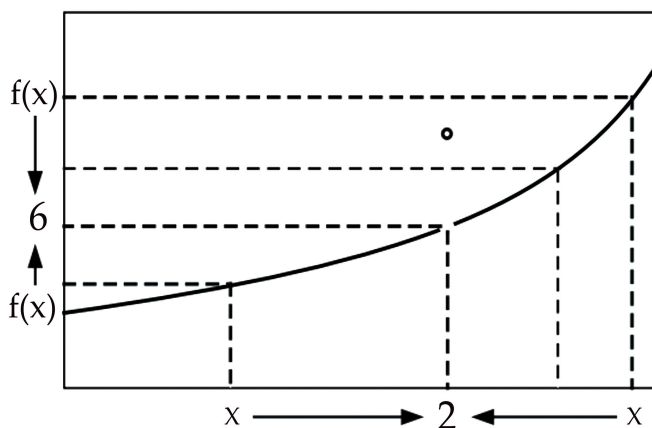
A limit looks at the behavior of $f(x)$ as x gets closer and closer to c . The truth is that when we are looking for limits, we are not interested in what happens to $f(x)$ when x actually equals c – we are only concerned with its behavior as x gets closer and closer to c .

Example: Consider the following function

$$f(x) = \begin{cases} x^2 + 2, & \text{when } x \neq 2. \\ 9 & \text{when } x = 2 \end{cases} \quad (6)$$

What happens as $x \rightarrow 2$.

Note that this function looks just like the function $f(x) = x^2 + 2$, except that $f(2) = 9$ here instead of 6. It can be seen from the graph on the next page that as x approaches 2, from both the positive and the negative side, the function $f(x)$ approaches 6. Therefore the limit of $f(x)$ as x approaches 2 is clearly 6 even though at $x = 2$ the function is clearly 9.



Left and Right Hand Limits

The last example is an example of a two-sided limit. It is considered two-sided because we get the same value for the limit whether we let x approach c “from the negative side” (left-hand limit) i.e. from values of x less than c or “from the positive side” (right-hand limit) i.e. from values of x greater than c . However, not all functions behave like this. Some functions approach different values depending on whether we let x approach c from the negative or the positive side. In general,

$$\lim_{x \rightarrow c} f(x) \text{ exists only if } \lim_{x \rightarrow c^-} f(x) = \lim_{x \rightarrow c^+} f(x) = L \quad (7)$$

In other words, the two-sided limit exists only if the left-hand and right-hand limits both exist and are equal.

Continuity

A function f is continuous at x_0 in its domain if for every sequence (x_n) with x_n in the domain of f for every n and $\lim x_n = x_0$, we have $\lim f(x_n) = f(x_0)$. We say that f is continuous if it is continuous at every point in its domain.

What does this say? It says that any time a sequence converges in the domain, the image of the sequence in the range also converges. In other words, we could either take the limit first, and then apply the function, or apply the function first, and then take the limits. Informally, f is continuous if:

$$\lim f(x_n) = f(\lim x_n)$$

This is a powerful definition because we have spent a lot of time studying sequences and limits, so we can use what we know to deduce results about continuity. In particular, we can use all the limit rules to avoid tedious calculations.

However, there is a $\varepsilon - \delta$ definition, similar to the definition of a limit, which goes as follows:

A function f is continuous at x_0 in its domain if for every $\varepsilon > 0$ there is a $\delta > 0$ such that whenever x is in the domain of f and $|x - x_0| < \delta$, we have $|f(x) - f(x_0)| < \varepsilon$. Again, we say f is continuous if it is continuous at every point in its domain.

The definition of continuity gives you a fair amount of information about a function, but this is all a waste of time unless you can show the function you are interested in is continuous. Fortunately for us, a lot of natural functions are continuous, and it is not too difficult to illustrate this is the case. There are three main techniques:

To show a function is continuous, we can do one of three things:

- (i) Show it satisfies the sequence definition of continuity
- (ii) Show it satisfies the $\varepsilon - \delta$ definition of continuity
- (iii) Decompose it into simpler functions we already know to be continuous

The most direct approach is (i): to show a function is continuous, we show it satisfies the definition of continuity. In other words, if we have a convergent sequence in the domain, then the image of the sequence converges (to the right limit). Note that this has to hold for every convergent sequence - you cannot show it works for just one. This is often a nice and clean approach for simple functions, as we can use the limit rules.

For example, suppose we have the function

$$f(x) = \frac{x^2 - 2x + 2}{x^4 + 1} \quad (8)$$

Note that the domain of this function is: the numerator and denominator are **polynomials**, hence defined for all real numbers. Moreover, as $x^4 \geq 0$, the denominator is always at least 1, so we never divide by 0.

Now suppose we have a convergent sequence (x_n) with $\lim x_n = x_0$. Then, using the limit rules, we have

Polynomial is an expression consisting of variables and coefficients which only employs the operations of addition, subtraction, multiplication, and non-negative integer exponents.

$$\begin{aligned}
\lim f(x_n) &= \lim \frac{x_n^2 - 2x_n + 2}{x_n^4 + 1} \\
&= \frac{\lim(x_n^2 - 2x_n + 2)}{\lim(x_n^4 + 1)} \\
&= \frac{\lim x_n^2 - \lim 2x_n + \lim 2}{\lim x_n^4 + \lim 1} \\
&= \frac{(\lim x_n)^2 - 2(\lim x_n) + 2}{(\lim x_n)^4 + 1} \\
&= \frac{x_0^2 - 2x_0 + 2}{x_0^4 + 1} = f(x_0)
\end{aligned} \tag{9}$$

Hence f satisfies the definition of continuity, and is thus continuous.

Thus it should be used in two situations: when we cannot compute the limit, or when we need some information about how continuous the function is.

For example, consider the function

$$f(x) = 4x^2 - 3x + 2 \tag{10}$$

Suppose we wish to show this is continuous using the $\varepsilon - \delta$ definition. Then we need to control the error $|f(x) - f(x_0)|$ in terms of $|x - x_0|$. We have

$$\begin{aligned}
|f(x) - f(x_0)| &= |(4x^2 - 3x + 2) - (4x_0^2 - 3x_0 + 2)| \\
&= |4x^2 - 4x_0^2 - 3x + 3x_0 + 2 - 2| \\
&= |4(x - x_0)(x + x_0) - 3(x - x_0)| \\
&= |(x - x_0)(4(x + x_0) - 3)| \\
&= |x - x_0| |4(x + x_0) - 3| \\
&\leq |x - x_0| (4|x + x_0| + 3)
\end{aligned} \tag{11}$$

where in the last inequality we use the triangle inequality. The first factor is in the form we want, as it is in terms of $x - x_0$. However, the second term also depends on x , so we must rewrite it in terms of $x - x_0$. We can then use the triangle inequality again to obtain an upper bound.

$$\begin{aligned}
|f(x) - f(x_0)| &\leq |x - x_0| (4|x + x_0| + 3) \\
&= |x - x_0| (4|x - x_0 + 2x_0| + 3) \\
&\leq |x - x_0| (4|x - x_0| + 8|x_0| + 3)
\end{aligned} \tag{12}$$

Thus we have bounded the error $|f(x) - f(x_0)|$ in terms of $|x - x_0|$. We now need to show that given any $\varepsilon > 0$, we can choose $\delta > 0$ small enough so that if $|x - x_0| < \delta$, then $|f(x) - f(x_0)| < \varepsilon$. Indeed, suppose $|x - x_0| < 1$. Then we have

$$|f(x) - f(x_0)| \leq |x - x_0| (4|x - x_0| + 8|x_0| + 3) \leq |x - x_0| (8|x_0| + 7) \tag{13}$$

If we further have

$$|x - x_0| < \frac{\varepsilon}{8|x_0| + 7} \tag{14}$$

Then

$$|f(x) - f(x_0)| \leq |x - x_0| (8|x_0| + 7) < \epsilon \quad (15)$$

as desired. Hence we may take $\delta = \min \left\{ 1, \frac{\epsilon}{8|x_0|+7} \right\}$ in the ϵ - δ definition, showing that f is continuous. Note that in this case δ depends on both ϵ and x_0 , but this is permitted.

The final method, of decomposing a function into simple **continuous functions**, is the simplest, but requires that you have a set of basic continuous functions to start with - somewhat akin to using limit rules to find limits. So, for example, if we know that both $g(x) = x$ and the constant function $h(x) = k$ (for $k \in \mathbb{R}$) are continuous, then we can show that

$$f(x) = \frac{x^2 - 2x + 2}{x^4 + 1} \quad (16)$$

is continuous, since it is the quotient of $f_1(x) = x^2 - 2x + 2$ and $f_2(x) = x^4 + 1$. Now x^2 and $2x$ are the products of continuous functions, hence continuous, and 2 is a continuous function, so $f_1(x) = x^2 - 2x + 2$ is continuous. As x^2 is continuous, x^4 is continuous (either since it is x^2 multiplied by itself, or x^2 composed with itself), and so $f_2(x) = x^4 + 1$ is continuous. As $f_2(x) \neq 0$ for all x , the quotient $f(x) = f_1(x) / f_2(x)$ is continuous everywhere.

Continuous function is a function for which sufficiently small changes in the input result in arbitrarily small changes in the output. Otherwise, a function is said to be a discontinuous function.

Partial Derivatives

Just as derivatives can be used to explore the properties of functions of 1 variable, so also derivatives can be used to explore functions of 2 variables.

Let $f(x, y)$ be a function with two variables. If we keep y constant and differentiate f (assuming f is differentiable) with respect to the variable x , we obtain what is called the partial derivative of f with respect to x which is denoted by

$$\frac{\partial f}{\partial x} \text{ or } f_x$$

Similarly If we keep x constant and differentiate f (assuming f is differentiable) with respect to the variable y , we obtain what is called the partial derivative of f with respect to y which is denoted by

$$\frac{\partial f}{\partial y} \text{ or } f_y$$

We might also define partial derivatives of function f as follows:

$$\begin{aligned}\frac{\partial f}{\partial x} &= \lim_{h \rightarrow 0} \frac{f(x+h, y) - f(x, y)}{h} \\ \frac{\partial f}{\partial y} &= \lim_{k \rightarrow 0} \frac{f(x, y+k) - f(x, y)}{k}\end{aligned}\quad (17)$$

We now present several examples with detailed solution on how to calculate partial derivatives.

Example: Find the partial derivatives f_x and f_y if $f(x, y)$ is given by:

$$f(x, y) = x^2 y + 2x + y \quad (18)$$

Solution: Assume y is constant and differentiate with respect to x to obtain

$$\begin{aligned}f_x &= \frac{\partial f}{\partial x} = \frac{\partial}{\partial x} [x^2 y + 2x + y] \\ &= \frac{\partial}{\partial x} [x^2 y] + \frac{\partial}{\partial x} [2x] + \frac{\partial}{\partial x} [y] = [2xy] + [2] + [0] = 2xy + 2\end{aligned}\quad (19)$$

Now assume x is constant and differentiate with respect to y to obtain

$$\begin{aligned}f_y &= \frac{\partial f}{\partial y} = \frac{\partial}{\partial y} [x^2 y + 2x + y] \\ &= \frac{\partial}{\partial y} [x^2 y] + \frac{\partial}{\partial y} [2x] + \frac{\partial}{\partial y} [y] = [x^2] + [0] + [1] = x^2 + 1\end{aligned}\quad (20)$$

Now that we have the brief discussion on limits out of the way we can proceed into taking derivatives of functions of more than one variable. Before we actually start taking derivatives of functions of more than one variable let's recall an important interpretation of derivatives of functions of one variable.

Recall that given a function of one variable, $f(x)$ the derivative, $f'(x)$ represents the rate of change of the function as x changes. This is an important interpretation of derivatives and we are not going to want to lose it with functions of more than one variable. The problem with functions of more than one variable is that there is more than one variable. In other words, what do we do if we only want one of the variables to change, or if we want more than one of them to change? In fact, if we are going to allow more than one of the variables to change there are then going to be an infinite amount of ways for them to change. For instance, one variable could be changing faster than the other variable(s) in the function. Notice as well that it will

be completely possible for the function to be changing differently depending on how we allow one or more of the variables to change. We will need to develop ways, and notations, for dealing with all of these cases.

Let's start with the function $f(x,y) = 2x^2y^3$ and let's determine the rate at which the function is changing at a point, (a,b) if we hold y fixed and allow x to vary and if we hold x fixed and allow y to vary.

We will start by looking at the case of holding y fixed and allowing x to vary. Since we are interested in the rate of change of the function at (a, b) and are holding y fixed this means that we are going to always have $y=b$ (if we did not have this then eventually y would have to change in order to get to the point). Doing this will give us a function involving only x 's and we can define a new function as follows:

$$g(x) = f(x, b) = 2x^2b^3 \quad (21)$$

Now, this is a function of a single variable and at this point all that we are asking is to determine the rate of change of $g(x)$ at $x=a$. In other words, we want to compute $g'(a)$ and since this is a function of a single variable we already know how to do that. Here is the rate of change of the function at (a, b) if we hold y fixed and allow x to vary.

$$g'(a) = 4ab^3$$

We will call $g'(a)$ the partial derivative of $f(x,y)$ with respect to x at (a,b) and we will denote it in the following way,

$$f_x(a, b) = 4ab^3$$

Now, let's do it the other way. We will now hold x fixed and allow y to vary. We can do this in a similar way. Since we are holding x fixed it must be fixed at $x=a$ and so we can define a new function of y and then differentiate this as we have always done with functions of one variable.

Here is the work for this,

$$h(y) = f(a, y) = 2a^2y^3 \Rightarrow h'(y) = 6a^2y^2 \quad (22)$$

In this case we call $h'(y)$ the partial derivative of $f(x,y)$ with respect to y at (a,b) and we denote it as follows,

$$f_y(a, b) = 6a^2b^2 \quad (23)$$

These two partial derivatives are sometimes called the first order partial derivatives. Just as with functions of one variable we can have derivatives of all orders. We will be looking at higher order derivatives.

The notation for partial derivatives is different than that for derivatives of functions of a single variable. With functions of a single variable we could denote the

derivative with a single prime. However, with partial derivatives we will always need to remember the variable that we are differentiating with respect to and so we will subscript the variable that we differentiated with respect to. We will shortly be seeing some alternate notation for partial derivatives as well.

Note as well that we usually do not use the (a,b) notation for partial derivatives. The more standard notation is to just continue to use (x,y). So, the partial derivatives will more commonly be written as,

$$f_x(x, y) = 4xy^3 \quad \text{and} \quad f_y(x, y) = 6x^2y^2$$

Now, as this quick example has shown taking derivatives of functions of more than one variable is done in pretty much the same manner as taking derivatives of a single variable. To compute $f_x(x, y)$ all we need to do is treat all the y 's as constants (or numbers) and then differentiate the x 's as we have always done. Likewise, to compute $f_y(x, y)$ we will treat all the x 's as constants and then differentiate the y 's as we are used to doing. Before we work any examples let's get the formal definition of the partial derivative out of the way as well as some alternate notation.

Since we can think of the two partial derivatives above as derivatives of single variable functions it should not be too surprising that the definition of each is very similar to the definition of the derivative for single variable functions.

Here are the formal definitions of the two partial derivatives:

$$f_x(x, y) = \lim_{h \rightarrow 0} \frac{f(x+h, y) - f(x, y)}{h} \quad f_y(x, y) = \lim_{h \rightarrow 0} \frac{f(x, y+h) - f(x, y)}{h}$$

Now let's take a quick look at some of the possible alternate notations for partial derivatives. Given the function $z=f(x, y)$ the following are all equivalent notations,

$$\begin{aligned} f_x(x, y) &= f_x = \frac{df}{dx} = \frac{d}{dx}(f(x, y)) = z_x = \frac{dz}{dx} = D_x f \\ f_y(x, y) &= f_y = \frac{df}{dy} = \frac{d}{dy}(f(x, y)) = z_y = \frac{dz}{dy} = D_y f \end{aligned} \quad (24)$$

For the fractional notation for the partial derivative notice the difference between the partial derivative and the ordinary derivative from single variable calculus.

$$\begin{aligned} f(x) &\Rightarrow f'(x) = \frac{df}{dx} \\ f(x, y) &\Rightarrow f_x(x, y) = \frac{df}{dx} \quad \& \quad f_y(x, y) = \frac{df}{dy} \end{aligned} \quad (25)$$

Total Derivative

A total derivative or full derivative of a function f of several variables, e.g., t, x, y , etc., with respect to an exogenous argument, e.g., t , is the limiting ratio of the change in

the function's value to the change in the exogenous argument's value (for arbitrarily small changes), taking into account the exogenous argument's direct effect as well as its indirect effects via the other arguments of the function.

The total derivative of a function is different from its corresponding partial derivative (δ). Calculation of the total derivative of f with respect to t does not assume that the other arguments are constant while t varies; instead, it allows the other arguments to depend on t . The total derivative adds in these indirect dependencies to find the overall dependency of f on t for example, the total derivative of $f(t, x, y)$ with respect to t is

$$\frac{df}{dt} = \frac{\partial f}{\partial t} \frac{dt}{dt} + \frac{\partial f}{\partial x} \frac{dx}{dt} + \frac{\partial f}{\partial y} \frac{dy}{dt} \quad (26)$$

Which simplifies to

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial x} \frac{dx}{dt} + \frac{\partial f}{\partial y} \frac{dy}{dt} \quad (27)$$

Consider multiplying both sides of the equation by the differential dt :

$$df = \frac{\partial f}{\partial t} dt + \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy \quad (28)$$

The result is the differential change df in, or total differential of, the function f . Because f depends on t , some of that change will be due to the partial derivative of f with respect to t . However, some of that change will also be due to the partial derivatives of f with respect to the variables x and y . So, the differential dt is applied to the total derivatives of x and y to find differentials dx and dy , which can then be used to find the contribution to df .

Differentiation with Indirect Dependencies

Suppose that f is a function of two variables, x and y . normally these variables are assumed to be independent. However, in some situations they may be dependent on each other. For example y could be a function of x , constraining the domain of f to a curve in R^2 . In this case the partial derivative of f with respect to x does not give the true rate of change of f with respect to changing x because changing x necessarily changes y . The total derivative takes such dependencies into account.

For example, suppose

$$f(x, y) = xy$$

The rate of change of f with respect to x is usually the partial derivative of f with respect to x ; in this case,

$$\frac{df}{dx} = y$$

However, if y depends on x , the partial derivative does not give the true rate of change of f as x changes because it holds y fixed.

Suppose we are constrained to the line

$$Y=x$$

then

$$f(x, y) = f(x, x) = x^2$$

In that case, the total derivative of f with respect to x is

$$\frac{df}{dx} = 2x$$

Instead of immediately substituting for y in terms of x , this can be found equivalently using the chain rule:

$$\frac{df}{dx} = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} \frac{dy}{dx} = y + x \cdot 1 = x + y \quad (29)$$

Notice that this is not equal to the partial derivative:

$$\frac{df}{dx} = 2x \neq \frac{\partial f}{\partial x} = y = x \quad (30)$$

While one can often perform substitutions to eliminate indirect dependencies, the chain rule provides for a more efficient and general technique. Suppose $M(t, p_1, \dots, p_n)$ is a function of time t and n variables p_i which themselves depend on time. Then, the total time derivative of M is

$$\frac{dM}{dt} = \frac{d}{dt} M(t, p_1(t), \dots, p_n(t)) \quad (31)$$

The chain rule for differentiating a function of several variables implies that

$$\frac{dM}{dt} = \frac{\partial M}{\partial t} + \sum_{i=1}^n \frac{\partial M}{\partial p_i} \frac{dp_i}{dt} = \left(\frac{\partial}{\partial t} + \sum_{i=1}^n \frac{dp_i}{dt} \frac{\partial}{\partial p_i} \right) (M) \quad (32)$$

This expression is often used in physics for a gauge transformation of the Lagrangian, as two Lagrangians that differ only by the total time derivative of a function of time and the n generalized coordinates lead to the same equations of motion. An interesting example concerns the resolution of causality concerning the Wheeler–Feynman time-symmetric theory. The operator in brackets (in the final expression) is also called the total derivative operator (with respect to t).

For example, the total derivative of $f(x(t), y(t))$ is

$$\frac{df}{dt} = \frac{\partial f}{\partial x} \frac{dx}{dt} + \frac{\partial f}{\partial y} \frac{dy}{dt} \quad (33)$$

Here there is no $\partial f / \partial t$ term since f itself does not depend on the independent variable t directly.

The total derivative via differentials

Differentials provide a simple way to understand the total derivative. For instance, suppose $M(t, p_1, \dots, p_n)$ is a function of time t and n variables p_i . Then, the differential of M is

$$dM = \frac{\partial M}{\partial t} dt + \sum_{i=1}^n \frac{\partial M}{\partial p_i} dp_i \quad (34)$$

This expression is often interpreted heuristically as a relation between infinitesimals. However, if the variables t and p_i are interpreted as functions, and $M(t, p_1, \dots, p_n)$ is interpreted to mean the composite of M with these functions, then the above expression makes perfect sense as an equality of differential 1-forms, and is immediate from the chain rule for the exterior derivative. The advantage of this point of view is that it takes into account arbitrary dependencies between the variables.

For example, if $p_1^2 = p_2 p_3$ then $2p_1 dp_1 = p_3 dp_2 + p_2 dp_3$. In particular, if the variables p_i are all functions of t , then

$$dM = \frac{\partial M}{\partial t} dt + \sum_{i=1}^n \frac{\partial M}{\partial p_i} \frac{\partial p_i}{\partial t} dt \quad (35)$$

Dividing through by dt gives the total derivative dM / dt .

The total derivative as a linear map

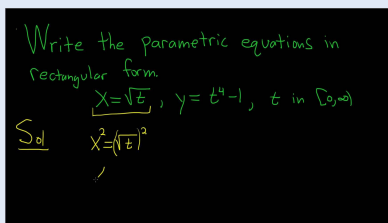
Let $U \subseteq \mathbb{R}^n$ be an open subset. Then a function $f: U \rightarrow \mathbb{R}^m$ is said to be (totally) differentiable at a point $p \in U$, if there exists a linear map $df_p: \mathbb{R}^n \rightarrow \mathbb{R}^m$ (also denoted $D_p f$ or $Df(p)$) such that

$$\lim_{x \rightarrow p} \frac{\|f(x) - f(p) - df_p(x - p)\|}{\|x - p\|} = 0 \quad (36)$$

The linear map df_p is called the (total) derivative or (total) differential of f at P . A function is (totally) differentiable if its total derivative exists at every point in its domain.

Note that f is differentiable if and only if each of its components $f_i: U \rightarrow \mathbb{R}$ is differentiable. For this it is necessary, but not sufficient, that the partial derivatives of each function f_j exist. However, if these partial derivatives exist and are continuous, then f is differentiable and its differential at any point is the linear map determined by the Jacobian matrix of partial derivatives at that point.

A partial derivative of a function of several variables is its derivative with respect to one of those variables, with the others held constant (as opposed to the total derivative, in which all variables are allowed to vary). Partial derivatives are used in vector calculus and differential geometry.



CORE CONCEPT: DIFFERENTIATION OF IMPLICIT FUNCTIONS

An implicit function is a function that is defined implicitly by an implicit equation, by associating one of the variables (the value) with the others (the arguments).

Sometimes functions are given not in the form $y = f(x)$ but in a more complicated form in which it is difficult or impossible to express y explicitly in terms of x . Such functions are called implicit functions.

Consider an expression such as

$$x^2 + y^2 - 4x + 5y - 8 = 0 \quad (37)$$

It would be quite difficult to re-arrange this so y was given explicitly as a function of x . We could perhaps, given values of x , use the expression to work out the values of y and thereby draw a graph. In general even if this is possible, it will be difficult. A function given in this way is said to be defined implicitly.

Example: Suppose we want to differentiate the implicit function

$$y^2 + x^3 - y^3 + 6 = 3y \quad (38)$$

with respect x .

We differentiate each term with respect to x :

$$\frac{d}{dx} (y^2) + \frac{d}{dx} (x^3) - \frac{d}{dx} (y^3) + \frac{d}{dx} (6) = \frac{d}{dx} (3y) \quad (39)$$

Differentiating functions of x with respect to x is straightforward. But when differentiating a function of y with respect to x we must remember the rule given in the next key point. We find

$$\frac{d}{dy} (y^2) \times \frac{dy}{dx} + 3x^2 - \frac{d}{dy} (y^3) \times \frac{dy}{dx} + 0 = \frac{d}{dy} (3y) \times \frac{dy}{dx} \quad (40)$$

that is

$$2y \frac{dy}{dx} + 3x^2 - 3y^2 \frac{dy}{dx} = 3 \frac{dy}{dx} \quad (41)$$

We rearrange this to collect all terms involving dy/dx together

$$3x^2 = 3\frac{dy}{dx} - 2y\frac{dy}{dx} + 3y^2\frac{dy}{dx} \quad (42)$$

Then

$$3x^2 = (3 - 2y + 3y^2) \frac{dy}{dx} \quad (43)$$

so that, finally

$$\frac{dy}{dx} = \frac{3x^2}{3 - 2y + 3y^2} \quad (44)$$

This is the expression for dy/dx .

Example: Suppose we want to differentiate, with respect to x , the implicit function

$$\sin y + x^2 y^3 - \cos x = 2y \quad (45)$$

We differentiate each term with respect to x .

$$\frac{d}{dx}(\sin y) + \frac{d}{dx}(x^2 y^3) - \frac{d}{dx}(\cos x) = \frac{d}{dx}(2y) \quad (46)$$

Recognize that the second term is a product and we will need the product rule. We will also use the chain rule to differentiate the functions of y . We find

$$\frac{d}{dy}(\sin y) \times \frac{dy}{dx} + \left\{ x^2 \frac{d}{dy}(y^3) + y^3 \frac{d}{dx}(x^2) \right\} + \sin x = \frac{d}{dy}(2y) \times \frac{dy}{dx} \quad (47)$$

so that

$$\cos y \frac{dy}{dx} + \left\{ x^2 \cdot \frac{d}{dy}(y^3) \frac{dy}{dx} + y^3 \cdot 2x \right\} + \sin x = 2 \frac{dy}{dx} \quad (48)$$

Tidying this up gives

$$\cos y \frac{dy}{dx} + x^2 3y^2 \frac{dy}{dx} + 2xy^3 + \sin x = 2 \frac{dy}{dx} \quad (49)$$

We now start to collect together terms involving dy/dx .

$$2xy^3 + \sin x = 2\frac{dy}{dx} - \cos y \frac{dy}{dx} - 3x^2y^2 \frac{dy}{dx}$$

$$2xy^3 + \sin x = (2 - \cos y - 3x^2y^2) \frac{dy}{dx} \quad (50)$$

So that, finally

$$\frac{dy}{dx} = \frac{2xy^3 + \sin x}{2 - \cos y - 3x^2y^2} \quad (51)$$

We have deliberately included plenty of detail in this calculation. With practice you will be able to omit many of the intermediate stages. The following two examples show how you should aim to condense the solution.

Jacobian Properties

Jacobian is a determinant which is defined for a finite number of functions of the same number of variables and in which each row consists of the first partial derivatives of the same function with respect to each of the variables.

Most mechanisms of interest to computer animation are too complex to allow an analytic solution. For these, the motion can be incrementally constructed. At each time step, the best way to change each joint angle in order to direct the current position and orientation of the end effector toward the desired configuration is computed. The computation is performed by forming the Jacobian matrix which is a matrix of partial derivatives.

In order to explain the Jacobian from a strictly mathematical point of view, consider the six arbitrary functions of equation (52), each of which is a function of six independent variables. Given specific values for the input variables, the x_i s, each of the output variables, the y_i s, can be computed by its respective function.

$$\begin{aligned} y_1 &= f_1(x_1, x_2, x_3, x_4, x_5, x_6) \\ y_2 &= f_2(x_1, x_2, x_3, x_4, x_5, x_6) \\ y_3 &= f_3(x_1, x_2, x_3, x_4, x_5, x_6) \\ y_4 &= f_4(x_1, x_2, x_3, x_4, x_5, x_6) \\ y_5 &= f_5(x_1, x_2, x_3, x_4, x_5, x_6) \\ y_6 &= f_6(x_1, x_2, x_3, x_4, x_5, x_6) \end{aligned} \quad (52)$$

The differentials of y_i can be written in terms of the differentials of x_i using the chain rule. This generates Equation (53)

$$\delta y_i = \frac{\delta f_i}{\delta x_1} \cdot \delta x_1 + \frac{\delta f_i}{\delta x_2} \cdot \delta x_2 + \frac{\delta f_i}{\delta x_3} \cdot \delta x_3 + \frac{\delta f_i}{\delta x_4} \cdot \delta x_4 + \frac{\delta f_i}{\delta x_5} \cdot \delta x_5 + \frac{\delta f_i}{\delta x_6} \cdot \delta x_6 \quad (53)$$

Equation (52) and (53) can be put in vector notation producing Equation (54) and Equation 5.

$$Y = F(X) \quad (54)$$

$$\delta Y = \frac{\partial F}{\partial X} \cdot \delta X \quad (55)$$

The 6x6 matrix of partial derivatives $\frac{\partial F}{\partial X}$ is called the Jacobian and is a function of the current values of the x_i . The Jacobian can be thought of mapping the velocities of X to the velocities of Y (Equation (56)). At any particular point in time, the Jacobian is a linear function of the x_i 's. At the next instant of time, X has changed and so has the linear transformation represented by the Jacobian.

$$\dot{Y} = J(X) \cdot \dot{X} \quad (56)$$

In applying the Jacobian to a linked appendage, the input variables, x_i 's, become the joint angles and the output variables, y_i 's, become the end effector position and orientation. In this case, the Jacobian relates the velocities of the joint angles to the velocities of the end effector position and orientation (Equation (57)).

$$V = J(\theta)\dot{\theta} \quad (57)$$

V is the vector of linear and rotational velocities and represents the desired change in the end effector. The desired change will be based on the difference between its current position/orientation to that specified by the goal configuration. These velocities are vectors in three-space so each has an x, y, and z component (Equation (58)) is a vector of joint angles velocities and are the unknowns of the equation (Equation (59)). J, the Jacobian, is a matrix which relates the two and is a function of the current pose (Equation (60)).

$$V = [v_x, v_y, v_z, \omega_x, \omega_y, \omega_z]^T \quad (58)$$

$$\dot{\theta} = [\dot{\theta}_1, \dot{\theta}_2, \dot{\theta}_3, \dots, \dot{\theta}_n]^T \quad (59)$$

$$J = \begin{bmatrix} \frac{\partial v_x}{\partial \theta_1} & \frac{\partial v_x}{\partial \theta_2} & \cdots & \frac{\partial v_x}{\partial \theta_n} \\ \frac{\partial v_y}{\partial \theta_1} & \frac{\partial v_y}{\partial \theta_2} & \cdots & \frac{\partial v_y}{\partial \theta_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial \omega_z}{\partial \theta_1} & \frac{\partial \omega_z}{\partial \theta_2} & \cdots & \frac{\partial \omega_z}{\partial \theta_n} \end{bmatrix} \quad (60)$$

Each term of the Jacobian relates the change of a specific joint with a change in the end effector. The rotational change in the end effector, ω , is merely the velocity of the joint angle about the axis of revolution at the joint under consideration. The linear change in the end effector is the cross product of the axis of revolution and a vector from the joint to the end effector. This is the linear direction of travel instantaneously induced at the end effector by the rotation at the joint. (See Figure 1).

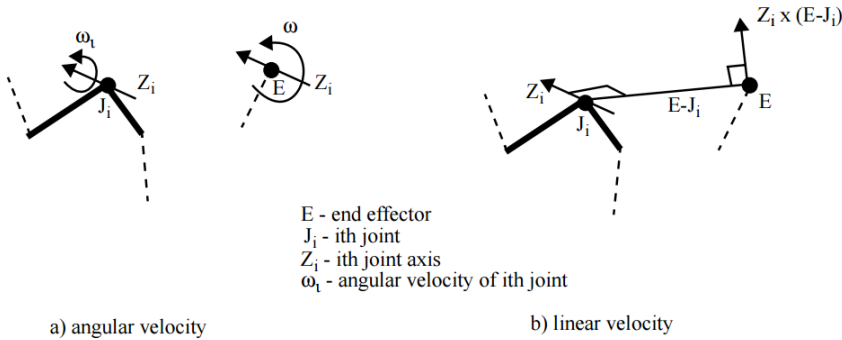


Figure 1: Angular and linear velocities induced by joint axis rotation.

The desired angular and linear velocities are computed from the difference between the current configuration of the end effector and the desired configuration. The angular and linear velocities of the end effector induced by the rotation of a specific joint axis are determined by the computations shown in Figure 1. The problem is to determine the best linear combination of velocities induced by the various joints that would result in the desired velocities of the end effector. By posing the problem in matrix form, the Jacobian is formed. In assembling the Jacobian, it is important to make sure that all of the coordinate values are in the same coordinate system. It is often the case that joint specific information is given in the coordinate system local to that joint. In forming the Jacobian matrix, this information must be converted into some common coordinate system such as the global inertial coordinate system or the end effector coordinate system. Various methods have been developed for computing the Jacobian based on attaining maximum computational efficiency given the required information in local coordinate systems but all methods produce the derivative matrix in a common coordinate system.

A Simple Example

Consider the simple three revolute joint, planar manipulator of Figure 2. In this example the objective is to move the end effector, E, to the goal position, G. The orientation of the end effector is of no concern in this example.

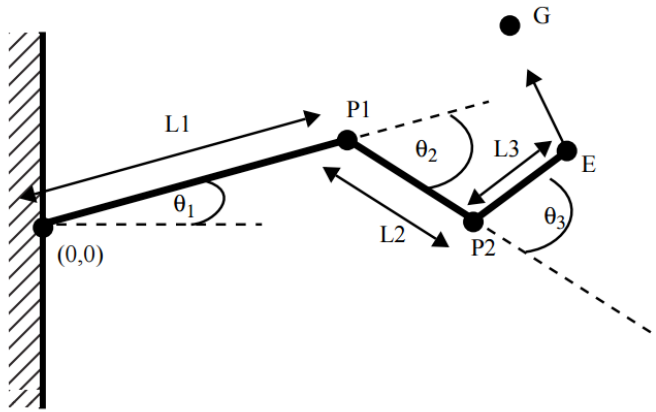


Figure 2: Planar, three joint manipulator.

The axis of rotation of each joint is perpendicular to the figure, coming out of the paper. The effect of an incremental rotation, g_i , of each joint can be determined by the cross product of the joint axis and the vector from the joint to the end effector, V_i (Figure 3). Notice that the magnitude of the g_i 's is a function of the distance between the locations of the joint and the end effector.

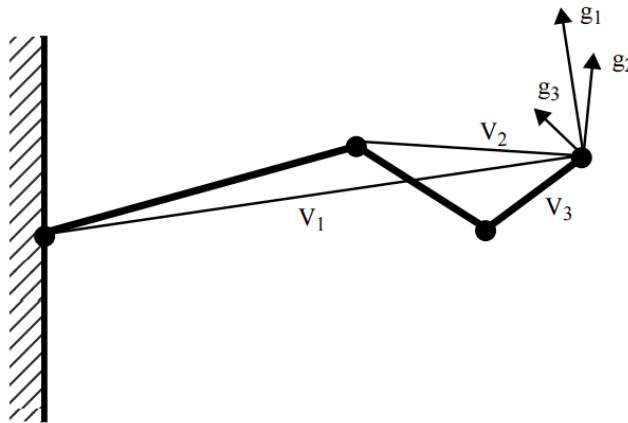


Figure 3: Instantaneous changes in position induced by joint angle rotations.

The desired change to the end effector is the difference between the current position of the end effector and the goal position. A vector of the desired change values is set equal to the Jacobian matrix multiplied by a vector of the unknown values: the change to the joint angles (Equation (61))

$$\begin{bmatrix} (G-E)_x \\ (G-E)_y \\ (G-E)_z \end{bmatrix} = \begin{bmatrix} ((0,0,1) \times E)_x & (0,0,1) \times (E-P1)_x & (0,0,1) \times (E-P2)_x \\ ((0,0,1) \times E)_y & (0,0,1) \times (E-P1)_y & (0,0,1) \times (E-P2)_y \\ ((0,0,1) \times E)_z & (0,0,1) \times (E-P1)_z & (0,0,1) \times (E-P2)_z \end{bmatrix} \cdot \begin{bmatrix} \dot{\theta}_1 \\ \dot{\theta}_2 \\ \dot{\theta}_3 \end{bmatrix} \quad (61)$$

Solution Using the Inverse Jacobian

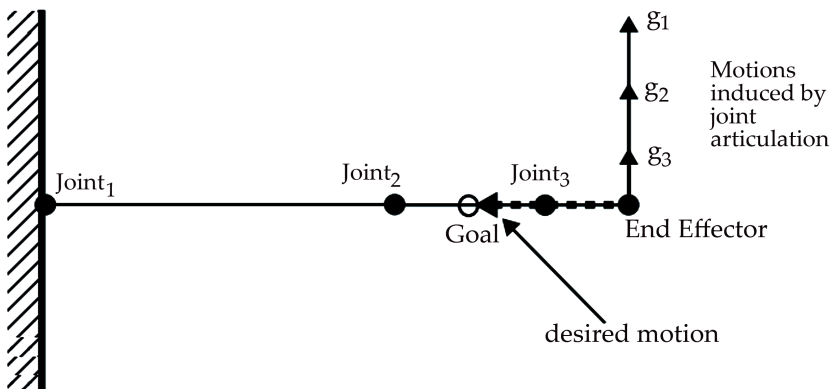
Once the Jacobian has been computed, then an equation in the form of Equation (62) must be solved.

$$V = J\dot{\theta} \quad (62)$$

In the case that J is square, the inverse of the Jacobian is needed to compute the joint angle velocities given the end effector velocities.

$$J^{-1}V = \dot{\theta} \quad (63)$$

If the inverse of the Jacobian (J^{-1}) does not exist, then the system is said to be singular for the given joint angles. A singularity occurs when a linear combination of the joint angle velocities cannot be formed to produce the desired end effector velocities. As a simple example of such a situation, consider a fully extended, planar arm with a goal position somewhere on the forearm. In such a case, a change in each joint angle would produce a vector perpendicular to the desired direction. Obviously, no linear combination of these vectors could produce the desired motion vector. Unfortunately, determining all of the singularities of a system cannot be determined simply by visually inspecting the possible geometric configurations of the linkage.



Problems with singularities can be reduced if the manipulator is redundant - when there are more degrees of freedom than there are constraints to be satisfied. In this case, the Jacobian is not a square matrix and there are an infinite number of solutions to the inverse kinematics problem. Because the Jacobian is not square, a conventional inverse does not exist. Instead, the pseudo inverse, J^+ , can be used (Equation (64)).

$$\begin{aligned}
 V &= J\dot{\theta} \\
 J^T V &= J^T J \dot{\theta} \\
 (J^T J)^{-1} J^T \dot{V} &= (J^T J)^{-1} J^T J \dot{\theta} \\
 J^+ V &= \dot{\theta}
 \end{aligned} \tag{64}$$

Equation (64) works because a matrix multiplied by its own transpose will be a square n by n matrix.

$J^+ = (J^T J)^{-1} J^T = J^T (J J^T)^{-1}$ is called the pseudo-inverse of J . It maps the desired velocities of the end effector to the required velocities of the joint angles.

$$\begin{aligned}
 J^+ V &= \dot{\theta} \\
 J^T (J J^T)^{-1} V &= \dot{\theta} \\
 \beta &= (J J^T)^{-1} V
 \end{aligned} \tag{65}$$

$$(J J^T) \beta = V \tag{66}$$

$$J^T \beta = \dot{\theta} \tag{67}$$

Gaussian elimination can be used to solve EQ 15 for β . This can then be substituted into Equation (67) to solve for $\dot{\theta}$.

The pseudo inverse solution computes one of many possible solutions. This solution minimizes joint angle rates. The configurations produced, however, do not necessarily correspond to what might be considered natural poses. In order to better control the kinematic model, a control expression can be added to the pseudo inverse Jacobian solution. The control expression is used to solve for control angle rates with certain attributes. The added solution contributes nothing to the desired **end effector** motion. The form for the control expression is shown in Equation (68). It is shown that expression does not add anything to the velocities (Equation (69)).

$$\dot{\theta} = (J^+ J - I)z \tag{68}$$

End effector is a device or tool that's connected to the end of a robot arm where the hand would be. The end effector is the part of the robot that interacts with the environment.

$$\begin{aligned}
 V &= J\dot{\theta} \\
 V &= J(J^+J - I)z \\
 V &= (JJ^+J - J)z \\
 V &= (J - J)z \\
 V &= 0 \cdot z \\
 V &= 0
 \end{aligned} \tag{69}$$

As a consequence, the control expression can be combined with the pseudo inverse Jacobian solution so that the given velocities are still satisfied.

In order to bias the solution toward specific joint angles, H is defined as in Equation (70) where θ_i are the current joint angles, θ_{ci} are the desired joint angles, α_i are the desired angle gains, ψ and is the ψ th norm (for even ψ). z is equal to the gradient of H , ∇H (Equation (71)).

$$H = \sum_{i=1}^n \alpha_i \cdot (\theta_i - \theta_{ci})^\psi \tag{70}$$

$$z = \nabla_{\theta} H = \frac{dH}{d\theta} = \psi \sum_{i=1}^n \alpha_i \cdot (\theta_i - \theta_{ci})^{\psi-1} \tag{71}$$

The desired angles and gains are input parameters. The gain indicates the relative importance of the associated desired angle. The higher the gain, the stiffer the joint. If the gain for a particular joint is high, then the solution will be such that the joint angle quickly approaches the desired joint angle. The control expression is added to the solution indicated by the conventional pseudo inverse of the Jacobian (Equation (72)). If all gains are zero, then the solution will reduce to the conventional pseudo inverse of the Jacobian.

$$\dot{\theta} = J^+ V + (J^+ J - I) \nabla_{\theta} H \tag{72}$$

Equation (73) can be solved by rearranging terms as shown in Equation (73).

$$\begin{aligned}
 \dot{\theta} &= J^+ V + (J^+ J - I) \nabla_{\theta} H \\
 \dot{\theta} &= J^+ V + J^+ J \nabla_{\theta} H - I \nabla_{\theta} H \\
 \dot{\theta} &= J^+ (V + J \nabla_{\theta} H) - \nabla_{\theta} H \\
 \dot{\theta} &= J^T (J J^T)^{-1} (V + J \nabla_{\theta} H) - \nabla_{\theta} H \\
 \dot{\theta} &= J^T [(J J^T)^{-1} (V + J \nabla_{\theta} H)] - \nabla_{\theta} H
 \end{aligned} \tag{73}$$

In order to solve this, set $\beta = (JJ^T)^{-1}(V + J\nabla_{\theta}H)$ so that Equation (73) becomes Equation (74). Use Gaussian elimination to solve for in Equation (75). Substitute the solution for into Equation (74) to solve for $\dot{\theta}$.

$$\dot{\theta} = J^T \beta - \nabla_{\theta} H \quad (74)$$

$$V + J\nabla_{\theta} H = (JJ^T) \beta \quad (75)$$

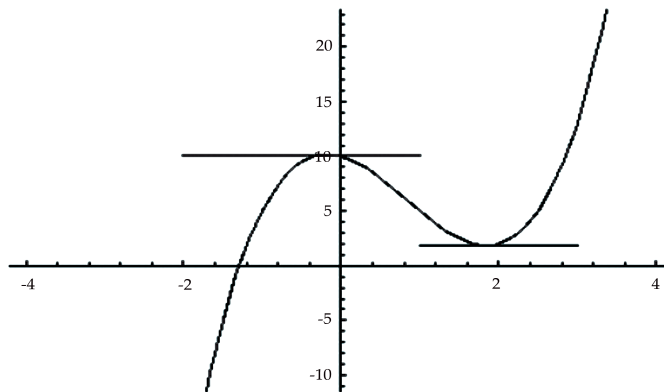
Simple Euler integration can be used at this point to update the joint angles. At the next time step, the Jacobian has changed so the computation must be performed again and another step taken. This process repeats until the end effector reaches the goal configuration within some acceptable (i.e., user defined) tolerance.

Maxima and Minima of Functions of Two Variables

The problem of determining the maximum or minimum of function is encountered in geometry, mechanics, physics, and other fields, and was one of the motivating factors in the development of the calculus in the seventeenth century.

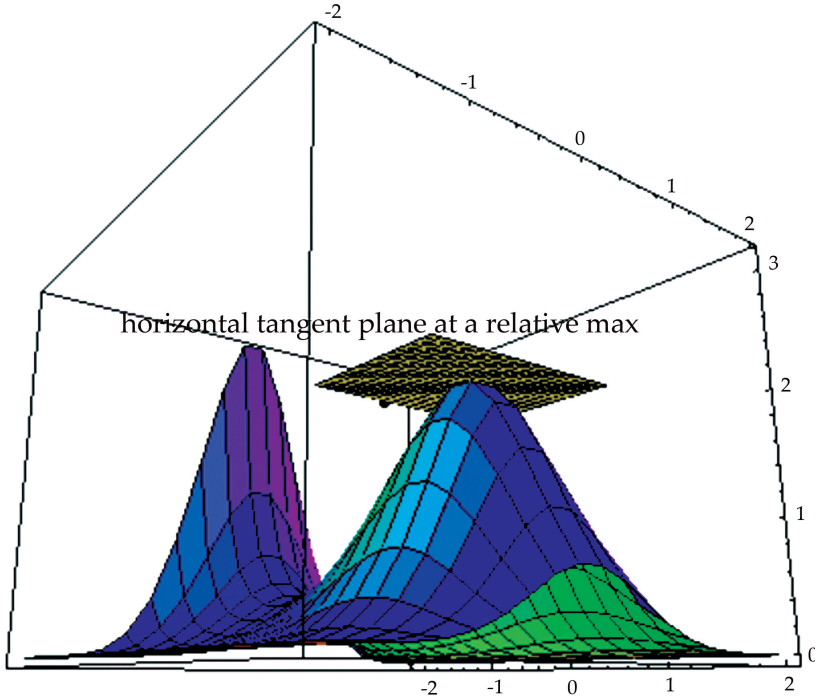
Let us recall the procedure for the case of a function of one variable $y=f(x)$. First, we determine points x_c where $f'(x)=0$. These points are called critical points. At critical points the tangent line is horizontal. This is shown in the figure.

The Jacobian generalizes the gradient of a scalar-valued function of multiple variables, which itself generalizes the derivative of a scalar-valued function of a single variable.



The second derivative test is employed to determine if a critical point is a relative maximum or a relative minimum. If $f''(x_c) > 0$, then x_c is a relative minimum. If $f''(x_c) < 0$, then x_c is a maximum. If $f''(x_c) = 0$, then the test gives no information.

The notions of critical points and the second derivative test carry over to functions of two variables. Let $z=f(x, y)$. Critical points are points in the xy -plane where the tangent plane is horizontal.



Since the normal vector of the tangent plane at (x,y) is given by

$$f_x(x, y)\mathbf{i} + f_y(x, y)\mathbf{j} - \mathbf{k} \quad (76)$$

The tangent plane is horizontal if its normal vector points in the z direction. Hence, critical points are solutions of the equations:

$$f_x(x, y) = 0 \quad \text{and} \quad f_y(x, y) = 0 \quad (77)$$

because horizontal planes have normal vector parallel to z -axis. The two equations above must be solved simultaneously.

Example: Let us find the critical points of

$$z = f(x, y) = \exp\left(-\frac{1}{3}x^3 + x - y^2\right) \quad (78)$$

The partial derivatives are

$$\begin{aligned} f_x(x, y) &= (-x^2 + 1)\exp\left(-\frac{1}{3}x^3 + x - y^2\right) \\ f_y(x, y) &= -2y\exp\left(-\frac{1}{3}x^3 + x - y^2\right) \end{aligned} \quad (79)$$

$f_x=0$ if $1-x^2=0$ or the exponential term is 0. $f_y=0$ if $-2y=0$ or the exponential term is 0. The exponential term is not 0 except in the degenerate case. Hence we require $1-x^2=0$ and $-2y=0$, implying $x=1$ or $x=-1$ and $y=0$. There are two critical points $(-1,0)$ and $(1,0)$.

The Second Derivative Test for Functions of Two Variables

How can we determine if the critical points found above are relative maxima or minima? We apply a second derivative test for functions of two variables.

Let (x_c, y_c) be a critical point and define

$$D(x_c, y_c) = f_{xx}(x_c, y_c)f_{yy}(x_c, y_c) - [f_{xy}(x_c, y_c)]^2 \quad (80)$$

We have the following cases:

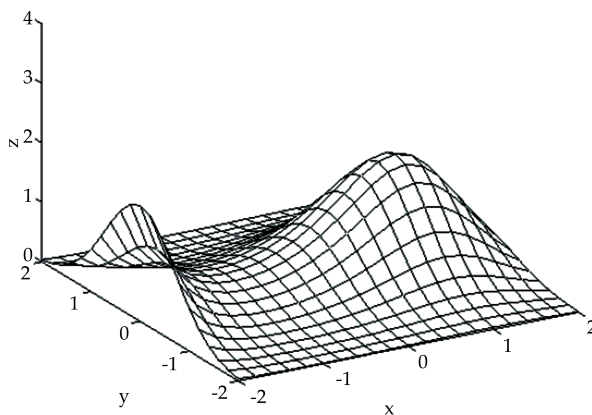
- If $D > 0$ and $f_{xx}(x_c, y_c) < 0$, then $f(x, y)$ has a relative maximum at (x_c, y_c) .
- If $D > 0$ and $f_{xx}(x_c, y_c) > 0$, then $f(x, y)$ has a relative minimum at (x_c, y_c) .
- If $D < 0$, then $f(x, y)$ has a saddle point at (x_c, y_c) .
- If $D = 0$, the second derivative test is inconclusive.

An example of a saddle point is shown below.

$$\begin{aligned} f_{xx}(x, y) &= (-2x + (1 - x^2)^2) \exp\left(-\frac{1}{3}x^3 + x - y^2\right) \\ f_{yy}(x, y) &= (-2 + 4y^2) \exp\left(-\frac{1}{3}x^3 + x - y^2\right) \\ f_{xy}(x, y) &= -2y(1 - x^2) \exp\left(-\frac{1}{3}x^3 + x - y^2\right) \end{aligned} \quad (81)$$

For $x=1$ and $y=0$, we have $D(1,0)=4\exp(4/3)>0$ with $f_{xx}(1,0)=-2\exp(2/3)<0$. Hence, $(1,0)$ is a relative maximum. For $x=-1$ and $y=0$, we have $D(-1,0)=-4\exp(-4/3)<0$. Hence, $(-1,0)$ is a saddle point

The figure show the plots of the surface $z=f(x, y)$.



Notice the relative maximum at $(x=1, y=0)$. $(x=-1, y=0)$ is a relative maximum if one travels in the y direction and a relative minimum if one travels in the x -direction. Near $(-1, 0)$ the surface looks like a saddle, hence the name.

Lagrange's Method of Multipliers

Lagrange multipliers are used in multivariable calculus to find maxima and minima of a function subject to constraints (like “find the highest elevation along the given path” or “minimize the cost of materials for a box enclosing a given volume”). It is a useful technique, but all too often it is poorly taught and poorly understood.

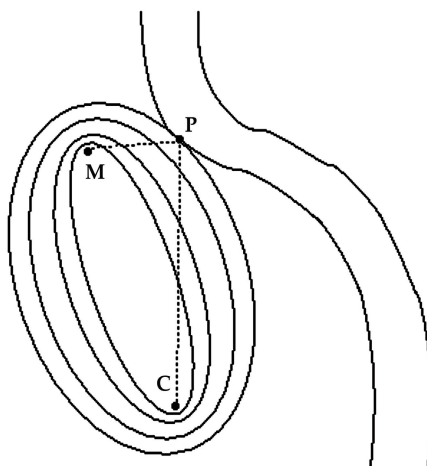
One of the most common problems in calculus is that of finding maxima or minima (in general, “extrema”) of a function, but it is often difficult to find a closed form for the function being extremized. Such difficulties often arise when one wishes to maximize or minimize a function subject to fixed outside conditions or constraints. The method of Lagrange multipliers is a powerful tool for solving this class of problems without the need to explicitly solve the conditions and use them to eliminate extra variables.

In general, Lagrange multipliers are useful when some of the variables in the simplest description of a problem are made redundant by the **constraints**.

Constraint is a condition of an optimization problem that the solution must satisfy.

The Mathematics of Lagrange Multipliers

In multivariable calculus, the gradient of a function h (written ∇h) is a normal vector to a curve (in two dimensions) or a surface (in higher dimensions) on which h is constant: $\mathbf{n} = \nabla h(\mathbf{P})$.



The length of the normal vector does not matter: any constant multiple of $\nabla h(\mathbf{P})$ is also a normal vector. In our case, we have two functions whose normal vectors are parallel, so

$$\nabla f(\mathbf{P}) = \lambda \nabla g(\mathbf{P})$$

The unknown constant multiplier λ is necessary because the magnitudes of the two gradients may be different. (Remember, all we know is that their directions are the same.)

In D dimensions, we now have $D+1$ equations in $D+1$ unknowns. D of the unknowns are the coordinates of $\mathbf{P}$ (e.g. x , y , and z for $D = 3$), and the other is the new unknown constant λ . The equation for the gradients derived above is a vector equation, so it provides D equations of constraint. Anyone can once got stuck on an exam at this point: do not let it happen to you. The original constraint equation $g(\mathbf{P}) = 0$ is the final equation in the system. Thus, in general, a unique solution exists.

As in many maximum/minimum problems, cases do exist with multiple solutions. There can even be an infinite number of solutions if the constraints are particularly degenerate: imagine if the milkmaid and the cow were both already standing right at the bank of a straight river, for example. In many cases, the actual value of the Lagrange multiplier is not interesting, but there are some situations in which it can give useful information (as discussed below).

Just set the gradient of the function you want to extremism equal to the gradient of the constraint function. You will get a vector's worth of (algebraic) equations, and together with the original constraint equation they determine the solution.

CONCEPT: DIFFERENTIAL CALCULUS

Differential calculus is a branch of calculus that deals with the concept of the derivative and its applications. It focuses on understanding how functions change as their input values (independent variable) change. Differential calculus provides tools for measuring rates of change, finding slopes of curves, and analyzing the behavior of functions locally around specific points.

Here are some key concepts and ideas in differential calculus:

- **Derivative:** The derivative of a function represents its rate of change at a particular point. It provides the instantaneous rate of change, or slope, of the function's graph at that point. The derivative of a function $f(x)$ with respect to its independent variable x is denoted as $f'(x)$ or dy/dx .
- **Differentiability:** A function is said to be differentiable at a point if its derivative exists at that point. In other words, the function's graph has a well-defined tangent line at that point.
- **Rules for Differentiation:** There are several rules and techniques for finding derivatives of various types of functions. Some important rules include the power rule, product rule, quotient rule, chain rule, and implicit differentiation.
- **Applications:** Differential calculus has numerous applications in various fields. For example, it's used to analyze motion, solve optimization problems, calculate marginal and instantaneous rates of change in economics, and study the behavior of physical quantities like velocity, acceleration, and electric fields.
- **Local Behavior of Functions:** Differential calculus enables the study of the local behavior of functions around specific points. The first derivative provides information about increasing or decreasing behavior and the concavity of a function. The second derivative helps determine points of inflection and the concavity of the graph.
- **Approximation:** Derivatives can be used to approximate the behavior of a function near a point. The tangent line at a point is often used to estimate the function's values nearby.
- **Higher Derivatives:** The second derivative is the derivative of the first derivative, and higher-order derivatives can be similarly defined. Higher-order derivatives provide information about the higher-order rates of change and concavity.

Differential calculus is fundamental to understanding the relationship between a function and its instantaneous behavior. It's the foundation for many advanced topics in calculus and is used extensively in physics, engineering, economics, biology, and other scientific disciplines to model and analyze real-world phenomena.

SUMMARY

- Differential calculus is the variation of a function with respect to changes in independent variables, especially the use of differentiation to calculate rates of change of a function and the maximum and minimum values of a function.
- Calculus approaches the paths of objects in motion as curves, or functions, and then determines the value of these functions to calculate their rate of change, area, or volume.
- The concept of continuity is an important first step in the analysis leading to differential and integral calculus. It is also an important analytical tool in its own right, with significant practical applications.
- The notation for partial derivatives is different than that for derivatives of functions of a single variable. With functions of a single variable we could denote the derivative with a single prime.
- The problem of determining the maximum or minimum of function is encountered in geometry, mechanics, physics, and other fields, and was one of the motivating factors in the development of the calculus in the seventeenth century.
- One of the most common problems in calculus is that of finding maxima or minima (in general, “extrema”) of a function, but it is often difficult to find a closed form for the function being extremized. Such difficulties often arise when one wishes to maximize or minimize a function subject to fixed outside conditions or constraints.

MULTIPLE CHOICE QUESTIONS

1. Two men on a 3-D surface want to meet each other. The surface is given by $f(x, y) = x^{-6} \cdot y^7 / (x + y)$. They make their move horizontally or vertically with the X-Y plane as their reference. It was observed that one man was initially at (200, 400) and the other at (100, 100). Their meet point is decided as (0, 0). Given that they travel in straight lines, will they meet?
 - a. They will meet
 - b. They will not meet
 - c. They meet with probability 1/2
 - d. Insufficient information
2. Velocity is a physical vector quantity; both magnitude and direction are needed to define it, also it is the measurement of the rate at which an object changes its position.
 - a. True
 - b. False
3. It is also an important analytical tool in its own right, with significant practical applications.
 - a. True
 - b. False
4. Limits and continuity sufficiently to avoid numerous common mistakes requires a care and depth of thought.
 - a. True
 - b. False
5. Polynomial is an expression consisting of variables and coefficients which only employs the operations of addition, subtraction, multiplication, and non-negative integer exponents.
 - a. True
 - b. False

Answers

1. (b) 2. (a) 3. (a) 4. (a) 5. (a)

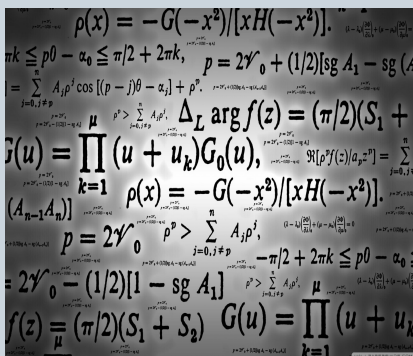
REVIEW QUESTIONS

1. What do you understand by the limits and continuity?
2. Elaborate the function of partial derivatives.
3. What are implicit functions? Explain.
4. Describe the properties of Jacobian.
5. What are Lagrange's multipliers? Explain.
6. Find all numbers x which satisfy $|x^2 - 9| = |x^2 - 5|$.

REFERENCES

1. Butcher, J., *Numerical Methods for Ordinary Differential Equations*, Second Edition, Wiley, Hoboken, NJ, 2008.
2. Enright, W. H., Continuous numerical methods for ODEs with defect control, *Journal of Computational and Applied Mathematics* 125 (2000), n. 1-2, 159–170.
3. Friedlander, S., and Pavlović, S., Blowup in a three-dimensional vector model for the Euler equations, *Comm. Pure Appl. Math.* 57 (6) (2004), 705–725.
4. Gowers, T., Barrow-Green, J., and Leader, I., Editors, *The Princeton Companion to Mathematics*, Princeton University Press, Princeton, NJ, 2008.
5. Hairer, E., Lubich, C., and Wanner, G., *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations*, Springer Series in Comput. Mathematics, Vol. 31, Springer-Verlag, 2002.

SITUATIONAL PROBLEMS INVOLVING RELATED RATES



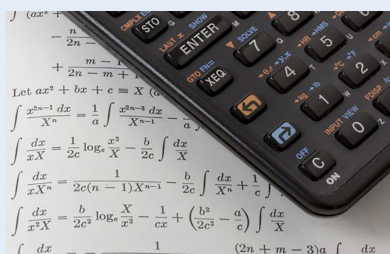
Related rates problems are applied problems where we find the rate at which one quantity is changing by relating it to other quantities whose rates are known.

After studying this chapter, you will understand the core concepts of:

- Higher-order Derivatives of Functions
- Chain Rule of Differentiation

INTRODUCTION

Calculus is primarily the mathematical study of how things change. One specific problem type is determining how the rates of two related items change at the same time. The keys to solving a related rates problem are identifying the variables that are changing and then determining a formula that connects those variables to each other. Once that is done, you find the derivative of the formula, and you can calculate the rates that you need. Related rate problems involve equations where a relationship exists between two or more derivatives. For example, you might want to find out the rate that the distance is increasing between two airplanes. Solving related rate problems has many real life applications. For example, a gas tank company might want to know the rate at which a tank is filling up, or an environmentalist might be concerned with the rate at which a certain marshland is flooding. Solving the problems usually involves knowledge of geometry and algebra in addition to calculus. Often, the “hard” part is the geometry or algebra — not the calculus, so you will want to make sure you brush up on those skills. Solving related rate problems also involves applications of the chain rule and implicit differentiation — where you differentiate both sides of the equation.



CORE CONCEPT: HIGHER-ORDER DERIVATIVES OF FUNCTIONS

A higher-order function (also functional, functional form or functor) is a function that does at least one of

the following:

- Takes one or more functions as arguments (i.e. procedural parameters),
- Returns a function as its result.

Let's start with a function $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ and only consider its second order partial derivatives. Take, for example,

$$f(x, y) = (x + y)e^y.$$

We can easily compute its two first-order partial derivatives.

$$f_x = \frac{\partial f}{\partial x} = e^y$$

$$f_y = \frac{\partial f}{\partial y} = e^y + (x + y)e^y = (1 + x + y)e^y$$

Each of these two functions, in turn, has two partial derivatives. Their partials with respect to x are

$$f_{xx} = \frac{\partial}{\partial x} \frac{\partial f}{\partial x} = \frac{\partial^2 f}{\partial x^2} = 0$$

$$f_{yx} = \frac{\partial}{\partial x} \frac{\partial f}{\partial y} = \frac{\partial^2 f}{\partial x \partial y} = e^y$$

Notice how the two different notations for partials indicate the order of taking derivatives differently. The partials with respect to y are

$$f_{xy} = \frac{\partial}{\partial y} \frac{\partial f}{\partial x} = \frac{\partial^2 f}{\partial y \partial x} = e^y$$

$$f_{yy} = \frac{\partial}{\partial y} \frac{\partial f}{\partial y} = \frac{\partial^2 f}{\partial y^2} = e^y + (1 + x + y)e^y = (2 + x + y)e^y$$

Thus, there are four second-order partial derivatives of this function f . Notice that in this example the two mixed partial derivatives are equal, that is $f_{yx} = f_{xy}$. These two mixed partials are usually equal. We would not prove the following theorem, but you will frequently use it.

Theorem 1

When the first and second partials of a function of two variables x and y are all continuous functions, the order that the partials are taken does not matter, that is, $f_{yx} = f_{xy}$, or in partial notation

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}.$$

Functions that have continuous first and second partial derivatives are called C^2 functions, or functions of class C^2 . The concept of second-order derivatives generalizes as you would expect it to for third- and higher order derivatives, and not just to functions $\mathbb{R}^2 \rightarrow \mathbb{R}$ but to scalar-valued functions $\mathbb{R}^n \rightarrow \mathbb{R}$ in general. The text has several examples, and you will be working out more in the exercises.

Preview of an application of second-order partial derivatives. You will recall from calculus of one variable that second derivatives were used in something called the second-derivative test for extreme values. You used it on critical points, that is, values of x where $f'(x)$ was 0. It said that if the second derivative was

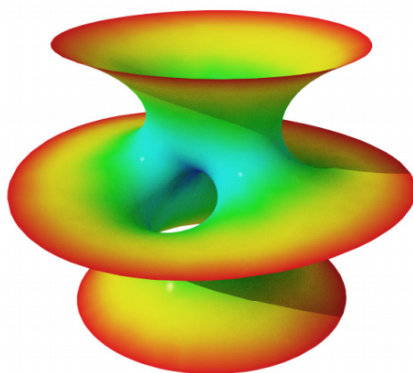
Partial Derivative of a function of several variables is its derivative with respect to one of those variables, with the others held constant (as opposed to the total derivative, in which all variables are allowed to vary).

positive at a critical point, then f had a minimum there, but if the second derivative was negative, then a maximum.

Things are more complicated in higher dimensions since there are more directions to consider. All the second order partial derivatives come into play. A particular matrix called the Hessian is filled with them, and it's used to determine the character of critical points.

Minimal Surfaces, another Application. A surface with a given boundary is called a minimal surface if among all surfaces with that boundary it has the least area. Of course, a circle bounds its interior, but that's not interesting since it's planar. You are interested in the non-planar ones. The first two were the catenoid and helicoid.

An interesting minimal surface is Costa's, illustrated in figure. Its boundary consists of three circular components.



You can make your own minimal surfaces by dipping bent clothes hangers in a soapy solution, but do not let there be any bubbles. A spherical bubble is not a minimal surface.

It turns out that if a minimal surface is the graph $z = f(x, y)$ of a function two variables, then

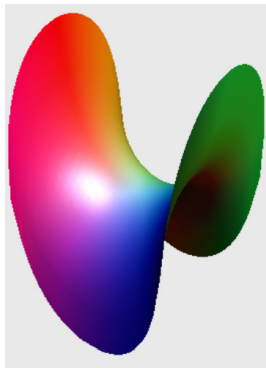
$$(1 + z_y^2)z_{xx} = (1 + z_x^2)z_{yy}$$

and conversely, functions whose derivatives satisfy that partial differential equation have graphs that are minimal surfaces.

Example 2. Heinrich Scherk constructed some embedded minimal surfaces in 1834. This surface has the equation

$$e^z \cos y = \cos x.$$

The part of the surface above the square $[-1.5, 1.5] \times [-1.5, 1.5]$ is illustrated in figure. As an exercise, verify that $(1 + z_y^2)z_{xx} = (1 + z_x^2)z_{yy}$ holds for this surface.



Higher-Order Derivatives of an Explicit Function

Let the function $y=f(x)$ have a finite derivative $f'(x)$ in a certain interval (a, b) , i.e. the derivative $f'(x)$ is also a function in this interval. If this function is differentiable, we can find the second derivative of the original function $f(x)$, which is denoted as

$$f'' = f' = \left(\frac{dy}{dx} \right)' = \frac{d}{dx} \left(\frac{dy}{dx} \right) = \frac{d^2y}{dx^2}.$$

Similarly, if f' exists and is differentiable, we can calculate the third Derivative of the function $f(x)$:

$$f''' = \frac{d^3y}{dx^3} = y''.$$

The derivatives of higher order (if they exist) are defined as

$$f^{(4)} = \frac{d^4y}{dx^4} = y^{(4)} = \left(f^{(3)} \right)', \dots, f^{(n)} = \frac{d^ny}{dx^n} = y^{(n)} = \left(f^{(n-1)} \right)'.$$

Thus, the notion of the n^{th} order derivative is introduced inductively by sequential calculation of n derivatives starting from the first order derivative. Transition to the next higher-order derivative is performed using the recurrence formula

$$y^{(n)} = \left(y^{(n-1)} \right)'.$$

In some cases, we can derive a general formula for the derivative of an arbitrary n^{th} order without computing intermediate derivatives. Some examples are considered below.

Note that the following linear relationships can be used for finding higher-order derivatives:

$$(u + v)^{(n)} = u^{(n)} + v^{(n)}, \quad (Cu)^{(n)} = Cu^{(n)}, \quad C = \text{const.}$$

Transformation is a general term for four specific ways to manipulate the shape of a point, a line, or shape. The original shape of the object is called the pre-image and the final shape and position of the object is the image under the transformation.

Higher-Order Derivatives of an Implicit Function

The n th order derivative of an implicit function can be found by sequential (n times) differentiation of the equation $F(x, y) = 0$. At each step, after appropriate substitutions and **transformations**, we can obtain an explicit expression for the derivative, which depends only on the variables x and y , i.e. the derivatives have the form

$$y' = f_1(x, y), \quad y'' = f_2(x, y), \dots, \quad y^{(n)} = f_n(x, y).$$

Higher-Order Derivatives of a Parametric Function

Consider a function $y=f(x)$ given parametrically by the equations

$$\begin{cases} x = x(t) \\ y = y(t) \end{cases}.$$

The first derivative of this function is given by

$$y' = y'_x = \frac{y'_t}{x'_t}.$$

Differentiating once more with respect to x , we find the second derivative:

$$y'' = y''_{xx} = \frac{(y'_x)'_t}{x'_t}.$$

Similarly, we define the derivatives of the third and higher order:

$$y''' = y'''_{xxx} = \frac{(y''_{xx})'_t}{x'_t}, \dots, \quad y^{(n)} = \underbrace{y^{(n)}_{xxx \dots x}}_n = \frac{\left(\underbrace{y_{xxx \dots x}}_{n-1}^{(n-1)} \right)'_t}{x'_t}.$$

Example 1:

Find y'' , if $y=x \ln x$.

Solution:

Calculate the first derivative using the product rule:

$$y' = (x \ln x)' = x' \cdot \ln x + x \cdot (\ln x)' = 1 \cdot \ln x + x \cdot \frac{1}{x} = \ln x + 1.$$

Now we can find the second derivative:

$$y'' = (\ln x + 1)' = \frac{1}{x} + 0 = \frac{1}{x}.$$

Example 2:

Find the second derivative of the function $y = \sqrt[4]{x+1}$.

Solution:

We start with first derivative:

$$y' = (\sqrt[4]{x+1})' = \left[(x+1)^{\frac{1}{4}} \right]' = \frac{1}{4} (x+1)^{-\frac{3}{4}} = \frac{1}{4\sqrt[4]{(x+1)^3}}.$$

Differentiate again to find the second derivative:

$$\begin{aligned} y'' = (y')' &= \left[\frac{1}{4} (x+1)^{-\frac{3}{4}} \right]' = \frac{1}{4} \cdot \left(-\frac{3}{4} \right) (x+1)^{-\frac{7}{4}} = -\frac{3}{16} \cdot \frac{1}{(x+1)^{\frac{7}{4}}} \\ &= -\frac{3}{16\sqrt[4]{(x+1)^7}}. \end{aligned}$$

Example 3:

Calculate y'' for the parabola equation $y^2=4x$.

Solution.

By implicit differentiation,

$$2yy'=4, \Rightarrow yy'=2.$$

Differentiating again and using the product rule, we obtain

$$y'y' + yy''=0.$$

Multiply both sides by y^2 :

$$y^2(y')^2 + y^3y'' = 0.$$

As $yy'=2$ and, hence, $(yy')^2=4$, we can write the last equation as

$$4 + y^3y''=0.$$

Then

$$y'' = -\frac{4}{y^3}.$$

Theorem 1 (Clairout's Theorem)

Suppose f is defined on a disk D that contains the point (a, b) . If the functions f_{xy} and f_{yx} are both continuous on D , the

$$f_{xy}(a, b) = f_{yx}(a, b)$$

This essentially says that for “nice” functions the mixed partial derivatives are equal, which means that the order in which we differentiate would not matter in most/all of the problems you will see here.

One example of a function whose mixed partials are different (so, it does not satisfy the hypotheses of Clairout's Theorem) is

$$f(x, y) = \begin{cases} \frac{xy(x^2 - y^2)}{x^2 + y^2}, & \text{if } (x, y) \neq (0, 0) \\ 0, & \text{if } (x, y) = (0, 0) \end{cases}$$

One can use the limit definition of the derivatives to show that $f_{xy}(0, 0) = 1$ but $f_{yx}(0, 0) = -1$

Matrix of partial derivatives

Matrix is a rectangular array of numbers, symbols, or expressions, arranged in rows and columns.

Partial derivatives of functions are usually displayed in a **matrix**, the so-called matrix of partial derivatives. If $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is a $\mathbb{R}^m$ -valued function of n variables, we arrange its partial derivatives as

$$Tf := \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & & \frac{\partial f_2}{\partial x_n} \\ \vdots & & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

We will denote this matrix by Tf . For example, if $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is real valued (only one component, $f = f_1$), then the matrix reduces to a vector

$$Tf = \left(\frac{\partial f_1}{\partial x_1}, \frac{\partial f_1}{\partial x_2}, \dots, \frac{\partial f_1}{\partial x_n} \right).$$

In this particular case, the vector of partial derivatives is called the gradient of f and it is alternatively denoted by Δf .

Example 1f

$$f(x, y) = x^2 y^3 + \sin(x + 2y) + y$$

Then its gradient is

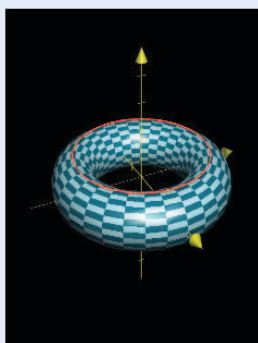
$$\nabla f = (2xy^3 + \cos(x + 2y), 3y^2x^2 + 2\cos(x + 2y) + 1)$$

and the matrix of second order partial derivatives is

$$\begin{pmatrix} 2y^3 - \sin(x + 2y) & 6xy^2 - 2\sin(x + 2y) \\ 6xy^2 - 2\sin(x + 2y) & 6yx^2 - 4\sin(x + 2y) \end{pmatrix}.$$

Observe that

$$\frac{\partial^2 f}{\partial y \partial x} = 6xy^2 - 2\sin(x + 2y) = \frac{\partial^2 f}{\partial x \partial y}.$$



CORE CONCEPT: CHAIN RULE OF DIFFERENTIATION

The chain rule can be used to derive some well-known differentiation rules. For example, the quotient rule is a consequence of the chain rule and the product rule. To see this, write the function $f(x)/g(x)$ as the product $f(x) \cdot 1/g(x)$.

The chain rule is a formula for computing the derivative of the composition of two or more functions. That is, if f and g are functions, then the chain rule expresses the derivative of their composition $f \circ g$ (the function which maps x to $f(g(x))$) in terms of the derivatives of f and g and the product of functions as follows:

$$(f \circ g)' = (f' \circ g) \cdot g'.$$

This can be written more explicitly in terms of the variable. Let $F = f \circ g$, or equivalently, $F(x) = f(g(x))$ for all x . Then one can also write

$$F'(x) = f'(g(x))g'(x).$$

The chain rule may be written, in Leibniz's notation, in the following way. If a variable z depends on the variable y , which itself depends on the variable x , so that y and z are therefore dependent variables, then z , via the intermediate variable of y , depends on x as well. The chain rule then states,

$$\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx}.$$

The two versions of the chain rule are related, if $z = f(y)$ and $y = g(x)$, then

$$\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx} = f'(y)g'(x) = f'(g(x))g'(x).$$

Solve Problems using the Chain Rule

The following problems require the use of the chain rule. The chain rule is a rule for differentiating compositions of functions. In the following discussion and solutions the derivative of a function $h(x)$ will be denoted by $D\{h(x)\}$ or $h'(x)$. Most problems are average. A few are somewhat challenging. The chain rule states formally that

$$D\{f(h(x))\} = f'(g(x))g'(x).$$

However, we rarely use this formal approach when applying the chain rule to specific problems. Instead, we invoke an intuitive approach. For example, it is sometimes easier to think of the functions f and g as "layers" of a problem. Function f is the "outer layer" and function g is the "inner layer." Thus, the chain rule tells us to first differentiate the outer layer, leaving the inner layer unchanged (the term $f'(g(x))$), then differentiate the inner layer (the term $g'(x)$). This process will become clearer as you do the problems. In most cases, final answers are given in the most simplified form.

Problem 1: Differentiate $y = (3x+1)^2$

Solution 1: The outer layer is "the square" and the inner layer is $(3x+1)$. Differentiate "the square" first, leaving $(3x+1)$ unchanged. Then differentiate $(3x+1)$. Thus,

$$\begin{aligned} D(3x+1)^2 &= 2(3x+1)^{2-1} D(3x+1) \\ &= 2(3x+1)(3) \\ &= 6(3x+1). \end{aligned}$$

Problem 2: Differentiate $y = \sin(5x)$

Solution 2: The outer layer is “the sine function” and the inner layer is $(5x)$. Differentiate “the sine function” first, leaving $(5x)$ unchanged. Then differentiate $(5x)$. Thus,

$$\begin{aligned} D\{\sin(5x)\} &= \cos(5x) D(5x) \\ &= \cos(5x)(5) \\ &= 5\cos(5x). \end{aligned}$$

Implicit Differentiation

Sometimes functions are given not in the form $y = f(x)$ but in a more complicated form in which it is difficult or impossible to express y explicitly in terms of x . Such functions are called implicit functions. We explain how these can be differentiated using implicit differentiation.

In order to master the techniques explained here it is vital that you undertake plenty of practice exercises so that they become second nature.

We look at how we might differentiate functions of y with respect to x . Consider an expression such as

$$x^2 + y^2 - 4x + 5y - 8 = 0$$

It would be quite difficult to re-arrange this so y was given explicitly as a function of x . We could perhaps, given values of x , use the expression to work out the values of y and thereby draw a graph. In general even if this is possible, it will be difficult. A **function** given in this way is said to be defined implicitly.

It will be necessary to use a rule known as the chain rule or the rule for differentiating a function of a function. In this unit we will refer to it as the chain rule. There is a separate unit which covers this particular rule thoroughly, although we will revise it briefly here.

Function is a relation between a set of inputs and a set of permissible outputs with the property that each input is related to exactly one output.

Solve Problems (including logarithmic, and inverse trigonometric functions) using Implicit Differentiation

Taking the derivatives of some complicated functions can be simplified by using logarithms. This is called logarithmic differentiation.

Example: Revision of the chain rule

Suppose we wish to differentiate $y = (5 + 2x)^{10}$ in order to calculate dy/dx .

We make a substitution and let $u = 5 + 2x$ so that $y = u^{10}$.

The chain rule states

$$\frac{dy}{dx} = \frac{dy}{du} \times \frac{du}{dx}$$

Now

$$\text{if } y = u^{10} \quad \text{then} \quad \frac{dy}{du} = 10u^9$$

and

$$\text{if } u = 5 + 2x \quad \text{then} \quad \frac{du}{dx} = 2$$

hence

$$\begin{aligned} \frac{dy}{dx} &= \frac{dy}{du} \times \frac{du}{dx} \\ &= 10u^9 \times 2 \\ &= 20u^9 \\ &= 20(5 + 2x)^9 \end{aligned}$$

So we have used the chain rule in order to differentiate the function $y = (5 + 2x)^{10}$.

In quoting the chain rule in the form $\frac{dy}{dx} = \frac{dy}{du} \times \frac{du}{dx}$ note that we write y in terms of u , and u in terms of x . i.e

$$y = y(u) \text{ and } u = u(x)$$

We will need to work with different variables. Suppose we have z in terms of y , and y in terms of x , i.e.

$$z = z(y) \text{ and } y = y(x)$$

The chain rule would then state:

$$\frac{dz}{dx} = \frac{dz}{dy} \times \frac{dy}{dx}$$

Example: Implicit differentiation

$$y^2 + x^3 - y^3 + 6 = 3y$$

with respect x .

We differentiate each term with respect to x :

$$\frac{d}{dx}(y^2) + \frac{d}{dx}(x^3) - \frac{d}{dx}(y^3) + \frac{d}{dx}(6) = \frac{d}{dx}(3y)$$

Differentiating functions of x with respect to x is straightforward. But when differentiating a function of y with respect to x we must remember the rule given in the previous key-point. We find

$$\frac{d}{dy}(y^2) \times \frac{dy}{dx} + 3x^2 - \frac{d}{dy}(y^3) \times \frac{dy}{dx} + 0 = \frac{d}{dy}(3y) \times \frac{dy}{dx}$$

That is

$$2y \frac{dy}{dx} + 3x^2 - 3y^2 \frac{dy}{dx} = 3 \frac{dy}{dx}$$

We rearrange this to collect all terms involving dy/dx together.

$$3x^2 = 3 \frac{dy}{dx} - 2y \frac{dy}{dx} + 3y^2 \frac{dy}{dx}$$

then

$$3x^2 = (3 - 2y + 3y^2) \frac{dy}{dx}$$

so that, finally,

$$\frac{dy}{dx} = \frac{3x^2}{3 - 2y + 3y^2}$$

This is our expression for dy/dx .

Example

Suppose we want to differentiate, with respect to x , the implicit function

$$\sin y + x^2 y^3 - \cos x = 2y$$

As before, we differentiate each term with respect to x .

$$\frac{d}{dx}(\sin y) + \frac{d}{dx}(x^2 y^3) - \frac{d}{dx}(\cos x) = \frac{d}{dx}(2y)$$

Recognize that the second term is a product and we will need the product rule. We will also use the chain rule to differentiate the functions of y . We find

$$\frac{d}{dy}(\sin y) \times \frac{dy}{dx} + \left\{ x^2 \frac{d}{dx}(y^3) + y^3 \frac{d}{dx}(x^2) \right\} + \sin x = \frac{d}{dy}(2y) \times \frac{dy}{dx}$$

So that

$$\cos y \frac{dy}{dx} + \left\{ x^2 \cdot \frac{d}{dy} (y^3) \frac{dy}{dx} + y^3 \cdot 2x \right\} + \sin x = 2 \frac{dy}{dx}$$

Tidying this up gives

$$\cos y \frac{dy}{dx} + x^2 3y^2 \frac{dy}{dx} + 2xy^3 + \sin x = 2 \frac{dy}{dx}$$

We now start to collect together terms involving dy/dx .

$$2xy^3 + \sin x = 2 \frac{dy}{dx} - \cos y \frac{dy}{dx} - 3x^2 y^2 \frac{dy}{dx}$$

$$2xy^3 + \sin x = (2 - \cos y - 3x^2 y^2) \frac{dy}{dx}$$

so that, finally

$$\frac{dy}{dx} = \frac{2xy^3 + \sin x}{2 - \cos y - 3x^2 y^2}$$

We have deliberately included plenty of detail in this calculation. With practice you will be able to omit many of the intermediate stages. The following two examples show how you should aim to condense the solution.

Derivatives of Inverse Trigonometric Functions

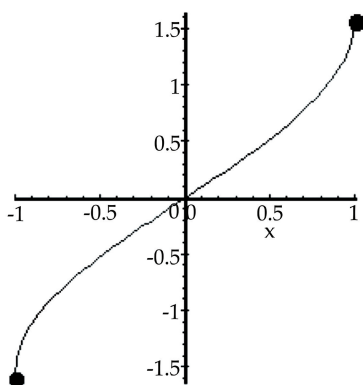
Inverse function
is a function that “reverses” another function: if the function f applied to an input x gives a result of y , then applying its inverse function g to y gives the result x , and vice versa. i.e., $f(x) = y$ if and only if $g(y) = x$.

Differentiation of inverse trigonometric functions is a small and specialized topic. However, these particular derivatives are interesting to us for two reasons. First, computation of these derivatives provides a good workout in the use of the chain rule, the definition of **inverse functions**, and some basic trigonometry. Second, it turns out that the derivatives of the inverse trigonometric functions are actually algebraic functions!! This is an unexpected and interesting connection between two seeming lay very different classes of functions.

It is possible to form inverse functions for restricted versions of all six basic trigonometric functions. One can construct and use an inverse cosecant function, for example. However, it is generally enough to consider the inverse sine and the inverse tangent functions. We will restrict our attention to these two functions.

$\frac{d}{dx} \arcsin(x) = \frac{1}{\sqrt{1-x^2}}$	$\frac{d}{dx} \arctan(x) = \frac{1}{x^2+1}$
--	---

The arctangent function is differentiable on the entire real line. The arcsine function is differentiable only on the open interval $(-1, 1)$. Even though both 1 and -1 are in the domain of the arcsine function, the arcsine is not differentiable at these points. From an algebraic point of view, we see that neither 1 nor -1 can be plugged into the derivative formula for arcsine...there's that thing about never dividing by zero again!! From a graphical point of view, the graph of the arcsine function has vertical tangent lines at the endpoints $(1, \pi/2)$ and $(-1, -\pi/2)$.



$$y = \arcsin(x)$$

$$-1 \leq x \leq 1$$

In order to verify the differentiation formula for the arcsine function, let us set

$$y = \arcsin(x).$$

We want to compute dy/dx . The first step is to use the fact that the arcsine function is the inverse of the sine function. Among other things, this means that

$$\sin(y) = \sin(\arcsin(x)) = x.$$

Next, differentiate both ends of this formula. We apply the chain rule to the left end, remembering that the derivative of the sine function is the cosine function and that y is a differentiable function of x .

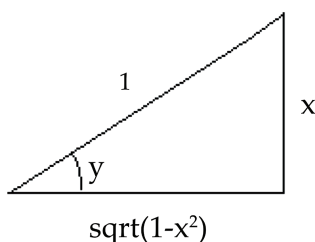
$\frac{d}{dx} \sin(y) = \frac{dy}{dx} \cos(y)$	$\frac{d}{dx} x = 1$
--	----------------------

The next step is to solve for dy/dx . (After all, this is the thing that we want to compute!)

$$\frac{dy}{dx} = \frac{1}{\cos(y)} = \sec(y)$$

This looks like progress, but it is not the answer. Remember, when we differentiate a function of x in terms of x (this is the meaning of the dx in d/dx), we must express our answer in terms of x . Therefore the question remains.

If $\sin(y)=x$, what is $\cos(y)$ in terms of x ?



$$\sin(y) = x$$

The key is to construct a “reference triangle” to record the relationship between x and y . This is a right triangle with the angle y (measured in radians) as one of the acute angles. The trigonometric functions of y can then be expressed as ratios of side lengths in this triangle. Note that we built our triangle in such a way that the side opposite to y has length x and so that the hypotenuse has length 1. This provides that

The length of the third side of the reference triangle is determined by the Pythagorean Theorem.

$$1^2 = x^2 + (\text{third side})^2$$

$$(\text{third side})^2 = 1 - x^2$$

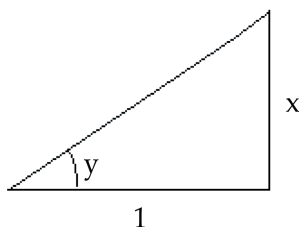
$$\text{third side} = \sqrt{1 - x^2}$$

The last step is to express the trigonometric functions of y in terms of ratios of side lengths in the reference triangle. In particular, the secant of y is equal to the hypotenuse length divided by the adjacent side length.

$$\frac{d}{dx} \arcsin(x) = \frac{dy}{dx} = \sec(y) = \frac{1}{\sqrt{1-x^2}}$$

Verification of the derivative formula for the arctangent function is left as an exercise.....an exercise that is highly recommended!! The procedure is the same as the

one that we used above. Begin by setting $y = \arctan(x)$ so that $\tan(y) = x$. Differentiating both sides of this equation and applying the chain rule, one can solve for dy/dx in terms of y . One wants to compute dy/dx in terms of x . A reference triangle is constructed as shown, and this can be used to complete the expression of the derivative of $\arctan(x)$ in terms of x .



$$\tan(y) = x$$

This completes our study of differentiation for now. We will investigate another general differentiation technique called implicit differentiation. Later still, we will learn how to differentiate exponential and logarithmic functions. For now, you should go to the Practice area and spend some time learning to use the many differentiation techniques that have been introduced. If there is one skill that we must develop for success in differential calculus.

The chain rule provides us a technique for finding the derivative of composite functions, with the number of functions that make up the composition determining how many differentiation steps are necessary.

If you find that you are having difficulty with differentiation, do not worry. You are not the first person to struggle with this technical skill. Contact your classmates. Discuss your difficulties. Contact your instructor.

We now turn to the exponential and logarithmic functions. We have already discussed “the” exponential function: $\exp(x) = e^x$. In order to differentiate the “other” exponential functions and the logarithmic functions, we must first compute the derivative of the inverse to the exponential function. Thus we turn our attention to the natural logarithm.

CONCEPTS OF INVERSE TRIGONOMETRIC FUNCTIONS

Inverse trigonometric functions, as a topic of learning, are closely related to the basic trigonometric functions. The domain and the range of the trigonometric functions are converted to the range and domain of the inverse trigonometric functions. In trigonometry, we learn about the relationships between angles and sides in a right-angled triangle. Similarly, we have inverse trigonometry functions. The basic trigonometric functions are $\sin$, $\cos$, $\tan$, cosec , $\sec$, and $\cot$. The inverse trigonometric functions on the other hand are denoted as $\sin^{-1}x$, $\cos^{-1}x$, $\cot^{-1}x$, $\tan^{-1}x$, $\operatorname{cosec}^{-1}x$, and $\sec^{-1}x$.

Inverse trigonometric functions have all the formulas of the basic trigonometric functions, which include the sum of functions, double and triple of a function. Here we shall try to understand the transformation of the trigonometric formulas to inverse trigonometric formulas.

SUMMARY

- Calculus is primarily the mathematical study of how things change. One specific problem type is determining how the rates of two related items change at the same time.
- Related rate problems involve equations where a relationship exists between two or more derivatives.
- Sometimes functions are given not in the form $y = f(x)$ but in a more complicated form in which it is difficult or impossible to express y explicitly in terms of x . Such functions are called implicit functions.
- Differentiation of inverse trigonometric functions is a small and specialized topic. However, these particular derivatives are interesting to us for two reasons.
- Later still, we will learn how to differentiate exponential and logarithmic functions. For now, you should go to the Practice area and spend some time learning to use the many differentiation techniques that have been introduced. If there is one skill that we must develop for success in differential calculus.

MULTIPLE CHOICE QUESTIONS

1. If $x > 0$, then which one is correct?
 - a. $x > \log(x + 1)$
 - b. $x < \log(x + 1)$
 - c. $x = \log(x + 1)$
 - d. $x \geq \log(x + 1)$
2. If $x > 0$, then which one is correct?
 - a. $\log(1 + x) > x - (x^2/2)$
 - b. $\log(1 + x) < x - (x^2/2)$
 - c. $\log(1 + x) \geq x - (x^2/2)$
 - d. $\log(1 + x) \leq x - (x^2/2)$
3. What will be the range of the function $f(x) = 2x^3 - 9x^2 - 24x + 5$ which increases with x ?
 - a. $x > 4$
 - b. $x > 4$ or $x < -1$
 - c. $x < -1$
 - d. Can't be determined
4. Calculus is primarily the mathematical study of how things change.
 - a. True
 - b. False
5. Related rate problems involve equations where a relationship exists between two or more derivatives.
 - a. True
 - b. False

Answers

1. (a) 2. (a) 3. (b) 4. (a) 5. (a)

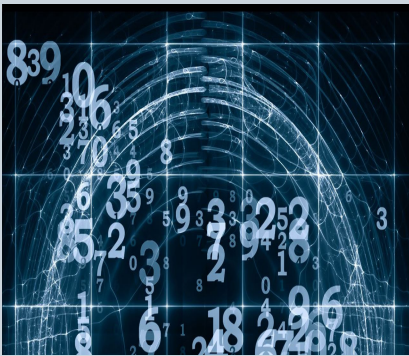
REVIEW QUESTIONS

1. Discuss about higher-order derivatives of an implicit function.
2. Explain the higher-order derivatives of a parametric function.
3. Describe the matrix of partial derivatives.
4. How to solve problems using the chain rule? Explain.
5. What do you mean by implicit differentiation?

REFERENCES

1. A. Abdesselam, the Grassmann-Berezin calculus and theorems of the matrix-tree type, *Adv. in Appl. Math.* 33 (2004), 51–70.
2. A. Bedini, S. Caracciolo, and A. Sportiello, Hyperforests on the complete hypergraph by Grassmann integral representation, *J. Phys. A* 41 (2008), 205003.
3. C.-H. Chang and C.-W. Ha, Central factorial numbers and values of Bernoulli and Euler polynomials at rationales, *Numer. Funct. Anal. Optim.* 30 (2009), 214–226.
4. D. Cvijović, Closed-form formulae for the derivatives of trigonometric functions at rational multiples of π , *Appl. Math. Lett.* 22 (2009), 906–909.
5. D. Cvijović, Derivative polynomials and closed-form higher derivative formulae, *Appl. Math. Comput.* 215 (2009), 3002–3006.
6. D. Cvijović, Values of the derivatives of the cotangent at rational multiples of π , *Appl. Math. Lett.* 22 (2009), 217–220.
7. K. N. Boyadzhiev, Derivative polynomials for $\tanh$, $\tan$, sech and $\sec$ in explicit form, *Fibonacci Quart.* 45 (2007), 291–303.
8. S. Caracciolo, A. D. Sokal, and A. Sportiello, Algebraic/combinatorial proofs of Cayleytype identities for derivatives of determinants and pfaffians, *Adv. in Appl. Math.* 50 (2013), 474–594.
9. V. S. Adamchik, On the Hurwitz function for rational arguments, *Appl. Math. Comput.* 187 (2007), 3–12.

CIRCULAR FUNCTIONS



The functions describing the horizontal and vertical positions of a point on a circle as a function of angle (cosine and sine). The circular functions may be thought of as a way to extend these extremely useful mathematical relationships to any triangle. To visualize circular functions, we first start with a unit circle, or a circle with a radius equal to one unit of measurement.

After studying this chapter, you will understand the core concepts of:

- Theory of Circular Functions
- Domain and Range of a Function

INTRODUCTION

Circular functions are defined such that their domains are sets of numbers that correspond to the measures (in radian units) of the angles of analogous trigonometric functions. The ranges of these circular functions, like their analogous trigonometric functions, are sets of real numbers. These functions are called circular functions because radian measures of angles are determined by the lengths of arcs of circles. In particular, trigonometric functions defined using the unit circle lead directly to these circular functions.

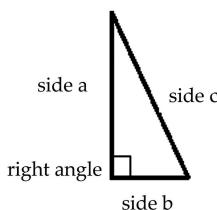
Circular functions are part of the set of trigonometric functions. Whereas trigonometric functions consist of domains that are sets of angles and ranges that are real numbers, circular functions have domains that are sets of numbers that correspond to the angles of the trigonometric functions (in radians). The graph of the equation $x^2 + y^2 = 1$ is a circle in the rectangular coordinate system. This graph is called the unit circle and has its center at the origin and has a radius of 1 unit. Trigonometric functions are defined so that their domains are sets of angles and their ranges are sets of real numbers.

You may have initially about the sine, cosine and tangent of an angle as the ratio of the sides of a right triangle. The sides that form the right angle are called legs, and the third side is called the hypotenuse. Given angle A, the three ratios are as follows:

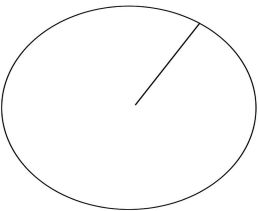
- $\sin A = (\text{length of side opposite angle A}) / (\text{length of hypotenuse})$
- $\cos A = (\text{length of side adjacent to angle A}) / (\text{length of hypotenuse})$
- $\tan A = (\text{length of side opposite angle A}) / (\text{length of side adjacent to angle A}) = \sin A / \cos A$

The circular functions may be thought of as a way to extend these extremely useful mathematical relationships to any triangle. To visualize circular functions, we first start with a unit circle, or a circle with a radius equal to one unit of measurement.

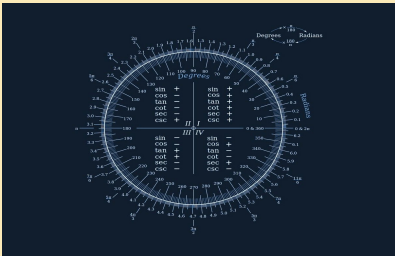
In this circle, draw an x-y coordinate plane, with the origin at the center of the circle.



Now, insert a right triangle in Quadrant I, so that the hypotenuse equals the radius of the circle, one leg of the right angle lies horizontally along the x-axis and the other leg is a vertical line that meets the hypotenuse lying on the circle. The angle will be formed by some length of the x-axis on one side and the hypotenuse on the other. The point of the circle where the hypotenuse and the vertical leg intersect is a rectangular coordinate point (x,y).



Unit circle
Radius equal to 1



CORE CONCEPT: THEORY OF CIRCULAR FUNCTIONS

Circular functions are commonly used in geological and structural engineering studies to create mathematical models of rock folds. This set of functions is considered “circular” because the

functions describe a point on a circle as the function of an angle, which is defined by radians.

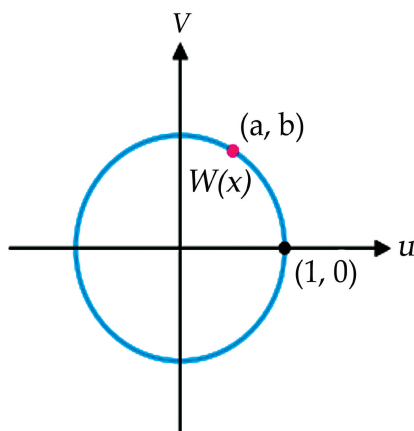
We saw that the wrapping function W pairs each real number x with an ordered pair of real numbers (a, b) , the coordinates of the circular point $W(x)$. We use this association to construct the six circular functions, also called circular functions: sine, cosine, tangent, cotangent, secant, and cosecant. The values of these functions for a real number x are denoted by $\sin x$, $\cos x$, $\tan x$, $\cot x$, $\sec x$, and $\csc x$, respectively.

These values are expressed in terms of the coordinates of the circular point $W(x)$ (a, b) as indicated

If x is a real number and (a, b) are the coordinates of the circular point $W(x)$, then

$$\begin{aligned} \sin x &= b & \csc x &= \frac{1}{b} \quad b \neq 0 \\ \cos x &= a & \sec x &= \frac{1}{a} \quad a \neq 0 \\ \tan x &= \frac{b}{a} \quad a \neq 0 & \cot x &= \frac{a}{b} \quad b \neq 0 \end{aligned}$$

The domain of both the sine and cosine functions is the set of real numbers $\mathbb{R}$. The range of both the sine and cosine functions is $[-1, 1]$. This is the set of numbers assumed by b , for sine, and a , for cosine, as the circular point (a, b) moves around the unit circle. The domain of cosecant is the set of real numbers x such that b in $W(x)$ (a, b) is not 0. Similar restrictions are made on the domains of the other three circular functions. We will have more to say about the domains and ranges of all six circular functions in subsequent sections.



Exact Values for Particular Real Numbers

We can evaluate any one of the six circular functions exactly, when it exists, for integer multiples of the real numbers $\pi/6$, $\pi/4$, $\pi/3$, and $\pi/2$. Which you should have memorized, and symmetry properties of the unit circle are central to the process.

We will show how a calculator can be used to evaluate the circular functions to 8 or more significant digits for arbitrary real numbers. You might ask why we don't go directly to the calculator. The answer is that there are many situations in which it is more desirable to work with exact forms, if available, than the corresponding decimal approximations that are produced by a calculator.

The real numbers include all the rational numbers, such as the integer -5 and the fraction $4/3$, and all the irrational numbers, such as $\sqrt{2}$ (1.41421356..., the square root of 2, an irrational algebraic number).

Example: Exact Evaluation of Circular Functions

Evaluate each circular function exactly for $x = \pi/3$.

Solution

$$W\left(\frac{\pi}{3}\right) = \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right)$$

Thus,

$$\sin \frac{\pi}{3} = b = \frac{\sqrt{3}}{2}$$

$$\csc \frac{\pi}{3} = \frac{1}{b} = \frac{1}{\sqrt{3}/2} = \frac{2}{\sqrt{3}}$$

$$\cos \frac{\pi}{3} = a = \frac{1}{2}$$

$$\sec \frac{\pi}{3} = \frac{1}{a} = \frac{1}{1/2} = 2$$

$$\tan \frac{\pi}{3} = \frac{b}{a} = \frac{\sqrt{3}/2}{1/2} = \sqrt{3}$$

$$\cot \frac{\pi}{3} = \frac{a}{b} = \frac{1/2}{\sqrt{3}/2} = \frac{1}{\sqrt{3}}$$

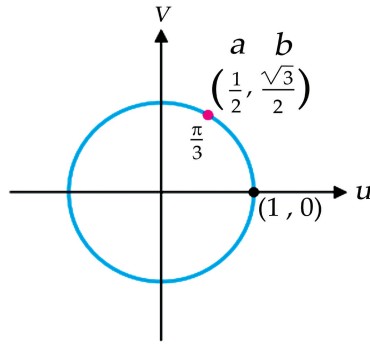


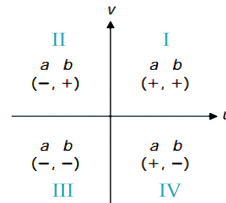
Figure 1: Exact Evaluation of Circular Functions.

Sign Properties

As a circular point $W(x)$ moves from quadrant to quadrant, its coordinates (a, b) undergo sign changes. Hence, the circular functions also undergo sign changes. It is important to know the sign of each circular function in each quadrant. Table 1 shows the sign behavior for each function. It is not necessary to memorize Table 1, since the sign of each function for each quadrant is easily determined from its definition (which should be memorized).

Table 1: Sign Properties

Circular function	Sign in quadrant			
	I	II	III	IV
$\sin x = b$	+	+	-	-
$\csc x = 1/b$	+	+	-	-
$\cos x = a$	+	-	-	+
$\sec x = 1/a$	+	-	-	+
$\tan x = b/a$	+	-	+	-
$\cot x = a/b$	+	-	+	-



Basic Identities

Returning to the definitions of the circular functions and noting that

$$\sin x = b \quad \text{and} \quad \cos x = a$$

We can obtain the following useful relationships among the six circular functions:

$$\csc x = \frac{1}{b} = \frac{1}{\sin x} \quad (1)$$

$$\sec x = \frac{1}{a} = \frac{1}{\cos x} \quad (2)$$

$$\cot x = \frac{a}{b} = \frac{1}{b/a} = \frac{1}{\tan x} \quad (3)$$

$$\tan x = \frac{b}{a} = \frac{\sin x}{\cos x} \quad (4)$$

$$\cot x = \frac{a}{b} = \frac{\cos x}{\sin x} \quad (5)$$

Because the circular points $W(x)$ and $W(-x)$ are symmetric with respect to the horizontal axis (Fig. 2), we have the following sign properties:

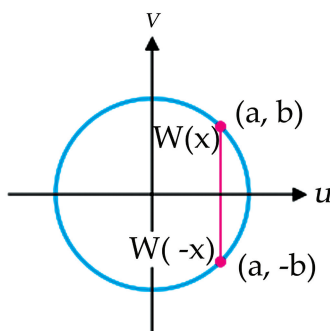


Figure 2: Symmetry property.

$$\sin(-x) = -b = -\sin x \quad (6)$$

$$\cos(-x) = a = \cos x \quad (7)$$

$$\tan(-x) = \frac{-b}{a} = -\frac{b}{a} = -\tan x \quad (8)$$

Finally, because (a, b) $(\cos x, \sin x)$ is on the unit circle $u^2 + v^2 = 1$, it follows that $(\cos x)^2 + (\sin x)^2 = 1$

which is usually written as

$$\sin^2 x + \cos^2 x = 1 \quad (9)$$

where $\sin^2 x$ and $\cos^2 x$ are concise ways of writing $(\sin x)^2$ and $(\cos x)^2$, respectively.

Calculator Evaluation

Evaluating circular functions for real numbers other than integer multiples of $\pi/6$, $\pi/4$, $\pi/3$, and $\pi/2$ is difficult without the use of a calculator. Using advanced mathematics, calculators are internally programmed to evaluate these functions automatically to an accuracy of 8 or more significant digits

If you look at the function keys on your calculator, you will find three keys labeled

SIN
COS
TAN

These keys are used to evaluate the sine, cosine, and tangent functions directly. A careful look at the function keys on your calculator also will reveal that there are no keys for cosecant, secant, and cotangent. Why is it not necessary to have these additional keys? Because of the reciprocal identities (1)–(3), we can use the function keys for sine, cosine, and tangent, along with the reciprocal function key

$$\boxed{x^{-1}} \quad \text{or} \quad \boxed{1/x}$$

to obtain $\csc x$, $\sec x$, and $\cot x$. Don't use the keys marked $\sin 1$, $\cos 1$, and $\tan 1$ to evaluate $\csc$, $\sec$, and $\cot$, respectively.

Example: Calculator Evaluation

Evaluate to 4 significant digits using a calculator:

(A) $\sin 2$ (B) $\tan (-1.612)$ (C) $\csc 3.2$

Solutions

(A) $\sin 2 = 0.9093$

(B) $\tan (-1.612) = -1.612$

(C) $\csc 3.2 = -1.132$

Angles and Their Measure

An angle is a measurement of a rotation or change of direction. In plane geometry, an angle is the measurement of the rotation required to move one ray onto another ray emanating from the same point. A ray is a 'half-line' beginning at some point P and extending indefinitely in one direction. By convention, counterclockwise rotations have positive measurements and clockwise rotations have negative measurements. An exception is that in plane geometry, sometimes all angles are considered to be positive, regardless of the direction of rotation.

When measuring rotations, various unit rotations may be used. A unit is a reference quantity with respect to which all other quantities may be compared in order to find their measure.

One unit of rotation is the revolution. If an object makes one revolution, it turns one time about some axis to return to its original orientation. If it turns one-half revolution, it will end up oriented opposite to its original orientation. When one says that a wheel is turning at a rate of 300 rpm, one is saying that the wheel turns through an angle of 300 revolutions per minute.

Number system is a way to represent numbers. We are used to using the base-10 number system, which is also called decimal. Other common number systems include base-16 (hexadecimal), base-8 (octal), and base-2 (binary).

Another traditional unit of rotation is the degree. The degree is one three hundred sixtieth of one revolution. A degree may be divided into 60 smaller units called minutes and the minute may be subdivided further into 60 smaller units called seconds. We can thank the ancient Babylonians for the degree measure of angles as well as the minutes and seconds we use to measure time. Whereas we use a base 10 **number system**, the ancient Babylonians used base 60. So when dividing up the circle they first divided it up into six equal parts using a common geometric construction, then subdivided that one-sixth part of a circle into 60 smaller parts to get the degree unit for measuring angles.

An angle is formed by rotating (in a plane) a ray m , called the initial side of the angle, around its endpoint until it coincides with a ray n , called the terminal side of the angle. The common endpoint V of m and n is called the vertex (see Fig 3). A counterclockwise rotation produces a positive angle, and a clockwise rotation produces a negative angle, as shown in Figure 4(a) and (b). The amount of rotation in either direction is not restricted.

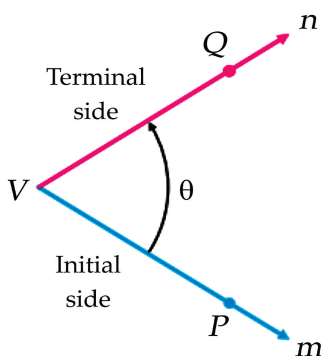


Figure 3: Angle θ or angle PVQ or $\angle V$.

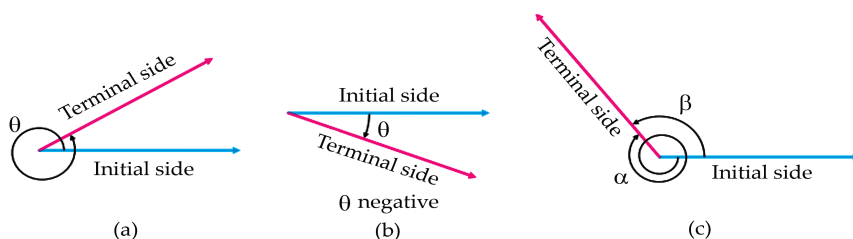


Figure 4: Angles and rotation.

Two different angles may have the same initial and terminal sides, as shown in Figure 4(c). Such angles are said to be coterminal.

An angle in a rectangular coordinate system is said to be in standard position if its vertex is at the origin and the initial side is along the positive x axis. If the terminal side of an angle in standard position lies along a coordinate axis, the angle is said to be a quadrantal angle. If the terminal side does not lie along a coordinate axis, then the angle is often referred to in terms of the quadrant in which the terminal side lies (Fig 5).

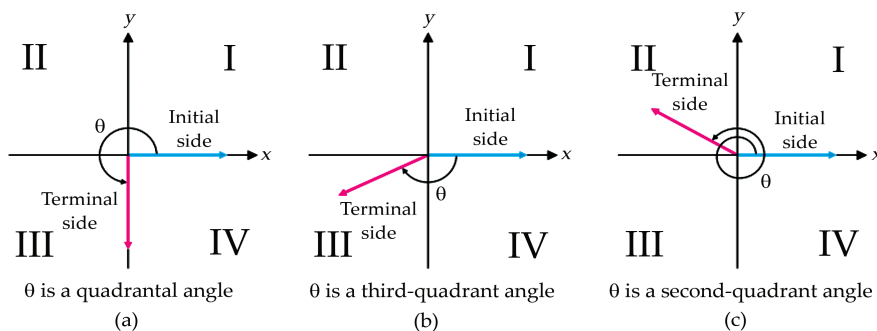


Figure 5: Angles in standard positions.

Degree and Radian Measure

Just as line segments are measured in centimeters, meters, inches, or miles, angles are measured in different units. The two most commonly used units for angle measure are degree and radian.

An angle formed by one complete rotation is said to have a measure of 360 degrees (360°). An angle formed by of a complete rotation is said to have a measure of 1 degree (1°). The symbol $^\circ$ denotes degrees.

Certain angles have special names. Figure 6 shows a straight angle, a right angle, an acute angle, and an obtuse angle.

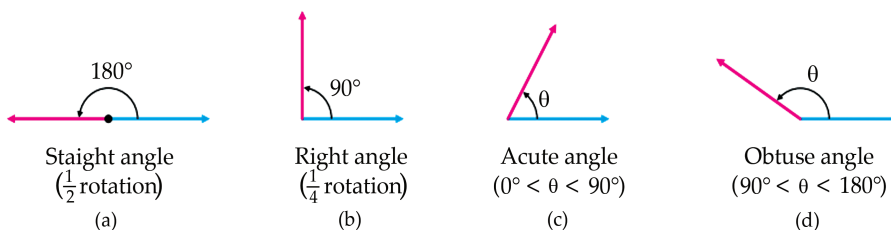


Figure 6: Types of angles.

Two positive angles are complementary if their sum is 90° ; they are supplementary if their sum is 180° . A degree can be divided further using decimal notation. For example, 42.75° represents an angle of degree measure 42 plus three-quarters of 1 degree. A degree can also be divided further using minutes and seconds just as an hour is divided into minutes and seconds. Each degree is divided into 60 equal parts called minutes, and each minute is divided into 60 equal parts called seconds.

Symbolically, minutes are represented by and seconds by Thus,

$12^\circ 23' 14''$

is a concise way of writing 12 degrees, 23 minutes, and 14 seconds. **Decimal degrees (DD)** are useful in some instances and degrees–minutes–seconds (DMS) are useful in others. You should be able to go from one form to the other, as demonstrated in Example.

Decimal degrees (DD)

express latitude and longitude geographic coordinates as decimal fractions and are used in many geographic information systems (GIS), web mapping applications such as OpenStreetMap, and GPS devices.

Example: From DMS to DD and Back

(A) Convert $21^{\circ}47'12''$ to decimal degrees.

(B) Convert 105.183° to degree-minute-second form.

Solution

$$(A) \ 21^{\circ}47'12'' = \left(21 + \frac{47}{60} + \frac{12}{3600}\right)^{\circ} = 21.787^{\circ}$$

$$\begin{aligned} (B) \ 105.183^{\circ} &= 105^{\circ}(0.183 \cdot 60)' \\ &= 105^{\circ}10.98' \\ &= 105^{\circ}10'(0.98 \cdot 60)'' \\ &= 105^{\circ}10'59'' \end{aligned}$$

Degree measure of angles is used extensively in engineering, surveying, and navigation. Another unit of angle measure, called the radian, is better suited for certain mathematical developments, scientific work, and engineering applications.

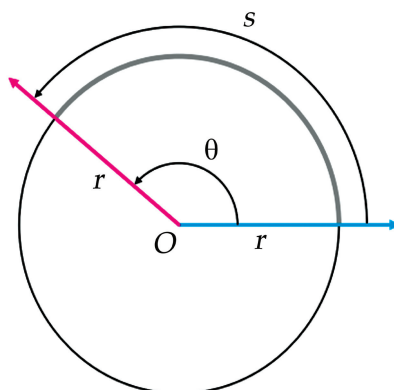
Radian Measure

If the vertex of an angle q is placed at the center of a circle with radius $r > 0$, and the length of the arc opposite q on the circumference is s , then the radian measure of q is given by

$$\theta = \frac{s}{r} \text{ radians}$$

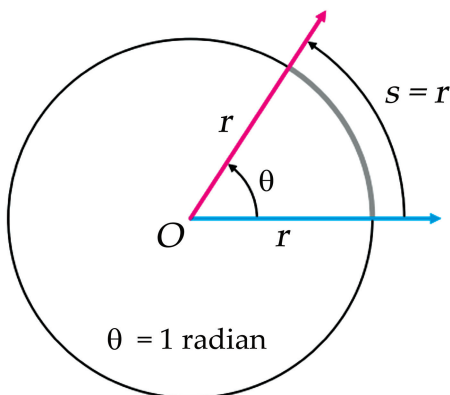
Also,

$$s = r\theta$$



If $s = r$, then

$$\theta = \frac{r}{r} = 1 \text{ radian}$$



Thus, one radian is the size of the central angle of a circle that intercepts an arc the same length as the radius of the circle. [Note: s and r must be measured in the same units. Also note that θ is being used in two ways: as the name of an angle and as the measure of the angle. The context determines the choice. Thus, when we write $\theta = s/r$, we mean the **radian measure** of angle θ is s/r .]

Example: Computing Radian Measure

What is the radian measure of a central angle opposite an arc of 24 meters in a circle of radius 6 meters?

Solution

$$\theta = \frac{s}{r} = \frac{24 \text{ meters}}{6 \text{ meters}} = 4 \text{ radians}$$

From Degrees to Radians, and Vice Versa

What is the radian measure of an angle of 180° ? A central angle of 180° is subtended by an arc one-half the circumference of a circle. Thus, if C is the circumference of a circle, then one-half the circumference is given by

$$s = \frac{C}{2} = \frac{2\pi r}{2} = \pi r \quad \text{and} \quad \theta = \frac{s}{r} = \frac{\pi r}{r} = \pi \text{ rad}$$

Hence, 180° corresponds to π rad. This is important to remember, since the radian measures of many special angles can be obtained from this correspondence. For example, 90° is

Radian measure of an angle drawn in standard position in the plane is equal to the length of arc on the unit circle subtended by that angle.

$180^\circ/2$; therefore, 90° corresponds to $\rho/2$ rad.

Table 2: From Degrees to Radians, and Vice Versa

Radians	$\pi/6$	$\pi/4$	$\pi/3$	$\pi/2$	π	$3\pi/2$	2π
Degrees				90	180		

The results in Table 2 are summarized in Figure 7 for easy reference. These correspondences and multiples of them will be used extensively in work that follows.

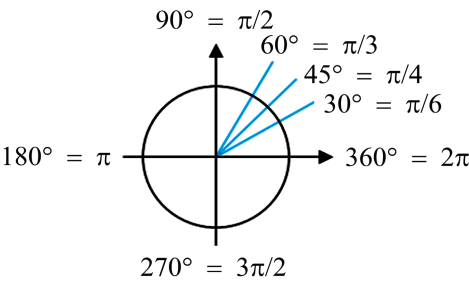


Figure 7: Radian–degree correspondences.

In general, the following proportion can be used to convert degree measure to radian measure, and vice versa.

Standard Position, Reference Angles and Coterminal Angles

The unit circle showed the placement of degrees and radians in the coordinate plane. Angles in the unit circle start on the x-axis and are measured counterclockwise about the origin. While these unit circle concepts are still in play, we will now not be “drawing” the unit circle in each diagram. The diagrams will now allow us to work with radii exceeding the unit one (as seen in the unit circle).

Standard Position

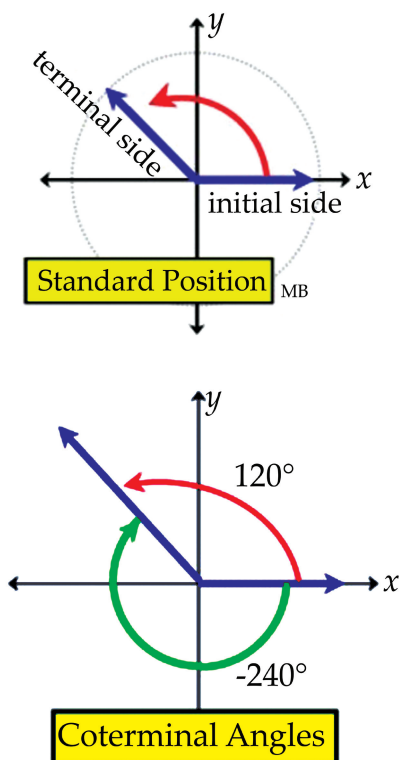
An angle is in standard position if its vertex is located at the origin and one ray is on the positive x-axis. The ray on the x-axis is called the initial side and the other ray is called the terminal side. If the terminal side of an angle lies “on” the axes (such as 0° , 90° , 180° , 270° , 360°), it is called a quadrant angle. The angle shown at the right is referred to as a Quadrant II angle since its terminal side lies in Quadrant II.

The angle is measured by the amount of rotation from the initial side to the terminal side. If measured in a counterclockwise direction the measurement is positive. If measured in a clockwise direction the measurement is negative. (A

Coterminal Angles are angles in standard position that have the same Initial Side and the same Terminal side. For example 45° , 405° and -315° are coterminal angles because all three angles have the same initial side (the x axis) and they share a same terminal side.

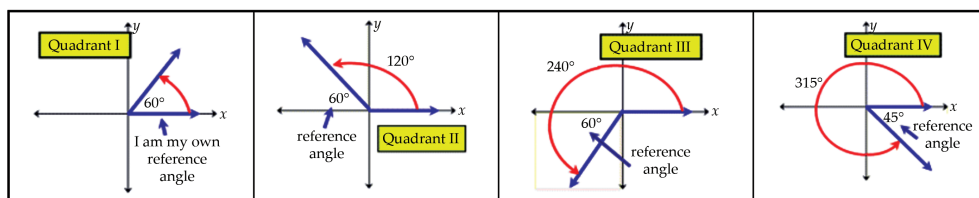
negative associated with an angle's measure refers to its “direction” of measurement, clockwise.)

If two angles in standard position have the same terminal side, they are called coterminal angles. In the diagram at the left, $+120^\circ$ and -240° are coterminal angles.



Reference Angles

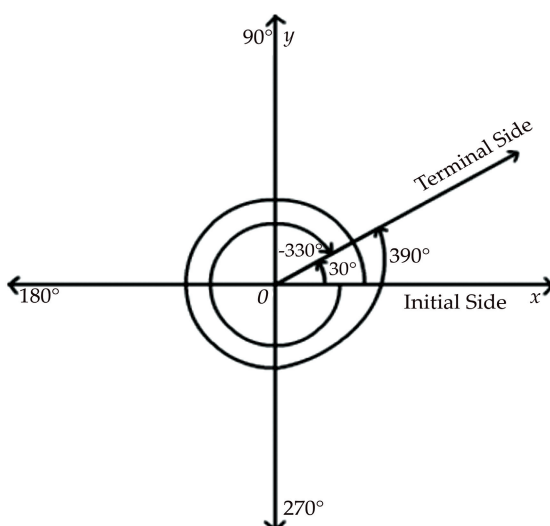
Associated with every angle drawn in standard position (except quadrantal angles) there is another angle called the reference angle. The reference angle is the acute angle (the smallest angle) formed by the terminal side of the given angle and the x-axis. Reference angles may appear in all four quadrants. Angles in quadrant I are their own reference angles.



Coterminal Angles

Coterminal angles are angles in standard position (angles with the initial side on the positive x-axis) that have a common terminal side. For example 30° , -330° and 390°

are all coterminal. Coterminal angles are angles in standard position (angles with the initial side on the positive x-axis) that have a common terminal side.



To find a positive and a negative angle coterminal with a given angle, you can add and subtract 360° if the angle is measured in degrees or 2π if the angle is measured in radians.

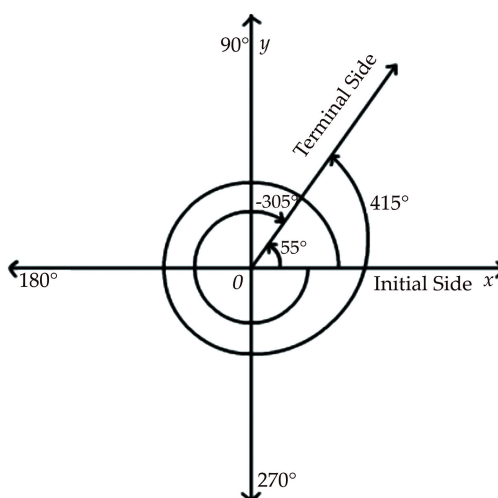
Example:

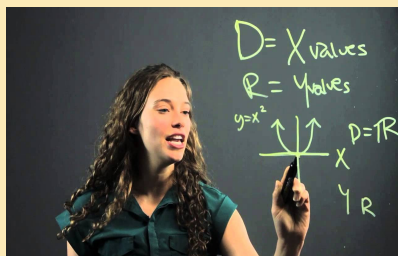
Find a positive and a negative angle coterminal with a 55° angle.

$$55^\circ - 360^\circ = -305^\circ$$

$$55^\circ + 360^\circ = 415^\circ$$

A -305° angle and a 415° angle are coterminal with a 55° angle.





CORE CONCEPT: DOMAIN AND RANGE OF A FUNCTION

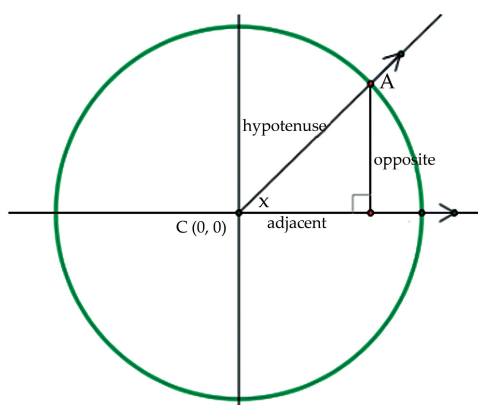
Functions are a correspondence between two sets, called the domain and the range. When defining a function, you usually state what kind of numbers the domain (x) and range ($f(x)$) values can be. But even if you say they are real numbers, that doesn't mean that all real numbers can be used for x . It also doesn't mean that all real numbers can be function values, $f(x)$. There may be restrictions on the domain and range. The restrictions partly depend on the type of function.

A mathematical function is like a black box that takes certain input values and generates corresponding output values. Not all functions work with all kinds of values. The particular set of values that a function can take as input is called its domain. The range of the function is the set of outputs that can be generated by a function, given its domain.

For the function $y = f(x)$, all the values that x can take become the domain of the function and all the output values that y take, become the range of the function.

In case of, circular functions are functions of angles in geometry, which means they take angle measures as input. They relate the angles of a triangle to the length of its sides.

There are three basic circular functions. As we can see in the image, if the point A on the circle is connected with its center, a right triangle is formed.



The three basic circular functions can be defined as

1. Sine (sin): $\sin(x)$ is defined as

$$\sin(x) = \frac{\text{opposite}}{\text{hypotenuse}}$$

The function $\csc(x)$ (cosecant) is the reciprocal of $\sin(x)$ and is defined as

$$\csc(x) = \frac{1}{\sin(x)} = \frac{\text{hypotenuse}}{\text{opposite}}$$

2. Cosine (cos): $\cos(x)$ is defined as

$$\cos(x) = \frac{\text{adjacent}}{\text{hypotenuse}}$$

The function $\sec(x)$ (secant) is the reciprocal of $\cos(x)$ and is defined as

$$\sec(x) = \frac{1}{\cos(x)} = \frac{\text{hypotenuse}}{\text{adjacent}}$$

3. Tangent (tan): $\tan(x)$ is defined as

$$\tan(x) = \frac{\text{opposite}}{\text{adjacent}}$$

The function $\cot(x)$ (cotangent) is the reciprocal of $\tan(x)$ and is defined as

$$\cot(x) = \frac{1}{\tan(x)} = \frac{\text{adjacent}}{\text{opposite}}$$

Domains and Ranges of Circular Functions

Now that you understand the range and domain of a function, let's have a look at the circular functions and define their domains and ranges. One more important point to note here is that the angle measures in radians are denoted as

$$180 \text{ degrees} = \pi \text{ radians}$$

Hypotenuse is the longest side of a right-angled triangle, the side opposite of the right angle.

1. Domain and Range of $\sin(x)$

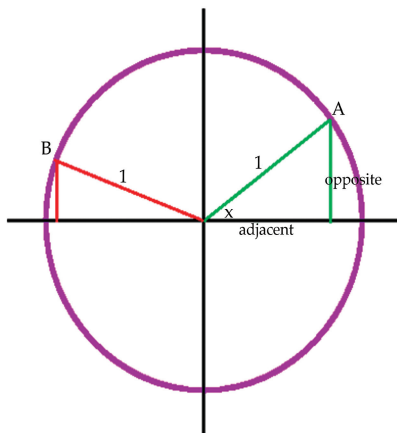
Let's imagine a circle with radius 1 and center at the origin. For any point on this circle, if we draw a right-angled triangle its **hypotenuse** will always be 1.

As we understand, that $\sin(x)$ is defined as,

$$\sin(x) = \frac{\text{opposite}}{\text{hypotenuse}}$$

For this unit circle, at any point, $\sin(x)$ is equal to opposite / 1. This measure of opposite can be defined for all the points

on the circle, indicating that the angle x can take any value. So, the domain of $\sin(x)$ is all real numbers.



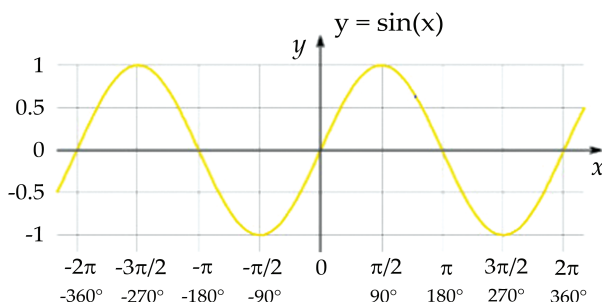
Also, the value of $\sin(x)$, depending on the point on the circle, can go to a maximum of 1 at $x = 90$ degrees and a minimum of -1 at $x = 270$ degrees. So, the range of $\sin(x)$ is -1 to 1.

In short, for $y = \sin(x)$:

Domain = $[+\infty, -\infty]$

Range = $[-1, +1]$

The function $y = \sin(x)$ can be plotted as shown in the image.



Domain and Range of $\csc(x)$

The $\csc(x)$ is the reciprocal of $\sin(x)$. Its domain and range can be found from the domain and range of $\sin(x)$. The $\csc(x)$ is not defined for those values of x for which $\sin(x) = 0$. Its domain is all real numbers excluding $x = 0$ degrees, 180 degrees, 360 degrees, and so on. Similarly, as the value of the range $\sin(x)$ lies between -1 to 1, the value of its reciprocal is either greater than or equal to 1 or lesser than or equal to -1.

Therefore, for the function $y = \csc(x)$

Domain = $\mathbb{R} - n\pi$

Range = $[-\infty, -1] \cup [1, +\infty]$

Domain and Range of Sine Inverse

The inverse of $\sin(x)$ is defined as

$$y = \sin^{-1} x$$

which means:

$$x = \sin(y)$$

In short, the inverse function of $\sin(x)$ is defined for all the points that correspond to a $\sin(x)$ value, which means that its domain is equal to the range of $\sin(x)$, -1 to 1. The range of this inverse function is the angles where $\sin(x)$ has a defined value, from -90 degrees to 90 degrees, as these values cover all the outputs of $\sin(x)$.

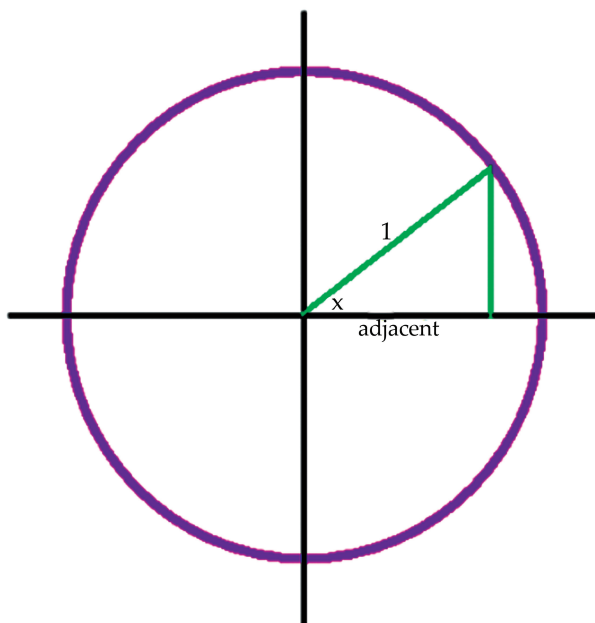
$$\text{For } y = \sin^{-1} x$$

$$\text{Domain} = [-1, 1]$$

$$\text{Range} = [-\pi/2, +\pi/2]$$

2. Domain and Range of $\cos(x)$

As we saw in the case of $\sin(x)$, for any point on a unit circle, $\cos(x)$ is equal to adjacent / 1.



The measure of adjacent can be defined for all the points on the circle, indicating that the angle x can take any value. So, the domain of $\cos(x)$ is all real numbers.

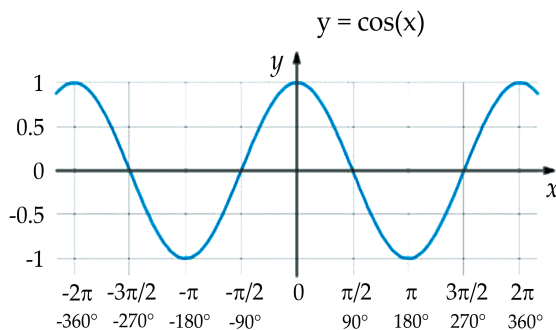
Also, the value of $\cos(x)$, depending on the point on the circle, can go to a maximum of 1 at $x = 0$ degrees and a minimum of -1 at $x = 180$ degrees. So, the range of $\cos(x)$ is from -1 to 1.

In short, for $y = \cos(x)$

Domain = $[-\infty, +\infty]$

Range = $[-1, +1]$

The function $y = \cos(x)$ can be plotted as seen in the graph.



Domain and Range of $\sec(x)$

The $\sec(x)$ is the reciprocal of $\cos(x)$. Its domain and range can be found out from the domain and range of $\cos(x)$. The $\sec(x)$ is defined for those values of x for which $\cos(x) \neq 0$. Its domain excludes $x = 90^\circ$, 270° , 450° , and so on. Similarly, as the value of $\cos(x)$ lies between -1 to 1 , the value of its reciprocal is either greater than or equal to 1 or lesser than or equal to -1 .

Therefore, for the function $y = \sec(x)$

Domain = $\mathbb{R} - \pi/2(n + 1)$

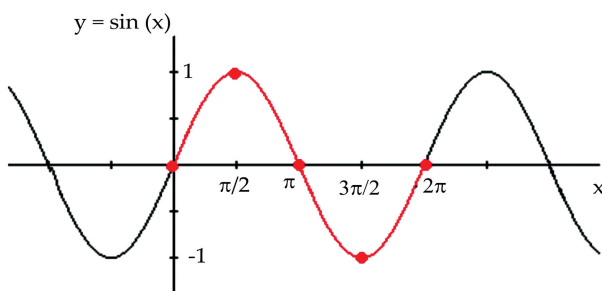
Range = $[-\infty, -1] \cup [1, +\infty]$

Graph the Six Circular Functions

The properties of the 6 circular functions: $\sin(x)$, $\cos(x)$, $\tan(x)$, $\cot(x)$, $\sec(x)$ and $\csc(x)$ are discussed. These include the graph, domain, range, asymptotes (if any), symmetry, x and y intercepts and maximum and minimum points.

Sine Function: $f(x) = \sin(x)$

- Graph

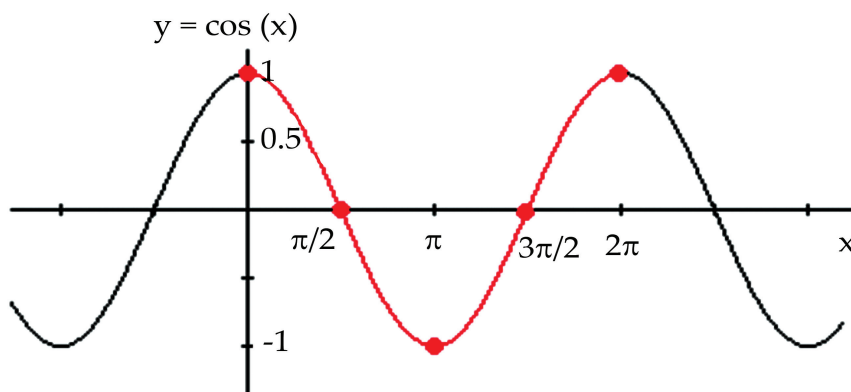


- Domain: all real numbers

- Range: $[-1, 1]$
- Period = 2π
- x intercepts: $x = k\pi$, where k is an integer.
- y intercepts: $y = 0$
- maximum points: $(\pi/2 + 2k\pi, 1)$, where k is an integer.
- minimum points: $(3\pi/2 + 2k\pi, -1)$, where k is an integer.
- symmetry: since $\sin(-x) = -\sin(x)$ then $\sin(x)$ is an odd function and its graph is symmetric with respect to the origin $(0, 0)$.
- intervals of increase/decrease: over one period and from 0 to 2π , $\sin(x)$ is increasing on the intervals $(0, \pi/2)$ and $(3\pi/2, 2\pi)$, and decreasing on the interval $(\pi/2, 3\pi/2)$.

Cosine Function: $f(x) = \cos(x)$

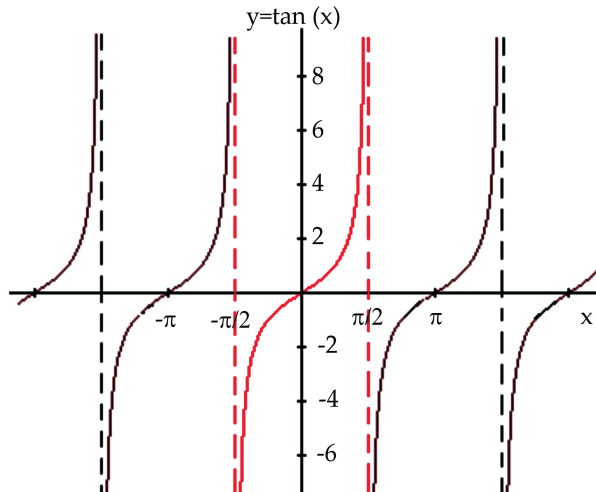
- Graph



- Domain: all real numbers
- Range: $[-1, 1]$
- Period = 2π
- x intercepts: $x = \pi/2 + k\pi$, where k is an integer.
- y intercepts: $y = 1$
- maximum points: $(2k\pi, 1)$, where k is an integer.
- minimum points: $(\pi + 2k\pi, -1)$, where k is an integer.
- symmetry: since $\cos(-x) = \cos(x)$ then $\cos(x)$ is an even function and its graph is symmetric with respect to the y axis.
- intervals of increase/decrease: over one period and from 0 to 2π , $\cos(x)$ is decreasing on $(0, \pi)$ and increasing on $(\pi, 2\pi)$.

Tangent Function: $f(x) = \tan(x)$

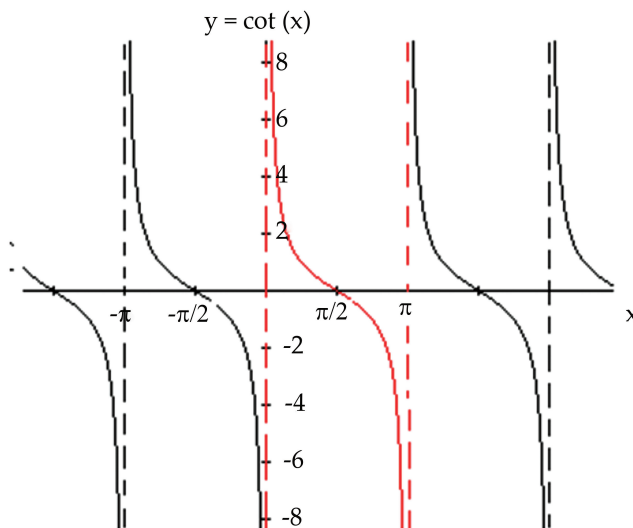
- Graph



- Domain: all real numbers except $\pi/2 + k\pi$, k is an integer.
- Range: all real numbers
- Period = π
- x intercepts: $x = k\pi$, where k is an integer.
- y intercepts: $y = 0$
- symmetry: since $\tan(-x) = -\tan(x)$ then $\tan(x)$ is an odd function and its graph is symmetric with respect to the origin.
- intervals of increase/decrease: over one period and from $-\pi/2$ to $\pi/2$, $\tan(x)$ is increasing.
- Vertical asymptotes: $x = \pi/2 + k\pi$, where k is an integer.

Cotangent Function: $f(x) = \cot(x)$

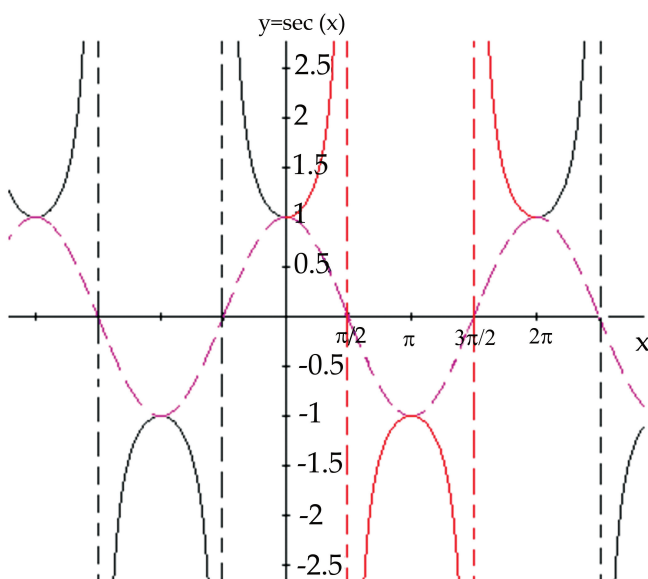
- Graph



- Domain: all real numbers except $k\pi$, k is an integer.
- Range: all real numbers
- Period = π
- x intercepts: $x = \pi/2 + k\pi$, where k is an integer.
- symmetry: since $\cot(-x) = -\cot(x)$ then $\cot(x)$ is an odd function and its graph is symmetric with respect to the origin.
- intervals of increase/decrease: over one period and from 0 to π , $\cot(x)$ is decreasing.
- Vertical asymptotes: $x = k\pi$, where k is an integer.

Secant Function: $f(x) = \sec(x)$

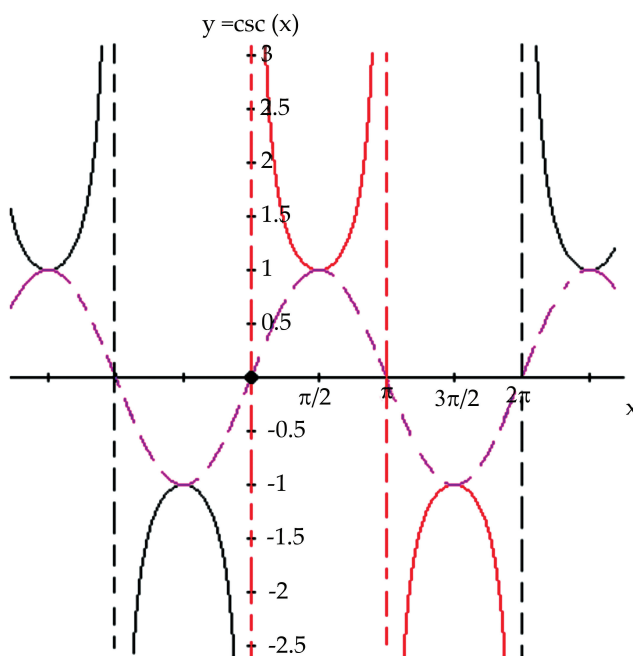
- Graph



- Domain: all real numbers except $\pi/2 + k\pi$, n is an integer.
- Range: $(-\infty, -1] \cup [1, +\infty)$
- Period = 2π
- y intercepts: $y = 1$
- symmetry: since $\sec(-x) = \sec(x)$ then $\sec(x)$ is an even function and its graph is symmetric with respect to the y axis.
- intervals of increase/decrease: over one period and from 0 to 2π , $\sec(x)$ is increasing on $(0, \pi/2) \cup (\pi/2, \pi)$ and decreasing on $(\pi, 3\pi/2) \cup (3\pi/2, 2\pi)$.
- Vertical asymptotes: $x = \pi/2 + k\pi$, where k is an integer.

Cosecant Function: $f(x) = \csc(x)$

- Graph



- Domain: all real numbers except $k\pi$, k is an integer.
- Range: $(-\infty, -1] \cup [1, +\infty)$
- Period = 2π
- symmetry: since $\csc(-x) = -\csc(x)$ then $\csc(x)$ is an odd function and its graph is symmetric with respect to the origin.
- intervals of increase/decrease: over one period and from 0 to 2π , $\csc(x)$ is decreasing on $(0, \pi/2) \cup (3\pi/2, 2\pi)$ and increasing on $(\pi/2, \pi) \cup (\pi, 3\pi/2)$.
- Vertical asymptotes: $x = k\pi$, where k is an integer.

CONCEPT: CIRCULAR FUNCTIONS

Circular functions, also known as trigonometric functions, are mathematical functions that describe the relationships between the angles of a right triangle and the ratios of the lengths of its sides. These functions have a rich history and find extensive applications in various fields, including geometry, physics, engineering, and more.

There are six primary circular functions, often abbreviated as “trig functions”:

Sine (sin): The sine of an angle in a right triangle is defined as the ratio of the length of the side opposite the angle to the length of the hypotenuse.

Cosine (cos): The cosine of an angle is defined as the ratio of the length of the side adjacent to the angle to the length of the hypotenuse.

Tangent (tan): The tangent of an angle is defined as the ratio of the length of the side opposite the angle to the length of the side adjacent to the angle.

Cosecant (csc): The cosecant is the reciprocal of the sine: $\csc(\theta) = 1/\sin(\theta)$.

Secant (sec): The secant is the reciprocal of the cosine: $\sec(\theta) = 1/\cos(\theta)$.

Cotangent (cot): The cotangent is the reciprocal of the tangent: $\cot(\theta) = 1/\tan(\theta)$.

These functions are most commonly defined on the unit circle, where the radius of the circle is 1. In this context, the coordinates of a point on the unit circle corresponding to an angle θ are $(\cos(\theta), \sin(\theta))$.

Key properties and relationships of circular functions:

The Pythagorean identity: $\sin^2(\theta) + \cos^2(\theta) = 1$.

The cofunction identities: These relate the trigonometric functions of complementary angles. For example, $\sin(\theta) = \cos(90^\circ - \theta)$ and vice versa.

The periodic nature: Circular functions are periodic, meaning they repeat their values after a certain interval. The period of $\sin$ and $\cos$ is 360° (or 2π radians), while the period of $\tan$, $\cot$, $\csc$, and $\sec$ is 180° (or π radians).

The relationships between circular functions and triangles, which allow for solving for unknown side lengths or angles in triangles.

Circular functions are extensively used to model waveforms, oscillations, rotational motion, and periodic phenomena in physics and engineering. They are also crucial in calculus, complex analysis, and other mathematical fields.

SUMMARY

- Circular functions are defined such that their domains are sets of numbers that correspond to the measures (in radian units) of the angles of analogous trigonometric functions.
- Circular functions are part of the set of trigonometric functions. Whereas trigonometric functions consist of domains that are sets of angles and ranges that are real numbers, circular functions have domains that are sets of numbers that correspond to the angles of the trigonometric functions (in radians).
- An angle is a measurement of a rotation or change of direction. In plane geometry, an angle is the measurement of the rotation required to move one ray onto another ray emanating from the same point.
- An angle in a rectangular coordinate system is said to be in standard position if its vertex is at the origin and the initial side is along the positive x axis.
- An angle is in standard position if its vertex is located at the origin and one ray is on the positive x-axis. The ray on the x-axis is called the initial side and the other ray is called the terminal side.
- A mathematical function is like a black box that takes certain input values and generates corresponding output values. Not all functions work with all kinds of values. The particular set of values that a function can take as input is called its domain.

MULTIPLE CHOICE QUESTIONS

1. If $\sin \theta$ and $\cos \theta$ are the roots of $ax^2 - bx + c = 0$, then the relation between a , b and c will be
 - a. $a^2 + b^2 + 2ac = 0$
 - b. $a^2 - b^2 + 2ac = 0$
 - c. $a^2 + c^2 + 2ab = 0$
 - d. $a^2 - b^2 - 2ac = 0$
2. If $\tan A = 1/2$ and $\tan B = 1/3$, then the value of $A + B$ is
 - a. $\pi/6$
 - b. π
 - c. 0
 - d. $\pi/4$
3. The value of $\cos 1^\circ \cos 2^\circ \cos 3^\circ \dots \cos 179^\circ$ is
 - a. $1/\sqrt{2}$
 - b. 0
 - c. 1
 - d. -1
4. An angle is a measurement of a rotation or change of direction.
 - a. True
 - b. False
5. An angle is in standard position if its vertex is located at the origin and one ray is on the positive x-axis.
 - a. True
 - b. False

Answers

1. (b) 2. (d) 3. (b) 4. (a) 5. (a)

REVIEW QUESTIONS

1. Define the basic theory of circular functions.
2. How to find exact values for particular real numbers? Explain.
3. Discuss on angles and their measurements.
4. Evaluate the degree and radian measure.
5. What is the standard position, reference angles and coterminal angles?

REFERENCES

1. Axler, Sheldon. Algebra and Trigonometry. Wiley: 2011, 1st ed. (contains worked solutions to problems, recommended for self-study; note that text and Axler's Precalculus cover nearly-identical content with slightly different focus)
2. Axler, Sheldon. Precalculus: A Prelude to Calculus. Wiley: 2012, 2nd ed. (contains worked solutions to problems, recommended for self-study)
3. Carl, S. and Jeff, Z. (2013) College Trigonometry. Lorain County Community College.
4. Delbaen, F. and Schachermayer, W. (2006) The Mathematics of Arbitrage. Springer, Berlin.
5. Honghai, L., George, M.C. and Dave, P.B. (2009) Fuzzy Qualitative Trigonometry. International Journal of Approximate Reasoning, 51, 71-88.
6. John, B. (2008) Geometry and Trigonometry. Engineering Mathematics Pocket Book (Fourth Edition), 105-148.
7. Karl, G. (2010) Operator Trigonometry of Multivariate Finance. Journal of Multivariate Analysis, 101, 374-384.
8. Kelley, W. Michael. The Humongous Book of Algebra Problems. ALPHA: 2008 (illustrates the important techniques of basic algebra through many series of example problems)
9. Lima, J.B. and Campello de Souza, R.M. (2013) Fractional Cosine and Sine Transforms over Finite Fields. Linear Algebra and Its Applications, 438, 3217-3230.

CONIC SECTIONS AND SYSTEMS



Conic section (or simply conic) is a curve obtained as the intersection of the surface of a cone with a plane. The three types of conic section are the hyperbola, the parabola, and the ellipse. The circle is a special case of the ellipse, and is of sufficient interest in its own right that it was sometimes called a fourth type of conic section.

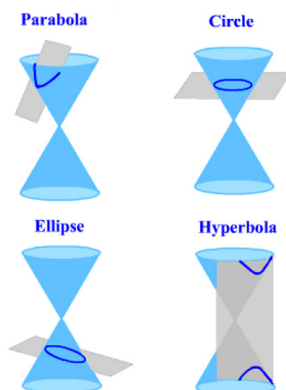
After studying this chapter, you will understand the core concepts of:

- Conic Sections
- Systems of Nonlinear Equations

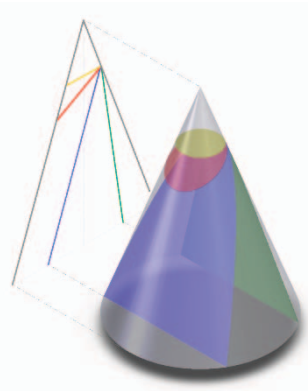
INTRODUCTION

A conic section, conic or a quadratic curve is a curve obtained from a cone's surface intersecting a plane. The three types of conic section are the hyperbola, the parabola, and the ellipse; the circle is a special case of the ellipse, though it was sometimes called as a fourth type. The ancient Greek mathematicians studied conic sections, culminating around 200 BC with Apollonius of Perga's systematic work on their properties.

A conic section is the intersection of a plane and a double right circular cone. By changing the angle and location of the intersection, we can produce different types of conics. There are four basic types: circles, ellipses, hyperbolas and parabolas. None of the intersections will pass through the vertices of the cone.



If the right circular cone is cut by a plane perpendicular to the axis of the cone, the intersection is a circle. If the plane intersects one of the pieces of the cone and its axis but is not perpendicular to the axis, the intersection will be an ellipse. To generate a parabola, the intersecting plane must be parallel to one side of the cone and it should intersect one piece of the double cone. And finally, to generate a hyperbola the plane intersects both pieces of the cone. For this, the slope of the intersecting plane should be greater than that of the cone.



The general equation for any conic section is

$Ax^2+Bxy+Cy^2+Dx+Ey+F=0$ where A, B, C, D, E and F are constants.

As we change the values of some of the constants, the shape of the corresponding conic will also change. It is important to know the differences in the equations to help quickly identify the type of conic that is represented by a given equation.

If $B^2 - 4AC$ is less than zero, if a conic exists, it will be either a circle or an ellipse.

If $B^2 - 4AC$ equals zero, if a conic exists, it will be a parabola.

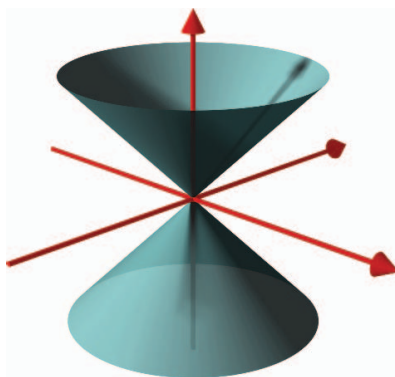
If $B^2 - 4AC$ is greater than zero, if a conic exists, it will be a hyperbola.



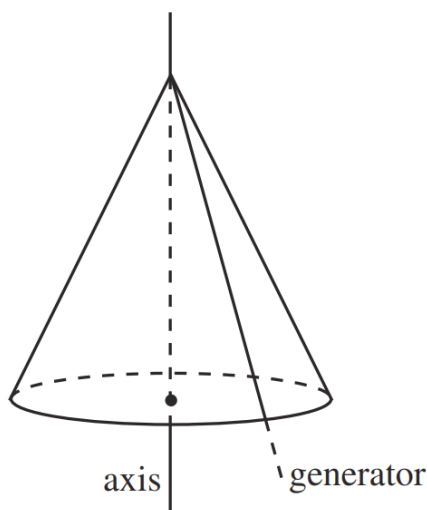
CORE CONCEPT: CONIC SECTIONS

A conic section is the intersection of a plane and a cone. In the Euclidean plane, the conic sections appear to be quite different from one another, but share many properties. By extending the geometry to a projective plane (adding a line at infinity) this apparent difference vanishes, and the commonality becomes evident.

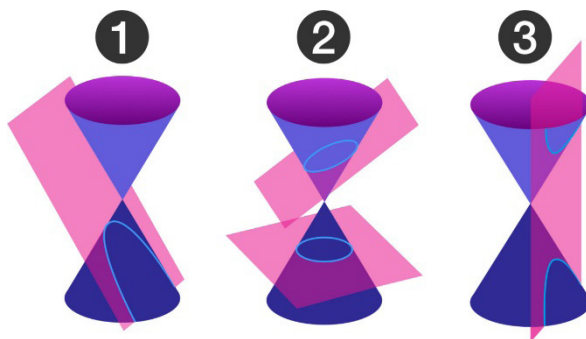
The conic sections, or conics, are curves obtained by making sections, or cuts, at particular angles through a cone. These curves were known to the ancient Greeks, who first explored their properties. But, ancient as they are, these curves have exceedingly modern applications — for example, the television dish aerial that brings sports and films into our homes, and the giant radio telescopes such as the one at Jodrell Bank in Cheshire, which look deep into space, all depend upon what is known as the ‘reflective property’ of one of the conic sections.



In the description below, we shall use diagrams showing just a part of a cone, but you should imagine that the cone extends infinitely far. We shall also talk about the axis of the cone, and a generator. The axis is the central line about which the cone is symmetric. A generator is a line which, when rotated about the axis, sweeps out the cone.



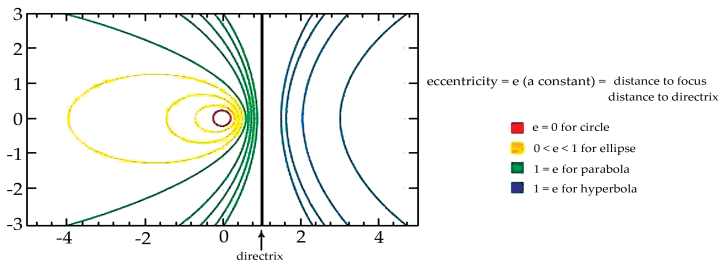
Conic section, also called conic, any curve produced by the intersection of a plane and a right circular cone. Depending on the angle of the plane relative to the cone, the intersection is a circle, an ellipse, a hyperbola, or a parabola. Special (degenerate) cases of intersection occur when the plane passes through only the apex (producing a single point) or through the apex and another point on the cone (producing one straight line or two intersecting straight lines).



The basic descriptions, but not the names, of the conic sections can be traced to Menaechmus, a pupil of both Plato and Eudoxus of Cnidus. Apollonius of Perga, known as the “Great Geometer,” gave the conic sections their names and was the first to define the two branches of the hyperbola (which presuppose the double cone). Apollonius’s eight-volume treatise on the conic sections, *Conics*, is one of the greatest scientific works from the ancient world.

Analytic Definition

Conics may also be described as plane curves that are the paths (loci) of a point moving so that the ratio of its distance from a fixed point (the focus) to the distance from a fixed line (the directrix) is a constant, called the eccentricity of the curve. If the eccentricity is zero, the curve is a circle; if equal to one, a parabola; if less than one, an ellipse; and if greater than one, a hyperbola. See the figure.



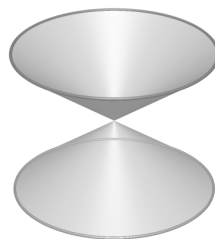
Every conic section corresponds to the graph of a second degree polynomial equation of the form $Ax^2 + By^2 + 2Cxy + 2Dx + 2Ey + F = 0$, where x and y are variables and A , B , C , D , E , and F are coefficients that depend upon the particular conic. By a suitable choice of coordinate axes, the equation for any conic can be reduced to one of three simple forms: $x^2/a^2 + y^2/b^2 = 1$, $x^2/a^2 - y^2/b^2 = 1$, or $y^2 = 2px$, corresponding to an ellipse, a hyperbola, and a parabola, respectively. (An ellipse where $a = b$ is in fact a circle.) The extensive use of coordinate systems for the algebraic analysis of geometric curves originated with René Descartes (1596–1650).

Conic Sections in Nature

Conic sections are important because they model important physical processes in nature. It can be shown that anybody under the influence of an inverse square law force must have a trajectory of one of the conic sections. Heavenly bodies attract each other with a gravitational force that is inversely proportional to the square of the distance between them. This law, Newton's Law of Gravitation, is an example of an inverse square law. Hence, the trajectories of heavenly bodies are conic sections (circle, ellipses, parabolas, and hyperbolas.). Coulomb's Law of the attraction of charged particles is another example of an inverse square law. Therefore, the trajectories of charged particles (e.g., electrons) are conic sections. So, from the large scale of the universe to the microscopic scale of the atom, conic sections occur in nature.

Inverse-square law is any physical law stating that a specified physical quantity or intensity is inversely proportional to the square of the distance from the source of that physical quantity.

The conic sections are the shapes that can be created when a plane intersects a double cone like the one below. In other words, the conic sections are the cross sections of a double cone.



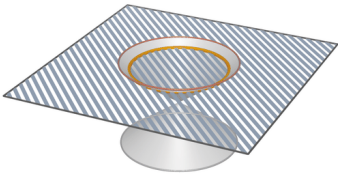
There are four primary conic sections - the circle, the parabola, the ellipse, and the hyperbola. These conic sections are shown below with their general equations. In geometry, you will focus on circles and parabolas. You will study ellipses and hyperbolas in more depth in future courses.

Conic Section	Equation	Picture
Circle	$(x-h)^2+(y-k)^2=r^2$	
Parabola	$(x-h)^2=4p(y-k)$ OR $(y-k)^2=4p(x-h)$	
Ellipse	$\frac{(x-h)^2}{a^2} + \frac{(y-k)^2}{b^2} = 1$	
Hyperbola	$\frac{(x-h)^2}{a^2} - \frac{(y-k)^2}{b^2} = 1$ OR $\frac{(y-k)^2}{a^2} - \frac{(x-h)^2}{b^2} = 1$	

Let’s look at some problems involving conic sections.

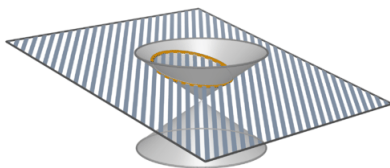
1. How is a circle created as the intersection of a double cone and a plane?

The cross section is taken parallel to the base of the cones, as shown below.



2. How is an ellipse created as the intersection of a double cone and a plane?

The intersection plane is at a slight angle with the plane containing the base, as shown below.



3. a) $\frac{(x-1)^2}{4} + \frac{(y+3)^2}{9} = 1$

Ellipse

b) $(x-4)^2 = 4(y-1)$

Parabola

c) $(x-2)^2 + (y-3)^2 = 16$

Circle

d) $\frac{(x-1)^2}{4} - \frac{(y+3)^2}{9} = 1$

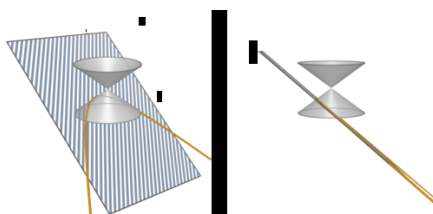
Hyperbola

Notice that the only equation where both the x and y terms are not squared is the parabola. The difference between the ellipse and hyperbola equations is with an ellipse the coefficients of x and y are the same sign while with a hyperbola the coefficients of x and y are different signs. The coefficients of the x and y terms are always exactly the same for a circle.

Example 1. Earlier, you were asked why conic sections are named accordingly. Conic sections are called conic sections because they are the cross sections of a double cone. The word conic comes from the word cone, and the word section comes from cross section.

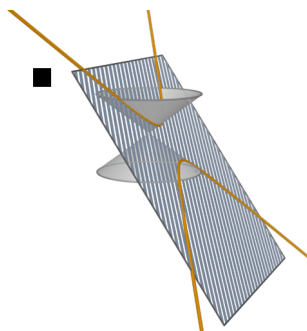
Example 2. How is a parabola created as the intersection of a double cone and a plane?

The intersection plane has the same slope as the lateral surface of the cone and passes through just one cone. Below, the same intersection is shown from two different perspectives.



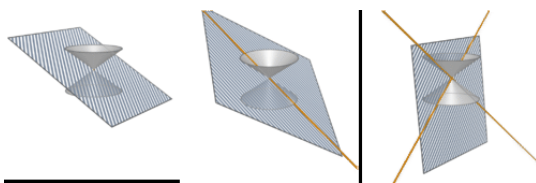
Example 3. How is an ellipse created as the intersection of a double cone and a plane?

The intersection plane is at an angle with the plane containing the base such that it passes through both cones, as shown below.



Example 4. Are there any other cross sections of a double cone besides the four primary conic sections?

There are three other cross sections -- a point, a line, and a pair of intersecting lines. These are referred to as the degenerate conics, because they are not as complex. The study of conic sections focuses on the four primary conic sections and not on the degenerate conics.

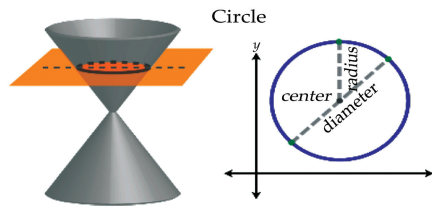


Greek Origins

The early history of conic sections is joined to the problem of “doubling the cube.” According to Eratosthenes of Cyrene, the people of Delos consulted the oracle of Apollo for aid in ending a plague (c. 430 BC) and were instructed to build Apollo a new altar of twice the old altar’s volume and with the same cubic shape. Perplexed, the Delians consulted Plato, who stated that “the oracle meant, not that the god wanted an altar of double the size, but that he wished, in setting them the task, to shame the Greeks for their neglect of mathematics and their contempt for geometry.” Hippocrates of Chio first discovered that the “Delian problem” can be reduced to finding two mean proportionals between a and $2a$ (the volumes of the respective altars)—that is, determining x and y such that $a:x = x:y = y:2a$. This is equivalent to solving simultaneously any two of the equations $x^2 = ay$, $y^2 = 2ax$, and $xy = 2a^2$, which correspond to two parabolas and a hyperbola, respectively. Later, Archimedes showed how to use conic sections to divide a sphere into two segments having a given ratio.

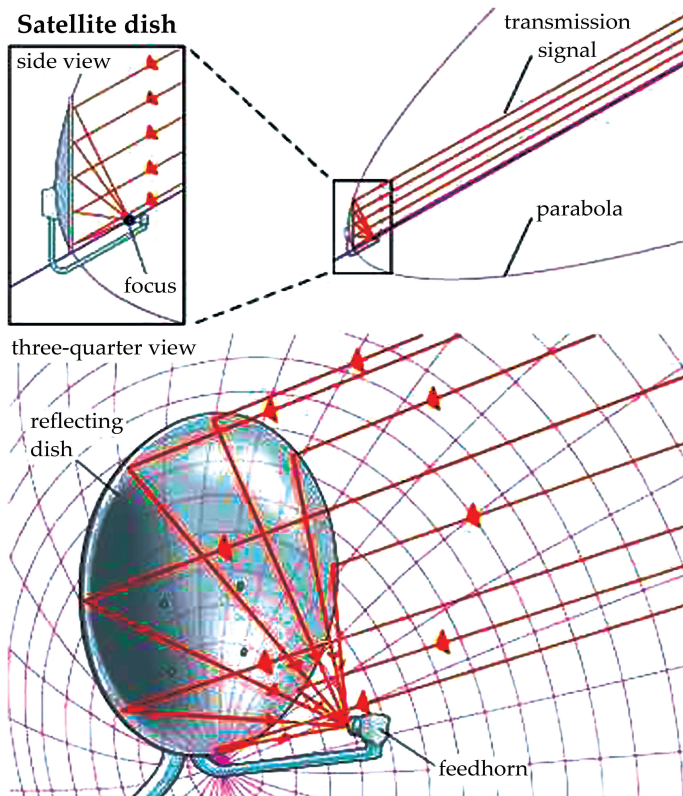
Diocles demonstrated geometrically that rays—for instance, from the Sun—that are parallel to the axis of a paraboloid of revolution (produced by rotating a parabola about its axis of symmetry) meet at the focus. Archimedes is said to have used this property to set enemy ships on fire. The focal properties of the ellipse were cited by Anthemius of Tralles, one of the architects for Hagia Sophia Cathedral in

Constantinople (completed in AD 537), as a means of ensuring that an altar could be illuminated by sunlight all day.



Post-Greek Applications

Conic sections found their first practical application outside of optics in 1609 when Johannes Kepler derived his first law of planetary motion: A planet travels in an ellipse with the Sun at one focus. Galileo Galilei published the first correct description of the path of projectiles a parabola in his *Dialogues of the Two New Sciences*. In 1639 the French engineer Girard Desargues initiated the study of those properties of conics that are invariant under projections. Eighteenth-century architects created a fad for whispering galleries such as in the U.S. Capitol and in St. Paul's Cathedral in London—in which a whisper at one focus of an ellipsoid (an ellipse rotated about one axis) can be heard at the other focus, but nowhere else. From the ubiquitous parabolic satellite dish (see the figure) to the use of ultrasound in lithotripsy, new applications for conic sections continue to be found.



Rectangular hyperbola is a hyperbola for which the asymptotes are perpendicular, also called an equilateral hyperbola or right hyperbola. This occurs when the semimajor and semiminor axes are equal.

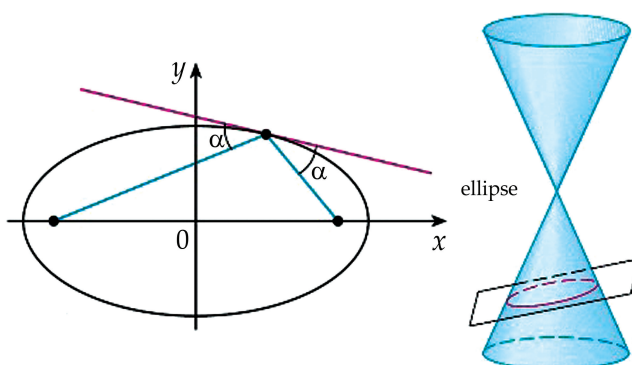
The Ellipse, Parabola and Hyperbola

Mathematicians, engineers and scientists encounter numerous functions in their work: polynomials, trigonometric and hyperbolic functions amongst them. However, throughout the history of science one group of functions, the conics, arise time and time again not only in the development of mathematical theory but also in practical applications. The conics were first studied by the Greek mathematician Apollonius more than 200 years BC.

Essentially, the conics form that class of curves which are obtained when a double cone is intersected by a plane. There are three main types: the ellipse, the parabola and the hyperbola. From the ellipse we obtain the circle as a special case, and from the hyperbola we obtain the **rectangular hyperbola** as a special case.

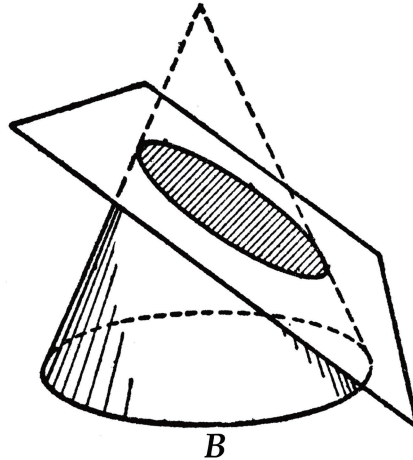
The Ellipse

Ellipse, a closed curve, the intersection of a right circular cone (see cone) and a plane that is not parallel to the base, the axis, or an element of the cone. It may be defined as the path of a point moving in a plane so that the ratio of its distances from a fixed point (the focus) and a fixed straight line (the directrix) is a constant less than one. Any such path has this same property with respect to a second fixed point and a second fixed line, and ellipses often are regarded as having two foci and two directrices. The ratio of distances, called the eccentricity, is the discriminant (q.v.; of a general equation that represents all the conic sections). Another definition of an ellipse is that it is the locus of points for which the sum of their distances from two fixed points (the foci) is constant. The smaller the distance between the foci, the smaller is the eccentricity and the more closely the ellipse resembles a circle.



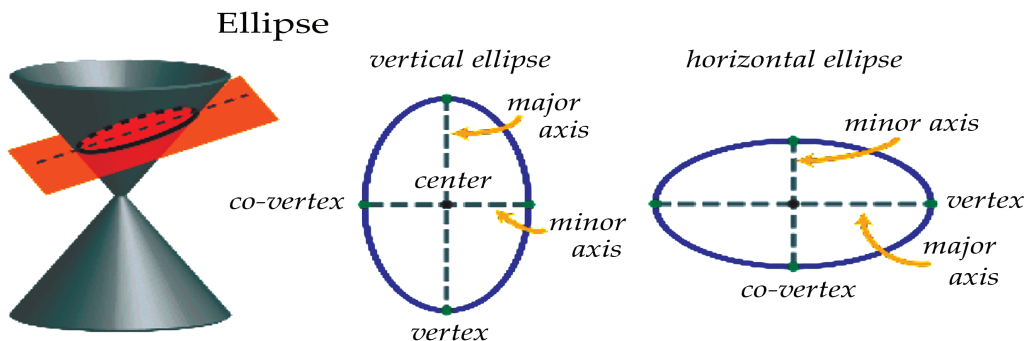
A straight line drawn through the foci and extended to the curve in either direction is the major diameter (or major axis) of the ellipse. Perpendicular to the major axis through the center, at the point on the major axis equidistant from the foci, is the minor axis. A line drawn through either focus parallel to the minor axis is a latus rectum (literally, “straight side”).

The ellipse is symmetrical about both its axes. The curve when rotated about either axis forms the surface called the ellipsoid (q.v.) of revolution, or a spheroid.

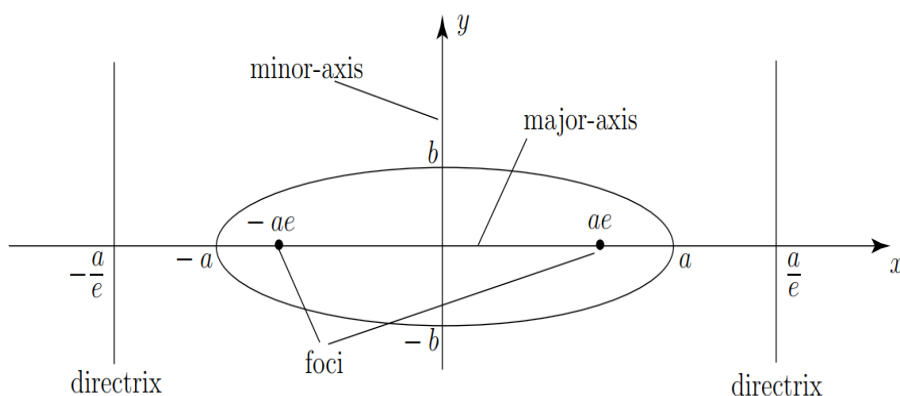


The path of a heavenly body moving around another in a closed orbit in accordance with Newton’s gravitational law is an ellipse. In the solar system one focus of such a path about the Sun is the Sun itself.

For an ellipse the centre of which is at the origin and the axes of which are coincident with the x and y axes, the equation is $x^2/a^2 + y^2/b^2 = 1$. The length of the major diameter is $2a$; the length of the minor diameter is $2b$. If c is taken as the distance from the origin to the focus, then $c^2 = a^2 - b^2$, and the foci of the curve may be located when the major and minor diameters are known. The problem of finding an exact expression for the perimeter of an ellipse led to the development of elliptic functions, an important topic in mathematics and physics.



We are all aware that the paths followed by the planets around the sun are elliptical. However, more generally the ellipse occurs in many areas of engineering. The standard form of an ellipse is shown in Figure 1.

**Figure 1.**

If $a > b$ (as we have drawn the figure) then the x -axis is called the major-axis and the y -axis is called the minor-axis. On the other hand if $b > a$ then the y -axis is called the major-axis and the x -axis is then the minor-axis. Two points, inside the ellipse are of importance; these are the foci. If $a > b$ these are located at coordinate positions $\pm ae$ (or at $\pm be$ if $b > a$) on the major-axis in which e , called the eccentricity, is given by

$$e^2 = 1 - \frac{b^2}{a^2} \quad (b < a) \quad \text{or by} \quad e^2 = 1 - \frac{a^2}{b^2} \quad (a < b)$$

The foci of an ellipse have the property that if light rays are emitted from one focus then on reflection at the elliptic curve they pass through at the other focus.

The standard Cartesian equation of the ellipse with its center at the origin is

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \quad (1)$$

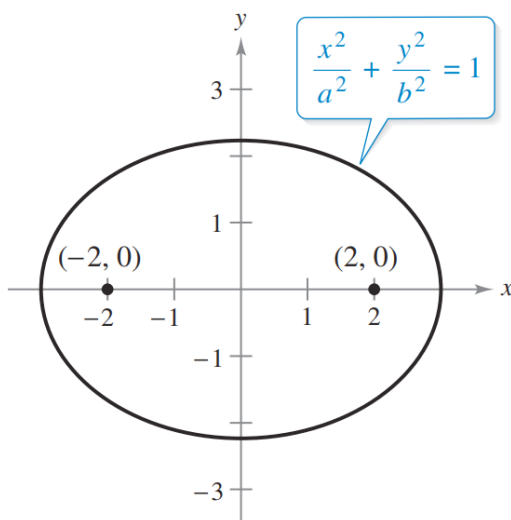
This ellipse has intercepts on the x -axis at $x = \pm a$ and on the y -axis at $\pm b$. The curve is also symmetrical about both axes. The curve reduces to a circle in the special case in which $a = b$.

The equation (1) gives the equation of the ellipse with its center at the origin. If the center of the ellipse has coordinates (α, β) and still has its axes parallel to the x and y -axes the standard equation becomes

$$\frac{(x - \alpha)^2}{a^2} + \frac{(y - \beta)^2}{b^2} = 1.$$

Example

Find the standard form of the equation of the ellipse that has a major axis of length 6 and foci at $(-2, 0)$ and $(2, 0)$ as shown in figure below.



Solution: Because the foci occur at $(-2, 0)$ and $(2, 0)$ the center of the ellipse is $(0, 0)$ and the major axis is horizontal. So, the ellipse has an equation of the form

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1. \quad (\text{Standard form, horizontal major axis})$$

Because the length of the major axis is 6, you have $2a = 6$, which implies that $a = 3$

Moreover, the distance from the center to either focus is $c = 2$. Finally, you have

$$b^2 = a^2 - c^2 = 3^2 - 2^2 = 9 - 4 = 5.$$

Substituting $a^2 = 3^2$ and $b^2 = (\sqrt{5})^2$ yields the following equation in standard form.

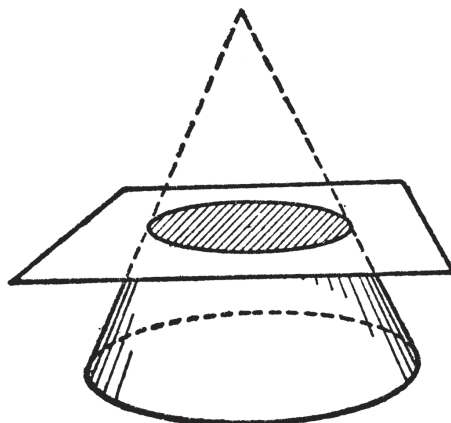
$$\frac{x^2}{3^2} + \frac{y^2}{(\sqrt{5})^2} = 1 \quad (\text{Standard form})$$

This equation simplifies to

$$\frac{x^2}{9} + \frac{y^2}{5} = 1.$$

The Circle

Circle, geometrical curve, one of the conic sections, consisting of the set of all points the same distance (the radius) from a given point (the center). A line connecting any two points on a circle is called a chord, and a chord passing through the center is called a diameter. The distance around a circle (the circumference) equals the length of a diameter multiplied by π . The area of a circle is the square of the radius multiplied by π . An arc consists of any part of a circle encompassed by an angle with its vertex at the center (central angle). Its length is in the same proportion to the circumference as the central angle is to a full revolution.



The circle is a special case of the ellipse; it occurs when $a = b = r$ so the equation becomes

$$\frac{x^2}{r^2} + \frac{y^2}{r^2} = 1 \quad \text{or, more commonly} \quad x^2 + y^2 = r^2$$

Here, the center of the circle is located at the origin $(0, 0)$ and the radius of the circle is r . If the center of the circle at a point (α, β) then we use the form:

$$(x - \alpha)^2 + (y - \beta)^2 = r^2$$

In Figure 2 we have drawn a number of circles

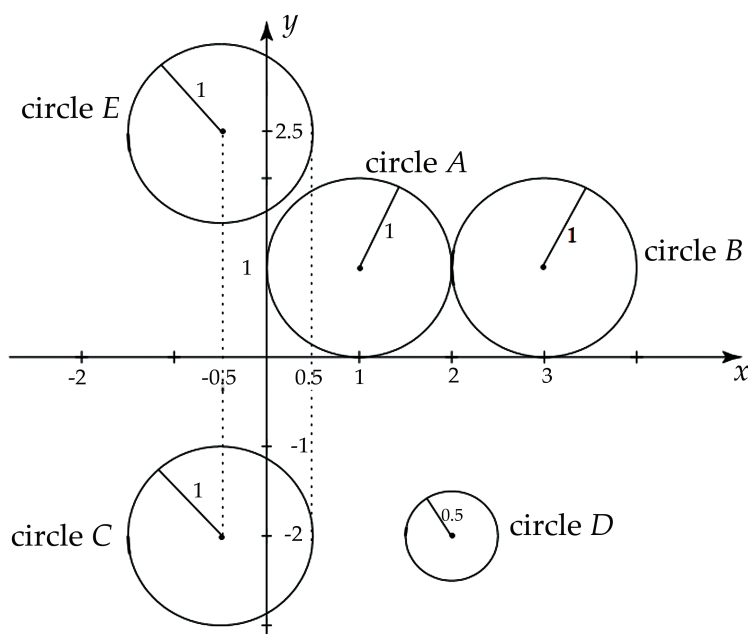


Figure 2.

The reader should confirm that the equations of these circles are

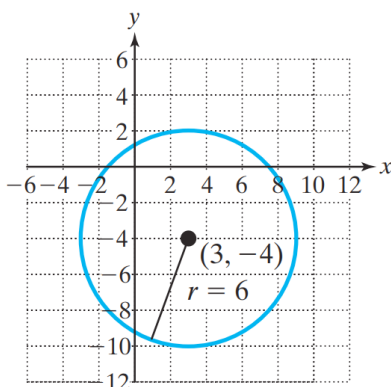
circle	equation
A	$(x - 1)^2 + (y - 1)^2 = 1$
B	$(x - 3)^2 + (y - 1)^2 = 1$
C	$(x + 0.5)^2 + (y + 2)^2 = 1$
D	$(x - 2)^2 + (y + 2)^2 = 0.25$
E	$(x + 0.5)^2 + (y - 2.5)^2 = 1$

Example

Find the center and radius of each circle. Then graph the circle.

$$(x - 3)^2 + (y + 4)^2 = 36$$

Solution:



$$(x - 3)^2 + (y + 4)^2 = 36$$

$$(x - 3)^2 + [y - (-4)]^2 = (6)^2$$

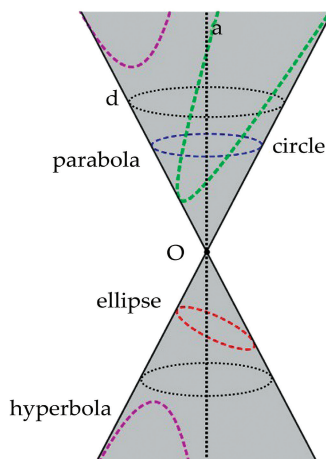
$$h = 3, k = -4, \text{ and } r = 6$$

The center is and the radius is 6.

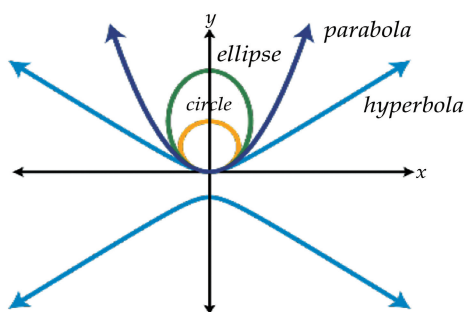
The Parabola

Parabola, open curve, a conic section produced by the intersection of a right circular cone and a plane parallel to an element of the cone. As a plane curve, it may be defined as the path (locus) of a point moving so that its distance from a fixed line (the directrix) is equal to its distance from a fixed point (the focus).

The vertex of the parabola is the point on the curve that is closest to the directrix; it is equidistant from the directrix and the focus. The vertex and the focus determine a line, perpendicular to the directrix, that is the axis of the parabola. The line through the focus parallel to the directrix is the latus rectum (straight side). The parabola is symmetric about its axis, moving farther from the axis as the curve recedes in the direction away from its vertex. Rotation of a parabola about its axis forms a paraboloid (q.v.).



The parabola is the path, neglecting air resistance and rotational effects, of a projectile thrown outward into the air. The parabolic shape also is seen in certain bridges, forming arches.



For a parabola the axis of which is the x axis and with vertex at the origin, the equation is $y^2 = 2px$, in which p is the distance between the directrix and the focus.

The standard form of the parabola is shown in Figure 3.

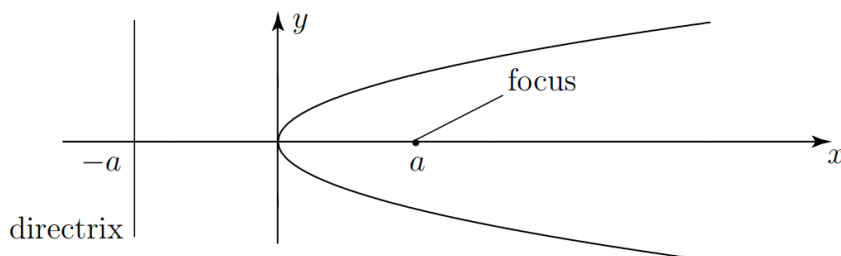


Figure 3.

Here the x -axis is the line of symmetry of the parabola. The standard equation of the parabola is

$$y^2 = 4ax$$

with focus at $(a, 0)$.

It can be shown that light rays parallel to the x -axis will, on reflection from the parabolic curve, come together at the focus. This is an important property and is used in the construction of some kinds of telescopes, of satellite dishes and of car headlights.

Example

Find the focus of the parabola whose equation is $y = -2x^2$

Solution: Because the squared term in the equation involves you know that the axis is vertical, and the equation is of the form

$x^2 = 4py$. (Standard form, vertical axis)

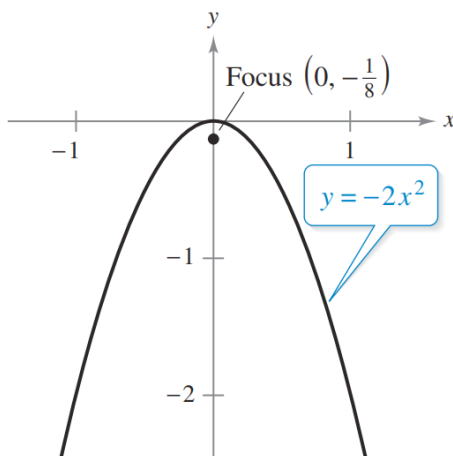
You can write the original equation in this form, as shown.

$$-2x^2 = y \quad (\text{Write original equation.})$$

$$x^2 = -\frac{1}{2}y \quad (\text{Divide each side by } -2)$$

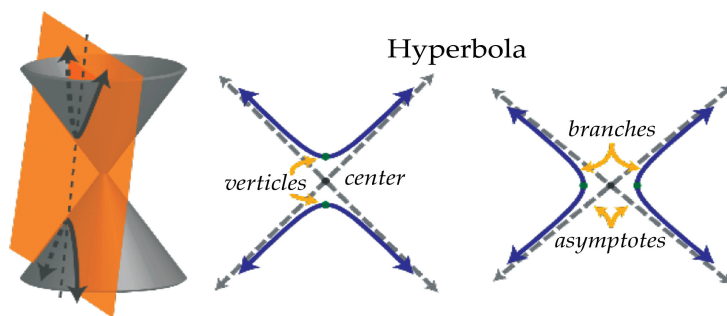
$$x^2 = 4\left(-\frac{1}{8}\right)y. \quad (\text{Write in standard form.})$$

So, $p = -1/8$. Because p is negative, the parabola opens downward and the focus of the parabola is $(0, p) = (0, -1/8)$ as shown in figure below.



The Hyperbola

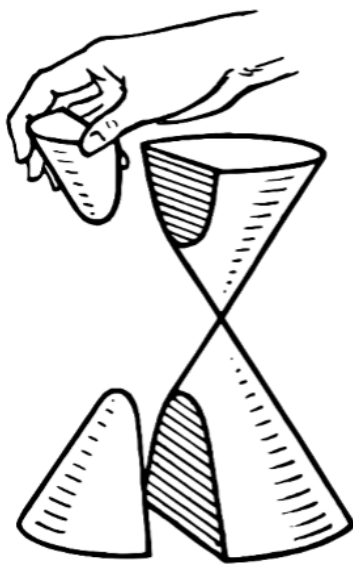
Hyperbola, two-branched open curve, a conic section, produced by the intersection of a circular cone and a plane that cuts both nappes (see cone) of the cone. As a plane curve it may be defined as the path (locus) of a point moving so that the ratio of the distance from a fixed point (the focus) to the distance from a fixed line (the directrix) is a constant greater than one. The hyperbola, however, because of its symmetry, has two foci. Another definition is that of a point moving so that the difference of its distances from two fixed points, or foci, is a constant. A degenerate hyperbola (two intersecting lines) is formed by the intersection of a circular cone and a plane that cuts both nappes of the cone through the apex.



A line drawn through the foci and prolonged beyond is the transverse axis of the hyperbola; perpendicular to that axis, and intersecting it at the geometric center of the hyperbola, a point midway between the two foci, lies the conjugate axis. The hyperbola is symmetrical with respect to both axes.

Two straight lines, the asymptotes of the curve, pass through the geometric center. The hyperbola does not intersect the asymptotes, but its distance from them becomes arbitrarily small at great distances from the center. The hyperbola when revolved about either axis forms a hyperboloid (q.v.).

A parabola is a graph of a quadratic function, $y = x^2$, for example. The line perpendicular to the directrix and passing through the focus (that is, the line that splits the parabola through the middle) is called the "axis of symmetry".



For a hyperbola that has its center at the origin of a Cartesian coordinate system and has its transverse axis lying on the x axis, the coordinates of its points satisfy the equation $x^2/a^2 - y^2/b^2 = 1$, in which a and b are constants.

The standard form of the hyperbola is shown in Figure 4(a).

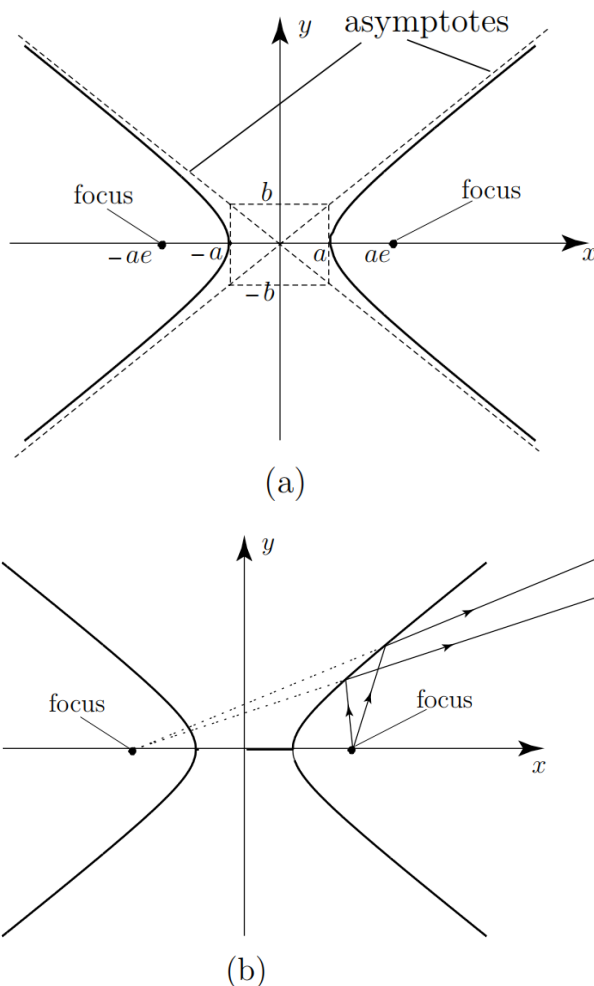


Figure 4.

This has standard equation

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$$

The eccentricity, e , is defined by

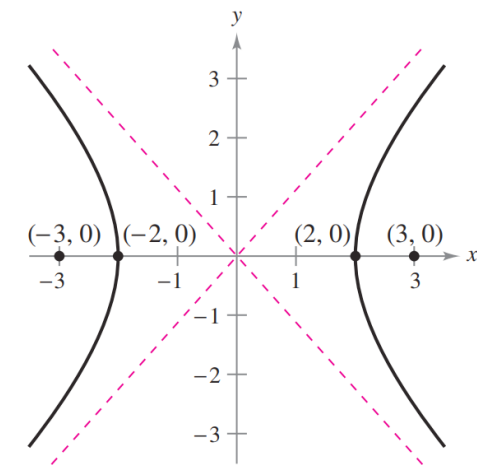
$$e^2 = 1 + \frac{b^2}{a^2} \quad (e > 1)$$

Note the change in sign when compared to the equivalent expressions for the ellipse. The lines $y = \pm b/a x$ are asymptotes to the hyperbola (these are the lines to which each branch of the hyperbola approach as $x \rightarrow \pm\infty$).

If light is emitted from one focus then on hitting the hyperbolic curve it is reflected in such a way as to appear to be coming from the other focus. See Figure 4(b). The hyperbola has fewer uses in applications than the other conic sections and so we will not dwell here on its properties.

Example

Find the standard form of the equation of the hyperbola with foci at $(-3, 0)$ and $(3, 0)$ and vertices at $(-2, 0)$ and $(2, 0)$ as shown in figure below.



Solution: From the graph, $c = 3$ because the foci are three units from the center. Also, $a = 2$ because the vertices are two units from the center. So,

$$b^2 = c^2 - a^2 = 3^2 - 2^2 = 9 - 4 = 5.$$

Because the transverse axis is horizontal, the standard form of the equation is

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1. \quad (\text{Standard form, horizontal transverse axis})$$

Finally, substituting $a^2 = 2^2$ and $b^2 = (\sqrt{5})^2$, you have

$$\frac{x^2}{2^2} - \frac{y^2}{(\sqrt{5})^2} = 1. \quad (\text{Write in standard form.})$$

General Conics

The conics we have considered above - the ellipse, the parabola and the hyperbola - have all been presented in standard form:- their axes are parallel to either the x or y -axes. However, conics may be rotated to any angle with respect to the axes: they clearly remain conics, but what equations do they have? It can be shown that the equation of any conic, can be described by the quadratic expression

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$$

where A, B, C, D, E, F are constants.

If not all of A, B, C are zero the graph of this equation is

- i. an ellipse if $B^2 < 4AC$ (circle if $A = C$)

from some approximate trial solution, a useful algorithm will improve the solution until some predetermined convergence criterion is satisfied.

Unlike most of the iterative methods used for linear systems, having a good first-guess for the solution of a nonlinear system usually is crucial in determining the success of an iterative process. Such methods, among which the Newton-Raphson method is the most celebrated, are called local methods.

Thus far, general-purposed global methods are not available. For special classes of problem, such as solving systems of polynomials, important progress has recently been made. The homotopy method, having connections with differential equations, is one such approach.

Another difficulty often associated with solving nonlinear equations is the detection of existence of one or multiple solutions. A nonlinear set of equations may have no (real) solutions at all. Contrariwise, it may have more than one solution. Ideally, one should resort to some other means, such as the degree theory, to determine that theoretically a nonlinear equation does have a solution before numerically finding the approximate solution. Applying a numerical method blindfold has the danger of being misled to a wrong answer even though the method behaves nicely.

The Newton-Raphson Method

The Newton Raphson method is for solving equations of the form $f(x) = 0$. We make an initial guess for the root we are trying to find, and we call this initial guess x_0 . The sequence $x_0, x_1, x_2, x_3, \dots$ generated in the manner described below should converge to the exact root.

Tangent line to a plane curve at a given point is the straight line that “just touches” the curve at that point.

To implement it analytically we need a formula for each approximation in terms of the previous one, i.e. we need x_{n+1} in terms of x_n . The equation of the **tangent line** to the graph $y = f(x)$ at the point $(x_0, f(x_0))$ is

$$y - f(x_0) = f'(x_0)(x - x_0)$$

The tangent line intersects the x-axis when $y = 0$ and $x = x_1$, so

$$-f(x_0) = f'(x_0)(x_1 - x_0)$$

Solving this for x_1 gives

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

and, more generally

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

You should memorize the above formula. Its application to solving equations of the form $f(x) = 0$, as we now demonstrate, is called the Newton Raphson method. It is guaranteed to converge if the initial guess x_0 is close enough, but it is hard to make a clear statement about what we mean by ‘close enough’ because this is highly problem specific. A sketch of the graph of $f(x)$ can help us decide on an appropriate initial guess x_0 for a particular problem.

Example

Let us solve $x^3 - x - 1 = 0$ for x .

In this case $f(x) = x^3 - x - 1$, so $f'(x) = 3x^2 - 1$. So the recursion formula (1) becomes

$$x_{n+1} = x_n - \frac{(x_n^3 - x_n - 1)}{3x_n^2 - 1}$$

Need to decide on an appropriate initial guess x_0 for this problem. A rough graph can help. Note that $f(1) = -1 < 0$ and $f(2) = 5 > 0$. Therefore, a root of $f(x) = 0$ must exist between 1 and 2. Let us take $x_0 = 1$ as our initial guess. Then

$$x_1 = x_0 - \frac{(x_0^3 - x_0 - 1)}{3x_0^2 - 1}$$

and with $x_0 = 1$ we get $x_1 = 1.5$.

Now

$$x_2 = x_1 - \frac{(x_1^3 - x_1 - 1)}{3x_1^2 - 1}$$

and with $x_1 = 1.5$ we get $x_2 = 1.34783$. For the next stage,

$$x_3 = x_2 - \frac{(x_2^3 - x_2 - 1)}{3x_2^2 - 1}$$

and with the value just found for x_2 , we find $x_3 = 1.32520$.

Carrying on, we find that $x_4 = 1.32472$, $x_5 = 1.32472$, etc. We can stop when the

digits stop changing to the required degree of accuracy. We conclude that the root is 1.32472 to 5 decimal places.

The Broyden Method

One of the drawbacks of using Newton's Method to solve a system of nonlinear equations $g(x) = 0$ is the computational expense that must be incurred during each iteration to evaluate the partial derivatives of g at $x^{(k)}$, and then solve a system of linear equations involving the resulting Jacobian matrix. The **algorithm** does not facilitate the re-use of data from previous iterations, and in some cases evaluation of the partial derivatives can be unnecessarily costly.

Algorithm is a procedure or formula for solving a problem, based on conducting a sequence of specified actions.

An alternative is to modify Newton's Method so that approximate partial derivatives are used, since the slightly slower convergence resulting from such an approximation is offset by the improved efficiency of each iteration. However, simply replacing the analytical Jacobian matrix $J_g(x)$ of g with a matrix consisting of finite difference approximations of the partial derivatives does not do much to reduce the cost of each iteration, because the cost of solving the system of linear equations is unchanged.

However, because $Jg(x)$ consists of the partial derivatives evaluated at an element of a convergent sequence, intuitively Jacobian matrices from consecutive iterations are “near” one another in some sense, which suggests that it should be possible to cheaply update an approximate Jacobian matrix from iteration to iteration, in such a way that the inverse of the Jacobian matrix, which is what is really needed during each Newton iteration, can be updated efficiently as well.

This is the case when an $n \times n$ matrix B has the form $B = A + u \otimes v$,

where u and v are given vectors in $\mathbb{R}^n$, and $u \otimes v$ is the outer product of u and v , defined by

$$u \otimes v = uv^T = \begin{bmatrix} u_1 v_1 & u_1 v_2 & \cdots & u_1 v_n \\ u_2 v_1 & u_2 v_2 & \cdots & u_2 v_n \\ \vdots & \ddots & & \vdots \\ u_n v_1 & u_n v_2 & \cdots & u_n v_n \end{bmatrix}$$

This modification of A to obtain B is called a rank-one update. This is because $u \otimes v$ has rank one, since every column

of $\mathbf{u} \otimes \mathbf{v}$ is a scalar multiple of $\mathbf{u}$. To obtain B^{-1} from A^{-1} , we note that if $A\mathbf{x} = \mathbf{u}$,

then

$$B\mathbf{x} = (A + \mathbf{u} \otimes \mathbf{v})\mathbf{x} = A\mathbf{x} + \mathbf{u}\mathbf{v}^T\mathbf{x} = \mathbf{u} + \mathbf{u}(\mathbf{v} \cdot \mathbf{x}) = (1 + \mathbf{v} \cdot \mathbf{x})\mathbf{u},$$

which yields

$$B^{-1}\mathbf{u} = \frac{1}{1 + \mathbf{v} \cdot A^{-1}\mathbf{u}} A^{-1}\mathbf{u}.$$

On the other hand, if $\mathbf{x}$ is such that $\mathbf{v} \cdot A^{-1}\mathbf{x} = 0$, then

$$BA^{-1}\mathbf{x} = (A + \mathbf{u} \otimes \mathbf{v})A^{-1}\mathbf{x} = AA^{-1}\mathbf{x} + \mathbf{u}\mathbf{v}^T A^{-1}\mathbf{x} = \mathbf{x} + \mathbf{u}(\mathbf{v} \cdot A^{-1}\mathbf{x}) = \mathbf{x},$$

which yields

$$B^{-1}\mathbf{x} = A^{-1}\mathbf{x}.$$

This takes us to the following more general problem: given a matrix C , we wish to construct a matrix D such that the following conditions are satisfied:

- $D\mathbf{w} = \mathbf{z}$, for given vectors $\mathbf{w}$ and $\mathbf{z}$
- $D\mathbf{y} = C\mathbf{y}$, if $\mathbf{y}$ is orthogonal to a given vector $\mathbf{g}$.

In our application, $C = A^{-1}$, $D = B^{-1}$, $\mathbf{w} = \mathbf{u}$, $\mathbf{z} = 1/(1 + \mathbf{v} \cdot A^{-1}\mathbf{u})A^{-1}\mathbf{u}$, and $\mathbf{g} = A^{-T}\mathbf{v}$.

To solve this problem, we set

$$D = C + \frac{(\mathbf{z} - C\mathbf{w}) \otimes \mathbf{g}}{\mathbf{g} \cdot \mathbf{w}}.$$

Then, if $\mathbf{g} \cdot \mathbf{y} = 0$, the second term in the definition of D vanishes, and we obtain $D\mathbf{y} = C\mathbf{y}$, but in computing $D\mathbf{w}$, we obtain factors of $\mathbf{g} \cdot \mathbf{w}$ in the numerator and denominator that cancel, which yields

$$D\mathbf{w} = C\mathbf{w} + (\mathbf{z} - C\mathbf{w}) = \mathbf{z}.$$

Applying this definition of D , we obtain

$$B^{-1} = A^{-1} + \frac{\left[\left(\frac{1}{1 + \mathbf{v} \cdot A^{-1}\mathbf{u}} A^{-1}\mathbf{u} - A^{-1}\mathbf{u} \right) \otimes \mathbf{v} \right] A^{-1}}{\mathbf{v} \cdot A^{-1}\mathbf{u}} = A^{-1} - \frac{A^{-1}(\mathbf{u} \otimes \mathbf{v})A^{-1}}{1 + \mathbf{v} \cdot A^{-1}\mathbf{u}}.$$

This formula for the inverse of a rank-one update is known as the Sherman-Morrison Formula.

We now return to the problem of approximating the Jacobian of $\mathbf{g}$, and efficiently obtaining its inverse, at each iterate $\mathbf{x}^{(k)}$. We begin with an exact Jacobian, $D_0 = J\mathbf{g}(\mathbf{x}^{(0)})$, and use D_0 to compute the first iterate, $\mathbf{x}^{(1)}$, using Newton's Method as follows:

$$\mathbf{d}^{(0)} = -D_0^{-1}\mathbf{g}(\mathbf{x}^{(0)}), \quad \mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \mathbf{d}^{(0)}.$$

Then, we use the approximation

$$g'(x_1) \approx \frac{g(x_1) - g(x_0)}{x_1 - x_0}.$$

Generalizing this approach to a system of equations, we seek an approximation D_1 to $J_g(\mathbf{x}^{(1)})$ that has these properties:

- $D_1(\mathbf{x}^{(1)} - \mathbf{x}^{(0)}) = \mathbf{g}(\mathbf{x}^{(1)}) - \mathbf{g}(\mathbf{x}^{(0)})$ (the Secant Condition)
- If $\mathbf{z} \cdot (\mathbf{x}^{(1)} - \mathbf{x}^{(0)}) = 0$, then $D_1 \mathbf{z} = J_g(\mathbf{x}^{(0)}) \mathbf{z} = D^0 \mathbf{z}$.

It follows from previous discussion that

$$D_1 = D_0 + \frac{\mathbf{y}_0 - D_0 \mathbf{d}^{(0)}}{\mathbf{d}^{(0)} \cdot \mathbf{d}^{(0)}} \otimes \mathbf{d}^{(0)},$$

where

$$\mathbf{d}^{(0)} = \mathbf{x}^{(1)} - \mathbf{x}^{(0)}, \quad \mathbf{y}^{(0)} = \mathbf{g}(\mathbf{x}^{(1)}) - \mathbf{g}(\mathbf{x}^{(0)}).$$

However, it can be shown that $\mathbf{y}_0 - D_0 \mathbf{d}^{(0)} = \mathbf{g}(\mathbf{x}^{(1)})$, which yields the simplified formula

$$D_1 = D_0 + \frac{1}{\mathbf{d}^{(0)} \cdot \mathbf{d}^{(0)}} \mathbf{g}(\mathbf{x}^{(1)}) \otimes \mathbf{d}^{(0)}.$$

Once we have computed D_0^{-1} , we can apply the Sherman-Morrison formula to obtain

$$\begin{aligned} D_1^{-1} &= D_0^{-1} - \frac{D_0^{-1} \left(\frac{1}{\mathbf{d}^{(0)} \cdot \mathbf{d}^{(0)}} \mathbf{g}(\mathbf{x}^{(1)}) \otimes \mathbf{d}^{(0)} \right) D_0^{-1}}{1 + \mathbf{d}^{(0)} \cdot D_0^{-1} \left(\frac{1}{\mathbf{d}^{(0)} \cdot \mathbf{d}^{(0)}} \mathbf{g}(\mathbf{x}^{(1)}) \right)} \\ &= D_0^{-1} - \frac{D_0^{-1} (\mathbf{g}(\mathbf{x}^{(1)}) \otimes \mathbf{d}^{(0)}) D_0^{-1}}{\mathbf{d}^{(0)} \cdot \mathbf{d}^{(0)} + \mathbf{d}^{(0)} \cdot D_0^{-1} \mathbf{g}(\mathbf{x}^{(1)})} \\ &= D_0^{-1} - \frac{(\mathbf{u}^{(0)} \otimes \mathbf{d}^{(0)}) D_0^{-1}}{\mathbf{d}^{(0)} \cdot (\mathbf{d}^{(0)} + \mathbf{u}^{(0)})}, \end{aligned}$$

where $\mathbf{u}^{(0)} = D_0^{-1} \mathbf{g}(\mathbf{x}^{(1)})$. Then, as D_1 is an approximation to $J_g(\mathbf{x}^{(1)})$, we can obtain our next iterate $\mathbf{x}^{(2)}$ as follows:

$$D_1 \mathbf{d}^{(1)} = -\mathbf{g}(\mathbf{x}^{(1)}), \quad \mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \mathbf{d}^{(1)}.$$

Repeating this process, we obtain the following algorithm, which is known as Broyden's Method:

Choose $\mathbf{x}^{(0)}$

$$D_0 = J_g(\mathbf{x}^{(0)})$$

$$\mathbf{d}^{(0)} = -D_0^{-1} \mathbf{g}(\mathbf{x}^{(0)})$$

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \mathbf{d}^{(0)}$$

$$k = 0$$

```

while not converged do
     $\mathbf{u}^{(k)} = D_k^{-1} \mathbf{g}(\mathbf{x}^{(k+1)})$ 
     $c_k = \mathbf{d}^{(k)} \cdot (\mathbf{d}^{(k)} + \mathbf{u}^{(k)})$ 
     $D_{k+1}^{-1} = D_k^{-1} - \frac{1}{c_k} [\mathbf{u}^{(k)} \otimes \mathbf{d}^{(k)}] D_k^{-1}$ 
     $k = k + 1$ 

     $\mathbf{d}^{(k)} = -D_k^{-1} \mathbf{g}(\mathbf{x}^{(k)})$ 
     $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{d}^{(k)}$ 
end

```

Note that it is not necessary to compute D_k for $k \geq 1$; only D_k^{-1} is needed. It follows that no systems of linear equations need to be solved during an iteration; only matrix-vector multiplications are required, thus saving an order of magnitude of computational effort during each iteration compared to Newton's Method.

Example

We consider the system of equations $\mathbf{g}(\mathbf{x}) = 0$, where

$$\mathbf{g}(x, y, z) = \begin{bmatrix} x^2 + y^2 + z^2 - 3 \\ x^2 + y^2 - z - 1 \\ x + y + z - 3 \end{bmatrix}.$$

We will begin with one step of Newton's Method to solve this system of equations, with initial guess $\mathbf{x}^{(0)} = (x^{(0)}, y^{(0)}, z^{(0)}) = (1, 0, 1)$.

As computed in a previous example,

$$J_{\mathbf{g}}(x, y, z) = \begin{bmatrix} 2x & 2y & 2z \\ 2x & 2y & -1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Therefore, the Newton iterate $\mathbf{x}^{(1)}$ is obtained by solving the system of equations $J_{\mathbf{g}}(\mathbf{x}^{(0)})(\mathbf{x}^{(1)} - \mathbf{x}^{(0)}) = -\mathbf{g}(\mathbf{x}^{(0)})$,

or, equivalently, $D_0 \mathbf{d}^{(0)} = -\mathbf{g}(\mathbf{x}^{(0)})$ where

$$D_0 = J_{\mathbf{g}}(\mathbf{x}^{(0)}) = \begin{bmatrix} 2x^{(0)} & 2y^{(0)} & 2z^{(0)} \\ 2x^{(0)} & 2y^{(0)} & -1 \\ 1 & 1 & 1 \end{bmatrix}, \quad \mathbf{d}^{(0)} = \begin{bmatrix} x^{(1)} - x^{(0)} \\ y^{(1)} - y^{(0)} \\ z^{(1)} - z^{(0)} \end{bmatrix},$$

$$\mathbf{g}(\mathbf{x}^{(0)}) = \begin{bmatrix} (x^{(0)})^2 + (y^{(0)})^2 + (z^{(0)})^2 - 3 \\ (x^{(0)})^2 + (y^{(0)})^2 - z^{(0)} - 1 \\ x^{(0)} + y^{(0)} + z^{(0)} - 3 \end{bmatrix}.$$

Substituting $(x^{(0)}, y^{(0)}, z^{(0)}) = (1, 0, 1)$ yields the system

$$\begin{bmatrix} 2 & 0 & 2 \\ 2 & 0 & -1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x^{(1)} - 1 \\ y^{(1)} \\ z^{(1)} - 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

This system has the solution $\mathbf{d}^{(0)} = (1/2, 1/2, 0)$ which yields $\mathbf{x}^{(1)} = (3/2, 1/2, 1)$.

Now, instead of computing $\mathbf{Jg}(\mathbf{x}^{(1)})$, we compute

$$\begin{aligned} D_1 &= D_0 + \frac{1}{\mathbf{d}^{(0)} \cdot \mathbf{d}^{(0)}} \mathbf{g}(\mathbf{x}^{(1)}) \otimes \mathbf{d}^{(0)} \\ &= \begin{bmatrix} 2 & 0 & 2 \\ 2 & 0 & -1 \\ 1 & 1 & 1 \end{bmatrix} + \frac{1}{1/2} \begin{bmatrix} 1/2 \\ 1/2 \\ 0 \end{bmatrix} \otimes \begin{bmatrix} 1/2 \\ 1/2 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} 2 & 0 & 2 \\ 2 & 0 & -1 \\ 1 & 1 & 1 \end{bmatrix} + 2 \begin{bmatrix} 1/4 & 1/4 & 0 \\ 1/4 & 1/4 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 5/2 & 1/2 & 2 \\ 5/2 & 1/2 & -1 \\ 1 & 1 & 1 \end{bmatrix}. \end{aligned}$$

This yields the system

$$\begin{bmatrix} 5/2 & 1/2 & 2 \\ 5/2 & 1/2 & -1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x^{(2)} - \frac{3}{2} \\ y^{(2)} - \frac{1}{2} \\ z^{(2)} - 1 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2} \\ 0 \end{bmatrix},$$

which has the solution $\mathbf{x}^{(2)} = (5/4, 3/4, 1)$. This happens to be the same second iterate computed by Newton's Method, though this is not the case for a general function $\mathbf{g}(\mathbf{x})$.

Sturm Sequences

Boundary condition is a condition that is required to be satisfied at all or part of the boundary of a region in which a set of differential conditions is to be solved.

The matrix elements of each transfer matrix are polynomials with respect to the eigenvalue λ , beginning with the basic matrix whose 11 element is of the first degree in λ , which is absent from the remaining elements. Multiplying transfer matrices raises the degree of the matrix elements, the 11 element always retaining the highest degree - the sum of the degrees of the factors - when the matrices have the form we are using. **Boundary conditions** eventually dictate that some combination of matrix elements gives zero, which means the vanishment of a polynomial whose degree is the length of the string, whose expression is evidently the characteristic equation of A .

If the boundary condition is independent of the length of the string, there is a series of characteristic equations, each having a similar structure yet differing by the degree of the polynomial involved. Moreover, the peculiar structure of the matrix M , arising from the tautology which was incorporated in its definition, interlaces the polynomials used at each stage. It is worth seeing how this is done, in the simple case where the boundary condition is that there be a terminal node.

$$M_{i+1} = \begin{bmatrix} a/ & -b & -c \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} P_{11}^i & P_{12}^i \\ P_{21}^i & P_{22}^i \end{bmatrix} \\ = \begin{bmatrix} (a/ - b)P_{11}^i & (a/ - b)P_{21}^i \\ P_{11}^i & P_{12}^i \end{bmatrix}$$

If the initial vector were (10), implying $x_1 = 0, x_2 = 1$ (any factor will do), and the final vector were (0, 1) implying that x_{i+1} were also zero, then the 11 element of M_{i+1} would vanish, or

$$(a/ - b)P_{11}^i - cP_{21}^i = 0,$$

$$a/ = b + c \frac{P_{21}^i}{P_{11}^i}$$

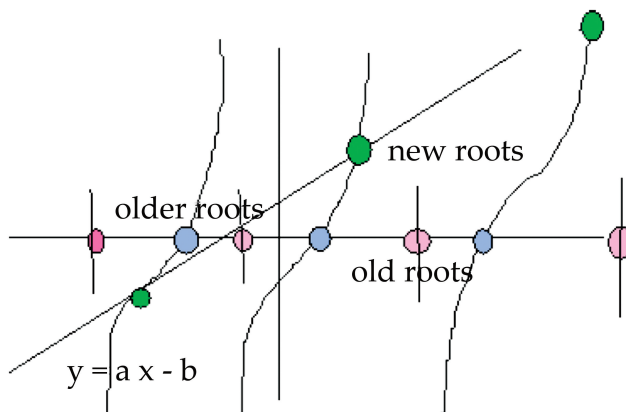


Figure 5: Interleaving of eigenvalues.

The new roots lie on a line determined by the diagonal element of A with slope likewise determined by A - the ratio of its diagonal to off-diagonal elements. The roots also lie on the quotient of two polynomials; the zeroes of the denominator are just the roots of the next smaller A , the roots of the numerator are those of the second previous A . Not both polynomials have roots in the same place because of the limit on the maximal thickness of a node. Thus zeros and poles of the quotient alternate, and the new roots lie between the poles so they are also separated. Thus all roots are distinct, and interleaved; this is called a Sturm sequence.

There are interesting details in the graph which has been sketched. The derivative of the quotient is nonzero and always has a consistent sign, making it monotone within the panels delimited by its poles. Consequently the intersecting line crosses

Eigenvalues are a special set of scalars associated with a linear system of equations (i.e., a matrix equation) that are sometimes also known as characteristic roots, characteristic values, proper values, or latent roots.

the quotient just once in each panel, assuring the interleaving of roots and keeping them real. The root cause of all this is the sign-symmetry of the matrix A; were it sign-antisymmetric, roots would have to run along in conjugate pairs. In fact, the interleaving property was already proved on more general grounds by Ledermann, so that we mainly have a more graphic presentation of the result.

The value where the line crosses the axis is the diagonal element of the bordering matrix, which will not be an **eigenvalue** although it will necessarily sit between some pair of old roots. It acts as a center of readjustment, crowding new eigenvalues toward the edges of the panels defined by the old eigenvalues while distancing them from still older eigenvalues. Eventually there will be very little change in eigenvalues far from that diagonal element since the line will intersect the quotient near its asymptotes, but finally one of the new eigenvalues will lie outside the range established previously.

Bairstow's Method

A real polynomial may have complex conjugate roots. In order to find the complex roots, most methods would have to begin at a complex starting point and be carried out in complex arithmetic. Bairstow's method avoids complex arithmetic.

The roots of a quadratic polynomial

$$d(x) = x^2 - rx - q$$

are obviously known. For a polynomial $p(x)$, we write

$$p(x) = p_1(x)(x^2 - rx - q) + Ax + B. \quad (2)$$

The coefficients of the remainder depends upon r and q . The idea of Bairstow's method is to determine r and q so that

$$A(r, q) = 0 \quad (3)$$

$$B(r, q) = 0 \quad (4)$$

Applying Newton's method to (3), we need to compute

$$\begin{bmatrix} r_{i+1} \\ q_{i+1} \end{bmatrix} = \begin{bmatrix} r_i \\ q_i \end{bmatrix} - \begin{bmatrix} \frac{\partial A}{\partial r} & \frac{\partial A}{\partial q} \\ \frac{\partial B}{\partial r} & \frac{\partial B}{\partial q} \end{bmatrix}_{(r_i, q_i)}^{-1} \begin{bmatrix} A(r_1, q_i) \\ B(r_1, q_i) \end{bmatrix}.$$

Upon differentiating, we observe that

$$0 \equiv \frac{\partial p(x)}{\partial r} = \frac{\partial p_1}{\partial r}(x)(x^2 - rx - q) - p_1(x)x + \frac{\partial A}{\partial r}x + \frac{\partial B}{\partial r} \quad (5)$$

$$0 \equiv \frac{\partial p(x)}{\partial q} = \frac{\partial p_1}{\partial q}(x)(x^2 - rx - q) - p_1(x) + \frac{\partial A}{\partial q}x + \frac{\partial B}{\partial q}. \quad (6)$$

Let $p_1(x)$ be further divided by $d(x)$ and denote

$$p_1(x) = p_2(x)(x^2 - rx - q) + \tilde{A}x + \tilde{B}.$$

Assuming the two roots x_0 and x_1 , of $d(x)$ are distinct, it follows that

$$p_1(x_i) = \tilde{A}x_i + \tilde{B}.$$

Substituting (6) we obtain

$$-x_i(\tilde{A}x_i + \tilde{B}) + \frac{\partial A}{\partial r}x_i + \frac{\partial B}{\partial r} = 0$$

$$-(\tilde{A}x_i + \tilde{B}) + \frac{\partial A}{\partial q}x_i + \frac{\partial B}{\partial q} = 0.$$

Form the second equations we have (since $x_0 \neq x_1$)

$$\frac{\partial A}{\partial q} = \tilde{A}, \quad \frac{\partial B}{\partial q} = \tilde{B}.$$

Form the first equations, we have

$$-\tilde{A}(rx_i + q) + x_i\left(\frac{\partial A}{\partial r} - \tilde{B}\right) + \frac{\partial B}{\partial r} = 0.$$

Therefore, we know

$$\frac{\partial A}{\partial r} = \tilde{B} = \tilde{A}r$$

$$\frac{\partial B}{\partial r} = \tilde{A}q.$$

The values of A , B can be obtained without difficulty: Suppose

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

$$p_1(x) = b_{n-2} x^{n-2} + \dots + b_0.$$

By comparing coefficients of (6.41), we obtain

$$b_{n-2} = a_n,$$

$$b_{n-3} = b_{n-2}r + a_{n-1},$$

$$b_{n-k} = b_{n-k+2}q + b_{n-k+1}r + a_{n-k+2}$$

$$A = b_1 q + b_0 r + a_1$$

$$B = b_0 q + a_0.$$

Similarly, by using (2), we can obtain the values of $\tilde{A}$ and $\tilde{B}$.

Variable is a symbol that represents a quantity in a mathematical expression, as used in many sciences.

Solving Systems of Nonlinear Equations

A system of equations where at least one equation is not linear is called a nonlinear system. There are several ways to solve systems of nonlinear equations:

- Substitution
- Elimination
- Using a Combination of methods
- Using absolute value

By Substitution

Solve the following nonlinear equations:

$$x^2 + y = 6$$

$$x - y = 14$$

To solve this system we first need to isolate one of the **variables**. In this example, we can use the second equation to solve for y ,

$$x - y = 14$$

$$y = x - 14$$

Now we can substitute this value of y in the second equation:

$$x^2 + y = 6$$

$$x^2 + (x - 14) = 6$$

$$x^2 + x - 14 - 6 = 0$$

$$x^2 + x - 20 = 0$$

By factoring we find two possible values of x :

$$(x + 5)(x - 4) = 0$$

$$x = -5 \text{ or } x = 4$$

From the first equation, we can use the value of x to find a value for y

$$\text{For } x = -5; y = x - 14 = -19$$

$$\text{For } x = 4; y = x - 14 = -10$$

As a result, the solution set for the nonlinear system is $\{(4, -10), (-5, -19)\}$

By Elimination

Solve the following nonlinear equations:

$$x^2 + y^2 = 34$$

$$x^2 - 2y^2 = 7$$

To solve this system we multiply the first equation by 2

$$(x^2 + y^2 = 34) \times 2$$

$$2x^2 + 2y^2 = 68$$

We then add that product to the second equation

$$2x^2 + 2y^2 = 68$$

$$+ x^2 - 2y^2 = 7$$

$$3x^2 = 75$$

Now we can solve for x

$$3x^2 = 75$$

$$x^2 = 25$$

$$x = 5$$

We can substitute this value of x into the first equation to find all possible values for y . Since we are substituting into a square, $x = 5$ and $x = -5$ will give us the same value:

$$y^2 = 34 - x^2$$

$$y^2 = 34 - 25$$

$$y^2 = 9$$

$$y = 3$$

The solution set for the nonlinear system is $\{(5,3), (5,-3), (-5,3), (-5,-3)\}$

Using a Combination of Methods

Solve the following nonlinear equations:

$$x^2 - 2xy + y^2 = 3$$

$$x^2 + xy + y^2 = 12$$

First we use the elimination method to find a value for y . We can multiply the first equation by -1 :

$$-x^2 + 2xy - y^2 = -3$$

Then we can add both equations

$$-x^2 + 2xy - y^2 = -3$$

$$x^2 + xy + y^2 = 12$$

$$3xy = 12$$

$$y = 4/x$$

Now substitute this value of y in the first equation:

$$x^2 + x(4/x) + (4/x)^2 = 12$$

$$x^2 + 4 + 16/x^2 = 12$$

Multiply both sides by x^2 :

$$(x^2)^2 + 4x^2 + 16 = 12x^2$$

$$(x^2)^2 - 8x^2 + 16 = 0$$

Using the quadratic equation, we can solve for x :

$$(x^2 - 4)(x^2 - 4) = 0$$

$$x^2 = 4$$

$$x = 2$$

We can then use that result to find the value for y :

$$y = 4/x$$

$$y = 2$$

The solution set for the nonlinear system is $\{(2,2), (-2,-2)\}$.

Using Absolute Value

Solve the following nonlinear equations:

$$x^2 + y^2 = 25$$

$$|x| + y = 5$$

First we use the substitution method and solve for the absolute value of x :

$$|x| + y = 5$$

$$|x| = 5 - y$$

We know from the definition of absolute value that $|x| \geq 0$. So for all x , $5 - y \geq 0$, and that can be rewritten as $5 \geq y$. In the first equation $|x|^2$ is the same as x^2 ; therefore, we can use substitution:

$$x^2 + y^2 = 25$$

$$(5 - y)^2 + y^2 = 25$$

$$(25 - 10y + y^2) + y^2 = 25$$

$$2y^2 - 10y = 0$$

$$2y(y - 5) = 0$$

We can now solve for y :

The behavior of a nonlinear system is described in mathematics by a nonlinear system of equations, which is a set of simultaneous equations in which the unknowns (or the unknown functions in the case of differential equations) appear as variables of a polynomial of degree higher than one or in the argument of a function which is not a polynomial of degree one.

$$y = 0 \text{ or } y = 5$$

We can use both of these values of y in the first equation:

$$\text{For } y = 0$$

$$x^2 + (0)^2 = 25$$

$$x^2 = 25$$

$$x = 5$$

$$\text{For } y = 5$$

$$x^2 + (5)^2 = 25$$

$$x^2 + 25 = 25$$

$$x = 0$$

The solution set for the nonlinear system is $\{(5,0), (0,5), (-5,0)\}$.

CONCEPT: CONIC SYSTEMS

It seems like there might be a slight confusion in your question. “CONIC system” isn’t a widely recognized term in mathematics or any specific field that I’m aware of. However, I can provide information on conic sections and systems if that’s what you’re referring to.

Conic Sections: Conic sections are the curves that result from the intersection of a plane and a cone. There are several types of conic sections: circles, ellipses, parabolas, and hyperbolas.

- **Circle:** A circle is formed when the plane intersects the cone perpendicular to its axis.
- **Ellipse:** An ellipse is formed when the plane intersects the cone at an angle, but not parallel to the base or perpendicular to the axis.
- **Parabola:** A parabola is formed when the plane intersects the cone parallel to one side of the cone.
- **Hyperbola:** A hyperbola is formed when the plane intersects the cone parallel to the axis but not touching the cone.

Conic Systems: If by “CONIC system” you’re referring to a system of equations involving conic sections, you might be talking about equations that describe the geometry of conic sections. Each type of conic section can be represented by an equation involving variables x and y .

For example:

- The equation of a circle centered at (h, k) with radius r is: $(x - h)^2 + (y - k)^2 = r^2$.
- The equation of an ellipse centered at (h, k) is: $((x - h)^2 / a^2) + ((y - k)^2 / b^2) = 1$, where a and b are the lengths of the semi-major and semi-minor axes, respectively.
- The equation of a parabola can have different forms depending on its orientation and vertex.
- The equation of a hyperbola can also have different forms depending on its orientation and center.

These equations can be used to analyze and solve problems related to conic sections in various fields like geometry, physics, and engineering. If you have a specific context in mind or if “CONIC system” refers to something more specific, please provide additional details so I can offer a more accurate response.

SUMMARY

- A conic section, conic or a quadratic curve is a curve obtained from a cone's surface intersecting a plane.
- A conic section is the intersection of a plane and a double right circular cone. By changing the angle and location of the intersection, we can produce different types of conics.
- Conic sections are important because they model important physical processes in nature. It can be shown that anybody under the influence of an inverse square law force must have a trajectory of one of the conic sections.
- Mathematicians, engineers and scientists encounter numerous functions in their work: polynomials, trigonometric and hyperbolic functions amongst them. However, throughout the history of science one group of functions, the conics, arise time and time again not only in the development of mathematical theory but also in practical applications.
- The matrix elements of each transfer matrix are polynomials with respect to the eigenvalue l , beginning with the basic matrix whose 11 element is of the first degree in l , which is absent from the remaining elements.
- The new roots lie on a line determined by the diagonal element of A with slope likewise determined by A - the ratio of its diagonal to off-diagonal elements. The roots also lie on the quotient of two polynomials; the zeroes of the denominator are just the roots of the next smaller A , the roots of the numerator are those of the second previous A .

MULTIPLE CHOICE QUESTIONS

1. Find the equation of circle with center at origin and radius 5 units.
 - a. $x^2+y^2=25$
 - b. $x^2+y^2=5$
 - c. $x^2=25$
 - d. $y^2=25$
2. Find the equation of circle with center at (2, 5) and radius 5 units.
 - a. $x^2+y^2+4x-10y+4=0$
 - b. $x^2+y^2-4x-10y+4=0$
 - c. $x^2+y^2+4x+10y+4=0$
 - d. $x^2+y^2+4x-10y-4=0$
3. Find the center of the circle with equation $x^2+y^2-4x-10y+4=0$.
 - a. (-2, 5)
 - b. (-2, -5)
 - c. (2, -5)
 - d. (2, 5)
4. Find the radius of the circle with equation $x^2+y^2-4x-10y+4=0$.
 - a. 25 units
 - b. 20 units
 - c. 5 units
 - d. 10 units
5. Find the equation of circle which pass through (5, 9) and center at (2, 5).
 - a. $x^2+y^2+4x-10y+4=0$
 - b. $x^2+y^2-4x-10y+4=0$
 - c. $x^2+y^2+4x+10y+4=0$
 - d. $x^2+y^2+4x-10y-4=0$

Answers

1. (a) 2. (b) 3. (d) 4. (c) 5. (b)

REVIEW QUESTIONS

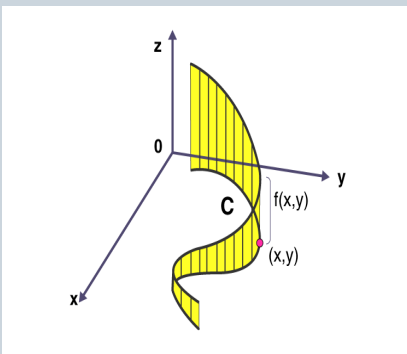
1. What are the four primary conic sections?
2. Find the center and radius of each of the following circles:
 - a. $x^2 + y^2 - 4x - 6y = -12$
 - b. $2x^2 + 2y^2 + 4x + 1 = 0$
3. Use Sturm's theorem to find the roots of $x^4 - 3x^2 + x - 2$.
4. The equation $9x^2 + 4y^2 - 36x + 24y - 1 = 0$ represents an ellipse. Find its center, the semi-major and semi-minor axes and the coordinate positions of the foci.
5. Find the equation of a circle of radius 3 which has its center at $(-1, 2.2)$
6. An asteroid has an elliptical orbit around the Sun. The major axis is of length 5×10^8 km. If the distance between the foci is 4×10^8 km find the equation of the orbit.

REFERENCES

1. Abbasbandy, S. 2003. Improving Newton-Raphson method for nonlinear equations by modified Adomian decomposition method. *Appl. Math. Comput.* 145: 887–893
2. Amat, S., Busquier, S. and Guitierrez, J.M. 2003. Geometric constructions of iterative functions to solve nonlinear equations. *J. Comput. Appl. Math.* 157: 197–205
3. Anton, H., I. Bivens, S. Davis. 2005. *Calculus*. 8th Edition. John Wiley & Sons.
4. Babajee, D.K.R., Dauhoo, M.Z., Darvishi, M.T., Karami, A. and Barati, A. 2010. Analysis of two Chebyshev-like third order methods free from second derivatives for solving systems of nonlinear equations. *J. Comput. Appl. Math.* 233: 2002–2012
5. Basu, D. 2008. From third to fourth order variant of Newton's method for simple roots. *Appl. Math. Comput.* 202: 886–892
6. Bi, W., Ren, H. and Wu, Q. 2008. New family of seventh-order methods for nonlinear equations. *Appl. Math. Comput.* 203: 408–412
7. Bi, W., Ren, H. and Wu, Q. 2009. Three-step iterative methods with eighth-order convergence for solving nonlinear equations. *J. Comput. Appl. Math.* 225: 105–112
8. Bi, W., Wu, Q. and Ren, H. 2009. A new family of eighth-order iterative methods for solving nonlinear equations. *Appl. Math. Comput.* 214: 236–245
9. Biazar, J. and Ghanbari, B. 2010. A new third-order family of nonlinear solvers for multiple roots. *Comput. Math. Appl.* 59: 3315–3319
10. Burden, R.L. and Faires, J.D. 2005. *Numerical Analysis*. Brooks/Cole Publishing Company, California
11. Chapra, S.C. and Canale, R.P. 1988. *Numerical Methods for Engineers*. McGrawHill Book Company, New York
12. Chun, C. 2005. Iterative methods improving Newton's method by the decomposition method. *Comput. Math. Appl.* 50: 1559–1568
13. Chun, C. 2007. A geometric construction of iterative functions of order three to solve nonlinear equations. *Comp. Math. Appl.* 53: 972–976
14. Chun, C. 2007. A one-parameter family of third-order methods to solve nonlinear equations. *Appl. Math. Comput.* 189: 126–130
15. Chun, C. 2007. Some improvements of Jarratt's method with sixth-order convergence. *Appl. Math. Comput.* 190: 1432–1437
16. D. A. Brannan, M. F. Esplen, J. J. Gray, *Geometry*, Cambridge University Press, 2002
17. Stroud, K.A. 2001. *Engineering Mathematics*. 5th Edition. PALGRAVE.

CHAPTER 8

LINE INTEGRALS



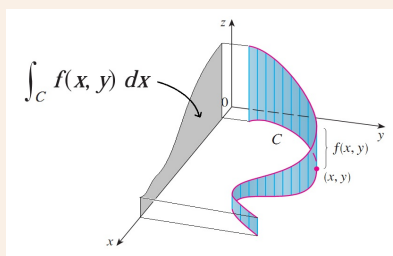
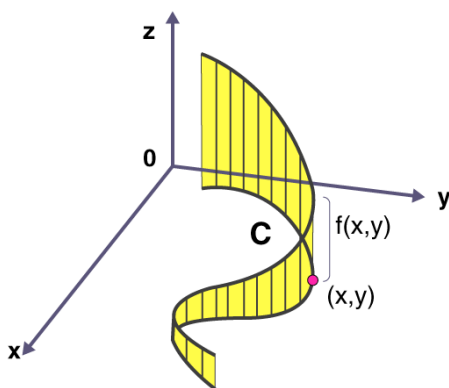
A line integral is represented as an integral in which a function is to be integrated along a curve. A line integral is also known as a path integral, curvilinear integral, or curve integral. Line integrals have several applications such as in electromagnetic, line integral is used to estimate the work done on a charged particle travelling along some curve in a force field defined by a vector field. In classical mechanics, line integral is used to compute the work performed on mass m moving in a gravitational field.

After studying this chapter, you will understand the core concepts of:

- line integral definition
- line integrals in green's theorem
- line integral in stokes's theorem
- line integrals in vector fields

INTRODUCTION

A line integral is integral in which the function to be integrated is determined along a curve in the coordinate system. The function which is to be integrated may be either a scalar field or a vector field. We can integrate a scalar-valued function or vector-valued function along a curve. The value of the line integral can be evaluated by adding all the values of points on the vector field.



CORE CONCEPT: LINE INTEGRAL DEFINITION

A line integral is an integral where the function to be integrated is evaluated along a curve. The terms path integral, curve integral, and curvilinear integral are also used; contour integral is used as well, although that is typically reserved for line integrals in the

complex plane.

Line integral is an integral in which the function to be integrated is evaluated along a curve. A line integral is also called the path integral or a curve integral or a curvilinear integral. We have so far integrated “over” intervals, areas, and volumes with single, double, and triple integrals. We now investigate integration over or “along” a curve—“line integrals” are really “curve integrals”.

As with other integrals, a geometric example may be easiest to understand. Consider the function $f = x + y$ and the parabola $y = x^2$ in the x - y plane, for $0 \leq x \leq 2$. Imagine that we extend the parabola up to the surface f , to form a curved wall or curtain, as in figure 1. What is the area of the surface thus formed? We already know one way to compute surface area, but here we take a different approach that is more useful for the problems to come.

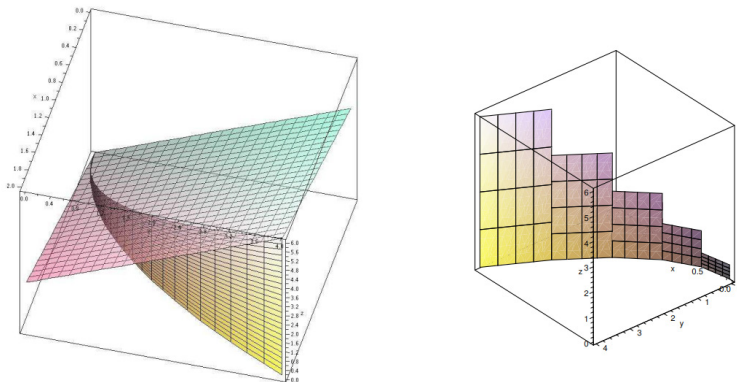


Figure 1. Approximating the area under a curve.

As usual, we start by thinking about how to approximate the area. We pick some points along the part of the parabola we're interested in, and connect adjacent points by straight lines; when the points are close together, the length of each line segment will be close to the length along the parabola. Using each line segment as the base of a rectangle, we choose the height to be the height of the surface f the line segment. If we add up the areas of these rectangles, we get an approximation to the desired area, and in the limit this sum turns into an integral.

Typically the curve is in vector form, or can easily be put in vector form; in this example we have $\mathbf{v}(t) = \langle t, t^2 \rangle$. The length of one of the straight line segments in the approximation is approximately $ds = |\mathbf{v}'| dt = \sqrt{1 + 4t^2} dt$, so the integral is

$$\int_0^2 f(t, t^2) \sqrt{1 + 4t^2} dt = \int_0^2 (t + t^2) \sqrt{1 + 4t^2} dt = \frac{167}{48} \sqrt{17} - \frac{1}{12} - \frac{1}{64} \ln(4 + \sqrt{17}).$$

This integral of a function along a curve C is often written in abbreviated form as

$$\int_C f(x, y) ds.$$

The work done by a force on an object is equal to the magnitude of the force times the distance the object moves; this assumes that the force is constant and in the direction of motion. We have already dealt with examples in which the force is not constant; now we are prepared to examine what happens when the force is not parallel to the direction of motion.

We have already examined the idea of components of force, in example 12.3.4: the component of a force F in the direction of a vector $\mathbf{v}$ is

$$\frac{\mathbf{F} \cdot \mathbf{v}}{|\mathbf{v}|^2} \mathbf{v},$$

the projection of F onto $\mathbf{v}$. The length of this vector, that is, the magnitude of the force in the direction of $\mathbf{v}$, is

$$\frac{\mathbf{F} \cdot \mathbf{v}}{|\mathbf{v}|},$$

the scalar projection of $\mathbf{F}$ onto $\mathbf{v}$. If an object moves subject to this (constant) force, in the direction of $\mathbf{v}$, over a distance equal to the length of $\mathbf{v}$, the work done is

$$\frac{\mathbf{F} \cdot \mathbf{v}}{|\mathbf{v}|} |\mathbf{v}| = \mathbf{F} \cdot \mathbf{v}.$$

Thus, work in the vector setting is still “force times distance”, except that “times” means “dot product”.

If the force varies from point to point, it is represented by a vector field $\mathbf{F}$; the displacement vector $\mathbf{v}$ may also change, as an object may follow a curving path in two or three dimensions. Suppose that the path of an object is given by a vector function $\mathbf{r}(t)$; at any point along the path, the (small) tangent vector $\mathbf{r}' \Delta t$ gives an approximation to its motion over a short time Δt , so the work done during that time is approximately $\mathbf{F} \cdot \mathbf{r}' \Delta t$; the total work over some time period is then

$$\int_{t_0}^{t_1} \mathbf{F} \cdot \mathbf{r}' dt.$$

It is useful to rewrite this in various ways at different times. We start with

$$\int_{t_0}^{t_1} \mathbf{F} \cdot \mathbf{r}' dt = \int_C \mathbf{F} \cdot d\mathbf{r},$$

abbreviating $\mathbf{r}' dt$ by $d\mathbf{r}$. Or we can write

$$\int_{t_0}^{t_1} \mathbf{F} \cdot \mathbf{r}' dt = \int_{t_0}^{t_1} \mathbf{F} \cdot \frac{\mathbf{r}'}{|\mathbf{r}'|} |\mathbf{r}'| dt = \int_{t_0}^{t_1} \mathbf{F} \cdot \mathbf{T} |\mathbf{r}'| dt = \int_C \mathbf{F} \cdot \mathbf{T} ds,$$

using the unit tangent vector $\mathbf{T}$, abbreviating $|\mathbf{r}'| dt$ as ds , and indicating the path of the object by C . In other words, work is computed using a particular line integral of the form we have considered. Alternately, we sometimes write

$$\begin{aligned} \int_C \mathbf{F} \cdot \mathbf{r}' dt &= \int_C \langle f, g, h \rangle \cdot \langle x', y', z' \rangle dt = \int_C \left(f \frac{dx}{dt} + g \frac{dy}{dt} + h \frac{dz}{dt} \right) dt \\ &= \int_C f dx + g dy + h dz = \int_C f dx + \int_C g dy + \int_C h dz, \end{aligned}$$

and similarly for two dimensions, leaving out references to z .

Arclength

If a curve C is described by the parametric equations

$$x = x(t) \quad \text{and} \quad y = y(t) \quad \text{for} \quad a \leq t \leq b$$

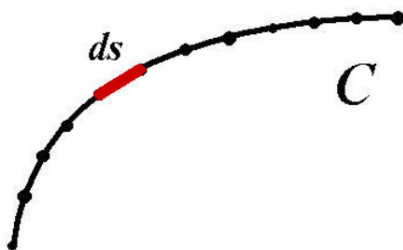
then the length along curve C is given by:

$$\text{Length}(C) = \int_a^b \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2} dt$$

The equation is simple enough to use. For example, to find the length along the segment of the circle $x = 2 \cos t$, $y = 2 \sin t$ for $0 \leq t \leq \frac{\pi}{8}$ we just calculate:

$$\text{Length} = \int_0^{\pi/8} \sqrt{(x'(t))^2 + (y'(t))^2} dt = \int_0^{\pi/8} \sqrt{4 \sin^2 t + 4 \cos^2 t} dt = \frac{\pi}{4}$$

The expression $\int_a^b \sqrt{(x'(t))^2 + (y'(t))^2} dt$ can be understood with either a geometric or physical interpretation. The usual geometric interpretation is to split up the curve C into small segments of length ds .



The length of a typical segment is approximated as a hypotenuse of a right triangle of base dx and height dy



and therefore, by the Pythagorean Theorem:

$$ds = \sqrt{(dx)^2 + (dy)^2} = \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2} dt$$

If we use the notation $\int_C ds$ to symbolize the limit of the sum of all the segments of length ds as t ranges from a to b then it becomes clearer that the length along C is represented by $\int_C ds = \int_a^b \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2} dt$.

For a more physical interpretation, we consider a particle moving along C with it's position t seconds given by $\vec{r} = \langle x(t), y(t) \rangle$. The velocity of the particle is $\vec{v} = \frac{d\vec{r}}{dt} = \langle x'(t), y'(t) \rangle$ and so the speed of the particle is:

$$|\vec{v}| = \sqrt{(x'(t))^2 + (y'(t))^2}$$

We can look at the formula :

$$ds = |\vec{v}| dt = \sqrt{(x'(t))^2 + (y'(t))^2} dt$$

as:

change in distance = (speed)(change in time)

therefore, if we look at the length of curve C , $\int_C ds$ as the sum of all the (speed)(time) expressions, so

$$\text{Total distance traveled} = \int_a^b |\vec{v}| dt = \int_a^b \sqrt{(x'(t))^2 + (y'(t))^2} dt$$

In three dimensions, the velocity vector has three coordinates

$$\vec{v} = \langle x'(t), y'(t), z'(t) \rangle$$

and the speed is $|\vec{v}| = \sqrt{(x'(t))^2 + (y'(t))^2 + (z'(t))^2}$ so the total distance along C is:

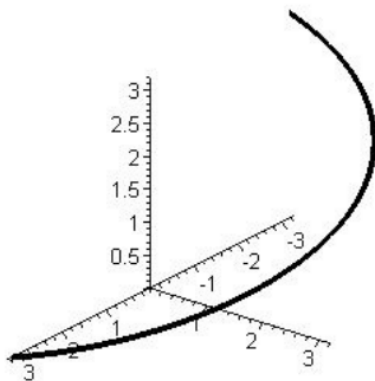
$$\int_C ds = \int_a^b |\vec{v}| dt = \int_a^b \sqrt{(x'(t))^2 + (y'(t))^2 + (z'(t))^2} dt$$

The notion of adding quantities up over a curve C applies not only to distance but to other physical quantities as well. Suppose, for example, we had electric charge distributed along this path C with a charge density at point $\vec{r} = \langle x, y, z \rangle$ of $f(x, y, z)$ coulombs per meter. Then, a small segment of length ds meters would hold a charge of $f(x, y, z) ds$ coulombs. The total charge along this path would be the limit of the sum of such quantities, which would give us:

$$\text{Total Charge along } C = \int_C f(x, y, z) ds$$

The computation of such an integral is similar to an arc length calculation. For example, suppose C is a helical segment, given by

$$\vec{r} = \langle x(t), y(t), z(t) \rangle = \langle \pi \cos t, \pi \sin t, t \rangle \quad 0 \leq t \leq \pi$$



Let us suppose that the linear charge density at any point (x, y, z) on this curve is given by $f(x, y, z) = \frac{1}{x^2 + y^2 + z^2}$ coulomb per meter. This would mean that as we move along the helix from $(\pi, 0, 0)$ at $t = 0$ to $(-\pi, 0, \pi)$ at $t = \pi$, the charge density gets smaller and smaller. The density at $\vec{r} = \langle \pi \cos t, \pi \sin t, t \rangle$ is:

$$f(x, y, z) = \frac{1}{x^2 + y^2 + z^2} = \frac{1}{\pi^2 \cos^2 t + \pi^2 \sin^2 t + t^2} = \frac{1}{\pi^2 + t^2}$$

The charge (in coulombs) along a segment of length ds is:

$$\begin{aligned}
 f(x, y, z) ds &= f(x, y, z) |\vec{v}| dt \\
 &= \frac{1}{\pi^2 + t^2} \sqrt{(x'(t))^2 + (y'(t))^2 + (z'(t))^2} dt \\
 &= \frac{1}{\pi^2 + t^2} \sqrt{(-\pi \sin t)^2 + (\pi \cos t)^2 + 1} dt \\
 &= \frac{\sqrt{\pi^2 + 1}}{\pi^2 + t^2} dt
 \end{aligned}$$

Therefore, the total charge is:

$$\int_C f(x, y, z) ds = \int_0^\pi \frac{\sqrt{\pi^2 + 1}}{\pi^2 + t^2} dt = \frac{\sqrt{\pi^2 + 1}}{\pi} \tan^{-1} 1 = \frac{\sqrt{\pi^2 + 1}}{4}$$

Work

Work is defined as the integral of force over distance.

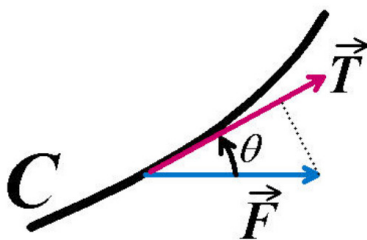
$$W = \int_a^b F(x) dx$$

This definition is only applicable if a force F is pushing something along a straight line. You might expect that if a force is acting along a curve C then the definition of work needs to be extended to be:

$$W = \int_C F(x, y, z) ds$$

since ds represents the change in distance along the curve.

However, there is a complication. Suppose the force is directed at an angle θ to the path C .



It is only the tangential component of the force that is doing work along the path. Therefore, we should be using $|\vec{F}| \cos \theta$ in our integral. If $\vec{T}$ is the unit tangent vector, then $|\vec{F}| \cos \theta = \vec{F} \cdot \vec{T}$ and the expression for the work done along path C is:

$$W = \int_C \vec{F} \cdot \vec{T} ds$$

The unit tangent vector $\vec{T}$ can be found by dividing the tangent vector $\vec{v} = \frac{d\vec{r}}{dt}$ by its length.

$$\vec{T} = \frac{\vec{v}}{|\vec{v}|}$$

We already know from that $ds = |\vec{v}|dt$ so:

$$\vec{T} ds = \frac{\vec{v}}{|\vec{v}|} |\vec{v}| dt = \vec{v} dt$$

Therefore, if the initial point on the curve occurs at $t = a$ and the final point is at $t = b$, the integral for work is given by:

$$W = \int_a^b \vec{F} \bullet \vec{v} dt$$

Note that $\vec{v} dt = \frac{d\vec{r}}{dt} dt = d\vec{r}$, so the integral for work can be written more succinctly as:

$$W = \int_C \vec{F} \bullet d\vec{r}$$

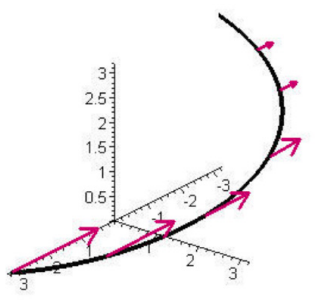
This is called a line integral whether the path C is a straight line or a curve.

Example: Let's consider the same helical segment C as:

$$\vec{r} = \langle x(t), y(t), z(t) \rangle = \langle \pi \cos t, \pi \sin t, t \rangle \quad 0 \leq t \leq \pi$$

and let's consider the work along this path by the force:

$$\vec{F} = (z - \pi)\vec{i} = \langle z - \pi, 0, 0 \rangle$$



The force is parallel to the x -axis at all points and it is shrinking to 0 as we approach the last point $(-\pi, 0, \pi)$ on the helical segment. At any point $(x, y, z) = (\pi \cos t, \pi \sin t, t)$, the force is $\vec{F} = (t - \pi, 0, 0)$ and $d\vec{r}$ is given by:

$$d\vec{r} = \frac{d\vec{r}}{dt} dt = \langle -\pi \sin t, \pi \cos t, 1 \rangle dt$$

Therefore,

$$\vec{F} \bullet d\vec{r} = \langle t - \pi, 0, 0 \rangle \bullet \langle -\pi \sin t, \pi \cos t, 1 \rangle dt = \pi(\pi - t) \sin t dt$$

and the work can be easily calculated using integration by parts:

$$W = \int_C \vec{F} \bullet d\vec{r} = \pi \int_0^\pi (\pi - t) \sin t dt = \pi^2$$

There are two interesting things to be observed about this integral. As you learned in Calculus, reversing the limits of integration changes the answer by a negative sign:

$$\pi \int_{\pi}^0 (\pi - t) \sin t \, dt = -\pi^2$$

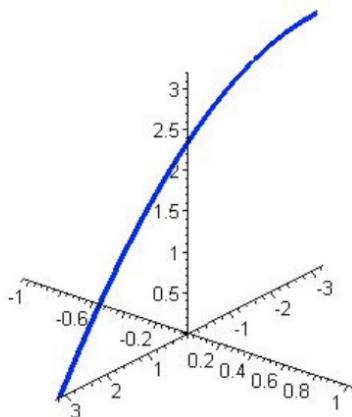
We interpret this as reversing direction along the path. If we start with $t = \pi$ then we are beginning at the point $(-\pi, 0, \pi)$ and if we stop with $t = 0$ then we are ending up at $(\pi, 0, 0)$. We use the notation $-C$ to denote the path C traversed in the opposite direction. Therefore,

$$\int_{-C} \vec{F} \bullet d\vec{r} = - \int_C \vec{F} \bullet d\vec{r}$$

The second interesting thing to point out is the effect of changing the path between $(\pi, 0, 0)$ and $(-\pi, 0, \pi)$. Consider the following path:

$$\langle x, y, z \rangle = \langle \pi - 2\pi t, 0, 2\pi t - \pi t^2 \rangle \quad \text{for } 0 \leq t \leq 1$$

Let's call this path Ω . Ω is a parabolic segment connecting $(\pi, 0, 0)$ to $(-\pi, 0, \pi)$.



At any point along this path:

$$\begin{aligned} \vec{F} \bullet d\vec{r} &= \langle z - \pi, 0, 0 \rangle \bullet \frac{d\vec{r}}{dt} dt \\ &= \langle 2\pi t - \pi t^2 - \pi, 0, 0 \rangle \bullet \langle -2\pi, 0, 2\pi - 2\pi t \rangle dt = 2\pi(\pi t^2 + \pi - 2\pi t) dt \end{aligned}$$

Therefore, the work done by $\vec{F}$ along Ω is:

$$\int_{\Omega} \vec{F} \bullet d\vec{r} = 2\pi \int_0^1 (\pi t^2 + \pi - 2\pi t) dt = \frac{2}{3}\pi^3$$

Note that $\int_C \vec{F} \bullet d\vec{r}$ is not the same as $\int_{\Omega} \vec{F} \bullet d\vec{r}$ even though C and Ω have the same starting point and the same ending point. We say that the integral is path dependent. An important result we will encounter is that, for certain vector fields, called conservative vector fields, the path connecting a point P to a point Q will not affect the answer. Since the path does make a difference for $\vec{F} = (z - \pi)\vec{i}$, this would be an example of a nonconservative vector field.

Example: Let $\vec{F} = \langle y - x, z - y, x - z \rangle$. Integrate this vector field over the straight line segment L connecting (0, 0, 1) to (1, 2, 2).

Solution:

The straight line L can be described by the equation:

$$\vec{r} = \langle 0, 0, 1 \rangle + t\langle 1, 2, 1 \rangle = \langle t, 2t, 1+t \rangle \quad \text{for } 0 \leq t \leq 1$$

At any point on this line, $\vec{F}$ is given by:

$$\vec{F} = \langle y - x, z - y, x - z \rangle = \langle 2t - t, (1+t) - 2t, t - (1+t) \rangle = \langle t, 1-t, -1 \rangle$$

The vector $d\vec{r}$ is given by:

$$d\vec{r} = \frac{d\vec{r}}{dt} dt = \langle 1, 2, 1 \rangle dt$$

We can now calculate the dot product $\vec{F} \bullet d\vec{r}$:

$$\vec{F} \bullet d\vec{r} = \langle t, 1-t, -1 \rangle \bullet \langle 1, 2, 1 \rangle dt = (t + 2(1-t) - 1) dt = (1-t) dt$$

The line integral is therefore equal to:

$$\int_L \vec{F} \bullet d\vec{r} = \int_0^1 (1-t) dt = \frac{1}{2}$$

It is not always necessary to use a parametric representation of the path we integrate over. In general, $d\vec{r} = \frac{d\vec{r}}{dt} dt = \left\langle \frac{dx}{dt}, \frac{dy}{dt}, \frac{dz}{dt} \right\rangle dt = \langle dx, dy, dz \rangle$.

If $\vec{F} = \langle F_1, F_2, F_3 \rangle$ then $\vec{F} \bullet d\vec{r} = F_1 dx + F_2 dy + F_3 dz$ and so the line integral can be written as:

$$\int_L \vec{F} \bullet d\vec{r} = \int_L F_1 dx + F_2 dy + F_3 dz$$

So, for example, the line integral we just did can be written as:

$$\int_L \vec{F} \bullet d\vec{r} = \int_L (y-x) dx + (z-y) dy + (x-z) dz$$

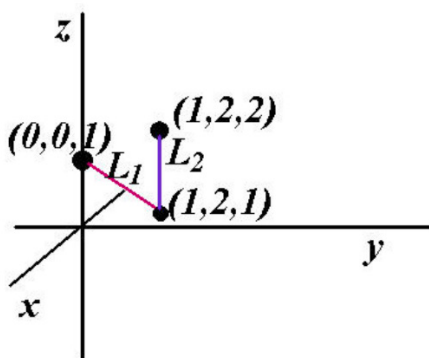
In the case of the line we just considered, where $x = t$, $y = 2t$ and $z = 1+t$, we have:

$$y = 2x \quad z = 1 + x$$

Therefore,

$$\begin{aligned} \int_L \vec{F} \bullet d\vec{r} &= \int_L (y-x) dx + (z-y) dy + (x-z) dz \\ &= \int_0^1 (2x-x) dx + ((1+x)-2x) 2 dx + (x-(1+x)) dx \\ &= \int_0^1 (1-x) dx \\ &= \frac{1}{2} \end{aligned}$$

It all comes to the same thing, but there are times that a description of the path in terms of x , y and z directly might be more convenient. For example, suppose L_1 denotes the straight line segment from (0, 0, 1) to (1, 2, 1) and L_2 denotes the straight line segment from (1, 2, 1) to (1, 2, 2) and let L_3 be the combined path along L_1 followed by L_2 .



The work done along the combined path L_3 is the sum of the work done along L_1 and the work done along L_2 .

$$\int_{L_3} \vec{F} \cdot d\vec{r} = \int_{L_1} \vec{F} \cdot d\vec{r} + \int_{L_2} \vec{F} \cdot d\vec{r}$$

Along L_1 , $y = 2x$ and $z = 1$ so $dy = 2 dx$ and $dz = 0$. Therefore:

$$\int_{L_1} \vec{F} \cdot d\vec{r} = \int_{L_1} (y - x) dx + (z - y) dy = \int_0^1 (2 - 3x) dx = \frac{1}{2}$$

Along L_2 , $x = 1$ and $y = 2$ at every point so $dx = dy = 0$.

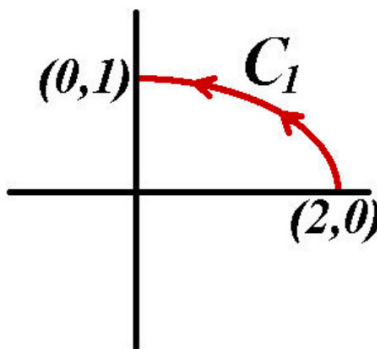
$$\int_{L_2} \vec{F} \cdot d\vec{r} = \int_{L_2} (x - z) dz = \int_1^2 (1 - z) dz = -\frac{1}{2}$$

Therefore,

$$\int_{L_3} \vec{F} \cdot d\vec{r} = \int_{L_1} \vec{F} \cdot d\vec{r} + \int_{L_2} \vec{F} \cdot d\vec{r} = \frac{1}{2} - \frac{1}{2} = 0$$

Please note that even though L and L_3 have the same initial point and the same final point, $\int_L \vec{F} \cdot d\vec{r}$ is not the same as $\int_{L_3} \vec{F} \cdot d\vec{r}$, so $\vec{F}$ is not a conservative vector field.

Example: Consider the path from $(2, 0)$ to $(0, 1)$ along the upper portion of the ellipse $\frac{x^2}{4} + y^2 = 1$. Let's call this path C_1 because a little while we are going to consider a different path, C_2 , connecting these two points.



Consider the following vector field:

$$\vec{\mathbf{F}} = \langle -x + y, x - 2y \rangle$$

and calculate the line integral:

$$\int_{C_1} \vec{\mathbf{F}} \bullet d\vec{\mathbf{r}}$$

Let's begin by describing the path with the equation $y = \frac{1}{2}\sqrt{4-x^2}$. It follows that $dy = \frac{-x}{2\sqrt{4-x^2}} dx$.

$$\begin{aligned} \vec{\mathbf{F}} \bullet d\vec{\mathbf{r}} &= (-x + y) dx + (x - 2y) dy \\ &= \left(-x + \frac{1}{2}\sqrt{4-x^2}\right) dx + \left(x - \sqrt{4-x^2}\right) \left(\frac{-x}{2\sqrt{4-x^2}}\right) dx \\ &= -\frac{x}{2} dx + \frac{2-x^2}{\sqrt{4-x^2}} dx \end{aligned}$$

To integrate over the path, we must be careful to start at $x = 2$ and end up at $x = 0$ if we want to integrate in the direction specified.

$$\int_{C_1} \vec{\mathbf{F}} \bullet d\vec{\mathbf{r}} = -\int_2^0 \frac{x}{2} dx + \int_2^0 \frac{2-x^2}{\sqrt{4-x^2}} dx = 1 - \int_0^2 \frac{2-x^2}{\sqrt{4-x^2}} dx$$

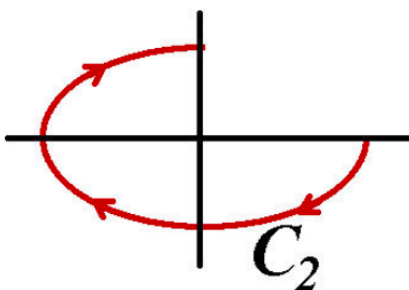
The remaining integral is somewhat tedious, but if you use trigonometric substitution with $x = 2 \sin \theta$, then you will find that $\int_0^2 \frac{2-x^2}{\sqrt{4-x^2}} dx = 0$ and therefore,

$$\int_{C_1} \vec{\mathbf{F}} \bullet d\vec{\mathbf{r}} = 1$$

The problem is actually much simpler if we use the parametric representation of this ellipse:

$$\begin{aligned} x &= 2 \cos t & y &= \sin t \\ \int_{C_1} \vec{\mathbf{F}} \bullet d\vec{\mathbf{r}} &= \int_{C_1} (-x + y) dx + (x - 2y) dy \\ &= \int_0^{\pi/2} (-2 \cos t + \sin t)(-2 \sin t dt) + (2 \cos t - 2 \sin t)(\cos t dt) \\ &= \int_0^{\pi/2} (2 \sin t \cos t + 2(\cos^2 t - \sin^2 t)) dt \\ &= \int_0^{\pi/2} (\sin 2t + 2 \cos 2t) dt \\ &= 1 \end{aligned}$$

Now, just for comparison, let's integrate from $(2, 0)$ to $(0, 1)$ along a different path. Let C_2 be the elliptical path along the other three quarters of the ellipse.



Since the parametric representation gives us the easiest integral, let's use that. To traverse path C_2 , we simply vary t from 2π to $\frac{\pi}{2}$.

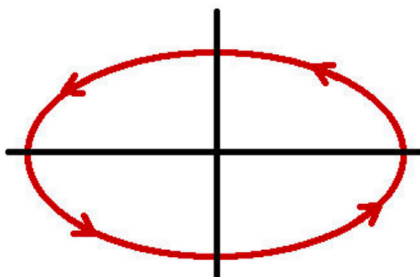
$$\int_{C_2} \vec{F} \cdot d\vec{r} = \int_{2\pi}^{\pi/2} (\sin 2t + 2 \cos 2t) dt = 1$$

Thus the choice of path has made no difference. Does this mean that the integral of $\vec{F}$ from $(2, 0)$ to $(0, 1)$ is path independent? The problem is that, for all we know, there may be some other path, call it C_3 from $(2, 0)$ to $(0, 1)$ where the answer for $\int_{C_3} \vec{F} \cdot d\vec{r}$ comes out differently. How can we resolve this? The answer lies in a line integral called a closed loop integral.

Closed Loop Integrals

Let's use the same vector field $\vec{F}$ and the same ellipse as the last problem, but this time, let's integrate around the entire ellipse. Thus, if we start at $(2, 0)$, we will end there as well. We can use the same parametric representation of the ellipse and simply adjust the limits of integration.

$$x = 2 \cos t \quad y = \sin t \quad 0 \leq t \leq 2\pi$$

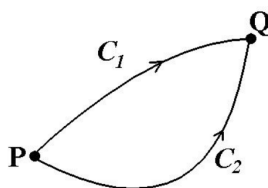


Let's call this path C . When the starting point is the same as the endpoint, the path is called a closed loop and we use the symbol $\oint$ for the integral.

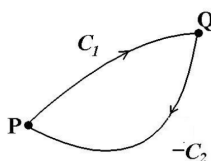
$$\oint_C \vec{F} \cdot d\vec{r} = \int_0^{2\pi} (\sin 2t + 2 \cos 2t) dt = 0$$

The integral of a vector field around a closed loop is called the circulation. It is not a coincidence that the circulation is 0. This is a reliable feature of path independent integrals. In general, suppose we have two different paths connecting point P to point Q . Call these paths C_1 and C_2 .

An **integral** is the continuous analog of a sum, which is used to calculate areas, volumes, and their generalizations.



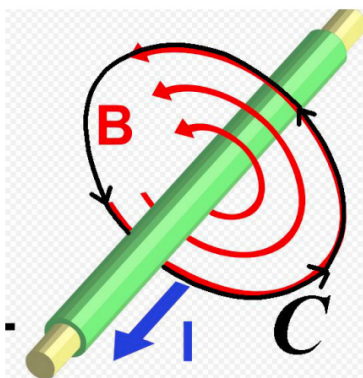
If we reverse direction along path C_2 , we get a closed loop C .



$$\oint_C \vec{F} \cdot d\vec{r} = \int_{C_1} \vec{F} \cdot d\vec{r} + \int_{-C_2} \vec{F} \cdot d\vec{r} = \int_{C_1} \vec{F} \cdot d\vec{r} - \int_{C_2} \vec{F} \cdot d\vec{r}$$

Thus, if the integral of a vector field is path independent and $\int_{C_1} \vec{F} \cdot d\vec{r}$ equals $\int_{C_2} \vec{F} \cdot d\vec{r}$ then the circulation $\oint_C \vec{F} \cdot d\vec{r}$ is 0. Conversely, if the circulation is always zero, no matter what closed loop we are considering, then the integrals along two different paths C_1 and C_2 connecting points P to Q will always come out to be the same as each other.

The circulation of a vector field is of interest for other reasons. For example, in physics the concept of circulation comes up in Ampere's Law which relates the current I along a wire to the magnetic field B around the wire.



If C is a closed loop around the wire, the circulation of the magnetic field B is proportional to the current I .

$$\oint_C \vec{B} \cdot d\vec{r} = \mu_0 I$$

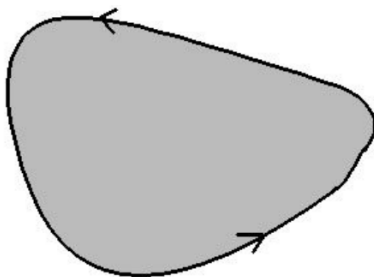
Another place we need this concept is in the study of fluids. If $\vec{v}$ is the velocity vector field of a fluid, one of the starting points of fluid theory is that the circulation is zero.

$$\oint_C \vec{v} \cdot d\vec{r} = 0$$

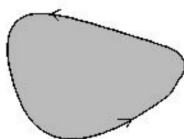
Circulation Per Unit Area

Closely associated with the circulation of a vector field is the circulation per unit area. More precisely, if $\vec{F}$ is a vector field and C is a closed loop then the circulation per unit area is defined as:

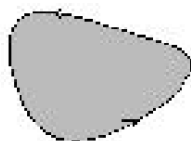
$$\frac{1}{\text{Area}} \oint_C \vec{F} \cdot d\vec{r}$$



where the area referred to is the area inside the closed loop. Now, suppose we shrink the closed loop down and make the area smaller.



Keep making the area smaller.



If we take the limit as the area goes to zero, we are shrinking the closed loop down to a point. The limit, if it exists, will be referred to as the circulation per unit area at a point.

$$\lim_{\text{Area} \rightarrow 0} \frac{1}{\text{Area}} \oint_C \vec{F} \cdot d\vec{r}$$

If a vector field is conservative, then the circulation per unit area is going to be zero, no matter how small the area, and therefore the circulation per unit area at a point can be expected to be zero at all points. The converse also turns out to be true, but we will discuss this when we get to a result called Stokes' Theorem.

Derivative Formula for Circulation per Unit Area

There is a simple formula for finding the circulation per unit area that is in terms of the partial derivatives of the coordinates of the vector field. There are two preliminary facts from calculus that we need to get this formula. The first is the Mean Value Theorem of Integrals. If $f(x)$ is continuous on the interval $[a, b]$, then there is some point c in this interval with the property that:

$$\int_a^b f(x) dx = f(c)(b - a)$$

We also need the limit definition of the derivative:

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x + h) - f(x)}{h}$$

Please note that there are other similar limits that also come out the same way:

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x) - f(x - h)}{h}$$

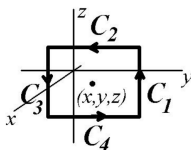
as well as:

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x + \frac{h}{2}) - f(x - \frac{h}{2})}{h}$$

You can use l'Hopital's Rule to see that each of these limits come out equal to $f'(x)$. Note that the very last limit can be written as:

$$f'(x) = \frac{f(x + \frac{\Delta x}{2}) - f(x - \frac{\Delta x}{2})}{\Delta x} + \mathcal{E} \quad \text{where} \quad \lim_{\Delta x \rightarrow 0} \mathcal{E} = 0$$

Let us begin our analysis of the circulation per unit area by considering a closed loop parallel to the yz plane. For simplicity, we will take this loop to be a rectangular path C . There is a point (x, y, z) in the center of this loop.



We will call the four segments of this rectangular path C_1 , C_2 , C_3 and C_4 .

The integral around the rectangular loop C is the sum of the line integrals along each of the four segments:

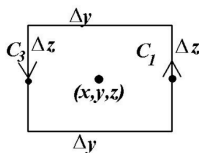
$$\oint_C \vec{F} \cdot d\vec{r} = \int_{C_1} \vec{F} \cdot d\vec{r} + \int_{C_2} \vec{F} \cdot d\vec{r} + \int_{C_3} \vec{F} \cdot d\vec{r} + \int_{C_4} \vec{F} \cdot d\vec{r}$$

We will assume that the four corners of this rectangle have the coordinates:

$$\begin{pmatrix} x, y - \frac{\Delta y}{2}, z - \frac{\Delta z}{2} \end{pmatrix} \quad \begin{pmatrix} x, y + \frac{\Delta y}{2}, z - \frac{\Delta z}{2} \end{pmatrix}$$

$$\begin{pmatrix} x, y + \frac{\Delta y}{2}, z + \frac{\Delta z}{2} \end{pmatrix} \quad \begin{pmatrix} x, y - \frac{\Delta y}{2}, z + \frac{\Delta z}{2} \end{pmatrix}$$

Let's start with path C_1 and C_3



$$\int_{C_1} \vec{\mathbf{F}} \bullet d\vec{\mathbf{r}} = \int_{C_1} F_1 dx + F_2 dy + F_3 dz = \int_{C_1} F_3 dz$$

This is true because x and y are constant along C_1 so $dx = dy = 0$ on this path. By the Mean Value Theorem for integrals, there is a point P with $P = (x, y + \frac{\Delta y}{2}, z^*)$ for z^* in the interval $[z - \frac{\Delta z}{2}, z + \frac{\Delta z}{2}]$ so that $\int_{C_1} F_3 dz = F_3(P)\Delta z$. By continuity, $F_3(P) = F_3(x, y + \frac{\Delta y}{2}, z) + \mathcal{E}_1$ where $\lim_{\Delta z \rightarrow 0} \mathcal{E}_1 = 0$ so

$$\int_{C_1} \vec{\mathbf{F}} \bullet d\vec{\mathbf{r}} = F_3\left(x, y + \frac{\Delta y}{2}, z\right) \Delta z + \mathcal{E}_1 \Delta z$$

Let's write this more concisely as:

$$\int_{C_1} \vec{\mathbf{F}} \bullet d\vec{\mathbf{r}} \approx F_3\left(x, y + \frac{\Delta y}{2}, z\right) \Delta z$$

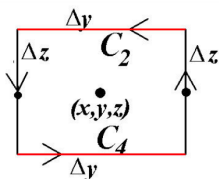
In the same way, we write the integral along C_3 as:

$$\int_{C_3} \vec{\mathbf{F}} \bullet d\vec{\mathbf{r}} \approx -F_3\left(x, y - \frac{\Delta y}{2}, z\right) \Delta z$$

Please note the negative sign in front of F_3 this time. This is because on C_3 , we start at $z + \frac{\Delta z}{2}$ and end at $z - \frac{\Delta z}{2}$ so the direction is the opposite of C_1 . Now, add these integrals together:

$$\begin{aligned} \int_{C_1} \vec{\mathbf{F}} \bullet d\vec{\mathbf{r}} + \int_{C_3} \vec{\mathbf{F}} \bullet d\vec{\mathbf{r}} &\approx F_3\left(x, y + \frac{\Delta y}{2}, z\right) \Delta z - F_3\left(x, y - \frac{\Delta y}{2}, z\right) \Delta z \\ &= \left(\frac{F_3\left(x, y + \frac{\Delta y}{2}, z\right) \Delta z - F_3\left(x, y - \frac{\Delta y}{2}, z\right) \Delta z}{\Delta y} \right) \Delta y \Delta z \\ &\approx \frac{\partial F_3}{\partial y} \Delta y \Delta z \end{aligned}$$

Next, we repeat this argument for paths C_2 and C_4



and the result is:

$$\int_{C_2} \vec{F} \bullet d\vec{r} + \int_{C_4} \vec{F} \bullet d\vec{r} \approx -\frac{\partial F_2}{\partial z} \Delta y \Delta z$$

Add the results together to get an approximation for $\oint \vec{F} \bullet d\vec{r}$

$$\oint \vec{F} \bullet d\vec{r} \approx \frac{\partial F_3}{\partial y} \Delta y \Delta z - \frac{\partial F_2}{\partial z} \Delta y \Delta z$$

The error in this approximation goes to 0 as Δy and Δz go to 0. Since the area inside C is $\text{Area} = \Delta y \Delta z$, the circulation per unit area is:

$$\frac{1}{\text{Area}} \oint \vec{F} \bullet d\vec{r} \approx \frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z}$$

Therefore,

$$\lim_{\text{Area} \rightarrow 0} \frac{1}{\text{Area}} \oint \vec{F} \bullet d\vec{r} = \frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z}$$

So, the significance of the expression:

$$\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z}$$

is that it represents the circulation per unit area at (x, y, z) in a plane parallel to the yz plane. What about the other planes? An analysis similar to what we have just done shows that:

$$\frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x}$$

represents the circulation per unit area at (x, y, z) in a plane parallel to the xz plane. Finally, the expression:

$$\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}$$

Parallel lines are coplanar infinite straight lines that do not intersect at any point.

represents the circulation per unit area at (x, y, z) in a **plane parallel** to the xy plane. We can package all these quantities very conveniently in a vector called the curl defined as:

$$\text{curl } \vec{F} = \left(\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z} \right) \vec{i} + \left(\frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x} \right) \vec{j} + \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) \vec{k}$$

Thus, $\vec{i} \bullet \text{curl } \vec{F}$ is the circulation per unit area in the plane perpendicular to the $\vec{i}$ vector, $\vec{j} \bullet \text{curl } \vec{F}$ is the circulation per unit area in the plane perpendicular to the $\vec{j}$ vector and $\vec{k} \bullet \text{curl } \vec{F}$ is the circulation per unit area in the plane perpendicular to the $\vec{k}$ vector. More generally, it can be shown that if $\vec{n}$ is any unit vector then $\vec{n} \bullet \text{curl } \vec{F}$ is the circulation per unit area in the

plane perpendicular to $\vec{n}$. There is a nice formula using determinants for the curl. We write the expression $\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z}$ as:

$$\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z} = \begin{vmatrix} \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ F_2 & F_3 \end{vmatrix}$$

Use this notation for each coordinate of the curl

$$\text{curl } \vec{F} = \vec{i} \begin{vmatrix} \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ F_2 & F_3 \end{vmatrix} - \vec{j} \begin{vmatrix} \frac{\partial}{\partial x} & \frac{\partial}{\partial z} \\ F_1 & F_3 \end{vmatrix} + \vec{k} \begin{vmatrix} \frac{\partial}{\partial x} & \frac{\partial}{\partial y} \\ F_1 & F_2 \end{vmatrix}$$

This can be written more compactly as a three by three determinant:

$$\text{curl } \vec{F} = \begin{vmatrix} \vec{i} & \vec{j} & \vec{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ F_1 & F_2 & F_3 \end{vmatrix}$$

This is commonly written as $\nabla \times \vec{F}$ since the three by three determinant has the same algebraic form as the cross product. When a vector field $\vec{F}$ is conservative, the circulation per unit area is zero at every point and in every plane, regardless of the orientation of the plane. Thus, we can tell if a vector field $\vec{F}$ is conservative simply by checking if $\nabla \times \vec{F}$ is the zero vector.

Example: Calculate the curl of $\vec{F} = \langle -x + y, x - 2y, 0 \rangle$

$$\begin{aligned} \nabla \times \vec{F} &= \begin{vmatrix} \vec{i} & \vec{j} & \vec{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ -x + y & x - 2y & 0 \end{vmatrix} \\ &= \vec{i}(0) - \vec{j}(0) + \vec{k}(1 - 1) \\ &= \langle 0, 0, 0 \rangle \end{aligned}$$

This verifies our observation that the integral of $\vec{F}$ seemed to be path independent.

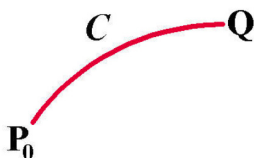
Example: Calculate the curl of $\vec{F} = \langle y - x, z - y, x - z \rangle$

$$\begin{aligned} \nabla \times \vec{F} &= \begin{vmatrix} \vec{i} & \vec{j} & \vec{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ y - x & z - y & x - z \end{vmatrix} \\ &= \vec{i}(0 - (-1)) - \vec{j}(1 - 0) + \vec{k}(0 - 1) \\ &= \langle 1, -1, -1 \rangle \end{aligned}$$

Since $\text{curl } \vec{F}$ is not the zero vector, this verifies our observation that the value of the line integral between two points might change if the path connecting the two points changes.

The Potential Function

Let us suppose that $\vec{F}$ represents force so that $\int_C \vec{F} \cdot d\vec{r}$ is work. If $\vec{F}$ is a conservative vector field then the integral is path independent. That is, if C connects point P_0 to Q then the value of $\int_C \vec{F} \cdot d\vec{r}$ would remain the same even if we changed the path C connecting P_0 to Q .

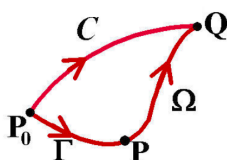


Since the path doesn't matter, we will use the notation:

$$\int_{P_0}^Q \vec{F} \cdot d\vec{r}$$

Suppose we took a side trip through point P to get from P_0 to Q . If Γ is a path from P_0 to P and Ω is a path from P to Q then:

$$\int_C \vec{F} \cdot d\vec{r} = \int_{\Gamma} \vec{F} \cdot d\vec{r} + \int_{\Omega} \vec{F} \cdot d\vec{r}$$



Equivalently:

$$\int_{P_0}^Q \vec{F} \cdot d\vec{r} = \int_{P_0}^P \vec{F} \cdot d\vec{r} + \int_P^Q \vec{F} \cdot d\vec{r}$$

Now, let us consider the work done from point $P_0 = (x_0, y_0, z_0)$ to $P = (x, y, z)$. If we regard the initial point (x_0, y_0, z_0) as a constant and the final point (x, y, z) as the variable then the work done is a function of (x, y, z) and will be denoted by $\phi(x, y, z)$

$$\phi(x, y, z) = \int_{P_0}^P \vec{F} \cdot d\vec{r} = \int_{(x_0, y_0, z_0)}^{(x, y, z)} \vec{F} \cdot d\vec{r}$$

$\phi(x, y, z)$ will be referred to as the potential function for $\vec{F}$. The integral sum $\int_{P_0}^Q \vec{F} \cdot d\vec{r} = \int_{P_0}^P \vec{F} \cdot d\vec{r} + \int_P^Q \vec{F} \cdot d\vec{r}$ that we considered can be written as:

$$\phi(Q) = \phi(P) + \int_P^Q \vec{F} \cdot d\vec{r}$$

and therefore,

$$\int_P^Q \vec{F} \cdot d\vec{r} = \phi(Q) - \phi(P)$$

Thus, the integral of a conservative vector field between any two points can be expressed as the change in its potential function. Therefore, given a vector field $\vec{F}$ it would be convenient to know how to get the formula for the potential function ϕ . Of course, we do have an integral definition $\phi(x, y, z) = \int_{P_0}^P \vec{F} \cdot d\vec{r}$ and this could be used. However, if ϕ is obtained by performing an antiderivative operation on the

coordinates of $\vec{F}$ it would make sense that we should be able to obtain the coordinates of $\vec{F}$ by performing some derivative operations on ϕ . Let's begin by looking at the partial derivative of $\phi(x, y, z)$ with respect to x .

$$\frac{\partial \phi}{\partial x} = \lim_{h \rightarrow 0} \frac{\phi(x+h, y, z) - \phi(x, y, z)}{h}$$

As we have just seen, $\phi(x+h, y, z) - \phi(x, y, z)$ would result from a line integral.

$$\begin{aligned} \phi(x+h, y, z) - \phi(x, y, z) &= \int_{(x,y,z)}^{(x+h,y,z)} \vec{F} \bullet d\vec{r} \\ &= \int_{(x,y,z)}^{(x+h,y,z)} F_1 dx + F_2 dy + F_3 dz \end{aligned}$$

Since the line integral of this vector field is path independent, we might as well take the path to be a straight line from (x, y, z) to $(x+h, y, z)$. This means that y and z are constant along the path, so $dy = dz = 0$ and we are left with:

$$\phi(x+h, y, z) - \phi(x, y, z) = \int_{(x,y,z)}^{(x+h,y,z)} F_1 dx$$

By the Mean Value Theorem of Integrals, there is a value x^* somewhere in the interval $[x, x+h]$ such that $\int_{(x,y,z)}^{(x+h,y,z)} F_1 dx = F_1(x^*, y, z) h$, and therefore:

$$\phi(x+h, y, z) - \phi(x, y, z) = \int_{(x,y,z)}^{(x+h,y,z)} F_1 dx = F_1(x^*, y, z) h$$

In exactly the same manner, we can show that:

$$\phi(x, y+h, z) - \phi(x, y, z) = \int_{(x,y,z)}^{(x,y+h,z)} F_2 dy = F_2(x, y^*, z) h$$

$$\phi(x, y, z+h) - \phi(x, y, z) = \int_{(x,y,z)}^{(x,y,z+h)} F_3 dz = F_3(x, y, z^*) h$$

where y^* is some value in $[y, y+h]$ and z^* is some value in $[z, z+h]$. If we now divide by h , we get:

$$\frac{\phi(x+h, y, z) - \phi(x, y, z)}{h} = F_1(x^*, y, z)$$

$$\frac{\phi(x, y+h, z) - \phi(x, y, z)}{h} = F_2(x, y^*, z)$$

$$\frac{\phi(x, y, z+h) - \phi(x, y, z)}{h} = F_3(x, y, z^*)$$

If we take the limit as h goes to 0, we are left with:

$$\frac{\partial \phi}{\partial x} = F_1 \quad \frac{\partial \phi}{\partial y} = F_2 \quad \frac{\partial \phi}{\partial z} = F_3$$

This implies that $\vec{F}$ is the gradient of the potential function.

$$\vec{F} = \langle F_1, F_2, F_3 \rangle = \left\langle \frac{\partial \phi}{\partial x}, \frac{\partial \phi}{\partial y}, \frac{\partial \phi}{\partial z} \right\rangle = \nabla \phi$$

We have just established that if the integral of $\vec{F}$ is path independent then $\vec{F}$ must be the gradient of its potential function. The converse is also true. That is, if $\vec{F} = \nabla \phi$ for some scalar-valued function ϕ then the integral of $\vec{F}$ between two points P and Q must be independent of the path connecting those two points. To understand why this true, take any path C between P and Q. C will be described by parametric equations of the form:

$$x = x(t) \quad y = y(t) \quad z = z(t) \quad \text{for } a \leq t \leq b$$

At any point t, ϕ will have the value $\phi(x(t), y(t), z(t))$ on the curve. By the Chain Rule for Partial Derivatives,

$$\frac{d}{dt}(\phi(x(t), y(t), z(t))) = \frac{\partial \phi}{\partial x} \frac{dx}{dt} + \frac{\partial \phi}{\partial y} \frac{dy}{dt} + \frac{\partial \phi}{\partial z} \frac{dz}{dt} = \nabla \phi \cdot \frac{d\vec{r}}{dt} = \vec{F} \cdot \frac{d\vec{r}}{dt}$$

Therefore,

$$\begin{aligned} \int_C \vec{F} \cdot d\vec{r} &= \int_a^b \vec{F} \cdot \frac{d\vec{r}}{dt} dt = \int_a^b \frac{d}{dt}(\phi(x(t), y(t), z(t))) dt \\ &= \phi(x(b), y(b), z(b)) - \phi(x(a), y(a), z(a)) = \phi(\mathbf{Q}) - \phi(\mathbf{P}) \end{aligned}$$

The final answer apparently comes out to $\phi(\mathbf{Q}) - \phi(\mathbf{P})$ regardless of which path C we have chosen. Notice that the line integral can now be written as:

$$\int_P^Q \nabla \phi \cdot d\vec{r} = \phi(\mathbf{Q}) - \phi(\mathbf{P})$$

This resembles the Fundamental Theorem of Calculus:

$$\int_a^b f'(x) dx = f(b) - f(a)$$

so the formula $\int_P^Q \nabla \phi \cdot d\vec{r} = \phi(\mathbf{Q}) - \phi(\mathbf{P})$ is essentially a generalization of the fundamental theorem to line integrals.

A **line integral** is an integral where the function to be integrated is evaluated along a curve.

Example: Calculate the potential function for $\vec{F} = \langle -x + y, x - 2y \rangle$ and use your result to recalculate the **line integral** of $\vec{F}$ over C_1 , the elliptical path from (2, 0) to (0, 1) that we considered earlier.

Solution: If ϕ is the potential function, then $\frac{\partial \phi}{\partial x} = -x + y$ and $\frac{\partial \phi}{\partial y} = x - 2y$. Take either one of these equations, it doesn't matter much which one, and integrate with respect to the

appropriate variable. That is, either integrate $-x + y$ with respect to x (holding y constant) or integrate $x - 2y$ with respect to y (holding x constant). Since it doesn't matter which we start with, let's choose $\frac{\partial \phi}{\partial y} = x - 2y$. This implies that:

$$\begin{aligned}\phi(x, y) &= \int (x - 2y) dy \quad (\text{holding } x \text{ constant}) \\ &= xy - y^2\end{aligned}$$

You may notice that I haven't written the usual constant of integration $+C$ at the end of this antiderivative, so $xy - y^2$ is not the most general expression for which $\frac{\partial \phi}{\partial y} = x - 2y$. Not only would $xy - y^2 + C$ also work for any constant C , but the following expressions would also work:

$$xy - y^2 + \sin x \quad xy - y^2 + e^x \quad xy - y^2 + \ln x + C$$

More generally,

$$xy - y^2 + G(x)$$

Since x was being held constant in our integration, x may appear in our "constant" of integration.

If $\frac{\partial \phi}{\partial y} = x - 2y$ where the only condition on ϕ then $xy - y^2 + G(x)$ is the answer for any function $G(x)$. However, there is one more condition, namely $\frac{\partial \phi}{\partial x} = -x + y$ and this will determine a specific expression for $G(x)$

$$\frac{\partial}{\partial x}(xy - y^2 + G(x)) = -x + y$$

$$y + G'(x) = -x + y \quad \text{so} \quad G'(x) = -x$$

If $G'(x) = -x$ then $G(x) = -\frac{1}{2}x^2 + C$ and therefore:

$$\phi(x, y) = xy - y^2 - \frac{1}{2}x^2 + C$$

Now, let's use this to integrate $\vec{F}$ from $(2, 0)$ to $(0, 1)$. Since the integral of this $\vec{F}$ is path independent, we can get our answer directly with ϕ

$$\int_{(2,0)}^{(0,1)} \vec{F} \cdot d\vec{r} = \phi(0, 1) - \phi(2, 0) = (-1 + C) - (-\frac{1}{2}2^2 + C) = 1$$

Notice how the $+C$ cancels out. The $+C$ will always cancel out when we evaluate $\phi(Q) - \phi(P)$ so, from now on, we might as well choose $C = 0$ for calculation of other potential functions.

Example: Find the potential function for $\vec{F} = \langle 3y^2 + y, x \rangle$ and calculate the line integral of this vector field over the straight line path from $(0, 0)$ to $(1, 4)$.

Solution: If there is a potential function ϕ then ϕ must satisfy the conditions:

$$\frac{\partial \phi}{\partial x} = 3y^2 + y \quad \frac{\partial \phi}{\partial y} = x$$

If $\frac{\partial \phi}{\partial y} = x$ the $\phi(x, y)$ must have the form $xy + G(x)$. If we now impose the condition $\frac{\partial \phi}{\partial x} = 3y^2 + y$ we get the equation $y + G'(x) = 3y^2 + y$ which leads to the contradictory result that $G'(x) = 3y^2$. $G'(x)$ cannot equal an expression depending on the y variable. We conclude that there is no potential function ϕ for this vector field.

We could have predicted this result had we calculated the curl of $\vec{F}$.

$$\nabla \times \vec{F} = \begin{vmatrix} \vec{i} & \vec{j} & \vec{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ 3y^2 + y & x & 0 \end{vmatrix} = \langle 0, 0, -6y - y \rangle$$

Had there been a function ϕ with $\nabla \phi = \vec{F}$ then:

$$\begin{aligned} \nabla \times \vec{F} &= \begin{vmatrix} \vec{i} & \vec{j} & \vec{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ \frac{\partial \phi}{\partial x} & \frac{\partial \phi}{\partial y} & \frac{\partial \phi}{\partial z} \end{vmatrix} \\ &= \left\langle \frac{\partial^2 \phi}{\partial y \partial z} - \frac{\partial^2 \phi}{\partial z \partial y}, \frac{\partial^2 \phi}{\partial z \partial x} - \frac{\partial^2 \phi}{\partial x \partial z}, \frac{\partial^2 \phi}{\partial x \partial y} - \frac{\partial^2 \phi}{\partial y \partial x} \right\rangle \\ &= \langle 0, 0, 0 \rangle \end{aligned}$$

So, a good general rule is that if $\nabla \times \vec{F}$ is not the zero vector then don't bother looking for a function ϕ with $\vec{F} = \nabla \phi$.

There are a couple of things to note about this last vector field. The first is that even though $\vec{F} = 3y^2 + y, x$ it doesn't mean that we can't calculate the line integral along the straight line path from the origin to $(1, 4)$. Let's call this path Γ . Along Γ , $y = 4x$ so,

$$\begin{aligned} \int_{\Gamma} (3y^2 + y) dx + x dy &= \int_0^1 (3(4x)^2 + 4x) dx + x(4 dx) \\ &= \int_0^1 (48x^3 + 8x) dx \\ &= 20 \end{aligned}$$

The second thing to note is even though $\vec{F} = \langle 3y^2 + y, x \rangle$ is not the gradient of any function, there is still a way to use potential functions to do the line integral. The vector field $\vec{F}$ can be split up in the following way:

$$\vec{F} = \langle 3y^2 + y, x \rangle = \langle 3y^2, 0 \rangle + \langle y, x \rangle$$

The vector field $\langle y, x \rangle$ does indeed have a potential function. A simple calculation shows that $\langle y, x \rangle = \nabla \phi$ where $\phi = xy$. So, $\vec{F}$ has split up into a nonconservative part and a conservative part. We can now do the calculation of $\int_{\Gamma} \vec{F} \cdot d\vec{r}$ using this fact:

$$\begin{aligned} \int_{\Gamma} \vec{F} \cdot d\vec{r} &= \int_{\Gamma} (\langle 3y^2, 0 \rangle + \nabla \phi) \cdot d\vec{r} \\ &= \int_{\Gamma} \langle 3y^2, 0 \rangle \cdot \langle dx, dy \rangle + \int_{\Gamma} \nabla \phi \cdot d\vec{r} \\ &= \int_0^1 3(4x)^2 dx + \phi(1, 4) - \phi(0, 0) \\ &= \int_0^1 48x^3 dx + (1)(4) - (0)(0) \\ &= 16 + 4 = 20 \end{aligned}$$

Next, let's try a more three dimensional example.

Example: Find the potential function $\phi(x, y, z)$ for $\vec{F} = \frac{k}{(x^2+y^2+z^2)^{3/2}} \langle x, y, z \rangle$ where k is a constant. The significance of this vector field is that it is the same format as the gravitational force or electrostatic force.

Solution: A straightforward, but slightly tedious calculation, shows that $\nabla \times \vec{F} = \vec{0}$ except at the origin. So, we should be able to find a function $\phi(x, y, z)$ such that:

$$\frac{\partial \phi}{\partial x} = \frac{kx}{(x^2 + y^2 + z^2)^{3/2}} \quad \frac{\partial \phi}{\partial y} = \frac{ky}{(x^2 + y^2 + z^2)^{3/2}} \quad \frac{\partial \phi}{\partial z} = \frac{kz}{(x^2 + y^2 + z^2)^{3/2}}$$

Let's start with the equation for $\frac{\partial \phi}{\partial x}$. Integrate with respect to x to get an initial general expression for ϕ

$$\phi(x, y, z) = \int \frac{kx}{(x^2 + y^2 + z^2)^{3/2}} dx$$

To calculate this integral, let $u = x^2 + y^2 + z^2$ and remember that y and z are being held constant in this operation.

$$\phi(x, y, z) = \frac{k}{2} \int u^{-3/2} du = -ku^{-1/2} + G(y, z) = \frac{-k}{\sqrt{x^2 + y^2 + z^2}} + G(y, z)$$

Next, we impose the condition

$$\frac{\partial \phi}{\partial y} = \frac{ky}{(x^2 + y^2 + z^2)^{3/2}} \quad \text{on} \quad \phi = \frac{-k}{\sqrt{x^2 + y^2 + z^2}} + G(y, z)$$

to obtain:

$$\frac{ky}{(x^2 + y^2 + z^2)^{3/2}} + \frac{\partial G}{\partial y}(y, z) = \frac{ky}{(x^2 + y^2 + z^2)^{3/2}}$$

and therefore $\frac{\partial G}{\partial y}(y, z) = 0$. From this we conclude that G is either constant or depends on z alone. Thus,

$$\phi(x, y, z) = \frac{-k}{\sqrt{x^2 + y^2 + z^2}} + H(z)$$

Finally, we impose the condition

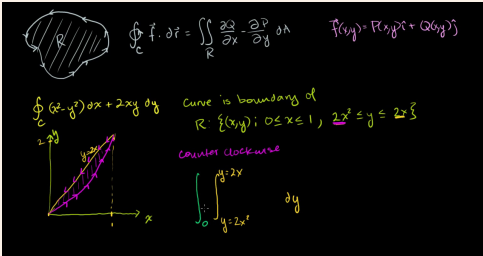
$$\frac{\partial \phi}{\partial z} = \frac{kz}{(x^2 + y^2 + z^2)^{3/2}} \quad \text{on} \quad \phi(x, y, z) = \frac{-k}{\sqrt{x^2 + y^2 + z^2}} + H(z)$$

and we obtain:

$$\frac{kz}{(x^2 + y^2 + z^2)^{3/2}} + H'(z) = \frac{kz}{(x^2 + y^2 + z^2)^{3/2}}$$

Therefore, H must be a constant. We might as well choose this constant to be 0 and we are left with the potential function:

$$\phi(x,y,z) = \frac{-k}{\sqrt{x^2+y^2+z^2}}$$



CORE CONCEPT: LINE INTEGRALS IN GREEN’S THEOREM

Green’s theorem relates a line integral around a simple closed curve C to a double integral over the plane region D bounded by C. It is the two-dimensional special case of Stokes’ theorem.

Green’s theorem gives a relationship between the line integral of a two-dimensional vector field over a closed path in the plane and the double integral over the region it encloses. The fact that the integral of a (two-dimensional) conservative field over a closed path is zero is a special case of Green’s theorem.

Green’s theorem is itself a special case of the much more general Stokes’ theorem. The statement in Green’s theorem that two different types of integrals are equal can be used to compute either type: sometimes Green’s theorem is used to transform a line integral into a double integral, and sometimes it is used to transform a double integral into a line integral.

We now come to the first of three important theorems that extend the Fundamental Theorem of Calculus to higher **dimensions**. They all share with the Fundamental Theorem the following rather vague description: To compute a certain sort of integral over a region, we may do a computation on the boundary of the region that involves one fewer integrations.

Note that this does indeed describe the Fundamental Theorem of Calculus and the Fundamental Theorem of Line Integrals: to compute a single integral over an interval, we do a computation on the boundary (the endpoints) that involves one fewer integrations, namely, no integrations at all.

If the vector field $F = \langle P, Q \rangle$ and the region D are sufficiently nice, and if C is the boundary of D (C is a closed curve), then

$$\iint_D \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA = \int_C P dx + Q dy,$$

Dimensions is a database of abstracts and citations and of research grants, which links grants to resulting publications, clinical trials and patents.

provided the integration on the right is done counter-clockwise around C .

To indicate that an integral $\int_C$ is being done over a closed curve in the counterclockwise direction, we usually write $\oint_C$. We also use the notation ∂D to mean the boundary of D oriented in the counterclockwise direction. With this notation, $\oint_C = \int_{\partial D}$.

We already know one case, not particularly interesting, in which this theorem is true: If F is conservative, we know that the integral $\oint_C \mathbf{F} \cdot d\mathbf{x} = 0$, because any integral of a conservative vector field around a closed curve is zero. We also know in this case that $\partial P/\partial y = \partial Q/\partial x$, so the double integral in the theorem is simply the integral of the zero function, namely, 0. So in the case that F is conservative, the theorem says simply that $0 = 0$.

EXAMPLE

We illustrate the theorem by computing both sides of

$$\int_{\partial D} x^4 dx + xy dy = \iint_D y - 0 dA,$$

where D is the triangular region with corners $(0, 0)$, $(1, 0)$, $(0, 1)$.

Starting with the double integral:

$$\iint_D y - 0 dA = \int_0^1 \int_0^{1-x} y dy dx = \int_0^1 \frac{(1-x)^2}{2} dx = -\frac{(1-x)^3}{6} \Big|_0^1 = \frac{1}{6}.$$

There is no single formula to describe the boundary of D , so to compute the left side directly we need to compute three separate integrals corresponding to the three sides of the triangle, and each of these integrals we break into two integrals, the “ dx ” part and the “ dy ” part. The three sides are described by $y = 0$, $y = 1 - x$, and $x = 0$. The integrals are then

$$\begin{aligned} \int_{\partial D} x^4 dx + xy dy &= \int_0^1 x^4 dx + \int_0^0 0 dy + \int_1^0 x^4 dx + \int_0^1 (1-y)y dy + \int_0^0 0 dx + \int_1^0 0 dy \\ &= \frac{1}{5} + 0 - \frac{1}{5} + \frac{1}{6} + 0 + 0 = \frac{1}{6}. \end{aligned}$$

Alternately, we could describe the three sides in vector form as $\langle t, 0 \rangle$, $\langle 1-t, t \rangle$. Note that in each case, as t ranges from 0 to 1, we follow the corresponding side in the correct direction. Now

$$\begin{aligned} \int_{\partial D} x^4 dx + xy dy &= \int_0^1 t^4 + t \cdot 0 dt + \int_0^1 -(1-t)^4 + (1-t)t dt + \int_0^1 0 + 0 dt \\ &= \int_0^1 t^4 dt + \int_0^1 -(1-t)^4 + (1-t)t dt = \frac{1}{6}. \end{aligned}$$

None of the integrations are difficult, but the second approach is somewhat tedious because of the necessity to set up three different integrals. In different circumstances, either of the integrals, the single or the double, might be easier to compute. Sometimes it is worthwhile to turn a single integral into the corresponding double integral,

sometimes exactly the opposite approach is best. Here is a clever use of Green's Theorem: We know that areas can be computed using double integrals, namely,

$$\iint_D 1 \, dA$$

computes the area of region D . If we can find P and Q so that $\partial Q/\partial x - \partial P/\partial y = 1$, then the area is also

$$\int_{\partial D} P \, dx + Q \, dy.$$

It is quite easy to do this: $P = 0$, $Q = x$ works, as do $P = -y$, $Q = 0$ and $P = -y/2$, $Q = x/2$.

EXAMPLE

An ellipse centered at the origin, with its two principal axes aligned with the x and y axes, is given by

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

We find the area of the interior of the ellipse via Green's theorem. To do this we need a vector equation for the boundary; one such equation is $\langle a \cos t, b \sin t \rangle$ ranges from 0 to 2π . We can easily verify this by substitution:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = \frac{a^2 \cos^2 t}{a^2} + \frac{b^2 \sin^2 t}{b^2} = \cos^2 t + \sin^2 t = 1.$$

Let's consider the three possibilities for P and Q : Using 0 and x gives

$$\oint_C 0 \, dx + x \, dy = \int_0^{2\pi} a \cos(t) b \cos(t) \, dt = \int_0^{2\pi} ab \cos^2(t) \, dt.$$

Using $-y$ and 0 gives

$$\oint_C -y \, dx + 0 \, dy = \int_0^{2\pi} -b \sin(t)(-a \sin(t)) \, dt = \int_0^{2\pi} ab \sin^2(t) \, dt.$$

Finally, using $-y/2$ and $x/2$ gives

$$\begin{aligned} \oint_C -\frac{y}{2} \, dx + \frac{x}{2} \, dy &= \int_0^{2\pi} -\frac{b \sin(t)}{2}(-a \sin(t)) \, dt + \frac{a \cos(t)}{2}(b \cos(t)) \, dt \\ &= \int_0^{2\pi} \frac{ab \sin^2 t}{2} + \frac{ab \cos^2 t}{2} \, dt = \int_0^{2\pi} \frac{ab}{2} \, dt = \pi ab. \end{aligned}$$

The first two integrals are not particularly difficult, but the third is very easy, though the choice of P and Q seems more complicated.

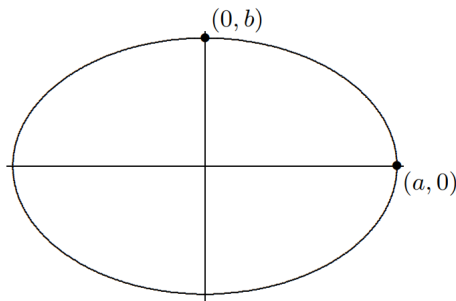


Figure 2. A “standard” ellipse, $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$.

Proof of Green’s Theorem

We cannot here prove Green’s Theorem in general, but we can do a special case. We seek to prove that

$$\oint_C P dx + Q dy = \iint_D \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA.$$

It is sufficient to show that

$$\oint_C P dx = \iint_D -\frac{\partial P}{\partial y} dA \quad \text{and} \quad \oint_C Q dy = \iint_D \frac{\partial Q}{\partial x} dA,$$

which we can do if we can compute the double integral in both possible ways, that is, using $dA = dy dx$ and $dA = dx dy$.

For the first equation, we start with

$$\iint_D \frac{\partial P}{\partial y} dA = \int_a^b \int_{g_1(x)}^{g_2(x)} \frac{\partial P}{\partial y} dy dx = \int_a^b P(x, g_2(x)) - P(x, g_1(x)) dx.$$

Here we have simply used the ordinary Fundamental Theorem of Calculus, since for the inner integral we are integrating a derivative with respect to y : an ant derivative of $\partial P/\partial y$ with respect to y is simply $P(x, y)$, and then we substitute g_1 and g_2 for y and subtract.

Now we need to manipulate $\oint_C P dx$. The boundary of region D consists of 4 parts, given by the equations $y = g_1(x)$, $x = b$, $y = g_2(x)$, and $x = a$. On the portions $x = b$ and $x = a$, $dx = 0 dt$, so the corresponding integrals are zero. For the other two portions, we use the parametric forms $x = t$, $y = g_1(t)$, $a \leq t \leq b$, and $x = t$, $y = g_2(t)$, letting t range from b to a , since we are integrating counter-clockwise around the boundary. The resulting integrals give us

$$\begin{aligned} \oint_C P dx &= \int_a^b P(t, g_1(t)) dt + \int_b^a P(t, g_2(t)) dt = \int_a^b P(t, g_1(t)) dt - \int_a^b P(t, g_2(t)) dt \\ &= \int_a^b P(t, g_1(t)) - P(t, g_2(t)) dt \end{aligned}$$

which is the result of the double integral times -1 , as desired.

Calculation of Area with Green's Theorem

If we choose special expressions F_1 and F_2 , we can relate area to the line integrals. For example, if $F_1 = 0$ and $F_2 = x$ then $\oint_C F_1 dx + F_2 dy = \iint_D \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right)$ becomes:

$$\oint_C x dy = \iint_D 1 dA = \text{Area}(\mathcal{D})$$

This is not the only choice for F_1 and F_2 that will result in area. For example, if $F_1 = y$ and $F_2 = 0$, then Green's Theorem becomes:

$$\oint_C y dx = \iint_D -1 dA = -\text{Area}(\mathcal{D})$$

Subtract these two integrals and the result is:

$$2 \text{Area}(\mathcal{D}) = \oint_C -y dx + x dy$$

Divide by 2 and we get:

$$\text{Area}(\mathcal{D}) = \frac{1}{2} \oint_C -y dx + x dy$$

Example: Use a line integral to find the area inside an ellipse.

Solution: The parametric equations $x = a \cos t$, $y = b \sin t$ for $0 \leq t \leq 2\pi$ describe the ellipse. By Green's Theorem:

$$\begin{aligned} \text{Area}(\text{Ellipse}) &= \frac{1}{2} \oint_C -y dx + x dy \\ &= \frac{1}{2} \int_0^{2\pi} (-b \sin t)(-a \sin t dt) + (a \cos t)(b \cos t dt) \\ &= \frac{1}{2} \int_0^{2\pi} ab (\sin^2 t + \cos^2 t) dt \\ &= \frac{ab}{2} \int_0^{2\pi} dt \\ &= \pi ab \end{aligned}$$

Stokes' Theorem to Evaluate

$$\iint_S \vec{\nabla} \times \vec{F} \cdot d\vec{S}$$

$$\vec{F}(x, y, z) = \tan^{-1}(x^2 y z^2) \hat{i} + x^2 y \hat{j} + x^2 z^2 \hat{k}$$

S is the part of the cone

$$x = \sqrt{y^2 + z^2}, 0 \leq x \leq 2$$

CORE CONCEPT: LINE INTEGRAL IN STOKES'S THEOREM

Stokes Theorem is a declaration about the integration of differential forms on manifolds, which both

generalizes and simplifies several theorems from vector calculus. As per this theorem, a line integral is related to a surface integral of vector fields.

Recall that one version of Green's Theorem is

$$\int_{\partial D} \mathbf{F} \cdot d\mathbf{r} = \iint_D (\nabla \times \mathbf{F}) \cdot \mathbf{k} dA.$$

Here D is a region in the x - y plane and $\mathbf{k}$ is a unit normal to D at every point. If D is instead an orientable surface in space, there is an obvious way to alter this equation, and it turns out still to be true:

Stokes's Theorem

Provided that the quantities involved are sufficiently nice, and in particular if D is orientable,

$$\int_{\partial D} \mathbf{F} \cdot d\mathbf{r} = \iint_D (\nabla \times \mathbf{F}) \cdot \mathbf{N} dS,$$

if ∂D is oriented counter-clockwise relative to $\mathbf{N}$.

Note how little has changed: $\mathbf{k}$ becomes $\mathbf{N}$, a unit normal to the surface, and dA becomes dS , since this is now a general surface integral. The phrase “counter-clockwise relative to $\mathbf{N}$ ” means roughly that if we take the direction of $\mathbf{N}$ to be “up”, then we go around the boundary counter-clockwise when viewed from. This description is inadequate. A slightly more complicated but general description is this: imagine standing on the side of the surface considered positive; walk to the boundary and turn left.

Example

Consider the cylinder $\mathbf{r} = \langle \cos u, \sin u, v \rangle$, $0 \leq u \leq 2\pi$, $0 \leq v \leq 2$, oriented outward, and $\mathbf{F} = \langle y, zx, xy \rangle$. We compute

$$\iint_D \nabla \times \mathbf{F} \cdot \mathbf{N} dS = \int_{\partial D} \mathbf{F} \cdot d\mathbf{r}$$

in two ways.

First, the double integral is

$$\int_0^{2\pi} \int_0^2 \langle 0, -\sin u, v-1 \rangle \cdot \langle \cos u, \sin u, 0 \rangle dv du = \int_0^{2\pi} \int_0^2 -\sin^2 u dv du = -2\pi.$$

The boundary consists of two parts, the bottom circle $\langle \cos t, \sin t, 0 \rangle$, with t ranging from 0 to 2π , and $\langle \cos t, \sin t, 2 \rangle$, with t ranging from 2π to 0. We compute the corresponding integrals and add the results:

$$\int_0^{2\pi} -\sin^2 t dt + \int_{2\pi}^0 -\sin^2 t + 2\cos^2 t = -\pi - \pi = -2\pi,$$

as.

An interesting consequence of Stokes's Theorem is that if D and E are two orientable surfaces with the same boundary, then

$$\iint_D (\nabla \times \mathbf{F}) \cdot \mathbf{N} dS = \int_{\partial D} \mathbf{F} \cdot d\mathbf{r} = \int_{\partial E} \mathbf{F} \cdot d\mathbf{r} = \iint_E (\nabla \times \mathbf{F}) \cdot \mathbf{N} dS.$$

Sometimes both of the integrals

$$\iint_D (\nabla \times \mathbf{F}) \cdot \mathbf{N} dS \quad \text{and} \quad \int_{\partial D} \mathbf{F} \cdot d\mathbf{r}$$

are difficult, but you may be able to find a second surface E so that

$$\iint_E (\nabla \times \mathbf{F}) \cdot \mathbf{N} dS$$

has the same value but is easier to compute.

In Green's Theorem we related a line integral to a double integral over some region. In this section we are going to relate a line integral to a surface integral

Proof of Stokes's Theorem

We can prove here a special case of Stokes's Theorem, which perhaps not too surprisingly uses Green's Theorem.

Suppose the surface D of interest can be expressed in the form $z = g(x, y)$, and let $\mathbf{F} = \langle P, Q, R \rangle$. Using the vector function $\mathbf{r} = \langle x, y, g(x, y) \rangle$ for the surface we get the surface integral

$$\begin{aligned} \iint_D \nabla \times \mathbf{F} \cdot d\mathbf{S} &= \iint_E \langle R_y - Q_z, P_z - R_x, Q_x - P_y \rangle \cdot \langle -g_x, -g_y, 1 \rangle dA \\ &= \iint_E -R_y g_x + Q_z g_x - P_z g_y + R_x g_y + Q_x - P_y dA. \end{aligned}$$

Here E is the region in the x - y plane directly the surface D .

For the line integral, we need a vector function for ∂D . If $\langle x(t), y(t) \rangle$ is a vector function for ∂E then we may use $\mathbf{r}(t) = \langle x(t), y(t), g(x(t), y(t)) \rangle$ to represent ∂D . Then

$$\int_{\partial D} \mathbf{F} \cdot d\mathbf{r} = \int_a^b P \frac{dx}{dt} + Q \frac{dy}{dt} + R \frac{dz}{dt} dt = \int_a^b P \frac{dx}{dt} + Q \frac{dy}{dt} + R \left(\frac{\partial z}{\partial x} \frac{dx}{dt} + \frac{\partial z}{\partial y} \frac{dy}{dt} \right) dt.$$

using the chain rule for dz/dt . Now we continue to manipulate this:

$$\begin{aligned} \int_a^b P \frac{dx}{dt} + Q \frac{dy}{dt} + R \left(\frac{\partial z}{\partial x} \frac{dx}{dt} + \frac{\partial z}{\partial y} \frac{dy}{dt} \right) dt \\ = \int_a^b \left[\left(P + R \frac{\partial z}{\partial x} \right) \frac{dx}{dt} + \left(Q + R \frac{\partial z}{\partial y} \right) \frac{dy}{dt} \right] dt \\ = \int_{\partial E} \left(P + R \frac{\partial z}{\partial x} \right) dx + \left(Q + R \frac{\partial z}{\partial y} \right) dy, \end{aligned}$$

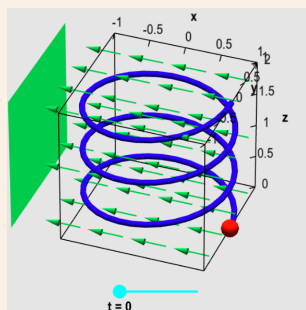
which now looks just like the line integral of Green's Theorem, except that the functions P and Q of Green's Theorem have been replaced by the more complicated $P + R(\partial z/\partial x)$ and $Q + R(\partial z/\partial y)$. We can apply Green's Theorem to get

$$\int_{\partial E} \left(P + R \frac{\partial z}{\partial x} \right) dx + \left(Q + R \frac{\partial z}{\partial y} \right) dy = \iint_E \left(\frac{\partial}{\partial x} \left(Q + R \frac{\partial z}{\partial y} \right) - \frac{\partial}{\partial y} \left(P + R \frac{\partial z}{\partial x} \right) \right) dA.$$

Now we can use the chain rule again to evaluate the derivatives inside this integral, and it becomes

$$\begin{aligned} \iint_E Q_x + Q_z g_x + R_x g_y + R_z g_x g_y + R g_{yx} - (P_y + P_z g_y + R_y g_x + R_z g_y g_x + R g_{xy}) dA \\ = \iint_E Q_x + Q_z g_x + R_x g_y - P_y - P_z g_y - R_y g_x dA, \end{aligned}$$

which is the same as the expression we obtained for the surface integral.



CORE CONCEPT: LINE INTEGRALS IN VECTOR FIELDS

A line integral (sometimes called a path integral) is the integral of some function along a curve. One can integrate a scalar-valued function along a curve, obtaining for example, the mass of a wire from its density. One can also integrate a certain type of vector-valued functions along a curve. These vector-valued functions are the ones where the input and output dimensions are the same, and we usually represent them as vector fields.

The abstraction of mathematical objects in one variable to higher dimensions can result in multiple kinds of generalizations. We run into this issue when we abstract the idea of a function: a function defined on a higher-dimensional domain can spit out a number, but it can also spit out a vector.

Recall that a function on a multidimensional domain that spits out a vector in the same space is called a vector field. Formally, a vector field is a function $F: \mathbb{R}^n \rightarrow \mathbb{R}^n$.

As we'll start in the two-dimensional case ($n = 2$) and easily generalize from there. So let C be a curve in the $x - y$ plane, and let $F: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ be a vector field. As always, our key intuition will be to think about taking a weighted sum over C , our weights were given by the numbers spit out by our scalar field f . Now, we have a function F that spits out a vector, so we need to reformulate what we mean by "weighting a chunk of curve." All, we can't do what we did and just multiply ds by F , because we would get a vector, and we want our integral to be a number. In order to remedy this, we turn to the dot product.

The multiplication of vectors can be done in two ways, i.e. dot product and cross product

Conceptualizing the Dot Product

Given vectors a and b , one way to think about the dot product $a \cdot b$ is that it is a measurement of how similar a and b are; i.e., how well they travel together. So if a and b are orthogonal (i.e. as "different" as possible), $a \cdot b = 0$, whereas if a and b are parallel, the magnitude of $a \cdot b$ will be as big as possible.

The first thing we'll do is parametrize C . We're going to have to do that eventually (remember, to compute stuff we have reduce our problem to a single variable integral), and it will make formulating the integral easier. So from now on, C is the image of the following function:

$$\mathbf{r}: [a, b] \subseteq \mathbb{R} \rightarrow \mathbb{R}^2 \quad \mathbf{r}(t) = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix}$$

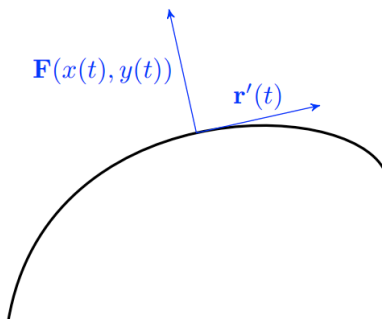
When we were looking at ds , this was like a mini tangent line to the curve. In terms of our parametrization $\mathbf{r}$, instead of calculating a mini tangent line, we can just calculate the tangent vector! So for any point $t \in [a, b]$, our analogue to ds will be $d\mathbf{r}$:

$$d\mathbf{r} = \frac{d\mathbf{r}}{dt} \cdot dt = \mathbf{r}'(t) dt = \begin{bmatrix} x'(t) \\ y'(t) \end{bmatrix} dt$$

Now, at any point $t \in [a, b]$, we can calculate the value of our vector field F at that point on the curve by inputting our position on the curve into the function: $F(\mathbf{r}(t)) = F(x(t), y(t))$. Remember that this is a vector. Our goal is to figure out how to give $d\mathbf{r}$ a weight. So what we can do is take the dot product of the vector spit out by the vector field at that point with the tangent vector to the curve:

$$\mathbf{F}(x(t), y(t)) \cdot \begin{bmatrix} x'(t) \\ y'(t) \end{bmatrix} dt$$

Visually, at every point on the curve, we have the following picture:



This represents one weighted chunk, so we add up all the chunks on the curve C to give us the line integral of the vector field! Notationally, we write:

$$\int_C \mathbf{F} \cdot d\mathbf{r} = \int_a^b \mathbf{F}(x(t), y(t)) \cdot \begin{bmatrix} x'(t) \\ y'(t) \end{bmatrix} dt$$

Note that by our interpretation of the dot product, this quantity gives a sort of measurement as to how well the curve flows with the vector field. If all of the tangent vectors are close in direction to the vector field vectors, our dot product will be larger and therefore the sum will be larger. If the tangent vectors are close to orthogonal to the vector field (so the curve is not following the flow), our dot products will be close to 0 and the line integral will be very small.

We can easily generalize this for a curve C which is the image of some function $\mathbf{r} : [a, b] \subseteq \mathbb{R} \rightarrow \mathbb{R}^n$ and a vector field $\mathbf{F} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ by just increasing the number of component functions each vector-valued function has.

CONCEPTS OF LINE INTEGRALS IN VECTOR FIELDS

A line integral in a vector field is a concept in multivariable calculus that involves integrating a vector field along a curve in a multi-dimensional space. It's used to calculate quantities like work done by a force, circulation of a fluid, or flux of a vector field through a curve.

Mathematically, let's consider a vector field $\mathbf{F} = (P, Q, R)$ and a curve C parametrized by a vector-valued function $\mathbf{r}(t) = (x(t), y(t), z(t))$, where 't' is a parameter that varies as you move along the curve. The line integral of $\mathbf{F}$ along the curve C is denoted as:

$$\int \mathbf{F} \cdot d\mathbf{r} = \int (P dx + Q dy + R dz)$$

where the dot ($\cdot$) represents the dot product, and dx , dy , dz are the differentials of the components of $\mathbf{r}(t)$ with respect to 't'.

The integral can be evaluated by substituting the components of $\mathbf{r}(t)$ into the expression for $\mathbf{F}$ and then integrating with respect to 't' over the range of the curve. The resulting value gives the cumulative effect of the vector field along the curve.

There are two common types of line integrals involving vector fields:

- **Scalar Line Integral:** This is the type of integral described above. It calculates the work done by a force field along a curve or the flux of a vector field through a curve. The scalar line integral is used when the vector field represents a force or a scalar quantity.
- **Vector Line Integral:** In some cases, you might be interested in preserving the vector nature of the integral. This is especially relevant when dealing with physical quantities like fluid flow and circulation. In this case, the integral preserves the vector components instead of summing them as a scalar. The vector line integral is denoted as:

$$\int \mathbf{F} \times d\mathbf{r} = (\int P dx + \int Q dy + \int R dz)$$

Here, each component of the vector field is integrated separately.

The choice of curve and parameterization is important when calculating line integrals. Often, the parameterization needs to be chosen in a way that makes the calculations easier. Additionally, the curve may need to be split into smaller segments if the vector field or the curve's behavior is not well-behaved throughout the entire range.

Line integrals in vector fields have applications in various scientific and engineering fields, such as physics (work, electric and magnetic fields), fluid dynamics (circulation, flux), and more generally in studying the behavior of vector fields along curves.

SUMMARY

- Line integral is an integral in which the function to be integrated is evaluated along a curve. A line integral is also called the path integral or a curve integral or a curvilinear integral.
- The notion of adding quantities up over a curve C applies not only to distance but to other physical quantities as well.
- It all comes to the same thing, but there are times that a description of the path in terms of x , y and z directly might be more convenient.
- Let's use the same vector field $\vec{F}$ and the same ellipse as the last problem, but this time, let's integrate around the entire ellipse.
- Closely associated with the circulation of a vector field is the circulation per unit area.
- There is a simple formula for finding the circulation per unit area that is in terms of the partial derivatives of the coordinates of the vector field.
- Let us suppose that $\vec{F}$ represents force so that $\int_C \vec{F} \cdot d\vec{r}$ is work. If $\vec{F}$ is a conservative vector field then the integral is path independent.
- Green's theorem gives a relationship between the line integral of a two-dimensional vector field over a closed path in the plane and the double integral over the region it encloses.
- If we choose special expressions F_1 and F_2 , we can relate area to the line integrals.
- We can prove here a special case of Stokes's Theorem, which perhaps not too surprisingly uses Green's Theorem.
- The abstraction of mathematical objects in one variable to higher dimensions can result in multiple kinds of generalizations.
- Given vectors a and b , one way to think about the dot product $a \cdot b$ is that it is a measurement of how similar a and b are; i.e., how well they travel together.

MULTIPLE CHOICE QUESTIONS

1. The area bounded by the curves $y^2 = 4x$ and $y = x$ is equal to
 - a. $1/3$
 - b. $8/3$
 - c. $35/6$
 - d. None of these
2. The area of the figure bounded by the curve $y = \log_e x$, the x-axis and the straight line $x = e$ is
 - a. $5-e$
 - b. $3+e$
 - c. 1
 - d. None of these
3. The area of the region bounded by the curve $x^2 = 4y$ and the straight line $x = 4y - 2$ is
 - a. $\frac{3}{8}$ sq. units
 - b. $\frac{5}{8}$ sq. units
 - c. $\frac{7}{8}$ sq. units
 - d. $\frac{9}{8}$ sq. units
4. The area enclosed between the graph of $y = x^3$ and the lines $x = 0$, $y = 1$, $y = 8$ is
 - a. 7
 - b. 14
 - c. $45/4$
 - d. None of these
5. The area enclosed between the graph of $y = x^3$ and the lines $x = 0$, $y = 1$, $y = 8$ is
 - a. 7
 - b. 14
 - c. $45/4$
 - d. None of these

Answers

1. (b) 2. (c) 3. (d) 4. (c) 5. (c)

REFERENCES

1. D. Cohen, E. Hairer. Linear energy-preserving integrators for Poisson systems. *BIT Numer. Math.* 51 (1) (2011) 91–101.
2. E. Celledoni, R.I. McLachlan, B. Owren, G.R.W. Quispel. Energy-Preserving Integrators and the Structure of B-series. *Found. Comput. Math.* 10 (2010) 673–693.
3. E. Celledoni, R.I. McLachlan, D. Mc Laren, B. Owren, G.R.W. Quispel, W.M. Wright. Energy preserving Runge–Kutta methods. *M2AN* 43 (2009) 645–649.
4. E. Hairer, C. Lubich, G. Wanner. *Geometric Numerical Integration. Structure Preserving Algorithms for Ordinary Differential Equations*, Second ed., Springer, Berlin, 2006.
5. E. Hairer. Energy preserving variant of collocation methods. *Journal of Numerical Analysis, Industrial and Applied Mathematics* 5,1-2 (2010) 73–84.
6. L. Brugnano, F. Iavernaro, T. Susca. Numerical comparisons between Gauss-Legendre methods and Hamiltonian BVMs defined over Gauss points. *Monografias de la Real Acedemia de Ciencias de Zaragoza* 33 (2010) 95–112.
7. L. Brugnano, F. Iavernaro. Geometric Integration by Playing with Matrices. *AIP Conference Proceedings* 1479 (2012) 16–19.
8. L. Brugnano, F. Iavernaro. Line Integral Methods which preserve all invariants of conservative problems. *Journal of Computational and Applied Mathematics* 236 (2012) 3905–3919.
9. L. Brugnano, F. Iavernaro. Recent Advances in the Numerical Solution of Conservative Problems. *AIP Conference Proc.* 1493 (2012) 175–182.
10. L. Brugnano, M. Calvo, J.I. Montijano, L. R´andez. Energy preserving methods for Poisson systems. *Journal of Computational and Applied Mathematics* 236 (2012) 3890–3904.
11. L. Brugnano. Blended Block BVMs (B3VMs): A Family of Economical Implicit Methods for ODEs. *Journal of Computational and Applied Mathematics* 116 (2000) 41–62.
12. M. Calvo, M.P. Laburta, J.I. Montijano, L. R´andez. Error growth in the numerical integration of periodic orbits, *Math. Comput. Simulation* 81 (2011) 2646–2661.
13. P. Betsch, P. Steinmann. Inherently Energy Conserving Time Finite Elements for Classical Mechanics. *Journal of Computational Physics* 160 (2000) 88–116.
14. P. Chartier, E. Faou, A. Murua. An algebraic approach to invariant preserving integrators: the case of quadratic and Hamiltonian invariants. *Numer. Math.* 103, 4 (2006) 575–590.

Index

A

abstraction of mathematical objects 243, 247
Algebra 2, 5
antiderivative 74, 80
Archimedes' spiral 52, 53
asymptotic behavior 73

C

calculus 73, 86
Cartesian system 46
Cauchy sequence 7, 8
Central angle 183
Circular functions 144, 145, 167
Comparison Test 79, 80, 81
Complex number 5, 6, 7, 8, 10, 23, 37
Complex plane 2, 3, 6, 28
Conic section 171, 174
conservative 219, 221, 225, 229, 230, 234, 236, 237, 247, 249
Constraint function 115
Continuous quantity 4
Convergence 2, 3, 4, 11, 12, 14, 19, 20, 22, 23, 24, 25, 26, 27, 29
convergent 75, 76, 77, 78, 80, 81
Coordinate system 144, 150
Curve integrals 212
Curve sketching 39

D

Decimal approximation 146
definite integral 67, 72, 81
Differentiable function 135
Differential calculus 87, 88, 89
Direct comparison test 19
Distributive law 6
Divergence 4, 14, 19, 22
divergent 75, 76, 77, 78, 79, 80, 81

E

Eccentricity 174, 180, 182, 189
Euclidean plane 173

F

Function 3, 4, 11, 12, 13, 24, 25, 26, 27, 28, 29, 30, 31, 32, 35
Fundamental Theorem of Calculus 232, 236, 239

G

Geometrical curve 183
Graph 15
Gravitational force 175
Green's theorem 236, 238, 247

H

Hyperbola 171, 172, 173, 174, 175, 176, 177, 178, 180, 187, 188, 189, 190, 191

Hypotenuse 144, 158

I

Implicit differentiation 127, 131, 137, 141

Improper integral 67

indeterminate form 68, 69, 70, 71, 85

Infinity 4, 16

Integral 3, 12, 14, 16, 17, 19, 20, 21

Integral calculus 89

Intersection 171, 172, 173, 174, 176, 177, 178, 180, 185, 187, 207

L

Lagrange multiplier 115

Laplace's method 3

Limit comparison test 19

Linear approximation 88

line integral 212, 214, 218, 220, 222, 223, 229, 231, 232, 233, 234, 236, 240, 242, 243, 245, 247

Logarithmic function 137

M

Mathematical discipline 90

Mathematical relationships 143, 144

Maxima and minima 40, 62

Mean Value Theorem of Integrals 226, 231

N

Natural function 93

Number system 149, 150

P

Partial derivative 95, 96, 97, 98, 99, 100, 101, 104, 105, 112, 119

Periodic limit cycle 4

Polar coordinate system 46, 47, 51, 52, 53, 60

Polar equations 52, 54

Polar grid 46, 47

Polynomial 2, 5, 6, 23, 24, 28, 31, 35

Population 4

positive continuous function 77

potential function 230, 232, 233, 234, 235

probability distributions 73

R

Radius of convergence 23, 25, 26, 27, 31, 35, 37

Ratio test 3, 22, 23, 25

Real number 4, 5, 7, 8, 9, 10, 37, 144, 145, 146, 148, 159, 160, 161, 162, 163, 164, 165, 167, 169

Rectangular coordinate 144

Rectangular hyperbola 180

Rolle's Theorem 32

S

Sequence 1, 3, 4, 7, 8, 11, 14, 15, 16, 17, 18, 19, 22, 37

Sequential calculation 125

Single prime 98

Stokes' theorem 236

T

Taylor's theorem 28, 31, 37

Technical skill 137

Total derivative 98, 99, 100, 101, 119

Trigonometric function 131, 134, 136, 142

V

Vector field 214, 236, 237, 247

vector-valued function 212, 245

W

Wrapping function 145

Core Concepts in Mathematics: Calculus

2nd Edition

Calculus is a field of study that focuses on understanding how things change. It provides a frame-work for modeling systems that experience change and allows us to make predictions based on these models. The development of calculus, particularly its applications in physics and engineering, has had a profound impact on the advancement of modern science. In fact, it played a pivotal role in the industrial revolution and subsequent scientific breakthroughs. No other branch of mathematics boasts such a rich history and cultural significance as calculus. From its foundational principles rooted in geometry, such as Pythagorean Theorem as part of vector calculus, to its application in measuring volumes inspired by Archimedes, calculus has been instrumental in addressing questions about our place in the universe. By analyzing velocities and forces, it enabled advancements in astronomy and aided our exploration of celestial bodies like planets, galaxies, and phenomena like black holes. The quest to understand our role within the vast universe itself has propelled us to embrace calculus. Moreover, calculus holds vital importance for our survival as it equips us with tools to develop climate models or predict population growth. It helps us comprehend disease spread dynamics and enables conflict resolution mechanisms alongside tackling economic crises effectively. Throughout history, humanity's reliance on calculus has only grown stronger as we continue pushing past boundaries and seeking greater insight into natural phenomena - making this field an indispensable pillar upon which our collective progress rests. This book, Core Concepts in Calculus 2nd edition is organized into Eight Chapters. The book provides a succinct account of the essentials of calculus, presented in an engaging style that is both readable and mathematically precise. All chapters are revised and restructured including newly added chapters.

Diana Dias is highly accomplished in calculus, having obtained a doctoral degree in the field. Her research interests span various areas of mathematics, such as algebraic topology and homotopy theory. Specifically, she is adept at working on problems related to combinatorial topology, differential geometry, and geometric group theory.

ISBN 978-1-98469-080-7

