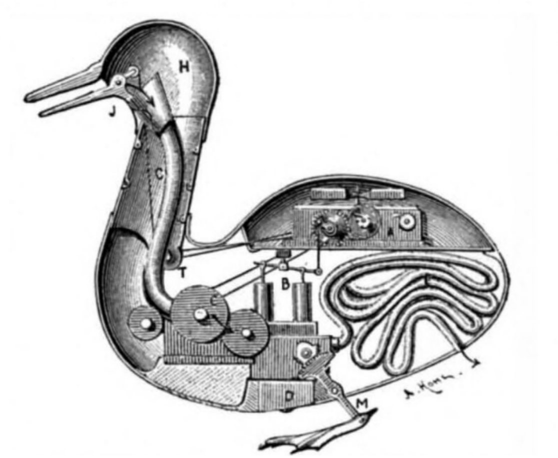
NICOLAS LAOS
Mathematician – Philosopher of Science
Knight of Dignity, Fellow of the Royal Society of Arts

MY LECTURES ON PURE AND APPLIED MATHEMATICS AND EPISTEMOLOGY



The Dignity Order Lectures Series
Vienna, 2023

Statement about this project:

This is a collection and systematic presentation of a series of lectures on pure and applied mathematics and epistemology that I wrote and presented during the academic year 2022–23 in the context of a relevant Laboratory that I organized for scholars and professional technocrats and in honor of the Grand Master of the Dignity Order, Professor Giuliano Di Bernardo, who has graciously offered me the spiritual charity of this prestigious Order. The first edition of these lectures of mine was prepared in December 2023 on the occasion of my Fellowship in the Royal Society of Arts (London). I am herewith officially stating that I personally maintain exclusive international copyright on the entire material contained in this publication (for every language).
Nicolas Laos
P.O. Box 9316, Postal Code 10032, Athens, Greece
Email: nlaosoffice@gmail.com

Cover image: Vaucanson's Automatic Duck (source: Wikimedia Commons: Author: A. Konby; https://commons.wikimedia.org/wiki/File:Digesting_Duck.jpg).

Dignity Orden zum Schutz der Menschenwurde
1220 Wien, Wagramer Strasse 19
Osterreich
https://www.dignityorder.com/en/the-order

“These lectures combine logical precision and intuition, and they provide a firm foundation in basic principles of pure and applied mathematics, as well as in mathematical philosophy.”
–**Dr. Giuliano Di Bernardo**, Professor Emeritus of Philosophy of Science and Logic at the Università degli Studi di Trento, Member of the Académie Internationale de Philosophie des Sciences (Brussels), and Founder and Grand Master of the Dignity Order.

Contents

Preface
Introduction: Mathematical Philosophy
Chapter 1: Mathematical Logic
Chapter 2: The Structure of Number Sets, Arithmetic, and Algebra
Chapter 3: Matrices and Applications in Input-Output Analysis and Linear Programming
Chapter 4: Basic Mathematical Economics, Political Economy, and a Vision of Scientific Totalitarianism
Chapter 5: Probability and Statistics
Chapter 6: Classical Euclidean Geometry, Analytic Geometry, and Trigonometry
Chapter 7: Vectors, Vector Spaces, Normed Vector Spaces, and Metric Spaces
Chapter 8: Infinitesimal Calculus: Functions, Limits, Continuity, the Topology of $\mathbb{R}^{\boldsymbol{n}}$, Differentiation, and Integration
Chapter 9: Complex Numbers and Complex Analysis
Chapter 10: Basic Principles of Ordinary Differential Equations
Bibliography
Nicolas Laos's Photographs with Other Scientists and Philosophers

List of Illustrations

Figure 0-1: Quantitative and qualitative aspects of modeling
Figure 2-1: Irrational numbers
Figure 2-2: Thales's Theorem
Figure 2-3: Cartesian coordinates
Figure 3-1: The feasible region
Figure 4-1: Market equilibrium price
Figure 5-1: A normal (or Gaussian) distribution
Figure 6-1: An angle
Figure 6-2: Measuring angles in radians
Figure 6-3: Opposite angles formed by intersecting straight angles
Figure 6-4: Angles formed by two parallel lines by a transversal
Figure 6-5: Exterior angle
Figure 6-6: A right-angled triangle
Figure 6-7: Proof of the Pythagorean Theorem
Figure 6-8: Slope and Distance
Figure 6-9: Circle

Figure 6-10: The number circle
Figure 6-11: Trigonometric functions
Figure 6-12: Ellipse
Figure 6-13: Hyperbola
Figure 6-14: Hyperbolic functions
Figure 6-15: Parabola
Figure 6-16: Spherical coordinates
Figure 6-17: Two-dimensional polar coordinates
Figure 6-18: The line of best fit
Figure 7-1: A vector in the xy-plane and its components
Figure 7-2: The dot product of two vectors
Figure 7-3: The cross product of two vectors
Figure 7-4: The shape of the unit sphere
Figure 7-5: A great circle passing through two points
Figure 7-6: Great circles and a spherical triangle
Figure 7-7: Latitude and Longitude on the Globe
Figure 7-8: Parallel lines in hyperbolic geometry
Figure 7-9: Triangles in hyperbolic geometry
Figure 8-1: Jump discontinuity
Figure 8-2: A homeomorphism
Figure 8-3: The real projective line
Figure 8-4: A tangent line to a curve
Figure 8-5: The slope of a tangent line
Figure 8-6: A concave-up function
Figure 8-7: A concave-down function
Figure 8-8: Rolle's Theorem
Figure 8-9: Lagrange's Mean Value Theorem
Figure 8-10: An optimization model
Figure 8-11: Radius of curvature
Figure 8-12: The geometric significance of a partial derivative
Figure 8-13: A closed curve
Figure 8-14: The geometric significance of the mean value theorem for a function in two variables
Figure 8-15: Saddle point
Figure 8-16: The integral as the limit of a sum
Figure 8-17: The area between two curves
Figure 8-18: The arc length of a curve
Figure 8-19: A solid of revolution
Figure 8-20: The method of infinitesimal calculus
Figure 8-21: The unit tangent vector and the principal unit normal vector
Figure 8-22: Tangent plane to a surface at a point

Figure 8-23: A linearly polarized electromagnetic wave
Figure 8-24: Green’s formula
Figure 8-25: Möbius strip
Figure 9-1: A complex number
Figure 9-2: The complex plane
Figure 10-1: A linear mechanical oscillator
Figure 10-2: RLC series circuit

List of Tables

Table 0-1: The two sets of considerations that are intrinsic to the fundamental ontology of the world
Table 1-1: Truth tables of a 2-element Boolean Algebra
Table 3-1: An input-output matrix
Table 3-2: A technological coefficient matrix
Table 3-3: A linear programming problem

Preface

The word “mathematics” comes from the Greek word “manthānein,” which means “to learn.” Mathematics is mainly about forming ways to see problems in order to solve them by combining logical rigor, imagination, and intuition. Furthermore, mathematics is a peculiar sense that enables us to perceive realities that would otherwise be inaccessible to us. In fact, mathematics is our sense for patterns, relations, and logical connections. Mathematics, in its essence, is not so much about calculating as about understanding, and, thus, it is a way of knowing, searching for truth, thinking, and developing technology.

In general, “truth”—the pursuit of which remains at the heart of scientific endeavor—can be defined as a set of relations that determine if, and the extent to which, the representation of reality within consciousness (that is, the knowledge of reality) is in concordance with the presence of reality itself (that is, with the nature of reality).

The development of mathematical intuition depends on learning the basic concepts (thus, creating a powerful intellectual toolbox), using our intellectual toolbox in order to solve problems, and thinking creatively (rather than simply memorizing mathematical tools).

I have written and presented these lectures in order to address the following audiences:

i. *Mathematics students:* Those who study mathematics can profitably use my present lectures as a self-contained, conceptual and methodic guide and compendium of pure and applied mathematics and as a supplement to their standard textbooks in the courses of algebra, linear algebra, geometry (including classical Euclidean geometry, analytic geometry, non-Euclidean geometries, and metric geometry), infinitesimal calculus (single-variable, multivariable, and vector calculus), differential equations, and real analysis.

ii. *Natural-science and social-science students:* I have written this series of lectures in order to enable one to understand the significance of mathematical modeling (including analytic and statistical methods) both in the context of the natural sciences and in the context of the social sciences. Therefore, this series of lectures can be useful for both natural-science and social-science students, helping them to better understand the importance of mathematics in their discipline and the mathematics courses included in their curriculum.

iii. *Philosophy students:* This series of lectures contains a systematic study of mathematical philosophy, philosophy of science, and the methodology of mathematics.

iv. *Any person who would like to enhance his/her ability to understand science in general, to get a better understanding of mathematics, and to fill cognitive gaps that he/she may have in mathematics and philosophy of science.*

Regarding my competence in mathematics, I would like to acknowledge the importance of the mathematical education and scientific guidance that I received from the following professors during my studies at the University of La Verne: the renowned research mathematician Professor Themistocles M. Rassias (Ph.D./University of California, Berkeley, former Chairman of the Department of Mathematics at the University of La Verne's Athens Campus and Professor at the National Technical University of Athens) taught me Calculus I, II & III, Advanced Calculus, Linear Algebra, Differential Equations, and Number Theory, and he supervised my research work in the foundations of mathematical analysis and differential geometry (a part of the research work and the dissertation that I completed at the University of La Verne under the supervision of Professor Themistocles M. Rassias was published in 1998 as the volume no. 24 of the scientifically advanced Series in Pure Mathematics of the World Scientific Publishing Company); the highly experienced applied mathematician Professor Christos Koutsogeorgis (Ph.D./City University of New York) taught me Discrete Mathematics, Abstract Algebra, and Probability Theory with mathematical statistics; and the distinguished IT Professor Chamberlain Foes (Ph.D./Portland State University) taught me PASCAL (programming language) and introduced me to mathematical informatics and management information systems.

Moreover, my cooperation with the prominent philosopher Dr. Giuliano Di Bernardo, who held the Chair of Philosophy of Science and Logic at the Faculty of Sociology of the University of Trento from 1979 until 2010, has helped me to explore several aspects of epistemology. Epistemology is the branch of philosophy that makes knowledge itself the subject matter of inquiry, and, therefore, every conscientious scholar has to be epistemologically sensitive and informed. Furthermore, epistemology is intimately related to ontology, also known as metaphysics, which investigates the nature of existence itself as well as the degree of existence of the phenomena that appear to us (and epistemology enables us to distinguish between theorems about models and theorems about reality; this distinction is very important in applied science, where models must be not only logically valid but also empirically validated).

Regarding my interdisciplinary studies and research work, I would like to acknowledge the contribution of the following professors to my education during my studies at the University of La Verne (1992–96): the historian Professor Vassilios Christides (Ph.D./Princeton University) taught me a comprehensive set of courses on the history of world civilization; the historian Professor Paul Angelides (Ph.D./Ohio State University) taught me the courses “U.S. Intellectual History” and “Development of American Democracy”; the political scientist Professor Blanca Ananiadis (Ph.D./University of Essex) taught me European politics and political institutions; and the sociologist Professor Gerasimos Makris (Ph.D./LSE) taught me Sociology. My studies in the history of civilization in general and in the history of science in particular have enabled me to articulate a typology of cultures, and, in this context, I have to mention that my approach to cultural issues, including science, is founded on certain aspects of classical philosophy and of what we call “modernity.”

My gratitude extends to the following scholars: the political scientist Dr. Hazel Smith (Professor of International Security at Cranfield University, UK, and Fellow of the Royal Society of Arts, London) and the economist and epistemologist Dr. Michael Nicholson (Professor of International Relations at the University of Sussex), who supervised my research work in the epistemology and the mathematical modeling of International Relations and Political Economy during 1997–99 at the University of Kent’s London Centre of International Relations; as well as my colleagues at the Faculty of Philosophy of the Theological Academy of Saint Andrew (Academia Teológica de San Andrés), Veracruz, Mexico, where I completed a series of Ph.D. courses (specifically, Methodology of Philosophical Investigation I & II, Theology and Philosophy I–IV, Selected Topics in Christian Philosophy I–IV, Seminar on Investigation in Christian Philosophy I–IV, and Interpretation of Philosophical Texts I & II), and the Dean of that Theological Academy, Metropolitan Dr. Daniel de Jesús Ruiz Flores of Mexico and All Latin America of the Ukrainian Orthodox Church (Iglesia Ortodoxa Ucraniana en México) helped me to explore and appreciate the interdisciplinary nature of the scholarly disciplines of theology and philosophy, and he signed my Doctoral Degree in Christian Philosophy.

In fact, I have systematically investigated theology and philosophy in order to investigate and analyze the meaning of reality, the dynamicity and the levels of the intentionality of human consciousness, and the general process of idealization. Moreover, my studies in theology and philosophy have helped me to study and understand the intellectual history of humanity and to study science in general and mathematics in particular within the context of the history of world civilization. The central theme of theology and philosophy is condensed in the meaning of the Greek word "logos," which means both language and thought, and refers to both the efficient cause and the final cause of the beings and the things that exist in the world. Indeed, ancient Greek and Roman scholars used the term "logos" in order to refer to the creative Nature, to the Norm of conduct, and to the Rule of discourse, and, gradually, the study of these three fundamental dimensions of reality was specialized in the context of particular scientific disciplines.

In the eleventh century C.E., the first "university" in the world was founded by an organized guild of students (*studiorum*) in Bologna. In fact, the founders of the University of Bologna created the word "uni-versity," and they invented an institution called "university" in order to give an adequate account of the "uni-verse," and, since the universe comes in many aspects, they thought that the study of each aspect of the universe requires the creation of a corresponding scholarly discipline. In fact, those students, acting as a mutual aid society, hired scholars to teach them liberal arts (grammar, logic, rhetoric, geometry, arithmetic, astronomy, and music), law, theology, and *ars dictaminis* (the composition of official letters and other epistolary documents). Thus, in the context of the "uni-versity," which reflects and gives an adequate account of the "uni-verse," each scholarly discipline informs and is informed by every other scholarly discipline, and this synthetic approach to knowledge underpins the classical ideal of education.

In this presentation of my lectures, I study and delineate the following topics: Mathematical Philosophy; Mathematical Logic; the Structure of Number Sets and the Theory of Real Numbers, Arithmetic and Axiomatic Number Theory, and Algebra (including the study of Sequences and Series); Matrices and Applications in Input-Output Analysis and Linear Programming; Probability and Statistics; Classical Euclidean Geometry, Analytic Geometry, and Trigonometry; Vectors, Vector Spaces, Normed Vector Spaces, and Metric Spaces; basic principles of non-Euclidean Geometries and Metric Geometry; Infinitesimal Calculus and basic Topology (Functions, Limits, Continuity, Topological Structures, Homeomorphisms, Differentiation, and Integration, including Multivariable Calculus and Vector Calculus); Complex Numbers and Complex Analysis; basic principles of Ordinary Differential Equations; as well as mathematical methods and mathematical modeling in the natural sciences (including physics, engineering, biology, and neuroscience) and in the social sciences (including economics, management, strategic studies, and warfare problems). The option of, firstly, presenting algebra, geometry, and mathematical analysis in a new, creative, and synthetic way (emphasizing a methodical and thorough conceptual study of the subject-matter) and, secondly, combining different branches of pure and applied mathematics as well as philosophy into one self-contained course gave rise to a unique, innovative project.

I originally wrote and presented these lectures during the academic year 2022–23 in the context of a Laboratory of Interdisciplinary Mathematics and Epistemology (for both scholars and professional technocrats) that I organized inspired by, and in honor of, my philosophy mentor Professor Giuliano Di Bernardo and with the support of an international private Masonic Lodge of literati that I have created and manage on the basis of the teachings and honors that I have received from Professor Giuliano Di Bernardo in the context of the Dignity Order (which is a private exclusive membership association for the defense of the dignity of humanity, and it was founded by Professor Giuliano Di Bernardo in 2012 under Austrian Law). As an independent scholar and consultant, with several informal international scholarly affiliations, I have the opportunity to consider and study various mathematical and methodological-epistemological problems as well as other analytical issues in the context of many projects in the fields of physics, engineering, biology, economics, management, social policy, and strategic studies. Furthermore, my inspiration for writing these lectures was enhanced by the legacy of the Royal Society of Arts (London), which approved my Fellowship in December 2023 (my Fellowship No. being 8289155), as well as by my experience as an instructor at the University of Indianapolis (Athens Campus, Greece, 2012–13), where I taught epistemological and methodological issues to students of International Relations, and as an analyst in financial-services, construction, IT, and shipping companies.

A Few Preliminary Thoughts

It is due to the intentionality, or the referentiality, of consciousness, or, in other words, due to the fact that consciousness is the consciousness of its contents, that the contents of consciousness become experiences for it. In fact, as the Austrian-German philosopher and mathematician Edmund Husserl (1859–1938) has taught, consciousness not only treats the presence of experiences within itself in a critical way, but also causes their presence, as it is implied by the term “intentionality.” Intentionality is not only the ability to refer to something, but also the ability to cause something. Given that, as the French philosopher Henri Bergson (1859–1941) has taught, intentionality consists of both the ability to refer and the ability to cause, we realize that the term “intentionality” expresses the dynamism of consciousness; and the dynamism of consciousness manifests itself in the manner in which consciousness intervenes in the reality of the world and restructures it.

Furthermore, regarding the creativity of human consciousness, it should be mentioned that the American neuroscientist Benjamin Libet and his collaborators clarified an aspect of free will through their discovery that humans consciously decide to act before they even think about making the decision to act. In his book *Mind Time*, Libet maintains that free will is not only an expression of the brain’s conscious activity, but it begins earlier in the unconscious mind, and it has a power of *veto* over whether or not the action takes place.

The evolutionary history of humanity is defined by the increase in brain size, as the latter made possible the rise of consciousness, both in its primitive form and in its higher order. Higher order consciousness, in turn, made possible the birth of language and intentionality or purposefulness. Consciousness is essentially linked to intentionality, through which human beings can access external reality and enter into relationships with each other. Undoubtedly, there are conscious states that are not intentional, and there are intentional states that are not part of our consciousness. Nevertheless, the connection between consciousness and intentionality plays a crucial role in understanding human beings and history.

Intentionality can operate according to a hierarchy of relations ranging from a minimum to a maximum. The levels of this hierarchy of relations are called “orders of intentionality,” in the terminology of the prominent British cognitive anthropologist Robin Dunbar (University of Oxford). In his book *The Human Story*, Dunbar has analyzed the development of the different orders of intentionality. Specifically, bacteria and certain insects have zeroth-order intentionality, while brain-equipped organisms are conscious of their mental states. For instance, brain-equipped organisms know when they are in danger or hungry. Therefore, brain-equipped organisms have first-order intentionality. First-order intentionality means that a being is self-aware, consciously referring to itself. However, there are also types of higher-order intentionality. Intentionality can be directed towards the beliefs of other people—we say that it is second-order intentionality. In other words, in the terminology of Robin Dunbar, we can distinguish the orders of intentionality as follows: most vertebrates can recall their mental states, at least in an elementary way, that is, by knowing that they know. Organisms that know that they know have first-order intentionality. Organisms that, moreover, know that someone else knows something have second-order intentionality. Organisms that, in addition, know that someone else knows that someone else knows something have third-order intentionality. As the number of subjects in the intentionality sequence increases, so does the number of hierarchical orders. This sequence can reflexively be extended indefinitely, but, in the context of their everyday life, most people rarely reach intentionality of an order higher than fourth, and they can very hardly rise to the fifth order—that is, to the following type of reasoning: "Theodore knows that Christina believes that George thinks that Nicolas supposes that Natasha intends to do something." Fourth-order intentionality is required, at a minimum, for the development of literature that goes beyond mere narrative, because, for example, an author wants his/her readers to believe that literary hero A thinks that literary hero B intends to do something. The same level of minimum skills is required for the development of science, since doing a scientific task requires asking whether the world could exist otherwise and going beyond the level of sensory experience, and then asking someone else to do the same.

As Robin Dunbar argues in his book *How Religion Evolved and Why It Endures*, the invention of religion by the species *Homo* is one of the earliest and most impressive manifestations of humanity's ascent to very high levels of intentionality, and, indeed, religion represents an extremely advanced and complex expression of humanity's creative capacity. In fact, the ability to conceive religion is an exclusive privilege of the human species. No other biological species living on Earth can formulate anything even remotely resembling religion. Since humans are a product of evolution, we must carefully investigate the factors that may have favored the emergence of our religious impulse.

In order to explain religion as a social activity and as a social institution, we need at least fourth-order (perhaps even fifth-order) intentionality, so that we can handle syllogisms of the following type: "John supposes (1) that Mary believes (2) that John believes (3) that there is a divine being intending (4) to influence people's future (because this divine being understands people's desires (5))." Until people can interact and form a community on the basis of fourth-order (or even fifth-order) intentionality, we cannot yet speak of a fully developed religion, but only of religious beliefs. The existence of a common belief—that is, the fact that there are things that mean the same to everyone—is the keystone of religion. Hence, a true communion of words, a sharing of words as a basic characteristic of any genuine dialogue, is a major underpinning of religion.

Based on the point that, in order to *understand* religion, one needs a well-formed language and at least fourth-order intentionality (while the *creation* of a religion requires at least fifth-order intentionality), we can determine when religion made its first appearance in the evolutionary history of hominids. Specifically, in view of the foregoing, we can argue that the first appearance of religion in the evolutionary history of hominids coincided with the time of the first appearance of language. Fifth-order intentionality, associated with *Homo sapiens*, manifested itself much later, when fifth-order intentionality in conjunction with a well-formed language equipped with advanced grammar and advanced syntax expressed religion as both a social institution and a metaphysical system.

Darwin's theory of evolution favors everything that can help the species to survive. As Giuliano Di Bernardo argues in his book *The Epistemological Foundation of Sociology* (Amazon, 2021), the evolutionary advantages that the human species has derived from religion are social cohesion, social control, creative imagination (especially regarding the conception of a better world), and creative management of existential anxiety. To achieve these goals, religion uses powerful means, such as belief in immortality, metaphysics, mysticism, and rituals.

However, in the context of the ancient Greek civilization, in the Aegean, certain Greek intellectuals became aware of and highlighted the fact that the human mind can discern and differentiate itself from the surrounding body of nature and can discern similarities in a multiplicity of events, abstract these from their settings, generalize them, and deduce therefrom other relationships consistent with further experience. "Abstraction" means getting rid of what we consider unnecessary details (so that, after getting rid of unnecessary details, things that were different because of unnecessary details become identical), and, therefore, we have a non-trivial concept of "identity," on the basis of which we study the "sameness" of certain things, or we look at certain things as if they were the same. "Composition" means that we combine certain abstract objects into bigger abstract objects, so that, when we have to deal with complex problems, we need to be able to divide ("analyze") the bigger problem into smaller problems, solve them separately, and then combine the solutions together. These concepts underpin "operational structuralism," which, in turn, underpins the development of modern mathematics (by the term "operation," we mean a rule according to which we can combine any two elements of a given system). The origin of "operational structuralism" can be traced back to ancient Greek philosophy. In the context of modern science and philosophy, the scholar that put operational structuralism within a rigorous mathematical-logical setting was René Descartes, the acknowledged founder of modern analytic geometry and of modern philosophy.

Philosophy and the scientific method were invented by the ancient Greek civilization. Initially, philosophy was developed within the context of the ancient Greek mystery cult of Orphism, but, gradually, it achieved its structural autonomy from religion; and a philosophical approach to religion (that is, a reflection on religion, which is something different from religion) gave rise to theology. Hence, with the invention of philosophy, the ancient Greek civilization created a method that enables humanity to rise to the highest levels of intentionality without having to resort to religion, as well as to secure for the human beings the evolutionary advantages offered by religion without being dependent on religion. In the context of philosophy, we study truth itself, what we can know, what makes an argument rational, valid, or fallacious, the reality of being, the relationship between consciousness and the world, moral criteria, and the interplay between different scholarly disciplines in the most abstract and most rigorous way possible. Thus, the invention of philosophy by the ancient Greek civilization made the ancient Greek civilization capable of becoming the inventor of science, too. For instance, the mathematical and philosophical problems suggested and studied by Aristotle, Plato, Zeno, and Pythagoras inspired and guided the mathematical works of Eudoxus, Archimedes, Apollonius of Perga, and Nicolas d'Oresme, who were leading pioneers of infinitesimal calculus, and, in turn, the latters' achievements inspired and guided the mathematical works of Torricelli, Cavalieri, Galileo, Kepler, Valerio, and Stevin, who made decisive contributions to the development of infinitesimal calculus and its applications, and, in turn, the latters' works inspired and guided Barrow and Fermat, who developed infinitesimal calculus even further and set the stage for the systematic and rigorous formulation of infinitesimal calculus by Newton and Leibniz. For a systematic study of the importance of ancient Greek thought for the development of science and philosophy, I strongly recommend the books written and edited by the British classical scholar, educationalist, and academic administrator Sir Richard Livingstone (1880–1960).

Based on the principles of abstraction and syllogism, mathematicians study the quantitative and the qualitative relations and the forms of a space (structured set), identify various connections in the processes that take place in reality, and they formulate them in the form of logical sentences written in symbols. The heuristic role of mathematics, that is, the articulation of new results, which then acquire empirical significance and confirmation or a new interpretation, is based on the correct representation of reality by mathematical models.

A "model" is intended as a carefully and methodically simplified analogue of real-world phenomena and situations, and its deductive structure helps scientists to explore the consequences of alternative assumptions. Given that scientific modeling aims to explain how things are and why things are the way they are, as well as to analyze and evaluate alternative assumptions, it contrasts, for instance, with the use of basic statistical methods solely to summarize empirical data. Furthermore, one can experiment with the model (by changing the assumptions) when it would be epistemologically, technically, and/or morally impossible and/or too risky to experiment with the real world.

In general, we should be aware that, usually, the object of scientific investigation is not an object of the real world but an ideal image of it. For instance, physics introduces many idealized objects to use in the idealization of physics problems; such as the following: (i) "Particle": this term refers to a fundamental and universal physical object, and, when physicists use this concept, they ignore the geometric dimensions of an object in comparison with the characteristic distances of the corresponding problem. (ii) "Rigid body": in this idealized object, all possible strains are ignored. (iii) "Elastic body": in this idealized object, the remnant strain is ignored. (iv) "Inelastic body": this idealized object is incapable of sustaining deformation without permanent change in size or shape, and, in this case, elastic deformation is ignored. Moreover, physics introduces idealized physical processes, too; such as the isochoric, isobaric, isothermal, and adiabatic processes. Similarly, in microeconomics and econometrics, the "model" of a real economic phenomenon reflects the essentials, allows only for the most important interrelations and interactions, and considers idealized actors rather than real actors. In particular, the mainstream of economic theory does not deal with real businesses or interests, or real markets, but it deals with theoretically representative firms, abstract markets, and generalities like the interest rate and the flow of money, and, therefore, it predicts general (rather than individual) behavior, and, more specifically, it predicts what the consequences of different kinds of behavior will be *under certain hypotheses*.

Scientific explanation is based on the fact that real objects and phenomena themselves are so complicated and interrelated that their study and quantitative investigation with due account for all aspects, interrelations, and interactions would lead to insurmountable mathematical difficulties. Therefore, a reasonable level of idealization of concrete problems characterizes every meaningful task in the context of applied science. If applied scientists did not idealize their problems, then they could not solve a single concrete problem in full. The simplifying assumptions vary from problem to problem, but a common feature of every scientific idealization consists of the methodical identification of non-essential, secondary interrelations and interactions and the decision to ignore them. Hence, a question of criteria arises. When, in what conditions, can an interrelation or interaction be characterized as non-essential and ignored, and when not? The answer depends on the method used in analyzing the solution to a problem and on the estimate method. However, two ways of idealization are most commonly used in applied science: the introduction of idealized objects (or idealized actors) and the decision to ignore non-essential interactions and processes (once we have clearly identified them as such).

In the context of my work in mathematical modeling, I have been using two categories of mathematical models: one category of mathematical models depends on the Italian physicist and engineer Galileo's method (consisting of intuition or resolution, demonstration, and experiment), and the other category of mathematical models depends on General System Theory (originally due to the Austrian biologist Ludwig von Bertalanffy). Applied firstly to celestial mechanics, Galileo's method is characterized by a mechanistic conception, according to which formal rules ("reasons") cause behavior (in an automatic way), and it is ideally suited for the study of classical physics. In fact, the estimation of a physical phenomenon consists of finding the fundamental law governing the phenomenon and, subsequently, numerically calculating the order of magnitude of the respective physical quantity. However, applied firstly in biology and, subsequently, in certain aspects of modern physics and ecological studies, as well as in behavioral and social sciences (where formal rules ("reasons") do not necessarily cause behavior), the "working attitude" of General System Theory is that of the "open system," delineated by Ludwig von Bertalanffy in his book *General System Theory* (originally published in 1968).

The closed system, reflecting the model of thought of classical physics, is axiomatic in a way that the object of scientific research is separated from the outer environment, and the outcome results from the initial conditions. From this perspective, scientific research is concerned with the analysis of the characteristics and the quantities of the elemental components, which are held in isolation for the purpose of study. Moreover, it is based on an additive methodology that underpins the deduction of the meaning of the whole from a specific *corpus* of knowledge of the character of its elementary parts. Thus, it is characterized by "reductionism." Thinking according to this model (the "machine model") has introduced both useful and misleading insights in the study of human systems. For instance, the "machine model" provides a rigorous reference system (specifically, a platform equipped with a ruler and a clock, enabling us to determine the position of the bodies under consideration and the course of time) as well as powerful analytical methods, but the actions of living things, in general, do not fit the conceptual models of classical physics.

On the other hand, General System Theory endorses and absorbs the clarity of thought and the rigor that characterize the "machine model," but it is based on the empirically verified fact that living beings and their organizations are not collections of isolated and uniform units, the sum of which accounts for a total phenomenon. Even though there is a structural continuity between inorganic matter and organic matter, life—by transforming inorganic matter into organic matter—implies an important differentiation in matter. Some characteristic differences between inorganic matter and organic matter are the following: Firstly, inorganic matter is governed by inertia, whereas organically structured living beings sense things, react to external stimuli, and move on their own. Secondly, inorganic matter reacts according to the laws of mechanics, but the reactions of organically structured living beings manifest peculiar qualitative features that are not strictly analogous to the stimuli that cause reaction, and they depend on organic relations that govern each living being according to its structural program. Thirdly, according to the Standard Model of particle physics, the minimal constituent matter elements of inorganic bodies are uniform—that is, subatomic particles are identical (so that no exchange of two identical particles, such as electrons, can lead to a new microscopic state)—but the minimal constituent matter elements of organic matter (such as DNA) are subject to differentiations, which underpin the actualization and the manifestation of the structural program of an organic being. Fourthly, inorganic bodies are connected with each other under specific conditions in order to form chemical compounds, which are always characterized by the same quantitative data (e.g., Antoine Laurent Lavoisier's "law of conservation of matter," Louis Proust's "law of constant composition," and John Dalton's "law of multiple proportions"), but organically structured living beings exchange some of their constituent elements with some of their environment's constituent elements in the context of a dynamical process that is called assimilation. Fifthly, inorganic bodies exist in definite and fixed quantities, but organically structured living beings (specifically, "parents") create new living beings (specifically, "offspring") similar to them in the context of the reproductive process. Sixthly, with few exceptions (such as radioactive nuclides, or nuclear species which are unstable structures that decay to form other nuclides by emitting particles and electromagnetic radiation), inorganic bodies are incapable of self-transformation, but organically structured living beings follow life cycles (namely, developmental stages that occur during an organism's lifetime).

The phenomena of the living world must be modeled as open systems, in which the "components" are sets of organized actions that are maintained by exchanges in the environment, and the issue of teleology (normative action) must be explicitly addressed in the models of human systems. Therefore, the postulates that refer to the dynamism of an open system and the rules that relate means to ends must explicitly find their place in any meaningful study of the social sciences. For this reason, deontic logic, approximation theory, stochastic processes, and a dynamical approach to structural analysis play an important role in the modeling of human systems.

The Meaning of a "Conceptual Study"

Mathematics plays a very important role in the world (and this is easily and thoroughly understood in applied mathematics); and, therefore, the student of mathematics must be properly instructed to understand the true nature of mathematics. The understanding of the true nature of mathematics is a key underpinning of the progress of civilization. As Euclid has taught, the true nature of mathematics is inextricably linked to deductive reasoning: from various hypotheses (often related to the perception of the world), we logically proceed to proofs.

However, hypothetico-deductive systems, especially when we rise to very high levels of abstraction, give rise to paradoxes, that is, contradictions of understanding, contradictions of logic, contradictions of semantics, and contradictions of thinking. Paradoxes have played an important role in the development of mathematics and logic. Some well-known mathematical paradoxes are the following: Zeno's paradox, Eubulides's heap ("sorites") paradox, Epimenides's "liar paradox," Hilbert's "Grand Hotel" paradox, Russell's paradox, etc.

In the context of hypothetico-deductive systems, we have to accept axiomatic truths, and, simultaneously, we have to be ready to concede that various problems stem from these axiomatic truths. On the one hand, we have to accept the existence of mathematical truths, and, on the other hand, we have to concede that mathematical truths give rise to paradoxes and comprehension problems. A way out of this uncomfortable situation was offered by the great philosopher and logician Ludwig Wittgenstein, who explained this uncomfortable situation, especially regarding the capacity of our evidentiary tools. Specifically, Wittgenstein maintains that the limits of our language together with our perceptual skills determine the limits of our thinking, since they construct the image (intellectual representation) of the world that we can perceive. As we rise to higher and higher levels of abstraction, we must be prepared to confront paradoxes. Moreover, another great philosopher and logician, Kurt Gödel, proved that, in the world, there exist propositions whose truth is valid but unprovable by (i.e., within the context of) the formal mathematical framework that we have established. In other words, as I shall explain in Chapter 1, Gödel proved that, in a formal mathematical framework, there exist mathematical propositions that are necessarily true and simultaneously unprovable by means of the tools provided by the given formal mathematical framework (and, therefore, they urge us to expand our conceptual and mathematical toolbox). Gödel has mathematically proved that a completely formalized system of arithmetic (like a machine) is either inconsistent (leading to a contradiction) or incomplete (lacking in its axiomatic foundations). Absolute, mechanistic rigor is impossible.

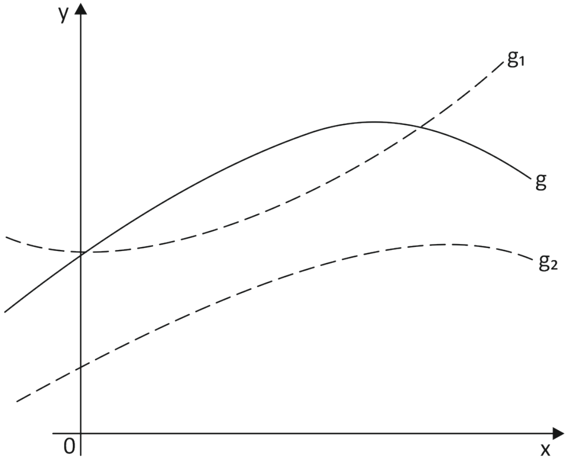
Conceptual knowledge (which is a formally constructed and linguistically expressed kind of knowledge), far from contradicting or excluding intuitive knowledge (which is a way of knowing that is more direct, immediate, and expressing a felt sense of things), is in a relationship of mutual complementarity with intuition, specifically, rational intuition (to which I shall refer in the Introduction). Conceptual knowledge is a necessary underpinning of rational intuition, and rational intuition, in turn, provides the mental readiness of the knowing subject to recognize and accept a truth that lies before him/her. This creative synthesis between conceptual knowledge and rational intuition underpins the ancient Greek notion of "epopteia," which means having seen an object in a comprehensive way ("global vision"). Moreover, as the philosopher Michael Dummett has pointedly argued, "intuition is not a special source of ineffable insight: it is the womb of articulated understanding" (Dummett, *Truth and Other Enigmas*, p. 214).

The value and the utility of mathematics do not derive from the "beauty" of mathematical formalism or from the complexity of mathematical abstractions, but from the fact that mathematics helps us to articulate representations of reality, which are useful in order to understand and/or restructure reality according to the intentionality of consciousness. I endorse the argument of the great French mathematician René Thom (1923–2002) that "what justifies the 'essential' character of a mathematical theory is its ability to provide us with a representation of reality"; and the possibility of abstracting mathematical entities from concrete situations derives from the fact that mathematics provides us with a model of the "real" (Thom, *Mathématiques essentielles*, pp. 2–3). Moreover, René Thom has brilliantly explained the term "real," which is the object of mathematical modeling, by arguing that, by the term "real," he means "both aspects of the reality of the external world—whether it is given to us by the immediate perception of the world around us, or by a mediated construction such as scientific vision" (*ibid*).

Thus, Thom's understanding of mathematics is not focused on formalism, but on a broad perspective of motion, form, and change of form, where "form" is interpreted according to Aristotle's hylozoism. According to formalism, mathematics is a game of symbols, bringing with it no more commitment to an ontology of objects or properties than chess or ludo, whereas, from the perspective of Aristotle's philosophy, mathematics is a body of propositions representing an abstract sector of reality. According to Aristotle's hylozoism, every being is composed, in an indissociable way, of matter and form, and matter is a substratum awaiting and needing to receive a form in order to become a substance, the substance of being. When Aristotle says that a being exists with regard to its substance, he refers to the "material" of which a being is composed, namely, to the "material cause" of a being. The "substantive" mode of being is complemented by form (i.e., by the "formal" mode of being), which is due to species. In his *Metaphysics*, Aristotle replaced the Platonic term "idea" with the concept of species. Form is a mode of being that is assumed by substance, and, due to its form, a being is even more sharply differentiated from every other being.

According to Thom, in the context of Aristotle's hylozoism, the notion of a bounded open set can exist as the substratum of being, whereas the notion of an unbounded open set cannot (Thom, "Les intuitions topologiques primordiales de l'aristotélisme," p. 396). Furthermore, following Aristotle's hylozoism, Thom maintains that, in mathematical modeling, the ideal of quantitative accuracy in description must always be pursued in conjunction with the ideal of qualitative accuracy in explanation. The ideal of qualitative accuracy in explanation refers to the elucidation of structure, that is, of the coherent link between the substance and the form of the phenomenon under study. In particular, Thom has considered the following case: Let us suppose that the experimental study of a phenomenon Φ gives an empirical graph g whose equation is $y = g(x)$, and that a researcher attempting to explain Φ has available two theories, say θ_1 and θ_2. In Figure 0-1, we see the empirical graph $y = g(x)$ of the phenomenon Φ, the graph $y = g_1(x)$ of theory θ_1, and the graph $y = g_2(x)$ of theory θ_2. Neither the graph $y = g_1(x)$ nor the graph $y = g_2(x)$ fits the graph $y = g(x)$ well. As shown in Figure 0-1, the graph $y = g_1(x)$ fits better *quantitatively*, in the sense that, over the interval considered, $\int|g - g_1|dx$ is smaller than $\int|g - g_2|dx$. On the other hand, Figure 0-1 clearly shows that the graph $y = g_2(x)$ fits better *qualitatively*, in the sense that it has the same shape and appearance as $y = g(x)$ (e.g., more specifically, in terms of monotonicity and curvature). Hence, René Thom argues that, in this situation, the researcher should retain θ_2 rather than θ_1 "even at the expense of a greater quantitative error," because " θ_2, which gives rise to a graph of the same appearance as the experimental result, must be a better clue to the underlying mechanisms of Φ than the quantitatively more exact θ_1 " (Thom, *Structural Stability and Morphogenesis*, p. 4).

Figure 0-1: Quantitative and qualitative aspects of modeling.



In view of the foregoing, there is a strong interplay between philosophy, logic, and mathematics; and mathematical education must include a deep understanding of mathematical concepts, the methodology of mathematics, and epistemology in general. The importance of the interaction between mathematical education and philosophical education becomes even clearer in the context of interdisciplinary mathematics.

The present series of my lectures on pure and applied mathematics and epistemology expresses my efforts to educate various groups of people in mathematical thinking and epistemology, starting from the basics. Moreover, these lectures aim to equip every aspiring person with a self-contained reference work for self-study in the fields of mathematics and epistemology.

I have taken great care in typing these lectures, which express my love for these subjects and my method of teaching in these fields. For any remaining typing errors in these lecture notes, I am wholly responsible, and I would deeply appreciate if they are brought to my notice by the readers.

Nicolas Laos
December 2023

Introduction: Mathematical Philosophy

Every scientific activity is based on consciousness, thinking, perception, memory, judgment, imagination, volition, emotion, attention, as well as intuition.

Consciousness can be construed as an existential state that allows one to develop the functions that are necessary in order to know both one's existential environment as well as the events that take place around oneself and within oneself. Thinking is based on symbols, which represent various objects and events, and it is a complex mental faculty characterized by the creation and the manipulation of symbols, their meanings, and their mutual relations. Perception is a process whereby a living organism organizes and interprets sensory-sensuous data by relating them to the results of previous experiences. In other words, perception is not static, but a developing attribute of living organisms; it is active in the sense that it affects the raw material of scattered and crude sensory-sensuous data in order to organize and interpret them; and it is completed with the reconstruction of the present (present sensory-sensuous data) by means of the past (data originating from previous experiences). Therefore, perception is intimately related to memory and judgment. Judgment is one's ability to compare and contrast ideas or events, to perceive their relations with other ideas or events, and to extract correct conclusions through comparison and contrast. Memory is one's ability to preserve the past within oneself—or, equivalently, the function whereby one retains and accordingly mobilizes preexisting impressions. Imagination is a mental faculty that enables one to form mental images, representations, that do not (directly) derive from the senses. Imagination is not subject to the principle of reality, as the latter is formed by the established institutions. Imagination develops because consciousness cannot conceive the absolute being in an objective way. Volition, or will, is one's ability to make decisions and implement them kinetically. Emotion or affect is the mental faculty that determines one's mood. Attention is a mental faculty that focuses conscious functions on particular stimuli in a selective way, and it operates as a link between perception and consciousness. Intuition means that consciousness conceives a truth (that is, it formulates a judgment about the reality of an object) according to a process of conscious processing that starts from a minimum empirical or logical datum and rises to a whole abstract system with which consciousness realizes that it is connected or to which consciousness realizes that it belongs (rational intuition, in particular, is intimately related to a type of subconscious thinking). In his *Republic*, Plato tries to define intuition as a fundamental capacity of human reason to comprehend the nature of the object of consciousness, and, in his works *Meno* and *Phaedo*, Plato understands intuition as the awareness of knowledge that previously existed in a dormant form within the mind. Moreover, David Hume, in his book entitled *A Treatise of Human Nature*, explains intuition as the power of the mind to recognize relationships (relations of time, place, and causation) without requiring further examination.

In general, philosophers are preoccupied with methodic and systematic investigations of the problems that originate from the reference of consciousness to the world and to itself. In other words, philosophers are preoccupied with the problems that originate from humanity's attempt to articulate a qualitative interpretation of the integration of the consciousness of existence into the reality of the world. The aforementioned problems pertain to the world itself, to consciousness, and to the relation between consciousness and the world.

It goes without saying that scientists are also preoccupied with similar problems. However, there are two important differences between philosophy and science. Firstly, from the perspective of science, it suffices to find and formulate relations and laws (generalizations) that, under certain conditions and to some extent, can interpret the objects of scientific research. Philosophy, on the other hand, moves beyond these findings and formulations in order to evaluate the objects of philosophical research and, ultimately, to articulate a *general method* and a *general criterion* for the explanation of every object of philosophical research. Whereas sciences consist of images and explanations of these images, philosophies are formulated by referring to wholes and by inducing wholes from parts. Hence, for instance, a philosopher will ask what is "scientific" about science, or what is the true meaning of science? Therefore, philosophy and science differ from each other with regard to the level of generality that characterizes their endeavors, and philosophy is a reflection on science. Secondly, as the French philosopher Pierre Hadot pointed out in his book *Philosophy As a Way of Life*, unlike the various scientific disciplines, philosophy is not merely a science, but it is a "way of life." More specifically, philosophy implies a conscious being's free and deliberate decision to seek truth for the sake of knowledge itself, since a philosopher is aware that knowledge is inextricably linked to the existential freedom and the ontological integration and completion of the human being. Furthermore, as I have already explained, philosophy is a reflection on science.

Knowledge

By the term "knowledge," we mean: (i) the mental action through which an object is recognized as an object of consciousness; (ii) the mental action through which consciousness conceives the substance of its object; (iii) the object whose image or idea is contained in consciousness; and (iv) that conscious content which is identified with the substance of the object of knowledge. Therefore, the term "knowledge" can be construed as a firm consideration of an object as something that corresponds to reality.

Logical knowledge, in particular, is a form of knowledge that derives from the rational faculty of consciousness, and it is characterized by indisputable and logically grounded truths (i.e., judgments about the reality of things). Rationality means the use of logical knowledge to attain goals. Logic is a theory of correct reasoning. Any relation between concepts is formulated by means of propositions. According to Aristotle's *Organon*, the "backbone" of any science is a set of propositions, so that, starting from the very primitive principles and causes, one can proceed to learn the rest. Aristotle's logic is focused on the notion of deduction (syllogism), which was defined by Aristotle, in the first book of his work entitled *Prior Analytics*, as follows: "A deduction is speech (*logos*) in which, certain things having been supposed, something different from those supposed results of necessity because of their being so"; each of the things "supposed" is a premise of the argument, and what "results of necessity" is the conclusion.

By the term "concept," we mean the set of all predicates of a thing (or of a set of conspecific things) that express the substance of the given thing (or of the given set of conspecific things). In the scholarly discipline of logic, the "intension" of a concept is the set of all predicates of the given concept, or the set of all those elements due to which and by means of which the given concept can be known and distinguished from every other concept. In other words, the intension of a concept is its formal definition. For instance, the properties of the three angles and the three sides of a geometric figure constitute the intension of the concept of a triangle. Moreover, in the scholarly discipline of logic, "extension" indicates a concept's range of applicability by naming the particular objects that it denotes. In other words, the extension of a concept encompasses all those things to which the given concept refers. For instance, the extension of the concept of a tree consists of all particular trees; the extension of the concept of a human being consists of all particular humans, etc.

By the term “genus” (plural: “genera”), we mean a concept whose extension includes other concepts, known as “species” or “kinds,” which fall within it. In other words, “genera” are concepts with an extension bigger than that of other concepts, whereas “species” or “kinds” are concepts with an extension smaller than that of other concepts. For instance, the concept of a geometric figure is a genus with regard to the concept of a triangle, whereas the concept of a triangle, which appertains to the concept of a geometric figure, is a kind with regard to the concept of a geometric figure.

Through the process of “abstraction,” we decrease the intension of concepts and increase their extension. Thus, due to abstraction, the concept of a human being can be gradually generalized into the following concepts: “vertebrate,” “mammal,” “animal,” “living being,” and “being”; “being” is the most general concept, in the sense that its intension is minimum and its extension is maximum. “Being,” to which every other concept is reducible, cannot be further analyzed into other concepts. Concepts of such general type, which are not susceptible to further analysis into simpler concepts, and to which other concepts are reducible, are called “categories.” Aristotle, in his book *Categories*, attempted to enumerate the most general species, or kinds, into which beings in the world are divided. In particular, in *Categories* (1b25), Aristotle lists the following as the ten highest categories of things “said without any combination”: “substance” (for instance, man, horse), “quantity” (for instance, four-foot, five-foot), “quality” (for instance, white, grammatical), “relation” (for instance, double, half), “place” (for instance, in the Lyceum, in the market-place), “date” (for instance, yesterday, last year), “posture” (for instance, is lying, is sitting), “state” (for instance, has shoes on, has armor on), “action” (for instance, cutting, burning), and “passion” (for instance, being cut, being burned).

No material object or system of objects—nor any connection or interaction that exists between them in material reality—is the direct object of mathematical study. In order for mathematical tools to be used to study the processes, the phenomena, and the individual objects that exist in reality, it is necessary to construct the corresponding mathematical models. By the term “mathematical model,” we mean a system of mathematical relations that symbolically describes the processes or the phenomena under study. For the construction of mathematical models, a variety of mathematical tools are used—such as: equations (algebraic, differential, and integral ones), graphs, matrices and determinants, relations of mathematical logic, geometric constructions, etc. In fact, the basic type of mathematical activity, the fundamental problem of mathematics, is the construction, the study, and the application of mathematical models.

No model can represent all the properties and all the relations of the original object. In other words, a model is a simplification, an approximate representation of the original object, and, therefore, an abstraction, but, simultaneously, a model highlights and describes an important pattern of the properties and the relations of the original object. The dialectical process of the knowledge of reality consists of two tasks: firstly, the replacement of existing models by others, which yield a more complete representation of the properties of the original object; and, secondly, the combined application of various models.

Mathematical Modeling

As I have already mentioned, mathematics is concerned with the construction of such models of objects (namely, of things, processes, and phenomena) that reflect the corresponding objects' quantitative and/or qualitative attributes as well as their spatial and structural peculiarities. For instance, geometry is the scientific study of the quantitative and the qualitative properties of spatial forms and relations (the criteria for equality of triangles provide instances of qualitative geometric knowledge, and the computation of lengths, areas, and volumes exemplifies quantitative geometric knowledge).

The constituent elements of a model are symbols and signs. Symbols are forms that express commonly accepted intentions and actions, and they can be organized into particular systems that are called codes, and the elements of such a code are called signs. In the context of mathematical modeling, the character of these signs can vary, since these signs can be schematic images (namely, shapes, drawings, and graphs), collections of numerical symbols, and elements of artificial or natural languages. Furthermore, symbols are subject to transformations according to specific symbol transformation rules. The symbols and their transformations are definitely interpreted in terms of the original objects. The combinations of symbols used and their transformations are dictated and determined by the properties of the original objects and by the relations selected and included in the corresponding model.

Mathematical models—which, with the help of the human senses, are directly extracted from material objects—usually express the primary simplest abstractions of a quantitative and spatial character, such as, for instance, enumeration, dimensions, form, position in space, etc. If a human being relies only on the sense organs, then he/she cannot achieve deep knowledge of anything. Nature, acting on the sense organs, can only produce in humans a limited set of sensations, impressions—namely, that type of knowledge which we call “empirical.”

The accumulation of empirical data constitutes the basis of generalizations and abstractions. The formulation of generalizations and abstractions provides the intellectual setting in which the application of mathematical tools becomes possible and meaningful. In the course of the historical development of mathematics, the construction of models of increasingly complex systems has been achieved, including systems that consist of multiple abstractions. With regard to its theoretical essence, mathematics can be construed as a science of modeling; and, therefore, both the reality of the world and the reality of consciousness are fundamental to mathematics.

According to such renowned mathematicians and logicians as Jacques Hadamard, Andrey Tikhonov, René Thom, Hermann Weyl, Ljubomir Iliev, Andrey Kolmogorov, and Leonid Kantorovich, the order of operations involved in the construction of mathematical models can be summarized as follows:

1. Determining and formulating the problem as clearly as possible.
2. Identification of the variable quantities that determine the process under study or are chosen for the study of the given problem.
3. Definition of the relations between these variables and the parameters on which the state of the process under study depends.
4. Formulation of a hypothesis (or hypotheses) about the nature of the conditions under study.
5. Construction of the model so that its properties coincide with the initially defined ones.
6. Conducting experimental tests.
7. Checking the hypothesis accepted for the construction of the model, and evaluating it according to the outcome of experimental tests.
8. Acceptance, rejection, or modification of the hypothesis on the basis of repeated experimental tests and conclusions.

In addition, regarding mathematical modeling, it should be mentioned that the value of mathematical modeling is not only based on quantitative accuracy but also on qualitative accuracy. By the “qualitative accuracy” of a mathematical model, I mean its ability to explain the characteristics of the structure of the phenomenon under study.

The symbolic language of mathematics is equipped with rules for handling concepts. In addition, the logical construction of mathematical models is rigorously determined in the context of, and my means of, a hypothetico-deductive system. In a “hypothetico-deductive” (or “axiomatic”) system, there are two requirements that must be met in order that we agree that a proof is correct: (i) acceptance of certain statements, called “axioms,” without proof, on the basis of their intrinsic merit, or because they are regarded as self-evident; and (ii) agreement on how and when one statement “follows logically” from another, that is, agreement on certain rules of reasoning. Inextricably linked to the aforementioned two requirements is the requirement that every person who applies hypothetico-deductive reasoning in a particular discourse understands the meaning of the words and the symbols that are used in that discourse. The more consistent and the more complete a hypothetico-deductive system is, the more its imposition is safeguarded. By the term “consistency,” we mean that the axioms of a hypothetico-deductive system neither contain nor produce contradictions. By the term “completeness,” we mean that the truth value of any proposition that belongs to a hypothetico-deductive system can be determined within the given hypothetico-deductive system (that is, according to the terms and the rules of the given hypothetico-deductive system). All these are philosophical questions.

In general, there is a close affinity between mathematics and philosophy. Mathematics, like philosophy, is created by consciousness. Mathematics provides a model of knowledge of a particular kind, and, in fact, philosophers have highlighted the particular nature of mathematical knowledge and have argued that all knowledge could possibly aspire to the particular nature of mathematical knowledge. According to the German mathematician and philosopher Friedrich Ludwig Gottlob Frege, unlike other kinds of knowledge, mathematical knowledge is characterized by rigor and objectivity, because mathematics is constituted as a logical system.

The Nature and the Structure of Mathematical Knowledge

Firstly, we have to consider mathematical Platonism, because Plato articulated a systematic philosophy founded on the principle of reasonableness in thought, rather than empirical rules, and he articulated a systematic theory of being (“ontology”). In fact, every philosophical activity is fundamentally concerned with the study of being. In the context of philosophy, the term “being” is almost always construed as a self-sufficient reality that is sustained either by being a closed system or by being an open system.

According to mathematical Platonism, numbers are forms, specifically, abstract, objectively existing objects. This thesis seems to be corroborated by the fact that numbers are not intrinsic characteristics of objects, but they are applicable to objects, and they seem to be the contents of objective truths, irrespective of any contingency and any particular object of the sensible world. From this perspective, numbers are objects themselves. In particular, according to mathematical Platonism, numbers are a peculiar kind of objects, since they exist objectively, but they cannot be grasped by the senses, they are not part of the material space-time, and they are not subject to the laws of material space-time. Far from negating the thesis that numbers are objects, the fact that numbers are not subject to the spatio-temporal structure of our sensible world corroborates the Platonic thesis that the world of forms is the reality *par excellence*, which underpins the logical constitution of our sensible world, which, in Platonic parlance, can be regarded as a "shadow" of the world of forms. This reasoning underpins the Platonic argument that, whereas the knowledge that is provided by the senses is subject to revision, the knowledge that is provided by forms, such as numbers, is incorrigible; and, therefore, reason ("logos"), which consists of thought and language, is superior to the senses. This is how mathematical Platonism explains the peculiar characteristics of the mathematical truth—namely, the certainty, the structural stability, and the necessity of the mathematical truth. From Plato's perspective, "truth" implies the concordance between a being or thing and its idea (the respective beingly being, or eternal and archetypal form), so that a being or thing is true if, and to the extent that, it is in concordance with its idea.

Mathematical Platonism is a variety of dualistic realism. In philosophy, the term "realism" refers to a philosophical model that is based on objectively existing objects, thus giving primacy to a consciousness-independent world, as opposed to "idealism," which gives primacy to the reality of consciousness. According to philosophical realism, the fact that experience furnishes consciousness with images—even unrelated to each other—of a reality that seems to lie outside the dominion of consciousness implies that the reality of the world is the cause of the particular images of the world that are present within consciousness. From the realist perspective, the principle of causality points us in the direction of the claim that the autonomous existence of reality is naturally and logically necessary. Even though the aforementioned reasoning is sound, dualistic realism, with its doubling of the world, leads to contradictions and logical gaps, especially regarding the existence of, and the relationship between, the world of forms and the world of "shadows," namely, their sensible images.

Aristotle attempted to overcome the contradictions and the logical gaps of Plato's dualistic realism by reformulating dualistic realism in a way that bridges the gap between the world of forms and the human mind. In particular, Aristotelianism highlights the structural mode of being.

The cohesive bond between substance and form is the structure of a being. The deepest reality of a being is its substance, while the external aspect and the existential otherness of that reality are the form of the given being—namely, an element that animates the given being—and these two elements (modes of being) concur with each other in the context of the structural mode of being. From the perspective of structuralism, Platonic realism corresponds to the *ante rem* structuralism ("before the thing"), in the sense that, according to Platonism, the ideational structure of mental life is a real but transcendent principle vis-à-vis the mind itself and the sensible world, and philosophical consciousness tries to partake and progress in the world of forms, while Aristotelian realism corresponds to the *in re* structuralism ("in the thing"), in the sense that, according to Aristotelianism, structures are held to exist inasmuch as they are exemplified by some concrete system, and the mind itself, not the world of forms, is a real and transcendent principle vis-à-vis the sensible world, and it conceives forms as abstractions. According to Plato's dualistic realism, forms are objectively existing objects, of which the objects of the sensible world are images, or "shadows." According to Aristotle's dualistic realism, forms are mental abstractions, the objects of the sensible world are material exemplifications of forms, forms are conceived by the mind, and the mind, rather than the world of forms itself, is transcendent to the sensible world. For this reason, Aristotle argued that the mind is the "entelechy"—that is, the program of actualization—of the body, generally, of the human organism.

According to mathematical Aristotelianism, mathematics refers to truths of the sensible world, in the sense that, even though numbers are not sensible things, they are properties of sensible things—specifically, abstract entities which can be predicated of sensible things. In other words, numbers are not objects themselves, they do not exist independently of objects, but they are features of objects, and they exist within objects. For instance, when we see ten people, the number ten is a property of the given collection of people that we see.

In the context of mathematical Aristotelianism, numbers are not self-subsistent forms, objects, but still numbers are properties of other things in an objective way. In general, according to Aristotle and according to Thomas Aquinas's variety of Aristotelianism (in the context of medieval scholasticism), consciousness is a passive mirror of reality, and truth refers to an objective correspondence between thinking consciousness and its object. But Descartes reversed the aforementioned relation between the intellect and its object, arguing that understanding (or intellection) is the basic reality, and that understanding is activated by conceiving itself; hence, Descartes's famous *dictum*: "cogito ergo sum," meaning "I think therefore I am." By assigning this active role to consciousness, Descartes emerged as the rigorous initiator and founder of modern philosophy.

Gradually, modern philosophy gave rise to a new general model, which is known as idealism. According to modern philosophical terminology, there are two general models whereby philosophers interpret the world: one gives primacy to the reality of the world, and it is known as philosophical realism, whereas the other gives primacy to the reality of consciousness, and it is known as philosophical idealism. According to idealism, the nature of consciousness is not totally different from or opposite to the nature of extra-conscious reality. The idealists' way of thinking can be summarized as follows: if the nature of reality was totally different from the nature of consciousness, then the human being would be unable to know reality. Thus, ultimately, idealism construes and studies the world not as something reflected in consciousness, but as an extension and a projection of consciousness outside itself and as part of consciousness.

In the nineteenth century, the German mathematician and philosopher Friedrich Ludwig Gottlob Frege departed from the traditional realist philosophy of mathematics, and, in contrast to mathematical Aristotelianism, he argued that, even though mathematical knowledge is objective, numbers are not objective, consciousness-independent properties of other things. According to Frege, any number n can be used in order to count any n-membered set, but the formulation of a claim concerning which number belongs to a set is determined by the way in which mathematical consciousness conceptualizes that set. For instance, consider the Tarot. The Tarot consists of 78 cards. Moreover, it has two distinct parts: the Major Arcana, consisting of 22 cards without suits, and the Minor Arcana, consisting of 56 cards divided into 4 suits of 14 cards each. Depending on whether we are thinking in terms of Tarot cards in general, or in terms of the Major Arcana Tarot cards, or in terms of the Minor Arcana Tarot cards, or in terms of the suits of the Minor Arcana Tarot cards, different numbers will belong to that particular set of cards. Hence, we have to decide if that particular set has the property 78, or the property 22, or the property 56, or the property 4. Similarly, a pair of shoes is one pair of shoes, but it consists of two shoes, and, therefore, we have to decide which number belongs to this physical object: the number one or the number two. Thus, according to Frege, numbers are not objective properties of objects, but objects acquire numbers as properties according to the ways in which consciousness thinks of the corresponding objects.

Frege’s argument about the active role of consciousness in mathematical creation—especially in light of Kant’s philosophy—may lead one to the conclusion that we have to do away with mathematical objectivity completely. Before explaining the way in which Frege prevented mathematical philosophy from sinking into arbitrary idealism, it is important to summarize Kant’s theses.

Immanuel Kant—who wrote the seminal book *Critique of Pure Reason* (1781/1787) and is one of the paradigmatic representatives of the European Enlightenment—formulated a theory of mathematical philosophy that is focused on the following question: given that mathematical knowledge is necessarily, intrinsically true, and, simultaneously, it is applicable to the sensible world—since the sensible world seems to conform to the laws of arithmetic, which transcend the sensible world—how is it possible to know something about the world that is necessarily true, or, in other words, how can we have knowledge of the world independently of recourse to experience? In order to tackle this question, Kant distinguished between two kinds of sentences: analytically true sentences and synthetically true sentences.

An analytically true sentence is necessarily true on purely logical grounds—that is, solely in virtue of its meaning—and, in reality, it elucidates meanings already implicit in the subject. For instance, the sentence “Pediatricians are medical doctors who specialize in the medical care of infants, children, adolescents, and young adults” is an analytic statement, because it is true by definition. By contrast, the sentence “Pediatricians are rich” is not necessarily true; since it is not part of the definition of a pediatrician that a pediatrician is rich, but it is part of the definition of a pediatrician that a pediatrician is a medical doctor who specializes in the medical care of infants, children, adolescents, and young adults. The sentence “Pediatricians are rich” is a synthetic statement.

The distinction between analytic and synthetic statements is based on whether we are dealing with one concept or two concepts. If you say that “Pediatricians are rich,” you are making a synthesis of two unrelated concepts—namely, the concept of being a medical doctor specialized in pediatrics and the concept of being rich. By contrast, if you say that “Pediatricians are medical doctors who specialize in the medical care of infants, children, adolescents, and young adults,” you are not synthesizing two unrelated concepts, but you are analyzing a feature of one concept—namely, the concept of being a pediatrician.

Furthermore, Kant made another important epistemological distinction in order to clarify the manner in which we know things to be true—specifically, he distinguished between *a priori* philosophical methods and *a posteriori* philosophical methods. The major attribute of the *a priori* methods is that they are based on primitive hypotheses usually intuitively conceived and axiomatically accepted, which deductively give rise to series of syllogisms, which, in turn, lead to ultimate conclusions, which are related to the preceding propositions in a logically rigorous way. For instance, we know that "pediatricians are medical doctors who specialize in the medical care of infants, children, adolescents, and young adults" *a priori*, that is, prior to any testing and any surveying. On the other hand, *a posteriori* philosophical methods are based on empirical research. For instance, the truth value of the statement that "pediatricians are rich" can only be determined *a posteriori*, that is, on the basis of doing some empirical research.

In view of the aforementioned Kantian epistemological distinctions, analytic statements are *a priori*, and synthetic statements are *a posteriori*. But mathematical knowledge exhibits the following peculiar feature: it is necessarily true, and, therefore, *a priori*, but, simultaneously, it is true of the world, and, therefore, *synthetic*. In fact, Kant observed the following peculiarity of mathematical knowledge: it is synthetic *a priori*. In other words, according to Kant, mathematical propositions, such as "$1 + 2 = 3$," are synthetic statements, abstractions from the sensing of objects, and, yet, they are *a priori*, in the sense that we do not need to do any experiments in order to verify them. Thus, Kant came up with the following question: how can we know things that are synthetic *a priori*? In order to answer this question, he developed a whole system of metaphysics that he called transcendental idealism and expounded in his *Critique of Pure Reason*.

Kant's metaphysical system is founded on the thesis that we do not know, and cannot know, the essence of things, the things-in-themselves, which he called "noumena"—meaning objects or events that exist independently of human sense and/or perception—but we can only know things as they appear to consciousness, which are called "phenomena." In Kant's philosophy, a phenomenon is a faded, dissolved declaration of the corresponding noumenon, the manner in which the corresponding noumenon (thing-in-itself) appears to an observer. According to Kant, phenomena have been put through a kind of mental filter, which is the way in which consciousness perceives the world, and mathematics is that kind of mental filter. In particular, Kant maintains that geometry is the spatial form through which consciousness perceives the world, and arithmetic—specifically, the one-dimensional sequence of numbers—is the temporal form through which consciousness perceives the world. Hence, according to Kant, we do not receive mathematics from the system of space-time itself, but we use mathematics, our spatio-temporal intuitions and intellectual "glasses," in order to understand and organize the world, and this is the reason why mathematics is *a priori*. Geometry is the way in which we organize space, and arithmetic is the way in which we organize time, and, when we combine geometry with arithmetic, we obtain the intellectual framework of the spatio-temporal world that we experience, which is relevant and meaningful because there is a structural continuity between the reality of the world and the reality of consciousness.

In his *Transcendental Aesthetic*, Kant refers to the followers of Newton's position as the "mathematical investigators" of nature, who contend that space and time "subsist" on their own; and he refers to the followers of Leibniz's position as the "metaphysicians of nature," who think that space and time "inhere" in objects and their relations. At the ontological level, Kant's position is that space and time do not exist independently of human experience, but they are "forms of intuition" (i.e., conditions of perception imposed by human consciousness). In this way, he managed to reconcile Newton's and Leibniz's arguments: he agrees with Newton that space is an irrefutable reality for objects in experience (i.e., for the elements of the phenomenal world, which are the objects of scientific inquiry), but he also agrees with Leibniz that space is not an irrefutable reality in terms of things-in-themselves. At the epistemological level, unlike David Hume, Kant argues that the axioms of Euclidean geometry are not self-evident or true in any logically necessary way. For Kant, the axioms of Euclidean geometry are logically synthetic, that is, they may be denied without contradiction, and, therefore, consistent non-Euclidean geometries are possible (as Lobachevski, Bolyai, and Riemann actually accomplished). However, Kant argues that the axioms of Euclidean geometry are known *a priori*, specifically, they depend on our intuition of space, that is, space as we can imaginatively visualize it.

After the publication of Kant's philosophical works, numerous attempts have been made to articulate methods of philosophical research that synthesize idealism and positivism, or that at least combine aspects of idealism and positivism with each other. Kant has correctly highlighted and elucidated the active role of consciousness in cognition, and the distinction between cognition and the object of cognition. The distinction between cognition and the object of cognition plays a central role in the so-called analytic philosophy. However, analytic philosophy may lead to an impasse, because it urges one to repeat the distinction between cognition and the object of cognition *ad infinitum* (forever). Inherent in analytic philosophy is the risk of using Kantian philosophy in an abortive way, in the sense that the attempt to define the presuppositions of the presuppositions of philosophy can continue *ad infinitum*, annihilating epistemology. To mitigate this risk, Kant resorted to a formalist view of idealism: Kant's *Critique* is characterized by formal idealism, in the sense that it maintains that the *form* of objects is due to consciousness, but not their *matter*. Furthermore, following Wittgenstein, and in order to avoid the excesses of analytic philosophy, particularly, scepticism, I would say that, at some point, a mature philosophical-scientific mind must make a final, epistemologically responsible decision, instead of transforming philosophy into a meaningless "language game." Wittgenstein has compared the sceptic with someone who looks for an object in a room and acts as follows: he opens a drawer and sees that it is not there; he closes the drawer, waits, and then he opens it again to see whether by chance the object is there; and he continues in this way, that is, he obsessively opens and closes a drawer looking for something that is not there. According to Wittgenstein, sceptical doubt is not true doubt, but an obsession, because true doubt, somehow, comes to an end. Regarding the reality of the external world, I should mention that the very fact that the object of cognition, the world, exhibits a kind of resistance to cognition (and, thus, consciousness has to try hard in order to know the world and impose the intentionality of consciousness on the word) implies that—even though, under certain conditions, the world is submissive to the intentionality of consciousness—the world is not merely a projection and an extension of consciousness.

The way in which Frege attempted to do justice to the objectivity of mathematics and to the reality of the world was logicism, which, as I mentioned earlier, brings together logic and arithmetic. Logicism resorts to Plato's philosophical realism regarding the objectivity of mathematics, but logicism differs from classical Platonism in two ways. Firstly, in contrast to classical Greeks, Frege and logicism in general regard arithmetic, rather than geometry, as the foundational branch of mathematics, because of the following two reasons: in the seventeenth century, Descartes's analytic geometry, adapting Viète's algebra to the study of geometric loci, showed that algebra can be used in order to model geometric objects in a systematic and rigorous way, thus establishing a correspondence between geometric curves and algebraic equations; and, in the nineteenth century, Nikolai Ivanovich Lobachevski, János Bolyai, and Bernhard Riemann invented rigorous and consistent alternatives to Euclidean geometry. Hence, for Frege and the logicists in general, the central problem in mathematical philosophy is to understand the meaning of a number. In particular, logicists endow arithmetic with the objectivity that characterizes Platonic forms, but they do so in an indirect way—through logic—trying, in a sense, to achieve a creative synthesis between Kant's transcendental idealism and Plato's philosophical realism. The role that logic plays in the "school" of logicism is the second issue with regard to which logicism differs from classical Platonism. In particular, Frege thought that we can do justice to mathematical Platonism, according to which arithmetic is about things that are forms, if we show that mathematics—particularly, arithmetic—is reducible to logic, and if we take a Platonic view of logic; hence, the name of this "school" of mathematical philosophy is logicism.

Frege fused logic and arithmetic by formulating a theory of numbers that is based on the concept of a class of objects and on structural linguistics. Hence, Frege synthesized Aristotle's work on logic and language with Plato's theory of forms. In particular, Frege thought as follows: Let us consider a variable x, meaning that x is either a symbol representing an unspecified term of a theory, or a basic object of a theory that is manipulated without referring to its possible intuitive interpretation. Thus, given a class of sentences that have the same form, we can capture their common form by replacing their specific subjects with a variable x. For instance, given sentences such as "Plato is a philosopher," "Aristotle is a philosopher," "Kant is a philosopher," "Frege is a philosopher," etc., which have the same form, we can replace the name of the subject with a variable x, thus formulating the sentence "x is a philosopher," which captures the common form of the aforementioned sentences. In this way, we obtain a class: all the things that can satisfy the sentence "x is a philosopher," whenever we replace x with a name, belong to the class of philosophers. Hence, Plato, Aristotle, Kant, Frege, and any other person whom we could substitute for x are members of the class of philosophers.

According to Frege's terminology, whereas propositions are declarative statements that are either true or false, such as the statement "Plato is a philosopher," a statement that contains a variable x and expresses a proposition as soon as a value is assigned to x is a propositional function, such as the statement "x is a philosopher." In other words, propositions and propositional functions differ from each other by the fact that propositional functions are ambiguous, in the sense that a propositional function contains a variable whose value is unassigned. A class is the extension of a propositional function; for instance, the collection of all philosophers constitutes the extension of the propositional function "x is a philosopher," and it is a class. Frege used the so defined concept of a class in order to refer to numbers and study the foundations of arithmetic.

According to Frege, numbers are classes. In his seminal book *Basic Laws of Arithmetic* (1893, 1903), Frege explained that any number n can be used in order to count any n-membered class. For instance, the number two can be thought of as the class of all two-membered things, namely, as the class of all pairs, independently of the nature of the objects that constitute each pair. Similarly, the number three can be thought of as the class of all triplets, namely, as the class of all those things which have three members; the number four can be thought of as the class of all quadruples, namely, as the class of all those things which have four members, etc. Collect all those things which have n members, and that, according to Frege, is the number n. Notice that this way of defining numbers is substantively different from the thesis that a number is a *property* of a collection of objects, because, according to Frege's conception of numbers, a number is a particular kind of *object*, it is a class. Frege built a whole system of logic on the aforementioned concept of a class.

In order to define the concept of a natural number, in particular, Frege defined, for every two-place relation R, the concept "x is an ancestor of y in the R-series," and this new relation is known as the "ancestor relation on R." The underlying idea can be easily grasped if we interpret Frege's two-place relation R as "x is the father of y in the R series." For instance, if a is the father of b, b is the father of c, and c is the father of d, then Frege's definition of "x is an ancestor of y in the fatherhood-series" ensures that a is an ancestor of b, c, and d, that b is an ancestor of c and d, and that c is an ancestor of d. More generally, given a series of facts of the form aRb, bRc, and cRd, Frege showed that we can define a relation R^* as "y follows x in the R-series." Thus, Frege formulated a rigorous definition of "precedes," and he concluded that a "natural number" is any number of the predecessor-series beginning with 0.

Scientific Creation and the Methodology of Mathematics

As I mentioned in the Preface, by the term "truth," I mean the set of those presuppositions which constitute the conditions under which the representation of reality within consciousness (i.e., the knowledge of the real) is consistent with the presence of reality (i.e., with the nature of the real).

By the second half of the twentieth century, the borders between the "schools" of logicism (which maintains that mathematical entities can be defined in the language of symbolic logic and implies a logical approach to truth), intuitionism (which maintains that mathematical entities are mental constructs and implies a constructivist approach to truth), and formalism (which maintains that mathematical entities, irrespective of any question about their essence, can be studied as terms of a formal language modulo the equivalence relation of “provable equality”) became blurred, and none of the aforementioned three “schools” of mathematical philosophy existed separately from the others. Thus, from the middle of the twentieth century onward, mathematicians became preoccupied with new, broader epistemological debates, which, in fact, have prevailed in every scientific discipline (both in the natural sciences and in the social sciences), and they center around the following two issues: (i) the difference between “truth as a discovery” and “truth as an invention”; and (ii) the determination of the degrees of truth and the difference between “correctness” and “fallacy.”

The French epistemologist Gaston Bachelard, in his books *Le nouvel esprit scientifique* (1934) and *La formation de l’esprit scientifique* (1938), pointed out that science is a mental process that aims to create concepts that contribute to an ever closer approach to reality. The phases through which consciousness passes in the context of scientific creation are the following: firstly, an intuitive general conception of its object; secondly, an analytic distinction of the individual elements that make up the given object, and, during this phase, a rigorous evaluation of those elements takes place; and, thirdly, a synthesis of the aforementioned elements, leading to the final interpretation of the scientific object in its entirety.

According to Bachelard, the “scientific object” is constructed by the scientific consciousness, and, therefore, rather than being seen in terms of dualism and opposition, empiricism and rationalism complement each other in the context of scientific creation; and both *a priori* methods (or reason) and *a posteriori* methods (or dialectic) are parts of scientific research.

In light of my claim that a synthesis between philosophical realism and idealism is required in order to formulate a proper ontology, reality is not merely an object whose various individual manifestations are grasped in a static way by a scientist’s consciousness. On the contrary, reality is an end towards which a scientist’s consciousness is directed in a dynamical way and with the aim of annihilating the distance between consciousness and reality. This process results in the objectification of a scientific theory generated by this very process, and this objectification is the essence of scientific creation. Therefore, science, including mathematics, is both an “invention” (referring to a conscious process of planning and producing something in order to meet a specific reason) and a “discovery” (referring to the provision of observational evidence and to the development of an initial understanding of some phenomenon usually pertaining to natural occurrences).

In the context of scientific creation, deduction, induction, and analogy are the methods by which a scientist's consciousness works, depending on the nature of the scientific object in question. In its essence, deduction differs from analysis, because analysis is merely concerned with the components of a unity, whereas deduction, being based on a set of axioms, expresses a deeper causal relationship between the derived result and its reference term(s). Moreover, in its essence, induction differs from synthesis, because synthesis is an essential way of transcending certain partial data which a scientist's consciousness initially takes into account together with the very pattern sought to be realized through synthesis, whereas induction is a formal variety of transcendence in the context of which consciousness moves, through generalizations, from a series of levels representing partial data to a level of unified consideration of the similarities exhibited by those partial data. Analogy is a mental process consisting of the transition from some partial data to some other partial data (by means of the identification of similarities or differences), and, although it does not provide conclusive evidence, it reinforces the element of inventiveness.

Deduction is particularly applicable to mathematics, which presupposes the existence of an ideal reality that is differentiated according to the axioms that underpin it. However, the ideal reality of mathematics can be viewed in a unified way thanks to synthetic processes that allow each individual aspect of mathematical reality to be autonomously valid in a particular field, while at the same time being related to the other aspects of mathematical reality. When the terms of a mathematical equation are given a meaning that refers to empirical data, we move from pure mathematics to applied mathematics. Induction is particularly applicable to experimental (and, generally, applied) science, which presupposes the existence of a sensible reality that appears in the form of individual experiences. Consciousness transcends these experiences by integrating them into a larger hypothetical reality, which is based on a model created by consciousness, and consciousness aims to confirm this model through empirical tests.

In view of the arguments that I have already put forward, truth is neither a pure essence nor a pure relation (or "correspondence")—it is a dynamical and rational contemplation of the world and of consciousness, as consciousness integrates and reintegrates itself into the world. Therefore, truth should be construed neither as a discovery alone nor as an invention alone, but as the outcome of the contact and the interaction between consciousness and the reality of the world. The integration of consciousness into the world is both a volitional act and an existential necessity. However, when conscious beings integrate themselves into the world, they do not only accept the reality of the world as a substantive presence, but they also attempt to understand and interpret the reality of the world. Even when consciousness cannot enter into and partake in the reality of a particular aspect of the world or of a particular situation, consciousness can create a pertinent concept. Hence, theoretical constructs play a necessary and major role in science. Moreover, Kant has masterfully proved that scientific laws are neither connatural to reality nor innate in it, but they are kinds of relations (specifically, hypothetico-deductive systems) through which consciousness understands and interprets reality. During the process of scientific explanation, the consciousness of a scientist creates new, more complete systems of relations (namely, hypothetico-deductive systems) in order to improve one's understanding and interpretation of reality, thus replacing older, scientifically degenerating systems of relations with new ones, which have a broader and deeper explanatory domain.

In the context of the relations between the classical logical values "true" and "false" (or "untrue"), we must distinguish the "false" not only from the "true" but also from the "erroneous," the "absurd," the "irrational," and the "fallacious." The term "erroneous" means a structural and automatic lapse in reasoning that cannot be corrected. The term "absurd" also means a definite error, but, although not amenable to correction itself, it may constitute a criterion for correcting a series of syllogisms in which we deliberately place it when we use it as an instrument of reference, especially in the context of the form of argument that is called *reductio ad absurdum* (where we try to establish a claim by showing that the opposite scenario would lead to absurdity or contradiction). The term "irrational" means the conclusion of a series of syllogisms that are not logically connected to each other, and the intellectually pathological nature of each proposition to which the irrational is reduced has a perverse effect on the entire series of syllogisms that results in (converges to) the irrational.

However, understanding the concept of the fallacious in the context of logic is somewhat more complex. The difference between the "fallacious" and the "false" can be understood if we have previously understood the difference between the "correct" and the "true." The "false" is the exact opposite of the "correct," but the contrast between the "true" and the "fallacious" is neither absolute nor insurmountable. The "fallacious" is an approximation of the "true," in the sense that it lacks the element of correctness but approaches the "true," tends to the "true." The "correct" is the unique conclusion of the generally understood "true," and it encompasses the "fallacious," which is a deviation from the "correct" but is subject to correction. Fallacy means intellectual wavering with a demand for truth, whereas correctness means the precise targeting of the truth. A fallacy prepares the consciousness to reach a truth, and the conception of truth as correctness refers to the culmination of the effort to reach a truth in an absolutely accurate manner.

The concept of truth as correctness was introduced at the beginning of the twentieth century by the French mathematician Jacques Hadamard.[1] In particular, if the equations and the data of a real system (i.e., of a physical or a social problem mathematically expressed) are such that (i) the model has a solution corresponding to the data, (ii) the solution is unique, and (iii) the solution is continuously dependent on the data (meaning that a small error in the data yields a small error in the estimation of the solution), then the solution is said to satisfy these three "Hadamard's restrictions," and then the equations of the model and the data give a well-posed problem. Given the well-posed problem that corresponds to a real system, we have to find its solution. Because of the third Hadamard's restriction (which is known as "stability"), we can use the "approximation principle" as a heuristic device. Hence, given the well-posed problem P, we search for an appropriate approximation P_n of the problem P such that the solution S_n, containing the index n, of P_n can be determined. Then the solution S of P is the limit of S_n as n tends to infinity, symbolically, $S = lim_{n\to\infty} S_n$. However, as the Soviet mathematician M. M. Lavrentiev, who was a prominent member of the Soviet Academy of Sciences (Siberian Department), argues in his book *Some Improperly Posed Problems of Mathematical Physics*, many real problems of mathematical physics give rise to problems that are not well-posed in the sense of Hadamard (principally, they fail to satisfy the condition of stability); and another prominent Soviet mathematician, Andrey N. Tikhonov, developed a method of regularization of "improperly posed" (or "ill-posed") problems, and this method is known as the "Tikhonov regularization." In fact, if a problem is not well-posed in the sense of Hadamard, then it needs to be reformulated for numerical treatment, and, typically, this involves including additional assumptions, such as appropriate continuity and differentiability properties of the mathematical expression of the corresponding well-posed problem. According to Tikhonov, a well-posed problem can be defined as follows: Let X and Y be some complete metric spaces, and let Af be a function whose domain is X and whose range is Y. Consider the equation

$$Af = g. \qquad (1)$$

[1] The concept of truth as correctness is intimately related to the concept of well-posedness.

We call the problem for the solution of (1) "well-posed according to Tikhonov" if the following conditions are fulfilled: (i) It is *a priori* known that the solution f exists for some class of data and belongs to some given closed set $M \subset X$, symbolically, $f \in M$. (ii) The solution is unique in a class of functions belonging to M. (iii) Arbitrarily small changes of the right-hand side g that do not carry the solution f out of M correspond to arbitrarily small changes in the solution f. If we denote by M_A the image of M after the application to the space X of the operator A, then the third Tikhonov's requirement can be restated as follows: the solution of the equation (1) depends continuously on the right-hand side g on the set M_A. In other words, Tikhonov changed Hadamard's notion of correctness by showing that an improperly posed problem can become well-posed by introducing a sufficiently "strong" norm in the data space Y or a sufficiently "weak" one in the space X. The mathematical concepts used in this definition will be completely clarified later in this book.

Revisiting ontology, and in view of the arguments that I have already put forward, we conclude that the structure governing the constitution of objective reality cannot be understood as opposed to the structure governing that kind of existence which we call consciousness and which is linked to objective reality. The structure of physical universe, the structure of biological universe, and the structure of mental universe (the universe of consciousness) are not "one," but are unified, and the structural continuity between the physical, the biological, and the mental is manifested in the energy field, as Pierre Teilhard de Chardin has pointed out, and its center of reference is consciousness. Hence, because of the presence of consciousness in the world, and because of the potential submissiveness of the world to the intentionality of consciousness, the fundamental ontology of the world involves two sets of considerations that work in tandem and are summarized in the following table:

Table 0-1: The two sets of considerations that are intrinsic to the fundamental ontology of the world.

Reality of the world	Reality of consciousness
Reality as immediacy	Reality as mediacy (mentally mediated)
Non-intentional action	Intentional action
Experience	Abstract thinking
Reality as a set of constraints	Reality as a set of opportunities

Based on the aforementioned arguments, I have articulated a model of human creativity and historical becoming, and I have called it the "dialectic of rational dynamicity." The dialectic of rational dynamicity, as a method for the operation of consciousness and as a model of the operation of reality in general, consists of the following five stages:

Stage I: Vision and Teleology. Consciousness forms a clear intellectual image of an existential state that it wants to achieve, and it is clearly oriented towards that intellectual image. Thus, in this stage, consciousness determines the teleology of its action, and, by extension, it gives meaning to the beings and the things that exist in the world. As Edmund Husserl has pointed out, every intentional act has, as part of its formation, a correlative "meaning" (implying "thought," or "what is thought about"), which is the object of the act.

Stage II: Strategy. In general, "strategy" refers to the orientation of a conscious being in the long term, within its environment. Consciousness makes the strategic decision to act upon the reality of the world and upon itself in accordance with its teleology—that is, in order to bring about intended changes.

Stage III: Planning. Consciousness articulates a plan: a method of deliberate, self-conscious activity, involving the consideration of outcomes before choosing among alternatives. The primary functions of planning are the following: (i) optimization (namely, improving efficiency of outcomes); (ii) balancing the agent's teleology (which is aimed at restructuring reality) and the goal of maintaining the continuity of existence (namely, offsetting systemic failures); (iii) widening the range of decision-making (namely, enhancing the consciousness of choice); and (iv) organizing and enriching codes and networks of communication.

Stage IV: Control. Consciousness continuously tries to maintain control over its action (and its consequences) in two ways: firstly, by intensifying its action (its intervention in the reality of the world and in itself) whenever its action is unreasonably sub-optimal (i.e., whenever it can improve its existential conditions even more, according to its strategic plan); secondly, by counterbalancing its original action (specifically, by reversing its original action and by following alternative paths of action) whenever the "negative externalities" of its original action, the costs of its original action (for the world in general and for itself in particular), tend to exceed a critical value that represents the maximum existential risks that consciousness is determined to undertake in order to continue acting in the same way. Additionally, it should be mentioned that the term "dialectic," in general, implies a transition from one state to another without the total elimination of the previous state, in the sense that the previous state leaves its traces in the new one. Therefore, according to the dialectic of rational dynamicity, an agent of change does not bring about a totally new state, which would be uncontrolled by the agent of change. In general, change cannot go beyond certain limits without running the risk of systemic collapse. For this reason, the dialectic of rational dynamicity highlights the importance of preventing uncontrolled systemic turbulence and of continuously maintaining control over the consequences of our actions.

Stage V: Development. Consciousness seeks to ensure and enhance its capabilities and to create favorable conditions for the continuation of its action in the future. However, consciousness realizes that the achievement of its ultimate goals is a work in progress. Thus, consciousness seeks to restructure the world according to the intentionality of consciousness—without, however, jeopardizing the possibility of future interventions in the reality of the world.

Chapter 1
Mathematical Logic

By the term “deductive system,” we mean a calculus endowed with an interpretation of its terms. In logic, a “calculus” is a collection of symbols equipped with a set of rules for their manipulation. When a calculus is equipped with an “interpretation” of its terms, that is, with a set of rules that makes its terms meaningful, it becomes a deductive system. A deductive system is called “pure” if the rules of the interpretation are sufficient to establish the truth or the falsity of its constituent statements. The statements of a pure deductive system are called “L-determinate,” where L stands for the relevant formal language (the truth value of an L-determinate statement is determined in L by an interpretation of the symbols in L). For instance, logic (the science of correct reasoning) is a pure deductive system. Therefore, truths derived from pure deductive systems are based on reason alone, and they are certain because they can never be empirically refuted. If a statement cannot be assigned a truth value only according to the rules of interpretation in the relevant deductive system, then it is called “non-L-determinate.” A non-L-determinate statement is called true or false not only on the basis of the rules of interpretation in the relevant deductive system, but also on the basis of a rule of disposition by reference to empirical data. Non-L-determinate statements for which a rule of disposition by reference to empirical data has been established are called “factual statements,” while the deductive systems in which they appear are called “applied.”

A “scientific theory” is a deductive system (pure or applied) that explains generalizations (i.e., “scientific laws”) or aims to criticize and change the structure of the world and/or consciousness.

In symbolic or mathematical logic, the following symbols are used:

$\wedge$ or &: conjunction (“and”);
$\vee$: disjunction (“or”);
$\neg$: negation (“not”);
$\rightarrow$ or $\Rightarrow$: material implication (“if . . . then . . .”);
$\leftrightarrow$ or $\Leftrightarrow$: biconditional (“if and only if”);
$\forall$: universal quantification (“for every”);
$\exists$: “there exists”;
$\exists!$: “there exists exactly one”;
$\nexists$: “there does not exist”;
$P(x)$: predicate letter (meaning that x (an object) has property P);
$|$: “such that”;

$\vdash$: turnstile ($x \vdash$ y means that x "proves" (i.e., syntactically entails) y; a sentence φ is "deducible" from a set of sentences Σ, expressed $\Sigma \vdash \varphi$, if there exists a finite chain of sentences $\psi_0, \psi_1, \psi_2, \dots, \psi_n$ where ψ_n is φ and each previous sentence in the chain either belongs to Σ, or follows from one of the logical axioms, or can be inferred from previous sentences; $\nvdash$ denotes the negation of $\vdash$);

$\vDash$: double turnstile ($x \vDash$ y means that x "models" (i.e., semantically entails) y; a sentence φ is a "consequence" (i.e., an ordered list) of a set of sentences Σ, expressed $\Sigma \vDash \varphi$, if every model of Σ is a model of φ);

$B \subseteq A$: B is a "subset" of A, meaning that every element of a set B is an element of a set A;

$B \subset A$: B is a "proper subset" of A, meaning that $B \subseteq A$ and there is at least one element of A that is not an element of B;

$x \leq y$: x is less than or equal to y;

$x < y$: x is strictly less than y;

$x \geq y$: x is greater than or equal to y;

$x > y$: x is strictly greater than y;

x^n: this operation is called "exponentiation" (pronounced as "x raised to the power of n"), and it means that x is multiplied by itself n times, where $n = 0,1,2,3, \dots$; $x^0 = 1$, $x^1 = x$, $x^2 = x \cdot x$, $x^3 = x \cdot x \cdot x$, etc.;

$x^{1/n}$: this operation is called the "nth root," and it is the number whose nth power equals the given number ($n \neq 0$); $x^{1/2} = \sqrt{x}$ is the square root, $x^{1/3} = \sqrt[3]{x}$ is the third root, etc.;

( ): brackets; they are used for convenience in grouping terms together (there are specific rules for removing brackets);

$\emptyset$: the empty set;

$\wp(X)$: the power set of a set X. The "power set" of a set X is the set of all the subsets of X, including the empty set and X itself. If a set has n elements, then the number of its subsets is 2^n, and the number of its proper subsets is $2^n - 1$. For instance, if $A = \{a, b\}$, then its power set is $\{\emptyset, \{a\}, \{b\}, \{a, b\}\}$, and its proper subsets are $\emptyset$, $\{a\}$, and $\{b\}$.

Sometimes, for emphasis, instead of the equality sign, namely, $=$, we use the symbol $\equiv$, which, in this case, means "identically equal."

The English mathematician and philosopher George Boole (1815–64) realized that arguments expressed in an ordinary language (e.g., in ordinary English) can be expressed in the notation of mathematical logic and then studied in the context of "propositional calculus." For instance, consider the following argument:

- If you want to learn mathematics, then you must study methodically.
- If you must study methodically, then you must be taught an effective method of studying.
- Therefore, if you want to learn mathematics, then you must be taught an effective method of studying.

The aforementioned argument involves various propositions, which we may present by letters as follows:

P: You want to learn mathematics.
Q: You must study methodically.
R: You must be taught an effective method of studying.

These propositions can be "true" or "false." The aforementioned argument can be formalized as follows:

$P \Rightarrow Q$

$Q \Rightarrow R$

$P \Rightarrow R$

where the two propositions above the dashed line are the "premises," and the one below the dashed line is the "conclusion." The reasoning process that leads from premises to a conclusion is called a "deductive process" or just a "deduction." A "theorem" is a formula inferred by means of a rule of inference in a finite number of steps from axioms and previously inferred formulae. Those propositions where truth value is dependent on the values of the variables in them are called "predicates" (hence, we talk about "predicate calculus").

It is important to distinguish between the terms "validity" and "truth" as they are used in logic. An argument, a reasoning process, or a deduction is said to be valid (i.e., logically correct) if the truth of the conclusion follows from the truth of the premises. Notice that, if the premises are both true, then the conclusion is logically necessarily true, too. Therefore, with one or more factually incorrect premises, an argument may still be valid, although its conclusion may be false. Furthermore, a valid argument based on false premises does not necessarily lead to a false conclusion. In other words, there is a significant difference between *logical* (i.e., procedural) correctness ("validity") and *factual* correctness. If an argument is valid (i.e., logically correct), and if its premises are true (i.e., if the facts on which it is based are true), then it is said to be "sound." In logic, we focus on the validity of arguments rather than on their soundness, and this fact explains the "instrumental" role of logic in philosophy and science.

In the context of logic, truth is a structural issue. Given a language L (i.e., a collection of symbols, letters, or words with arbitrary meanings that are governed by rules and are used for communication), a structure S is an ordered pair $\langle D, I\rangle$ where: D is a non-empty set denoting the domain of discourse (it is a non-empty set of any entities), and I is an interpretation, that is, a rule that assigns to each individual element of L an element of D, and to each n-place predicate of L a subset of D^n (where D^n denotes the set of n-tuples taken from D).

A "Boolean algebra" is the six-tuple

$\langle A, \wedge, \vee, \neg, 0, 1\rangle$,

consisting of a set A equipped with two binary operations: $\wedge$ (called "meet" or "and") and $\vee$ (called "join" or "or"), a unary operation $\neg$ (called "complement" or "not"), and two elements 0 and 1 in A (called "bottom" and "top," respectively, and denoted by the symbols $\bot\, and \top$, respectively), such that the truth value of a true sentence is 1, the truth value of a false sentence is 0, and, for all elements a, b, and c of A, the following axioms hold:

i. Associativity:
$a \vee (b \vee c) = (a \vee b) \vee c$; $a \wedge (b \wedge c) = (a \wedge b) \wedge c$.

ii. Commutativity:
$a \vee b = b \vee a$; $a \wedge b = b \wedge a$.

iii. Absorption:
$a \vee (a \wedge b) = a$; $a \wedge (a \vee b) = a$.

iv. Identity:
$a \vee 0 = a; a \wedge 1 = a$.

v. Distributivity:
$a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c)$; $a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c)$.

vi. Complements:
$a \vee \neg a = 1$ and $a \wedge \neg a = 0$.

For instance, the 2-element Boolean algebra has only two elements, namely, 0 (or "False") and 1 (or "True"), and it is defined by the rules mentioned in Table 1-1.

Table 1-1: Truth tables of a 2-element Boolean Algebra.

a	b	$a\wedge b$	$a\vee b$	a	$\neg a$
0	**0**	0	0	**0**	1
1	**0**	0	1	**1**	0
0	**1**	0	1		
1	**1**	1	1		

Let $\boldsymbol{U} = \{\boldsymbol{u_1}, \boldsymbol{u_2}, \dots, \boldsymbol{u_m}, \dots\}$ be the original alphabet consisting of variables (arguments). A "Boolean function" is a rule that maps each argument in its domain to exactly one value in its range where the allowable values of range and the allowable values of domain are just one of two variables, namely, "true" (symbolized by $\mathbf{1}$) and "false" (symbolized by $\mathbf{0}$). In order to define any Boolean function, we must specify its value for each possible value of its inputs. For instance, the Boolean functions "not" ($\neg$), "and" ($\wedge$), and "or" ($\vee$) are defined as follows:

$$\boldsymbol{NOT(x)} = \begin{cases} \boldsymbol{1\ if\ x\ is\ 0} \\ \boldsymbol{0\ if\ x\ is\ 1} \end{cases}$$

$$\boldsymbol{AND(x,y)} = \begin{cases} \boldsymbol{1\ if\ both\ x\ and\ y\ are\ 1} \\ \boldsymbol{0\ otherwise} \end{cases}$$

$$\boldsymbol{OR(x,y)} = \begin{cases} \boldsymbol{1\ if\ at\ least\ one\ of\ x\ and\ y\ is\ 1} \\ \boldsymbol{0\ otherwise} \end{cases}$$

(the corresponding "truth tables" are shown in Table 1-1).

De Morgan's Laws

The following pair of transformation rules is known as De Morgan's laws (named after the nineteenth-century British mathematician Augustus De Morgan), and it is originally due to Aristotle:

The negation of a disjunction is the conjunction of the negations: $\boldsymbol{not(A\ or\ B) = (notA)and(notB)}$.

The negation of a conjunction is the disjunction of the negations: $\boldsymbol{not(A\ and\ B) = (notA)or(notB)}$.

Basic Principles of Predicate Calculus

As I have already mentioned, George Boole developed a purely symbolic system for deduction in a rigorous language of predicates (or relations, or properties), and, thus, Predicate Calculus (henceforth, PC) emerged. The formal system PC involves the following:

i. The alphabet of PC: a countable set of variables (or arguments): $v_1, v_2, v_3, \ldots$ and a two-place predicate letter P; two logical connectives: ¬ and ∧; one quantifier symbol: ∃; three improper symbols: the left parenthesis, the comma, and the right parenthesis, namely, (,), but quite often we may also use brackets [and] as well as the symbol | standing for "such that."

ii. These symbols are used in order to build the (well-formed) formulae of PC, according to the following rules:
   a. If x, y are individual variables, then $P(x, y)$ is a formula of PC.
   b. If φ, ψ are formulae of PC, then so are $(\varphi \wedge \psi)$ as well as $\neg\varphi$ and $\neg\psi$.
   c. If x is an individual variable and φ is a formula, then so is $\exists x\varphi$.
   d. Something is a formula of PC only by virtue of the aforementioned conditions (a), (b), and (c).

Remark: The alphabet contains only the logical symbols ¬, ∧, and ∃, because the other usual symbols can be defined in terms of these three as follows:
$(\varphi \vee \psi)$ is defined as $\neg(\neg\varphi \wedge \neg\psi)$,
$(\varphi \rightarrow \psi)$ is defined as $\neg(\varphi \wedge \neg\psi)$,
$(\varphi \leftrightarrow \psi)$ is defined as $((\varphi \rightarrow \psi) \wedge (\psi \rightarrow \varphi))$, and
$\forall x\varphi$ is defined as $\neg\exists x\neg\varphi$.

A variable is said to be "bounded" if it is determined by a quantifier; otherwise, it is said to be "free." For instance, in the formula $\exists xP(x, y)$, x is bounded, and y is free. If a formula of PC contains no free variables, then it is said to be a "sentence."

By an "interpretation," we mean the task of giving a certain meaning to the undefined terms of a formal system. Consider, for instance, the following sentences of PC:

i. $\forall x\forall y(P(x, y) \rightarrow P(x, y))$,
ii. $((P(x, y) \wedge P(y, z)) \rightarrow P(x, z))$, and
iii. $\forall y\exists xP(x, y)$.

If we interpret P as the ancestor relation over the domain of all (living and dead) people (and if we assume that such a relation is biologically determined in a definite way), then: the sentence i means that, "if x is an ancestor of y, then x is an ancestor of y, for every x and y," namely, it is a tautology; the sentence ii means that, "if x is an ancestor of y, and if y is an ancestor of z, then x is an ancestor of z"; the sentence iii means that, "for every y, there exists an ancestor x." Thus, the sentences i, ii, and iii are true. However, if we interpret P as $<$ ("strictly less than") over the natural numbers, then the sentence iii is false. Moreover, if we interpret P as "the father of" over the domain of human beings, then the sentence ii is false. We can easily notice that the sentence i will remain true for any interpretation of P; such sentences of PC are said to be "universally valid" (and they are tautological in character).

A "formal system" is obtained by choosing a finite set of axioms (or schemes of axioms, i.e., selected formulae) and a finite set of rules of inference in a given language. In the case of PC, we have the following axioms and the following rule of inference (φ, ψ, χ are formulae, $x, y, y_1, \dots, y_n, \dots$ are variables, and $\varphi(y)$ is the result of substituting y for all free occurrences of x in $\varphi(x)$):

Axioms of Predicate Calculus:

i. $\forall y_1 \dots \forall y_n\ (\varphi \to (\psi \to \varphi))$.

ii. $\forall y_1 \dots \forall y_n\ ((\varphi \to (\psi \to \chi)) \to ((\varphi \to \psi) \to (\varphi \to \chi)))$.

iii. $\forall y_1 \dots \forall y_n\ ((\neg\varphi \to \neg\psi) \to ((\neg\varphi \to \psi) \to \varphi))$.

iv. $\forall y_1 \dots \forall y_n\ (\forall x(\varphi \to \psi) \to (\varphi \to \forall x\psi))$, provided that φ has no free occurrence of x.

v. $\forall y_1 \dots \forall y_n\ ((\varphi \to \psi) \to (\forall y_1 \dots \forall y_n \varphi \to \forall y_1 \dots \forall y_n \psi))$.

vi. $\forall y_1 \dots \forall y_n\ (\forall x \varphi(x) \to \varphi(y))$, provided that, as we substitute the free occurrences of x in $\varphi(x)$ with y, the y's are free in $\varphi(y)$, that is, they are not determined by quantifiers already occurring in φ.

The rule of inference for predicate calculus is *modus ponens* ("method of affirming," or proof by affirming the antecedent), which has the following form: from φ and $(\varphi \to \psi)$, infer ψ. In other words, if a conditional statement ("if φ then ψ") is accepted, and the antecedent (φ) holds, then the consequent (ψ) may be inferred.

Remark: Another very famous and important rule of inference (logical argument) is *modus tollens* ("method of denying," or proof by contrapositive), which has the following form: from $\neg\psi$ and $(\varphi \to \psi)$, infer $\neg\varphi$.

A "theorem" is a formula inferred by means of a rule of inference in a finite number of steps from axioms and previously inferred formulae. Hence, we are faced with the problem of determining that finite set of axioms (or schemes of axioms) from which the rule of inference will give only true sentences.

Formalism, Structuralism, and Mathematical Modeling

The formalist approach to mathematics maintains that, in order to analyze a mathematical text, it suffices to study its formal devices, mainly, its syntax. Hence, according to formalism, mathematical statements are statements about the consequences of the manipulation of strings (i.e., alphanumeric sequences of symbols, usually presented in the form of equations) using established rules of inference (by a "rule of inference," we mean a logical form consisting of a function that takes premises, analyzes their syntax, and returns a conclusion). In other words, according to formalism, mathematics does not consist of propositions representing an abstract sector of reality, but it is actually a game of symbols, without bringing with it more ontological commitments than, for instance, chess.

In the 1930s, the great Austrian mathematician and logician Kurt Gödel undertook to evaluate the logical rigor of formalism. Broadly speaking, Gödel considered a statement of the type

$P =$ "This statement is false,"

which leads to the following complicated situation: if $P =$ "This statement is false" is true, then it is false, but the sentence asserts that it is false, and, if it is, indeed, false, then it must be true, and so on. The earliest study of problems pertaining to self-reference in logic is due to the seventh-century B.C.E. Greek philosopher and logician Epimenides, who formulated the classical "liar paradox." Gödel's Incompleteness Theorem shows that such complicated situations can occur in any theory that is consistent and comprehensive enough to contain elementary arithmetic as the latter has been encoded by Peano's axioms for natural numbers (see Chapter 2). Consequently, logic is necessary and capable of organizing every mathematical and, generally, scientific theory, but logic is not sufficient to completely organize itself. According to Gödel, human consciousness, in general, and thought processes, in particular, are not merely algorithmic. Gödel established the following argument mathematically: Either the human mind (even within the realm of pure mathematics) infinitely surpasses any finite machine (algorithmic process), or else there exist absolutely undecidable arithmetic propositions (see: Shanker, ed., *Gödel's Theorem in Focus*).

Formalism rightly stresses the importance of syntax and, particularly, of logical consistency, but it cannot stand as a general theory of the epistemology of mathematics or any other scientific discipline. Therefore, we have to turn from formalism to structuralism. Structuralism is concerned with the analysis of the underlying structures in a text. The structure of a mathematical text can be explained and described as follows: Let C denote the set of all basic conceptual objects (i.e., the "universe" of concepts), R the set of all basic conceptual relations, and A the set of the axioms of a structure. Then the corresponding structure is denoted by $\mathcal{S}(C,R,A)$. A segment of a structure is a set of concepts, definitions, and judgments of the given structure, it satisfies the axioms of the given structure as well as some additional conditions, and it is denoted by $\bar{\mathcal{S}}(\bar{C},\bar{R},\bar{A})$. Suppose that a phenomenon of the sensible world has been described by a structure $\mathcal{S}(C,R,A)$, or by a segment of this structure. Both the phenomenon and its mathematical model can be regarded as two isomorphic models, since the original phenomenon is initially modeled by our perception of it. More precisely, it is modeled by the initial reference of our consciousness to it, and its mathematical model is $\mathcal{S}(C,R,A)$ or a segment of $\mathcal{S}(C,R,A)$.

The creation of isomorphisms between mathematics and other scientific disciplines or human activities is called mathematical modeling. Thus, mathematical modeling consists of two stages: (i) the formulation of the mathematical model of the object that one studies—that is, the transformation of the given problem into a mathematical one—and (ii) the solution to the corresponding mathematical problem, namely, the processing of the information that is contained in the given problem by means of mathematics and mathematical informatics.

Regarding the logical-mathematical modeling of problems that belong to the realm of the social sciences, in particular, the value-system of the society in which behavior is studied must somehow find its place in the framework of action employed in the relevant analysis (see: Parsons, *The Structure of Social Action*). The French philosopher Louis Lavelle (one of the greatest French metaphysicians of the twentieth century) has argued that every value is an object of a desire and of a judgment. Thus, in the philosophy of the social sciences, by the term "values," we mean needs that arise in consciousness and must be addressed by consciousness. For instance, needs to know, to reap, to sustain, to socialize, to individuate, to control, to act, and so on. Consciousness selects some concrete values-needs which it projects onto the world, thus transforming them into historical objects, and, finally, the values that have been historically objectified (specifically, have become social and institutional events) influence consciousness, shaping the subject's existential conditions.

The transition from the natural sciences to the social sciences is an upward-moving process known as "emergentism." The specific nature of the social sciences clearly emerges when we examine them from an ontological point of view. In his book *Le Regole dell'Azione Sociale*

(Milano: Il Saggiatore, 1983), Giuliano Di Bernardo (specifically, in the chapter entitled "La fondazione del sociale") shows that social reality is constructed by humanity through "constitutive rules." In particular, Di Bernardo (ibid) maintains that, based on constitutive rules, language, and the collective self, social reality has a dual ontology: one that is "visible," observable, made up of objects from the external world, such as houses, monuments, and money; the other is "invisible," made up, respectively, of housing *regulations*, the *aesthetics* of monuments, and the *significance* of money.

As Giuliano Di Bernardo maintains in his book *The Epistemological Foundation of Sociology* (p. 48), "values and norms are closely related to action,", and, indeed, "values, norms, and action are a unit, which can only be broken down analytically, in order to gain a better understanding of the different parts of which it consists." Furthermore, Giuliano Di Bernardo (*ibid*, p. 50), referring to the "construction of social reality through constitutive rules," succinctly maintains that "the social being (fact) is constituted not only by the visible ontology (of physics and biology) but also by the invisible ontology (of the normative)."

The consciousness of existence that not only functions as a witness or an observer, but also functions as a judge is what we call "moral consciousness." The logic of moral consciousness, that is, the logic of ethics, is called "deontic logic" (the word "deontic" derives from the Greek word "deon," which means "what is binding" or "proper"). Ethics is concerned with what good as a concept is and with what we should and should not do. Deontic logic is concerned with the manner in which we can represent those things that we should and should not do logically. Some formal analogies between deontic notions and "pure" (*alethic*) modalities ("necessity," "possibility," and "impossibility") were studied during the Middle Ages (especially in the context of fourteenth-century Aristotelianism) in terms of the following equivalences (where the sentential operator O denotes the concept of obligation, the sentential operator P denotes the concept of permission, the sentential operator F denotes the concept of prohibition, and p is the corresponding sentence):

i. $P(p) \leftrightarrow \neg O(\neg p)$,
ii. $O(p) \leftrightarrow \neg P(\neg p)$,
iii. $O(p) \leftrightarrow F(\neg p)$,
iv. $F(p) \leftrightarrow O(\neg p)$.

In his book *Elementa juris naturalis*, Gottfried Wilhelm Leibniz developed the first modern system of deontic logic based on modal logic, and he called the deontic categories of the obligatory (*debitum*), the permitted (*licitum*), and the prohibited (*illicitum*) "modalities of law" (*juris* *modalia*). The system of "standard deontic logic" developed by the Finnish philosopher Georg Henrik von Wright (1916–2003) can be axiomatized by adding the following axioms to the standard axiomatization of classical propositional calculus (modal logic):

i. $(\vDash p) \rightarrow (\vDash O(p))$, meaning that, "if p is a tautology, then it ought to be that p" (i.e., contradictions are not permitted);

ii. $O(p \rightarrow q) \rightarrow (O(p) \rightarrow O(q))$, meaning that, "if it ought to be that p implies q, then, if it ought to be that p, it ought to be that q."

iii. $O(p) \rightarrow P(p)$, meaning that, "if it ought to be that p, then it is permitted that p" (equivalently, "if it is not permitted that p, then it is not obligatory that p").

The sentential operator $F(p)$, meaning "it is forbidden that p," can be formally defined as $O(\neg p)$ or as $\neg P(p)$.

In general, deontic logic builds a bridge between logical rigor and ethics. For a systematic study of deontic logic, one should read the following books by Giuliano Di Bernardo: *Introduzione alla logica dei sistemi normativi*, Bologna: Il Mulino, 1972; *Le regole dell'azione sociale*, Milano: Il Saggiatore, 1983; and *Normative Structures of the Social World*, Amsterdam: Rodopi, 1988.

Proof: A Theme in Need of a Focus

In its broadest sense, science is a system of behavior by means of which humans become masters of their environment. For this reason, no human society can exist without science. In a narrower sense, science is not so much a system of behavior as a system of knowledge which, specifically, aims to conceptualize, describe, and interpret the phenomena of the macrocosm and the microcosm according to a clearly determined and robust method, as well as to create the necessary intellectual tools for understanding the aforementioned phenomena, logic, and mathematics. Science builds knowledge through logic and testable explanations and predictions. Thus, science contrasts prejudice, superstition, personal opinion, subjective political beliefs, and, generally, irrational passions.

In mathematics, a "proof" is a verification of a proposition by a chain of logical deductions from a set of axioms. As already explained above, by a "proposition," we mean a statement that is either true or false; by a "predicate," we mean a proposition whose truth value depends on the value of a variable; and, by an "axiom," we mean a proposition that is assumed to be true (because we think that it is reasonable, that is, worthy enough to be declared true). We can choose any propositions as axioms, provided, however, that the axiomatic system that we create is consistent and complete: a set of axioms is “consistent” if no proposition (in the given axiomatic system) can be proved to be *both* true and false, and a set of axioms is “complete” if it can be used to prove that *every* proposition (in the given axiomatic system) is either true or false (and, hence, in such an axiomatic system, every problem becomes solvable). However, as I have already mentioned, Kurt Gödel, in the 1930s, proved that there is no such axiomatic system (if it is to contain arithmetic), and, therefore, we must make compromises as to the range of validity and the explanatory power of each axiomatic system, but we must always follow the rules of logic in order to avoid contradictions.

The first systematization of logic is due to the ancient Greek philosopher and scientist Aristotle, and, for this reason, the phrase “Aristotelian logic” is still commonly used. Aristotle’s works on logic were grouped together by ancient commentators under the title *Organon* (“Instrument”). In particular, the *Organon* comprises the following logical treatises written by Aristotle: (i) *Categories*, (ii) *On Interpretation*, (iii) *Prior Analytics*, (iv) *Posterior Analytics*, (v) *Topics*, and (vi) *On Sophistical Refutations*. The title *Organon*, meaning instrument, implies that logic is an instrument and a method used by philosophy and science, and, in particular, according to both Aristotle and the later Peripatetics, the ultimate purpose of correct reasoning is to create correct social relationships and to enable people to correctly communicate the results of philosophical and scientific research with each other. In the third century B.C.E., the Greek Stoic philosopher and logician Chrysippus founded a propositional calculus, studying implication, conjunction, and disjunction, and, in the early twentieth century, the Austrian philosopher Ludwig Wittgenstein brilliantly studied the problems of communication in a comprehensive and systematic way.

Logic underpins “mathematical proof.” As Steve Halperin (*Introduction to Proof in Analysis*, p. 9) has pointedly argued, by the term “mathematical proof,” one should understand “a sequence of statements which establish that certain assumptions (the hypotheses) imply that a certain statement (the conclusion) is true,” and the statements that constitute a proof must satisfy the following requirements: (i) “each is clear and unambiguous”; (ii) “each is true, and its truth follows immediately from the truth of the preceding statements and the hypotheses”; and (iii) “the final statement is the conclusion.” Thus, in the context of a mathematical proof, we may use several techniques, such as direct proof (involving arguing step by step, starting from what we know until we have demonstrated the truth of some conclusion), mathematical induction, counterexamples (since a single counterexample suffices to prove that a statement claiming necessity and universality is wrong), *reductio ad absurdum* (i.e., the form of argument that attempts to prove a statement by proving that the negation of the given statement leads to absurdity or contradiction), proof by contraposition (i.e., inferring a conditional statement from its contrapositive; the contrapositive of the statement "if A, then B" being the statement "if not B, then not A"), etc. In addition, the concept of a mathematical proof is inextricably linked to the concept of a definition, that is, to a deep and rigorous understanding of the essence of the object under consideration.

In ancient Greece, Socrates, Plato, and Aristotle developed a way of thinking that is based on "universal definitions." The classical Greek way of thinking consists of understanding the whole as a whole and of thinking upon thinking itself, thus leading contemplation to a level that supersedes mere practical thinking and spontaneity. This way of thinking opens the mind to the world of philosophy, scientific rigor, and genuine strategy.

Chapter 2
The Structure of Number Sets, Arithmetic, and Algebra

The attempts of nineteenth-century mathematicians to found mathematical analysis in a rigorous way were based on real numbers, which also needed a rigorous foundation. Numbers are abstract objects, concepts. Simultaneously, they are intimately related to the world, since we organize the world with them (that is, we count, we measure, and we form scientific theories with numbers). In order to understand the concept of a number, we have to keep in mind that what we count are not "things," but "sets of things."

The history of set theory and of non-numerical mathematics, in general, can be traced back to the era of classical Greece, but the first systematic inquiry into the foundations of set theory was due to the German mathematician Georg Ferdinand Ludwig Philipp Cantor (1845–1918). However, before Cantor, George Peacock (1791–1858), Augustus De Morgan (1806–71), and George Boole (1815–64) had already made significant contributions to the formalization of non-numerical mathematics.

According to Cantor, by the term "set," we should understand a well-defined gathering together into a whole of definite, distinguishable objects of perception or of our thought that are called elements of the set. By the term "well-defined," Cantor means that, given any object and any set, the given object is either an element of the given set or not an element of the given set. By the terms "definite" and "distinguishable," Cantor means that no two elements of a set are the same.

The empty set is denoted by $\emptyset$. The empty set has no elements. If a set has only one element, then it is called a "singleton."

If every element of a set B is an element of a set A, then B is said to be a "subset" of A, and we write $B \subseteq A$. Every set is a subset of itself. If A is an arbitrary set, then $\emptyset \subseteq A$; that is, the empty set is a subset of every set. Two sets A and B are "equal" if and only if $A \subseteq B$ and $B \subseteq A$, and then we write $A = B$. If two sets A and B satisfy the condition $B \subseteq A$ and there is at least one element of A that is not an element of B, then B is said to be a "proper subset" of A, and we write $B \subset A$. If $B \subseteq A$ or $B \subset A$, then A is said to be a "superset" of B. When in a particular situation all the sets under consideration are subsets of a fixed set, this fixed set, which is the superset of every set under consideration, is called the "universal set," or the "universe of discourse."

If the elements of a set are sets themselves, then the set is called a "set of sets," "family of sets," "collection of sets," or "class of sets." For instance, $\mathcal{C} = \{\{x\}, \{y, z\}\}$ is a class of sets (notice that x is something different from $\{x\}$).

If A and B are two arbitrary sets, then we define their

i. "union":
$A \cup B = \{every\ x\ such\ that\ x\ belongs\ to\ at\ least\ one\ of\ A\ and\ B\}$;
and

ii. "intersection":
$A \cap B = \{every\ x\ such\ that\ x\ belongs\ to\ both\ A\ and\ B\}$.

Two sets are called "(relatively) disjoint" if their intersection is the empty set.

However, in 1901, the British philosopher and mathematician Bertrand Russell proved that every set theory that contains an unrestricted comprehension principle leads to contradictions. In other words, the "universal set" is not a set. The aforementioned contradictory situation is known as Russell's paradox.

Russell's Paradox: Let U be the collection of all sets:
$U = \{x | x\ is\ a\ set\}$.
Then U is not a set.

We can prove Russell's paradox by *reductio ad absurdum*. Assume, for the sake of contradiction, that U is a set. However, any ordinary mathematical set (e.g., of numbers, functions, etc.) is not a member of itself and can be naturally regarded as a member of a smaller universe of sets. In particular, let V be an arbitrary set and $V \notin V$. Then, by the definition of U, $V \in U$. Moreover, because U is a set, either $U \in U$ or $U \notin U$. If $U \notin U$, then, because $V \in U$, it follows that $U \in U$. But, if $U \in U$, then, again because $V \in U$, where $V \notin V$, it follows that $U \notin U$. Therefore, in both of these cases, we reach a contradiction, and, in this way, we prove that U is not a set (U is called "Russell's class").■

The German mathematician, logician, and philosopher Friedrich Ludwig Gottlob Frege (1848–1925) believed that the foundational problems of mathematics could be solved and overcome by reformulating Aristotelian logic in a "Platonic" way, in order, in this way, to equip mathematics with epistemologically and ontologically robust underpinnings. Thus, Frege argued that mathematical truths are reducible to logical truths, and that logic is equivalent to Plato's world of ideas. In order to understand the manner in which Frege attempted to reduce all mathematics to logic and to a rigorous conception of set theory, we have to understand the manner in which he defined the concept of a number as a set of mutually equivalent sets. In particular, Frege has explained that any number n can be used in order to count any n-membered set. For instance, the number two can be thought of as the set of all 2-membered sets, or as the set of all pairs, independently of the nature of the objects that constitute each pair. Similarly, the number three can be thought of as the set of all triplets, the number four can be thought of as the set of all quadruples, and so on.

In particular, in order to define the concept of a natural number $(0,1,2,3,\ldots,n,n+1,\ldots)$, Frege defined, for every 2-place relation R, the concept "x is an ancestor of y in the R-series," and this new relation is known as the "ancestor relation on R." The underlying idea can be easily grasped if we interpret Frege's 2-place relation R as "x is the father of y in the R-series." For instance, if a is the father of b, b is the father of c, and c is the father of d, then Frege's definition of "x is an ancestor of y in the fatherhood-series" ensures that a is an ancestor of b, c, and d, that b is an ancestor of c and d, and that c is an ancestor of d. More generally, given a series of facts of the form aRb, bRc, and cRd, Frege showed that we can define a relation R^* as "y follows x in the R-series." Thus, Frege formulated a rigorous definition of "precedes," and he concluded that a "natural number" is any number of the predecessor-series beginning with 0.

Using the concept of a "predecessor," the American mathematician John von Neumann (1903–57) has proposed an even more accurate definition of a "natural number." According to von Neumann, instead of defining a natural number n as the set of all n-membered sets, a natural number n should be defined as a particular n-membered set—more specifically, as the set of its predecessors. For instance, the number two having two predecessors, zero and one, we can think of the number two as the set $\{0,1\}$, where zero has no predecessor. Therefore, zero can be thought of as the empty set, denoted by $\emptyset$. The number one has only one predecessor, zero. Therefore, we can think of the number one as $\{\emptyset\}$, namely, as the singleton of the empty set. Thus, von Neumann formulated the modern definition of "ordinal numbers." In particular, given the "successor operation," which is defined as

$succesor(n) = n \cup \{n\}$,

the set of von Neumann natural numbers, namely, the ordinal numbers, denoted by ω, is defined as follows:

i. $\emptyset \in \omega$.
ii. If $n \in \omega$, then $succesor(n) \in \omega$.
iii. Nothing belongs to ω unless it can be constructed using the preceding rules.

Thus, we obtain the following definitions:

$0 = \emptyset.$
$1 = successor(0) = \emptyset \cup \{\emptyset\} = \{\emptyset\} = \{0\}.$
$2 = successor(1) = \{\emptyset\} \cup \{\{\emptyset\}\} = \{\emptyset, \{\emptyset\}\} = \{0,1\}.$
$3 = successor(2) = \{\emptyset, \{\emptyset\}\} \cup \{\{\emptyset, \{\emptyset\}\}\} = \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\} = \{0,1,2\}.$
$\vdots$

The set-theoretical approach to modern mathematics is necessarily based on the acceptance of the following axioms (see: Halmos, *Naive Set Theory*):
Axiom 1 ("Axiom of Extensionality"): For every set A and for every set B, $A = B$ if and only if, for every element x, it holds that $x \in A$ if and only if $x \in B$.
Axiom 2 ("Axiom of Foundation"): Infinite descending membership chains of sets ($X_1 \ni X_2 \ni X_3 \ni \cdots$) are forbidden; that is, we cannot have a set X_1 that has an element X_2 that has an element X_3, and so on, forever. Any descending membership chain of sets, where each term of the chain belongs to a previous term of the given chain, must be finite (and the last element of any such chain must be the empty set).
Axiom 3 ("Axiom of Specification"): If $\varphi(x)$ is a formula, whose truth value ("True" or "False") depends on x, then, for every set A, there exists a set B such that, for every element x, it holds that $x \in B$ if and only if $x \in A$ and $\varphi(x)$ is true. In other words, given a set A, the "container" B whose elements are the elements of A that satisfy $\varphi(x)$ is a set (this axiom, by forbidding unrestricted comprehension, shields modern mathematics against Russell's paradox).
Axiom 4 ("Axiom of Pairing"): For every set A and for every set B, there exists a set C such that, for every x, it holds that $x \in C$ if and only if $x \in A$ or $x \in B$ (meaning that any two things in mathematics can also be paired up). In other words, for any sets A and B, there exists a set $\{A, B\}$ that contains exactly A and B.
Axiom 5 ("Axiom of Union"): Given any set X whose elements are sets, we can create the set whose elements are all the members of all the sets that belong to X.
Axiom 6 ("Axiom of Replacement"): We can take a set X and form another set by replacing the elements of X with other sets according to any definite rule.
Moreover, later in this chapter, I shall refer to the "Axiom of Choice," which allows us to perform a lot of operations on infinite sets that mirror things that are intuitively obvious on finite sets.

Let X be a set of elements $a, b, \ldots$ Suppose that there is a binary relation expressed by $a \prec b$, defined between certain pairs (a, b) of elements of X, and satisfying the following properties:
$a \prec a$;
if $a \prec b$ and $b \prec a$, then $a = b$;
if $a \prec b$ and $b \prec c$, then $a \prec c$ (transitivity).
Then X is said to be "partially ordered" (or "semi-ordered") by the relation $\prec$.
Let X be a partially ordered set with elements $a, b, \ldots$ If $a \prec c$ and $b \prec c$, then c is said to be an "upper bound" for a and b. If, furthermore, $c \prec d$ whenever d is an upper bound for a and b, we call c the "least upper bound," or the "supremum," of a and b, and we write $sup(a, b)$. This element of X is unique if it exists. In a similar way, we define the "greatest lower bound," or the "infimum," of a and b, and we denote it by $inf(a, b)$.
A partially ordered set X is said to be "linearly ordered" (or "totally ordered") if, for every pair (a, b) in X, either $a \prec b$ or $b \prec a$ holds. A subset of a partially ordered set X is itself partially ordered by the relation that partially orders X; and the subset may even be linearly ordered by this relation. If X is a partially ordered set and A is a subset of X, then an element $m \in X$ is said to be an upper bound of A if $a \prec m$ for every $a \in A$. An element $m \in X$ is said to be "maximal" if the relations $m \in X$ and $m \prec x$ imply that $m = x$ (the maximum is the largest number of the set, while the supremum is the smallest upper bound of the set). In a similar way, we define a "minimal element."

The Number Sets ℕ, ℤ, ℚ, ℚ˜, and ℝ

In this section, we shall study the structure of the number sets ℕ, ℤ, ℚ, ℚ˜, and ℝ.

The Natural Numbers

ℕ: the "natural numbers," namely, the positive integers 1,2,3, ..., which are used to count objects, and 0. For any natural numbers m, n, and k, the following equalities hold true:

i. $m + n = n + m$,
ii. $m + (n + k) = (m + n) + k$,
iii. $mn = nm$,
iv. $m(nk) = (mn)k$,

v. $m(n+k) = mn + mk$,

vi. $m \cdot 1 = m$.

Equalities (i) and (iii) express the "commutative law" of addition and multiplication respectively; equalities (ii) and (iv) express the "associative law" of addition and multiplication respectively; and equality (v) is known as the "distributive law" of multiplication over addition. The aforementioned laws underlie all computations. If a natural number m is divisible by a natural number n, then m is said to be a "multiple" of the number n, and n, in turn, is said to be the "divisor" of the number m. If m is a multiple of the number n, then there is a natural number k such that $m = kn$. For instance, 18 is divisible by 3, and we write $18 = 6 \cdot 3$. In this case, $m = 18$ (the "dividend"), $n = 3$ (the "divisor"), and $k = 6$ (the "quotient"). If a natural number m is not exactly divisible by a natural number n, that is, if there is no natural number k such that $kn = m$, then we consider "division with a remainder." For instance, 33 divided by 2 equals 16 ("partial quotient") with a remainder of 1, and therefore $33 = 16 \cdot 2 + 1$.

For any two natural numbers a and b, there exists a unique natural number n such that $a \cdot n = b$ if and only if a is a divisor of b, and then we write $n = b \div a \equiv \frac{b}{a}$. "Even numbers" are divisible by 2 without remainders, whereas "odd numbers" are not evenly divisible by 2. Notice that odd numbers end in the digit 1, 3, 5, 7, or 9, whereas all the numbers ending with 0, 2, 4, 6, or 8 are even numbers.

The greatest common divisor (denoted by gcd) of two natural numbers a and b is the largest natural number that divides both a and b, and the Euclidean Algorithm for computing $gcd(a,b)$ is as follows:

i. If $a = 0$, then $gcd(a,b) = b$.

ii. If $b = 0$, then $gcd(a,b) = a$.

iii. If a and b are both non-zero natural numbers ($a > b$), then we write a in quotient remainder form, namely, $a = b \cdot q + r$, and, subsequently, we compute $gcd(b,r)$ using the Euclidean Algorithm since $gcd(a,b) = gcd(b,r)$. For instance, if $a = 280$ and $b = 120$, then we can compute $gcd(a,b)$ as follows: firstly, we use long division to find that $\frac{280}{120} = 2\, with\, a\, remainderof\ 40$, which can be written as $280 = 120 \times 2 + 40$; secondly, we compute $gcd(120{,}40) = 40\, with\, a\, remainderof\ 0$; and, therefore, $gcd(280{,}120) = 40$.

Let a and b be both non-zero natural numbers. Moreover, let $lcm(a,b)$ denote the least common multiple of a and b (i.e., $lcm(a,b)$ is the smallest natural number that is evenly divisible by both a and b). Then

$gcd(a,b) = \frac{a \cdot b}{lcm(a,b)} \Leftrightarrow lcm(a,b) = \frac{a \cdot b}{gcd(a,b)}$.

If a natural number has only two divisors, a unity (one) and the number itself, then it is called a "prime number"; if it has more than two divisors, then it is called a "composite number." For instance, 2, 3, 5, and 7 are prime numbers, but 9 is not a prime number (9 is a composite number, because the divisors of 9 are 1, 3, and 9). Notice that 2 is the only even prime number, and that, except for 2 and 5, all prime numbers end in the digit 1, 3, 7, or 9. All numbers have prime factors. For instance, the prime factors of 10 are 2 and 5, since $10 = 2^1 \times 5^1$; the prime factors of 11 are 1 and 11, since $11 = 1^1 \times 11^1$; the prime factors of 100 are 2 and 5, since $100 = 2^2 \times 5^2$, etc.

The Italian mathematician and glottologist Giuseppe Peano (1858–1932) has organized the natural numbers as an axiomatic system on the basis of the following axioms, known as the "Peano axioms":

i. 0 is a natural number, symbolically: $0 \in \mathbb{N}$.
ii. If n is a natural number, then the successor of n (i.e., $successor(n) = n + 1$) is also a natural number.
iii. If two natural numbers have the same successor, then the two natural numbers are identical.
iv. 0 is not the successor of any natural number.
v. "Induction Axiom": If X is a set containing both 0 and the successor of every natural number belonging to X, then every natural number belongs to X.

The "Induction Axiom" gives rise to and underpins the principle of "Mathematical Induction," which is a mathematical proof technique for propositions: Suppose that P is a proposition defined on the natural numbers $\mathbb{N}$, such that:

i. $P(1)$ is true, that is, P holds true for 1;
ii. $P(n+1)$ is true whenever $P(n)$ is true.

Then P is true for every natural number. In this case, P is the "inductive hypothesis." By completing the aforementioned two steps of mathematical induction, we prove that P is true for every natural number.

Example: Let P be the proposition that the sum of the first n natural numbers is
$\frac{1}{2}n(n+1)$, namely: $P(n) = 1 + 2 + 3 + \cdots + n = \frac{1}{2}n(n+1)$. We can prove that P is true for every natural number $n \in \mathbb{N}$ using mathematical induction as follows:

Basis step: The proposition holds for $n = 1$, because $1 = \frac{1}{2}(1)(1+1)$. Hence, $P(1)$ is true.

Induction step: We assume that $P(n)$ is true, and we add $n+1$ to both sides of $P(n)$, obtaining

$$1+2+3+\cdots+n+(n+1)=\frac{1}{2}n(n+1)+(n+1)=$$
$$\frac{1}{2}[n(n+1)+2(n+1)]=\frac{1}{2}[(n+1)(n+2)],$$

which is $P(n+1)$. Hence, $P(n+1)$ is true whenever $P(n)$ is true. By the principle of mathematical induction, P is true for every natural number $n \in \mathbb{N}$.

The Fundamental Theorem of Arithmetic: Every natural number greater than 1 can be uniquely represented as a product of prime numbers, up to the order of the factors. This theorem can be derived from Book VII, propositions 30, 31, and 32, and Book IX, proposition 14, of Euclid's *Elements*. In other words, this theorem states that every natural number $n > 1$ can be represented in exactly one way as a product of prime numbers:

$$n = p_1^{n_1} p_2^{n_2} \dots p_k^{n_k} = \prod_{i=1}^{k} p_i^{n_i}$$

where $p_1 < p_2 < \cdots < p_k$ are prime numbers, and n_i are natural numbers greater than zero. This representation is commonly known as the "canonical representation of a natural number" (it can be extended to include 1 by the convention that the "empty product" is equal to 1; the "empty product" corresponds to $k = 0$).

Proof: The existence of prime factorization can be proved using mathematical induction: In the basis step, we see that the statement is true for $n = 2$, since 2 is a prime number. Suppose that the statement is true for all n such that $1 < n < k$, so that we can write every n (where $1 < n < k$) as a product of primes. We can prove that the statement is true for $n = k$ as follows: If k is prime, then the case is obvious. If k is not prime, then it is a composite number, and we can factor it as follows:

$k = x \times y$, where $1 < x, y < k$.

Hence, by induction, we can argue that x and y can be written as the product of primes, meaning that k can also be written as the product of primes.

The uniqueness of prime factorization can be proved by *reductio ad absurdum*, using Euclid's Lemma, which states that, if a prime p divides the product ab of two natural numbers a and b, then p must divide at least one of those natural numbers a or b (Euclid's Lemma can be proved using mathematical induction). For the sake of contradiction, suppose that there exist two distinct prime factorizations for the same natural number n ($>$ 1), so that:

$$n = p_1^{n_1} p_2^{n_2} \dots p_k^{n_k} = q_1^{m_1} q_2^{m_2} \dots q_j^{m_j}. \qquad (1)$$

Moreover, suppose that n (> 1) is the least natural number that has two distict prime factorizations. Notice that

$$n = p_1^{n_1} p_2^{n_2} \dots p_k^{n_k} = q_1^{m_1} q_2^{m_2} \dots q_j^{m_j}$$

implies that p_1 divides $q_1^{m_1} q_2^{m_2} \dots q_j^{m_j}$ (p_1 divides the left side, so it divides the right side); and, therefore, by Euclid's Lemma, p_1 divides some q_i. Without loss of generality, let that q_i be q_1. Because both p_1 and q_1 are primes, the fact that p_1 divides q_1 implies that $p_1 = q_1$. Since $p_1 = q_1$, we can delete p_1 and q_1 in equation (1), so that now (1) gives us two distinct factorizations of some natural number strictly smaller than n, contradicting the assumed minimality of n; *quod erat demonstrandum*

A set is said to be "countable" (or "denumerable") if you can make a list of its members, and, by a "list," we mean that you can find a first member, a second member, a third member, and so on, and, thus, assign to each member a natural number of its own, perhaps going on forever. Obviously, the natural numbers are countable (you can assign each natural number to itself).

The Integral Numbers

$\mathbb{Z}$: the "integral numbers," or the negative integers, zero, and the positive integers:

$$\dots - 3, -2, -1, 0, 1, 2, 3, \dots$$

The notation $\mathbb{Z}$ for the set of integers derives from the German word "Zahlen," which means "numbers."

From the perspective of ancient mathematicians, numbers are things by means of which we count, but modern mathematical analysis, founded on Cartesianism, understands numbers mainly as positions on the number line. Let us draw a straight line l and mark on it a point 0 that will be taken as the origin. Then we choose a unit segment $0P$, where P is a natural number that lies to the right of 0, and, in this way, we specify the positive direction. In other words, the unit segment $0P$ determines the direction of the number line and corresponds to the positive unity $+1$ (or simply 1). Let us, for instance, take the number 4. Laying off the unit segment from the point 0 in the given direction four times, we obtain the point Q that corresponds to the number 4. Let us now lay off four unit segments from the zero point in the direction opposite to the specified. We then get the point Q', which is symmetric to the point Q about the origin 0. The point

Q' corresponds to the number -4. Thus, the numbers 4 and -4 are said to be "opposite." By analogy, we can define any other integer (positive or negative). In general, the numbers situated on the number line l in the specified direction are said to be "positive," whereas the numbers located on the number line in the direction opposite to the given one are said to be "negative." Hence, the natural numbers and their opposites (the opposite of the number zero being the same number) form together the set of integral numbers (integers), which is denoted by $\mathbb{Z}$.

If a point X of the line l corresponds to some number r, then this number is said to be the "coordinate of the point X," and, in this case, we write $X(r)$.

The "absolute value" of the number r is denoted by $|r|$. The absolute value of any positive number is the number itself. The absolute value of any negative number is equal to its opposite number. The absolute value of the number zero is zero.

The sum of two negative numbers is a negative number. In order to find the absolute value of a sum, it is necessary to add together the absolute values of the addends. The sum of two numbers having unlike signs is a number that has the same sign as the addend with greatest absolute value; and, in order to find the absolute value of their sum, it is necessary to subtract the smaller value from the larger one (so that, for instance, $5 + (-3) = 5 - 3 = 2$).

In order to subtract one number from another, it is necessary to add to the minuend a number that is the opposite of the subtrahend.

The product (resp. quotient) of two negative numbers is a positive number. The product (resp. quotient) of two numbers having unlike signs is a negative number. In order to find the absolute value of a product (resp. quotient), it is necessary to multiply (resp. divide) the absolute values of these numbers.

When we divide two integers, we get an equation of the following form:

$\frac{a}{b} = q \, with \, remainder \, r,$

where a is the dividend, b is the divisor, q is the quotient, and r is the remainder. Sometimes, we are only interested in the remainder in the division of a by b, and, for these cases, there is an operator called the "modulus operator" (abbreviated as mod). If the difference of two integers a and b is divisible by n, then a and b are said to be congruent with respect to the modulus n, and this is symbolically expressed as follows:

$$a \equiv b(modn)$$

(and then we read "a is congruent to b modulo n"), and each of the numbers a and b is said to be a residue $(modn)$ of the other (the "modulus" is the remainder of the division of one number by another; for instance, 9 divided by 4 equals 2 and there is a remainder of 1, so that we write $9 \equiv 4 = 1$, whereas, for instance, $k \equiv k = 0$ for all $k \in \mathbb{N}$). Our intuition for the integers $modn$ should be a circle with the integers 0 through $n - 1$ arranged on it. Notice that:

$$a \equiv b(modn) \Leftrightarrow n|b - a$$

where $n|b - a$ means that n divides $b - a$, in which case a and b have the same remainder when we divide them by n (this notation and much of the elementary theory of congruences are due to the German mathematician Carl Friedrich Gauss). For instance, $5 \equiv 2(mod3)$ because $3|5 - 2$; and $4 \equiv -1(mod5)$ because $5|4 - (-1)$.

Fundamental principles of the theory of congruences:

I. If $a \equiv b(modn)$ and $a \equiv c(modn)$, then $b \equiv c(modn)$.

II. If $a \equiv a', b \equiv b', c \equiv c', etc.(modn)$, then $a \pm b \pm c \pm \cdots \equiv a' \pm b' \pm c' \pm \cdots (modn)$.

III. If $a \equiv a'(modn)$, then $ka \equiv ka'(modn)$.

IV. If $a \equiv a'(modn)$ and $b \equiv b'(modn)$, then $ab \equiv a'b'(modn)$.

V. If $a \equiv a', b \equiv b', c \equiv c', etc.(modn)$, then $abc \ldots \equiv a'b'c' \ldots (modn)$.

VI. If $ka \equiv kb(modn)$, then $a \equiv b\left(mod\frac{n}{d}\right)$ where d is the greatest common divisor of k and n.

By a “linear congruence,” we mean a congruence of the form

$ax \equiv b(modn)$, where $a, b, n \in \mathbb{Z}$ and $n > 0$. (1)

A solution to (1) is an integer x that satisfies (1) and is a least residue $(modn)$ (that is, $0 \leq x \leq n - 1$). The congruence relation (1) has a solution if the (“unknown”) integers x (where $0 \leq x \leq n - 1$) and k satisfy the equation

$ax = b + kn$ (2)

(keep in mind that, in congruences, we work only with integers). Moreover, notice that

$ax \equiv b(modn) \Leftrightarrow ax - ny = b$,

where $ax - ny = b$ is the corresponding linear Diophantine equation (by a “Diophantine equation,” we mean an equation involving only sums, products, and powers in which all the constants are integers and the only solutions of interest are integers; they are named in honor of the third-century C.E. Greek mathematician Diophantus, who developed a systematic theory of such equations). The linear congruence (1) has a solution precisely when $gcd(a, n)|b$, that is, precisely when b is a multiple of $d = gcd(a, n)$, and, in this case, (1) is equivalent to

$\frac{a}{d}x \equiv \frac{b}{d}\left(mod\frac{n}{d}\right)$.

We shall prove this theorem and its ramifications shortly.

If $gcd(a,n) = d|b$, then (1) has a solution x_0, and, given a solution x_0, we can construct infinitely many solutions to (1) of the form $x = x_0 + \lambda\frac{n}{d}$, where λ is any integer. If x_0 is one solution to (1), then all the solutions to (1) are described by

$$x \equiv x_0 \left(mod\frac{n}{d}\right)$$

where $d = gcd(a,n)$.

Example 1: Consider the linear congruence

$4x \equiv 8(mod5)$.

In this case (where $a = 4$ and $n = 5$), we are allowed to divide both sides by 4, because $gcd(4,5) = 1$. Thus,

$4x \equiv 8(mod5) \Rightarrow \frac{4}{4}x \equiv \frac{8}{4}(mod5) \Rightarrow x \equiv 2(mod5)$,

so that, according to the above definition of a linear congruence (specifically, according to relation (2)), $x = 2 + 5k$.

Example 2: Consider the linear congruence

$4x \equiv 2(mod5)$.

In this case, we cannot divide both sides by 4, because $\frac{2}{4}$ is not an integral number, and, therefore, it is not allowed in linear congruences. In this case, we can work as follows: Since $4x$ can be treated as $2(2x)$, and $gcd(2,5) = 1$, we can divide both sides by 2 to obtain:

$2x \equiv 1(mod5)$,

so that, according to the above definition of a linear congruence (specifically, according to relation (2)), $2x = 1 + 5n \Leftrightarrow x = \frac{1+5n}{2}$; and we have: if $n \geq 0$, then possible values of x are $\frac{1}{2}, 3, \frac{11}{2}, 8, \frac{21}{2}, 13, \ldots$, whereas, if $n < 0$, then possible values of x are $-2, -\frac{9}{2}, -7, -\frac{19}{2}, \ldots$ If we are looking for integral values of x, then possible solutions include $(3,8,13,\ldots)$ and $(-2,-7,-12,\ldots)$.

The linear congruence $ax \equiv b(modn)$ has a solution (for x) if and only if $gcd(a,n)$ divides b, in which case the congruence has $d = gcd(a,n)$ *incongruent* solutions.

Proof: Let x_0 be a solution to the given linear congruence, so that

$ax_0 \equiv b(modn)$. (1)

Due to (1), n divides $ax_0 - b$; symbolically, $n|ax_0 - b$. (2)

By the definition of divisibility, (2) implies that

$ax_0 - b = ny$ for some integral number y. Therefore,

$ax_0 - ny = b$. (3)

Equation (3) implies that b can be written as a linear combination of a and n, and, for this reason, b is a multiple of $gcd(a,n)$, meaning that

$gcd(a,n)$ divides b, as required. Now, we shall prove the reverse as follows:

Suppose that $gcd(a,n)|b$, and let $d = gcd(a,n)$. Hence, $b = dk$ for some integral number k. Let us consider two integral numbers x_0 and y_0 such that

$$ax_0 - ny_0 = d. \tag{4}$$

If we multiply both sides of equation (4) by k, then we obtain:

$$a(kx_0) - n(ky_0) = dk = b. \tag{5}$$

If we set $kx_0 = x_1$ and $ky_0 = y_1$, then (5) becomes

$$ax_1 = b + ny_1 \Rightarrow ax_1 - b = ny_1 \Rightarrow n|ax_1 - b.$$

Because n divides $ax_1 - b$, it holds that $ax_1 \equiv b(modn)$, where x_1 is the solution to the given congruence; *quod erat demonstrandum.*

Finally, we have to prove that, given the solution x_0, the linear congruence has $d = gcd(a,n)$ incongruent solutions. In other words, we know that $ax_0 \equiv b(modn)$. Let

$$x_m = x_0 + m\left(\frac{n}{d}\right), \tag{6}$$

where $0 \leq m \leq d-1$, and, as before, $d = gcd(a,n)$. These are the required $d = gcd(a,n)$ incongruent solutions to the given congruence. To show that they are indeed solutions to the given linear congruence, we work as follows: We multiply (6) by a to obtain:

$$ax_m = ax_0 + m\frac{a}{d}n.$$

Notice that d is a divisor of a, and, thus, $\frac{a}{d}$ is an integral number, so that, given that $d = gcd(a,n)$, we obtain:

$$ax_m = ax_0 + m\frac{a}{d}n \equiv b(modn),$$

meaning that all of these x_m's are indeed solutions to our original congruence, $ax \equiv b(modn)$. We can show that these solutions are incongruent as follows: Suppose that

$$x_i \equiv x_j(modn). \tag{7}$$

We shall show that this implies that $i = j$. Equations (6) and (7) imply that

$$x_0 + i\left(\frac{n}{d}\right) \equiv x_0 + j\left(\frac{n}{d}\right)(modn) \Rightarrow i\left(\frac{n}{d}\right) \equiv j\left(\frac{n}{d}\right)(modn) \Rightarrow (i-j)\left(\frac{n}{d}\right) = np$$

(i.e., a multiple of n), meaning that $\frac{i-j}{d}$ is an integral number. Given that $0 \leq m \leq d-1$ and $\frac{i-j}{d}$ is an integral number, it holds that $i - j = 0 \Leftrightarrow i = j$, meaning that the only case in which the solutions x_i and x_j are congruent modulo n is when they are the same (i.e., when $i = j$); ; *quod erat demonstrandum.*

Remark: The above theorem provides a criterion by which we can decide whether a linear congruence has solutions and how many solutions it has, and, in fact, the structure of the proof of this theorem shows us the way in which we can find the solutions to a linear congruence (provided that it has solutions).

Fermat's Little Theorem: If p is a prime, and if a is any integer prime to p, so that $gcd(a,p) = 1$, then

$$a^{p-1} \equiv 1(modp)$$

(thus, this theorem, also known as "Fermat's Little Theorem," provides an important primality test).

Proof: Let us consider the numbers $a, 2a, 3a, \ldots, (p-1)a, pa$. All of these numbers are incongruent to each other modulo p, as can be easily shown by *reductio ad absurdum*, and, therefore, their residues modulo p form the set $\{0,1, \ldots, p-1\}$. Hence,
$a \cdot 2a \cdot 3a \cdot \ldots \cdot (p-1)a \equiv [1 \cdot 2 \cdot 3 \cdot \ldots \cdot (p-1)](modp) \Leftrightarrow a \cdot 2a \cdot 3a \cdot \ldots \cdot (p-1)a \equiv (p-1)!\,(modp) \Leftrightarrow a^{p-1}(p-1)! \equiv (p-1)!\,(modp)$, and, dividing both sides by $(p-1)!$, which is prime to p, we obtain $a^{p-1} \equiv 1(modp)$, *quod erat demonstrandum.*

Fermat's Little Theorem implies the following "primality test": given an integer n, we can test whether it is prime by checking whether $a^{n-1} \equiv 1(modn)$ for any integer a not divisible by n. If this congruence holds, then n is likely to be prime; this is a necessary but not sufficient condition. For instance, in order to test if 23 is a prime number, we need to calculate $a^{22} \equiv 1(mod23)$ for different integral values of a, and, indeed, we shall always get the 22 nd power of a to be congruent to 1 modulo 23. Howerver, for instance, the number 561 passes the aforementioned primality test (that is, satisfies Fermat's Little Theorem), in the sense that $a^{560} \equiv 1(mod561)$, but it is a composite number $(561 = 3 \times 11 \times 17)$. The composite numbers that pass the aforementioned primality test (that is, satisfy Fermat's Little Theorem) are called "Carmichael numbers" (named after the American mathematician Robert Daniel Carmichael).

Number theory and, especially, prime numbers have important applications in cryptography. In the context of cryptography, there is a plaintext (i.e., an intelligible message) that is converted into a ciphertext (i.e., an unintelligible message) according to an encryption algorithm, and this ciphertext is transmitted on the internet and is received by a receiver who will use the decryption algorithm (which is the opposite to the encryption algorithm) in order to convert the ciphertext into the original plaintext. Thus, in cryptography, the computer converts information into a single number (representing one's message), say m. In order to be computationally secure, many encryption algorithms are based on prime numbers because of the following reason: generally, multiplying two large prime numbers can be very fast, but it is very difficult to do the reverse (it is extremely computer-intensive to find the prime factors of large numbers).

The Rational Numbers

$\mathbb{Q}$: the "rational numbers," namely, the set of all numbers of the form $\frac{p}{q}$ such that the numbers p and q are integers, $q \neq 0$, and the greatest common divisor (gcd) of the integers p and q is ± 1 (that is, p and q are relatively prime integers). In other words, the integral and the fractional numbers (both positive and negative) form together the set of rational numbers, which is denoted by $\mathbb{Q}$. The notation $\mathbb{Q}$ for the set of rational numbers derives from the Italian word "quoziente," which means "quotient."

By the term "common fraction," we refer to a number of the form $\frac{m}{n}$, where m and n are integral numbers, and $n \neq 0$. The number m is called the "numerator" of the fraction, and the number n is called the "denominator" of the fraction. In particular, n may be equal to 1. In this case, we usually write m rather than $\frac{m}{1}$. In other words, any integral number can be represented in the form of a common fraction whose denominator is 1.

Two fractions $\frac{a}{b}$ and $\frac{c}{d}$ are regarded to be equal if $ad = bc$. The "basic property of fractions" states the following: the fractions $\frac{a}{b}$ and $\frac{am}{bm}$ are equal. Therefore, if the numerator and the denominator of a given fraction are multiplied or divided by the same positive integer, then an equivalent fraction is obtained (namely: $\frac{a}{b} = \frac{am}{bm}$). Taking advantage of the basic property of fractions, we may sometimes replace a given fraction with another equivalent fraction but with a smaller numerator and a smaller denominator by dividing all common factors out of the numerator and the denominator. This operation is called "reduction of a fraction to its lowest terms," or simply "reduction of a fraction." In general, reduction of a fraction is always possible if its numerator and denominator are not relatively prime numbers. If the numerator and the denominator are relatively prime numbers, then the fraction is called "irreducible."

The addition of common fractions is defined in the following way:

$\frac{a}{b} + \frac{c}{d} = \frac{ad+bc}{bd}$.

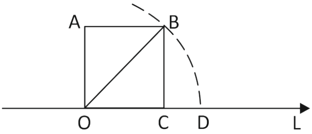
The subtraction of common fractions is defined in the following way:
$\frac{a}{b}-\frac{c}{d}=\frac{ad-bc}{bd}$.
The multiplication of common fractions is defined in the following way:
$\frac{a}{b}\cdot\frac{c}{d}=\frac{ac}{bd}$.
The division of common fractions is defined in the following way:
$\frac{a}{b}\div\frac{c}{d}=\frac{a/b}{c/d}=\frac{ad}{bc}$.
A fraction $\frac{m}{n}$ is called a "proper fraction" if its numerator is less than the denominator; and it is called an "improper fraction" if its numerator is greater than the denominator.
Let us consider an improper fraction $\frac{m}{n}$. Since m is greater than n, there are two numbers p and r (with r less than n) such that $m=pn+r$, so that: $\frac{m}{n}=\frac{pn+r}{n}=\frac{pn}{n}+\frac{r}{n}=p+\frac{r}{n}$. Since the remainder is always less than the divisor, $\frac{r}{n}$ is a proper fraction. Hence, we have succeeded in representing the improper fraction $\frac{m}{n}$ in the form of a sum of a natural number p and a proper fraction $\frac{r}{n}$. This operation is called the "reduction of an improper fraction to a mixed number." A number consisting of an integer and a fraction is called a "mixed number." For instance, in order to locate the mixed number $3\frac{1}{8}$ on the number line, we think as follows: laying off the unit segment ($0P=+1$) from the point zero in the given (positive) direction three times and then $\frac{1}{8}$th part of this unit segment, we obtain the point Q that exactly corresponds to the mixed number $3\frac{1}{8}$ (the coordinate of the point Q is $3\frac{1}{8}$).

The Irrational Numbers

$\mathbb{Q}^{\sim}$: the "irrational numbers," or the set of all numbers that cannot be written as the quotient of two relatively prime integers. For instance, we can prove that $\sqrt{2}\in\mathbb{Q}^{\sim}$ by *reductio ad absurdum* as follows: For the sake of contradiction, suppose that $\sqrt{2}=\frac{p}{q}$ where $p,q\in\mathbb{Z}$, the greatest common divisor of the integers p and q is ± 1, and $q\neq 0$. Then
$\sqrt{2}=\frac{p}{q}\Rightarrow 2=\frac{p^2}{q^2}\Rightarrow p^2=2q^2\Rightarrow p=2k$,
where k is an appropriate integer; therefore $4k^2=2q^2\Rightarrow q^2=2k^2$; but then the greatest common divisor of the integers p and q is 2, which contradicts the hypothesis.

The German mathematician Richard Dedekind (1831–1916) observed that there exist infinitely many points on the straight number line L that correspond to no rational number. Thus, the domain of rational numbers is insufficient if we want to arithmetically follow up all phenomena on the straight line. Therefore, new numbers must be created in such a way that the domain of all numbers will gain the same "completeness" or "continuity" as the straight line. In fact, Dedekind observed that there exist infinitely many cuts that are not produced by rational numbers. For instance, as shown in Figure 2-1, construct a square $OABC$ on the unit segment OC (i.e., the length of OC is equal to one) and lay off in the positive direction a line segment OD equal in length to the diagonal OB; then (according to the Pythagorean Theorem, which we shall study shortly) it is clear that D is a point that does not correspond to any rational number—it, in fact, corresponds to $\sqrt{2}$.

Figure 2-1: Irrational numbers.



The history of irrational numbers goes back to the Pythagorean mathematicians, who had demonstrated that there exist lengths incommensurable with a given unit of length. In the seventh century B.C.E., Thales of Miletus (a Greek mathematician, astronomer, and philosopher from Miletus, in Ionia, Asia Minor) officially initiated a new approach to mathematics. In contrast to the mathematics of other civilizations, such as the Babylonians and the Egyptians, Thales's approach to mathematics is based on the thesis that scientific propositions are not recipes for practical tasks—that is, techniques whose validity is determined by the method of trial and error—but they should be explained and proved. In other words, Thales attempted to endow mathematics with rigor—which, in this case, means logical validity.

In the context of Thales's rigorous mathematics, by the term "line segment," we mean a part of a line that is bounded by two distinct endpoints, and contains every point on the line between the endpoints. Let us consider the line segments $a_1, a_2, a_3, \dots, a_n$ and the non-zero line segments $b_1, b_2, b_3, \dots, b_n$. The line segments $a_1, a_2, a_3, \dots, a_n$ are said to be "proportional" to $b_1, b_2, b_3, \dots, b_n$, respectively, if
$\frac{a_1}{b_1} = \frac{a_2}{b_2} = \frac{a_3}{b_3} = \cdots = \frac{a_n}{b_n}$.

Thus, two arbitrary line segments a and c are proportional to two other arbitrary line segments b and d, respectively, if and only if b and d are non-zero, and it holds that

$$\frac{a}{b} = \frac{c}{d}. \tag{1}$$

Any equality between two ratios, such as (1), is said to be a "proportion" with terms a, b, c, and d, as shown above.

Assume that AB is a non-zero straight line segment, and that P is a point on AB. Then we say that the point P "divides internally" the straight line segment AB in a ratio λ, where $\lambda \geq 0$, if it holds that

$$\frac{PA}{PB} = \lambda.$$

If this is the case, then

$\frac{PA}{PA+PB} = \frac{\lambda}{\lambda+1} \Leftrightarrow PA = \frac{\lambda}{\lambda+1}AB$, which implies the uniqueness of P.

Similarly, we say that a point Q "divides externally" the straight line segment AB in a ratio λ, where $\lambda > 0$, if the points A, B, and Q are collinear, Q is external to AB, and it holds that

$$\frac{QA}{QB} = \lambda.$$

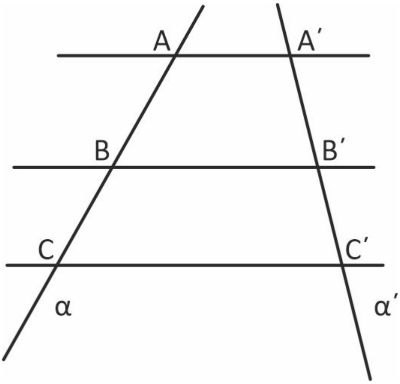
If this is the case, then $\frac{QA}{|QA-QB|} = \frac{\lambda}{|\lambda-1|}$ (given that $QA \neq QB$, it holds that $\lambda \neq 1$), so that

$QA = \frac{\lambda}{|\lambda-1|}$ AB, which implies the uniqueness of Q.

Thales's Theorem: If parallel straight lines intersect two straight lines, then they define proportional straight line segments on them. For instance, if parallel straight lines l_1, l_2, and l_3 intersect straight lines a and a' at points A, B, C and A', B', C' respectively, as shown in Figure 2-2, then

$$\frac{AB}{A'B'} = \frac{AC}{A'C'} = \frac{BC}{B'C'}.$$

Figure 2-2: Thales's Theorem.



Corollary 1: Every straight line that is parallel to the bases of a trapezoid divides, internally or externally, the non-parallel sides of the given trapezoid in equal ratios.

Corollary 2: Every straight line that is parallel to one side of a triangle divides, internally or externally, the other two sides of the given triangle in equal ratios.

Corollary 3: If two triangles have a common angle, and if they have parallel opposite sides, then they are said to be in Thales position, and then they are similar triangles and have proportional sides.

In the sixth century B.C.E., Pythagoras and his school (the so-called "Pythagoreans") endorsed Thales's approach to mathematics. From the Pythagorean perspective of mathematics, the relations between the objects of the world (e.g., magnitudes) correspond to the relations between natural (and, generally, integral) numbers. However, it was soon realized that things are not so simple, since it was realized that there exist magnitudes that do not have a common measure. According to the Pythagoreans, two objects (magnitudes) are "commensurable" (that is, they have a common measure) if and only if there is a magnitude of the same kind that is contained an integral number of times in both of them. In other words, two magnitudes are "commensurable" if and only if their ratio is a rational number. However, the Pythagoreans encountered "incommensurable" magnitudes: magnitudes whose ratio is an irrational number. For instance, as shown in Figure 2-1, the length of a diagonal of a unit square (i.e., of a square with sides measuring 1 unit) is, according to the Pythagorean Theorem, equal to $\sqrt{2}$, which is an irrational number. Similarly, a circle's circumference and its diameter are incommensurable (that is, π, the ratio of a circle's circumference to its diameter, is an irrational number). The awareness that there exist incommensurable magnitudes compelled ancient Greek mathematicians to inquire into the relations between incommensurable magnitudes. This event marked a major crisis in ancient mathematics.

According to ancient Greek mathematicians, quantities (magnitudes) are continuous and uniform objects, which are best represented by straight line segments. Their division into parts, or their measurement in terms of a "unit of measurement" (i.e., a definite magnitude of a quantity), meanwhile, represents the notion of discreteness. Ancient Greek mathematicians used the term "ratio of magnitudes" in order to refer to the relation between two magnitudes that can be measured in terms of a common unit of measurement. Thus, the ancient Greek concept of a ratio is most similar to the more abstract modern concept of a number. In the context of ancient Greek mathematics, the objects of mathematics were quantities (represented by straight line segments), and the ratio between two quantities was a meta-object, or something that was used in order to study mathematical objects without being treated as a mathematical object itself. In other words, in the context of ancient Greek mathematics, a ratio (a number) was construed as a measuring relationship between two quantities, and such a measuring relationship could be built up (and, hence, proved) in finitely many steps, using a common unit of measurement. Nevertheless, the discovery of incommensurable ratios demonstrated that a ratio could not be interpreted as a measuring relationship in the aforementioned way. In fact, as a result of the discovery of incommensurable ratios, the concept of a ratio (or a number) acquired its conceptual autonomy, and, instead of being treated as a meta-object, it started being treated as an object of mathematics. Therefore, ancient Greek mathematicians had to transcend the system of mathematics that was based on commensurable ratios. Notice that a commensurable ratio could easily become an object of mathematical theory, since it is a rational number, and, therefore, it can be constructed in finitely many steps, whereas the decimal representation of an irrational number neither terminates nor infinitely repeats but extends forever without regular repetition.

In the fourth century C.E., Theon, one of the most important Greek mathematicians and commentators of Euclid's and Ptolemy's works, attempted to solve the problems that were generated as a result of the aforementioned crisis in the foundations of ancient Greek mathematics. In particular, Theon started from an extremely small (infinitesimal) unit square such that the ratio between any of its sides and any of its diagonals is equal to 1 (given that it is infinitely small); symbolically, if a_1 is the

length of any of the sides of the given infinitesimal unit square, and if δ_1 is the length of any of the diagonals of the given infinitesimal unit square, then $\frac{\delta_1}{a_1} = 1$ (rather than $\sqrt{2}$). Subsequently, Theon formulated a recursive sequence of squares defined by the following rule:
$a_n = \delta_{n-1} + a_{n-1}$ and $\delta_n = 2a_{n-1} + \delta_{n-1}$,
so that the ratio between a diagonal and a side of a unit square approaches its actual value, that is,
$\frac{\delta_n}{a_n} \to \sqrt{2}$. The aforementioned recursive sequence yields $\frac{1}{1}, \frac{3}{2}, \frac{7}{5}, \frac{17}{12}, \frac{41}{29}, \frac{99}{70}, \ldots$ beginning with $\frac{\delta_1}{a_1} = \frac{1}{1}$, and, thus, Theon's rule provides an algorithm for successive approximations to the square root of 2. Theon explained that he started from the case in which $\frac{\delta_1}{a_1} = 1$ because, just as the sperm of a living organism encompasses subsequent properties of the given organism, any ratio "spermatically" (i.e., at the infinitesimal level) encompasses the unit.

Theon's aforementioned reasoning is underpinned by Aristotle's concept of a "potential infinity." The concept of modern mathematics that is semantically most similar to Aristotle's concept of a "potential infinity" is the convergence of a sequence of natural numbers. Thus, from the perspective of ancient Greek mathematics, infinity is not a being (i.e., it is not an actual state); it cannot be simultaneously considered in its whole extension, but it can only be considered as a becoming (i.e., a process). In this way, the concept of an infinite approach helps us to overcome the contradiction between incommensurable ratios and commensurable ratios, since we can think of an incommensurable ratio infinitely approaching a commensurable ratio (and vice versa). Similarly, the concept of an infinite approach helps us to overcome the contradiction between broken lines and curves, as well as the contradiction between continuity and discreteness. This reasoning is endorsed by Euclid; in his *Elements*, he does not consider infinitely extended straight lines, but he always works with straight line segments which, as he says, can be extended as much as one needs.

However, several intellectuals have used infinite processes in a way that is not rigorous. For instance, they have attempted to compute the length of the circumference of a circle by considering an inscribed polygon whose number of sides increases indefinitely. Therefore, the length of each side of such a polygon decreases indefinitely, so that a triangle whose base is a side of the given polygon and whose vertex (i.e., the "top" corner opposite its base) is the center of the given circle could become such that its base coincides with the given circle's circumference. To what extent is such a shape a triangle, and beyond which point does an arc become a chord? One may argue that these changes happen when a straight line segment becomes infinitely small, but then one may counter-argue that, by becoming infinitely small, a straight line segment is not "something" any more, and it becomes "nothing." Hence, how is it possible that an infinite series of "nothing" ("no-things") gives "something," such as a circle? The aforementioned example indicates the problems that are generated as a result of the use of infinite processes in computations.

The aforementioned crisis in the foundations of mathematics was overcome by Eudoxus's theory of proportions and by the method of exhaustion, which derives from Eudoxus's theory of proportions, and it was used by Archimedes. The method of exhaustion was originally developed in the fifth century B.C.E. by the Athenian intellectual Antiphon, and it was put in a rigorous scientific setting shortly afterwards by the Greek mathematician and astronomer Eudoxus of Cnidus, who used it in order to calculate areas and volumes. The Greek mathematician Euclid (the acknowledged father of "Euclidean geometry") and the Greek mathematician, physicist, and engineer Archimedes made extensive use of the method of exhaustion in order to prove several mathematical propositions. For instance, Archimedes used the method of exhaustion in order to compute the area of a circle by approximating the area of a circle from above and below by circumscribing and inscribing regular polygons of an increasingly larger number of sides (so that sides become "infinitesimals," or infinitely small): each of the polygons is a union of triangles, so it is easily verified that the area of a circle of radius r and circumference C is equal to the area of a triangle whose altitude is equal to r and whose base is equal to $C = 2\pi r$. Then, given that the area of a triangle is equal to half of the product of its base and altitude, we obtain the formula for the computation of the area of a circle: $\frac{1}{2}(rC) = \frac{1}{2}(r2\pi r) = \pi r^2$. Moreover, Archimedes was able to calculate the length of various tangents to the spiral (i.e., to a curve emanating from a point moving farther away as it revolves around the point).

Archimedes was very careful in the use of infinite processes; he approximated π by using the fact that the circumference of a circle is bounded by the perimeter of an *inscribed* polygon and by the perimeter of a *circumscribed* polygon. According to Eudoxus and Archimedes, there is always a ratio between any two magnitudes, and we can always make any magnitude smaller or greater than a given magnitude, so that the ratio between two magnitudes a and b is the same as the ratio between two other magnitudes c and d if and only if, for any natural numbers m and n, it holds that

$$ma \gtreqless nb \Rightarrow mc \gtreqless nd, \tag{2}$$

meaning that both of these ratios are characterized by the same placement property (i.e., ordering) with regard to other numbers. In (2), the equality sign ($=$) refers to commensurable ratios, whereas the inequality signs ($\gtrless$) refer to incommensurable ratios. These ideas of Eudoxus and Archimedes indicate that ancient Greek mathematicians discovered not only incommensurable magnitudes but also incommensurable numbers. Eudoxus's aforementioned theory of proportions underpins Archimedes's method of exhaustion for solving geometric problems, and Archimedes's method of exhaustion underpins modern infinitesimal calculus.

It is important to notice that the way in which Eudoxus solved the problem of the existence of incommensurable ratios (specifically, his attempt to study the conundrum of irrationality that appears to exist in elementary geometry in a scientifically rigorous way) marks a shift away from the traditional constructivist approach to mathematics towards formalism. In other words, Eudoxus does not explain what a ratio is (as a mathematical object), but he states only when two ratios are similar to each other. The constructivist approach to mathematics allows us to determine what an object is by being able to construct it, whereas the formalist approach to mathematics is not concerned with the substance of the mathematical object under consideration, and is concerned only with the relations between the mathematical object under consideration and other mathematical objects. Moreover, the ideas of Eudoxus and Archimedes are conceptually very similar to Dedekind's cuts.

Fusing geometry and arithmetic is an arduous task. In order to understand the difficulties that originate from fusing geometry and arithmetic, let us consider, for instance, the famous irrational number $\sqrt{2}$, which was discovered by Pythagoreans when they attempted to compute the length of a diagonal of a unit square.

The Pythagoreans realized that the diagonal of a unit square is not commensurable with the side of the given square, but, by keeping geometry and arithmetic separate from each other (that is, by refusing to identify numbers with lengths of straight line segments), ancient Greek mathematicians could argue as follows: given a straight line segment whose length is one, we can construct a straight line segment whose length is $\sqrt{2}$ (as shown in Figure 2-1). In general, irrational numbers are geometrically constructible (and, hence, geometrically explicable and manageable), even though, from the perspective of arithmetic, irrational numbers are ideal quantities, in the sense that the calculation of irrational numbers (such as $\sqrt{2}$) is an infinite process (namely, irrational numbers have infinitely many decimal digits).

On the other hand, having endorsed the Cartesian approach to mathematics, mathematicians in the nineteenth century realized that they had to clarify some still ambiguous fundamental concepts (such as that of a real number), to formulate new methods of doing mathematics in a logically rigorous way, and to create a rigorous theory of the arithmetic continuum—specifically, a rigorous theory of real numbers and their arithmetic.

The Real Numbers

$\mathbb{R}$: the "real numbers," or the set that is formed by the union of the set $\mathbb{Q}$ of all rational numbers and the set $\mathbb{Q}^{\sim}$ of all irrational numbers; symbolically: $\mathbb{R} = \mathbb{Q} \cup \mathbb{Q}^{\sim}$.

Based on and following the methodology of the fifth volume of Euclid's *Elements* (that is, the mathematical work of Eudoxus), Richard Dedekind formulated the modern theory of real numbers. He began with the following three properties of rational numbers:

i. If $a > b$ and $b > c$, then $a > c$.
ii. If a and c are two distinct (rational) numbers, then there exist infinitely many distinct numbers lying between a and c.
iii. If a is any definite (rational) number, then all numbers of the system $\mathbb{Q}$ fall into two classes, A_1 and A_2, each of which contains infinitely many individuals; A_1 contains all numbers a_1 that are $< a$, while A_2 contains all numbers a_2 that are $> a$; the number a itself may be assigned at will to A_1 or A_2, being, respectively, the greatest number of A_1 or the least number of A_2.

Then Dedekind stated three properties of the points on a straight number line L:

i. If p lies to the right of q and q to the right of r, then p lies to the right of r; and q is said to lie between p and r.
ii. If p and r are two distinct points, then there always exist infinitely many points lying between p and r.
iii. If p is a definite point on L, then all points on L fall into two classes, P_1 and P_2, each of which contains infinitely many individuals; P_1 contains all the points p_1 that lie to the left of p, while P_2 contains all the points p_2 that lie to the right of p; the point p itself may be assigned at will to P_1 or P_2. In any case, every point of P_1 lies to the left of every point of P_2.

Each such division (or partition) of the set $\mathbb{Q}$ of all rational numbers defines a "cut," called the "Dedekind's cut." However, after having observed that every rational number effects a "cut" in the set of rationals, Dedekind considered the inverse question: if, by a given criterion, the set of rationals is divided into two subsets A and B so that every number in A is less than every number in B, is there always a greatest rational in A or a smallest rational in B? Dedekind immediately realized that the number line should be "continuous," or unbroken, in the intuitive sense. Like Eudoxus and Cantor before him, he developed theoretical concepts for the purpose of filling the gaps in the ordered set of rationals so that the final geometric picture is a continuous, straight number line. However, the answer to the last question is in the negative: when A has no maximum rational and B has no minimum rational, there is, indeed, a gap in the rational series (or a puncture in the number line) which must be filled. In that case, the cut (A, B) is said to define (or to be) an irrational number (as shown, for instance, in Figure 2-1). Hence, the set $\mathbb{R}$ of all real numbers is called the "(arithmetic or geometric) continuum" or the "straight line of real numbers."

In modern mathematical notation, the set of all real numbers x such that $a \leq x \leq b$ is said to be a "closed interval," denoted by $[a, b]$, of the real line $\mathbb{R}$, while the set of all real numbers x such that $a < x < b$ (which does not include its endpoints) is said to be an "open interval," denoted by (a, b), of the real line $\mathbb{R}$. The intervals $[a, b) = \{x \in \mathbb{R} | a \leq x < b\}$ and $(a, b] = \{x \in \mathbb{R} | a < x \leq b\}$ are neither open nor closed, but they are sometimes called "half-open" or "half-closed." Notice that $(a, a) = \emptyset$, and $[a, a] = \{a\}$. Moreover, we define the intervals:

$(a, \infty) = \{x \in \mathbb{R} | a < x\}$,
$[a, \infty) = \{x \in \mathbb{R} | a \leq x\}$,
$(-\infty, a) = \{x \in \mathbb{R} | x < a\}$,
$(-\infty, a] = \{x \in \mathbb{R} | x \leq a\}$.

By the term "interval," we generally mean a set of points with the property that, if x and y are distinct points of the set, every point between x and y is also a point of the set (if the points x and y are included, then the interval is closed; otherwise, it is open).

A real number b is said to be an "upper bound" of a non-empty subset S of $\mathbb{R}$ if every member of the set S is less than or equal to the number b, symbolically, if $x \leq b \, \forall x \in S$. If this is the case, then S is said to be "bounded from above." For instance, if $S = \{2,4,6,8,10\}$, then 10 is an upper bound of S, and every real number greater than 10 is also an upper bound of S. Notice that, if a set is bounded from above, then it has infinitely many upper bounds, and that an upper bound of such a set need not be a member of the given set. For instance, the number 10 is an upper bound of the open interval $(2,10)$, but $10 \notin (2,10)$. On the other hand, the set ℕ of all natural numbers has no upper bound.

The least of all upper bounds of a set is said to be the "least upper bound" (often denoted by $l.u.b.$), or the "supremum" (often denoted by sup). Hence, a real number b is defined to be the $l.u.b.$ of a set S if b is an upper bound of S (i.e., $x \leq b \, \forall x \in S$), and if , given any other upper bound c of S, $b < c$; and then we write $sup(S) = b$. For instance, if $S = \{2,4,6,8,10\}$, then $sup(S) = 10$. On the other hand, the set ℕ of all natural numbers has no supremum. The supremum, when it exists, is unique for a set.

A real number a is said to be a "lower bound" of a non-empty subset S of ℝ if every member of the set S is greater than or equal to the number a, symbolically, if $x \geq a \, \forall x \in S$. If this is the case, then S is said to be "bounded from below." For instance, if $S = \{2,4,6,8,10\}$, then 2 is a lower bound of S, and every real number less than 2 is also a lower bound of S. Notice that, if a set is bounded from below, then it has infinitely many lower bounds, and that a lower bound of such a set need not be a member of the given set. For instance, the number 2 is a lower bound of the open interval $(2,10)$, but $2 \notin (2,10)$. On the other hand, the set ℤ of all integral numbers has no lower bound.

The greatest of all lower bounds of a set is said to be the "greatest lower bound" (often denoted by $g.l.b.$), or the "infimum" (often denoted by inf). Hence, a real number a is defined to be the $g.l.b.$ of a set S if a is a lower bound of S (i.e., $x \geq a \, \forall x \in S$), and if , given any other lower bound d of S, $a > d$; and then we write $inf(S) = a$. For instance, if $S = \{2,4,6,8,10\}$, then $inf(S) = 2$. On the other hand, the set ℤ of all integral numbers has no infimum. The infimum, when it exists, is unique for a set.

A set is said to be "bounded" if it is both bounded from above and bounded from below. In other words, a set S is bounded if there exist two real numbers a and b such that $a \leq x \leq b \, \forall x \in S$. If this is the case, then $x \in [a, b] \, \forall x \in S$, meaning that, for any bounded set S, there exist two real numbers a and b such that $S \subseteq [a, b]$.

Notice that the empty set, $\emptyset$, is a subset of every set, and, $\forall a, b \in$ ℝ, $\emptyset \subseteq [a, b]$. Therefore, $\emptyset$ is a bounded set. Because of the fact that $\emptyset \subseteq [a, b]$ for any real numbers a and b, every real number is a lower bound of $\emptyset$, and every real number is an upper bound of $\emptyset$, meaning that $\emptyset$ does not have a supremum or an infimum.

Moreover, notice that, for an arbitrary singleton $A = \{x\}$, $sup(A) = x = inf(A)$. Thus, every singleton is a bounded set in which $supremum = infimum$.

If the supremum of a set belongs to the given set, then it is said to be the "maximum element" of the given set. If the infimum of a set belongs to the given set, then it is said to be the "minimum element" of the given set. For instance, 5 is the maximum element of the set (closed interval) $[-3,5]$, and -3 is the minimum element of this set. However, the set (open interval) $(-3,5)$ does not have a maximum element or a minimum element.

Assume that ε is a positive real number—that is, $\varepsilon > 0$. Moreover, consider the open interval $N = (p - \varepsilon, p + \varepsilon)$. Hence, $p \in (a, b) \subseteq (p - \varepsilon, p + \varepsilon)$. If this is the case, then $(p - \varepsilon, p + \varepsilon)$ is called the " ε-neighborhood" of the point p, and it is denoted by $N_\varepsilon(p)$. In other words, the ε-neighborhood of a point p on the real line is the set of all those real numbers which are within an ε distance of p on either side of it; p is the midpoint or the center of $N_\varepsilon(p)$; and ε is the radius of $N_\varepsilon(p)$. In other words, a subset N of $\mathbb{R}$ is said to be a neighborhood of a real number p if there exists an open interval (a, b) containing p and itslf contained in N, symbolically, $p \in (a, b) \subseteq N$. Then the set $\mathbb{N}$ of all natural numbers, the $\mathbb{Z}$ of all integers, the set $\mathbb{Q}$ of all rational numbers, and the set $\mathbb{Q}^{\sim}$ of all irrational numbers are not neighborhoods of any of their elements, whereas the set $\mathbb{R}$ of all real numbers itself is a neighborhood of each of its elements.

We shall use the notation $N_\varepsilon^{'}(p)$ in order to denote the "deleted neighborhood," consisting of $N_\varepsilon(p)$ with the point p deleted. In terms of the real line $\mathbb{R}$, a deleted neighborhood is an interval on $\mathbb{R}$ with the center point removed.

Given a set S, a real number p is said to be an "interior point" of S if S is a neighborhood of p; symbolically: if $p \in (a, b) \subseteq S$. Obviously, an interior point of a set S belongs to S. The set of all interior points of a given set S is called the "interior" of S, and it is denoted by $Int(S)$. In general, a point $p \in \mathbb{R}^n$ is said to be an "interior point" of U if some neighborhood (open ball) $N_\varepsilon(p)$ with center p is contained in U. For instance, if $S = [2,5]$, then $\frac{7}{2}$ is an interior point of S, whereas neither 2 nor 5 is an interior point of S, because $[2,5]$ is not a neighborhood of 2 and 5. The interior of the closed interval $[2,5]$ is the open interval $(2,5)$. The points of the "boundary" of a set S are those points on the edge of S separating the interior of S from its exterior; and, more formally, we can say that a point p is a "boundary point" of a set S if and only if every neighborhood of p contains at least one point that belongs to S and one point that does not belong to S.

A real number p is called a "closure point" of a set $S \subseteq \mathbb{R}$ if every neighborhood of p contains a point of S. The set of all closure points of S is called the "closure" of S, and it is denoted by $Cls(S)$. Therefore, every point of $S \subseteq \mathbb{R}$ is a closure point of S.

A real number p is called an "accumulation point," a "limit point," or a "cluster point" of S if every deleted neighborhood of p contains at least one point of S ; symbolically: if $S \cap N'_\varepsilon(p) \neq \emptyset \, \forall \varepsilon > 0$ (in other words, every neighborhood of p contains at least one point of S other than p). For instance, if $A = [a, b]$ and $B = (a, b)$, then every member of A is an accumulation point of A and of B, since, for instance, $\forall \varepsilon > 0$, the neighborhood $(a - \varepsilon, a + \varepsilon)$ of a contains infinitely many elements of A and of B. Moreover, every real number is an accumulation point of the set $\mathbb{Q}$ of all rational numbers as well as of the set $\mathbb{R}$ of all real numbers, since, for instance, given an arbitrary real number p, $\forall \varepsilon > 0$, the neighborhood $(p - \varepsilon, p + \varepsilon)$ contains infinitely many real numbers as well as infinitely many rational numbers. On the other hand, the set $\mathbb{N}$ of all natural numbers, the set $\mathbb{Z}$ of all integral numbers, and the empty set have no accumulation point. Furthermore, no finite set has any accumulation point, because, if, for instance, $A = \{a_1, a_2, a_3, \dots, a_n\}$, and if p is an arbitrary real number, we can construct a sufficiently small neighborhood N with center p such that N contains no point of A; therefore, p, which is an arbitrary real number, is not an accumulation point of A.

Every accumulation point of a set is also a closure point of that set, but not conversely. For instance, given the set $A = \left\{\frac{1}{n} | n \in \mathbb{N} - \{0\}\right\}$, $0 = inf(A)$ and $0 \notin A$, and, therefore, 0 is an accumulation point of A, but 1 is a closure point of A without being an accumulation point of A, since a neighborhood $(1 - \varepsilon, 1 + \varepsilon)$, where $\varepsilon > 0$, does not contain a member of A other than 1.

The following theorem, known as the Bolzano–Weierstrass Theorem, can be easily deduced from the principles of the Dedekind's cuts:

Bolzano–Weierstrass Theorem: If a set S contains infinitely many points of the real line, and if it is entirely contained in an open interval (a, b), then at least one point of that interval is a point of accumulation of S. In other words, every bounded infinite set of real numbers has at least one accumulation point. Indeed, if we define a Dedekind's cut (P_1, P_2) with $a \in P_1$ and $b \in P_2$, then there is a poit ξ such that, however small be ε, $\xi - \varepsilon \in P_1$ and $\xi + \varepsilon \in P_2$, so that the interval $(\xi - \varepsilon, \xi + \varepsilon)$ contains infinitely many points of S, and, therefore, ξ is a point of accumulation of S; in fact, this point may coincide with a or b, as, for instance, when $a =$

0, $b = 1$, and S consists of the points $1, \frac{1}{2}, \frac{1}{3}, \ldots$, in which case 0 is the sole point of accumulation.

The Bolzano–Weierstrass Theorem is important, because, by guaranteeing the sufficiency of the density of $\mathbb{R}$, it provides a rigorous way of proving the convergence of infinite sequences of real numbers. Bernard Bolzano (1781–1848), an Italian-Czech mathematician, philosopher, theologian, and Catholic priest, and Karl Weierstrass (1815–97), a German mathematician, were pioneering advocates of rigor in mathematical analysis.

We can think of infinity in two ways: either as an indefinite quantity whose size has exceeded all limits or as a definite quantity that we imagine growing continuously, but the latter always remains less than what we call actual infinity. Thus, we come up with two types of infinity: one is absolute and static, corresponding to the notion of an indefinite quantity, and the other is dynamical, corresponding to a definite yet potentially continuously growing quantity. Furthermore, it is important to understand the difference between infinity *per se* and an infinite quantity. Infinity *per se* is an idea of pure reason, whereas an infinite quantity is a constructive concept based on the idea of infinity. Infinity *per se* does not coincide with any empirical (or theoretical) quantity, but it is a critical intellectual capability that enables us to characterize any particular quantity as finite by conceiving a quantitative space that contains the entire finite and is based on the notion of transcendence. The negation of the finite—and, more precisely, the conception of a process of transcending any particular quantity—refers us to the idea of infinity. Whereas infinity *qua* infinity is an idea of pure reason, an infinite quantity manifests itself through several concepts, so that, for instance, we refer to the infinite quantity (or the infinite size) of the natural numbers, of a line, of a plane, of a geometric space, etc.

A "real number" is a quantity x that has a "decimal expansion":

$x = n + 0.d_1 d_2 d_3 \ldots$,

where n is an integer, each d_i is a digit between 0 and 9 ($i = 1,2,3, \ldots$), and no infinite sequence of 9's appears (0.999 ... with an infinite sequence of 9 's is exactly the same number as 1). The aforementioned representation implies that

$$n + \frac{d_1}{10} + \frac{d_2}{100} + \cdots + \frac{d_k}{10^k} \leq x < n + \frac{d_1}{10} + \frac{d_2}{100} + \cdots + \frac{d_k}{10^k} + \frac{1}{10^k},$$

for all positive integers k.

Exponents, factorials, and logarithms: Let a be a real number. Then the product $a \cdot a \cdot a \ldots$ (n times) is denoted by a^n, where n is called the "exponent," and a is called the "base." Therefore, the following properties of exponents hold $\forall a, b \in \mathbb{R}$:

i. $a^n a^m = a^{n+m}$,
ii. $(a^n)^m = a^{nm}$,
iii. $\frac{a^n}{a^m} = a^{n-m}$,
iv. $a^0 = 1$, and
v. $\left(\frac{a}{b}\right)^n = \frac{a^n}{b^n}$.

A "factorial" is a function in mathematics denoted with the symbol ! that multiplies a positive integer n by every number that precedes it:
$n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1$.
For instance, $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$. Notice that $0! = 1$, and $1! = 1$. In other words, $n!$ ("n factorial") is the product of all natural numbers from 1 to n.
The number of "permutations" (ordered arrangements) of n elements taken m at a time is $P_m^n = \frac{n!}{(n-m)!}$, and the number of "combinations" of n elements taken m at a time is $C_m^n = \frac{n!}{(n-m)!m!} = \frac{P_m^n}{m!}$. Notice that the term "permutation" means the number of ways we can arrange a set of objects in a specific order (in this case, the order of the objects matters), whereas the term "combination" means the number of ways we can select a subset of objects from a larger set without taking the order of the objects into consideration. For instance, consider a group of 10 persons. If we want to form a subgroup, a subcommittee, of 3 persons from this group, then this is a combination problem (since the order we select persons for the subcommittee doesn't change the subcommittee we form), and, therefore, we apply the formula $C_3^{10} = \frac{10!}{(10-3)!3!} = 120$ (there are 120 ways to select a group of 3 persons from a group of 10 persons). If we want to select a President, a Vice President, and a Secretary from this group (of 10 persons), then this is a permutation problem (since the order we select them changes the role that they perform), and, therefore, we apply the formula $P_3^{10} = \frac{10!}{(10-3)!} = 720$ (there are 720 ways to select a President, a Vice President, and a Secretary from a group of 10 persons). But if we simply want to find the number of ways that a group of 10 persons can arrange themselves in a row of 10 chairs, then the answer is 10! (the 10 persons can arrange themselves in a row in 10! ways).
Intimately related to the concepts of an exponent and an index is the concept of a logarithm, which is the inverse function to exponentiation.

The "logarithm" of a number a is the exponent to which another fixed number, the base b, must be raised to produce the number a; symbolically:
$log_b a = x \Leftrightarrow b^x = a$,
where b and a are positive numbers with $b \neq 1$.
For instance, $log_{10} 1{,}000 = 3$, since $10^3 = 1{,}000$, and $log_3 81 = 4$, since $3^4 = 81$. The method of logarithms was originally developed by the Scottish mathematician, physicist, and astronomer John Napier (1550–1617), who published his book *Mirifici Logarithmorum Canonis Descriptio* (*Description of the Wonderful Rule of Logarithms*) in 1614.
The following properties of the logarithm can be easily verified (since they derive from the properties of exponents):

i. $log_b(xy) = log_b x + log_b y$,
ii. $log_b\left(\frac{x}{y}\right) = log_b x - log_b y$,
iii. $log_b x^k = k log_b x$,
iv. $log_b 1 = 0$,
v. $log_b b^x = x = b^{log_b x}$,

where b, x, y are positive, with $b \neq 1$, and k any real number.
Equation-solving principle: If x, y, and b are positive real numbers with $b \neq 1$, then
$x = \mathrm{y} \Rightarrow log_b x = log_b y$, and, conversely,
$log_b x = log_b y \Rightarrow x = y$.
Therefore, we can solve exponential equations (i.e., equations in which the unknown is in the exponent) by taking the logarithm of both sides of the equation. For instance, let us solve the exponential equation $5^{2x} = 21$ for x, using log base of 10: $5^{2x} = 21 \Rightarrow log(5^{2x}) = log21 \Rightarrow 2x \cdot log5 = log21 \Rightarrow 2x = \frac{log21}{log5} \Rightarrow x = \frac{\frac{log21}{log5}}{2} \approx 0.9458$.
Change of base rule: We may change a logarithm in one base to a logarithm in another base according to the following rule:
$log_b x = \frac{log_a x}{log_a b}$.
The number e and the "natural logarithm": Now, let us consider exponential expressions that represent phenomena that change continuously, such as the concept of compound interest. By the term "compound interest," we mean the interest calculated on the principal (the invested/borrowed initial sum) and the interest accumulated over the corresponding period of time (i.e., compound interest differs from simple interest, where interest is not added to the principal when we calculate the interest during the next period of time). Let P denote the principal, r denote the interest rate, n denote the number of times interest is

compounded per year, t denote time (in years), and A denote the amount (including principal and interest) of an investment or a loan. Then:

$$A = P\left(1 + \frac{r}{n}\right)^{nt}$$

(this is the formula of compound interest). For instance, using the formula of compound interest, let us examine the return on an $\$1$ investment for one year at an investment rate of 100%. Moreover, progressively, we shall compound our investment more frequently and observe the result (we chose $P = \$1$ and $r = 100\%$ in order to illustrate the situation with the easiest numbers). Therefore, if compounded:

annually ($n = 1$), then $A = \left(1 + \frac{1}{1}\right)^1 = \$2.00$;

semiannually ($n = 2$), then $A = \left(1 + \frac{1}{2}\right)^2 = \$2.25$;

quarterly ($n = 4$), then $A = \left(1 + \frac{1}{4}\right)^4 \approx \$2.441$;

monthly ($n = 12$), then $A = \left(1 + \frac{1}{12}\right)^{12} \approx \$2.613$;

daily ($n = 365$), then $A = \left(1 + \frac{1}{365}\right)^{365} \approx \$2.714$;

hourly ($n = 8{,}760$), then $A = \left(1 + \frac{1}{8{,}760}\right)^{8{,}760} \approx \$2.718$.

The problem of compound interest was systematically investigated by the Swiss mathematician Jacob Bernoulli (1655–1705), who observed that, in the above situation, as n increases (that is, as compounding intervals become smaller), $\left(1 + \frac{1}{n}\right)^n$ approaches a limit (the "force of interest"), specifically, it approaches an irrational number that is denoted by the letter e, in order to honor the Swiss mathematician Leonhard Euler. Notice that $e = \sum_{n=0}^{\infty} \frac{1}{n!} = \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \cdots \approx 2.718$, meaning that, with continuous compounding, the value of the aforementioned investment will reach approximately $\$2.718$. Euler proved that the number e is irrational by showing that its simple continued fraction expansion is infinite (by a "continued fraction," we mean an expression obtained through an iterative process of representing a number as the sum of its integral part and the reciprocal of another number, then writing this other number as the sum of its integral part and another reciprocal, etc.).

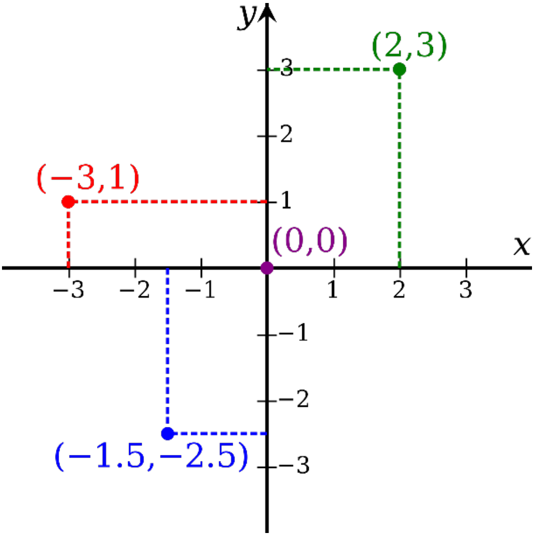
In case the base $b = e = \sum_{n=0}^{\infty} \frac{1}{n!} \approx 2.718$, which is known as "Euler's number," then $log_e a$ is written as lna, and it is said to be the "natural logarithm" of a. Notice that $log_e a = lna$ is called the "natural logarithm" because many processes can be described mathematically using it; such as: the rate at which your money will grow if you apply an interest rate continuously over a period of time; the population of a colony of rabbits that reproduce at a constant rate; the population of seeds in a sun flower; the decay rate of a radioactive isotope; etc.

Ordered Pairs and the Cartesian Product

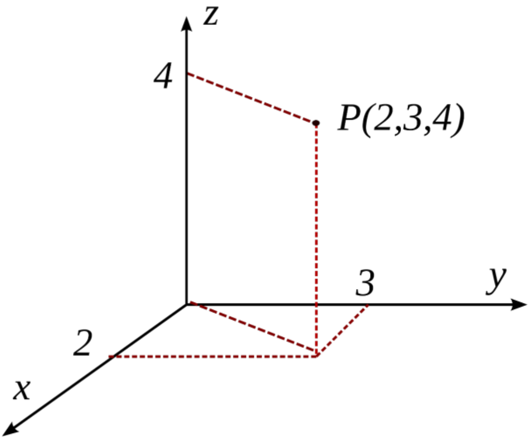
The Fundamental Property of Ordered Pairs: For any ordered pairs $(w,x) and\ (y,z)$, it holds that:
$(w,x)=(y,z)\Leftrightarrow w=y\ \&\ x=z$,
and, in this case, the two ordered pairs are called "equal."
The "Cartesian product" (also known as the "direct product") $A\times B$ of two sets A and B is the set of all ordered pairs (a,b) such that $a\in A\ and\ b\in B$; symbolically:
$A\times B=\{(a,b)|a\in A\ \&\ b\in B\}$.
For instance, if $A=\{1,2\}$ and $B=\{1,3\}$, then the Cartesian product $A\times B$ is the set $\{(1,1),(1,3),(2,1),(2,3)\}$. In general, the Cartesian product of the sets $A_1,A_2,\dots,A_n$, denoted by $A_1\times A_2\times\dots\times A_n$ is the set of all ordered n-tuples of the form $(a_1,a_2,\dots,a_n)$, where a_i is an element of $A_i(i=1,2,\dots,n)$.
Remark: It is easily checked that, for any sets A, B, and C, we have:
$A\times(B\cup C)=(A\times B)\cup(A\times C)$,
$A\times(B\cap C)=(A\times B)\cap(A\times C)$.
If $A=\emptyset$ or $B=\emptyset$, then $A\times B=\emptyset$.
$A\times B=B\times A\Leftrightarrow A=B$.
Let $A\times B=\{(a,b)|a\ \&\ b\ are\ real\ numbers\}$. Then $A\times B$ is the set of all points in a plane whose coordinates are (a,b). Thus, $A\times B$ is the Cartesian plane
$\mathbb{R}^2=\mathbb{R}\times\mathbb{R}$,
as shown, for instance, in Figure 2-3. In this case, each point P in the plane represents an ordered pair (a,b) of real numbers, and vice versa. In other words, the vertical line through P meets the x-axis at a, and the horizontal line through P meets the y-axis at b. Thus, we can understand the relationship between set theory, mathematical analysis, and geometry. In other words, a two-dimensional coordinate system consists of the horizontal axis (namely, the x-axis) and the vertical axis (namely, the y-axis), and the intersection of the two axes is the origin $O(0,0)$ of the coordinate system (by the term "axis," we mean a straight line with respect to which a body or structure is symmetrical). By analogy, we can define an n-dimensional coordinate system for any $n\geq 2$ ($n=2,3,4,5,\dots$), using n axes of reference at right angles to each other.

Figure 2-3: Cartesian coordinates.

A. The Cartesian plane $\mathbb{R}^2$ *(source: Wikimedia Commons: Author: K. Bolino; https://commons.wikimedia.org/wiki/File:Cartesian-coordinate-system_v2.svg).*



B. The Cartesian space $\mathbb{R}^3$ *(source: Wikimedia Commons: Author: Jhncls; https://commons.wikimedia.org/wiki/File:Coordinaten.svg).*



As noted above, the set $\mathbb{R}$ of all real numbers is called the real line, or the continuum. A set of pairs of real numbers is called a "number plane," and it is denoted by $\mathbb{R}^2$. As already mentioned, the set $\mathbb{R}$ can be represented geometrically as a horizontal number line. A geometric representation of the set $\mathbb{R}^2$ is the coordinate plane xOy, defined as two perpendicular number lines with a common origin O and the same scale (the number of units represented by a unit length along an axis is called the "scale"). The point $O(0,0)$ is called the "origin of coordinates." If P_0 is a point in the coordinate plane, then, by projecting it on the coordinate lines Ox and Oy, we find the coordinates of the projections x_0 and y_0 (notice: if you drop a perpendicular from a point to a line or plane, then the point you reach on that line or plane is called the projection of the point onto the line or plane). The coordinates are called, respectively, the "abscissa" (i.e., the x-coordinate) and the "ordinate" (i.e., the y-coordinate) of the point P_0, and the straight lines Ox and Oy are respectively called the "axis of abscissas" and the "axis of ordinates". Hence, to the point P_0 there corresponds one pair of numbers (x_0, y_0); conversely, given a pair of numbers (x_0, y_0), we mark the points x_0 and y_0 on the coordinate lines (axes) Ox and Oy, respectively, and, drawing through these points straight lines parallel to the coordinate lines (axes) Ox and Oy, we find the point of their intersection P_0. By analogy, we work in $\mathbb{R}^n$.

In general, the use of coordinate systems implies that space itself is encoded by n-tuples (i.e., by sequences, ordered lists, of n numbers), and, specifically, that the two-dimensional space, the "plane," is encoded by pairs of numbers, so that the conception of space becomes subordinate to the conception of arithmetic.

The "absolute value" (also known as the "modulus" or the "magnitude") of a real number x is denoted by $|x|$, and it is defined as follows:

$$|x| = \begin{cases} x \; if x \geq 0 \\ -x \; if x < 0 \end{cases}.$$

Therefore, the absolute value of any real number is always non-negative, and it may be thought of as that real number's distance from zero along the real line ("arithmetic continuum"). The aforementioned definition implies the following:

i. $|x|$ is the distance between the point x and zero (i.e., the "origin") on the real line. Hence, for instance, $|x| < 2$ means that the distance between x and the origin is less than 2, so that x lies between -2 and $+2$ on the real line, that is, $-2 < x < 2$. In general, whenever $|x| < a$, it holds that $-a < x < a$. Moreover, $\sqrt{x^2} = |x|$.

ii. $|x| = |-x|$ ("evenness," namely, "reflection symmetry" of the graph).

iii. $|x| \geq x$ and $|x| \geq -x$.

Notice that $|x| = |y|$ does not necessarily imply that $x = y$.

The absolute value of any real number has the following properties:

i. $|xy| = |x||y|$, and, generally,
$|x_1 x_2 \dots x_n| = |x_1||x_2| \dots |x_n|$.

ii. $|x + y| \leq |x| + |y|$ (triangle inequality: its geometric interpretation is that, for any triangle, the sum of the lengths of any two sides is greater than or equal to the length of the remaining side; equality only happens in the degenerate case when the sides are collinear and the triangle has zero area), and, generally,
$|x_1 + x_2 + \cdots + x_n| \leq |x_1| + |x_2| + \cdots + |x_n|$ (subadditivity).

iii. $|x - y| \geq |x| - |y|$.

iv. $|x| - |y| \leq \big||x| - |y|\big| \leq |x - y|$ (reverse triangle inequality).

v. $|x - y| < k \Rightarrow y - k < x < y + k$.

The concept of an absolute value was originally articulated by the French mathematician Jean-Robert Argand (1768–1822), who used the French term "module" (meaning "unit of measure"), which was borrowed into English as the Latin equivalent "modulus." The notation $|x|$ was introduced by the German mathematician Karl Weierstrass (1815–97).

Relations and Functions between Sets

Let A and B be two arbitrary sets. Then a "relation" between A and B, denoted by R, is defined to be a subset of the Cartesian product $A \times B$; symbolically: $R \subseteq A \times B$. The "domain" of relation R is defined by $D_R = \{a|(a,b) \in R\}$, and the "range" of relation R is defined by $R_R = \{b|(a,b) \in R\}$. If R is a relation from A to B, then the relation from B to A is called the "inverse" of R, and it is defined by $R^{-1} = \{(b,a)|(a,b) \in R\}$. A relational proposition is often denoted by aRb, where R relates a term a to a term b. Hence, a relation of two terms proceeds, somehow, from one to the other.

If R_1 is a relation from a set A to a set B, and if R_2 is a relation from B to a set C, then their "composition" denoted by $R_1 \circ R_2$ is a relation from A to C, symbolically:
$R_1 \circ R_2 = \{(a,c) \in A \times C | for\ some\ b \in B, (a,b) \in R_1\ \&\ (b,c) \in R_2\ with\ a \in A, c \in C\}$.

If R_1 and R_2 are relations such that $R_1 \subseteq R_2$, then R_2 is said to be an "extension" of R_1, and R_1 is said to be a "restriction" of R_2.

A relation R on a set A is "reflexive" if (a, a) is an element of R for every $a \in A$; it is "symmetric" if (a, b) is an element of R whenever (b, a) is an element of R; and it is "transitive" if (a, c) is an element of R whenever (a, b) and (b, c) are elements of R. A relation R on a set A is "antisymmetric" if, whenever a and b are distinct, then (a, b) is an element of R only if (b, a) is not an element of R. For instance, if $A = \{u, v, w\}$ and R is a relation on A, then:
$R = \{(u, v), (v, u), (u, u), (v, v), (v, w), (w, w)\}$ is a reflexive relation on A;
$R = \{(u, v), (v, u), (w, w)\}$ is a symmetric relation on A;
$R = \{(u, v), (v, w)(u, w), (v, v)\}$ is a transitive relation on A;
$R = \{(u, w), (v, v), (u, v), (u, u)\}$ is an antisymmetric relation on A.
Let A and B be two arbitrary sets. A relation $f \subseteq A \times B$ is called a "function," "mapping," or "transformation," denoted by $f: A \rightarrow B$, if it assigns to each element $a \in A$ exactly one element $b \in B$. The set A is called the "domain" of the function f and is denoted by D_f, while the set B is called the "codomain" of the function f. The set of all elements of B that are related to the elements of A via f is called the "range" (or "codomain") of the function f, and it is denoted by R_f, meaning that the range of f is the image of A by f:
$f(A) = \{f(a) | a \in A\}$.
Notice that the Axiom of Replacement, to which I referred earlier in this chapter, allows us to construct new sets from old ones by specifying a rule for generating the elements of the new set. Now, we shall state the Axiom of Choice.
The Axiom of Choice: Let $X = \{A_i, i \in I\}$ be a non-empty family of non-empty pairwise disjoint sets. Then there exists a set A consisting of exactly one element from each A_i. In other words, there exists a function f defined on X with the property that, for each $A_i \in X$, $f(A_i) \in A_i$; and the function f is then called a "choice function."
The acceptance of this axiom by mathematicians guarantees the existence of mathematical objects that are obtained by a series of choices. Thus, the Axiom of Choice can be viewed as an extension of a finite process (choosing objects from sets) to infinite settings. The Axiom of Choice was originally formulated in 1904 by the German logician and mathematician Ernst Zermelo in order to ensure that, whenever infinite sets play a role, the formulation of theorems is simple and relevant to the sets under consideration.
By the term "graph" of a function $f: A \rightarrow B$, we mean the set $\{x, f(x)\}$, where $x \in A$. In other words, the "graph" of a function $f(x)$ is the set of all points in a coordinate system that correspond to ordered pairs in $f(x)$.

Vertical line test: Imagine a vertical line sweeping across a graph. Assume that the vertical line at any position intersects the graph in more than one point. Then the graph is not the graph of a function.

If c is a positive constant, then:

i. The graph of $y = f(x) + c$ is the graph of f raised by c units.
ii. The graph of $y = f(x) - c$ is the graph of f lowered by c units.
iii. The graph of $y = f(x + c)$ is the graph of f shifted c units to the left. In fact, if we analyze the x-values, we can see a pattern, and we realize that the new x that we need in order to obtain $f(0)$ is the one that makes $f(x + c) = f(0)$, namely, $-c$. We can generalize this result as follows:
$f(x_{new} + c) = f(x) \Rightarrow x_{new} + c = x \Rightarrow x_{new} = x - c$,
meaning that the new x-values are the old x-values translated $-c$ units (that is, c units to the left, since that direction is the negative direction).
iv. The graph of $y = f(x - c)$ is the graph of f shifted c units to the right.

The graph of $y = -f(x)$ is the graph of f reflected about the x-axis.

If $c > 1$, then the graph of $y = cf(x)$ is the graph of f stretched by a factor of c. If $0 < c < 1$, then the graph of $y = cf(x)$ is the graph of f flattened out by a factor of c.

Operations with functions:

i. $(f \pm g)(x) = f(x) \pm g(x)$.
ii. $(f \cdot g)(x) = f(x) \cdot g(x)$.
iii. $\left(\frac{f}{g}\right)(x) = \frac{f(x)}{g(x)}, g(x) \neq 0$.
iv. Composite functions of functions f and g:
$(g \circ f)(x) = g\big(f(x)\big)$ and
$(f \circ g)(x) = f\big(g(x)\big)$.

Two functions $f: A \to B$ and $g: A \to B$ are called "equal" if $f(x) = g(x), \forall x \in A$, and they are called "different" if there is at least one $x_0 \in A$ such that $f(x_0) \neq g(x_0)$.

A function f is said to be "odd" if $f(-x) = -f(x)$ for every x in the domain of f. The graph of an odd function has symmetry about the origin. For instance, $y = x^3$ is an odd function. A function f is said to be "even" if $f(-x) = f(x)$ for every x in the domain of f. The graph of an even function has symmetry about the y-axis. For instance, $y = |x|$ is an even function.

A function $f: X \to Y$ is called "one-to-one" (or "injective," or an "injection," or a "monomorphism") if

$f(x_1) = f(x_2) \Rightarrow x_1 = x_2, \forall x_1, x_2 \in X$;

that is, a function is "one-to-one" if each x value in the domain is assigned a different y value, so that no two ordered pairs have the same second component. If more than one element of X has the same f-image in Y, then the function $f: X \to Y$ is said to be "many-to-one."

Horizontal line test: Imagine a horizontal line sweeping down the graph of a function. Assume that the horizontal line at any position intersects the graph in more than one point. Then the function is not one-to-one, and its inverse is not a function.

A function $f: X \to Y$ is called "into" if there exists at least one element of Y that is not the f-image of an element of X. In other words, for any into function $f: X \to Y$, the range set $f(X)$ is a proper subset of Y; symbolically, $f(X) \subset Y$.

If the range of a function f is the whole codomain of f, then f is said to be "onto" (or "surjective," or a "surjection," or an "epimorphism"). In other words, for any onto function $f: X \to Y$, $f(X) = Y$.

If a function is both one-to-one and onto, then it is called "bijective," or a "bijection," or an "one-to-one correspondence."

For instance:

i. If A is a subset of X, then the restriction to A of the identity mapping id_x, defined by $A \ni x \to x \in A$, is an injection j_A, called the "natural injection."
ii. The identity mapping of any set is bijective.
iii. The function $f: X \times Y \to Y \times X$ defined by $(x, y) \to (y, x)$, where $x \in X \, and \, y \in Y$, is bijective.
iv. The function $f(x) = x^2$, where $x \in \mathbb{R}$, is not injective, since $f(-x) = f(x) = x^2$, but the restriction to $\mathbb{R}^+$ (the set of all positive real numbers) of f is injective.
v. $f: \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^3$ is an one-to-one and onto mapping, that is, a bijection from $\mathbb{R}$ to $\mathbb{R}$.

Inverse functions: By the "inverse function" of a function f, we mean a function that undoes the operation of f, and it is denoted by f^{-1}. The inverse of f exists if and only if f is bijective. Given a function $f: X \to Y$, its inverse $f^{-1}: Y \to X$ assigns each element $y \in Y$ to the unique element $x \in X$ such that $f(x) = y$. In other words, two functions with exactly reverse assignments are said to be "inverse functions." Thus, a function $f: X \to Y$ is "invertible" if there exists a function $g: Y \to X$ such that $g(f(x)) = x$ for all $x \in X$ and $f(g(y)) = y$ for all $y \in Y$; and then the function g is called the inverse of f. Given an one-to-one function $y = f(x)$, you can find its inverse function as follows: (i) Interchange x and y in this equation. (ii) Solve the resulting equation for y, and then replace y with f^{-1}. (iii) Define the domain of f^{-1} to be equal to the range of f. For instance: if $f(x) = x + a$, then $f^{-1}(y) = y - a$; if $f(x) = a - x$, then $f^{-1}(y) = a - y$; if $f(x) = mx$, then $f^{-1}(y) = \frac{y}{m}$ provided that $m \neq 0$; if $f(x) = x^k$, then $f^{-1}(y) = \sqrt[k]{y} = y^{1/k}$ with $x, y \geq 0$ if k is even, and integer $k > 0$; and, if $f(x) = a^x$, then $f^{-1}(y) = log_a y$, where $y > 0$, and $a > 0$.

Sequences and Series

A “sequence” is a function whose domain is the set of positive integers (i.e., 1,2,3, ...). The functional values (i.e., the range elements) are called the terms of the sequence. In other words, a sequence is a set of numbers arranged in a definite order. Sequences test our logical skills and play an important role in the study of functions, spaces, and other mathematical structures, using the convergence properties of sequences.

Arithmetic progression: An “arithmetic progression” is a sequence of numbers in which each term after the first is found by adding a constant to the preceding term. This constant is called the “common difference” and is symbolized by d. Thus, the formula for the nth term in an arithmetic progression with first term a_1 and common difference d is:

$a_n = a_1 + (n - 1)d$.

Geometric progression: A “geometric progression” is a sequence of numbers in which each term after the first is found by multiplying the preceding term by a constant. This constant is called the “common ratio” and is symbolized by r. Thus, the formula for the nth term in a geometric progression with first term a_1 and common ratio r is:

$a_n = a_1 r^{n-1}$.

Arithmetic and geometric series: Associated with any sequence $a_1, a_2, a_3, \ldots$ is a “series”

$$a_1 + a_2 + a_3 + \cdots$$

which is the sum of all the terms in the sequence. A series that is associated with an arithmetic progression is called an “arithmetic series.” A series that is associated with a geometric progression is called a “geometric series.”

The sum of the first n terms of an arithmetic series is given as the following formula:

$S_n = \frac{n}{2}(a_1 + a_n) = \frac{n}{2}[2a_1 + (n - 1)d]$.

The sum of the first n terms of a geometric series is given as the following formula:

$S_n = \frac{a_1 - a_1 r^n}{1 - r} = \frac{a_1 - a_n r}{1 - r} = \frac{a_1(r^n - 1)}{r - 1}$.

Infinite sequences: Let us consider an infinite list of numbers: $a_1, a_2, a_3, \ldots$ This infinite list can be symbolized as follows:

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

(n goes from 1 to infinity); and this is what we call an "infinite sequence" (for simplicity and if there is no likelihood of confusion, we may symbolize a sequence by (a_n), or, sometimes, even simply by a_n). If the numbers in a sequence $\{a_n\}_{n=1}^{\infty}$ get arbitrarily close to some fixed number a, then we say that this sequence tends to the limit a, and we write $a_n \to a$ as $n \to \infty$, or $lim_{n\to\infty} a_n = a$.

For instance, $\left\{\frac{1}{n}\right\}_{n=1}^{\infty} = 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$ gets closer and closer to 0, and the sequence $\left\{1 + \frac{1}{2^n}\right\}_{n=1}^{\infty} = 1\frac{1}{2}, 1\frac{1}{4}, 1\frac{1}{8}, 1\frac{1}{16}, \ldots$ gets closer and closer to 1. However, not all sequences tend to a limit. For instance, the alternating sequence $\{(-1)^{n+1}\}_{n=1}^{\infty} = +1, -1, +1, -1, \ldots$ does not approach any particular number (it just bounces back and forth between $+1$ and -1), and the sequence $\{n\}_{n=1}^{\infty} = 1,2,3, \ldots$ doesn't tend to a limit at all.

Given an infinite sequence $\{a_n\}_{n=1}^{\infty}$, the statement that

$a_n \to a$ as $n \to \infty$

is equivalent to the statement that

$|a_n - a| \to 0$,

that is, $|a_n - a|$ gets arbitrarily close to 0. More formally, we can say the following: $a_n \to a$ as $n \to \infty$ if and only if, for every real number $\varepsilon > 0$, there exists a natural number n such that, for every natural number $m \geq n$, it holds that $|a_m - a| < \varepsilon$. This definition is absolutely crucial to calculus and, generally, "real analysis" (i.e., the analysis of the system of real numbers). This definition can be explained as follows: from some point ($m \geq n$) onward, all the members of the sequence $\{a_n\}_{n=1}^{\infty}$ are within a distance ε from a (symbolically, $|a_m - a| < \varepsilon$), and, since we can take the positive real number ε as small as we want, this condition means that a_n gets arbitrarily close to a as $n \to \infty$ (the ε-neighborhood of a can become arbitrarily small). For instance, using this formal definition, we can formally prove that the sequence $\left\{\frac{1}{n}\right\}_{n=1}^{\infty}$ gets arbitrarily close to 0 as follows: We have to prove that

$$(\forall\varepsilon > 0)(\exists n \in \mathbb{N})(\forall m \geq n)\left[\left|\frac{1}{m} - 0\right| < \varepsilon\right] \Leftrightarrow (\forall\varepsilon > 0)(\exists n \in \mathbb{N})(\forall m \geq n)\left[\frac{1}{m} < \varepsilon\right].$$

Let $\varepsilon > 0$ be given and arbitrary. Then we must find an $n \in \mathbb{N}$ such that, $\forall m \geq n$, $\frac{1}{m} < \varepsilon$. Let us choose any n such that $n > \frac{1}{\varepsilon}$. Then, if $m \geq n$,

$\frac{1}{m} \leq \frac{1}{n} < \varepsilon$, and this proves the statement that the sequence $\left\{\frac{1}{n}\right\}_{n=1}^{\infty}$ gets arbitrarily close to 0. Notice that the choice of n depended on ε, so that, the smaller the ε is, the bigger the n has to be, meaning that, for this sequence, the more we want $\frac{1}{m}$ to be close to 0, that is, the smaller the ε in the inequality $\left|\frac{1}{m} - 0\right| < \varepsilon$, the further out in the sequence we have to go, that is, the bigger the n is before we are within the required neighborhood of zero.

Let us consider another example. In terms of the above formal definition, we can prove that the sequence $\left\{\frac{n}{n+1}\right\}_{n=1}^{\infty} = \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \ldots$ tends to 1 as $n \to \infty$ as follows: This is intuitively obvious, but, working in the same way as before, we have to show that

$(\forall \varepsilon > 0)(\exists n \in \mathbb{N})(\forall m \geq n)\left[\left|\frac{m}{m+1} - 1\right| < \varepsilon\right]$.

Let $\varepsilon > 0$ be given and arbitrary. Then we must find an $n \in \mathbb{N}$ such that, $\forall m \geq n$, $\left|\frac{m}{m+1} - 1\right| < \varepsilon$. Let us choose any n such that $n > \frac{1}{\varepsilon}$. Then, if $m \geq n$, $\left|\frac{m}{m+1} - 1\right| = \left|\frac{-1}{m+1}\right| = \frac{1}{m+1} < \frac{1}{m} \leq \frac{1}{n} < \varepsilon$, and this proves that the sequence $\left\{\frac{n}{n+1}\right\}_{n=1}^{\infty}$ gets arbitrarily close to 1.

Cauchy sequence: A sequence $\{a_n\}_{n=1}^{\infty}$ is said to be a “Cauchy sequence” if and only if, for every real number $\varepsilon > 0$, there exists a natural number m such that, for every natural number $n \geq m$ and for every natural number k, it holds that $|a_{n+k} - a_n| < \varepsilon$. The intuition behind this definition is that the terms of a Cauchy sequence become arbitrarily close to each other (i.e., the distance ε between two terms of this sequence becomes arbitrarily small) as the sequence progresses. For instance, using the above definition, we can show that the sequence $\left\{\frac{1}{2^n}\right\}_{n=1}^{\infty}$ is a Cauchy sequence as follows: If $a_n = \frac{1}{2^n}$, then $|a_{n+k} - a_n| = \left|\frac{1}{2^{n+k}} - \frac{1}{2^n}\right| = \frac{1}{2^n}\left|\frac{1}{2^k} - 1\right| = \frac{1}{2^n}\left(1 - \frac{1}{2^k}\right) < \varepsilon$, where $k \in \mathbb{N}$, and $\varepsilon > 0$, for $\frac{1}{2^n} < \varepsilon$, so that $2^n > \frac{1}{\varepsilon} \Rightarrow n > \frac{ln\frac{1}{\varepsilon}}{ln2}$. Let $m > \frac{ln\frac{1}{\varepsilon}}{ln2}$. Then m is a natural number such that $|a_{n+k} - a_n| < \varepsilon$ for all $n \geq m$, and $k \in \mathbb{N}$.

Subsequences: Let (a_n) be a sequence, and let $n_1, n_2, \ldots, n_i, \ldots$ with $n_{i+1} > n_i$ and $i = 1,2,3, \ldots$ be a set of positive integers. Then

$$(a_{n_i})$$

is said to be a “subsequence” of (a_n), and, if $a_{n_i} \to l'$ as $i \to \infty$, then l' is said to be a “subsequential limit” of (a_n). For instance, the sequences (a_{2n-1}), (a_{2n}), (a_n^2), and (a_n^3) are subsequences of (a_n).

Theorem 1: Every accumulation point of a subsequence of a sequence is also an accumulation point of the sequence.

Proof: Let (a_{n_i}) be a subsequence of the sequence (a_n), $n_{i+1} > n_i$ and $i = 1,2,3,\dots$ Suppose that p is an accumulation point of (a_{n_i}). Then, by definition, $a_{n_i} \in (p - \varepsilon, p + \varepsilon) \Rightarrow a_n \in (p - \varepsilon, p + \varepsilon)$, $\varepsilon > 0$. Thus, p is an accumulation point of the sequence (a_n).

Remark: The converse is not necessarily true. For instance, consider the subsequence $\{1,2,3,4,\dots\}$ of the sequence $\{1,1,1,2,1,3,1,\dots\}$. In this case, 1 is an accumulation point of the sequence, but the subsequence has no accumulation point.

Theorem 2: Let (a_n) be an arbitrary bounded sequence, meaning that its range $R = \{a_n | n \in \mathbb{N}\}$ is a bounded set. Then (a_n) contains a convergent subsequence.

Proof: If R is a finite set, then there exists at least one element $r \in R$ such that $r_{n_i} = r$ for all $i = 1,2,3,\dots$, where n_i are positive integers and $n_{i+1} > n_i$. Hence, (a_n) contains a subsequence (r_{n_i}) that converges to r.

Now, suppose that R is infinite. Then, by the Bolzano–Weierstrass Theorem (proved in Chapter 2), R has at least one accumulation point, say p. Let us take an element $p_1 \in R$ such that $0 < |p_1 - p| < 1$, and then let us take an element $p_2 \in R$ such that

$$p_2 \neq p_1, \text{ and } 0 < |p_2 - p| < \frac{1}{2}. \tag{1}$$

Similarly, let $p_{n+1} \in R$ such that $p_{n+1} \neq p_i$, where $i = 1,2,3,\dots,n$, and

$$0 < |p_{n+1} - p| < \frac{1}{n+1}. \tag{2}$$

Therefore, by induction, for all n, there exist $p_n \in R$ such that, for all n, all p_n’s are distinct and

$$0 < |p_n - p| < \frac{1}{n}.$$

Let m be the least integer such that $m \geq \frac{1}{\varepsilon}$ for any $\varepsilon > 0$, so that

$$\varepsilon \geq \frac{1}{n} \text{ for all } n \geq m. \tag{3}$$

From (2) and (3), we obtain

$$|p_n - p| < \varepsilon \text{ for all } n \geq m,$$

meaning that $p_n \to p$ as $n \to \infty$. Thus, (a_n) has the so defined convergent subsequence (p_n), *quod erat demonstrandum*.

Cauchy’s General Principle of Convergence: A sequence (a_n) converges if and only if it is a Cauchy sequence.

Proof: Firstly, suppose that (a_n) converges to l. Then, $\forall\varepsilon > 0, \exists m \in \mathbb{N}$ such that
$|a_n - l| < \frac{\varepsilon}{2}, \forall n \geq m.$
Let us consider $|a_{n+k} - a_n| = |a_{n+k} - l - (a_n - l)| \leq |a_{n+k} - l| + |a_n - l| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \Rightarrow |a_{n+k} - a_n| < \varepsilon$, $\forall n \geq m$, meaning that (a_n) is a Cauchy sequence.

Now, we shall prove the converse. Suppose that (a_n) is a Cauchy sequence, that is, $\forall\varepsilon > 0, \exists m \in \mathbb{N}$ such that, $\forall n \geq m$ and $k \in \mathbb{N}$, it holds that

$$|a_{n+k} - a_n| < \varepsilon. \qquad (1)$$

Then

$$a_m - \varepsilon < a_n < a_m + \varepsilon, \forall n \geq m. \qquad (2)$$

If $b = max\{|a_1|, |a_2|, \dots, |a_m|\}$, then, due to (2), for all n, it holds that
$|a_n| < b + \varepsilon,$
meaning that (a_n) is a bounded sequence. Therefore, (a_n) has a convergent subsequence. Let (a_{n_i}) be a convergent subsequence that converges to a. Then, $\forall\varepsilon > 0, \exists m_1 \in \mathbb{N}$ such that, $\forall n_i \geq m_1$, it holds that
$|a_{n_i} - a| < \varepsilon.$
Moreover, due to (1),
$|a_n - a_p| < \varepsilon, \forall n \geq m$ and $\forall p \geq m,$
so that, for $m_2 = max\{m, m_1\}$, we obtain
$|a_n - a| = |a_n - a_{n_i} + a_{n_i} - a| \leq |a_n - a_{n_i}| + |a_{n_i} - a| < 2\varepsilon$, $\forall n \geq m_2$, meaning that a_n converges to a, that is, (a_n) is convergent, *quod erat demonstrandum*.

Monotone sequences: A sequence (a_n) is called "increasing" if $a_m \leq a_n$ for all $m < n \in \mathbb{N}$; and it is called "strictly increasing" if $a_m < a_n$ for all $m < n \in \mathbb{N}$. By analogy, a sequence (a_n) is called "decreasing" if $a_m \geq a_n$ for all $m < n \in \mathbb{N}$; and it is called "strictly decreasing" if $a_m > a_n$ for all $m < n \in \mathbb{N}$. A sequence that is increasing or decreasing is called "monotone."

Completeness Axiom of the Real Numbers: If (a_n) is any monotone and bounded sequence in $\mathbb{R}$, then (a_n) converges.

Without loss of generality, suppose that (a_n) is increasing and bounded from above (we work similarly in case it is decreasing and bounded from below). Then the set $A = \{a_n | n \in \mathbb{N}\}$ has a supremum $sup(A) = s$, so that, $\forall\varepsilon > 0, \exists p \in A$ such that $s - \varepsilon \leq a_p \leq s$. Because, by hypothesis, (a_n) is increasing, it also holds that, $\forall\varepsilon > 0, \exists p \in \mathbb{N}$ such that, $\forall n \geq p$, it holds that $s - \varepsilon \leq a_p \leq a_n \leq s$, meaning that $a_n \to s$ as $n \to \infty$.

Infinite series: By a (real) sequence, we mean a function $f: \mathbb{N} \rightarrow \mathbb{R}$ whose images are $a_1, a_2, a_3, \dots, a_n, \dots$ Let us consider a sequence of real numbers $a_n, n \in \mathbb{N}$. From this sequence, we can create a new sequence $s_n, n \in \mathbb{N}$, as follows:

$s_1 = a_1$
$s_2 = a_1 + a_2$
$s_3 = a_1 + a_2 + a_3$
$\vdots$
$s_n = a_1 + a_2 + a_3 + \cdots + a_n = \sum_{k=1}^{n} a_k$.

This sequence $s_n, n \in \mathbb{N}$, whose general term is $s_n = \sum_{k=1}^{n} a_k$, is said to be the "sequence of the partial sums" of the sequence $a_n, n \in \mathbb{N}$. The real numbers $s_1, s_2, s_3, \dots, s_n$ are, respectively, called the first partial sum, the second partial sum, . . ., the nth partial sum.

A "series" of real numbers, symbolized by $\sum_{n=1}^{\infty} a_n = a_1 + a_2 + a_3 + \cdots + a_n + \cdots$, is defined to be the ordered pair (a_n, s_n) where $a_n, n \in \mathbb{N}$, is a sequence of real numbers, and $s_n = a_1 + a_2 + a_3 + \cdots + a_n$, $n \in \mathbb{N}$. Each term of the sequence a_n , $n \in \mathbb{N}$, is called a "term" of the corresponding series, and each term of the sequence $s_n, n \in \mathbb{N}$, is called a "partial sum" of the series $\sum_{n=1}^{\infty} a_n$.

The founders of the modern theory of infinite series are Isaac Newton and James Gregory in the seventeenth century, and the Bernoulli family mathematicians (Jacob, John, Nicolaus, and Daniel), Leonhard Euler, and Joseph-Louis Lagrange in the eighteenth century. In fact, the eighteenth-century mathematicians were thinking of infinite series as infinite polynomials (mathematical expressions consisting of variables, coefficients, and the operations of addition, subtraction, multiplication, and non-negative integral exponents), and they tried to develop an arithmetic system of infinite polynomials.

The basic idea in the study of infinite series is that an infinite summation of numbers can have a finite sum. Some of the early work on series was motivated by paradoxes related to the concept of infinity, with which many ancient Greek mathematicians were preoccupied. In the fifth century B.C.E., the Greek mathematician and philosopher Zeno posed the following paradox: Consider a race between the legendary Greek hero Achilles and a tortoise over 100 meters. Suppose that the tortoise starts 80 meters ahead, and Achilles can run 10 times as fast as the tortoise. Then, after 10 sec., when Achilles will have run 80 meters, reaching the point where the tortoise started, the tortoise will have run only 8 meters farther. Then it will take Achilles 1 sec. more to cover that distance, but, during the same time, the tortoise will have run 0.8 meters farther. Then it will take Achilles 0.1 sec. to reach this third point, while the tortoise moves ahead by 0.08 meters, etc. Thus, whenever Achilles reaches somewhere the tortoise has been, the tortoise is still ahead, and it seems that the tortoise will stay ahead. In fact, Zeno's paradox can be resolved as follows: the total time that it would take Achilles to catch up, in seconds, is $10 + 1 + 0.1 + 0.01 + 0.001 + \cdots$, which is an infinite series. But this infinite series is equal to 11.111…, which is a finite number. In particular, let $x = 0.1 + 0.01 + 0.001 + \cdots$. In fact, $0.1 + \frac{x}{10} = 0.1 + 0.01 + 0.001 + \cdots$, and, therefore, $x = 0.1 + \frac{x}{10} \Rightarrow 10x = 1 + x \Rightarrow 9x = 1 \Rightarrow x = \frac{1}{9}$. Hence, the time for Achilles to catch up is $11\frac{1}{9}$ sec.

A series $\sum_{n=1}^{\infty} a_n$ is said to "converge" to a number l, and we write $\sum_{n=1}^{\infty} a_n = l$, if and only if $lim_{n\to\infty} s_n = l$, where s_n is the nth partial sum of the corresponding sequence a_n, $n \in \mathbb{N}$. For instance, consider the infinite geometric series $\sum_{n=1}^{\infty} \frac{1}{2^n} = \frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2^n} + \cdots$ The nth partial sum of this sequence is

$$s_n = \frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2^n} = \frac{\frac{1}{2}\left[\left(\frac{1}{2}\right)^n - 1\right]}{\frac{1}{2} - 1} = 1 - \frac{1}{2^n}$$

(and it is intuitively obvious that, as n increases, $\frac{1}{2^n}$ decreases, and the difference $1 - \frac{1}{2^n}$ approaces 1). Given that $lim_{n\to\infty} s_n = 1$, we conclude that $\sum_{n=1}^{\infty} \frac{1}{2^n} = 1$.

A series $\sum_{n=1}^{\infty} a_n$ is said to "diverge" to $\pm\infty$ if and only if $lim_{n\to\infty} s_n = \pm\infty$, respectively.

A series $\sum_{n=1}^{\infty} a_n$ is said to "diverge," or to be an "alternating series," if and only if $lim_{n\to\infty} s_n$ does not exist.

For instance, consider the series $\sum_{n=1}^{\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} + \cdots$, which is known as the "harmonic series," because $\frac{2}{a_n} = \frac{1}{a_{n-1}} + \frac{1}{a_{n+1}}$ $\forall n \geq 2$. We can prove that the harmonic series diverges to $+\infty$ as follows:

Instead of considering all the partial sums s_n (where $n = 1,2,3, \ldots$), let us look to the following sequence of partials sums (each time, we consider the sum of 2^n terms, where $n = 1,2,3, \ldots$):

$s_2 = 1 + \frac{1}{2}$,

$s_4 = 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) > 1 + \frac{1}{2} + \left(\frac{1}{4} + \frac{1}{4}\right) = 1 + \frac{2}{2}$,

$s_8 = 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) + \left(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}\right) > 1 + \frac{1}{2} + \left(\frac{1}{4} + \frac{1}{4}\right) + \left(\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}\right) = 1 + \frac{3}{2}$, and, in general, we observe the following pattern: when we take the sum of 2^n terms, the corresponding partial sum is bigger than $1+\frac{n}{2}$; symbolically, the general pattern of this sequence of partial sums can be captured by the inequality

$$s_{2^n} > 1+\frac{n}{2}$$

and, thus, $lim_{n\to\infty}s_{2^n} > lim_{n\to\infty}\left(1+\frac{n}{2}\right)$. The right-hand side diverges to infinity, and, therefore, the limit of s_{2^n} also diverges to infinity, and, then, the whole s_n diverges to infinity (even if just some of the partial sums diverge to infinity, the whole series diverges to infinity). Notice that the sequence $\left\{\frac{1}{n}\right\}_{n=1}^{\infty}$ converges to 0, but the series $\sum_{n=1}^{\infty}\frac{1}{n}$ diverges to $+\infty$. In fact, in the case of sequences, "convergence" refers to the behavior of the terms of a sequence, whereas, in the case of series, "convergence" refers to the behavior of the sum of the terms.

Operations with series:

1. Addition: $\sum_{n=1}^{\infty} a_n + \sum_{n=1}^{\infty} b_n = \sum_{n=1}^{\infty}(a_n + b_n)$.
2. Subtraction: $\sum_{n=1}^{\infty} a_n - \sum_{n=1}^{\infty} b_n = \sum_{n=1}^{\infty}(a_n - b_n)$.
3. Scalar multiplication: $c\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{\infty} ca_n$, where $c \in \mathbb{R}$.

Basic propositions regarding the convergence of series:

1. If $\sum_{n=1}^{\infty} a_n$ converges to a number, then the sequence s_n, $n \in \mathbb{N}$, is bounded.
2. If the sequence s_n, $n \in \mathbb{N}$, is not bounded, then $\sum_{n=1}^{\infty} a_n$ does not converge to a number.
3. If $\sum_{n=1}^{\infty} a_n$ converges to a number, then $lim_{n\to\infty}a_n = 0$ (but not conversely, as, for instance, the case of the harmonic series indicates).
4. The series $\sum_{n=1}^{\infty} a_n$ and $\sum_{n=1}^{\infty} a_{n+k}$ have the same behavior with regard to convergence.
5. If $\sum_{n=1}^{\infty} a_n$ with $a_n > 0\ \forall n \in \mathbb{N}$, and if the sequence s_n, $n \in \mathbb{N}$, is not bounded, then $\sum_{n=1}^{\infty} a_n = +\infty$.
6. If $\sum_{n=1}^{\infty} a_n$ with $a_n < 0\ \forall n \in \mathbb{N}$, and if the sequence s_n, $n \in \mathbb{N}$, is not bounded, then $\sum_{n=1}^{\infty} a_n = -\infty$.

Comparison Test: Suppose that we have two series $\sum_{n=1}^{\infty} a_n$ and $\sum_{n=1}^{\infty} b_n$ such that $a_n > 0$ and $b_n > 0$ for all $n \in \mathbb{N}$. If $a_n \leq b_n$, then:

i. If $\sum_{n=1}^{\infty} b_n$ is convergent, then so is $\sum_{n=1}^{\infty} a_n$.
ii. If $\sum_{n=1}^{\infty} a_n$ diverges to $+\infty$, then so does $\sum_{n=1}^{\infty} b_n$.

Proof: If $s_n = \sum_{k=1}^{n} a_k$ and $t_n = \sum_{k=1}^{n} b_k$ are the terms of the sequences of partial sums, then, since they are summations of finitely many positive terms, it is straightforward that $0 < s_n \leq t_n < \infty$, as well as that the sequences (s_n) and (t_n) are both increasing.

Suppose that $\sum_{n=1}^{\infty} b_n$ is convergent, and that $\sum_{n=1}^{\infty} b_n = b$. Then $lim_{n\to\infty} t_n = b$. Hence, the sequence (s_n) is increasing and bounded from above by b. Therefore, as I have already shown in the study of sequences, (s_n) converges, and so does $\sum_{n=1}^{\infty} a_n$.

Now, suppose that $\sum_{n=1}^{\infty} a_n$ diverges to $+\infty$. Then $lim_{n\to\infty} s_n = +\infty$. Hence, $lim_{n\to\infty} t_n = +\infty$, which implies that $\sum_{n=1}^{\infty} b_n$ diverges to $+\infty$, and the proof of the Comparison Test is complete.

The Limit Comparison Test: Suppose that we have two series $\sum_{n=1}^{\infty} a_n$ and $\sum_{n=1}^{\infty} b_n$ such that $a_n > 0$ and $b_n > 0$ for all $n \in \mathbb{N}$. If $lim_{n\to\infty} \frac{a_n}{b_n} = c$ for some positive real number c (i.e., $c > 0$, and $c < \infty$), then either both series converge, or both series diverge to $+\infty$, or both series are alternating series.

Proof: Because $0 < c < \infty$, we can find two positive (finite) numbers m and M such that $m < c < M$. Given that $lim_{n\to\infty} \frac{a_n}{b_n} = c$, the definition of the limit of a sequence implies that, for a sufficiently large n, the quotient $\frac{a_n}{b_n}$ must be close to c, and, therefore, there must exist a positive integer N such that, if $n > N$, it holds that

$m < \frac{a_n}{b_n} < M.$

Multiplying through by b_n, we obtain

$b_n m < a_n < b_n M$, provided that $n > N$.

Hence, if $\sum_{n=1}^{\infty} b_n$ diverges, then so does $\sum_{n=1}^{\infty} mb_n$, and, since $b_n m < a_n$ for all sufficiently large n, then the Comparison Test implies that $\sum_{n=1}^{\infty} a_n$ also diverges. Similarly, if $\sum_{n=1}^{\infty} b_n$ converges, then so does $\sum_{n=1}^{\infty} Mb_n$, and, since $a_n < b_n M$ for all sufficiently large n, then the Comparison Test implies that $\sum_{n=1}^{\infty} a_n$ also converges, and the proof of the Limit Comparison Test is complete.

Cauchy Criterion: A series $\sum_{n=1}^{\infty} a_n$ converges if and only if the sequence (s_n) of its partial sums is a Cauchy sequence (it is the same as the Cauchy criterion for sequences).

Fibonacci sequence: The Fibonacci sequence (named after the medieval Italian mathematician Fibonacci) is a sequence in which each number is the sum of the two preceding ones. Thus, starting from 0 and 1, the Fibonacci sequence begins

$$0,1,1,2,3,5,8,13,21,34,55,89,144,233,\ldots$$

The Fibonacci numbers may be defined by the following recurrence relation:

$F_0 = 0, F_1 = 1$, and $F_n = F_{n-1} + F_{n-2}, \forall n > 1$.

Applications of Fibonacci numbers include computer algorithms (such as the Fibonacci search technique, which narrows down possible locations with the aid of Fibonacci numbers), and Fibonacci numbers appear in several biological settings (such as branching in trees, the arrangement of leaves on a stem, the fruit sprouts of a pineapple, the arrangement of a pine cone's bracts, the flowering of an artichoke, etc.). The Fibonacci sequence diverges to infinity, since, starting with $n = 6$, we see that $F_n > n$; and, therefore, given any number $M > 0$, $F_n > M$ for all $n > max\{\lceil M \rceil, 6\}$, where $\lceil x \rceil$ denotes the "ceiling function," which maps x to the smallest integer greater than or equal to x.

The Cardinality of Number Sets

Two sets A and B are "equinumerous" or "have the same cardinality" if their elements can be correlated one-to-one in such a way that each element of either corresponds to exactly one of the other, namely, if there exists a bijection from A to B; then we write $A =_c B$. A set A is countable if and only if either $A = \emptyset$ or A accepts an "enumeration," namely, there exists an onto function $\varepsilon: \mathbb{N} \to A$ such that
$A = \{\varepsilon(0), \varepsilon(1), \varepsilon(2), \dots\}$.
We can prove the theorem that a union of a countable collection of countable sets $A = \cup_n A_n$, where $n \in I \subseteq \mathbb{N}$, is a countable set as follows: Assume that I is infinite (if I is finite, then we work analogously), so that I can be replaced by $\mathbb{N}$. Then the given countable collection of countable sets may be designated by

$$A = \cup_{n=0}^{\infty} A_n = A_0 \cup A_1 \cup A_2 \cup \dots$$

Without loss of generality, assume that each A_n is non-empty. Then we can find an enumeration $\varepsilon^n: \mathbb{N} \to A_n$ for each A_n . Setting
$a_i^n = \varepsilon^n(i)$,
we obtain
$A_n = \{a_0^n, a_1^n, \dots\}$,
and we can construct a table containing every element of A as follows:

$$\begin{gathered} A_0: a_0^0 a_1^0 a_2^0 \dots \\ A_1: a_0^1 a_1^1\, a_2^1 \dots \\ A_2: a_0^2 a_1^2 a_2^2 \dots \\ \vdots \end{gathered}$$

Therefore, collecting the aforementioned elements diagonally, we obtain $A = \{a_0^0, a_0^1, a_1^0, a_2^0, a_1^1, \dots\}$, which proves the theorem.
We can prove the theorem that, if the sets $A_1, A_2, A_3, \dots, A_n$ are countable, then their Cartesian product $A_1 \times A_2 \times \dots \times A_n$ is a countable set as follows: By definition, if $A_i, i = 1,2, \dots, n,$ is empty, then the

corresponding Cartesian product is empty. Otherwise, for two sets A and B, we have the enumeration of B given by
$B = \{b_0, b_1, b_2, \dots\}$,
so that
$A \times B = \cup_{n=0}^{\infty} (A \times \{b_n\})$,
and each $A \times \{b_n\}$ is equinumerous to A (and, therefore, countable) by the correspondence $x \to (x, b_n)$, which proves the theorem.
As I have already mentioned, the set $\mathbb{N}$ of natural numbers is countable. The set $\mathbb{Z}$ of integers is also countable, since we can define a bijection from $\mathbb{Z}$ to $\mathbb{N}$ as follows: send 0 to 1, send negative integers to odd natural numbers, and send positive integers to even natural numbers. Moreover, the set $\mathbb{Q}$ of rational numbers is countable. In fact, notice that $\mathbb{Q}$ is a set of tuples of integers, since every rational number is of the form a/b where a and b are integers, and the set of tuples of integers is countable. However, one may ask if there exists a rational number for every tuple of integers (a and b must be coprime). The answer is that, if A is a subset of B, and if B is countable, then so is A, and, in this case, $\mathbb{Q}$ is a subset of the set of the tuples of integers. However, the set $\mathbb{Q}^{\sim}$ of irrational numbers is uncountable (i.e., it contains too many elements to be countable): it suffices to consider an irrational number, such as $\sqrt{2}$, and think that all the infinitely many products of $\sqrt{2}$ by all the infinitely many rational numbers are irrational numbers. Even though both $\mathbb{Q}$ and $\mathbb{Q}^{\sim}$ are infinite sets, the set $\mathbb{Q}^{\sim}$ is much larger than the set $\mathbb{Q}$. Given that $\mathbb{R} = \mathbb{Q} \cup \mathbb{Q}^{\sim}$ and $\mathbb{Q}^{\sim}$ is uncountable, the set $\mathbb{R}$ of real numbers is uncountable.

Real Equations and Algebra

By the term “equation,” we mean a statement that two quantities are equal. For instance, $1{,}000m = 1km$. More often, an equation contains an unknown quantity that is represented by a symbol, and we try to find the value of this unknown quantity. By the term “algebra,” we refer to methods and techniques for solving equations. In fact, the core of the study of structures in mathematics consists of taking numbers and putting them into equations in the form of “variables”; and the rules for manipulating these equations are contained in algebra. Moreover, in the context of algebra, we study multidimensional numbers, such as matrices and vectors (see chapters 3 and 7).
The word “algebra” derives from the Arabic word “al-Jabr,” meaning “transformation.” It refers to a methodology developed by the Persian mathematician Al-Khwarizmi, who lived in Baghdad early in the Islamic era. Al-Khwarizmi was interested in solving algebraic equations, and his

method consists in applying a transformation to the given equation in order to put it into a standard form for which the solution method is known.

Equations requiring multiplication and division:

i. We can solve the equation $\frac{x}{12} = 4$ as follows: multiplying each side by 12, we get $\frac{x}{12} \times 12 = 4 \times 12 \Rightarrow x = 48$. Check: when $x = 48$, the left-hand side of the given equation becomes $\frac{48}{12} = 4$. The right-hand side of the given equation is equal to 4. Therefore, the solution is correct.

ii. We can solve the equation $6x = 3$ as follows: dividing each side by 6, we get $\frac{6x}{6} = \frac{3}{6} \Rightarrow x = \frac{1}{2}$. Check: when $x = \frac{1}{2}$, the left-hand side of the given equation becomes $6 \times \frac{1}{2} = 3$. The right-hand side of the given equation is equal to 3. Therefore, the solution is correct.

Equations requiring addition and subtraction:

i. We can solve the equation $x - 2 = 4$ as follows: adding 2 to each side, we get $x - 2 + 2 = 4 + 2 \Rightarrow x = 6$. The operation of adding 2 to each side is the same as transferring -2 to the right-hand side, but, in so doing, the sign is changed from a minus to a plus. Hence, $x - 2 = 4 \Leftrightarrow x = 4 + 2 \Leftrightarrow x = 6$. Check: when $x = 6$, the left-hand side of the given equation becomes $6 - 2 = 4$. The right-hand side of the given equation is equal to 4. Therefore, the solution is correct.

ii. We can solve the equation $x + 18 = 30$ as follows: subtracting 18 from each side, we get $x + 18 - 18 = 30 - 18 \Rightarrow x = 12$. Alternatively, moving $+18$ to the right-hand side (changing the sign from a plus to a minus), we get $x = 30 - 18 \Leftrightarrow x = 12$. Check: when $x = 12$, the left-hand side of the given equation becomes $12 + 18 = 30$. The right-hand side of the given equation is 30. Therefore, the solution is correct.

Equations containing the unknown quantity on both sides: In equations of this kind, we group all the terms containing the unknown quantity on one side of the equation and the remaining terms on the other side.

i. We can solve the equation $4x + 3 = 6x + 11$ as follows: transferring $6x$ to the left-hand side and $+3$ to the right-hand side, we get $4x - 6x = 11 - 3 \Rightarrow -2x = 8 \Rightarrow x = -\frac{8}{2} = -4$. Check: when $x = -4$, the left-hand side becomes $4(-4) + 3 = -13$, and the right-hand side becomes $6(-4)+11=-13$. Therefore, the solution is correct.

ii. We can solve the equation $7x-2=5x+8$ as follows: $7x-5x=8+2 \Rightarrow 2x=10 \Rightarrow x=5$. Check: when $x=5$, the left-hand side becomes $7\times 5-2=33$, and the right-hand side becomes $5\times 5+8=33$. Therefore, the solution is correct.

Equations containing brackets: When an equation contains brackets, we remove these first, and then we solve according to the aforementioned methods. For instance, $3(2x-1)=9 \Rightarrow 6x-3=9 \Rightarrow 6x=12 \Rightarrow x=2$. Check: when $x=2$, the left-hand side is $3(2\times 2-1)=9$, and the right-hand side is 9. Therefore, the solution is correct.

Equations containing fractions: When an equation contains fractions, we multiply each term of the equation by the least common multiple of the denominators. For instance, we can solve the equation $\frac{x}{3}+\frac{2}{5}=\frac{5x}{2}-1$ as follows: The least common multiple of the denominators 3, 5, and 2 is 30. Multiplying each term by 30 gives $\frac{x}{3}\times 30+\frac{2}{5}\times 30=\frac{5x}{2}\times 30-1\times 30 \Rightarrow 10x+12=75x-30 \Rightarrow -65x=-42 \Rightarrow x=\frac{42}{65}$. The solution may be verified by the check method shown in the previous examples.

Simultaneous equations: Consider the two following equations:

$$\begin{Bmatrix} ax+by=c \\ px+qy=r \end{Bmatrix}.$$

Each equation contains the unknown quantities x and y. The solutions to the equations are the values of x and y that satisfy both equations. Equations such as these are called “simultaneous equations” (or a “system of equations”).

i. We can solve the simultaneous equations

$$4x+5y=14 \qquad (*)$$

$$x+2y=11 \qquad (**)$$

as follows: If we multiply equation $(**)$ by 4, we shall have the same coefficient of x in both equations:

$$4x+8y=44 \qquad (***)$$

We can now eliminate x by subtracting equation $(*)$ from equation $(***)$:

$$4x+8y=44$$

$$4x+5y=14$$

$$3y=30$$

Hence, $y=10$. In order to find x, we substitute $y=10$ in either of the original equations. Therefore, substituting for

$y = 10$ in equation $(*)$, we get $4x + 5 \times 10 = 14 \Rightarrow x = -9$. In order to check these values, it suffices to substitute them in equation $(**)$.

ii. We can solve the simultaneous equations

$$5x + 7y = 15 \qquad (*)$$

$$4x + \frac{8}{5}y = 24 \qquad (**)$$

as follows: the same coefficient of x can be obtained in both equations if equation $(*)$ is multiplied by 4 (the coefficient of x in equation $(**)$) and equation $(**)$ is multiplied by 5 (the coefficient of x in equation $(*)$). Multiplying equation $(*)$ by 4, we get

$$20x + 28y = 60 \qquad (***)$$

Multiplying equation $(**)$ by 5, we get

$$20x + 8y = 120 \qquad (****)$$

Subtracting equation $(***)$ from equation $(****)$, we get $-20y = 60 \Rightarrow y = -3$.

Substituting for $y = -3$ in equation $(*)$, we get $x = \frac{36}{5}$. In order to check these values, it suffices to substitute them in equation $(**)$.

iii. We can solve the simultaneous equations

$$7x + 4y = 20 \qquad (*)$$

$$3x - 2y = 3 \qquad (**)$$

as follows: in this system of equations, it is easier to eliminate y, since the same coefficient of y can be obtained in both equations by multiplying equation $(**)$ by 2. In fact, multiplying equation $(**)$ by 2, we get

$$6x - 4y = 6 \qquad (***)$$

Adding equations $(*)$ and $(***)$, we get $13x = 26 \Rightarrow x = 2$. Substituting for $x = 2$ in equation $(*)$, we get $y = \frac{3}{2}$. In order to check these values, it suffices to substitute them in equation $(**)$.

iv. We can solve the simultaneous equations

$$\frac{x}{5} - \frac{y}{3} = \frac{1}{10} \qquad (*)$$

$$\frac{3x}{4} - \frac{2y}{3} = \frac{2}{3} \qquad (**)$$

as follows: first, we shall clear each equation of fractions. In equation $(*)$, the least common multiple of the denominators is 30. Hence, by multiplying equation $(*)$ by 30, we get

$$6x - 10y = 3 \qquad (***)$$

In equation $(**)$, the least common multiple of the denominators is12. Hence, by multiplying equation $(**)$ by 12, we get

$9x - 8y = 8$ $\qquad (****)$

We now proceed in the usual way. Multiplying equation $(***)$ by 6, we get

$36x - 60y = 18$ $\qquad$ (A)

Multiplying equation $(****)$ by 4, we get

$36x - 32y = 32$ $\qquad$ (B)

Subtracting equation (B) from equation (A), we get $-28y = -14 \Rightarrow y = \frac{1}{2}$. Substituting for $y = \frac{1}{2}$ in equation $(***)$, we get $x = \frac{8}{6} = \frac{4}{3}$. Therefore, the solutions are $y = \frac{1}{2}$ and $x = \frac{4}{3}$. Since equation $(***)$ came from equation $(*)$, we must do the check in equation $(**)$. Indeed, $\frac{3(4/3)}{4} - \frac{2(1/2)}{3} = \frac{2}{3}$.

Factoring Models

Common factor: $ax + ay = a(x + y)$.
Difference of squares: $x^2 - y^2 = (x + y)(x - y)$.
Trinomial (leading coefficient 1): $x^2 + (a + b)x + ab = (x + a)(x + b)$.
Perfect square trinomial: $x^2 + 2xy + y^2 = (x + y)^2$.
General trinomial: $(ac)x^2 + (ad + bc)x + bd = (ax + b)(cx + d)$.
Sum of cubes: $a^3 + b^3 = (a + b)(a^2 - ab + b^2)$.
Difference of cubes: $a^3 - b^3 = (a - b)(a^2 + ab + b^2)$.

Real Polynomials

A function of a single variable x is said to be a "polynomial" on its domain if it can be put in the following form:

$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$,

where $a_n, a_{n-1}, \ldots, a_1, a_0$ are constants. Hence, every polynomial can be expressed as a finite sum of monomial terms of the form $a_k x^k$, in which the variable is raised to a non-negative integral power. Notice that $x^0 = 1$, and so $a_0 x^0 = a_0$. For the aforementioned polynomial with $a_n \neq 0$:
the numbers a_i (where $0 \leq i \leq n$) are called "coefficients";
a_n is the "leading coefficient";
$a_n x^n$ is the "leading term";
a_0 is the "constant term" or the "constant coefficient";
a_1 is the "linear coefficient";

a_1x is the "linear term";
when the leading coefficient, a_n, is equal to 1, the polynomial is said to be "monic";
the non-negative integer n is the "degree" of the polynomial, and we write $\deg(p) = n$.
A "constant polynomial" has only one term, specifically, a_0. A non-zero constant polynomial has degree 0, and, by convention, the "zero polynomial" (with all coefficients vanishing) has degree $-\infty$.
A "zero" of a polynomial $p(x)$ is any number r for which $p(r)$ takes the value 0. Hence, when $p(r) = 0$, we say that r is a "root," or a "solution" of the equation $p(x) = 0$.
Let
$p(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ and
$q(x) = b_0 + b_1x + b_2x^2 + \cdots + b_mx^m$
be two arbitrary polynomials. Then we can operate with them as follows:
Sum: $(p + q)(x) = (a_0 + b_0) + (a_1 + b_1)x + (a_2 + b_2)x^2 + \cdots$
Difference: $(p - q)(x) = (a_0 - b_0) + (a_1 - b_1)x + (a_2 - b_2)x^2 + \cdots$
Product of a constant and a polynomial: $(cp)(x) = ca_0 + ca_1x + ca_2x^2 + \cdots$
Product of two polynomials: $(p \cdot q)(x) = a_0b_0 + (a_0b_1 + a_1b_0)x + (a_0b_2 + a_1b_1 + a_2b_0)x^2 + \cdots + (a_0b_k + a_1b_{k-1} + \cdots + a_ib_{k-i} + \cdots + a_kb_0)x^k + \cdots + (a_nb_m)x^{m+n}$.
Composition of two polynomials: $(p \circ q)(x) = p\big(q(x)\big)$, so that we replace each occurrence of x in the expression for $p(x)$ with $q(x)$.
Notice that we divide one polynomial by another in a manner similar to the division of two integers. Firstly, we arrange the terms of the dividend and the divisor in descending powers of x. If a term is missing, then we write 0 as its coefficient. Then, we divide the first term of the dividend by the first term of the divisor to obtain the first term of the quotient. Next, we multiply the entire divisor by the first term of the quotient, and we subtract this product from the dividend. We use the remainder as the new dividend, and we repeat the same process until the remainder is of lower degree than the divisor. As with the division of numbers,
$dividend = (divisor)(quotient) + remainder.$
Remainder Theorem: If a polynomial $p(x)$ is divided by $x - b$, then the remainder is $p(b)$.
Proof: Let $q(x)$ and r be, respectively, the quotient and the remainder when $p(x)$ is divided by $x - b$. Then, given that
$dividend = (divisor)(quotient) + remainder,$
it holds that, for any x,
$p(x) = (x - b)q(x) + r.$

If $x = b$, then $p(b) = r$.■

Factor Theorem: Given an arbitrary polynomial function $y = p(x)$, b is a zero of $y = p(x)$ if and only if $x - b$ is a factor of $p(x)$.

Proof. It can be easily verified using the Remainder Theorem.■

Remark: The usefulness of the Factor Theorem can be illustrated by the following examples. *Example 1:* We can write a polynomial function (in factored form) of degree 3 with zeros -1, 4, and 3 as follows: $p(x) = (x + 1)(x - 4)(x - 3)$. *Example 2:* If $p(x) = x(x + 3)(x - 7)^2$, then the zeros of this function are 0, -3, and 7; and, in particular, 7 is a "zero of multiplicity 2," since there are two factors of $x - 7$. In general, the "multiplicity of a zero" b is given by the highest power of $x - b$ that is a factor of $p(x)$.

The real number zeros of $y = p(x)$ are also the x-intercepts in the graph of $y = p(x)$. If b is a real number zero with multiplicity n of $y = p(x)$, then the graph of $y = p(x)$ crosses the x-axis at $x = b$ if n is odd, whereas the graph turns around and stays on the same side of the x-axis at $x = b$ if n is even. Hence, the x-intercepts can be obtained from the Factor Theorem, and the behavior of the graph at an x-intercept, say $(b, 0)$, is determined by the multiplicity of zero b, that is, by the highest power of $(x - b)$ that is a factor of $p(x)$. For instance, if $p(x) = (x + 1)(x - 2)^2$, then, by setting $x = 0$, we realize that the y-intercept is $(0,4)$. Because $(x + 1)$ is a factor with an odd exponent, it holds that $(-1,0)$ is an x-intercept at which the graph crosses the x-axis. Because $(x - 2)^2$ is a factor with an even exponent, it holds that $(2,0)$ is an x-intercept at which the graph touches the x-axis and then turns around.

In fact, the fundamental problem in algebra consists in finding ways of solving polynomial equations; specifically, we seek formulae for zeros/roots in terms of the coefficients of the corresponding polynomial. A well-known example is the "quadratic formula." If we have the quadratic equation $ax^2 + bx + c = 0$, where $a \neq 0$, then we have the formula

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

where the expression $b^2 - 4ac$ is known as the "discriminant," meaning that, if we have a number r such that $r^2 = b^2 - 4ac \Leftrightarrow r = \sqrt{b^2 - 4ac}$, then

$x_1 = \frac{-b+r}{2a}$ and $x_2 = \frac{-b-r}{2a}$

are the solutions to $ax^2 + bx + c = 0$.

If $f(x) = ax^2 + bx + c$ where $a, b, c \in \mathbb{Q}$, then the value of the discriminant shows how many roots $f(x) = 0$ has, and it explains the behavior of the quadratic polynomial $ax^2 + bx + c$; specifically:

- If $b^2 - 4ac < 0$, then the quadratic equation has no real roots (its roots are conjugate complex numbers, which will be studied in Chapter 9). Hence, if the discriminant is negative, then $f(x)$ never crosses the x-axis (the "roots" are where the graph crosses the x-axis). A quadratic expression $ax^2 + bx + c$ is always positive if and only if the discriminant is negative, that is, $b^2 - 4ac < 0$, and $a > 0$. A quadratic expression $ax^2 + bx + c$ is always negative if and only if the discriminant is negative, that is, $b^2 - 4ac < 0$, and $a < 0$.
- If $b^2 - 4ac = 0$, then the quadratic equation has one repeated real, rational root.
- If $b^2 - 4ac > 0$ and is a perfect square (i.e., a positive integer that is obtained by multiplying an integer by itself), then the quadratic equation has two distinct real, rational roots.
- If $b^2 - 4ac > 0$ and is not a perfect square, then the quadratic equation has two distinct real, irrational roots.

Notice that, when the discriminant is positive, the quadratic function crosses the x-axis twice, so that it has two real roots (and then the function's sign will be the same as that of a when x is less than the smaller root or greater than the larger root, and the opposite of that of a when x is between the roots).

Vieta's formulae: If x_1 and x_2 are the roots of the quadratic equation $ax^2 + bx + c = 0$, then

$$x_1 + x_2 = -\frac{b}{a}$$

and

$$x_1 x_2 = \frac{c}{a}$$

(we can, thus, find the roots x_1 and x_2 of the quadratic equation by solving the aforementioned system of equations).

Now, let us try to find the roots of a third-degree polynomial. The first thing that we have to do is to find at least one root of the given cubic equation. Then we must divide that polynomial by the factor that we have found out by hit and trial, so that we ultimately come up with the roots of a quadratic equation. For instance, consider the cubic equation

$x^3 - 6x^2 + 11x - 6 = 0$.

By considering the factors of -6, namely, $1, 2, 3, -1, -2, -3, \ldots$, we notice that 1 satisfies the above equation, and then we divide this cubic equation by $x - 1$. Thus, we obtain the quotient $x^2 - 5x + 6$, which can be factored as follows: $(x-2)(x-3)$. The three roots of the given cubic equation are $x = 1$, $x = 2$, and $x = 3$.

For a cubic equation, Vieta's formulae can be formulated as follows: If x_1, x_2, and x_3 are the roots of the cubic equation $ax^3 + bx^2 + cx + d = 0$, then

$x_1 + x_2 + x_3 = -\frac{b}{a}$,

$x_1x_2 + x_1x_3 + x_2x_3 = \frac{c}{a}$,

and

$x_1x_2x_3 = -\frac{d}{a}$.

If a function $y = f(x)$ satisfies an equation of the form

$p_0(x)y^n + p_1(x)y^{n-1} + \cdots + p_{n-1}(x)y + p_n(x) = 0$,

where $p_0(x), \dots, p_n(x)$ are polynomials in x, then it is said to be an "algebraic function." In other words, an algebraic function is a function that can be defined as the root of a polynomial equation. If a function can be expressed as the quotient of two polynomials,

$f(x) = \frac{p(x)}{q(x)}$,

then it is called a "rational algebraic function."

Polynomials play a very important role in everyday life. For instance, the content of a shopping basket can be described in terms of a polynomial, engineers design roller coasters using polynomial functions with the quadratic equation, economists and businessmen use polynomials in order to model the growth rate and forecast revenues, etc.

The Cauchy–Schwarz–Bunyakovsky Inequality: If a_i and b_i are any real numbers ($i = 1,2., \dots, n$), then

$$\sum_{i=1}^{n} (a_i b_i) \leq \left(\sum_{i=1}^{n} a_i^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^{n} b_i^2\right)^{\frac{1}{2}}$$

with equality if and only if the sequences $(a_1, a_2, \dots, a_n)$ and $(b_1, b_2, \dots, b_n)$ are proportional, namely, there is a constant λ such that $a_k = \lambda b_k$ for each $k \in \{1,2, \dots, n\}$. This inequality can be easily proved by thinking as follows: For any $x \in \mathbb{R}$, we have

$\sum_{i=1}^{n}(a_i x + b_i)^2 \geq 0 \Leftrightarrow (\sum_{i=1}^{n} a_i^2)x^2 + 2x\sum_{i=1}^{n} a_i b_i + \sum_{i=1}^{n} b_i^2 \geq 0$.

The left-hand side of the last inequality is a quadratic polynomial in x. Because it cannot have two distinct real roots, its discriminant is non-positive, namely, $(\sum_{i=1}^{n} a_i b_i)^2 \leq (\sum_{i=1}^{n} a_i^2)(\sum_{i=1}^{n} b_i^2)$, *quod erat demonstrandum*.

The Minkowski Inequality: If a_i and b_i are any real numbers ($i = 1,2., \dots, n$), then

$$\left[\sum_{i=1}^{n}(a_i+b_i)^2\right]^{\frac{1}{2}} \leq \left(\sum_{i=1}^{n}a_i^2\right)^{\frac{1}{2}} + \left(\sum_{i=1}^{n}b_i^2\right)^{\frac{1}{2}}$$

with equality if and only if the sequences $(a_1, a_2, \dots, a_n)$ and $(b_1, b_2, \dots, b_n)$ are proportional. This inequality can be easily proved by thinking as follows: By the Cauchy–Schwarz–Bunyakovsky Inequality, $[\sum_{i=1}^{n}(a_i+b_i)^2]^{\frac{1}{2}} = [\sum_{i=1}^{n}(a_i^2+2a_ib_i+b_i^2)]^{\frac{1}{2}} \leq \left[\sum_{i=1}^{n}a_i^2 + 2(\sum_{i=1}^{n}a_i^2)^{\frac{1}{2}}(\sum_{i=1}^{n}b_i^2)^{\frac{1}{2}} + \sum_{i=1}^{n}b_i^2\right]^{\frac{1}{2}} = (\sum_{i=1}^{n}a_i^2)^{\frac{1}{2}} + (\sum_{i=1}^{n}b_i^2)^{\frac{1}{2}}$, *quod erat demonstrandum*.

Fixed Points of Functions

By a "fixed point" of a function f, we mean a point a such that $f(a) = a$, that is, a belongs to both the domain and the range of f. In other words, a fixed point of a function is a point at which the input to the function is equal to the output of the function (and, therefore, fixed points play an important role in equilibrium analysis). For instance, the function $f: \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x + 1$ has no fixed point, whereas the function $f: \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x$ has infinitely many fixed points (in fact, every real number is a fixed point of this function).

By definition, the fixed points of a function f are the solutions of $f(x) = x$ or the roots of $f(x) - x$. For instance, we can find the fixed points of $f(x) = \sqrt[3]{x}$ as follows: we set $f(x) = x$, and, therefore, $\sqrt[3]{x} = x \Leftrightarrow x^{1/3} = x \Leftrightarrow x = x^3 \Leftrightarrow 0 = x^3 - x \Leftrightarrow x(x^2 - 1) = 0$, and each of these factors, namely, x and $(x^2 - 1)$, must be set equal to 0. Thus, the first fixed point is $x = 0$, and the other fixed points are $x^2 - 1 = 0 \Leftrightarrow x^2 = 1 \Leftrightarrow x = \pm 1$. Hence, $f(x) = \sqrt[3]{x}$ has three fixed points: -1, 0, and $+1$.

One of the reasons why fixed points play a significant role in mathematical analysis is that the existence of solutions to systems of equations is equivalent to the existence of fixed points of appropriate functions. If we want to show that $f(x) = 0$ for some x, then this is equivalent to showing that $f(x) + x = x$, which means that the function F defined by $F(x) = f(x) + x$ has a fixed point.

Chapter 3
Matrices and Applications in Input-Output Analysis and Linear Programming

Matrices are often used in physics, statistics, and economics, and they are particularly useful when they are used in connection with systems of linear equations. For instance, let us considers the following linear simultaneous equations:

$$4x + 5y = 14$$
$$x + 2y = 11$$

By arranging the coefficients of x and y in the way in which they occur in the equations, we obtain the array

$$\begin{pmatrix} 4 & 5 \\ 1 & 2 \end{pmatrix},$$

which is an example of a matrix.

In general, consider the following rectangular array

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix},$$

consisting of m rows (i.e., horizontal n-tuples) and n columns (i.e., vertical m-tuples). This is called an "$m \times n$ matrix," usually denoted by $A = (a_{ij})$. If the number of rows in the matrix is m and the number of columns is n, then the matrix is said to be of order $m \times n$. The term "matrix" was introduced by the nineteenth-century English mathematician James Sylvester, but it was his friend the mathematician Arthur Cayley who developed the algebra of matrices in the 1850s.

Types of matrices:

i. *Row matrix.* This is a matrix having only one row; for instance, the following is a row matrix:
$$\begin{pmatrix} 4 & 5 \end{pmatrix}.$$
ii. *Column matrix.* This is a matrix having only one column; for instance, the following is a column matrix:
$$\begin{pmatrix} 5 \\ 2 \end{pmatrix}.$$
iii. *Null matrix.* This is a matrix with all its elements zero.
iv. *Square matrix.* This is a matrix having the same number of rows and columns.

v. *Diagonal matrix.* This is a square matrix in which all the elements are zero except the main diagonal elements (the main diagonal in a matrix always runs from upper left to lower right, so that the main diagonal of a matrix $A = (a_{ij})$ is the list of entries a_{ij} where $i = j$); for instance, the following is a diagonal matrix: $\begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}$.

vi. *Identity matrix.* This is a diagonal matrix in which the main diagonal elements are equal to 1 (an identity matrix is usually denoted by I); for instance, the following is an identity matrix: $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

Two matrices are "equal" if and only if their corresponding elements are equal.

Addition and Subtraction of Matrices: Two matrices may be added or subtracted provided that they are of the same order. Addition of matrices is done by adding together the corresponding elements of each of the two matrices. For instance:

$$\begin{pmatrix} 4 & 5 \\ 1 & 2 \end{pmatrix} + \begin{pmatrix} 3 & 6 \\ 2 & 4 \end{pmatrix} = \begin{pmatrix} 4+3 & 5+6 \\ 1+2 & 2+4 \end{pmatrix} = \begin{pmatrix} 7 & 11 \\ 3 & 6 \end{pmatrix}.$$

In general, the sum of two $m \times n$ matrices A and B is an $m \times n$ matrix C whose elements are $c_{ij} = a_{ij} + b_{ij}$, where $a_{ij} \in A$, $b_{ij} \in B$, $1 \leq i \leq m$, and $1 \leq j \leq n$.

Properties of the addition of matrices: If A, B, and C are $m \times n$ matrices, then:

$A + B = B + A$;

$A + (B + C) = (A + B) + C$;

$A = B \Leftrightarrow A + C = B + C$;

the equation $X + B = A$ has a unique solution $X = A - B$; and

$-(A + B) = -A - B$.

Subtraction of matrices is done in a similar way except the corresponding elements are subtracted. For instance:

$$\begin{pmatrix} 4 & 5 \\ 1 & 2 \end{pmatrix} - \begin{pmatrix} 3 & 6 \\ 2 & 4 \end{pmatrix} = \begin{pmatrix} 4-3 & 5-6 \\ 1-2 & 2-4 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ -1 & -2 \end{pmatrix}.$$

Multiplication of Matrices:

i. Scalar multiplication: A matrix may be multiplied by a number as follows:

$$4\begin{pmatrix} 5 & -2 \\ 1 & 8 \end{pmatrix} = \begin{pmatrix} 4 \times 5 & 4 \times (-2) \\ 4 \times 1 & 4 \times 8 \end{pmatrix} = \begin{pmatrix} 20 & -8 \\ 4 & 32 \end{pmatrix}.$$

In general, given a matrix $A = (a_{ij})$ and a real number k, $kA = (ka_{ij})$.

ii. General Matrix Multiplication: Two matrices can only be multiplied by each other if the number of columns in the one is equal to the number of rows in the other. Multiplication of matrices is done by multiplying a row by a column as follows: $\begin{pmatrix} 4 & 5 \\ 1 & 2 \end{pmatrix} \times \begin{pmatrix} 3 & 6 \\ 2 & 4 \end{pmatrix} = \begin{pmatrix} 4 \times 3 + 5 \times 2 & 4 \times 6 + 5 \times 4 \\ 1 \times 3 + 2 \times 2 & 1 \times 6 + 2 \times 4 \end{pmatrix} = \begin{pmatrix} 22 & 44 \\ 7 & 14 \end{pmatrix}$.

The product of an $m \times n$ matrix $A = (a_{ij})$ and an $n \times p$ matrix $B = (b_{ij})$ is an $m \times p$ matrix $C = AB = (c_{ij})$ whose (i, j) entry is $c_{ij} = \sum_{k=1}^{n} a_{ik}\, b_{kj}$, where $1 \leq i \leq m$ and $1 \leq j \leq p$. Thus, if A is an $m \times n$ matrix, and if B is $n \times p$ matrix, then the product AB is an $m \times p$ matrix in which the element that corresponds to the ith row and the jth column of AB is found by multiplying each element in the ith row of A by the corresponding element in the jth column of B and adding the results.

Properties of the multiplication of matrices:

$A(BC) = (AB)C$;

$A(B + C) = AB + AC$;

$(B + C)A = BA + CA$.

Inverting a Matrix: An n-square matrix A is said to be "invertible" or "non-singular" if there exists an n-square matrix B with the following property:

$AB = BA = I_n$,

where I_n is the n-square identity matrix, namely, the $n \times n$ matrix with ones along the main diagonal and zeros elsewhere. If this is the case, then the matrix B is called the inverse of A, and the notation A^{-1} is used to designate B. If no such B exists, then A is said to be "singular." If

$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$,

then

$A^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$.

In general, A^{-1} can be found in the following way: (i) Write the augmented matrix $[A|I]$, where I is the identity matrix with the same dimension as A. (ii) Using elementary row operations, replace matrix $[A|I]$ with a matrix of the form $[I|B]$. (iii) Then A^{-1} is exactly this matrix B. For instance, if

$A = \begin{pmatrix} 3 & 0 & 2 \\ 2 & 0 & -2 \\ 0 & 1 & 1 \end{pmatrix}$,

then we can construct A^{-1} as follows:

We form the augmented matrix:

$$\left(\begin{array}{ccc|ccc} 3 & 0 & 2 & 1 & 0 & 0 \\ 2 & 0 & -2 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right).$$

We add the second row to the first row to obtain

$$\left(\begin{array}{ccc|ccc} 5 & 0 & 0 & 1 & 1 & 0 \\ 2 & 0 & -2 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right).$$

Then we divide the first row by 5 to obtain

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 0.2 & 0.2 & 0 \\ 2 & 0 & -2 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right).$$

Now, let's take two times the first row and subtract it from the second row to obtain

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 0.2 & 0.2 & 0 \\ 0 & 0 & -2 & -0.4 & 0.6 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right).$$

We multiply the second row by $-\frac{1}{2}$ to obtain

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 0.2 & 0.2 & 0 \\ 0 & 0 & 1 & 0.2 & -0.3 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right).$$

Now, we swap the second and the third rows to obtain

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 0.2 & 0.2 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0.2 & -0.3 & 0 \end{array}\right).$$

Finally, we subtract the third row from the second row to obtain

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 0.2 & 0.2 & 0 \\ 0 & 1 & 0 & -0.2 & 0.3 & 1 \\ 0 & 0 & 1 & 0.2 & -0.3 & 0 \end{array}\right),$$

and, thus, we constructed

$$A^{-1} = \begin{pmatrix} 0.2 & 0.2 & 0 \\ -0.2 & 0.3 & 1 \\ 0.2 & -0.3 & 0 \end{pmatrix}.$$

Transposition of Matrices: The "transpose" of a matrix A is denoted by A^T, and it is the matrix obtained by writing the rows of A, in order, as columns; that is, if $A = (a_{ij})$ is an $m \times n$ matrix, then $A^T = (a_{ij}^T)$ is the $n \times m$ matrix where $a_{ij}^T = a_{ji}$, for all i and j. For instance, if

$A = \begin{pmatrix} 1 & 7 \\ 4 & 3 \end{pmatrix}$, then $A^T = \begin{pmatrix} 1 & 4 \\ 7 & 3 \end{pmatrix}$.

If a square matrix A is such that $A = A^T$, then it is called "symmetric" (its elements are symmetric with respect to its main diagonal).

Determinants: The determinant of a matrix A is a scalar assigned to A, and it is denoted by $det(A)$. Given a matrix

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

its determinant is

$$det(A) = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}.$$

Notice that a matrix A has an inverse if and only if the determinant of A is not zero.

Solution to simultaneous equations using matrices: Let us consider a system of two linear equations with two unknowns:

$$\begin{Bmatrix} a_1x + b_1y = c_1 \\ a_2x + b_2y = c_2 \end{Bmatrix},$$

which gives rise to the following three matrices:

$$A = \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix}, B = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}, \text{ and } X = \begin{pmatrix} x \\ y \end{pmatrix}.$$

Thus, the original system of linear equations can be reformulated as follows:

$$AX = B \Leftrightarrow X = A^{-1}B,$$

where A is the matrix of the system's coefficients, X is the matrix of the system's unknowns, and B is the matrix of the system's constant terms. The system has a unique solution if and only if the determinant $\det(A) = a_1b_2 - b_1a_2 \neq 0$, and that solution is:

$$x = \frac{B_x}{det(A)} = \frac{\begin{vmatrix} c_1 & b_1 \\ c_2 & b_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}} = \frac{c_1b_2 - b_1c_2}{a_1b_2 - b_1a_2}$$

and

$$y = \frac{B_y}{det(A)} = \frac{\begin{vmatrix} a_1 & c_1 \\ a_2 & c_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}} = \frac{a_1c_2 - c_1a_2}{a_1b_2 - b_1a_2}$$

where the numerators B_x and B_y are obtained by substituting the column of constant terms in place of the column of coefficients of the corresponding unknown in the matrix of coefficients. If $\det(A) = 0$, then the system has either no solution or an infinite number of solutions.

Consider the 3-square matrix

$$A = \begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{pmatrix}.$$

The determinant of A is

$$\det(A) = \begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix} = a_1 b_2 c_3 + b_1 c_2 a_3 + c_1 a_2 b_3 - a_3 b_2 c_1 -$$

$b_3 c_2 a_1 - c_3 a_2 b_1$.

Moreover, it can be easily shown that

$$\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix} = a_1 \begin{vmatrix} b_2 & c_2 \\ b_3 & c_3 \end{vmatrix} - b_1 \begin{vmatrix} a_2 & c_2 \\ a_3 & c_3 \end{vmatrix} + c_1 \begin{vmatrix} a_2 & b_2 \\ a_3 & b_3 \end{vmatrix}.$$

Let us consider a system of 3 linear equations with 3 unknowns:

$$\begin{cases} a_1 x + b_1 y + c_1 z = d_1 \\ a_2 x + b_2 y + c_2 z = d_2 . \\ a_3 x + b_3 y + c_3 z = d_3 \end{cases}$$

The aforementioned system has a unique solution if and only if the determinant of the matrix of coefficients is not zero:

$$\det(A) = \begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix} \neq 0.$$

In this case, the unique solution to the given system can be expressed as quotients of determinants as follows:

$$\begin{cases} x = \dfrac{B_x}{\det(A)} \\ y = \dfrac{B_y}{\det(A)} \\ z = \dfrac{B_z}{\det(A)} \end{cases}$$

where the numerators B_x, B_y, and B_z are obtained by substituting the column of constant terms for the column of coefficients of the corresponding unknown in the matrix of coefficients, so that:

$$B_x = \begin{vmatrix} d_1 & b_1 & c_1 \\ d_2 & b_2 & c_2 \\ d_3 & b_3 & c_3 \end{vmatrix}, B_y = \begin{vmatrix} a_1 & d_1 & c_1 \\ a_2 & d_2 & c_2 \\ a_3 & d_3 & c_3 \end{vmatrix}, \text{and } B_z = \begin{vmatrix} a_1 & b_1 & d_1 \\ a_2 & b_2 & d_2 \\ a_3 & b_3 & d_3 \end{vmatrix}.$$

If $\det(A) = 0$, then the system has either no solution or an infinite number of solutions.

Advances in computing power have significantly contributed to the application of matrix algebra in several scientific disciplines, such as physics and mathematical economics.

The Application of Matrices in Input-Output Analysis

In general, a “computational problem” is a binary relation between inputs and outputs, and, in particular, we specify which outputs are correct for the given inputs by means of specific predicates.

The major economic tasks that every society must accomplish pertain to decision-making about an economy’s inputs and outputs. In economics, the term “input” refers to goods or services used by firms in their production processes. Thus, by means of its technology, an economy combines inputs to produce outputs. In economics, the term “output” refers to the various useful goods or services that are either employed in further production or consumed.

The acknowledged founder of “input-output analysis” is the Russian-American economist Wassily Leontief, who won the Nobel Prize in Economics in 1973. An input-output matrix is a square matrix, say $A = (a_{ij})$, whose entries a_{ij} represent the amount of input i required per unit of output j. A column of such a matrix depicts the inputs needed for the achievement of a specific output. Therefore, from the perspective of economics, it can be considered as a “production technique.” Hence, an input-output matrix is a “constellation” of production techniques. If the list of inputs is complete, including factor inputs, then the input-output matrix contains techniques for the production of the factor services as well.

Input-output analysis is used in order to analyze inter-industry relations, thus explaining inter-dependencies and complexities of the economic system as well as the conditions for maintaining equilibrium between supply and demand. The inputs of one industry are the outputs of another industry, and vice versa. An input is obtained (purchased), and an output is produced. Hence, “input” represents the expenditure of a firm, and “output” represents the (sales) revenue of a firm. The sum of the money values of inputs is the total cost of a firm, and the total money value of the output is the total revenue of a firm. Input-output analysis implies that, in a state of equilibrium, the money value of the aggregate output of the whole economy must be equal to the sum of the money values of the inter-industry inputs and the sum of the money values of the inter-industry outputs. For instance, coal is an input for steel industry, and steel is an input for coal industry, but both coal and steel are the outputs of their respective industries. An important part of economic activity consists of the production of intermediate goods and services (inputs) for further use in producing final goods and services (outputs).

Let us divide the economic system into the “inter-industry sectors” and the “final-demand sectors,” each of which can be divided into different sub-sectors. The total output of any inter-industry sector can be used as an input by other inter-industry sectors, by the given sector (i.e., by itself), as well as by final-demand sectors. Prices, consumer demand, and factor supply (i.e., the availability of factors of production for purchase by producers) are given. Moreover, we assume that there are no externalities (by "externalities," we mean the indirect effects that have an impact on the consumption and the production opportunities of others, but the price of the product does not take those externalities into account; for instance, a traditional example of a negative externality is pollution, and the research and development (R&D) activities are traditionally associated with positive externalities, since they have positive effects beyond those enjoyed by the producer). Furthermore, an input-output model is based on the following assumptions: (i) No two products are produced jointly, and, therefore, each industry produces only one homogeneous product. (ii) Each producing sector satisfies the properties of a linear homogeneous production function (i.e., the production of each sector is subject to constant returns of scale: its inputs increase at the same rate as its outputs). (iii) The combinations of inputs are employed in rigidly fixed proportions (there are fixed input coefficients of production).

For instance, in Table 3-1, we see the input-output matrix of a four-sector economy, which, specifically, consists of three inter-industry sectors, namely, X_1, X_2, and X_3, as well as one final-demand sector. The rows of the input-output matrix (i.e., the rows of Table 3-1) inform us about the products of X_1, X_2, and X_3 that are used as intermediate products (inputs) by the corresponding inter-industry sector as well as for final consumption by the government and the households. The columns of the input-output matrix (i.e., the columns of Table 3-1) inform us about the total inputs (from all sectors) utilized by each inter-industry sector for its production.

Table 3-1: An input-output matrix.

Total output of the sectors	X_1	X_2	X_3	Final demand
X_1	X_{11}	X_{12}	X_{13}	F_1
X_2	X_{21}	X_{22}	X_{23}	F_2
X_3	X_{31}	X_{32}	X_{33}	F_3
Labor input	L_1	L_2	L_3	

Given Table 3-1, the rows, which represent consumption functions, can be written as follows:

$$X_1 = X_{11} + X_{12} + X_{13} + F_1$$
$$X_2 = X_{21} + X_{22} + X_{23} + F_2$$
$$X_3 = X_{31} + X_{32} + X_{33} + F_3$$
$$L = L_1 + L_2 + L_3$$

or, equivalently:

$$X_i = \sum X_{ij} + F_i$$
$$L = \sum L_i$$

where i and j vary from 1 to 3 (since, in this example, there are three inter-industry sectors).

Moreover, given Table 3-1, the columns, which represent production functions, can be written as follows:

$$X_1 = X_{11} + X_{21} + X_{31} + L_1$$
$$X_2 = X_{12} + X_{22} + X_{32} + L_2$$
$$X_3 = X_{13} + X_{23} + X_{33} + L_3$$

(and, thus, each entry a_{ij} in an input-output matrix represents the amount of input i required per unit of output j).

In Table 3-2, we see the corresponding technological coefficient matrix, for the same example (vertical interpretation: proportion of the corresponding commodity produced by the corresponding sector; horizontal interpretation: proportion of the corresponding commodity used by the corresponding sector).

Table 3-2: A technological coefficient matrix

Total output of the sectors	X_1	X_2	X_3	Final demand
X_1	$a_{11}X_1$	$a_{12}X_2$	$a_{13}X_3$	F_1
X_2	$a_{21}X_1$	$a_{22}X_2$	$a_{23}X_3$	F_2
X_3	$a_{31}X_1$	$a_{32}X_2$	$a_{33}X_3$	F_3
Labor input	L_1	L_2	L_3	

Given that we have assumed that the input requirements are fixed, the amount of input i that is required in order to produce one unit of output j is given by $a_{ij} = X_{ij}/X_j$, and we have:

$$\begin{aligned} X_1 &= a_{11}X_1 + a_{12}X_2 + a_{13}X_3 + F_1 \\ X_2 &= a_{21}X_1 + a_{22}X_2 + a_{23}X_3 + F_2 \\ X_3 &= a_{31}X_1 + a_{32}X_2 + a_{33}X_3 + F_3 \\ L &= l_1X_1 + l_2X_2 + l_3X_3 \end{aligned}$$

or, equivalently:

$$X_i = \sum a_{ij} X_j + F_i$$
$$L = \sum l_i X_i$$

for $i = 1,2,3$ (since, in this example, there are three inter-industry sectors). Therefore, if we define the matrices

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix},$$

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix},$$

and

$$F = \begin{pmatrix} F_1 \\ F_2 \\ F_3 \end{pmatrix},$$

then we have:

$$X = AX + F$$
$$L = \sum l_i X_i$$

(and, thus, we can determine the optimum level of production for the given economic network). Notice that F indicates that the inter-industry sectors not only satisfy each other's needs, but they also satisfy some outside demands (on the other hand, in case of a "closed system," $F = 0$).

If I is the 3-square identity matrix, then, given the above definitions of the matrices X, A, and F, we can formulate the matrix equation

$X = AX + F \Leftrightarrow X - AX = F \Leftrightarrow (I - A)X = F \Leftrightarrow X = (I - A)^{-1}F,$

from which we can get the values of X_1, X_2, and X_3 that correspond to a state of equilibrium between supply and demand (and, thus, we can avoid both oversupplying and undersupplying the market with the corresponding commodities).

The Application of Matrices in Linear Programming

By the term “linear programming,” we mean a method to achieve the best outcome (e.g., to maximize profit, minimize cost, etc.) in a mathematical model whose requirements are represented by linear functions. The first contributions to linear programming are due to the Soviet mathematician and economist Leonid Vitalyevich Kantorovich (1912–86), who won the Nobel Prize in Economics in 1975. Moreover, one of the acknowledged founders of linear programming is the American mathematician George Bernard Dantzig (1914–2005), who managed to make significant contributions to industrial engineering, operations research, economics, statistics, and computer science. In fact, input-output analysis is a special and very important case of linear programming.

The “canonical form” of linear programming is the following: given a system of m linear constraints (or linear inequalities) with n variables, we wish to find non-negative values (i.e., ≥ 0) of these variables that will satisfy the constraints and will maximize a function of these variables; symbolically: given m linear inequalities and/or equalities

$$\sum_j a_{ij}\, x_j \leq b_i, i = 1,2, \dots, m, and\ j = 1,2, \dots, n, \qquad (*)$$

we wish to find those values of x_j which satisfy the constraints $(*)$ and the condition that $x_j \geq 0$ (for $j = 1,2, \dots, n$) and simultaneously maximize the linear function

$$z = \sum c_j x_j\, , j = 1,2, \dots, n. \qquad (**)$$

For instance, consider a problem where we wish to maximize the gross profit of an industry (or of a firm offering several product lines) that produces n commodities, and, thus, has n sectors of production. In this case, $(*)$ and $(**)$ can be interpreted as follows: z denotes an overall performance measure (specifically, total gross profit); x_j denotes the level of activity j ($j = 1,2, \dots, n$), specifically, the output of the jth sector of production (i.e., the produced quantity of the jth commodity); c_j denotes the performance measure coefficient for activity j, specifically, the gross profit per unit of output in the jth sector of production (so that the total gross profit in the jth sector of production is $c_j x_j$); b_i denotes the available quantity of resource (input) i ($i = 1,2, \dots, m$); and a_{ij} denotes the quantity of resource (input) i consumed by each unit of activity j (i.e., required per unit of output j).

In matrix form, the constrained maximization problem $(**)$ can be rewritten as follows:

$$z_{max} = \begin{pmatrix} c_1 & c_2 & \cdots & c_n \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix},$$

under the constraints

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \leq \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix},$$

and

$x_j \geq 0$ for $j = 1,2,\ldots,n$. More simply, given the above concepts, we can write:

$$\left.\begin{array}{c} maxz = cx \\ under\ the\ constraints \\ Ax \leq b \\ x_j \geq 0 \end{array}\right\}. \qquad (***)$$

Regarding the geometric significance of $(***)$, notice that the constraints $Ax \leq b$ and $x_j \geq 0$ define a convex polyhedron P_n in $\mathbb{R}^n$, and such P_n is called the "feasible region" of the corresponding model, meaning the region of all the feasible solutions to the corresponding problem. In general, a polyhedron P_n in $\mathbb{R}^n$ is the set of all points $x \in \mathbb{R}^n$ that satisfy a finite set of linear inequalities. Moreover, a set Q in $\mathbb{R}^n$ is called "convex" if, for any two points x and y in Q, the line segment joining them lies entirely within Q (and, in particular, a convex polyhedron is a polyhedron for which a line connecting any two non-coplanar points on the surface of the polyhedron always lies in the interior of the polyhedron); symbolically: $\forall x, y \in Q$, the "convex combination" $kx + (1-k)y \in Q$ for any k such that $0 \leq k \leq 1$. The goal of constrained maximization in the context of linear programming is to choose that feasible combination $(x_1, x_2, \ldots, x_n)$ of actions that maximize a given function $z = cx$. This occurs at the maximum point $(x_1^*, x_2^*, \ldots, x_n^*)$ of the feasible region.

The constrained maximization problem $(***)$ is known as the "primal problem," while the so-called "dual problem" is the corresponding constrained minimization problem where, given a system of m linear constraints (linear inequalities) with n variables, we wish to find non-negative values (i.e., ≥ 0) of these variables that will satisfy the constraints and will minimize a function (e.g., a cost function) of these variables; symbolically (if, for instance, z represents total cost, c represents cost per unit of output, and b represents the required level of output), we obtain the following model:

$$\left.\begin{array}{c} minz = cx \\ under\ the\ constraints \\ Ax \geq b \\ x_j \geq 0 \end{array}\right\}. \qquad (****)$$

For instance, using the "dual problem," we can create models of constrained cost minimization in economics and business management.

Firms seek to minimize cost subject to the constraint that they produce at least b units of output, so that the firm's cost minimization problem is given by $(****)$. In general, linear programming (also known as linear optimization) is useful for guiding quantitative decisions in business planning, microeconomics, industrial engineering, and in several other problems of the social and the natural sciences.

Example: Let us consider the following linear-programming problem:

$$maxz = 4x + 5y$$

under the constraints

$$x + y \leq 20$$
$$3x + 4y \leq 72$$
$$x, y \geq 0$$

where $z = 4x + 5y$ is the objective function. We work as follows: Firstly, we have to draw the straight lines that represent the constraints, thus, finding x-intercepts and y-intercepts. For the constraint
$x + y \leq 20$,
we set $y = 0$ to find the x-intercept, which is $x + 0 = 20 \Rightarrow x = 20$, and, therefore, the x-intercept is the point $(20,0)$. For the same constraint, we set $x = 0$ to find the y-intercept, which is $0 + y = 20 \Rightarrow y = 20$, and, therefore, the y-intercept is the point $(0,20)$. For the constraint
$3x + 4y \leq 72$,
by setting $y = 0$, we obtain $3x + 0 = 72 \Rightarrow x = 24$, and, therefore, the the x-intercept is the point $(24,0)$; and, by setting $x = 0$, we obtain $0 + 4y = 72 \Rightarrow y = 18$, and, therefore, the y-intercept is the point $(0,18)$. Moreover, regarding the other given constraints, we notice that $x = 0$ is the y-axis, and $y = 0$ is the x-axis. In Figure 3-1, the shaded region (a convex polyhedron) is the feasible region, and the points that lie within it satisfy all the given constraints simultaneously. In order to find the intersection point Cin Figure 3-1 we have to solve the following system of equations:

$$x + y = 20 \qquad (1)$$
$$3x + 4y = 72 \qquad (2)$$

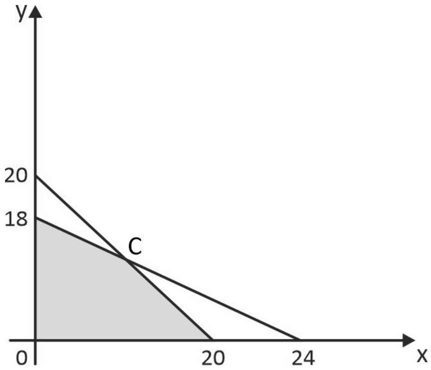
and, therefore, we multiply equation (1) by -3 and then add it to equation (2) to obtain $y = 12$, and $x = 8$. Hence, the point C in Figure 3-1 is $(8,12)$. Now, given Figure 3-1, we substitute all the corner points into the objective function $z = 4x + 5y$ in order to find the maximum one. Hence, we obtain the following results:

Table 3-3: A linear programming problem.

Corner points	Value of z
(20,0)	80
(0,18)	90
(8,12)	92

Consequently, as indicated in Table 3-3, the maximum value of the objective function is 92, and the corner point that corresponds to this value is $(8,12)$, meaning that the optimal solution to the given linear-programming problem is $(x, y) = (8,12)$.

Figure 3-1: The feasible region.



Chapter 4
Basic Mathematical Economics, Political Economy, and a Vision of Scientific Totalitarianism

By the term “economy,” we refer to a system for making decisions about the use of limited resources so that goods and services can be produced and cosumed. By the term “market,” we refer to a system in which two or more parties participate in order to engage in economic transactions.

Standard economic analysis is based on the concept of rationality. In general, in the social sciences, “rationality” means that social behavior can be seen in terms of actors pursuing goals. The “rationality postulate” implies the following: (i) actors have well ordered preference systems over the set of outcomes (i.e., of alternative actions), namely, for all pairs c_i and c_j, there is a preference relation R such that either c_iRc_j (the actor prefers c_i to c_j), or c_jRc_i (the actor prefers c_j to c_i), or both (the actor is indifferent); (ii) each actor’s preference system is substantially independent of the other social variables; and (iii) each actor acts to maximize one’s utility index (according to the principle of utility, an action is good in so far as it tends to promote happiness for moral agents, and the meaning of happiness depends on one’s ethics; for instance, according to Plato, societal happiness stems from citizens treating each other justly, leading virtuous lives, and each fulfilling their social function). In particular, one can formulate a decreasing sequence of numbers (these numbers are called “utilities,” u_n) where the largest number is assigned to the most preferred outcome, the second largest number to the next outcome in the preference order, etc. The function that maps consequences to numbers representing an actor’s preference over those outcomes is said to be a “utility function.” The most well-known utility function is the von Neumann–Morgenstern utility function, which is defined as follows: the actor considers a set of all conceivable states of the world and assesses the likelihood of each state S by assigning a probability $p(S)$ to it, so that the expected utility $U_e(A)$ for an action A can be calculated by multiplying the probability $p(S)$ of each state’s occurring by the utility $u\big(C(S,A)\big)$ of the outcome that results from the given state of the world and the given action, and then summing these products over all the possible states:

$U_e(A) = \sum_{all\,S} p(S)u\big(C(S,A)\big)$;

the actor chooses A such that $U_e(A)$ is maximized.

"Marginal utility" measures the change in the satisfaction that a consumer gets from consuming one more unit of a commodity; symbolically:

$Marginal\ Utility = \frac{\Delta TU}{\Delta Q}$,

where ΔTU denotes the change in total utility, and ΔQ denotes the change in the quantity consumed of the corresponding commodity. "Total utility" measures the total amount of satisfaction that a consumer gets from all the units that he/she consumes of a commodity.

By analogy, "marginal cost" measures the change in the total cost that comes from making or producing one additional unit (of output); symbolically:

$Marginal\ Cost = \frac{\Delta C}{\Delta Q}$,

where ΔC denotes the change in total cost, and ΔQ denotes the change in output. "Total cost" is the sum of the expenses that a producer needs to make in order to achieve a specific level of output.

"Productivity" measures how much output can be produced with a given set of inputs, and "marginal productivity" measures the change in output as a result of one additional unit of input; symbolically:

$Marginal\ Productivity = \frac{\Delta Y}{\Delta X}$,

where ΔX denotes the change in the firm's use of the input, and ΔY denotes the change in the quantity of output produced.

Economic Equilibrium and Economic Planning

As shown in Figure 4-1, the demand curve is drawn with the price on the vertical axis (y-axis) and quantity demanded on the horizontal axis (x-axis), thus obtaining a downward-sloping curve, meaning that, as price decreases, the quantity demanded will increase. Moreover, as shown in Figure 4-1, the supply curve is drawn with the price on the vertical axis (y-axis) and quantity supplied on the horizontal axis (x-axis), thus obtaining an upward-sloping curve, meaning that, as price increases, the quantity supplied will increase.

In a competitive market—which is based on the assumptions that the number of economic actors is so big that none of them can influence prices significantly by varying one's demand or supply, and that economic actors can freely enter and exit each trade and industry—economic equilibrium is achieved by trial and error within the context of a competitive market. In particular, according to this economic model, the conditions of "economic equilibrium" are the following:

i. The individual condition of equilibrium: it refers to the maximization of the individuals' utility, profit, or income from the ownership of productive resources. The consumers maximize the total utility that they derive from their income by spending it in such a way that the marginal utility of the quantity of commodity x_i obtainable for a unit of income (expressed in money) is equal for all commodities. Thus, given commodities $x_1, x_2, \dots, x_n$, the utility-maximizing rule for consumers can be formulated as follows:

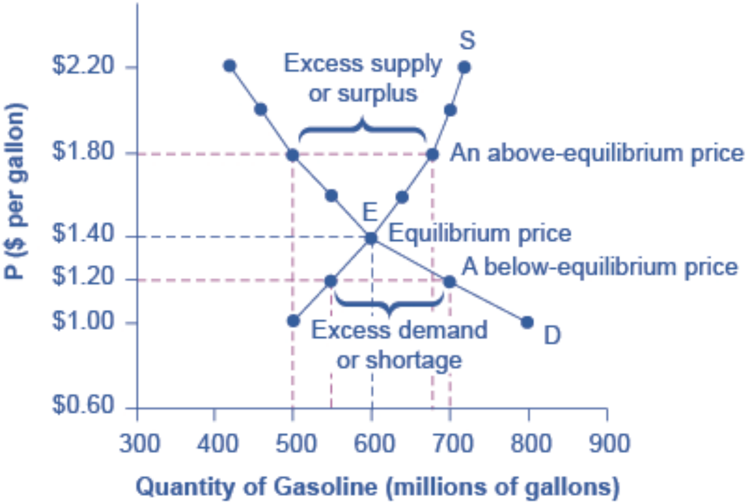
$$\frac{MU\ of\ x_1}{Price\ of\ x_1} = \frac{MU\ of\ x_2}{Price\ of\ x_2} = \cdots = \frac{MU\ of\ x_n}{Price\ of\ x_n}$$

where MU denotes marginal utility. The producers maximize their profit in two ways: firstly, by optimizing the combination of factors of production (i.e., by combining the factors of production in such a way that the marginal productivity of the quantity of factor of production x_i that can be purchased for a unit of money is equal for all factors of production) and, secondly, by optimizing the scale of output. If prices cannot be manipulated by particular actors, but are given by the market itself (as independent parameters), then the minimum cost curve of the producer is given (since the prices of the factors of production are given), and, therefore, the optimum scale of output is attained when the marginal cost is equal to the price of the product (which is given by the market itself). The owners of the fundamental productive resources (namely, labor, capital, and natural resources) maximize their income by selling the services of these resources to the highest bidder.

ii. The social condition of equilibrium: it refers to the assumption that the incomes of the consumers are equal to their receipts from selling the services of the productive resources that they own, plus entrepreneurs' profits. When the economic system is in a state of equilibrium, entrepreneurs' profits are equal to zero, since the marginal cost is equal to the price of the product (which is given by the market itself). By "zero entrepreneurs' profits," economists mean that, in a state of equilibrium, workers, managers, lenders, and owners of resources are earning their equilibrium returns, and this situation does not mean that there are in fact no profits, but that profits are expressed as differences in the remuneration earned by different economic actors (such as, for instance, the difference between the remuneration for providing managerial services or leadership and the remuneration for providing basic labor skills). This condition changes substantially when particular economic actors can manipulate prices and, generally, the conditions of economic activity by creating oligopolistic or monopolistic conditions.

iii. The structural condition of equilibrium: The equilibrium prices are determined by the condition that the demand for each commodity is equal to the supply of the corresponding commodity, as shown in Figure 4-1. The French economist Léon Walras (1834–1910) has explained this process as follows: On the basis of a (historically) given random set of prices, the economic actors strive to satisfy the individual condition of equilibrium and optimize their positions. To each commodity there correspond a quantity demanded and a quantity supplied. If, for each commodity, the quantity demanded and the quantity supplied are equal, then the entire situation is settled, and the prices are the equilibrium prices. But, if the quantities demanded and the quantities supplied diverge, the competition of the buyers and the sellers will alter the prices. When supply exceeds demand for a good/service, the price of this good/service tends to fall, and, when demand exceeds supply of a good/service, the price of this good/service tends to rise. As a result, the economic actors get a new set of prices, which serves as a new basis for the economic actors' attempt to satisfy the individual condition of equilibrium and optimize their positions. The individual condition of equilibrium being carried out, the economic actors get a new set of quantities demanded and quantities supplied. If, for each commodity, demand and supply are not equal, then prices will change again, and the economic actors will get another set of prices, which serves as a new basis for the economic actors' attempt to satisfy the individual condition of equilibrium and optimize their positions; and so on.

Figure 4-1: Market equilibrium price: in this example, the supply curve (S) and the demand curve (D) intersect at the equilibrium point E, representing a price of $1.40 and a quantity of 600 (Source: Wikimedia Commons: Author: OpenStax College; https://openstax.org/details/books/principles-microeconomics).



Assuming that freedom of choice in consumption and that freedom of choice of occupation are maintained, and assuming that the preferences of consumers, as expressed by their demand prices, guide production and the allocation of resources, the major goals of a rational and scientifically rigorous Central Economic Planning Authority (CEPA) are the following:

i. Minimization of the average cost of production: The managers who run existing plants and those who are engaged in building new plants must be guided and controlled by the CEPA in order to combine factors of production in such a way that the marginal productivity of the quantity of factor of production x_i that can be purchased for a unit of money is equal for all factors of production. In other words, given inputs (factors of production) $x_1, x_2, \ldots, x_n$ employed in a productive activity, the CEPA ensures that the following cost-minimizing rule is fulfilled

$$\frac{MP\ of\ x_1}{Price\ of\ x_1} = \frac{MP\ of\ x_2}{Price\ of\ x_2} = \cdots = \frac{MP\ of\ x_n}{Price\ of\ x_n}$$

where MP denotes marginal productivity. In this way, the CEPA ensures that each commodity is produced with a minimum sacrifice of alternatives, and it tries to eliminate social waste.

ii. Optimization of the scale of output: The managers of plants and the leaders of whole industries must be guided and controlled by the CEPA in order to determine the scale of production in such a way that the marginal cost is equal to the price of the product. In this way, the CEPA ensures that the marginal significance of each preference that is satisfied is equal to the marginal significance of the alternative preferences, which have been sacrificed, and, thus, the CEPA maintains a well defined hierarchy of preferences.

iii. The maintenance of an objective price structure: In the model of a competitive market, there is an objective price structure, in the sense that, as a result of the parametric function of prices, there is only one set of prices, which satisfies the structural equilibrium condition, that is, it equalizes the demand for and the supply of each commodity. The same objective must be consciously and intentionally maintained by the CEPA, but, whereas, according to the model of a competitive market, the parametric function of prices derives merely from the weak and fragile assumption that the number of competing economic actors is too large to enable any one to influence prices by one's own action, the CEPA can and should ensure the imposition of the parametric function of prices by imposing rational and scientifically rigorous price controls on strategic resources and, generally, on goods and services of critical social importance, in conjunction with appropriate monetary and fiscal policies.

iv. Rational control of the production process: The huge economic progress that took place during the nineteenth and the twentieth centuries was mainly a consequence of scientific, technological, and organizational innovations that (as they were integrated into the production process) increased the productivity of a combination of factors of production, or created new economic goods and services. However, given the contradictions of the capitalist system, the results of the integration of scientific, technological, and organizational innovations into the economy are not homogeneous. Companies that innovate make a direct profit or increase their profitability, but this profit (or increase in their profitability) is a temporary phenomenon, as free competition will tend to equate the price of the product with the average cost of production. On the other hand, companies that use outdated factors of production or outdated production models, and companies that produce competitive economic goods that can be easily substituted with others (by competitors) in the market, suffer losses which lead to a devaluation of the capital invested in them. In the competitive market regime, due to the parametric function of prices and the freedom of entry and exit enjoyed by private companies in every sector of the economy, any innovation is inevitably associated with a reduction in the value of some old investments, since, in principle, there is no way of reacting against a given innovation. What entrepreneurs can do to respond to their competitors' innovations is to try to innovate in their own companies, causing, in turn, losses for their competitors. Moreover, innovative companies need to constantly strive to innovate, because free competition tends to nullify the profitability of existing innovations (due to the freedom of entry of new competitors in each sector of the economy), so the more a company leads in the field of innovation the more profitable it becomes.

Nevertheless, as the prominent American economist, diplomat, and economic consultant John Kenneth Galbraith (1908–2006) has pointed out, industrial planning is inextricably linked to the size of the industrial complex, and size is not only a particular underpinning and provider of profits, but also the general underpinning and provider of technology and innovation. Furthermore, due to the inherent contradictions of capitalism, in the free competitive market, there emerge several phenomena that oppose free competition, such as the following: (i) monopolies, (ii) monopsonies, (iii) oligopolies, (iv) oligopsonies, and (v) groups of companies (i.e., gentlemen's agreements, cartels, concerns, pools, and trusts).

When the size of some business units increases so much that they can nullify both the efficiency of the parametric function of prices (thus being able to exert some control over prices) and the freedom of entry of new firms and new investors in a sector of the economy in general, such companies develop a strong tendency to prevent any development that could bring about a devaluation of the capital already invested. Therefore, when a firm is not forced by market competition to innovate, it will only innovate when the old invested capital is depreciated or if the reduction in production costs that is achieved by the immediate implementation of an innovation exceeds the devaluation of the capital already invested. As J. K. Galbraith has aptly explained, this delay in actualizing available possibilities to improve the economy works to the detriment of social interest. In addition, the British economist Lionel Robbins (1898–1984), who was made a life peer as Baron Robbins of Clare Market in the City of Westminster in 1959, has pointed out that the attempt of certain capitalist elites to maintain the value of their invested capital may lead them to prevent the entry of new producers who find the prospects of one economic sector more attractive than the prospects of other economic sectors, as well as to postpone or cancel the implementation of technical improvements that reduce costs and, consequently, reduce the price paid by the consumer.

In any case, the ruling capitalist elite seeks to keep the general development of innovation under control and to manage innovations according to its own particular interests, thus coming into conflict not only with the social interest, but also with a rival capitalist elite which wants to become the new ruling capitalist elite by displacing the previous one. As a result of the contradictions of the capitalist system, the protection of monopoly privileges and specific investments contradicts economic progress, in the sense that it hinders the reduction of prices and the improvement of the quality of economic goods and services, and it is a major source of imperialist rivalry between the great powers of the international system.[2] When the pressure of scientific, technological, and organizational innovations for structural change is far greater than the tendency of some capitalist elites to maintain the value of old investments and their control over economic dynamics, an economic crisis ensues. This crisis is exacerbated, at a later stage, by the

[2] See also: Mavroudeas, “Periodising Capitalism”; Warren, *Imperialism*.

intensification of stock speculation, which manifests itself through a bear market for old investments and a bull market for new investments (innovations). The CEPA has to correct the aforementioned structural flaws of the competitive market system, to implement an efficient policy of innovation, and to ensure and impose a rational and scientifically rigorous production model.

The Financial System

According to the standard functional definition of "money," four functions have been ascribed to money—namely: medium of exchange, unit of account, store of value, and standard of deferred payment. The stock of money held in an economy is held for various reasons: firstly, money is held in order to facilitate exchange (i.e., it is to be spent rather than saved), and, secondly, it may be held as an asset (i.e., to be saved rather than spent).

If the supply of money falls below the level that is necessary to support the growth of the economy, then the growth of the economy will be held below its potential. On the other hand, if the supply of money is above the level that is necessary to support the potential growth of the economy in real terms, then the growth of the economy in money terms will be greater than the growth in real terms, and this, other things equal, will manifest itself in inflation. The "central bank" is a public institution that is responsible for implementing and managing the monetary policy of a country, or of a group of countries, and it controls the money supply.

In an economy, there will always exist two groups of economic agents: (i) surplus units, namely, those whose revenue exceeds their current expenditure during a given period of time, and (ii) deficit units, namely, those whose expenditure exceeds their current revenue in a given period of time. Therefore, some mechanism is required to ensure that the surplus funds are channeled to the deficit units.

The surplus units can lend their excess funds directly to the deficit units. For instance, a person can buy company or government securities through a public issue. However, it is very often the case that a surplus unit will lend its excess funds to a financial institution ("financial intermediary"), which will then on-lend these funds by itself, buying company stocks, government bonds, or other assets in which it invests. Thus, instead of a direct contractual relationship between the provider and the user of the funds, there are two contractual relationships: (i) the surplus unit lends to or acquires a financial claim on the financial intermediary, and (ii) the financial intermediary lends to or acquires a financial claim on the ultimate borrower, the user of the funds. Financial intermediation facilitates the reconciliation of the differing needs of lender and borrower by means of: (i) maturity transformation (since a financial intermediary can borrow short and lend long), (ii) aggregation (i.e., by collecting together a large number of relatively small amounts), and (iii) risk transformation. The most important financial intermediaries are commercial banks, investment banks, insurance companies, mutual funds, hedge funds, pension funds, venture capitals, savings and loans associations, credit unions, mutual savings banks, and consumer finance companies.

In economics, by the term "interest," we refer to the profit return on investment. The money that is invested is called the "principal." The percentage return per annum is called the "rate per cent." Thus, if P stands for the principal, T stands for the time in years, R stands for the rate per cent per annum, and I stands for the interest, then

$$I = \frac{PRT}{100}$$

where P and I must be in the same monetary units. This formula can be transposed to give P, R, and T in terms of the other letters:

$T = \frac{100I}{PR}$,
$R = \frac{100I}{PT}$, and
$P = \frac{100I}{RT}$.

Compound interest is different from simple interest in that the interest which is added also attracts interest. If a sum of P monetary units is invested at $r\%$ per annum for n years, then the value or amount after n years is

$P\left(1 + \frac{r}{100}\right)^n$.

For instance, the value of $\$2{,}500$ invested at 5% compound interest after eight years (i.e., $P = \$2{,}500$, $r = 5$, and $n = 8$) will be

$P\left(1 + \frac{r}{100}\right)^n = \$2{,}500\left(1 + \frac{5}{100}\right)^8 = \$3{,}693.$

The mathematical formula of compound interest and regular deposits, which underpins banking transactions, can be formulated as follows: assume that you borrow an amount P of money (the "principal") at an (annual) interest rate of $r > 0$, and that, at the end of each year, you have to pay back a fixed amount (a "deposit") d. Let A_n be the total amount of money owed after n years. The formula for computing A_n in terms of P

(the principal of the loan), r (the interest rate of the loan), and d (the loan deposits) is the following (where $0 < r \leq 1$, e.g. $5\% = 0.05$):

$$A_n = A_{n-1}(1+r) - d$$
$$= P(1+r)^n - d(1+r)^{n-1} - d(1+r)^{n-2} - \cdots - d$$

$= P(1+r)^n - d[(1+r)^{n-1} + (1+r)^{n-2} + \cdots + (1+r) + 1]$, which includes a geometric series, and $r \neq 0$; so that the initial condition is $A_0 = P$; at the end of the first year, you owe P (the principal) plus an interest equal to rP minus the deposit you have agreed to pay each year. Therefore, $A_1 = P + rP - d = P(1+r) - d$; by analogy, at the end of the second year, you owe $A_2 = A_1(1+r) - d = P(1+r)^2 - d(1+r) - d$, etc. By allowing the owners of large sums of money to lend (that is, trade) money on interest, we give them power to immunize themselves against loss (in fact, this is the ultimate purpose of charging interest on loans: to immunize the lender of money against loss), while socializing loss and risks. Therefore, if the level of interest is formed spontaneously, merely as a result of competition, it may give rise to important irrationalities and inefficiencies, and it can lead to the establishment of an exceptionally privileged financial oligarchy. As against this situation, a rational and scientifically rigorous Central Economic Planning Authority (CEPA) should systematically control and guide the banking system in order to implement the optimal financial and monetary policy in accordance with the CEPA's economic plan. In other words, the CEPA must ensure that credit is connected with the CEPA's policy for the rational organization of the resources of enterprises and for the achievement of rational settlements between enterprises.

The net present value (NPV) of an investment project consists of calculating the amount by which the value of that investment project exceeds its cost. If i is the cost of capital (which, for convenience, is assumed to be fixed for the project under consideration), then the NPV is defined as follows:

$$NPV = \frac{X_1}{1+i} + \frac{X_2}{(1+i)^2} + \cdots + \frac{X_n}{(1+i)^n} - C_0$$

where X_t $(t = 1,2,\ldots,n)$ denotes the cash flow that corresponds to year t, C_0 is the capital cost of the investment project in year 0, and n is the lifetime (in years) of the investment project. Notice that, in the NPV formula, i is the "discount rate," that is, the company's cost of capital (specifically, the company's interest rate and loan payments or dividend payments to shareholders); when a company uses both debt and equity to fund operations, i is the weighted average cost of capital. Hence, according to the Italian-American economist Franco Modigliani (who was awarded the Nobel Prize in Economics in 1985) and the American economist Merton Miller, under certain conditions (in particular, if we assume that there is total information transparency and total rationality), the intrinsic or real value of a company can be considered to be the net present value of all the investment projects of that company. Furthermore, if we divide the intrinsic or real value of a company by the total number of outstanding shares issued by that company, we can find the real or intrinsic value per share (for the given company).

Whereas the term "stock" means a share in the ownership of a company, the term "bond" means debt. In fact, a bond is a debt instrument issued for a period of more than one year with the purpose of raising capital by borrowing. By the term "maturity," we mean the date on which a debt becomes due for payment. The "face value" (also known as the "par value" or "principal") is the amount of money a holder of a fixed income security will receive back once the given security matures. The "coupon" is the amount that a holder of a fixed income security will receive as interest payments. The coupon is expressed as a percentage of the par value. "Yield" is a figure that shows the return one gets on a bond.

$$Current\ Yield = \frac{coupon\ amount}{market\ price}$$

(when we buy a bond at par, yield is equal to the coupon, and, when price changes, so does the yield). For instance, suppose that a bond has a par value of $\$1{,}000$ and that its coupon rate is equal to 6%. Since the market price of a bond changes, an investor may purchase a bond at a discount (i.e., less than par value) or a premium (i.e., more than par value). In particular, if an investor buys this 6% coupon rate bond for a discount of $\$900$, then the investor earns an annual interest income of $(\$1{,}000 \times 6\%) = \$60$, and the current yield is $\$60/\$900 = 6.67\%$. Notice that the annual cash flow of $60 is fixed, regardless of the price paid for the bond. On the other hand, if an investor buys this 6% coupon rate bond at a premium of $\$1{,}100$, then the investor earns again an annual interest income of $(\$1{,}000 \times 6\%) = \$60$, but, in this case, the current yield is $\$60/\$1{,}100 = 5.45\%$.

A "zero-coupon bond" is a type of bond that makes no coupon payments but, instead, is issued at a considerable discount to par value. For instance, a zero-coupon bond with a $\$1{,}000$ par value and ten years to maturity might be trading at $\$600$. In case of a zero-coupon bond,

$$Y = \left(\frac{M}{P}\right)^{1/N} - 1$$

where Y denotes the yield to maturity, M denotes the value of the given zero-coupon bond at the time of maturity (i.e., the par value), P denotes the current price of this bond (present value), and N denotes the years to maturity.

In general, as we have seen, in bonds, when price goes up, yield goes down, and vice versa. The factor that influences a bond more than any other is the level of prevailing interest rates in the economy. When interest rates rise, the prices of bonds in the market fall, and, thus, we see an increase in the yield of the older bonds, which are brought into line with the newer bonds being issued with a higher coupon. On the other hand, when interest rates fall, the prices of bonds in the market rise, thereby lowering the yield of the older bonds and bringing them into line with the newer bonds being issued with a lower coupon. Moreover, another important factor that influences a bond is the issuer's default risk. In fact, investors try to determine if the bond rating agencies are going to change the issuer's rating. Rating changes may be prompted by changes in such factors as: financial ratios, Gross National Product, inflation, etc.

In 1911, the American economist Irving Fisher expressed the "quantity theory of money" in what is known as the equation (actually, identity) of exchange:

$$MV = PQ$$

where M is the quantity of money in the economy, V is the velocity of the circulation of money (i.e., the amount of nominal Gross National Product each year divided by the money stock), P is the general price level (i.e., the average value of each transaction), and Q is aggregate output (i.e., the physical volume of transactions during the given time period, so that $Gross\ National\ Product = PQ$). Thus, according to Fisher, if we assume that, at least in the short-run, both V and Q are constant (given that the velocity of circulation is determined by institutional factors, such as the payments interval for wages, and Q is determined by the productive capacity of the economy), then a change in the money supply, M, results in an equal percentage change in the price level P.

The previous equation implies that

$$M = \frac{PQ}{V}$$

and, since V is (assumed to be) constant, $1/V$ can be replaced by a constant k. Additionally, when the money market is in equilibrium, the demand for money, M_d, is equal to M. Hence,

$$M_d = kPQ$$

which means that, according to Fisher's model, the demand for money is a function of income and does not depend on interest rates.

However, in practice, the velocity of the circulation of money, V, is not constant, even in the short-run, and, especially, during periods of recession. Therefore, the English economist John Maynard Keynes extended Fisher's equation of exchange by pointing out that there are three motives for holding money: (i) Transactions motive: money is a medium of exchange, and, as income rises, people conduct more transactions and hold more money. (ii) Precautionary motive: people hold money for emergencies, and money demand is again expected to rise with income. (iii) Speculative motive: money is also a way for people to store wealth, and, under the speculative motive, the demand for money is negatively related to the interest rate. Moreover, Keynes modeled the demand for money as the demand for the real (as opposed to the nominal) quantity of money (real balances), M/P. According to Keynes, the demand for real money balances is a function of both income and interest rates:

$$\frac{M}{P} = f(Q, i)$$

where Q denotes output or income and i denotes the interest rate (and, hence, the velocity of the circulation of money fluctuates with the interest rate).

The level of interest rates can indeed be treated as a monetary target, but it is important to determine the extent to which interest rates are a major factor in decisions of either businesses, consumers, or governments. For instance, if an economy is characterized by important structural inefficiencies, then an increase in the supply of money (other things equal), instead of boosting economic growth, may lead to an increase in inflation and money incomes.

Moreover, it is worth mentioning that central banks have at their disposal a number of policy instruments that can affect certain intermediate targets, such as the money supply, interest rates, etc. The three major instruments of monetary policy are:

i. *Open market operations:* this is the activity of a central bank in buying or selling government bonds to influence the money supply, interest rates, and bank reserves. In fact, if securities are bought (by the central bank), the money paid out by the central bank increases commercial-bank reserves, and the money supply increases. On the other hand, if securities are sold (by the central bank), then money supply decreases.

ii. *Discount-rate policy:* given that the discount rate is the interest rate charged by the central bank on a loan that it makes to a commercial bank, it follows that the central bank can increase the discount rate to reduce the money supply, whereas the central bank can reduce the discount rate to increase the money supply.

iii. *Reserve-requirements policy:* by the term "required reserves," we mean that portion of deposits that a bank sets aside in the form of vault cash or non-interest-earning deposits with the central bank. In fact, if the central bank wants to tighten money overnight, then it can raise reserve requirements, whereas, if the central bank wants to ease credit conditions (and, thus, increase the money supply), then it can cut reserve requirements.

Economic Cycle, Economic Crisis, and Political Economy

By the term "economic cycle," or "business cycle," we refer to economic fluctuations between periods of economic expansion and economic contraction. In other words, an economic cycle is the circular movement of an economic system as it moves from expansion to contraction and back again. The four stages that characterize the economic cycle (or business cycle) are the following:

i. *Expansion:* during this stage, the economy experiences relatively rapid growth, interest rates tend to be low, and production increases. The economic indicators associated with growth (e.g., employment and wages, business profits and output, aggregate demand, and the supply of goods and services) tend to increase through the expansionary stage, but, at some point, the increase in the money supply may cause inflationary pressures.

ii. *Peak:* during this stage, growth hits its maximum rate, and prices and economic indicators may stabilize for a short period of time before they start to decrease.

iii. *Contraction:* during this stage, growth slows, employment decreases, and prices stagnate. As demand decreases, businesses may not immediately adjust production levels, causing a situation characterized by oversupply and falling prices. If the contraction of economic activity continues, then it may turn into a "recession."

iv. *Trough:* during this stage, the economy hits a minimum point, with supply and demand hitting bottom before recovery.

By the term "economic crisis," in general, we refer to sudden interruptions in the (re)production of the economy. Irrespective of the romantic aspects of Karl Marx's communist vision, which are irrelevant to my work and thought, Karl Marx, under the influence of British and German scientific economic theories, articulated a brilliant critical method of understanding and restructuring social reality. Irrespective of the romantic, egalitarian teleology of Marx's thought, of which I disapprove, I maintain that Marx's analysis of capitalism is very useful in order to overcome class conflict towards a hierarchical and scientifically organized society, in line with Plato's vision and the principles of cybernetics. Additionally, I would like to mention that, in contrast to romantic populism, Vladimir Lenin (1870–1924) conceded that, historically, scientific socialism comes from educated intellectuals, from an elite, and the reason why it cannot easily spread among the working class is that it is too complicated, and, therefore, he highlighted the importance of education. Furthermore, in view of Darwin's theory of evolution, we have to keep in mind that, since man is a thinking ape who developed civilization as a negation of nature (in a Promethean sense) and as a manifestation of the intentionality of human consciousness, humanity's attempt to transcend its ape origins and to rise to better and better and more successful levels of being is an open process, a constant quest, and a constant existential gamble.

The way I have delineated my conception of the Central Economic Planning Authority (CEPA) and its specific roles and goals suggests that I am proposing a specific model of market socialism, where markets for goods, services, money, and capital exist but are thoroughly controlled, systematically directed, and totally constrained by a powerful ruling scientific elite in the context of a vertical-hierarchical political system based on Plato's vision of the ideal republic.

Marx has correctly and mathematically rigorously analyzed the endogenous causes of the capitalist crises. I shall briefly present Marx's explanation of economic crises, and afterwards I shall briefly present a scientifically rigorous, "technocratic" approach to economic planning.

Karl Marx, in his seminal book *The Capital* (published as three volumes in 1867, 1885, and 1894), articulated a structural theory of economic crises in the capitalist system. He started from the "labor theory of value," which was originally formulated by the Scottish economist and philosopher Adam Smith and the British political economist and politician David Ricardo. According to the labor theory of value, only people can *create* value. Machines or production charts *have* value (in particular, they have use value), in the sense that they are useful things, but machines and production charts on their own cannot do anything, until someone does something with them.

Marx thought as follows: Let us divide the average working day into two component parts, V and S, where V denotes "variable capital," and S denotes "surplus value." Variable capital means that proportion of capital which is equal to the cost of labor, so that it is invested in wages (V is sufficient for the purchase of labor power). Surplus value means the additional capital that is produced, and it is the source of profit (i.e., the accumulated product of the unpaid labor time of the producers). The ratio

$$\frac{S}{V}$$

is called the "rate of surplus value." Marx defined the "rate of profit" as follows:

$$\frac{S}{C+V}$$

where S is the surplus value, V is the variable capital (i.e., the wages paid for the production of a commodity), and C is the constant capital. By the term "constant capital," Marx means the value of goods and materials (e.g., machinery, raw materials, etc.) required to produce a commodity (Marx used the term "constant capital" in accordance with the labor theory of value). Thus, Marx formulated the "law of the tendency of the rate of profit to fall" over an economic cycle due to competition.

A capitalist (or, generally, an owener of a business) cannot fritter away the entire surplus value in luxury expenditure, but he/she has to reinvest a significant proportion of the surplus value in order to protect himself/herself from competition and in order to get an advantage over competitors. Moreover, because of technical improvements (such as machinery), represented by C (the "constant capital"), the productivity of labor increases. Over time, as a capitalist invests, the ratio

$$\frac{C}{V}$$

increases (more and more machinery is working together with the individual laborer). Hence, there is a tendency to replace living labor (i.e., V) with "dead labor" (i.e., C). If we divide both the numerator and the denominator of the rate of profit by V, we obtain

$$\frac{\frac{S}{V}}{\frac{C}{V}+\frac{V}{V}}=\frac{\frac{S}{V}}{\frac{C}{V}+1}$$

where the ratio C/V increases over an economic cycle, as we have just explained. Therefore, if the ratio S/V increases at a lower speed than the ratio C/V, then the rate of profit decreases. This result is Marx's "law of the tendency of the rate of profit to fall" over an economic cycle due to competition. It goes without saying that there are counter-acting influences that can delay the onset of this effect (e.g., through technology, imperialism, foreign trade, labor intensification, an expansionary monetary policy, etc.), but the underlying structure of the economy is reflected by Marx's "law of the tendency of the rate of profit to fall" over an economic cycle due to competition. In other words, capitalist crises are due to the very nature, the intrinsic contradictions, of the capitalist system. During capitalist crises, business bankruptcies and consolidations occur, so that C decreases, and the "reserve army of unemployed" increases. The increase in unemployment causes a downward pressure on the level of wages, and, therefore, V decreases. As C and V decline, the rate of profit recovers, and a new cycle of accumulation begins until a new crisis occurs, and so on.

In view of the foregoing, a policy of economic planning is necessary. My conception of a Central Economic Planning Authority (CEPA), as I described it earlier in this chapter, represents an updated, modern version of Plato's vision of a republic ruled by the "epaiontes" (i.e., "those with real understanding," the "genuine experts," "those who perceive things according to their nature"). Advances in mathematics and technology combined with an aristocratic ethic can make this ideal practical, consistent, and effective. Thus, Plato's political theory should be merged with cybernetics, which reflects a conception of a "universal organizational science," which would be capable of combining and coordinating all the individual scientific disciplines. Cybernetics is a transdisciplinary systematic study of regulatory and purposive systems (their structures, constraints, and possibilities). Hence, cybernetics has been defined as "the art of governing or the science of government" (according to the French physicist and mathematician André-Marie Ampère), "the art of steersmanship" (according to the English psychiatrist Ross Ashby), "the study of systems of any nature which are capable of receiving, storing, and processing information so as to use it for control" (according to the Soviet mathematician Andrey Kolmogorov), "the science and art of the understanding of understanding" (according to Rodney E. Donaldson, the first president of the American Society for Cybernetics), "the art of creating equilibrium in a world of constraints and possibilities" (according to the American philosopher Ernst von Glasersfeld), as well as "a branch of mathematics dealing with problems of control, recursiveness, and information, focuses on forms and the patterns that connect" (according to the English anthropologist and linguist Gregory Bateson).

The sixth-century B.C.E. Ionian Greek philosopher and mathematician Pythagoras was, arguably, the first person who called himself a "philosopher." In particular, Diogenes Laertius, in his *Lives of Eminent Philosophers* (Book VIII, Chapter 1: Pythagoras, 8) writes the following:

> Sosicrates in his *Successions of Philosophers* says that, when Leon the tyrant of Phlius asked him [namely, Pythagoras] who he was, he said, "A philosopher," and that he compared life to the Great Games, where some went to compete for the prize and others went with wares to sell, but the best as spectators; for similarly, in life, some grow up with servile natures, greedy for fame and gain, but the philosopher seeks truth.

Moreover, Diogenes Laertius, in his *Lives of Eminent Philosophers* (Book V, Chapter 1: Aristotle, 20) writes that, when Aristotle was asked what advantage he had ever gained from philosophy, Aristotle's response was the following: "This, that I do without being ordered what some are constrained to do by their fear of the law." From Aristotle's perspective, philosophy—expressing a continuous and systematic quest for knowledge, which is dialectically directed towards the ultimate knowledge—enables one to understand the underlying order and harmony of the world and, thus, to act rationally without coercion.

Philosophy being the most general approach to knowledge and truth, Plato, in his *Republic*, aptly proposed a model of polity based on the concept of the "philosopher king," a theoretical ruler who combines philosophical knowledge and temperament with political skill, power, and authority. This political vision, being based on the supreme and noblest epistemological, moral, and aesthetic values, and not on particular economic/social interests, aims at the scientifically and morally optimal organization and governance of human beings and at the guidance of science and education by philosophy, by a ruling philosophical elite, and not by self-interested individual social actors, or the capitalist class, or irrational passions. Philosophy enables one to reason and argue in the most abstract, the most comprehensive way and to consciously choose a value system and, thus, a way of life and a type of humanity. Consequently, a genuinely philosophical mindset is a necessary prerequisite for genuine political leadership and statesmanship. As Plato has correctly argued, politics separated from philosophy is a counterfeit of politics.

The constitutive and the regulative rules of a polity shape a dominant ethos, which differentiates a genuine political community from any coalition of self-interested actors; and genuine thought, that is, thought as understood in the context of science and philosophy (which is a reflection on science), is the source of correct and optimal rules. My thinking on these issues shares the conviction of the philosopher Giuliano Di Bernardo, who, in his books *Liberalismo contro Totalitarismo* and *The*

Future of Homo Sapiens, argues for an updated variety of enlightened, genuinely aristocratic (that is, spiritually aristocratic) totalitarianism. The vertical and technocratic hierarchical system that we propose should not be confused with other historical models, such as those of the tyrant, the dictator, the monarch, and any similar models, because it is analogous to the government of philosophers delineated by Plato in his *Republic*.

Scientific Totalitarianism: A Theme in Need of a Focus

The ancient Greek *polis* (city-state) had a unique, distinctive characteristic on the basis of which and due to which the institution of the *polis* was differentiated from other forms of organized collective behavior, and it gave rise to the notions of political art, political virtue, and political science. The unique, distinctive characteristic of the ancient Greek *polis* consisted of a collective attempt to institute a community whose *telos*, or existential purpose, was to live in harmony with the principle of truth, as we read in Aristotle's *Nicomachean Ethics*, X. From the aforementioned philosophical perspective, we can talk meaningfully about "politics" and "civilization" only when the ultimate goal of collective life is "truth," which, according to Plato and Aristotle, implies the imitation of true being, that is, of that mode of existence which is free from corruption, alterations, and annihilation (Plato, *Republic*, II, IV, VII, and X; Aristotle, *Nicomachean Ethics*, II–VI). This is the reason why logic and, especially, the kind of knowledge that is represented by mathematics play a key role in Plato's and Aristotle's thought.

In the context of Plato's and Aristotle's philosophical works, genuine politics refers to an existential goal of the human being, and, therefore, genuine politics is a collective struggle that is aimed at the truthfulness of human existence. In other words, the *telos* of politics is to enable humanity to exist authentically through and within a social system. This aspiration is the core of classical Greek political thought.

In order to clarify the arguments that genuine politics consists of the pursuit of truth and that truth consists of the imitation of true being, we need to understand the meaning of "truth" (in Greek, "aletheia") and "reason" (in Greek, "logos") in the context of classical Greek thought. In terms of the Greek word "aletheia," everything that exists is manifested as an entity in the world, that is, the truth of anything/anyone is ultimately determined by its/one's participation in the logical constitution of the world, and the Greek term "logos" refers to the disclosure of this fact. The event of disclosure speaks about and declares the existence of an entity in the world, and it refers to a conscious being that is aware of the event of disclosure. Hence, truth emerges from the relationship between a disclosed entity and the viewer and agent of this disclosure; and “logos” is the event of disclosure and the elucidation of the way in which disclosure takes place.

The “existent” is disclosed through its form, or species, that is, through its distinctive way of being. For instance, the form of a robot “says” to its viewer that the given object is a robot (and not, for instance, a flower). However, the Greek term “logos” refers not only to the individual form of each being or thing that exists in the world, but also to the overall configuration of the world, that is, to the way in which beings and things that exist in the world relate to each other. This is the reason why the Greek term “logic” derives from the Greek term “logos.” Furthermore, according to ancient Greek aesthetics, the overall formation of the entities that exist in the word has “kallos,” which means beauty, as we read in Plato’s works *Timaeus* (29a–d, 47b–c), *Republic* (443d, 500c), *Phaedrus* (246–251, 247c–d), and *Laws* (734a–741a), as well as in Aristotle’s works *Physics* (265a25), *Politics* (1289b25), and *Nicomachean Ethics* (1181b21). The Greek noun “kallos” (beauty) is semantically related to the Greek verb “kalo” (*καλῶ*), meaning “attract” and “invite.” By viewing and contemplating the way of the overall formation of the entities in the world, ancient Greek philosophers identified the harmony and, hence, the beauty of the world. Therefore, they called the universe “cosmos,” which, in Greek, is semantically related to the Greek noun “cosmema,” meaning “jewel,” “ornament,” and “embellishment.”

The “logos” of the entities that exist in the world consists of the way in which they participate in the corresponding species/form and of the way in which they relate to each other in the context of the cosmic harmony and order. The “logos” of the cosmic entities that belong to the same species/form is common to all of them, and it is incorrigible and eternal, independent of the characteristics of particular entities. For instance, every particular rose and every particular lion will perish, and, eventually, they will be annihilated. But the form of a rose, namely, its “logos,” or the way of its participation in existence, which makes it what it is (the given plant), and the form of a lion, its “logos,” or the way of its participation in existence, which makes it what it is (the given animal), are not susceptible to corruption, but they are incorrigible and eternal. Moreover, the set of the fundamental relations (i.e., the structure) in which every particular plant and every particular animal participate (e.g., the way of a plant’s sowing, vegetation, and blossoming, and the way of an animal’s birth, development, and reproduction) is an integral, incorrigible, and eternal whole. Hence, “logos” means participation in the corresponding (eternal and incorrigible) form that makes existents what they are as well as participation in the formation of the entire cosmos; and this idea underpins my conception of scientific totalitarianism in general as well as my conception of a hierarchical, organic society in particular. As I have already mentioned, it goes without saying that I reject every variety of totalitarianism that is based on biological racism, chauvinism,[3] religion, romanticism, and/or particular class interests. Moreover, I am aware of the traumatic experiences left in Europe by attempts by essentially irrelevant persons to pursue totalitarian politics. I advocate a concrete vision of scientific totalitarianism based on Plato's political thought, interdisciplinary mathematics, and epistemology.

True being, that is, the way of being eternal and incorrigible, is the event of participation in the "logos," and, therefore, it is clear what one must do if he/she "seeks . . . to be immortal" (Plato, *Symposium*, 207d1–2): he/she must imitate the "logos" of the relations of participation in the formation of the cosmos. For instance, he/she must understand and organize society as an event of participation in the order, the harmony, and the decency of the relations that constitute the eternal cosmic beauty. This is the essence of my conception of scientific totalitarianism in general as well as of my conception of a hierarchical, organic society in particular.

[3] Regarding chauvinism, in particular, it should be mentioned that, in the twentieth century, it was reinforced by English, German, and American geopolitics (e.g., by such geopoliticians as Halford John Mackinder, Karl Ernst Haushofer, and Nicholas J. Spykman). It is worth mentioning that G. Nicolas and C. Guanzini have used the evocative symbolism of the Second Horseman of the Apocalypse in order to articulate their argument that even the most renowned and venerated of geopolitical theorists and political geographers had "attempted to vindicate war through teaching the love of the Mother Earth" which was most powerfully and emotionally expressed in an aggressive, nationalist love of the Mother Country. For more details, see: G. Nicolas and C. Guanzini, "Ancient History for the Future: The Political Role of Geography," Video English Version G. Parker, University of Lausanne, Ératosthène, 1993. Additionally, geopolitics has been used by particular Western bureaucracies in order to undermine Russia's imperial, multiethnic tradition and structure, and, during the Cold War, geopolitics was also used as an ideological weapon against the Soviet bloc and, generally, against the cosmopolitan aspect of socialism. Finally, it should be mentioned that White Russian émigrés (i.e., Russians who emigrated from the former Russian Empire in the wake of the Bolshevik Revolution (1917) and the Russian Civil War (1917–23), and who were in opposition to the Bolsheviks) developed a peculiar variety of chauvinism and fascism by combining Western theories of geopolitics, mysticism, and religious doctrines (the Russian intellectual Ivan Ilyin is a characteristic representative of this ideological current).

Chapter 5
Probability and Statistics

First of all, it should be clarified that, by the term "quantitative analysis," we mean the study of phenomena by means and on the basis of any type of quantitative information. Such an inquiry takes place by applying suitable methods that determine the nature of the available information and the phenomena under consideration. Quantitative methods mainly include methods that derive from mathematical analysis, mathematical programming, probability theory, and statistics.

In fact, statistics emerged from the constant efforts of humankind to deal with situations of uncertainty in which they lived. In these situations, the element of luck always appeared as a key determining factor which prevented the identification of the existence of systematicness in the manifestations of various phenomena and in the formulation of relations between them. Aristotle was the first philosopher to offer a systematic account of "luck" and to include it as a significant topic in both physics and ethics (Aristotle, *Physics*, 2:4–6, and *Metaphysics*, 7:7–9). A method is called statistical if it relates facts and hypotheses of some kind. Hence, statistics investigates and develops methods for evaluating hypotheses in reference to empirical facts.

In general, luck is involved in all things where actors do not hold full control over the outcome of action. One of the basic attributes of the statistical method is the fact that it refers to properties of populations instead of individual cases. Statistics examines a unit only in its capacity as a member of a population. The statistical method can be applied in order to solve any problem related to the definition of overall behavior, based on individual observations expressed numerically. The concept of luck is commonly used in statistics in order to display all the possible outcomes given a very large sample and the probability of each outcome. In science, "probabilities," often called chances or stochastic processes, are relative frequencies in series of events, or tendencies or propensities in the systems that give rise to those events. By the term "frequency," we refer to the number of times each measurement occurs.

Probability theory is primarily concerned with the issue of uncertainty. In fact, "probability," usually denoted by p, is a quantitative measure of uncertainty. It is a number between 0 and 1, where 0 indicates impossibility and 1 indicates certainty. Assume that we take any very large number, N, out of a series of cases in which an event, A, is in question, and that A happens on pN occasions (where $0 \leq p \leq 1$). The probability of the event A is said to be p (the certainty of the corresponding proposition increases as the number N of specimen cases selected increases). Furthermore, the following corollaries and extensions may be added to the aforementioned definition of a probability: (i) If the probability of an event is p, then, out of N cases in which it is in question, it will happen pN times, where N is any very large number (where $0 \leq p \leq 1$). (ii) If the probability of an event is p, then the probability of its failing is $1 - p$.

Probability theory is based on set theory. By the term "experiment," we mean a process that leads to one of several possible outcomes. By the term "outcome," we mean an observation or measurement. The "sample space" is the set of all possible outcomes of an experiment. An "event" is a subset of a sample space—or, in other words, a set of basic outcomes. Thus, we say that the event "occurs" if the corresponding experiment gives rise to a basic outcome belonging to the event. Therefore, we obtain the following formula:

$$Probability\ of\ event\ A = \frac{n(A)}{n(S)},$$

where $n(A)$ is the number of elements in the set of the event A, and $n(S)$ is the number of elements in the sample space S. For instance, roulette as it is played in Las Vegas or Atlantic City consists of a wheel that has 36 numbers, numbered 1 through 36, and the numbers 0 and 00 (double zero). Therefore, in this case, the sample space, S, consists of 38 numbers, and the probability of winning a single number that you bet is $P = 1/38$.

When the sets corresponding to two events are disjoint (their intersection is the empty set), then these events are called "mutually exclusive."

The axiomatic definition of probability is the following: Let E be a space of elementary events (i.e., the space of outcomes of experiments, or the space of states of a system, since the state of a system can be construed as the outcome of an experiment). The "probability of an event" $A \subseteq E$ is denoted by $p(A)$, and it is defined as a single number that corresponds to A and has the following properties:

(P1) $p(A) \geq 0$;

(P2) for each pair of mutually exclusive events, $A, B \subseteq E$, it holds that
$p(A \cup B) = p(A) + p(B)$;

(P3) $p(E) = 1$ (i.e., the total probability, after adding all possibilities, is equal to one).

Remark: For each $A, B \subseteq E$, $p(A \cup B) = p(A) + p(B) - p(A \cap B)$; but, in case A and B are mutually exclusive, it holds that $p(A \cap B) = 0$, so we obtain (P2).

By the term “conditional probability,” we mean the probability of event A conditional upon the occurrence of event B. Assume that we investigate the probability of an event A given that we know that an event B has occurred, and that event B influences the probability of event A. The “conditional probability” of event A given the occurrence of event B is defined as the quotient of the probability of the intersection of A and B over the probability of event B; symbolically:
$P(A|B) = \frac{P(A\cap B)}{P(B)}$,
where $P(A|B)$ denotes the probability of A conditioned on B, and we assume that $P(B) \neq 0$. The aforementioned formula for the computation of conditional probability is known as Bayes’s Law, since it was originally formulated by the eighteenth-century English statistician and philosopher Thomas Bayes. Notice that A is independent of B if $P(A|B) = P(A)$; that is, knowing that B occurred does not change the probability that A occurred. Thus, according to Bayes’s Law, two events A and B are independent of each other if and only if
$P(A \cap B) = P(A)P(B)$.
Bayes’s Law provides a method of revising existing predictions or theories (specifically, updating probabilities) given new additional evidence. In fact, Bayes’s Law implies that the interpretation of any risk assessment depends on an estimate of the base rate, and the corresponding base rate, which is never known with complete certainty at the time of the assessment, is a Bayesian “prior probability.”
Probability theory has several significant applications in the natural sciences and in the social sciences. For instance, in genetics, probability is a measurement tool that helps us to predict the chances of an offspring being inherited with a particular trait of interest (assuming Mendel’s laws of inheritance). The sum law helps us to find the probability of two or more events occurring as long as they are mutually exclusive: the probability of the occurrence of one event or the other, of two mutually exclusive events, is the sum of their individual probabilities; that is, if A and B do not share any outcome, then $p(A \cup B) = p(A) + p(B)$. The product law helps us to find the probability of two or more events occurring as long as they are independent of each other: if A and B are independent of each other, then $P(A \cap B) = P(A)P(B)$. Moreover, probability theory helps us to estimate the chances of success or failure of a business project, an investment, or product launch.
By the term “random variable,” we refer to a function from the outcomes of an experiment to the set of real numbers. A “probability distribution function” specifies the probabilities associated with the values of the

corresponding random variable. The "expected value" of a discrete random variable X is denoted by $E(X)$, and it is defined as follows:

$$E(X) = \sum x_i p(x_i)$$

where x_i denotes the ith value of the random variable X ($i = 1,2,3,\ldots$), $p(x_i)$ denotes the probability of x_i, and the symbol $\sum$ denotes the sum of all products $x_i p(x_i)$.

One of the most important methods that is used to discover, describe, and explain "typical" behavior of mass data is the "arithmetic mean." The formula is

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$$

where $\bar{X}$ denotes the arithmetic mean, $\sum_{i=1}^{N} X_i$ denotes the summation of the values of the individual observations X_i under consideration ($i = 1,2,\ldots,N$), and N is the total number of items in the series that have been summated. It is worth noticing that arithmetic means are often "weighted" averages, in the sense that, when averaging values, it is sometimes logically necessary to assign more importance to some than to others (by multiplying each value with a suitable statistical weight), so that particular values may be more influential in determining the "typical" value than others. Formally, the weighted arithmetic mean of a non-empty finite set of data $\{X_1, X_2, \ldots, X_N\}$ with corresponding non-negative weights $\{w_1, w_2, \ldots, w_N\}$ is

$$\bar{X} = \frac{\sum_{i=1}^{N} w_i X_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 X_1 + w_2 X_2 + \cdots + w_N X_N}{w_1 + w_2 + \cdots + w_N}$$

(the weights can be in the form of decimals, whole numbers, percentages, etc.). For instance, if $x_1, x_2, x_3, \ldots$ are the measured observations and $f_1, f_2, f_3, \ldots$ are the corresponding frequencies, then the arithmetic mean is

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + f_3 x_3 + \cdots}{f_1 + f_2 + f_3 + \cdots}$$

(this is the arithmetic mean of a frequency distribution). Moreover, notice that a consumer price index (CPI) is typically calculated as a weighted average of the price change of the goods and the services covered by the index (in this case, the weights are meant to reflect the relative importance of the goods and the services as measured by their shares in the total consumption of households).

Weighted aggregative price index: Firstly, having chosen a base year, we obtain the prices of a list of commodities and some measure of the importance of each commodity that is relevant to the purpose of the index. The importance may be measured by the quantity of each commodity sold, consumed, or produced (other weights may be devised in special situations). The data used for weighting purposes must refer to the base period. Secondly, we multiply each of the commodity prices by the corresponding weight. Thirdly, we summate the so obtained values for each time period. Fourthly, we divide each total by the base total, and we multiply by 100 to reduce the index to percentage form. If we denote the base year quantities used for weights by $q_0, q_0', q_o'', \dots, q_o^n$, then the aforementioned method of computing a weighted aggregative price index is given by the following formula, which is known as the "Laspeyres Price Index":

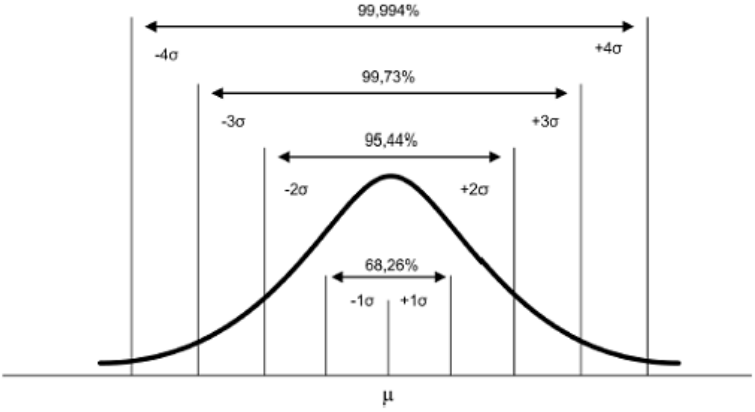
$$
\begin{aligned}
&Weighted\ aggregative\ price\ index\\
&= \frac{p_1 q_0 + p_1' q_0' + p_1'' q_0'' + \cdots + p_1^n q_0^n}{p_0 q_0 + p_0' q_0' + p_0'' q_0'' + \cdots + p_0^n q_0^n} \times 100\\
&= \frac{Sum\ of (Price\ at\ observation\ period \times Base\ quantity)}{Sum\ of\ (Price\ at\ base\ period \times Base\ quantity)} \times 100
\end{aligned}
$$

(do not be confused by the notations: the numerator is simply the total expenditures for all commodities at the observation period using base quantities, and the denominator is simply the total expenditures for all commodities at the base period using base quantities).

Whereas the mean is the average value of a set of data, the "median" is the middle value in a set of data. Hence, in order to find the median of a data set, one must arrange the data from least to greatest, and find the data point located in the middle: if there is an odd number of data, then the median is the middle point in the array, but, if there is an even number of data in the array, then the median is the average of the two middle data points in the array. The "mode" is the value that appears most frequently in a set of data.

As I have already mentioned, by the term "probability distribution," we mean a statistical function that describes all the possible values and likelihoods that a random variable can take within a given range. A probability distribution is called a "normal distribution," or a "Gaussian distribution," if it is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean (as shown in Figure 5-1). In the normal distribution, its mean (average), median (midpoint), and mode (most frequent observation) are all equal to each other; and these values all represent the peak, or highest point, of the distribution. In graphical form, the normal distribution appears as a "bell curve," as shown in Figure 5-1. In other words, the "normal curve" is bell-shaped and perfectly symmetric (centered on the mean).

Figure 5-1: A normal (or Gaussian) distribution (source: Wikimedia Commons: Author: Thais Monteiro Peres; https://commons.wikimedia.org/wiki/File:Curva_Gaussiana.png).



One of the most important methods that are used to discover, describe, and explain “risk” or “uncertainty” is the “standard deviation,” which is a quantity expressing by how much the members of a database (i.e., the data under consideration) differ from the arithmetic mean of the given database. The formula of the standard deviation is:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N}}$$

where: firstly, we calculate the arithmetic mean $\bar{X}$ of the values X_i $(i = 1,2,\dots,N)$ under consideration; secondly, we record the deviation of each value X_i from the arithmetic mean, namely, $x_i = X_i - \bar{X}$; thirdly, we square these deviations (we compute x_i^2); fourthly, we summate the squared deviations and divide by N (thus finding the “variance” of our data); fifthly, we extract the square root to obtain σ (i.e., $standard\ deviation = \sqrt{variance}$). However, the aforementioned formula for the standard deviation is used when N is the entire population of the species or kind under consideration; if we do not have the entire population, we use the following formula for the standard deviation:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

where n is the size of the sample (i.e., the number of the point data that are contained in the database that we use), X_i is the ith point of the sample ($i = 1,2,\dots,n$), and $\bar{X}$ is the arithmetic mean of the sample (namely, of the database that we use).

The normal curve's standard deviation tells us what percentage of observations falls within a specific distance from the mean. As shown in Figure 5-1, when we have a normal curve, the area below the curve contains 100% of all observations; approx. 68% of all observations fall within one standard deviation from the mean; approx. 95% of all observations fall within two standard deviations from the mean; and approx. 99% of all observations fall within three standard deviations from the mean.

When we have two sets of data and we want to find how strong a relationship is between them, we use "Pearson's correlation coefficient" (PCC), also known as Pearson's r. In other words, PCC calculates the level of change in one variable due to the change in the other. When applied to a sample of the variables x and y, PCC is commonly represented by r_{xy}.

Given paired data

$\{(x_1, y_1), \dots, (x_n, y_n)\}$, consisting of n pairs, r_{xy} is defined as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

n is the sample size,

x_i are the values of the x-variable in the sample,

$\bar{x}$ is the mean of the values of the x-variable,

y_i are the values of the y-variable in the sample, and

$\bar{y}$ is the mean of the values of the y-variable.

PCC returns values between -1 and 1, symbolically,

$-1 \leq r_{xy} \leq 1$,

where: 1 indicates a strong (actually, perfect) positive relationship, -1 indicates a strong (actually, perfect) negative relationship, and a result of zero indicates no relationship at all. In general, a positive correlation between two variables means that both the variables move in the same direction, whereas a negative correlation between two variables means that both the variables move in opposite directions. In the numerator of the formula of correlation, we calculate how far away we are from the mean *and* if we are above or below the mean, whereas, in the denominator of the formula of correlation, we calculate only how far away we are from the mean. Notice that, in the denominator of the formula of correlation, we take the square roots of some numbers being squared, because the square root of a square is conceptually equivalent to the absolute value, and "absolute value" means "distance." In particular, in this case, the absolute value tells us how far away we are from the corresponding mean value (this process is known as "standardizing," that is, dividing through by magnitude).

For instance, in biology, the relation between independent or the predictor variables and outcome or the dependent variable is explored using correlation analysis. In this way, one can explain how the risk factors or the predictor variables account for the possibility of the occurrence of a disease or presence of a phenotype. The disease outcome or the dependent variable is associated with biological factors (e.g., age and gender), lifestyle variables, psychological variables, and genetic factors (genetic mutations), and correlation tests help us to understand such "risk factors–disease" relationships. Moreover, correlation is an important part of statistical analysis in economics and social policy, and it helps us to understand economic and social phenomena and trends.

Poisson process: A "general random process" (such as the temperature in a room) is random but varies continuously with time, whereas a "Poisson process" refers to a random process that is discrete (namely, a "random point process") and occurs at particular times (e.g., it may describe people arriving at a bus stop, telephone users making telephone calls, etc.). In a Poisson process, events are characterized by a constant mean rate (i.e., these events are random, but, over a certain period of time, they have a known mean rate), and events happen independently (of each other). The Poisson distribution is

$$P_X(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where: $P_X(k)$ denotes the probability, for the Poisson process X, that k events happen in the time period of interest, λ denotes the expected number of events over the time interval of interest (so that $\lambda = rate \times time$), and $k! = k \cdot (k-1) \cdot (k-2) \cdot \ldots \cdot 2 \cdot 1$. This distribution was first introduced by the French mathematician and physicist Siméon Denis Poisson (1781–1840). The mean value of the Poisson distribution equals λ; and the variance of the Poisson distribution also equals λ.

For instance, suppose that you are fishing for 2 hours at a spot where on average people catch 2.8 fishes per hour, and you want to know the probability of catching 5 fishes. This is a random variable that is "Poisson distributed" with $\lambda = 2.8 \times 2 = 5.6$, so that $P_X(k = 5) = \frac{5.6^5 e^{-5.6}}{5!}$.

Chapter 6
Classical Euclidean Geometry, Analytic Geometry, and Trigonometry

Geometry is the scientific study of the quantitative and the qualitative properties of spatial forms and relations (the criteria for equality of triangles provide instances of qualitative geometric knowledge, and the computation of lengths, areas, and volumes exemplifies quantitative geometric knowledge).

Around 300 B.C.E., Euclid published the definitive treatment of Greek geometry and number theory in his thirteen-volume *Elements*, building on the experience and the achievements of previous Greek mathematicians: on the Pythagoreans for Books I–IV, VII, and IX, on Archytas for Book VIII, on Eudoxus for Books V, VI, and XII, and on Theaetetus for Books X and XIII. The axiomatic method used by Euclid is the prototype for the entire field of "pure mathematics," which is "pure" in the sense that we need only pure thought, no physical experiments, in order to verify that the statements are correct—that is, we need only to check the reasoning in the demonstrations. All mathematical theorems are conditional statements—namely, statements of the form

If (hypothesis) *then* (conclusion).

Put simply, one condition (hypothesis) implies another (conclusion). In particular, in a given mathematical system, the only statements that are called "theorems" are those statements for which a proof has been supplied. By a "proof," we mean a list of statements that is endowed with a justification for each statement, and it ends up with the conclusion desired. The following are the six types of justifications allowed for statements in proofs: (i) "by hypothesis . . ."; (ii) "by axiom . . ."; (iii) "by theorem . . ."; (iv) "by definition . . ."; (v) "by step . . ."; (vi) "by rule . . . of logic"; and a justification may involve several of the aforementioned types.

In particular, Euclid articulated:

i. *A set of definitions, such as the following:*
   - A point is that which has no part or magnitude (i.e., it does not have a concrete size).
   - A line is length without breadth.
   - The ends of a line are points.
   - A straight line is a line that lies evenly with the points on itself.
   - A surface is that which has length and breadth only.

- The edges of a surface are lines.
- A plane surface is a surface that lies evenly with the straight lines on itself.

ii. *A set of fundamental rules (axioms):*

- Things that are equal to the same thing are equal to each other.
- If equals are added to equals, then the wholes are equal.
- If equals are subtracted from equals, then the remainders are equal.
- Things that coincide with each other are equal to each other.
- The whole is greater than the part.
- Things that are double of the same things are equal to each other.
- Things that are halves of the same things are equal to each other.

iii. *A set of fundamental propositions (postulates):*

- Postulate 1: a straight line may be drawn from one point to any other point. Given two distinct points, there is a unique straight line that passes through them.
- Postulate 2: a terminated straight line can be produced indefinitely.
- Postulate 3: a circle can be drawn with any center and any radius.
- Postulate 4: all right angles are equal to each other.
- Postulate 5 (known as the Parallel Postulate): if a line segment intersects two straight lines forming two interior angles on the same side that sum to less than two right angles, then the two lines, if extended indefinitely, meet on that side on which the angles sum to less than two right angles.

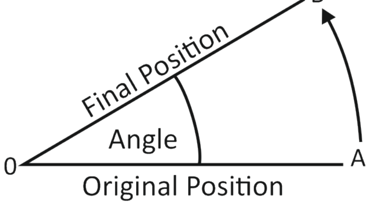
According to Euclidean geometry, space is three-dimensional and isotropic (i.e., it has the same value when measured in different directions). This scientific conception of space clashes with several mythical and folk perceptions of space, according to which space is connected with a form of temporality, and it is unisotropic (for instance, the “upward” and the “forward” directions are evaluated as superior to the “downward” and the “backward” directions). The Euclidean perception of space, combined with the concept of gravity, found its fullest expression in Isaac Newton’s calculus and mechanics.

In view of Euclid’s geometric treatises and the subsequent development of geometry as a scientific discipline, geometry is “an axiomatic in which we ignore all representation, and in which the word ‘space’ designates a structure, i.e., a system of axioms and deductions” (Saddo Ag Almouloud, "Demonstration in Geometry: Historical and Philosophical Perspectives," *Quantitative Research Journal*, vol. 8, Special Edition: Philosophy of Mathematics, 2020, p. 562). In other words, in mathematics, by the term "space," we mean a non-empty set endowed with some mathematical structure. In general, in mathematics, the term "structure" refers to a class of mathematical objects described by axioms. Moreover, sometimes mathematicians use the term "structure" in order to refer to the description of the way in which an object could be reconstructed from simpler objects of the same kind.

Euclidean Geometry

The two most basic geometric concepts are those of an angle and of a straight line. An angle may be considered to be an amount of a rotation or turning. In Figure 6-1, the line OA has been rotated about O in an anti-clockwise direction, until it takes up the position OB. The angle through which the line has turned is the amount of opening between the lines OA and OB. If the line OA is rotated until it returns to its original position, then it will have described one revolution. Angles are usually measured in degrees, minutes, and seconds as follows: $60\,seconds = 1\,minute$, $60\,minutes = 1\,degree$, and $360\,degrees = 1\,revolution$. For instance, an angle of 32 degrees 18 minutes and 3 seconds is written as follows: $32^o 18' 3''$. A "right angle" is the $\frac{1}{4}$th of a revolution, and, therefore, it contains 90^o. An "acute angle" is less than 90^o. An "obtuse angle" lies between 90^o and 180^o. A "reflex angle" is greater than 180^o. "Complementary angles" are angles whose sum is 90^o. "Supplementary angles" are angles whose sum is 180^o.

Figure 6-1: An angle.

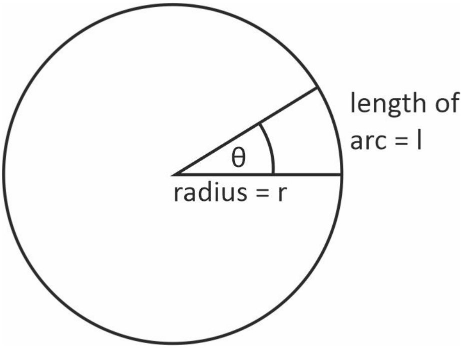


While we usually measure angles in degrees, we can also measure angles in radians. Referring to Figure 6-2,

$$angle\ in\ radians = \frac{length\ of\ arc}{radius\ of\ circle}$$

so that $\theta\ radians = \frac{l}{r} \Leftrightarrow l = r\theta.$

Figure 6-2: Measuring angles in radians.



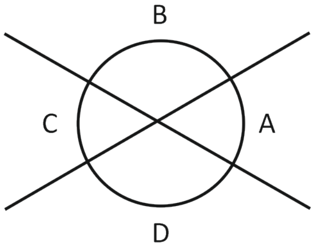
In geometry, the abstraction of a straight line can be attributed to mathematical intuition. According to the ancient Greek mathematician Euclid, an arbitrary straight line can be construed as a “length without breadth” that is perceived as a whole. Furthermore, there are points on every straight line, each point on the straight line corresponds to a real number, and the straight line is complete. For this reason, it is known as the arithmetic or geometric continuum. In fact, the ancient Greek mathematicians’ awareness of the existence of real numbers was developed with reference to geometric processes, in the sense that they construed a real number either as a completed process of combining units or monads (that is, as a rational number) or as an incomplete process of measuring non-commensurable quantities (that is, as an irrational number).

Properties of angles and straight lines:

i. The total angle of a straight line is 180^o(i.e., π radians).

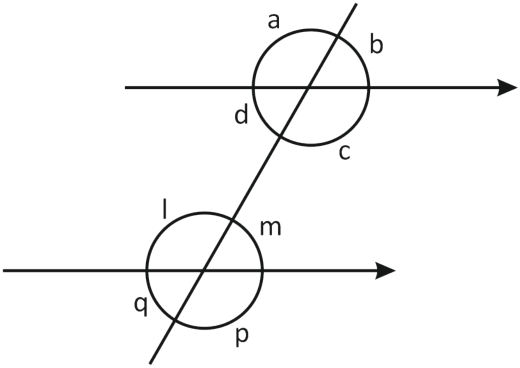
ii. When two straight lines intersect, the opposite angles are equal, as shown in Figure 6-3, where $\angle A = \angle C$ and $\angle B = \angle D$.

Figure 6-3: Opposite angles formed by intersecting straight lines.



iii. If two parallel lines are cut by a transversal, then, as shown in Figure 6-4: the corresponding angles are equal (i.e., $a = l$, $b = m$, $c = p$, and $d = q$); the alternate angles are equal (i.e., $d = m$ and $c = l$); and the interior angles are supplementary (i.e., $d + l = 180^o$ and $c + m = 180^o$). Conversely, if two straight lines are cut by a transversal, then the lines are parallel if one of the following conditions is satisfied: (i) two corresponding angles are equal; (ii) two alternate angles are equal; (iii) two interior angles are supplementary.

Figure 6-4: Angles formed by two parallel lines cut by a transversal.



Types of triangles on the basis of their angles and their sides:

i. An "acute-angled" triangle has all its angles less than 90^o.

ii. A "right-angled" triangle has one of its angles equal to 90^o. The side opposite to the right angle is the longest side, and it is called the "hypotenuse."

iii. An “obtuse-angled” triangle has one angle greater than 90^o.

iv. A “scalene” triangle has all three sides of different length.

v. An “isosceles” triangle has two sides and two angles equal. The equal angles lie opposite to the equal sides.

vi. An “equilateral” triangle has all its sides and angles equal. Each angle of an equilateral triangle is equal to 60^o.

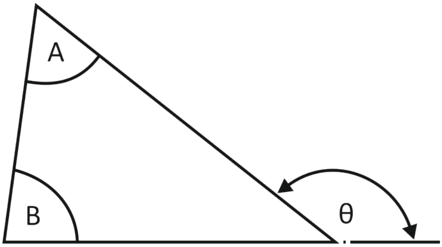
Angle properties of triangles:

i. The sum of the angles of a triangle is equal to 180^o

ii. In every triangle, the greatest angle is opposite to the longest side, and the smallest angle is opposite to the shortest side. Moreover, in every triangle, the sum of the lengths of any two sides is always greater than the length of the third side.

iii. When the side of a triangle is produced, the exterior angle so formed is equal to the sum of the opposite interior angles. For instance, in Figure 6-5, $\angle\theta = \angle A + \angle B$.

Figure 6-5: Exterior angle.



iv. In an isosceles triangle, the perpendicular (drawn from the point where the two equal sides meet) to the base bisects the angle between the two equal sides. Moreover, it bisects the base of the triangle.

Two triangles are said to be “congruent” if they are equal in every respect, both with regard to their corresponding angles and with regard to their corresponding sides (if that is the case, then their areas are equal). If one side and two angles in one triangle are, respectively, equal to one side and two similarly located angles in another triangle, then these triangles are congruent. Moreover, if two sides and the angle between them in one triangle are, respectively, equal to two sides and the angle between them in another triangle, then these triangles are congruent. Given two right-angled triangles, if their hypotenuses are equal to each other and one other side in each triangle are also equal to each other, then these right-angled triangles are congruent.

Two triangles are said to be “similar” if they are equi-angular. Two triangles are equi-angular if and only if their corresponding sides are proportional (by “corresponding sides,” we mean the sides opposite to the equal angles). For instance, given two triangles $\triangle ABC$ and $\triangle XYZ$ such that $\angle A = \angle X$, $\angle B = \angle Y$, and $\angle C = \angle Z$, then
$\frac{AB}{XY} = \frac{AC}{XZ} = \frac{BC}{YZ}$ (and conversely).

Areas of triangles: The area of any triangle is:
$area = \frac{1}{2} \times base \times height.$

Triangles having equal bases and equal heights are equal in area. Moreover, the areas of congruent triangles are equal.

Angle Bisector Theorem: The internal bisector of an angle of a triangle divides the opposite side in the ratio of the sides containing the angle (the converse is also true).

A “median” of a triangle is a line segment that joins a vertex to the midpoint of the side that is opposite to that vertex. The three medians of a triangle intersect at a point called the “centroid.” Notice that the area of a triangle is divided into half by a median (hence the name).

One of the most important geometric theorems is the Pythagorean Theorem, which states that, in every right-angled triangle, the square of the hypotenuse is equal to the sum of the squares of the other two sides. As mentioned earlier, the Pythagorean Theorem led Greek mathematicians to prove the existence of irrational numbers. The Pythagorean Theorem can be proved in an algebraic way, using the concept of a locus, as follows.

Pythagorean Theorem: Consider a right-angled triangle $\triangle ABC$, whose hypotenuse is c, and whose other two sides are a and b, as shown in Figure 6-6. Then
$a^2 + b^2 = c^2.$

Proof: Given the triangle shown in Figure 6-6, we create four triangles identical to it, and we use them in order to form a square with side lengths $a + b$ as shown in Figure 6-7. The area of this square is
$A = (a + b)(a + b).$

Figure 6-6: A right-angled triangle.

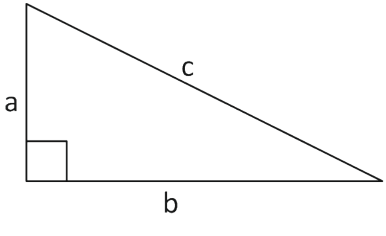
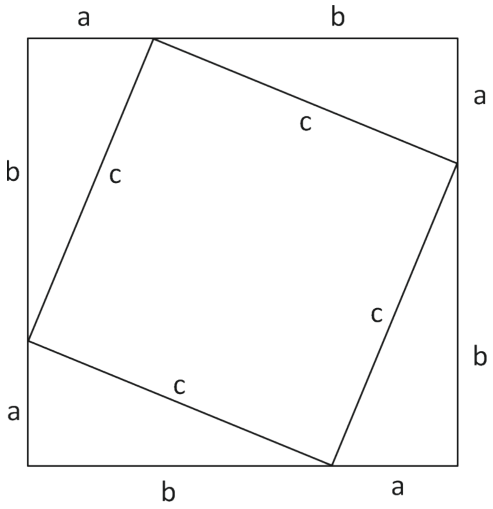


Figure 6-7: Proof of the Pythagorean Theorem.



In Figure 6-7, inside the big square, the hypotenuses of the four identical triangles form another smaller square, whose area is equal to c^2. Each of the four triangles has an area of $\frac{ab}{2}$. In general, notice that, given an arbitrary rectangle $ABCD$ whose height is h, and whose base is b, its area is equal to hb. Therefore, if we draw a diagonal from one vertex, say diagonal AC, it will break the rectangle into two congruent, or equal, triangles, and the area of each of these triangles is half the area of the rectangle, that is, $\frac{hb}{2}$. The area of all four of the triangles that are shown in Figure 6-7 is equal to $4\frac{ab}{2} = 2ab$. Adding up the areas of the smaller square and of the four triangles, we obtain
$A = c^2 + 2ab$.
Hence, given that, as we have shown, $A = (a + b)(a + b)$, it holds that
$(a + b)(a + b) = c^2 + 2ab \Leftrightarrow a^2 + b^2 = c^2$.■

Quadrilaterals and Polygons: A “quadrilateral” is any four-sided figure. Given that a quadrilateral can be split up into two triangles, the sum of its angles is 360^o.
A “parallelogram” has both pairs of opposite sides parallel. If the base of a parallelogram is equal to b and its height is equal to h, then its area is given by the following formula: $A = bh$. Parallelograms having equal

bases and equal heights are equal in area. A parallelogram has the following properties: (i) the sides that are opposite to each other are equal in length; (ii) the angles that are opposite to each other are equal; (iii) the diagonals bisect each other; (iv) the diagonals each bisect the parallelogram.

A "rectangle" is a parallelogram with all its angles equal to 90^o. If the length of a rectangle is equal to l and its width is equal to w, then its area is equal to lw, and its perimeter is equal to $2l + 2w$. A rectangle has all the properties of a parallelogram, but in addition the diagonals are equal in length.

A "rhombus" is a parallelogram with all its sides equal in length. It has all the properties of a parallelogram, but in addition it has the following properties: (i) the diagonals bisect at right angles; (ii) the diagonal bisects the angle through which it passes. If the lengths of the diagonals of a rhombus are d_1 and d_2, then its area is $A = \frac{d_1 \times d_2}{2}$.

A "square" is a rectangle with all its sides equal in length. If the length of each side of a square is equal to a, then its area is equal to a^2, and its perimeter is equal to $4a$. A square has all the properties of a parallelogram, a rectangle, and a rhombus.

A "trapezoid" is a quadrilateral having only one pair of parallel sides (as opposed to a parallelogram, which has both pairs of opposite sides parallel). The parallel sides are called the "bases" of the trapezoid, while the other two sides are called the "legs" of the trapezoid. If the bases (parallel sides) of a trapezoid are equal to a and b, respectively, and if its height is equal to h, then its area is equal to $\frac{1}{2}h(a + b)$.

By the term "polygon," we refer to any plane closed figure bounded by straight lines. A "convex polygon" (e.g., Figure 3-1) has no interior angle greater than 180^o, whereas a "re-entrant polygon" has at least one angle greater than 180^o. In a convex polygon having n sides, the sum of the interior angles is $(2n - 4)$ right angles, and the sum of the exterior angles is 360^o.

Analytic Geometry and Trigonometric Functions

Analytic geometry signifies the introduction of coordinates into geometry in a systematic way—specifically, by unifying aspects of algebra and aspects of geometry. In analytic geometry, geometric theorems are proved using coordinates, algebraic equations, and trigonometry; and analytic geometry is based on the axiomatization of the set $\mathbb{R}$ of real numbers. Moreover, the development of analytic geometry through the algebraization of geometry set the stage for the development of infinitesimal calculus.

The first pioneers of analytic geometry were the second-century B.C.E. Greek astronomer and mathematician Hipparchus of Nicaea, who introduced coordinates for the sphere (in the context of his studies of the night sky), and the third-century B.C.E. Greek geometer Apollonius of Perga, who introduced coordinates for the study of conic sections.

Ancient Greek mathematicians, such as Apollonius of Perga, were the first to observe that circles, ellipses, hyperbolas, and parabolas result from the intersection of a cone by an adequate plane. A cone is defined to be a three-dimensional geometric shape that tapers smoothly from a flat circular base to a point called the vertex (or apex). A circle is produced when the cone is cut by a plane that is parallel to the base of the cone. An ellipse is produced when the cone is cut by a plane that is not parallel to the base of the cone or the side of the cone, and it cuts only one nappe of the cone. A hyperbola is produced when the intersecting plane cuts both nappes of the cone. A parabola is produced when the oblique section of the cone is parallel to the slant height (the height of a cone from the vertex to the periphery, rather than the center, of the base). In the Middle Ages, the use of coordinates in mathematics and analytic geometry was further analyzed and developed by the fourteenth-century French Catholic bishop, philosopher, and mathematician Nicolas d'Oresme.

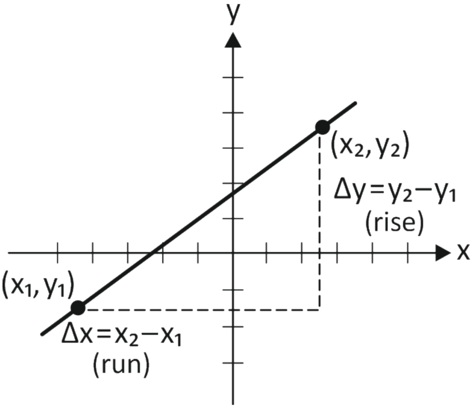
By the term "locus," we mean a set of all the points that satisfy a specific rule. Moreover, the path drawn by a point moving according to a given rule is called the "locus of the point." Thus, using the concept of a locus, we can study geometric problems through algebra. In analytic geometry, we put traditional (Euclidean) geometry on the Cartesian plane. René Descartes has pointed out that "any problem in geometry can easily be reduced to such terms that knowledge of lengths of certain straight lines is sufficient for its construction" (René Descartes, "On Analytic Geometry," translated by David E. Smith and Marcia L. Latham, in *A Source Book in Mathematics*, edited by David E. Smith, New York: Dover, 1959, p. 397). In particular, according to Descartes, "just as arithmetic consists of only four or five operations, namely, addition, subtraction, multiplication, division, and the extraction of roots, which may be considered a kind of division, so in geometry," we can find required lines by merely adding or subtracting other lines; or else, by working as follows (*ibid*, pp. 397–98):

> . . . taking one line which I shall call unity in order to relate it as closely as possible to numbers, and which can in general be chosen arbitrarily, and having given two other lines, to find a fourth line which shall be to one of the given lines as the other is to unity (which is the same as multiplication);

> or, again, to find a fourth line which is to one of the given lines as unity is to the other (which is equivalent to division); or, finally, to find one, two, or several mean proportionals between unity and some other line (which is the same as extracting the square root, cube root, etc., of the given line).

Consider two points $P(x_1, y_1)$ and $Q(x_2, y_2)$ on the xy-plane and connect them with a straight line segment as shown in Figure 6-8.

Figure 6-8: Slope and Distance.



The x-coordinate of point P is x_1, the x-coordinate of point Q is x_2, and the distance between x_1 and x_2 is $x_2 - x_1$; in order to avoid the use of plus and minus signs, we can use the absolute value $|x_2 - x_1|$. The y-coordinate of point P is y_1, the y-coordinate of point Q is y_2, and the distance between y_2 and y_1 is $y_2 - y_1$; in order to avoid the use of plus and minus signs, we can use the absolute value $|y_2 - y_1|$. Therefore, the horizontal distance between points P and Q is $x_2 - x_1$, and the vertical distance between points P and Q is $y_2 - y_1$. Now, consider the right-angled triangle that is defined by the points $P(x_1, y_1)$, $Q(x_2, y_2)$, and the point R (the intersection between the horizontal side and the vertical side): the three sides of this right-angled triangle are the hypotenuse PQ, the horizontal side, which is $x_2 - x_1$, and the vertical side, which is $y_2 - y_1$. The "slope," or "gradient," of the straight line segment PQ, denoted by m_{PQ}, is the quotient of the "rise" over the "run," comparing how much one travels vertically ("up and down") versus how much one travels horizontally. Thus, it relates the steepness or inclination of the straight line segment PQ to the coordinates; symbolically:

$$slope = m_{PQ} = \frac{rise}{run} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\Delta y}{\Delta x}$$

(see Figure 6-8; the Greek letter Δ is used to indicate change).

In Figure 6-8, the distance between points P and Q, denoted by d_{PQ}, is given by (and, indeed, is a version of) the Pythagorean Theorem. Therefore, in Figure 6-8,

$$\left(d_{PQ}\right)^2 = (run)^2 + (rise)^2 \Leftrightarrow d_{PQ} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

It can be easily verified that the midpoint of the straight line segment joining points (x_1, y_1) and (x_2, y_2) is $\left(\frac{x_1+x_2}{2}, \frac{y_1+y_2}{2}\right)$.

All points (x, y) in $\mathbb{R}^2$ satisfying the equation $y = mx + b$ form a straight line, and m is the slope of the straight line. For the slope m of the straight line passing through the points (x_1, y_1) and (x_2, y_2), we have:

i. If $x_1 = x_2$, m is undefined (the line is vertical).

ii. If $x_1 \neq x_2$, then $m = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$.

Two non-vertical straight lines y_1 and y_2, with slopes m_1 and m_2, respectively, are parallel if and only if $m_1 = m_2$ (i.e., their slopes are equal), and they are perpendicular if and only if $m_1 m_2 = -1$ (i.e., the product of their slopes is -1).

In order to find the equation of a non-vertical straight line, we work as follows:

i. we find a point (x_1, y_1) on the line;

ii. we find the slope m of the line;

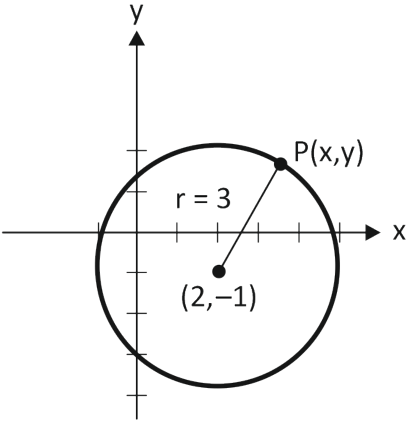
iii. we write the equation of the line as follows:
$y - y_1 = m(x - x_1)$; this equation is called the "point-slope" form of the equation of a line.

For instance, let us find the equation of the straight line passing through the points $(5, -0.5)$ and $(10, 9.5)$. Firstly, we define the point $(x_1, y_1) = (5, -0.5)$. Secondly, we find the slope of the required line: $m = \frac{9.5-(-0.5)}{10-5} = 2$. Thirdly, we find the equation of the required line: $y - y_1 = m(x - x_1) \Rightarrow y - (-0.5) = 2(x - 5) \Rightarrow y = 2x - 10.5$.

Circle

As we can see in Figure 6-9, a circle with center $O(v, w)$ and radius r is the set of all points in the xy-plane whose distance from O is r (in Figure 6-9, $O(v, w) = O(2, -1)$, and $r = 3$).

Figure 6-9: Circle.

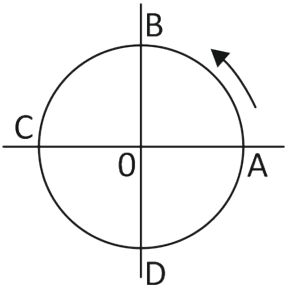


If (x, y) is a point on the circle with center $O(v, w)$ and radius r, then the distance formula implies that
$r = \sqrt{(x-v)^2 + (y-w)^2} \Leftrightarrow r^2 = (x-v)^2 + (y-w)^2$,
which is the standard form of the equation of a circle with center (v, w) and radius r. The circumference of a circle of radius r is $C = 2\pi r$, and the area of a circle of radius r is $A = \pi r^2$, where $\pi \approx 3.14$ is Archimedes's constant (the ratio of the circle's circumference to its diameter). Archimedes approximated π by using the fact that the circumference of a circle is bounded by the perimeter of an inscribed polygon and the perimeter of a circumscribed polygon. In particular, he used a 96-sided inscribed polygon and a 96-sided circumscribed polygon to find the following approximation:
$3 + \frac{10}{71} < \pi < 3 + \frac{10}{70}$.
It is worth mentioning that the degenerate possibilities for a circle are the following: a point or no graph at all.
The study of the circle underpins trigonometry. The term "trigonometry" appeared for the first time in the book *Trigonometria* by Bartholomaeus Pitiscus (1561–1613) in 1595, and it literally means measuring (and, more broadly, studying) "trigons" ("trigon" being the Latin word for "triangle"). The acknowledged founder of trigonometry is the ancient Greek astronomer and mathematician Hipparchus of Nicaea (ca. 190–ca. 120 B.C.E.). Moreover, around 100 C.E., another Greek mathematician, Menelaus of Alexandria, published a series of treatises on chords.

Trigonometric Functions

In the context of analytic geometry, we can also study the basic trigonometric functions on the unit circle (specifically, on a circle whose center is $(0,0)$ and whose radius $r = 1$).

Figure 6-10: The number circle.

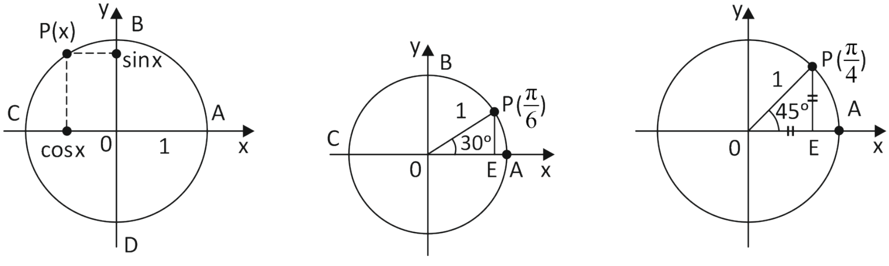


Consider a circle of unit radius, as shown in Figure 6-10, and let point A (the right-hand endpoint of the horizontal diameter) be a reference point. Let an anti-clockwise motion round the circle be a positive direction, and a clockwise motion be a negative direction. A circle of unit radius with a reference point and the direction of tracing specified is called the "number circle." Given an arbitrary point P of the number circle, there are infinitely many arcs beginning at the point A and terminating at the point P. One of these arcs is the shortest arc connecting the points A and P, and all the other arcs are obtained from the shortest arc by adding or subtracting an integral number of complete revolutions. Hence, every point P of the number circle is associated with an infinite set of numbers that consists of the values of all the arcs beginning at the point A and terminating at the point P (the lengths of the arcs are taken with the plus or the minus sign according as the motion from the point A to the point P is anti-clockwise or clockwise, respectively).

The circumference of the circle of unit radius is equal to 2π. Therefore, the lengths of all the arcs terminating at the given point P differ from one another by a multiple of 2π, so that the general form of these quantities is $x + 2\pi a$, where $a \in \mathbb{Z}$, and x is the length of the shortest arc connecting the points A and P. Thus, for every real number x, there is a point $P(x)$ of the number circle such that the length of the arc AP is x, and every point P of the circle corresponds to an infinite set of numbers of the form $x + 2\pi a$, where $a \in \mathbb{Z}$, and x is the length of one of the arcs connecting the points A and P. For instance, the point of reference, namely, A (Fig. 6-10), corresponds to the number 0, that is, $A = A(0)$, and, since the length of the circumference of the unit circle is 2π, it follows that $\frac{\pi}{2}$ is the length of an arc equal to one-fourth of the circumference. Hence, if we lay off an arc of length equal to a quarter of the circumference from the point A in the positive direction, then we obtain a point $B = B\left(\frac{\pi}{2}\right)$, as shown in Figure 6-10. By analogy, we can find the point corresponding to $-\frac{\pi}{2}$ by starting from the point A in the negative direction and covering the path of length $\frac{\pi}{2}$, thus arriving at the point $D = D\left(-\frac{\pi}{2}\right)$, as shown Figure 6-10.

Assume that the center of the number circle coincides with the origin $O(0,0)$ of the rectangular coordinate system XOY, as shown in Figure 6-11. Let x be an arbitrary real number. Then, on the number circle, we find the point $P(x)$ that corresponds to x. The ordinate of the point $P(x)$ is called the "sine" of the number x (denoted by $sinx$), the abscissa of the point $P(x)$ is called the "cosine" of the number x (denoted by $cosx$), the ratio $\frac{sinx}{cosx}$ is called the "tangent" of the number x (denoted by $tanx$), and the ratio $\frac{cosx}{sinx}$ is called the "cotangent" of the number x (denoted by $cotx$).

Figure 6-11: Trigonometric functions.



Notice that the reference point A on the number circle corresponds to the number 0, that is, $A = A(0)$. Since the abscissa and the ordinate of this point are 1 and 0, respectively, we have $cos0 = 1$, $sin0 = 0$, and $tan0 = \frac{sin0}{cos0} = 0$. The point B of intersection of the circle and the positive ray of the axis OY corresponds to the number $\pi/2$. Since the abscissa and the ordinate of the point B are 0 and 1, respectively, we have $\cos\left(\frac{\pi}{2}\right) = 0$ and $\sin\left(\frac{\pi}{2}\right) = 1$, whereas $\tan\left(\frac{\pi}{2}\right)$ is not defined. Similarly, as shown in Figure 6-11, given the coordinates of the points C and D, we realize that $cos\pi =$

-1, $sin\pi = 0$, $tan\pi = 0$, $\cos\left(\frac{3\pi}{2}\right) = 0$, $sin\left(\frac{3\pi}{2}\right) = -1$, and $tan\left(\frac{3\pi}{2}\right)$ is not defined. The parametrization of the unit circle can be written as follows:

$$(cos\theta, sin\theta)$$

where $0 \leq \theta \leq 2\pi$.

The domain of $y = sinx$ is $(-\infty, +\infty)$, and its range is $[-1,1]$. The function $y = sinx$ is odd (and, therefore, it is symmetric about the origin), its y-intercept is $(0,0)$, and its x-intercepts are $x = n\pi$, where n is an integer.

The domain of $y = cosx$ is $(-\infty, +\infty)$, and its range is $[-1,1]$. The function $y = cosx$ is even (and, therefore, it is symmetric about the y-axis), its y-intercept is $(0,1)$, and its x-intercepts are $x = \frac{\pi}{2} + n\pi$, where n is an integer.

We can summarize the basic definitions and the basic formulae of trigonometry as follows:

$Sine$: $\sin\theta = \frac{opposite\ side}{hypotenuse}$;

$Cosine$: $cos\theta = \frac{adjacent\ side}{hypotenuse}$;

$Tangent$: $tan\theta = \frac{opposite\ side}{adjacent\ side}$;

$Cosecant$: $csc\theta = \frac{hypotenuse}{opposite\ side} = \frac{1}{sin\theta}$;

$Secant$: $sec\theta = \frac{hypotenuse}{adjacent\ side} = \frac{1}{cos\theta}$;

$Cotangent$: $cot\theta = \frac{adjacent\ side}{opposite\ side} = \frac{1}{tan\theta}$;

and the basic trigonometric identities:

$sin^2a + cos^2a = 1$ (this is an expression of the Pythagorean theorem in terms of trigonometric functions);

$1 + tan^2a = \frac{1}{cos^2a}$;

$1 + cot^2a = \frac{1}{sin^2a}$;

$sin(-a) = -sina$;

$cos(-a) = cosa$;

$\sin(a \pm b) = sina \cdot cosb \pm cosa \cdot sinb$;

$\cos(a \pm b) = cosa \cdot cosb \mp sina \cdot sinb$;

$sina + sinb = 2sin\frac{1}{2}(a + b) \cdot cos\frac{1}{2}(a - b)$;

$sina - sinb = 2sin\frac{1}{2}(a - b) \cdot cos\frac{1}{2}(a + b)$;

$cosa + cosb = 2cos\frac{1}{2}(a + b) \cdot cos\frac{1}{2}(a - b)$;

$cosa - cosb = -2sin\frac{1}{2}(a + b) \cdot sin\frac{1}{2}(a - b)$;

$sin2a = 2sina \cdot cosa$;
$cos2a = cos^2a - sin^2a$;
$sin\frac{1}{2}a = \pm\sqrt{\frac{1-cosa}{2}}$,
$cos\frac{1}{2}a = \pm\sqrt{\frac{1+cosa}{2}}$,
sin $(a \pm \pi/2) = \pm cosa$, and cos $(a \pm \pi/2) = \mp sina$ (the graphs of $sine$ and $cosine$ have the same shape, but the only difference is a shift of $y = cosx$ to $y = sinx$ by $\frac{\pi}{2}$ units to the right).
The inverse trigonometric functions are denoted as follows: $arcsinx \equiv sin^{-1}x$ ($y = arcsinx \Leftrightarrow x = siny$) , $arccosx \equiv cos^{-1}x$ ($y = arccosx \Leftrightarrow x = cosy$) , $arctanx \equiv tan^{-1}x$ ($y = arctanx \Leftrightarrow x = tany$), and $arccotx \equiv cot^{-1}x$ ($y = arccotx \Leftrightarrow x = coty$).
In order to calculate the angle A subtended at the center of a circle of radius r by a chord of length a, we use the cosine rule

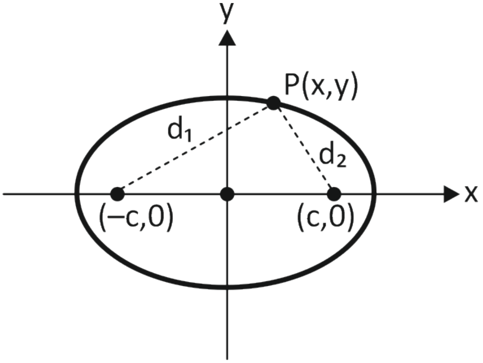
$$A = arccos\left(\frac{b^2 + c^2 - a^2}{2bc}\right)$$

where b and c are the sides of the angle A, and $b = c = r$ (i.e., A is bounded by two radii), $a = chord\ length$, and $A = angle\ subtended\ at\ the\ center\ of\ the\ circle$.

Ellipse

As we can see in Figure 6-12, an “ellipse” is the set of all points in a plane the sum of whose distances from two fixed points (“foci”) is constant. Foci: $(-c, 0)$ and $(c, 0)$. Notice that, if the two foci coincide, then we receive a circle. The Greek word ellipse, literally meaning “omission,” was first applied by Apollonius of Perga, because, in the case of an ellipse, the conic section of the cutting plane makes a smaller angle with the base than does the side of the cone.

Figure 6-12: Ellipse.



The standard form of the equation of an ellipse with center at the origin and foci on the x-axis is

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1.$$

By setting $y=0$, we find that the x-intercepts are $(-a,0)$ and $(a,0)$. By setting $x=0$, we find that the y-intercepts are $(0,-b)$ and $(0,b)$. The larger segment from $(-a,0)$ to $(a,0)$ is called the "major axis," while the "minor axis" is the segment from $(0,-b)$ to $(0,b)$. The endpoints of the major axis are called the "vertices of the ellipse"; vertices: $(-a,0)$ and $(a,0)$.

If the foci are placed on the y-axis at $(0,-c)$ and $(0,c)$, then the standard form of the equation of an ellipse is

$$\frac{x^2}{b^2}+\frac{y^2}{a^2}=1.$$

In this case, the major axis is along the y-axis, the foci are $(0,c)$ and $(0,-c)$, and the vertices are $(0,a)$ and $(0,-a)$.

Given the definition of an ellipse, the degenerate possibilities for an ellipse are the following: a point or no graph at all.

In our solar system, many bodies revolve in elliptical orbits around a larger body that is located at one focus. In the seventeenth century, Johannes Kepler, based on Apollonius's mathematical study of the ellipse, articulated a rigorous explanation of planetary motions.

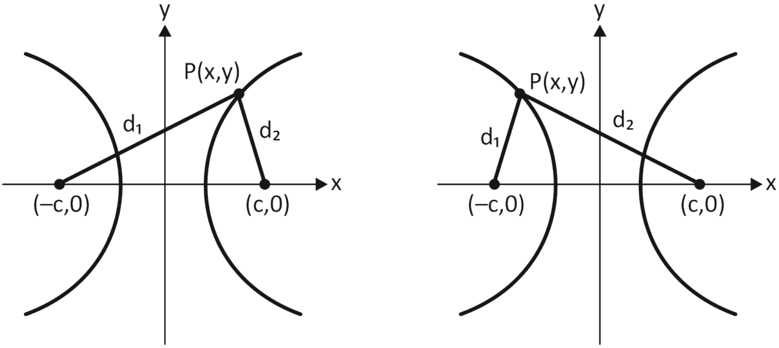
Moreover, regarding the ellipse, it should be mentioned that it has a reflection property that causes any ray or wave that originates at one focus to strike the ellipse and pass through the other focus. In terms of acoustics, the aforementioned property implies that, in a room with an elliptical ceiling, even a slight noise made at one focus can be heard at the other focus, but, if people are standing between the foci, then they hear nothing. Such rooms are known as whispering galleries.

As regards architecture, it should be mentioned that ornamental arches are often elliptical in shape; in other words, arches whose main purpose is beauty and not strength are often elliptical in shape.

Hyperbola and hyperbolic functions

As we can see in Figure 6-13, a “hyperbola” is the set of all points in a plane the difference of whose distances from two fixed points (“foci”) is a positive constant (the Greek word hyperbola literally means “extravagance”). Hence, the distances between the foci and a point on the figure maintain a *constant difference* for a hyperbola and a *constant sum* for an ellipse.

Figure 6-13: Hyperbola.



Given the definition of a hyperbola, the degenerate possibilities for a hyperbola are two intersecting straight lines.

The standard form of a hyperbola with center at the origin and foci on the x-axis is

$$\frac{x^2}{a^2}-\frac{y^2}{b^2}=1.$$

By setting $y=0$, we find that the x-intercepts are $(-a,0)$ and $(a,0)$. The line segment joining these two points is called the “transverse axis.” The endpoints of the transverse axis are called the “vertices of the hyperbola.” By setting $x=0$, we find that there are no y-intercepts. The line segment from $(0, b)$ to $(0, -b)$ is called the "conjugate axis." In order to determine the significance of b, we write

$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$ as $y = \frac{\pm bx}{a}\sqrt{1 - \frac{a^2}{x^2}}$.

As $|x|$ tends to infinity, $1 - \frac{a^2}{x^2}$ tends to 1, and, therefore, the graph of the hyperbola approaches the lines

$y = \pm\frac{b}{a}x$.

These lines are called the "asymptotes of the hyperbola" (they are the diagonals of a rectangle of dimensions $2a$ by $2b$).

If the foci are placed on the y-axis at $(0, -c)$ and $(0, c)$, then the standard form of the equation of a hyperbola is

$\frac{y^2}{a^2} - \frac{x^2}{b^2} = 1$,

and, in this case, the asymptotes are given by

$y = \pm\frac{a}{b}x$.

Hyperbolic functions: Hyperbolic functions are analogues of the ordinary trigonometric functions, but hyperbolic functions are defined using the hyperbola rather than the circle. The hyperbolic functions are defined as follows:

$hyperbolic\ sine: sinhx = \frac{1}{2}(e^x - e^{-x})$,

$hyperbolic\ cosine: coshx = \frac{1}{2}(e^x + e^{-x})$,

$hyperbolic\ tangent: tanhx = \frac{sinhx}{coshx}$,

$hyperbolic\ cotangent: cothx = \frac{coshx}{sinhx}$,

$hyperbolic\ secant: sechx = \frac{1}{coshx}$, and

$hyperbolic\ cosecant: cschx = \frac{1}{sinhx}$.

Basic formulae of hyperbolic functions:

i. $cosh(-x) = coshx$,

ii. $sinh(-x) = -sinhx$,

iii. $e^x = sinhx + coshx$,

iv. $e^{-x} = coshx - sinhx$,

v. $tanh(-x) = -tanhx$,

vi. $cosh^2x - sinh^2x = 1$,

vii. $sech^2x + tanh^2x = 1$,

viii. $coth^2x - csch^2x = 1$,

ix. $cosh2x = cosh^2x + sinh^2x$,

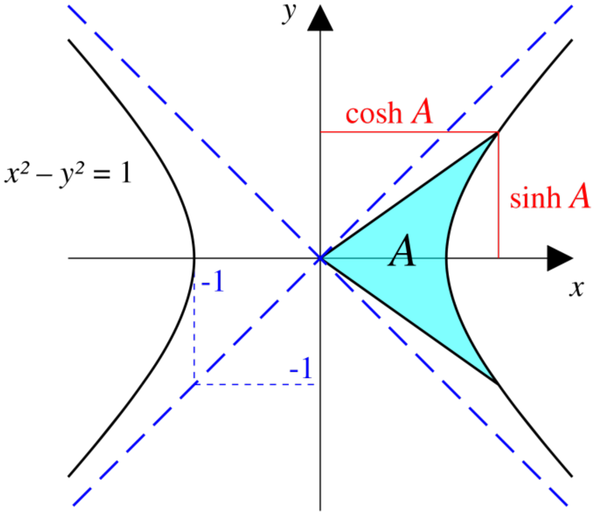
x. $sinh2x = 2sinhxcoshx$,

xi. $sinh(x+y) = sinhxcoshy + coshxsinhy$, and

xii. $cosh(x+y) = coshxcoshy + sinhxsinhy$.

Just as the points $(cost, sint)$ form the unit circle (defined by $x^2 + y^2 = 1$), the points $(cosht, sinht)$ form the right half of the unit hyperbola (defined by $x^2 - y^2 = 1$), as shown in Figure 6-14.

Figure 6-14: Hyperbolic functions (source: Wikimedia Commons: Author: Marco Polo; https://commons.wikimedia.org/wiki/File:Hyperbolic_functions.svg).

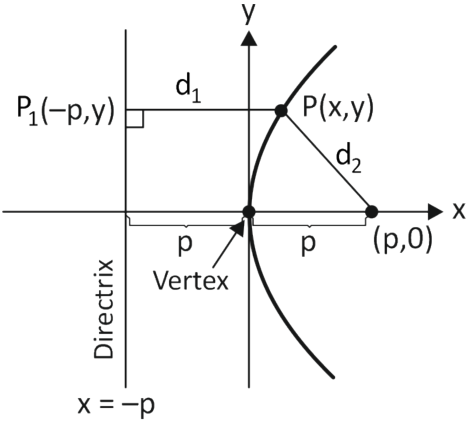


In mechanics, hyperbolic functions are used in order to describe the shape of electric lines freely hanging between two poles and any idealized hanging chain or cable supported only at its ends and hanging under its own weight. In particular, the "catenary" is a curve that describes the shape of a flexible hanging chain or cable, and its equation in Cartesian coordinates is $y = acosh\left(\frac{x}{a}\right) = \frac{a}{2}\left(e^{\frac{x}{a}} + e^{-\frac{x}{a}}\right)$. Moreover, catenaries and related curves are used in the design of bridges and arches, so that forces do not result in bending moments, and, in the offshore oil and gas industry, the term "catenary" refers to a steel catenary riser (a pipeline suspended between a production platform and the seabed that adopts an approximate catenary shape).

Parabola

As we can see in Figure 6-15, a "parabola" is the set of all points in a plane that are equidistant from a fixed line ("directrix") and a fixed point ("focus") not on the line (the word "parabola" derives from the Greek terms "parā," meaning "beside," and "bolē," meaning "a throw," and, therefore, "parabola" literally means "para-beside"—that is, placing side by side).

Figure 6-15: Parabola.



The standard form of the equation of a parabola with directrix $x = -p$ and focus at $(p, 0)$ is
$4px = y^2$.
The line through the focus that is perpendicular to the directrix is called the "axis of symmetry." In this case, the axis of symmetry is the x-axis, and the parabola opens to the right. The point on the axis of symmetry that is midway between the focus and the directrix is called the "vertex," and the vertex is the turning point of the parabola. The standard form of the equation of a parabola with directrix $x = p$ and focus at $(-p, 0)$ is
$-4px = y^2$,
and, in this case, the parabola opens to the left.
Obviously, the axis of symmetry of a parabola may be the y-axis. If the directrix is $y = -p$ and the focus is at $(0, p)$, then the standard form of the equation of a parabola is
$x^2 = 4py$,

and the parabola opens upward. If the directrix is $y = p$ and the focus is at $(0, -p)$, then the standard form of the equation of a parabola is
$x^2 = -4py$,
and the parabola opens downward.
As regards the parabola in general, it should be mentioned that it has a reflection property that causes any ray or wave that originates at the focus and strikes the parabola to be reflected parallel to the axis of symmetry. Thus, for instance, flashlights and searchlights use a parabolic reflector with the bulb located at the focus. Additionally, due to the reflection property of a parabola, any ray or wave that comes into a parabolic reflector parallel to the axis of symmetry is directed to the focus point. For this reason, radars, radio antennas, and reflecting telescopes operate according to this principle. In astronomy, the parabola features in both the construction of telescopes and in the motion of comets around the Sun. Finally, due to their great strength, parabolic arches are used extensively in bridges, cathedrals, and elsewhere in architecture and engineering, especially in case we have equally spaced load.

Volumes and Surface Areas

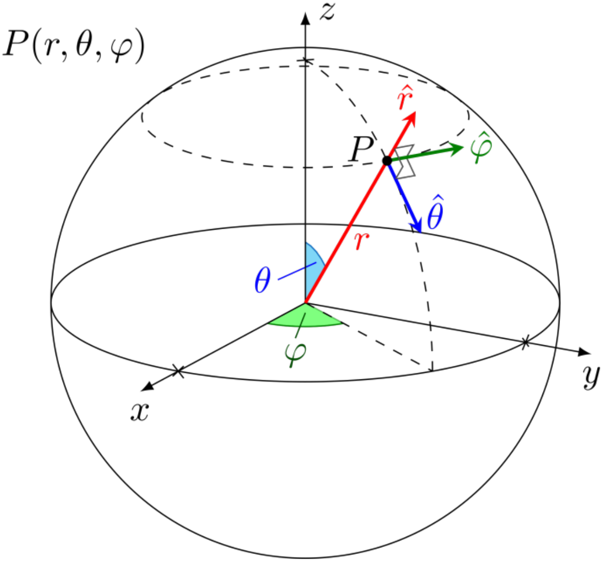
By the term “volume,” we mean the amount of three-dimensional space enclosed by a closed surface. The volume of any solid having a uniform cross-section is equal to:
cross-sectional area$\times$length of solid.
The surface area of any solid having a uniform cross-section is equal to:
curved surface$+$ends; namely:
perimeter of cross-sections$\times$length of solid$+$total area of ends.
The volume of a sphere with radius r is equal to
$\frac{4}{3}\pi r^3$,
and its surface area is equal to
$4\pi r^2$.
The volume of a cylinder whose height is h and whose base is a circle with radius r is equal to
$\pi r^2 h$,
and its surface area is equal to
$2\pi rh + 2\pi r^2 = 2\pi r(h + r)$.
The volume of a cone whose vertical height is h and whose base is a circle with radius r is equal to
$\frac{1}{3}\pi r^2 h$,
and, if l is its slant height, then its surface area is equal to

$\pi rl + \pi r^2$.
The volume of a pyramid whose height is h and whose base's area is equal to A is given by the following formula:
$V = \frac{1}{3}Ah$.
The surface area of a pyramid is equal to the sum of the areas of the corresponding triangles plus the area of the base.

Spherical Coordinates and Polar Coordinates

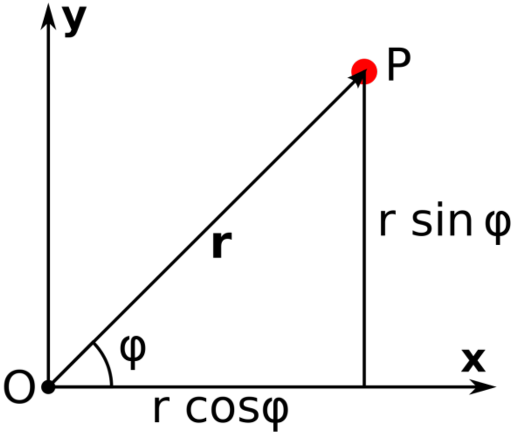
Until now, we have determined the position of a point P by the lengths of its Cartesian (rectangular) coordinates. As I explained in Chapter 2, in the Cartesian or rectangular coordinate system, we have three axes, x, y, and z, which are perpendicular to each other, and we can define any point by taking the number of units in the x-direction, the number of units in the y-direction, and the number of units in the z-direction, through the corresponding projections.
As shown in Figure 6-16, in the spherical coordinate system, we have again three mutually perpendicular coordinates, r, θ, and φ, where: the radial line r is the shortest distance between the origin of the coordinate system and the given point P, θ (known as the "polar angle" or the "inclination") is defined to be the angle between the z-axis and the radial line r, and φ (known as the "azimuthal angle" or the "azimuth") is the angle between the orthogonal projection of the radial line r onto the reference xy-plane (which is orthogonal to the z-axis and passes through the origin) and the x-axis (which is orthogonal to the z-axis and to the y-axis).
As shown in Figure 6-16, the relation between the spherical coordinate system and the rectangular coordinate system is the following:
We can convert spherical coordinates (r, θ, φ) to rectangular coordinates (x, y, z), using the following formulae:
$x = rsin\theta cos\varphi$,
$y = rsin\theta sin\varphi$, and
$z = rcos\theta$.
Moreover, we can convert rectangular coordinates (x, y, z) to spherical coordinates (r, θ, φ), using the following formulae:
$r = \sqrt{x^2 + y^2 + z^2}$,
$\theta = cos^{-1}\frac{z}{\sqrt{x^2+y^2+z^2}} = arccos\frac{z}{\sqrt{x^2+y^2+z^2}}$, and
$\varphi = tan^{-1}\frac{y}{x} = arctan\frac{y}{x}$.

Figure 6-16: Spherical coordinates (source: Wikimedia Commons: Author: Cristian Quinzacara; https://commons.wikimedia.org/wiki/File:Coordenadas_esf%C3%A9ricas_01.svg).



Using the aforementioned methodology and reasoning regarding spherical coordinates, we can define points in the two-dimensional polar coordinate system, where $(x, y) = (rcos\varphi, rsin\varphi)$, as shown in Figure 6-17. For instance, the equations of lines and conic sections can be expressed in polar coordinates through the relation $(x, y) = (rcos\varphi, rsin\varphi)$.

Figure 6-17: Two-dimensional polar coordinates (source: Wikimedia Commons: Author: WillowW; https://commons.wikimedia.org/wiki/File:Polar_coordinate_components.svg).



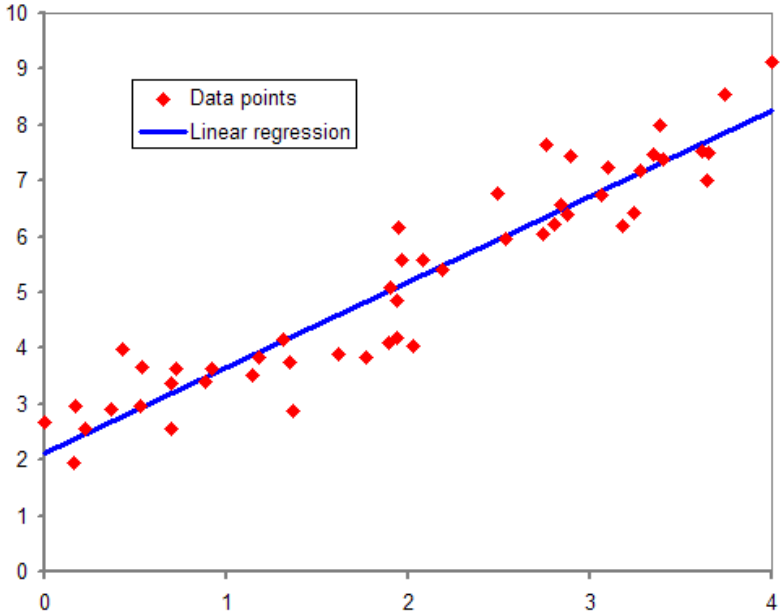
Hence, the graph of the equation $r = f(\varphi)$ in polar coordinates is the same as the graph of the parametric equations (with parameter φ) $x = f(\varphi)cos\varphi$ and $y = f(\varphi)sin\varphi$ in Cartesian coordinates; and, conversely, the graph of a given equation in x and y is the same as the graph of the equation in r and φ obtained by substituting x and y with $rcos\varphi$ and $rsin\varphi$, respectively.

A Note on the Line of Best Fit

A “scatter plot” is a type of mathematical diagram that uses Cartesian coordinates, and it provides a visual and statistical means to test the strength of a relationship between two variables. The “line of best fit” is a (straight) line that is used to express a relationship in a scatter plot of different data points, and it minimizes the distance between it and the data under consideration, as shown in Figure 6-18. In Figure 6-18, we can see that the general trend of the data points is going up to the right, indicating a positive correlation. When we draw a line of best fit, we do not want to draw it so high that all of the data points are below that line, nor do we want to draw it so low that all of the data points are above that line, but we want to draw a line of best fit that comes as close to those data points as possible.

In other words, a line of best fit, also known as a “trend line” or a “line of regression,” is a line that best displays the trend of a group of points on a scatter plot, and it is used to predict the behavior of data using the slope of that line.

Figure 6-18: The line of best fit (source: Wikimedia Commons: Author: Amatulic; https://commons.wikimedia.org/wiki/File:Normdist_regression.png).



A basic way of approximating the line of best fit is to use a ruler and try to draw that line in such a way that it comes as close to the given data points as possible (as shown in Figure 6-18). Assume, for instance, that, given a scatter plot, two points that lie on the line of best fit are $A(1,-3)$ and $B(7,5)$. Firstly, we have to find the slope of this line: $m = \frac{rise}{run} = \frac{y_2-y_1}{x_2-x_1} = \frac{5-(-3)}{7-1} = \frac{8}{6} = \frac{4}{3}$. Now, we shall use the slope-intercept form, $y = mx + b$, in order to find the equation of this line. Given that $m = \frac{4}{3}$, we obtain $y = \frac{4}{3}x + b$, and we have to determine the value of b. Let us use the point $A(1,-3)$ in order to determine the value of b (of course, we shall find the

same result if we use the point $B(7{,}5)$ in order to determine b). When $y = -3$, $x = 1$, and, therefore, we have:
$y = \frac{4}{3}x + b \Rightarrow -3 = \frac{4}{3}(1) + b \Rightarrow b = -\frac{13}{3}$.
Hence, in this case, the equation of the line of best fit is
$y = \frac{4}{3}x - \frac{13}{3}$.
Thus, using analytic geometry, we can find the line of best fit, which is an intelligent guess or approximation on a set of data aiming to identify and describe the relationship between given variables.

Chapter 7
Vectors, Vector Spaces, Normed Vector Spaces, and Metric Spaces

The discipline of mathematics that deals with matrices (covered in Chapter 3) and vectors (and, more generally, with vector spaces and linear transformations) is called Linear Algebra. In this chapter, we shall complete our study of the basic concepts of linear algebra, and we shall study the basic principles of metric spaces and metric geometry.

Fields and Vectors

In mathematics, a "field" is an algebraic structure that has two binary operations, usually called "addition" and "multiplication," and both of them are always commutative. Fields model number systems (since numbers can be added or multiplied, and, therefore, subtracted and divided, too, and various relationships hold true between them). A "field" is a structured set

$$(F, 0,1, +,\cdot)$$

that satisfies the following properties:

(F1) $0,1 \in F, 0 \neq 1$, and $+$ and $\cdot$ are binary functions (operations) on F.

(F2) Addition $+$ satisfies the following identities:

i. $(x+y)+z = x+(y+z)$,

ii. $x+y = y+x$,

iii. $x+0 = x$,

and, for every x, there exists some x' such that $x + x' = 0$.

(F3) Multiplication $\cdot$ satisfies the following identities:

i. $(x\cdot y)\cdot z = x\cdot (y\cdot z)$,

ii. $x\cdot y = y\cdot x$,

iii. $x\cdot 1 = x$,

and, for every x, there exists some x'' such that $x \cdot x'' = 1$.

(F4) Both addition and multiplication satisfy the identity

$x\cdot (y+z) = x\cdot y + x\cdot z$.

Remark: The axioms of a field imply that any field F satisfies the following:

i. For every x, there exists a unique x' such that $x + x' = 0$; and then $x' = -x$ (called the "additive inverse" of x). Moreover, for every $x \neq 0$, there exists a unique x'' such that $x\cdot x'' = 1$; and then $x'' = x^{-1}$ (called the "multiplicative inverse" of x).

ii. $x \cdot 0 = 0$.

iii. $x \cdot y = 0 \Rightarrow x = 0 \; or \; \mathrm{y} = 0$.

iv. $(-x) \cdot y = -(x \cdot y)$.

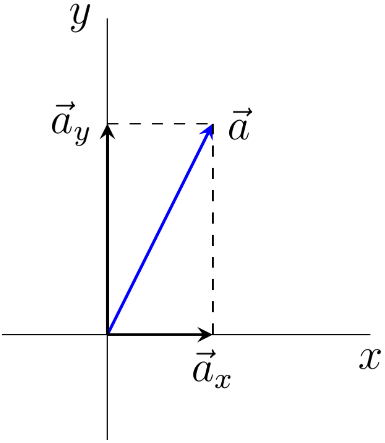
v. A field is a set F that is closed under the operations of addition and multiplication.

Familiar examples of fields are the set $\mathbb{Q}$ of all rational numbers and the set $\mathbb{R}$ of all real numbers. Notice that the set $\mathbb{Z}$ of all integers is not a field, because not every element of $\mathbb{Z}$ has a multiplicative inverse (in fact, only 1 and -1 have multiplicative inverses in $\mathbb{Z}$).

A "scalar" is a quantity that can be specified by determining only its magnitude. However, the quantities that are specified by determining both magnitude and direction are called "vectors." In other words, a "vector" is a quantity that has both a direction and a magnitude of length; therefore, it is graphically denoted by an oriented line segment ("arrow"). In physics, vectors are very useful, because they can visually represent position, displacement, velocity, and acceleration. Moreover, vector graphics are used in computers, since they can be scaled to a larger size without losing any image quality.

If the coordinates of a point P in the coordinate plane are (x, y), and if we denote the origin of the coordinate system by $O(0,0)$, then a vector OP is denoted by $\overrightarrow{OP}$, since the length OP represents the magnitude, and the arrow represents the direction, as shown in Figure 7-1.

Figure 7-1: A vector in the xy-plane and its components (source: Wikimedia Commons: Author: JozumBjada; https://commons.wikimedia.org/wiki/File:Vector_in_2D_space_and_its_decomposition.png).



The column vector (matrix) corresponding to $\overrightarrow{0P}$ is $\begin{pmatrix} x \\ y \end{pmatrix}$.

Since the coordinates of point P are (x, y), the length from $O(0,0)$ to P is $\sqrt{x^2 + y^2}$, according to the Pythagorean Theorem. Notice that, frequently, we do not need to use arrows in order to indicate that letters represent vectors (in particular where there is no likelihood of confusion).

The operations between vectors are based on matrix algebra. For instance, given two vectors $\overrightarrow{OA} = \begin{pmatrix} p \\ q \end{pmatrix}$ and $\overrightarrow{OB} = \begin{pmatrix} r \\ s \end{pmatrix}$, their sum is a vector $\overrightarrow{OC}$ such that $\overrightarrow{OC} = \overrightarrow{OA} + \overrightarrow{OB} = \begin{pmatrix} p \\ q \end{pmatrix} + \begin{pmatrix} r \\ s \end{pmatrix} = \begin{pmatrix} p + r \\ q + s \end{pmatrix}$.

In general, we can define the following vector operations:

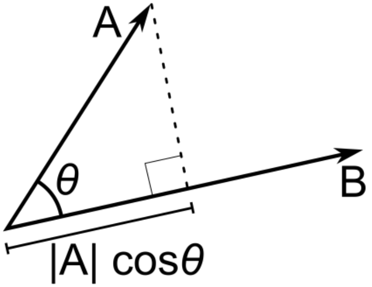
Vector addition: $\vec{u} + \vec{v} = (u_1 + v_1, u_2 + v_2, \dots, u_n + v_n)$. For instance, given two vectors $\vec{u}$ and $\vec{v}$ in $\mathbb{R}^2$, draw $\vec{u}$ (with its tail, that is, initial point, anywhere), and then draw $\vec{v}$ with its tail at the head (that is, the terminal point) of $\vec{u}$. Then $\vec{u} + \vec{v}$ is defined to be that vector that goes from the tail of $\vec{u}$ to the head of $\vec{v}$.

Scalar multiplication: $k\vec{u} = (ku_1, ku_2, \dots, ku_n)$, where: $\vec{u} = (u_1, u_2, \dots, u_n)$ is a vector in $\mathbb{R}^n$, and k is a real number (scalar). For

instance, given a vector $\vec{u}$, $2\,\vec{u}$ is a vector pointing in the same direction as $\vec{u}$ and twice as long as $\vec{u}$, whereas $-1.5\,\vec{u}$ is a vector pointing in the opposite direction from $\vec{u}$ and 1.5 times as long as $\vec{u}$. Two vectors are "parallel" if one of them is a scalar multiple of the other.
Negation: $-\vec{u} = (-1)\vec{u} = (-u_1, -u_2, \ldots, -u_n)$. To subtract vectors, switch the direction of the vector that is being subtracted, arrange the two vectors from "head to tail," and draw a resultant vector from the tail of the first vector to the head of the second vector; symbolically:
$\vec{u} - \vec{v} = \vec{u} + (-\vec{v})$.
Dot Product (or Scalar Product or Inner Product):
$\vec{u} \cdot \vec{v} = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n = \sum_{i=1}^{n} u_i v_i$ (i.e., the Euclidean inner product of two vectors $\vec{u}$ and $\vec{v}$ in $\mathbb{R}^n$ is the real number determined by multiplying the correspondent components of $\vec{u}$ and $\vec{v}$ and then summing the resulting products), where: $\vec{u} = (u_1, u_2, \ldots, u_n)$ and $\vec{v} = (v_1, v_2, \ldots, v_n)$ are vectors in $\mathbb{R}^n$.
Norm (Length): $||\vec{u}|| = \sqrt{\vec{u} \cdot \vec{u}} = \sqrt{u_1^2 + u_2^2 + \cdots + u_n^2}$
(specifically, the norm of a vector is the distance of the vector from the origin), where: $\vec{u} = (u_1, u_2, \ldots, u_n)$ is a vector in $\mathbb{R}^n$. When we divide a vector by its norm, we turn it into a "unit vector," and this process is called "normalization."
Notice that, as a result of the Cauchy–Schwarz–Bunyakovsky inequality, the absolute value of the dot product of two vectors is less than or equal to the product of their lengths; symbolically:
$||\vec{u}||\,||\vec{v}|| \geq |\vec{u} \cdot \vec{v}|$,
with equality if and only if there is a scalar λ such that $\vec{u} = \lambda\vec{v}$ or if one of the vectors is zero. This inequality can be easily proved as follows (method of C. C. Pugh): Notice that, $\forall \lambda \in \mathbb{R}$, the dot product
$(\lambda\vec{u} + \vec{v}) \cdot (\lambda\vec{u} + \vec{v})$
is always greater than or equal to zero; and consider the following polynomial of λ:
$(\lambda\vec{u} + \vec{v}) \cdot (\lambda\vec{u} + \vec{v}) = \lambda^2 ||\vec{u}||^2 + 2\lambda(\vec{u} \cdot \vec{v}) + ||\vec{v}||^2$.
This polynomial (which is of the form $a\lambda^2 + b\lambda + c$) must always be greater than or equal to zero, and, thus, it must have a non-positive discriminant, meaning that $(\vec{u} \cdot \vec{v})^2 \leq ||\vec{u}||^2 ||\vec{v}||^2$; *quod erat demonstrandum*.
The dot product is an operation on vectors that enables us to find the angle between two vectors, and, when we talk about the angle *between* two vectors, we are picturing the vectors *with their tails at the same point*. Thus, if θ is the angle between two vectors $\vec{u}$ and $\vec{v}$, then the dot product
$\vec{u} \cdot \vec{v} = ||\vec{u}||\,||\vec{v}|| cos\theta,$

where $||\vec{u}||$ denotes the norm of $\vec{u}$, and $||\vec{v}||$ denotes the norm of $\vec{v}$, as shown in Figure 7-2.

Figure 7-2: The dot product of two vectors (source: Wikimedia Commons: Author: Mazin07; https://commons.wikimedia.org/wiki/File:Dot_Product.svg).



Cross Product of two vectors in a 3-dimensional space: Consider two vectors $\vec{u} = (u_1, u_2, u_3)$ and $\vec{v} = (v_1, v_2, v_3)$, and let $\vec{\imath}$, $\vec{\jmath}$, and $\vec{k}$ be the unit vectors of the three coordinate axes, respectively. Then the cross product of $\vec{u}$ and $\vec{v}$ is a vector given by the following determinant:

$$\vec{u} \times \vec{v} = \begin{vmatrix} \vec{\imath} & \vec{\jmath} & \vec{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} = \begin{vmatrix} u_2 & u_3 \\ v_2 & v_3 \end{vmatrix} \vec{\imath} - \begin{vmatrix} u_1 & u_3 \\ v_1 & v_3 \end{vmatrix} \vec{\jmath} + \begin{vmatrix} u_1 & u_2 \\ v_1 & v_2 \end{vmatrix} \vec{k} =$$

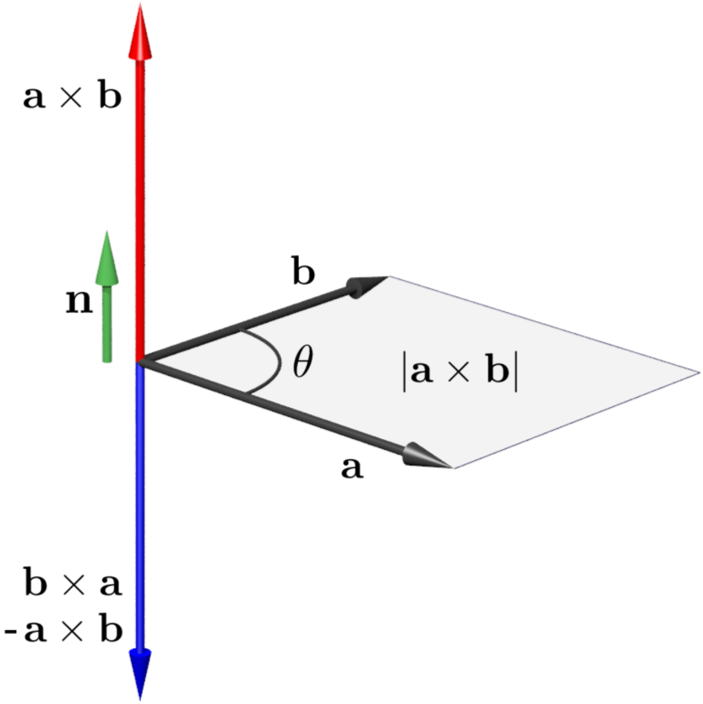
$$(u_2v_3 - u_3v_2)\vec{\imath} - (u_1v_3 - u_3v_1)\vec{\jmath} + (u_1v_2 - u_2v_1)\vec{k}.$$

The geometric significance of this operation is that, if θ is the angle between $\vec{u}$ and $\vec{v}$ with $0 \leq \theta \leq \pi$, then

$$\vec{u} \times \vec{v} = ||\vec{u}||\,||\vec{v}||(sin\theta)\vec{n},$$

where $\vec{n}$ is a unit vector perpendicular to the plane containing $\vec{u}$ and $\vec{v}$ (with your right hand, point your index finger along vector $\vec{u}$, and point your middle finger along vector $\vec{v}$; then $\vec{n}$ goes in the direction of your extended thumb), as shown in Figure 7-3. Obviously, if the vectors $\vec{u}$ and $\vec{v}$ are parallel (i.e., if the angle θ between them is either 0^o or 180^o), then $\vec{u} \times \vec{v}$ is equal to the zero vector.

Figure 7-3: The cross product of two vectors (source: Wikimedia Commons: Author: Svjo; https://commons.wikimedia.org/wiki/File:Cross-product-povray.png).


$\mathbf{a} \times \mathbf{b}$
$\mathbf{n}$
$\mathbf{b}$
θ
$|\mathbf{a} \times \mathbf{b}|$
$\mathbf{a}$
$\mathbf{b} \times \mathbf{a}$
$-\mathbf{a} \times \mathbf{b}$

The magnitude of the cross product ($|\vec{u} \times \vec{v}|$) can be interpreted as the positive area of the parallelogram having $\vec{u}$ and $\vec{v}$ as its sides. Whilst the resultant of the dot product of two vectors $\vec{u}$ and $\vec{v}$ is a scalar quantity, the cross product of two vectors $\vec{u}$ and $\vec{v}$ is a third vector whose direction is perpendicular to both $\vec{u}$ and $\vec{v}$ (the direction is given by the aforementioned right-hand rule).

Notice that we can write the equation of a straight line in three dimensions using vector notation as follows: Let $\vec{a}$ and $\vec{b}$ be the radius (or position) vectors of two points A and B, respectively, with respect to some origin. Then the condition for an arbitrary point P with radius vector $\vec{r}$ to lie on the straight line going through A and B is that the vectors $\vec{r} - \vec{a}$ and $\vec{b} - \vec{a}$

be parallel, that is, $\vec{r}-\vec{a}=\lambda(\vec{b}-\vec{a})$; and, therefore, if λ is regarded as a parameter, then the equation of the straight line takes the form
$\vec{r}=\vec{a}+\lambda(\vec{b}-\vec{a})$.
The parameter λ can be eliminated by taking the cross product of the aforementioned expression with $(\vec{b}-\vec{a})$, thus obtaining
$(\vec{r}-\vec{a})\times(\vec{b}-\vec{a})=0\Rightarrow\vec{r}\times(\vec{b}-\vec{a})=\vec{a}\times\vec{b}$.
By analogy, we can write the equation of a plane in vector form as follows: Let $\vec{a}$, $\vec{b}$, and $\vec{c}$ be the radius vectors of three given points A, B, and C, respectively. In order to find the equation of the plane going through these three points, we think as follows: since the vectors $\vec{r}-\vec{a}$, $\vec{b}-\vec{a}$, and $\vec{c}-\vec{a}$ are coplanar, the required equation is
$\vec{r}-\vec{a}=\lambda(\vec{b}-\vec{a})+\mu(\vec{c}-\vec{a})$,
where λ and μ are parameters. In order to eliminate the parameters λ and μ, we firstly take the cross product of the aforementioned expression with $\vec{c}-\vec{a}$ and then the dot product with $\vec{b}-\vec{a}$, thus obtaining
$[(\vec{r}-\vec{a})\times(\vec{c}-\vec{a})]\cdot(\vec{b}-\vec{a})=0$.

Vector (or Linear) Spaces

The most abstract definition of a vector is that a vector is an element of a "vector (or linear) space," which, in turn, can be defined as follows: let U be a set endowed with two operations: addition and scalar multiplication, defined in the following way:
$+: U\times U\to U$ defined by $(u,v)\in U\times U\to u+v\in U$ for all $u,v\in U$, that is, U is "closed under addition";
$\cdot: k\times U\to U$ defined by $(k,u)\in K\times U\to k\cdot u\in U$ for every $k\in K$ (where K is a field, such as $\mathbb{R}$) and for every $u\in U$, that is, U is "closed under scalar multiplication." Of course, $0\in U$, since, for every $u\in U$, $(-1)u\in U$, and, therefore, $u-u\in U\Rightarrow 0\in U$. As a result of the aforementioned definition, we say that U under the operations of $+$ (addition) and $\cdot$ (scalar multiplication) forms a "vector space" (or "linear space") over the field K; and, therefore, a "vector" can be defined as an element of such a U.
For instance, we can prove that, if
$V=\{ax^2+bx+c|a,b,c\in\mathbb{R}\}$,
then V is a vector space over $\mathbb{R}$ as follows:
Step 1: 0$=0x^2+0x+0\in V$.
In other words, $0\in V$.
Step 2: Let

$$\begin{cases} v_1 = a_1x^2 + b_1x + c_1 \\ v_2 = a_2x^2 + b_2x + c_2 \end{cases}.$$

Then $v_1 + v_2 = (a_1 + a_2)x^2 + (b_1 + b_2)x + (c_1 + c_2) \in V$.

In other words, V is closed under addition.

Step 3: Let $v = ax^2 + bx + c$ with $a, b, c \in \mathbb{R}$.

Then $kv = (ka)x^2 + (kb)x + (kc) \in V$.

In other words, V is closed under scalar multiplication.

Therefore, $V = \{ax^2 + bx + c | a, b, c \in \mathbb{R}\}$ is a vector space over $\mathbb{R}$. In other words, the set of all real quadratic polynomials forms a vector space over $\mathbb{R}$.

On the other hand, we can prove that a sphere S is not a vector space as follows: let v be a vector belonging to the sphere S. If we multiply v by an adequate number k, then kv does not belong to S any more (it "pierces" the sphere). Hence, a sphere is not a vector space (it is not closed under scalar multiplication). This example helps us to understand why no bounded set, in general, is a vector space.

Linearly Independent Vectors: Let V be a vector space over K. The vectors $v_1, v_2, \dots, v_n$ of V are "linearly independent" if and only if every time $k_1v_1 + k_2v_2 + \cdots + k_nv_n = 0 \Rightarrow k_1 = k_2 = \cdots = k_n = 0$.

For instance, the vectors $v_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, $v_2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, $v_3 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$, and $v_4 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$ are linearly independent, since

$$k_1v_1 + k_2v_2 + \cdots + k_nv_n = 0$$
$$\Rightarrow \begin{pmatrix} k_1 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & k_2 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ k_3 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & k_4 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$
$$\Rightarrow \begin{pmatrix} k_1 & k_2 \\ k_3 & k_4 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \Rightarrow k_1 = k_2 = k_3 = k_4 = 0.$$

Linearly Dependent Vectors: Let V be a vector space over K. The vectors $v_1, v_2, \dots, v_n$ of V are "linearly dependent" if and only if $k_1v_1 + k_2v_2 + \cdots + k_nv_n = 0$ for some $k_i \neq 0$, where $i = 1,2, \dots, n$.

For instance, the vectors $v_1 = (0,1)$, $v_2 = (1,0)$, and $v_3 = (1,1)$ are linearly dependent.

Basis: Let V be a vector space over K. The vectors $v_1, v_2, \dots, v_n$ form a "basis" of V if and only if these vectors are linearly independent and generate (or span) V; that is, every vector of V must be expressed in terms of $v_1, v_2, \dots, v_n$. For instance, if $V = \{a + bx + cx^2 | a, b, c \in \mathbb{R}\}$, then $v_1 = 1$, $v_2 = x$, and $v_3 = x^2$ form a basis of V, because: (i) v_1, v_2, and v_3 are linearly independent, since no vector from $\{1, x, x^2\}$ can be written in terms of the other vectors; and (ii) $\{1, x, x^2\}$ generate V, since, for any $v \in V$, it holds that $v = k + lx + mx^2 = k \cdot 1 + lx + mx^2$. Every (non-zero) vector space over a field K has at least one basis (actually, it has many different bases). We shall see that all bases of a finite-dimensional vector space have the same length (i.e., the same number of elements), and this length is said to be the "dimension" of the corresponding vector space.

A list of vectors $(v_1, \dots, v_n)$ is linearly independent if and only if every vector $v \in span(v_1, \dots, v_n)$, that is, every vector belonging to the vector space spanned by $(v_1, \dots, v_n)$, can be uniquely written as a linear combination of $(v_1, \dots, v_n)$: If $(v_1, \dots, v_n)$ is a linearly independent list of vectors, then, for the sake of contradiction, suppose that there are two ways of writing $v \in span(v_1, \dots, v_n)$ as a linear combination of $(v_1, \dots, v_n)$, say

$v = k_1 v_1 + \cdots + k_n v_n$, and

$v = k_1' v_1 + \cdots + k_n' v_n$.

Subtracting these two equations by parts, we obtain $0 = (k_1 - k_1')v_1 + \cdots + (k_n - k_n')v_n$. Thus, every vector belonging to the vector space spanned by the linearly independent list of vectors $(v_1, \dots, v_n)$ can be uniquely written as a linear combination of $(v_1, \dots, v_n)$. Now, we shall prove the converse: Suppose that, for every $v \in span(v_1, \dots, v_n)$, there are unique $k_1, \dots, k_n \in F$ (where F is a field) such that $v = k_1 v_1 + \cdots + k_n v_n$. Then the only way in which the zero vector $v = 0$ can be written as a linear combination of $v_1, \dots, v_n$ is with $k_1 = \cdots = k_n = 0$, and this fact implies that $(v_1, \dots, v_n)$ are linearly independent.

If V is a finite-dimensional vector space, and if $(v_1, \dots, v_m)$ is a linearly independent list of vectors that spans V, then, given any list $(w_1, \dots, w_n)$ that also spans V, it holds that $m \leq n$. Notice that a list of vectors is linearly independent if and only if removing any vector from the list yields a list whose span is strictly smaller than that of the original list, and, therefore, a linearly independent list is minimal for its span (such a list does not have any linear redundancies). On the other hand, a spanning set for a vector space V is generally a list of vectors in V such that every vector of V is in the span of the list, so that the last proposition means that spanning sets have to be at least as large as linearly independent sets ("bases"). Indeed, this proposition can be verified as follows: Consider an arbitrary list of vectors $A_0 = (w_1, \dots, w_n)$ such that $V = span(A_0)$. At the kth step of the procedure, construct a new list A_k by replacing some vector w_{jk} with the vector v_k such that A_k still spans V. Repeating the same process for every v_k, we obtain a new list A_m of length n that contains each of the vectors $v_1, \dots, v_m$, and, therefore, $m \leq n$.

Using the last proposition, we can prove that every vector space V has the following invariant property: the number of vectors in every basis of V remains the same (and, thus, this number is said to be the "dimension" of

V, and it is denoted by $dim(V)$). Indeed, if $(v_1, \dots, v_m)$ and $(w_1, \dots, w_n)$ are two bases of V, then, due to the last proposition, we have $m \leq n$, since $(v_1, \dots, v_m)$ is linearly independent (given that we assumed that it is a basis of V), and $n \leq m$, since $(w_1, \dots, w_n)$ is linearly independent (given that we assumed that it is a basis of V); and, therefore, $m = n$.

By the definition of the direct (or Cartesian) product of two sets, it can be easily verified that, if $X_1, X_2, \dots, X_n$ are finite-dimensional vector spaces over the same field, then $X_1 \times X_2 \times \dots \times X_n$ is finite-dimensional and
$dim(X_1 \times X_2 \times \dots \times X_n) = dim(X_1) + dim(X_2) + \cdots + dim(X_n)$.

Let A and B be non-empty subsets of a vector space V. The "sum" of A and B, denoted by $A + B$, is the set of all possible sums of elements from both sets: $A + B = \{a + b | a \in A, b \in B\}$.

By the definition of a basis of a vector space, it can be easily verified that, if X and Y are two subspaces of a vector space V over a field F, and if B_X is a basis of X and B_Y is a basis of Y, then $B_X \cup B_Y$ is a basis of $X + Y$. Notice that the union $B_X \cup B_Y$ may contain linearly dependent elements, and, therefore, if X and Y are subspaces of a vector space V, then
$dim(X + Y) = dim(X) + dim(Y) - dim(X \cap Y)$.

Direct sum decompositions: Let U and W be subspaces of a vector space V. Then V is said to be the "direct sum" of U and W, and we write $V = U \oplus W$, if and only if $V = U + W = \{v = u + w | u \in U, w \in W\}$ and $U \cap W = \{0\}$. In other words, the "direct sum" is a way of adjoining two (or more) vector spaces in order to obtain a larger vector space, and the condition that $U \cap W = \{0\}$ implies that every such $v \in V$ has a unique expression as $v = u + w$ with $u \in U, w \in W$. Hence, given two subspaces U and W of a vector space V, $V = U \oplus W$ if and only if, for every $v \in V$, there exist unique vectors $u \in U$ and $w \in W$ such that $v = u + w$.

Norms and Normed Vector Spaces

When we study vector spaces, we must keep in mind that the term "space" signifies a collection of vectors that interact in a certain way, which is determined by the corresponding structure (e.g., by a set of operations, by a norm, etc.). We can define a norm in an abstract way as follows: given a vector (or linear) space X over $\mathbb{R}$, a "norm" $\|\cdot\|$ for X is a function on X that assigns to each element a real number (symbolically: $\|\cdot\|: X \to \mathbb{R}$) satisfying the following properties:
for every $x \in X$:

i. $\|x\| \geq 0$,
ii. $\|x\| = 0$ if and only if $x = 0$,
iii. $\|kx\| = |k|\|x\|$ for any scalar k, and,

for every $x, y \in X$,

iv. $\|x + y\| \leq \|x\| + \|y\|$ (the triangle inequality, implying that the shortest path between two points is a line segment; equality holds whenever one of x and y is a non-negative multiple of the other).

A vector (or linear) space that is equipped with a norm $\|\cdot\|$ is denoted by $(X, \|\cdot\|)$ and is called a "normed vector space" (or "normed linear space"). Different norms can be defined on the same vector space, thus giving rise to different normed vector spaces.

Example 1: $(\mathbb{R}, |\cdot|)$. The set of real numbers ($\mathbb{R}$) is a normed vector space with norm given by the absolute value (or modulus), that is,
$\|x\| = |x|$,
and we call this the "usual norm" for $\mathbb{R}$.

Example 2: $(\mathbb{R}^n, \|\cdot\|_2)$. The set of ordered n-tuples of real numbers ($\mathbb{R}^n$) is a normed vector space with norm $\|\cdot\|_2$ defined as follows:
for any real vector $x = (k_1, k_2, \dots, k_n)$,
$\|x\|_2 = \sqrt{|k_1|^2 + |k_2|^2 + \cdots + |k_n|^2}$,
and we call this the "Euclidean norm" (notice that, in this case, the only norm property that provides any difficulty to verify is the triangle inequality; we can show that $\|x\|_2$ satisfies the triangle inequality by using the Cauchy–Schwarz–Bunyakovsky Inequality and the Minkowski Inequality).

Example 3: $(\mathbb{R}^n, \|\cdot\|_1)$. The set of ordered n-tuples of real numbers ($\mathbb{R}^n$) is a normed vector space with norm $\|\cdot\|_1$ defined as follows:
for any real vector $x = (k_1, k_2, \dots, k_n)$,
$\|x\|_1 = |k_1| + |k_2| + \cdots + |k_n|$.

Example 4: $(\mathbb{R}^n, \|\cdot\|_\infty)$. The set of ordered n-tuples of real numbers ($\mathbb{R}^n$) is a normed vector space with norm $\|\cdot\|_\infty$ defined as follows:
for any real vector $x = (k_1, k_2, \dots, k_n)$,
$\|x\|_\infty = max\{|k_i|, where\ i = 1,2, \dots, n\}$,
and we call this the "supremum (or uniform) norm" for $\mathbb{R}^n$.

Example 5: $(\mathcal{B}(X), \|\cdot\|_\infty)$. For any non-empty set X, we denote by $\mathcal{B}(X)$ the set of bounded real functions on X. Notice that a function f on some set X with real values is said to be "bounded" if the set of its values is bounded—that is, if there exists a real number M such that, for every $x \in X$, it holds that $|f(x)| \leq M$.

$\mathcal{B}(X)$ is a real vector space under the pointwise definitions of addition and scalar multiplication. Moreover, $\mathcal{B}(X)$ is a normed vector space with norm $\|\cdot\|_\infty$ defined by
$\|f\|_\infty = sup\{|f(x)|, where\ x \in X\}$,
and we call this the "supremum (or uniform) norm" for $\mathcal{B}(X)$. Notice that Example 4 is the special case when $X = \{1,2, \dots, n\}$.

Example 6: l_2-space, also known as the "Hilbert (sequence) space." This is a generalization of the Euclidean n-space. The set l_2, whose elements are sequences of scalars (real numbers) $x = \{k_1, k_2, \dots, k_n, \dots\}$ such that $\sum |k_n|^2$ is convergent, is a real vector space under the pointwise definitions of addition and scalar multiplication, and it is a normed vector space with norm $\|\cdot\|_2$ defined by
$\|x\|_2 = \sqrt{\sum_{i=1}^{\infty} |k_i|^2}$;
where the only norm property that provides any difficulty to verify is the triangle inequality, and we can show that $\|x\|_2$ satisfies the triangle inequality by using the Cauchy–Schwarz–Bunyakovsky Inequality and the Minkowski Inequality.
In an arbitrary normed vector space $(X, \|\cdot\|)$, the set

$$S(0;1) = \{x \in X \; such \; that \; \|x\| = 1\}$$

is called the "unit sphere"; the set

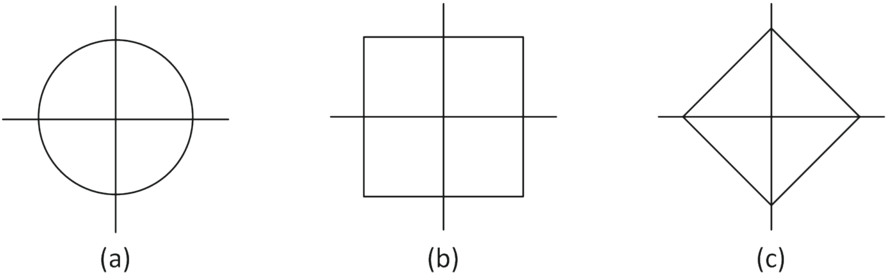
$$B[0;1] = \{x \in X \; such \; that \; \|x\| \leq 1\}$$

is called the "closed unit ball"; and the set

$$B(0;1) = \{x \in X \; such \; that \; \|x\| < 1\}$$

is called the "open unit ball." In Figure 7-4, we consider the shape of the unit sphere in several coordinate space examples: (a) in $(\mathbb{R}^2, \|\cdot\|_2)$, where $S\big((0,0);1\big) = \{(k,l) \; such \; that \; k^2 + l^2 = 1\}$ (i.e., here, we have the graph of $\sqrt{x^2 + y^2} = 1$, which is the unit circle); (b) in$(\mathbb{R}^2, \|\cdot\|_\infty)$, where $S\big((0,0);1\big) = \{(k,l) \; such \; that \; max\{|k|, |l|\} = 1\}$ (i.e., here, we have the infinity-norm for two elements, which is the maximum value of the two elements, and we require that it is equal to one, so that we end up with the square with the corners at $(1,1)$, $(1,-1)$, $(-1,-1)$, and $(-1,1)$); and (c) in $(\mathbb{R}^2, \|\cdot\|_1)$, where $S\big((0,0);1\big) = \{(k,l) \; such \; that \; |k| + |l| = 1\}$ (i.e., here, we have the one-norm for two values, which is the sum of their absolute values, and we require that one is the magnitude, and, therefore: in the first quadrant, we have the graph of the equation $y = 1 - x$; in the second quadrant, we have the graph of the equation $y = 1 + x$, since x is negative there, and we change the sign; in the third quadrant, we have the graph of the equation $y = -x - 1$; and, in the fourth quadrant, we have the graph of the equation $y = x - 1$).

Figure 7-4: The shape of the unit sphere in: (a) $(\mathbb{R}^2, \|\cdot\|_2)$*, (b)* $(\mathbb{R}^2, \|\cdot\|_\infty)$*, and (c)* $(\mathbb{R}^2, \|\cdot\|_1)$.



Linear Transformations

Linear transformations are transformations (functions) that preserve the operations of vector addition and scalar multiplication. Thus, a transformation T is "linear" if and only if

i. $T(\vec{u}+\vec{v}) = T(\vec{u}) + T(\vec{v})$ and

ii. $T(c\vec{u}) = cT(\vec{u})$, where c is a scalar quantity.

Remark: If T is a linear transformation, then $T(\vec{0}) = \vec{0}$.

Example 1: Recall that, when we multiply an $m \times n$ matrix by an $n \times 1$ column vector (which is an element of $\mathbb{R}^n$), we receive an $m \times 1$ column vector (which is an element of $\mathbb{R}^m$). If A is any $m \times n$ matrix, then the mapping $T: \mathbb{R}^n \to \mathbb{R}^m$ which is matrix-vector multiplication

$$T(\vec{x}) = A\vec{x}$$

is a linear transformation. In fact, every linear transformation can be expressed as a matrix transformation.

Example 2: Projection is a linear transformation. In particular, in $\mathbb{R}^2$, a projection is a linear transformation $T: \mathbb{R}^2 \to \mathbb{R}^2$, which takes every vector in the plane into a vector in the plane. The "vector projection" of $\vec{v}$ onto $\vec{u}$ is denoted by $proj_{\vec{u}}\vec{v}$, and it is defined as follows:

$$proj_{\vec{u}}\vec{v} = \left(\frac{\vec{v}\cdot\vec{u}}{\|\vec{u}\|^2}\right)\vec{u}$$

where the operator $\cdot$ denotes the dot product, and $\|\vec{u}\|$ is the length of $\vec{u}$. This formula indicates that the new vector is going in the direction of $\vec{u}$ (notice that the vector projection is the vector produced when one vector is resolved into two component vectors, one that is parallel to the second vector and one that is perpendicular to the second vector). The "scalar projection" of of $\vec{v}$ onto $\vec{u}$ is equal to

$$v_1 = ||\vec{v}\,||cos\theta$$

where θ is the angle between $\vec{v}$ and $\vec{u}$ (notice that the scalar projection is the length of the vector projection); and recall that $cos\theta = \frac{\vec{v}\cdot\vec{u}}{||\vec{v}\,||||\vec{u}\,||}$.

Example 3: Rotation is a linear transformation. In particular, in $\mathbb{R}^2$, we write $Rot_\theta: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ for the linear transformation that rotates vectors in $\mathbb{R}^2$ counter-clockwise through the angle θ about the origin of the Cartesian coordinate system. Its matrix is

$$\begin{pmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{pmatrix}$$

and, to perform the rotation on a plane point with standard coordinates $\vec{v} = (x, y)$, it should be written as a column vector and multiplied by $Rot_\theta: \mathbb{R}^2 \rightarrow \mathbb{R}^2$, namely:

$$Rot_\theta\vec{v} = \begin{pmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} xcos\theta - ysin\theta \\ xsin\theta + ycos\theta \end{pmatrix}.$$

The kernel of a linear transformation: The “kernel” (or “null space”) of a linear transformation is the subset of the domain that is transformed into the zero vector. In formal notation, the kernel of a linear transformation $T: V \rightarrow W$ is denoted by $ker(T)$, and it is the set of all input vectors $\vec{v} \in V$ such that $T(\vec{v}) = \vec{0}$. The kernel is a measure of injectivity. In fact, since the kernel consists of the elements sent to $\vec{0}$, the dimension of the kernel tells us how much the corresponding linear transformation shrinks the source space into the target space. Hence, a linear transformation is injective if and only if its kernel is trivial, that is, if and only if its kernel is the singleton of $\vec{0}$.

Eigenvectors and eigenvalues: In linear algebra, we often need to know which vectors have their directions unchanged by a linear transformation. An “eigenvector” (or “characteristic vector”) is such a vector. Hence, an eigenvector $\vec{v}$ of a linear transformation T is merely scaled by a constant factor λ when the linear transformation is applied to it; symbolically, $T(\vec{v}) = \lambda\vec{v}$. The corresponding “eigenvalue” (or “characteristic value”) is the multiplying factor λ. In other words, if T is a linear transformation from a vector space V over a field F into itself and $\vec{v}$ is a non-zero vector in V, then $\vec{v}$ is an eigenvector of T if $T(\vec{v})$ is a scalar multiple of $\vec{v}$, that is, $T(\vec{v}) = \lambda\vec{v}$ where λ is a scalar in F, and then λ is said to be the eigenvalue associated with $\vec{v}$.

Let A be an $n \times n$ matrix, and let $X \in \mathbb{R}^n$ be a non-zero vector for which

$$AX = \lambda X$$

for some scalar λ. Then λ is said to be the eigenvalue of the matrix A, and X is said to be an eigenvector of A associated with λ. If this is the case, then

$$AX - \lambda X = 0 \Leftrightarrow (A - \lambda I)X = 0$$

where I is the corresponding identity matrix. Therefore, when we have to find eigenvectors, we have to find the non-trivial solutions to this homogeneous system of equations. The expression (determinant)

$det(A - \lambda I)$

is a polynomial called the "characteristic polynomial" of A; that is, it contains the eigenvalues as roots, and it is invariant under matrix similarity. In other words, the solutions to the "characteristic equation"

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

are the eigenvalues.

For instance, let us find the eigenvalus of the matrix

$A = \begin{pmatrix} 2 & 2 \\ 5 & -1 \end{pmatrix}$,

and then let us find the corresponding eigenvectors.

In order to find the eigenvalues of A, we have to find those λ for which $det(A - \lambda I) = 0$. In this case,

$$\begin{aligned} det(A - \lambda I) &= det\left(\begin{pmatrix} 2 & 2 \\ 5 & -1 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) \\ &= det\left(\begin{pmatrix} 2 & 2 \\ 5 & -1 \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}\right) = \begin{vmatrix} 2-\lambda & 2 \\ 5 & -1-\lambda \end{vmatrix} \\ &= (2-\lambda)(-1-\lambda) - 10 = \lambda^2 - \lambda - 12 \end{aligned}$$

meaning that the eigenvalues of A are the solutions to the quadratic equation $\lambda^2 - \lambda - 12 = 0$, namely, $\lambda_1 = -3$ and $\lambda_2 = 4$.

Case I: Let $\lambda = -3$. Then

$Ax = \lambda x \Rightarrow Ax = -3x.$ (1)

If we write

$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$,

then, given the definition of matrix A, we obtain:

$Ax = \begin{pmatrix} 2 & 2 \\ 5 & -1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2x_1 + 2x_2 \\ 5x_1 - x_2 \end{pmatrix}.$ (2)

Moreover,

$-3x = \begin{pmatrix} -3x_1 \\ -3x_2 \end{pmatrix}.$ (3)

Because of equation (1), (2) is equal to (3), and, therefore, we get

$$\begin{pmatrix} 2x_1 + 2x_2 \\ 5x_1 - x_2 \end{pmatrix} = \begin{pmatrix} -3x_1 \\ -3x_2 \end{pmatrix}$$

meaning that

$2x_1 + 2x_2 = -3x_1 \Rightarrow 5x_1 = -2x_2 \Rightarrow x_1 = -\frac{2}{5}x_2,$

and

$5x_1 - x_2 = -3x_2.$

This result means that there are infinitely many solutions to the equation $Ax = -3x$, but they all satisfy the condition that the first entry x_1 is $-\frac{2}{5}$ times the second entry x_2. All the solutions to this equation have the following pattern:

$$\begin{pmatrix} 2t \\ -5t \end{pmatrix} = t\begin{pmatrix} 2 \\ -5 \end{pmatrix}$$

where t is any real number. The non-zero vectors x that satisfy equation (1) are the eigenvectors associated with the eigenvalue $\lambda = -3$. One such eigenvector is

$\vec{v}_1 = \begin{pmatrix} 2 \\ -5 \end{pmatrix}$,

and every other eigenvector associated with the eigenvalue $\lambda = -3$ is a scalar multiple of $\vec{v}_1$, that is, $\vec{v}_1$ spans this set of eigenvectors.

Case II: Let $\lambda = 4$. Then, in a similar way, we can find eigenvectors associated with the eigenvalue $\lambda = 4$ by solving the equation

$Ax = 4x$,

which implies that

$$\begin{pmatrix} 2x_1 + 2x_2 \\ 5x_1 - x_2 \end{pmatrix} = \begin{pmatrix} 4x_1 \\ 4x_2 \end{pmatrix}$$

meaning that

$2x_1 + 2x_2 = 4x_1 \Rightarrow x_1 = x_2$,

and

$5x_1 - x_2 = 4x_2$.

This means that the set of eigenvectors associated with the eigenvalue $\lambda = 4$ is spaned by the vector

$\vec{v}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

Isomorphisms: Let U and V be two vector spaces over the same field K. Then a linear transformation $T: U \to V$ is called an "isomorphism" if and only if T is one-to-one and onto; and, in this case, the vector spaces U and V are said to be "isomorphic." In general, in mathematics, an "isomorphism" is a bijective function (one-to-one correspondence) between two structures that preserves the operations of the structures.

If V_n is an n-dimensional vector space over $\mathbb{R}$ with basis $\{v_1, v_2, \ldots, v_n\}$, then let us define a mapping $f: V_n \to \mathbb{R}^n$ by

$f(a_1v_1 + \cdots + a_nv_n) \to (a_1, \ldots, a_n)$.

Then it can be easily shown that $f: V_n \to \mathbb{R}^n$ is linear, one-to-one, and onto. Hence, an n-dimensional vector space V_n over $\mathbb{R}$ is isomorphic to $\mathbb{R}^n$.

Hyperplanes

Notice that, in a three-dimensional space, a plane is given by a linear equation such as $ax + by + cz + d = 0$ where a, b, and c are the components of the normal vector $\vec{n} = (a, b, c)$, which is perpendicular to the plane or to any vector parallel to the plane. Moreover, in a three-dimensional space, a straight line can be specified as the intersection of two planes, and, thus, it is given by two such linear equations; specifically: $\{(x, y, z) \in \mathbb{R}^3 | a_1x + b_1y + c_1z = d_1 and a_2x + b_2y + c_2z = d_2\}$.

Given a non-zero vector $\vec{n} = (a, b, c)$ and a point $p_1 = (x_1, y_1, z_1)$, if $p = (x, y, z)$ is an arbitrary point of $\mathbb{R}^3$, then, expanding the "scalar equation," we obtain

$0 = a(x - x_1) + b(y - y_1) + c(z - z_1) = ax + by + cz - (ax_1 + by_1 + cz_1)$,

and, by setting $ax_1 + by_1 + cz_1 = d$, we obtain the "linear equation" $ax + by + cz = d$. Conversely, if (x_1, y_1, z_1) lies on the plane with linear equation $ax + by + cz = d$, where $d = ax_1 + by_1 + cz_1$, then we obtain the scalar equation $a(x - x_1) + b(y - y_1) + c(z - z_1) = 0$. Each can be written as a vector equation:

the scalar form can be written as $0 = \vec{n} \cdot (p - p_1)$, and

the linear form can be written as $\vec{n} \cdot p = \vec{n} \cdot p_1$.

Both equations describe the plane in $\mathbb{R}^3$ through the point p_1 and with normal vector $\vec{n}$.

Generally, a k-dimensional plane, usually called a "hyperplane," in an n-dimensional space is the geometric locus of the points whose coordinates satisfy a system of $n - k$ linear equations, such as:

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n + b_1 &= 0 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n + b_2 &= 0 \\ &\vdots \\ a_{n-k,1}x_1 + a_{n-k,2}x_2 + \cdots + a_{n-k,n}x_n + b_{n-k} &= 0 \end{aligned}\right\} \qquad (1)$$

provided that these equations are consistent and independent. Each of these equations represents an $(n - 1)$-dimensional hyperplane, and all together determine the common points of $n - k$ hyperplanes. Hence, a k-dimensional hyperplane is determined by the intersection of $n - k$ hyperplanes of dimension $n - 1$.

An important property of a k-dimensional hyperplane is the fact that it is a k-dimensional space. For instance, a 3-dimensional hyperplane is the ordinary 3 -dimensional space. Therefore, we can generalize results concerning n-dimensional spaces to $(n + 1)$-dimensional spaces.

If equations (1) are consistent and independent, then, by simple algebraic techniques, we can choose k out of the n variables x_i and express the remaining $n - k$ variables as functions of these, namely:

$$\left.\begin{aligned} x_{k+1} &= c_{11}x_1 + c_{12}x_2 + \cdots + c_{1k}x_k + d_1 \\ x_{k+2} &= c_{21}x_1 + c_{22}x_2 + \cdots + c_{2k}x_k + d_2 \\ &\vdots \\ x_n &= c_{n-k,1}x_1 + c_{n-k,2}x_2 + \cdots + c_{n-k,k}x_k + d_{n-k} \end{aligned}\right\}$$

where the variables $x_1, x_2, \dots, x_k$ admit arbitrary values, and the rest of the x_i's are determined by these. Hence, the position of a point in a k-dimensional hyperplane is determined by k coordinates.

Spherical Geometry and Hyperbolic Geometry

Using analytic geometry, we can define a "solid sphere" with center (x_0, y_0, z_0) and radius r as a solid bounded by a surface given by the locus of all points (x, y, z) such that $(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2 = r^2$. The straight line that joins any point of this surface with the center is called a "radius," and a straight line drawn through the center and terminated both ways by this surface is called a "diameter."

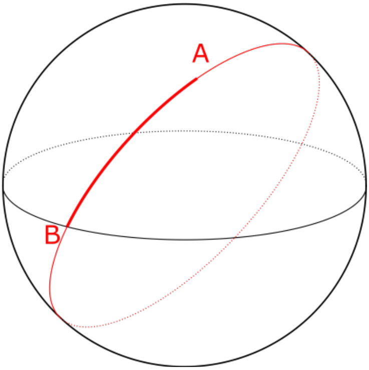
Moreover, the equation of the surface of a solid sphere with center C and radius a can be expressed in vector form as follows: Let O be the origin of the Cartesian coordinate system. Let $\vec{c}$ be a vector such that its tail is O, its head is C, and its magnitude is OC. Let P be an arbitrary point on the surface of the solid sphere, and let $\vec{r}$ be a position vector whose magnitude is OP (i.e., it indicates the location of a point on the surface of the solid sphere with respect to the origin of the Cartesian coordinate system). Obviously, $CP = a$ (the radius). Then the vector equation of the surface of a solid sphere is

$|\vec{r} - \vec{c}|^2 = a^2$,

and, thus, a point P lies on the solid sphere if and only if its position vector $\vec{r}$ satisfies this condition.

The section of the surface of a solid sphere made by any plane is a "circle." The section of the surface of a solid sphere by a plane is called a "great circle" if the plane passes through the center of the solid sphere, and it is called a "small circle" if the plane does not pass through the center of the solid sphere. Hence, the radius of a great circle is equal to the radius of the corresponding solid sphere, and the radius of a small circle is less than the radius of the corresponding solid sphere. See, for instance, Figure 7-5.

Figure 7-5: A great circle passing through two points (source: Wikimedia commons: Author: HaEr48; https://commons.wikimedia.org/wiki/File:Great_circle_passing_through_two_points.svg).



Through the center of a solid sphere and any two points on the surface, we can draw a plane, and, in fact, this plane is unique, unless the two points are the extremities of a diameter of the solid sphere, in which case infinitely many such planes can be drawn. Therefore, only one great circle can be drawn through two given points on the surface of a solid sphere, unless the points are the extremities of a diameter of the solid sphere. When only one great circle can be drawn through two given points, the great circle is uniquely divided at the two points, and the shorter of the two arcs is said to be the "arc of a great circle joining the two points," such as, for instance, the arc AB in Figure 7-5.

The "axis" of any circle of a solid sphere is that diameter of the solid sphere which is perpendicular to the plane of the circle, and the extremities of the axis are called the "poles" of the circle. The poles of a great circle are equidistant from the plane of the circle, whereas the poles of a small circle are not equidistant from the plane of the circle. But a pole of a circle is always equidistant from every point of the circumference of the circle.

The arc of a great circle that is drawn from a pole of a great circle to any point in its circumference is a quadrant (a quarter of a circle). The angle subtended at the center of a solid sphere by the arc of a great circle joining the poles of two great circles is equal to the inclination of the planes of the great circles. The angle between two great circles is defined as the "angle of inclination of the planes of the circles." Two great circles bisect each other.

Assume that the the arcs of great circles join a point P on the surface of a solid sphere with two other points A and B on the surface of the solid sphere, which are not at opposite extremities of a diameter, in such a way that each of these arcs is equal to a quadrant. Then P is a pole of the great circle through A and B.

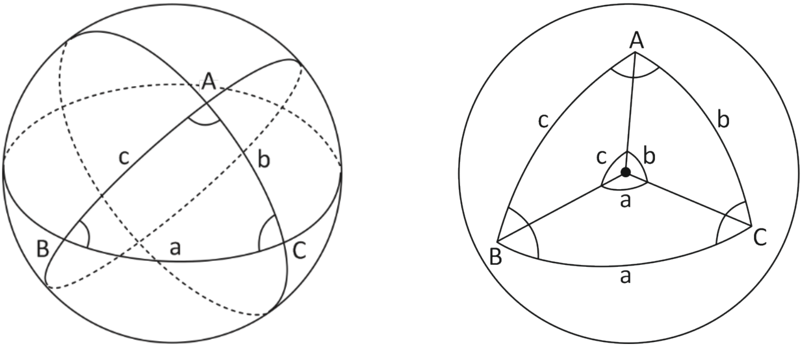
The great circles that pass through the poles of a given great circle are said to be "secondaries" to the given great circle. The angle between any two great circles is measured by the arc that they intercept on the great circle to which they are secondaries.

Assume that, from a point P on the surface of a solid sphere, there can be drawn two arcs of great circles, so that they are not parts of the same great circle, and the corresponding planes are at right angles to the plane of a given circle (i.e., the line in which they intersect is perpendicular to the plane of the given circle, and, therefore, it is the axis of the given circle). Then that point P is a pole of the given circle.

In summary, on the surface of a solid sphere, the "lines" can be interpreted as geodesics: a "geodesic" is the shortest path between two points on a curved surface (i.e., the equivalent of a Euclidean straight line in the context of spherical geometry); like, for instance, on the surface of the Earth (e.g., airplanes, wishing to minimize the time that they spend on the air, do not follow Euclidean straight lines, but they follow shortest curves known as geodesics). In spherical geometry, "great circles," or "geodesics," are intersections with planes through the center of the sphere. Thus, it is not unconditionally true that, given any two points, there is a unique line through them, because, if one chooses two points on the surface of a solid sphere that are opposite, or "antipodal," then there is a whole family of great circles that go through them.

Suppose that the angular point of a solid angle is made the center of a solid sphere. Then the planes that form the solid angle cut the solid sphere in arcs of great circles, and the figure that is formed on the surface of the solid sphere is called a "spherical triangle" if it is formed by the meeting of three plane angles, that is, if it is bounded by three arcs of great circles, as shown, for instance, in Figure 7-6. The three arcs of great circles that form a spherical triangle are called the "sides" of the spherical triangle, and the angles formed by the arcs at the points where they meet are called the "angles" of the spherical triangle. The angles of a spherical triangle are the inclinations of the plane faces that form the solid angle.

Figure 7-6: Great circles and a spherical triangle.



On the plane, the sum of the interior angles of any triangle is exactly π radians (i.e., 180^o). However, on the surface of a solid sphere, the corresponding sum varies, but it is always greater than π radians. If the angles at each vertex of a spherical triangle are α, β, and γ, then the positive quantity

$$E = \alpha + \beta + \gamma - \pi$$

is called the "spherical excess" of the triangle. If r is the radius of the solid sphere on which a spherical triangle resides, and if the angles are measured in radians, then the area of a spherical triangle is equal to r^2E, where E is the spherical excess, defined above; and, in degrees, the formula for the area of a spherical triangle is $\pi r^2 E/180^o$.

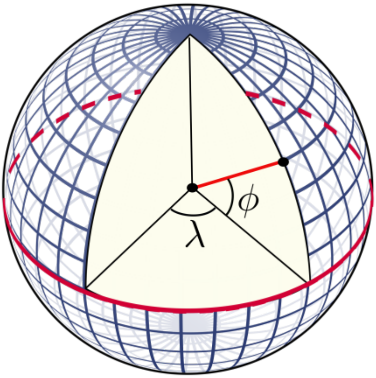
The radius is the distance from the center of the solid sphere to any point on its surface. Thus, given a solid sphere with center (x_0, y_0, z_0) and radius r, the distance from its center to an arbitrary point (x, y, z) on its surface is $r = \sqrt{(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2}$. However, the shortest distance between two points on the surface of a solid sphere is the so-called "great-circle distance": the shortest distance between point $a = (a_1, a_2, a_3)$ and point $b = (b_1, b_2, b_3)$ on the surface of a solid sphere of radius $r > 0$ is part of the great circle lying in a plane that intersects the surface of the solid sphere and contains the points $a = (a_1, a_2, a_3)$ and $b = (b_1, b_2, b_3)$ as well as the center of the solid sphere. In particular, if $a = (a_1, a_2, a_3)$ and $b = (b_1, b_2, b_3)$ are points on a sphere of radius $r > 0$ centered at the origin of Euclidean 3-space, then the distance from a to b along the surface of the sphere is

$$d(a, b) = r \cdot arccos\left(\frac{a \cdot b}{r^2}\right) = r \cdot arccos\left(\frac{a_1b_1 + a_2b_2 + a_3b_3}{r^2}\right)$$

as can be easily seen by considering the plane through a, b, and the origin. If θ is the angle between the vectors a and b, then $a \cdot b = r^2 cos\theta$, and the short arc joining a and b has length $r\theta$.

On the surface of the Earth, lines of longitude, also called meridians (i.e., lines running North-South that measure angular distance from the Prime Meridian, i.e., they measure distance East-West), and lines of latitude, also called parallels (i.e., lines running East-West that measure distance from the Equator, i.e., they measure distance North-South), are used as reference points, as shown in Figure 7-7 (zero degrees latitude is the line designating the Equator; and zero degrees longitude is known as the Greenwich Prime Meridian). Meridians coincide with points of the same longitude, and parallels coincide with points of the same latitude. By the term "great circle," we mean the largest circle that circumnavigates the Earth and is centered at the center of the Earth. A great circle divides the Earth in half, and, thus, the Equator is a great circle, but no other latitudes. All lines of latitude, except for the Equator, are "small circles." All lines of longitude are "great circles." The shortest distance between any two points on the Earth's surface lies along a great circle.

Figure 7-7: Latitude and Longitude on the Globe (source: Wikimedia Commons: Author: Peter Mercator; https://commons.wikimedia.org/wiki/File:Latitude_and_longitude_graticule_on_a_sphere.svg).



First of all, we know that the circumference of a circle is given by the formula $C = 2\pi r$, and an arc length is a fraction of a circle; and such a fraction is equal to $\frac{\theta}{360^o}$. Hence, the formula for the computation of an arc length is

$l = \frac{\theta}{360^o} \times 2\pi r.$ (1)

When we have to find the distance between two points on the Earth's surface, we use formula (1) with the angle θ being the angular distance from the center of the Earth. The radius of the Earth is approximately $6{,}371km$. Therefore: the formula for finding the distance between two points with the same longitude is

$$d(x, y) = \frac{Angular\ distance}{360^o} \times 2\pi \times 6{,}371km$$

where the angular distance is the angle between the two points relative to the center of the Earth; and the formula for finding the distance along a parallel between two points with the same latitude is

$$d(x, y) = \frac{Angular\ distance}{360^o} \times 2\pi \times 6{,}371km \times cos\theta$$

where θ is the latitude, and the angular distance is the angle between the two points relative to the center of the small circle of the parallel on which they are located.

The two most common non-Euclidean geometries are spherical geometry, also known as Riemannian geometry (named after the German mathematician Bernhard Riemann), and hyperbolic geometry, also known as Lobachevskian geometry (named after the Russian mathematician Nikolai Ivanovich Lobachevski). I have already clarified the following: Euclidean geometry exists on surfaces that have constant zero curvature, and, in Euclidean geometry, Euclid's parallel postulate holds (i.e., through any given point not on a line, passes exactly one line parallel to that line in the same plane), and the sum of angles of a triangle is always equal to π radians (180^o); whereas Riemannian geometry (that is, geometry on the sphere or on the ellipsoid) exists on surfaces that have constant positive curvature, and, in Riemannian geometry, there are no parallel lines (instead, there exist geodesics, which intersect each other), and the sum of angles of a triangle is always strictly greater than π radians (180^o). On the other hand, Lobachevskian geometry (i.e., hyperbolic geometry), which is based on hyperbolic functions, exists on surfaces that have constant negative curvature, and, in Lobachevskian geometry, there exist infinitely many lines that pass through a point P and are parallel to a given line, as indicated in Figure 7.8: in fact, there is a pair of lines through P parallel to a given line l that form an angle, and every line through P and in the interior of this angle is parallel to l. Moreover, in Lobachevskian geometry, the sum of angles of a triangle is always strictly less than π radians (180^o), as indicated in Figure 7.9 (whereas Riemannian geometry is characterized by "fat triangles," Lobachevskian geometry is characterized by "thin angles").

In geometry, we must have in advance not only the concept of space but also the very fundamental concepts for constructions in space; and, indeed, geometry gives them nominal definitions, and geometric axioms provide the means which are necessary in order to determine them. Our choice among the different geometries is based, on a case-by-case basis, on experimental facts and practical needs.

The geometry of the surface of a solid sphere is a non-Euclidean geometry focused on the coordinate representation of the sphere, namely, on the equation

$x^2 + y^2 + z^2 = k,$

and the transition from spherical geometry to hyperbolic geometry is based on a small but crucial modification. In particular, hyperbolic geometry is a non-Euclidean geometry focused on the following equation:

$x^2 + y^2 - z^2 = k,$

so that: for $k = 0$, this equation yields a cone ($x^2 + y^2 = z^2$); for $k = 1$, this equation yields a hyperboloid (i.e., what we get when we rotate a hyperbola around the z axis); and, for $k = -1$, this equation yields another hyperboloid (with two branches, one opening upward, and the other opening downward), and, in fact, this hyperboloid is considered to be the most important hyperbolic analogue of the sphere (it best captures Lobachevski's thought). Hence, according to the Italian mathematician Eugenio Beltrami (1835–1900), the hyperbolic plane is the surface $x^2 + y^2 - z^2 = -1$, and, in this geometry, the analogues of "straight lines" are obtained by taking a plane through the origin (as we did in the case of spherical geometry) and cutting the aforementioned surface with such a plane, thus obtaining hyperbolic lines.

Figure 7.8: Parallel lines in hyperbolic geometry.

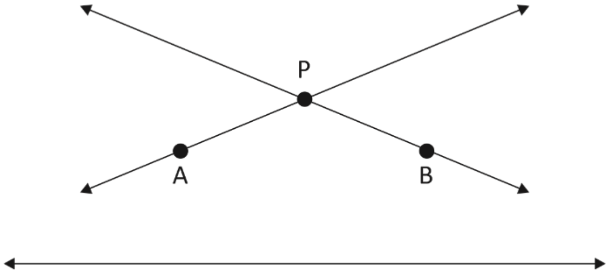
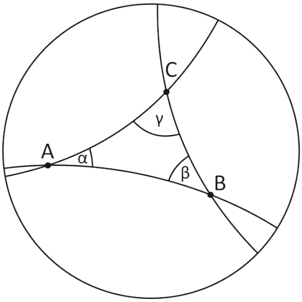


Figure 7.9: Triangles in hyperbolic geometry.



Metrics and Metric Spaces

In a vector space V_n over the field of real numbers, we can define the distance between points $x = (\alpha_1, \alpha_2, \dots, \alpha_n)$ and $y = (\beta_1, \beta_2, \dots, \beta_n)$ by

$$|x - y| = \left[\sum_{k=1}^{n} (\alpha_k - \beta_k)^2\right]^{\frac{1}{2}}$$

and, thus, obtain the n-dimensional Euclidean space $\mathbb{R}^n$.

A "norm" measures the size of a single thing (specifically, the length of a vector, as measured from the origin), but a "metric" (or "distance function") is a more general concept and mesures distances between pairs of things (specifically, the distance between two arbitrary points). A "metric," or "distance function," on an arbitrary set X is a real-valued function d defined on $X \times X$ that has the following properties for all x, y, and z:

(D1) $d(x, y) \geq 0; d(x, y) = 0 \Leftrightarrow x = y$;

(D2) $d(x, y) = d(y, x)$;

(D3) $d(x, y) \leq d(x, z) + d(z, y)$.

Properties (D1), (D2), and (D3) are known, respectively, as the "positive definite" property, the "symmetric property," and the "triangle inequality." In other words, a metric on X is a real-valued function that is positive definite and symmetric and satisfies the triangle inequality. If we allow $d(x, y) = 0$, then the metric is sometimes called "semi-metric" or "pseudometric." A set X endowed with a metric is called a "metric space." The systematic study of metric spaces (spaces with a metric) was initiated by the French mathematician Maurice Fréchet in the 1900s.

For instance, given two typical points $p = (p_1, p_2)$ and $q = (q_1, q_2)$ of $\mathbb{R}^2$, the Euclidean metric is given by

$$d_E(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}.$$

The Euclidean metric on $\mathbb{R}^n$ is defined by

$$d_E(p,q) = \left[\sum_{i=1}^{n} (p_i - q_i)^2\right]^{\frac{1}{2}}$$

(the only metric property that provides any difficulty to verify is the triangle inequality; we can show that $d_E(p,q)$ satisfies the triangle inequality by using the Cauchy–Schwarz–Bunyakovsky Inequality and the Minkowski Inequality).

It is possible to define more than one metric on the same set X, and, in general, different metrics define different metric spaces on X. Two metrics, somewhat different from d_E, on $\mathbb{R}^n$ are the following:

$d_m(p,q) = max\{|p_i - q_i|, i = 1,2,\dots,n\}$, and

$d_\Sigma(p,q) = \sum_{i=1}^{n}|p_i - q_i|$.

For any non-empty set X, the "discrete metric" is defined by

$d(x,y) = \begin{cases} 1 \; if \; x \neq y \\ 0 \; if \; x = y \end{cases}$'

which specifies that the distance from a point to itself is equal to 0, while the distance between any two distinct points is equal to 1. Notice that a metric space is called "discrete" if and only if each $x \in X$ is an "isolated point," meaning that there exists a neighborhood of x that does not contain any other points of X. It is clear that the discrete metric on any non-empty set defines a discrete metric space.

All norms are metrics, but normed vector spaces have a richer structure than general metric spaces. If you have a norm, then you can define a metric by saying that the distance (metric) between vectors $\vec{u}$ and $\vec{v}$ is the size of $\vec{u} - \vec{v}$, namely:

$$d(\vec{u},\vec{v}) = \|\vec{u} - \vec{v}\|$$

which is the metric induced by the corresponding norm.

Let (X,d) be a metric space. The "open (metric) ball" of radius $r > 0$ centered at a point $p \in X$ is usually denoted by $B_r(p)$, and it is defined by

$$B_r(p) = \{x \in X | d(x,p) < r\}$$

(i.e., a subset of points in X that satisfy $d(x,p) < r$). The "closed (metric) ball" is usually denoted by $B_r[p]$, and it is defined by

$$B_r[p] = \{x \in X | d(x,p) \leq r\}$$

(i.e., a subset of points in X that satisfy $d(x,p) \leq r$).

In a metric space (X,d), the "(metric) sphere" of radius $r > 0$ centered at a point $p \in X$ is usually denoted by $S_r(p)$, and it is defined by

$$S_r(p) = \{x \in X | d(x,p) = r\}$$

(i.e., a subset of points in X that satisfy $d(x,p) = r$).

For instance, in $\mathbb{R}^3$, a ball is a three-dimensional ("solid") figure bounded by a sphere, which is a two-dimensional figure (i.e., in $\mathbb{R}^3$, a two-dimensional sphere is the surface of a three-dimensional ball; and a three-dimensional ball is also called a "solid sphere"). The 0-sphere is the pair of points at the ends of a line segment, which can be construed as the 1-ball (i.e., an 1-ball is a line segment). The 1-sphere is a circle, that is, the circumference of a disc, which can be construed as the 2-ball (i.e., a 2-ball is a disc). The 2-sphere is the boundary of a 3-ball in $\mathbb{R}^3$.

In terms of metrics, in Figure 7-4 (a), (b), (c), we see the unit sphere in $(\mathbb{R}^2, d_E)$, $(\mathbb{R}^2, d_m)$, and $(\mathbb{R}^2, d_\Sigma)$, respectively.

A ball of n dimensions is called an n-ball, and it is bounded by an $(n-1)$-sphere (i.e., a sphere of $(n-1)$ dimensions is the boundary of a ball of n dimensions). Thus, an n-dimensional closed ball is determined by $n+1$ independent variables: the n coordinates of its center, and its radius. For instance, in the n-dimensional Euclidean space $(\mathbb{R}^n, d_E)$, a closed ball of center $(a_1, \dots, a_n)$ and radius r is analytically expressed as

$(x_1 - a_1)^2 + \cdots + (x_n - a_n)^2 \leq r^2$,

while the corresponding sphere is

$(x_1 - a_1)^2 + \cdots + (x_n - a_n)^2 = r^2$.

Notice that a geometry of three-dimensional closed balls may be regarded as a four-dimensional geometry, so that a three-dimensional closed ball may be regarded as a point of a four-dimensional space.

Let (X, d) be a metric space, and let $x \in X$. A subset A of X is said to be a "neighborhood" of x with respect to the metric d if and only if there exists an $\varepsilon > 0$ such that $B_\varepsilon(x) \subseteq A$, that is, if and only if A contains an open ball of radius ε centered at x. Hence, given a metric space (X, d), a subset A of X is said to be "open" in (X, d) if and only if, for every $p \in A$, there exists an $\varepsilon > 0$ such that $B_\varepsilon(p) \subseteq A$; that is, a set is "open" if and only if it is a neighborhood of each of its points. The name "open ball" is justified by the fact that it can be easily verified that an open ball is an open set according to the aforementioned definition. Moreover, by the definition of an open set, it can be easily verified that, in an arbitrary metric space (X, d), the union of any collection of open sets is open, and the intersection of any finite collection of open sets is open. An infinite intersection of open sets may result in a non-open set. For instance, $\bigcap_{n=1}^{\infty} \left(-\frac{1}{n}, \frac{1}{n}\right) = \{0\}$ is an infinite intersection of open sets that results in a non-open set, the singleton of zero.

Let A be a subset of a metric space (X, d). Consider the "complement" of A with respect to X, also known as the "set difference" of X and A, namely, $X - A = A^{\sim}$, consisting of the elements of X that do not belong to A. By De Morgan's laws, the complement of the union of two sets is the same as the intersection of their complements; and the complement of the intersection of two sets is the same as the union of their complements. Thus, we can define a “closed” subset of a metric space (X,d) as the complement of an open subset of (X,d). In other words, a subset A of a metric space (X,d) is said to be “closed” in (X,d) if and only if its complement, namely, $X-A=A^{\sim}$, is open in (X,d). By the definition of a closed set and the above properties of an open set, we can easily verify the following properties of a closed set (by applying De Morgan’s laws): in an arbitrary metric space (X,d), the union of any finite collection of closed sets is closed, and the intersection of any collection of closed sets is closed. An infinite union of closed sets may result in an open set. For instance, $\cup_{n=2}^{\infty}\left[\frac{1}{n},1-\frac{1}{n}\right]=(0,1)$.

Notice that, given an arbitrary metric space (X,d), the sets X and $\emptyset$ are considered to be both open and closed in (X,d).

We can formulate an alternative definition of a closed set in a metric space using the concept of an accumulation point (see also Chapter 2). Let (X,d) be a metric space, let $A\subseteq X$, and let $a\in A$. Then a is an “accumulation point” (or a “cluster point” or a “limit point”) of A if and only if, for every $\varepsilon>0$, $A\cap(B_{\varepsilon}(a)-\{a\})\neq\emptyset$; that is, if and only if, for every $\varepsilon>0$, there is at least one point of A, other than a itself, within distance ε of a. Hence, a subset A of a metric space (X,d) is “closed” with respect to the metric d if and only if every accumulation point of A is a member of A. Moreover, the closure of any subset A of an arbitrary metric space (X,d) is closed in (X,d) (the closure of A is the set consisting of all the points of A together with all the accumulation points of A).

In a discrete metric space, every subset is both open and closed. Recall that the discrete metric says that $d(x,x)=0$, and $d(x,y)=1$ whenever $x\neq y$. In a discrete metric space, consider an open ball of radius $0<r<1$, namely, $B_{0<r<1}(x)$. Then, due to the definition of the discrete metric, $B_{0<r<1}(x)$ contains only the point at which it is centered, that is, $B_{0<r<1}(x)=\{x\}$. Thus, in a discrete metric space, any point x in a set A has an open ball containing it (since we can always construct an open ball that only contains x), but, in this case, all sets are open, and, therefore, their complements are also open, while they are also closed as complements of open sets.

Given a metric space (X,d) and a non-empty subset A of X, the “diameter” of A is given by

$diam(A)=sup\{d(x,y)|x\in A,y\in A\}$,

and, therefore, a set A is “bounded” if and only if $diam(A)<\infty$. If $A\subseteq B$, then $diam(A)\leq diam(B)$. If A contains only one element, then

$diam(A) = 0$. If (X, d) is a metric space, then we can define the "bounded metric" d_b for X generated by d as follows:
$d_b(x, y) = \frac{d(x,y)}{1+d(x,y)}$ or $d_b(x, y) = min\{1, d(x, y)\}$.
Obviously, every non-empty subset of a bounded set is bounded. Moreover, it is easily checked that the union of two bounded sets is bounded.
For an arbitrary ball $B_r(a)$ of radius r, it holds that
$diam\big(B_r(a)\big) \leq 2r$.
A simple example of a metric space where the diameter of a ball is not equal to twice the radius is the following: Consider the discrete metric d on a set X, that is,
$d(x, y) = \begin{cases} 0, if\ x = y \\ 1, if\ x \neq y \end{cases}$'
and consider the ball of radius $r = \frac{1}{2}$ centered at x. Then $B_r(x) = \{x\}$, and, since, by definition, $diam(A) = sup\{d(a, b)|a, b \in A\}$ for any set A, the diameter of $B_r(x) = \{x\}$ is equal to zero.
Moreover, it is evident that, if a point x does not belong to an open ball $B_r(a)$, then $d\big(x, B_r(a)\big) \geq d(a, x) - r$.
In a discrete metric space, every set is bounded, since, in a discrete metric space, we have only two distances, namely, 0 and 1, and, therefore, if we take any two points x and y in a discrete metric space, then the distance $d(x, y)$ is always less than 2, symbolically, $d(x, y) < 2$.

Continuity and uniform continuity: Given metric spaces (X, d_1) and (Y, d_2), a mapping (function) $f: X \rightarrow Y$ is said to be "continuous" at a point $x_0 \in X$ if and only if, given $\varepsilon > 0$, there exists a $\delta > 0$ such that
$d_2\big(f(x), f(x_0)\big) < \varepsilon$ whenever $d_1(x, x_0) < \delta$ and $x_0 \in X$.
Geometrically, the aforementioned definition of the continuity of the mapping f at x_0 means that $f(x)$ belongs to the open ball $B_\varepsilon\big(f(x_0)\big)$ in the metric space (Y, d_2) when x belongs to the open ball $B_\delta(x_0)$ in the metric space (X, d_1). Equivalently, we can say that f is (d_1, d_2)-continuous at $x_0 \in X$ if and only if, whenever (x_k) is a sequence in X for which
$x_k \overset{d_1}{\rightarrow} x_0$ as $k \rightarrow \infty$,
then the sequence
$f(x_k) \overset{d_2}{\rightarrow} f(x_0\ )$ as $k \rightarrow \infty$.
Hence, a function f is continuous at a point x_0 if and only if the range of f over the neighborhood of x_0 shrinks to a single point $f(x_0)$ as the width of

the neighborhood around x_0 shrinks to zeo. Intuitively, “continuous” at a point means “joined” at that point, and the continuity of a function means that the function has a gapless graph. If f is continuous at every point of a subset A in X, then we say that f is continuous on A.

We say that a mapping (function) $f: X \to Y$ is “uniformly continuous” on X if, for every $\varepsilon > 0$, there exists a $\delta > 0$ such that $d_1(x, y) < \delta$ implies that $d_2\big(f(x), f(y)\big) < \varepsilon$. “Uniform continuity” is a stronger condition than “continuity,” because “continuity” is defined at a point x_0, whereas “uniform continuity” is defined on a set X: in case of “continuity,” the point x_0 is part of the definition’s data, and it is kept fixed, just as f itself, whereas “uniform continuity” requires the existence of a single $\delta > 0$ that works for the whose set X, and not only in a neighborhood of x_0. For a function to be continuous, we can check “one x at a time,” so that, for *each* x, we pick an ε and then we define a δ that depends on both x and ε so that $d_2\big(f(x), f(y)\big) < \varepsilon$ whenever $d_1(x, y) < \delta$; but, if we want uniform continuity, then we must choose an ε and then define a δ that is good *for all* the x values under consideration. Thus, uniform continuity implies continuity (since uniform continuity is a global property), but not all continuous functions are uniformly continuous (continuity is a local property).

Example 1: The function $f: \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$ is continuous but not uniformly continuous. Firstly, we can prove that $f(x) = x^2$ is continuous at $x \in \mathbb{R}$ as follows: Let $\varepsilon > 0$. Then a $\delta > 0$ must be found such that $|x - x_0| < \delta \Rightarrow |f(x) - f(x_0)| < \varepsilon$ for $x, x_0 \in \mathbb{R}$. In other words, by definition, $f(x)$ is continuous at $x = x_0$ if, for any real number $\varepsilon > 0$, we can find a real number $\delta > 0$ such that $x \in (x_0 - \delta, x_0 + \delta) \Rightarrow |f(x) - f(x_0)| < \varepsilon$. For $x \in (x_0 - \delta, x_0 + \delta)$, we have that

$$|x - x_0| < \delta \Rightarrow |x| < |x_0| + \delta,$$

and

$$|f(x) - f(x_0)| = |x^2 - x_0^2| = |(x + x_0)(x - x_0)| = |x + x_0||x - x_0| \leq (|x| + |x_0|)\delta \leq \delta(2|x_0| + \delta).$$

Hence, for any $\varepsilon > 0$, if we choose δ such that $\delta(2|x_0| + \delta) < \varepsilon$, then the condition of continuity of $f(x) = x^2$ at x_0 is satisfied. Now, we shall prove that we cannot establish the uniform continuity of $f(x) = x^2$ on $\mathbb{R}$, by *reducto ad absurdum*. For the sake of contradiction, suppose that $f(x) = x^2$ is uniformly continuous over $\mathbb{R}$. Then, by definition, for every $\varepsilon > 0$, there exists a $\delta > 0$ such that $|x - y| < \delta \Rightarrow |x^2 - y^2| < \varepsilon$. But, if, say, $\varepsilon = 1$, then, if such a δ existed and $y = x + \frac{\delta}{2}$, we would obtain

$$\left|x^2 - \left(x + \frac{\delta}{2}\right)^2\right| < 1 \text{ for all } x \in \mathbb{R},$$

which would mean that $\left|x\delta + \frac{\delta^2}{4}\right| < 1$ for any real number x, which is *not* true for sufficiently large values of x; *quod erat demonstrandum*.

Example 2: We can prove that the function $f(x) = \sqrt{x}$ defined on $[0, \infty)$ is uniformly continuous as follows: Let $\varepsilon > 0$, $|x - y| < \delta$, and $\delta = \varepsilon^2$. Then we have: $|f(x) - f(y)| = |\sqrt{x} - \sqrt{y}| = \sqrt{|\sqrt{x} - \sqrt{y}|^2} \leq \sqrt{|\sqrt{x} + \sqrt{y}||\sqrt{x} - \sqrt{y}|} = \sqrt{|x - y|} < \sqrt{\delta} = \varepsilon$, and, therefore, we have a δ that satisfies the definition of uniform continuity; *quod erat demonstrandum*.

Isometric embeddings, isometries, and embeddings with distortion: Given two arbitrary metric spaces (X, d_1) and (Y, d_2), a mapping $f: X \rightarrow Y$ is called an "isometric embedding" if and only if

$d_2(f(x), f(y)) = d_1(x, y)$ for all $x, y \in X$. (1)

In other words, if the distance between the transformed versions of two points ("image") is the same as the distance between the original two points ("pre-image"), then such a transformation is said to be an "isometric embedding." An isometric embedding is an injective mapping that preserves the distances between elements exactly, but it is not necessarily surjective. If an injective mapping preserves the distances between elements exactly, thus satisfying condition (1), and if it is surjective, then it is said to be an "isometry." Notice that, if $f: X \rightarrow Y$ is an isometry, then the inverse mapping $f^{-1}: Y \rightarrow X$ is an isometry of Y onto X. Therefore, an isometry is an isomorphism for metric spaces. In fact, if two metric spaces are isometric, then, as metric spaces (that is, as regards their metric structure), they are structurally identical.

A mapping $T: \mathbb{R}^n \rightarrow \mathbb{R}^n$ that maps every point $p \in \mathbb{R}^n$ to $p + a$ for a fixed $a \in \mathbb{R}^n$ is called a "translation." Moreover, notice that an orientation preserving linear mapping $T: \mathbb{R}^n \rightarrow \mathbb{R}^n$ that carries a set

$\{e_1, e_2, \dots, e_n\}$

of orthogonal unit vectors at 0 to another set

$\{e_1', e_2', \dots, e_n'\}$

of orthogonal unit vectors at 0 in such a way that

$T(e_i) = e_i'$,

where $i = 1,2, \dots, n$, is called a "rotation" (about 0). It is easily verified that translations and rotations are isometries.

Given two arbitrary metric spaces (X, d_1) and (Y, d_2), a mapping $f: X \rightarrow Y$ is an "embedding with distortion α" if there exists a constant $r > 0$ such that, for all $x, y \in X$,

$$r \cdot d_1(x,y) \leq d_2\big(f(x),f(y)\big) \leq \alpha r \cdot d_1(x,y). \qquad (2)$$

More precisely, the distortion of an embedding f is the infimum of all α such that f satisfies condition (2). The scaling by r, in condition (2), implies that we only care about approximately preserving the ratio between distances. Thus, isometric embeddings are embeddings with distortion equal to 1.

In general, given a metric space, consider the problem of finding a host metric space from within some class of "simpler" and "more convenient" metric spaces into which the original metric space can be embedded while preserving pairwise distances as much as possible. This is a key and fundamental problem in the theory of algorithms in general and in the algorithmic study of metric spaces in particular, since this process of simplification and approximation can provide the researcher with a new set of efficient algorithmic tools. In order to quantify the extent to which an embedding (generally, an injection between metric spaces) preserves distances (and, thus, the extent to which it is structurally faithful and informationally accurate), we consider the (multiplicative) distortion. In particular, if f is an embedding from the metric space (X,d_1) into another metric space (Y,d_2), then we define:

$$expansion(f) = sup_{x,y\in X}\frac{d_2\big(f(x),f(y)\big)}{d_1(x,y)}$$

$$contraction(f) = sup_{x,y\in X}\frac{d_1(x,y)}{d_2\big(f(x),f(y)\big)}$$

(where $x \neq y$), and then $distortion(f)$ is defined as the product of $expansion(f)$ and $contraction(f)$, symbolically:

$$distortion(f) = expansion(f) \times contraction(f)$$

(notice that the lowest distortion we can hope for is 1, in which case all distances are preserved exactly, and the embedding is called isometric). Low-distortion embeddings have been used in several computer science applications.

Connectedness: The intuitive meaning of a metric space X being connected is that it constitutes one piece, meaning that it cannot be represented as the union of two separated sets A and B. As I have already mentioned, two sets A and B are said to be disjoint if their intersection is the empty set. However, there is a stronger condition on A and B than disjointness, and this condition is known as "separation." By "separated sets" A and B, we mean that $Cls(A) \cap \mathrm{B} = \emptyset$ and $A \cap Cls(B) = \emptyset$, where Cls denotes "closure": each set is disjoint from the other's closure (obviously, any two separated sets are automatically disjoint). Hence, separated sets not only do not overlap but do not even touch each other.

For instance, let us consider the metric space $\mathbb{R}$ of all real numbers endowed with the usual metric $d(x,y)=|x-y|$ for all $x,y\in\mathbb{R}$. The sets $A=(-1,0)$, $B=\{0\}$, and $C=(0,1)$ are pairwise disjoint. But $Cls(A)\cap \mathrm{B}=[-1,0]\cap\{0\}=\{0\}\neq\emptyset$, and, therefore, the sets A and B are not separated. Similarly, it can be shown that the sets B and C are not separated. However, $Cls(A)\cap \mathrm{C}=[-1,0]\cap(0,1)=\emptyset$, and $A\cap Cls(C)=(-1,0)\cap[0,1]=\emptyset$, and, therefore, the sets A and C are separated in this metric space.

Notice that, given two non-empty sets $A,B\subseteq(X,d)$, where (X,d) is a metric space,

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

meaning the minimum distance between the elements in sets A and B. The condition of separation is not as strong as requiring that the distance between separated sets should be positive. For instance, the distance between the separated sets $[0,1)$ and $(1,2]$ is zero.

A set (or a metric space) is "connected" if and only if it is not possible to be represented as the union of two separated sets A and B. By a "domain," we mean a non-empty connected open set in a metric space; and a bounded domain together with all its boundary points is said to be a "region." If a set (or a metric space) is not connected, then it is said to be "disconnected." For instance, the hyperbola $H=\{(x,y)\in\mathbb{R}^2|x^2-y^2=1\}$ is disconnected, since the sets $H_1=\{(x,y)\in\ H|x>0\}$ and $H_2=\{(x,y)\in\ H|x<0\}$ form a disconnection of H.

The real line $\mathbb{R}$ is connected, as it can be easily shown by invoking the Dedekind Cut Axiom: Suppose that $\mathbb{R}=U\cup V$, where U and V are two non-empty sets such that $U\cap V=\emptyset$. Without loss of generality, let $u\in U$, $v\in V$, and $u<v$. Let $X=\{u_i\in U|u_i<v\}$ and $s=sup(X)$. Then s may or may not belong to U. If s does not belong to U, then $s\in Cls(U)$. If $s\in U$, then $s<v$, so that all points between s and v belong to V, and s is a limit point of U. Hence, either $Cls(U)\cap V\neq\emptyset$ or $U\cap Cls(V)\neq\emptyset$, *quod erat demonstrandum*.

Complete metric spaces: Let u_n be a sequence in which the difference between any two terms becomes arbitrarily small as the index of the term increases. As I mentioned in Chapter 2, such a sequence is called a "Cauchy sequence." In formal notation, a sequence u_n in a metric space (U,d) is a Cauchy sequence if and only if, for every $\varepsilon>0$, there exists an integer N such that $d(u_n,u_m)<\varepsilon$ for all $n,m\geq N$. Recall that, in the context of the real number system, every convergent sequence is a Cauchy sequence, and every Cauchy sequence converges. However, a Cauchy sequence may not converge in the field $\mathbb{Q}$. For instance the square root algorithm for the approximation of $\sqrt{8}$ gives the following sequence of rational numbers: $2, 2.8, 2.82, 2.828, 2.8284, \ldots$ This is a Cauchy sequence, but it does not converge in the field of rational numbers, since $\sqrt{8} \notin \mathbb{Q}$.

The definition of a Cauchy sequence is importat for the study of metric spaces because it is based on the concept of a metric (distance function). A metric space M is called "complete" if every Cauchy sequence of points in M has a well-defined limit that is also in M. In other words, a metric space (X, d) is said to be "complete" if every Cauchy sequence in (X, d) converges to a point of X. A normed vector space that is complete as a metric space is called a "Banach space" (named after the Polish mathematician Stefan Banach). Notice that $\mathbb{R}$ with the usual norm is complete (this completeness property derives from the fact that any subset A of $\mathbb{R}$ that is bounded from above has a supremum in $\mathbb{R}$).

If A is a closed subset of X, where (X, d) is a complete metric space, then (A, d) is also a complete metric space; because: if (x_n) is a Cauchy sequence in (A, d), then it is a Cauchy sequence in (X, d), so that it converges to some $\xi \in X$, and, since A is given to be closed, $\xi \in A$. Moreover, if $A \subseteq X$, and if (A, d) is a complete metric space, then A is a closed subset of X; because: if (x_n) is a sequence of elements of A that converges to $\xi \in X$, then we must show that $\xi \in A$, and, indeed, this is the case, since (x_n) is a Cauchy sequence (since it converges), and the fact that (A, d) is a complete metric space implies that (x_n) converges to a limit in A, that is, $\xi \in A$.■

The Cantor Intersection Theorem for Complete Metric Spaces: Let (X, d) be a complete metric space. Suppose that (x_n) is a sequence of points in X, and that (r_n) is a sequence of positive real numbers such that $r_n \to 0$ as $n \to \infty$, so that we obtain the closed balls $\ldots B[x_{n+1}, r_{n+1}] \subseteq B[x_n, r_n] \subseteq \cdots \subseteq B[x_1, r_1]$. Then the intersection of these closed balls is non-empty, and, more precisely, there exists a point ξ such that $\cap_{n=1}^{\infty} B[x_n, r_n] = \{\xi\}$.

Proof: Firstly, we shall show that the sequence (x_n) given in this theorem is a Cauchy sequence. Let $\varepsilon > 0$, and let $n' \in \mathbb{N}$ be chosen such that $r_{n'} < \varepsilon$. Then, if $m, n \geq n'$ with $m \geq n$, it holds that $B[x_m, r_m] \subseteq B[x_n, r_n]$, and, therefore, $d(x_m, x_n) \leq r_m < r_{n'} < \varepsilon$. Hence, indeed, (x_n) is a Cauchy sequence in (X, d). Since (X, d) is a complete metric space, (x_n), being a Cauchy sequence, converges to some $\xi \in X$. Because, whenever $n \geq m$, it holds that $(x_n) \subseteq B[x_m, r_m]$, for any $m \in \mathbb{N}$, it holds that $\xi \in \cap_{n=1}^{\infty} B[x_n, r_n]$. Now, we shall show that the aforementioned intersection

of closed balls contains only the point ξ. For the sake of contradiction, let $\xi, x \in \cap_{n=1}^{\infty} B[x_n, r_n]$. Then, for any $n \in \mathbb{N}$, it holds that
$d(\xi, x) \leq d(\xi, x_n) + d(x_n, x) < 2r_n$,
and, by taking the limit as $n \to \infty$, we realize that $d(\xi, x) = 0$ (since, by hypothesis, $r_n \to 0$ as $n \to \infty$). Therefore, $\xi = x$, so that $\cap_{n=1}^{\infty} B[x_n, r_n] = \{\xi\}$.■

Characterization of complete metric spaces: A necessary and sufficient condition that a metric space (X, d) is complete is that, for any decreasing sequence $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$ of non-empty closed sets with diameters approaching 0, the intersection $\cap_i A_i$ is non-empty.

Note: If $diam(A_i) \to 0$ as $i \to \infty$, then $\cap_i A_i$ is either empty or contains exactly one point due to Cantor's Intersection Theorem.

Proof (according to the method of G. Cantor): Suppose that (X, d) is a complete metric space. Let x_i be a point in A_i, $i = 1,2,3, \dots$ If we consider any $\varepsilon > 0$, then there is a λ large enough so that $diam(A_\lambda) < \varepsilon$, that is, $d(x_i, x_j) < \varepsilon$ for $i, j > \lambda$. Therefore, (x_i) is a Cauchy sequence. Because (X, d) is a complete metric space, (x_i) converges to some $x \in X$. Claim that $x \in A_i$. Indeed, if we consider a specific set A_j, it holds that $i \geq j \Rightarrow x_i \in A_j$; because A_j is closed, it follows that $x \in A_j$.

Conversely, given a Cauchy sequence (x_i) in (X, d), we must show that it converges to a point in X. Consider a set B_i containing the points $x_i, x_{i+1}, \dots$, so that $B_1 \supseteq B_2 \supseteq \cdots$ Additionally, $diam(B_i) \to 0$ because (x_i) is a Cauchy sequence. If $A_i = Cls(B_i)$, then $A_i \supseteq A_{i+1}$, and $diam(A_i) \to 0$, $i = 1,2,3, \dots$ By hypothesis, the A_i's are non-empty, and, due to Cantor's Intersection Theorem, there exists an element $\xi \in \cap_i A_i$. Hence, the sequence (x_i) converges to the point ξ; *quod erat demonstrandum.*

Corollary: $\mathbb{R}^n$ equipped with the Euclidean metric d_E is a complete metric space.

Proof: Firstly, we shall prove that $(\mathbb{R}, d_E)$ is a complete metric space.

Let (x_n) be a Cauchy sequence. Then, by definition, for any $\varepsilon > 0$, there exists a natural number n' such that $d(x_m, x_n) = |x_m - x_n| < \varepsilon$ for all $m, n \geq n'$.

For $\varepsilon = \frac{1}{2} > 0$, let n_0 be the smallest natural number such that $|x_m - x_n| < \frac{1}{2}$ for all $m, n \geq n_0$.

For $\varepsilon = \frac{1}{2^2} > 0$, let n_1 be the smallest natural number such that $|x_m - x_n| < \frac{1}{2^2}$ for all $m, n \geq n_1$.

Continuing in the same way for $\varepsilon = \frac{1}{2^{k+1}} > 0$, we realize that there exists a smallest natural number n_k such that $|x_m - x_n| < \frac{1}{2^{k+1}}$ for all $m, n \geq n_k$. Thus, $n_0 < n_1 < n_2 < \cdots < n_k < \cdots$ Then we come up with the sequence (x_{n_k}), which is a subsequence of (x_n). Now, we shall show that (x_{n_k}) is convergent. For this purpose, let us consider closed intervals $I_k = \left[x_{n_k} - \frac{1}{2^k}, x_{n_k} + \frac{1}{2^k}\right]$, and then $I_{k+1} = \left[x_{n_{k+1}} - \frac{1}{2^{k+1}}, x_{n_{k+1}} + \frac{1}{2^{k+1}}\right]$, for any k. Now, we shall show that the sequence (I_k) is decreasing, that is, the diameter of I_k tends to 0 as $k \to \infty$. Since (x_{n_k}) is a subsequence of a Cauchy sequence, namely, of (x_n), it follows that (x_{n_k}) is also a Cauchy sequence, and, therefore, for $\varepsilon = \frac{1}{2^{k+1}} > 0$, there must exist a natural number n'' such that $|x_{n_{k+1}} - x_{n_k}| < \frac{1}{2^{k+1}}$ for all $n_{k+1}, n_k \geq n''$, so that $-\frac{1}{2^{k+1}} < x_{n_{k+1}} - x_{n_k} < \frac{1}{2^{k+1}}$. Then, we observe the following:

$$-\frac{1}{2^{k+1}} < x_{n_{k+1}} - x_{n_k}, \tag{1}$$

and

$$x_{n_{k+1}} - x_{n_k} < \frac{1}{2^{k+1}}. \tag{2}$$

Inequality (1) implies that $x_{n_k} - \frac{1}{2^{k+1}} < x_{n_{k+1}} \Rightarrow x_{n_k} - \frac{1}{2^k} + \frac{1}{2^k} - \frac{1}{2^{k+1}} < x_{n_{k+1}} \Rightarrow x_{n_k} - \frac{1}{2^k} + \frac{2-1}{2^{k+1}} < x_{n_{k+1}} \Rightarrow x_{n_k} - \frac{1}{2^k} + \frac{1}{2^{k+1}} < x_{n_{k+1}} \Rightarrow x_{n_k} - \frac{1}{2^k} < x_{n_{k+1}} - \frac{1}{2^{k+1}} \Rightarrow x_{n_k} - \frac{1}{2^k} < x_{n_{k+1}} - \frac{1}{2^{k+1}} < x_{n_{k+1}} + \frac{1}{2^{k+1}}$. (3)

Inequality (2) implies that $x_{n_{k+1}} < x_{n_k} + \frac{1}{2^{k+1}} \Rightarrow x_{n_{k+1}} < x_{n_k} + \frac{1}{2^k} - \frac{1}{2^k} + \frac{1}{2^{k+1}} \Rightarrow x_{n_{k+1}} < x_{n_k} + \frac{1}{2^k} - \frac{2-1}{2^{k+1}} \Rightarrow x_{n_{k+1}} < x_{n_k} + \frac{1}{2^k} - \frac{1}{2^{k+1}} \Rightarrow x_{n_{k+1}} + \frac{1}{2^{k+1}} < x_{n_k} + \frac{1}{2^k}$. (4)

Combining inequalities (3) and (4), we obtain:
$x_{n_k} - \frac{1}{2^k} < x_{n_{k+1}} - \frac{1}{2^{k+1}} < x_{n_{k+1}} + \frac{1}{2^{k+1}} < x_{n_k} + \frac{1}{2^k} \Rightarrow \left[x_{n_{k+1}} - \frac{1}{2^{k+1}}, x_{n_{k+1}} + \frac{1}{2^{k+1}}\right] \subseteq \left[x_{n_k} - \frac{1}{2^k}, x_{n_k} + \frac{1}{2^k}\right] \Rightarrow I_{k+1} \subseteq I_k$ for all k. Hence, (I_k) is a decreasing sequence of closed intervals. Moreover, $diam(I_k) = sup_{x,y \in I_k} d(x,y) = sup_{x,y \in I_k}|x-y| = x_{n_k} + \frac{1}{2^k} - \left(x_{n_k} - \frac{1}{2^k}\right) = x_{n_k} + \frac{1}{2^k} - x_{n_k} + \frac{1}{2^k} \Rightarrow diam(I_k) = \frac{2}{2^k}$, and, thus, $diam(I_k) = \frac{1}{2^{k-1}}$, which tends to 0 as $k \to \infty$. By Cantor's Intersection Theorem, $\cap_k I_k$ will have exactly one point, say ξ, that is, $\cap_k I_k = \{\xi\}$. Then ξ must belong to I_k for all k, that is, $\xi \in \left[x_{n_k} - \frac{1}{2^k}, x_{n_k} + \frac{1}{2^k}\right]$ for all k, meaning that $|x_{n_k} - \xi| \leq \frac{1}{2^k}$ for all k. Hence, $|x_{n_k} - \xi| \to 0$ as $k \to \infty$, so that $x_{n_k} \to \xi$ as $k \to \infty$, that is,

(x_{n_k}) is a convergent subsequence of (x_n), and, therefore, (x_n) is also convergent, which proves the completeness of $(\mathbb{R}, d_E)$.

Now, we shall use the completeness of $(\mathbb{R}, d_E)$ in order to deduce the completeness of $(\mathbb{R}^n, d_E)$. Recall that $(\mathbb{R}^n, d_E)$ is the set of all real n-tuples $(x_1, x_2, \dots, x_n)$ endowed with the metric

$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$,

where $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_n)$ are elements of $\mathbb{R}^n$. In order to prove the completeness of $(\mathbb{R}^n, d_E)$, it suffices to prove that every Cauchy sequence in $(\mathbb{R}^n, d_E)$ is convergent. Let (x^n) be a Cauchy sequence in $\mathbb{R}^n$, whose terms are real n-tuples, namely:

the first term of (x^n) is $x^1 = (x_1^1, x_2^1, \dots, x_n^1)$;

the second term of (x^n) is $x^2 = (x_1^2, x_2^2, \dots, x_n^2)$;

$$\vdots$$

the pth term of (x^n) is $x^p = \left(x_1^p, x_2^p, \dots, x_n^p\right)$;

$$\vdots$$

the qth term of (x^n) is $x^q = \left(x_1^q, x_2^q, \dots, x_n^q\right)$;

$$\vdots$$

Since (x^n) is a Cauchy sequence, it holds that, given any $\varepsilon > 0$, there exists an $r \in \mathbb{N}$ such that, for all $p, q \geq r$,

$d(x^p, x^q) < \varepsilon \Rightarrow \sqrt{\left(x_1^p - x_1^q\right)^2 + \cdots + \left(x_n^p - x_n^q\right)^2} < \varepsilon \Rightarrow \left(x_1^p - x_1^q\right)^2 + \cdots + \left(x_n^p - x_n^q\right)^2 < \varepsilon^2$. Snce this is a summation of positive numbers, each term is less than ε^2, namely, for all $p, q \geq r$, and for each $i = 1,2, \dots, n$, it holds that

$\left(x_i^p - x_i^q\right)^2 < \varepsilon^2 \Rightarrow \left|x_i^p - x_i^q\right| < \varepsilon$,

by taking the square root. Hence, we have proved that, for each $i = 1,2, \dots, n$, $\left(x_i^p\right)$ is a Cauchy sequence in $\mathbb{R}$ (i.e., each component is a Cauchy sequence). Due to the completeness of $(\mathbb{R}, d_E)$, the sequence $\left(x_i^p\right)$ converges to some x_i for each $i = 1,2, \dots, n$. Set $x = (x_1, x_2, \dots, x_n)$, which obviously belongs to $\mathbb{R}^n$. Given any $\varepsilon > 0$, we have $\frac{\varepsilon}{\sqrt{n}} > 0$. Set $\frac{\varepsilon}{\sqrt{n}} = \varepsilon'$. Given this ε', and since $\left(x_i^p\right)$ converges to x_i for each $i = 1,2, \dots, n$, it holds that, for each $i = 1,2, \dots, n$, there exists an $r_i \in \mathbb{N}$ such that $\left|x_i^p - x_i\right| < \varepsilon'$ for all $p \geq r_i$. Let $r = max\{r_1, r_2, \dots, r_n\}$. Then

$\left|x_i^p - x_i\right| < \varepsilon' = \frac{\varepsilon}{\sqrt{n}}$ for all $p \geq r$ and for all i. Hence,

$d(x^p,x) = \sqrt{\left(x_1^p - x_1\right)^2 + \left(x_2^p - x_2\right)^2 + \cdots + \left(x_n^p - x_n\right)^2} <$
$\sqrt{\frac{\varepsilon^2}{n} + \frac{\varepsilon^2}{n} + \cdots + \frac{\varepsilon^2}{n}} = \sqrt{n\frac{\varepsilon^2}{n}} = \varepsilon$, for all $p \geq r$.
In other words, $d(x^p,x) < \varepsilon$ for all $p \geq r$, and this means that (x^n) converges to x, which, in turn, means that $(\mathbb{R}^n, d_E)$ is complete.■

Compactness: In real analysis, the property known as the "local compactness of $\mathbb{R}$" states that every bounded sequence has a convergent subsequence (this result was proved in Chapter 2). The local compactness of $\mathbb{R}$ highlights the significance of bounded closed intervals, since every sequence (x_n) in a bounded closed interval $[a,b]$ has a subsequence $\left(x_{n_k}\right)$ that is convergent to a point in $[a,b]$. Given a metric space (X,d), a subset A of X is said to be "compact" if and only if every sequence (x_n) in A has a subsequence $\left(x_{n_k}\right)$ that is convergent to a point of A.
For instance, given a metric space (X,d), any finite subset $A = \{x_k | k = 1,2,\ldots,n\}$ is a compact set: in any infinite sequence formed from members of A, at least one of the values must appear infinitely many times, because, if each value appears only finitely many times, then the sequence itself would be a finite number of values appearing finitely many times, that is, it would be finite, which contradicts the assumption that it is infinite; and, therefore, one of the values appears infinitely many times. Then the subsequence that is formed by the value that appears infinitely many times is obviously convergent to a point of A.
On the other hand, for instance, in $\mathbb{R}$ with the usual metric, the subset $(0,1]$ is not compact: the sequence $\left(\frac{1}{n}\right)$ converges to 0, which is not a point of $(0,1]$, and, thus, all subsequences of $\left(\frac{1}{n}\right)$ converge to 0, and no subsequence of $\left(\frac{1}{n}\right)$ converges to a point of $(0,1]$.
Consider a metric space (X,d) and a subset A of X. A collection of subsets $\{U_k\}$ of X is called a "cover" for A if $A \subseteq \cup_k U_k$. For instance, if U_1 is the set of all odd numbers $(1,3,5,\ldots)$, if U_2 is the set of all even numbers $(0,2,4,6,\ldots)$, and if $C = \{U_1, U_2\}$, then every element of the set $U = \{0,1,2,3,4,5,6\}$ belongs either to U_1 or to U_2, that is, $U \subset U_1 \cup U_2$, and, therefore, C is a cover of the set U. Moreover, another simple example is the following: the collection $C = \{(-n,n) | n \in \mathbb{N}\}$ is an open cover of $\mathbb{R}$, since $\mathbb{R} \subseteq \cup_{n \in \mathbb{N}} (-n,n)$.
Any subcollection of $\{U_k\}$ that is itself a cover for A is called a "subcover" for A. A cover is called "finite" if it contains only a finite number of sets.

A cover $\mathcal{C}$ of a set S is said to be an "open cover" of S if each member of $\mathcal{C}$ is an open set.

Let us consider the closed interval $U = \{x \in \mathbb{R} | 0 \leq x \leq 1\}$. If $\varepsilon > 0$ is fixed, then the collection $\mathcal{C} = \{(\alpha - \varepsilon, \alpha + \varepsilon) | \alpha \in U\}$ is an open cover of U. This open cover provides many subcovers. For instance, we may choose the family $\mathcal{C}' = \{(\beta - \varepsilon, \beta + \varepsilon) | \beta \in \{x \in \mathbb{Q} | 0 \leq x \leq 1\}\}$, which is an open cover of U, since every irrational number $x \in [0,1]$ can be approximated to within ε by some rational number, and $\mathcal{C}'$ is a subset of $\mathcal{C}$, meaning that $\mathcal{C}'$ is a subcover of U. Similarly, the family $\mathcal{C}'' = \{(\gamma - \varepsilon, \gamma + \varepsilon) | \gamma \in \{x \in \mathbb{Q}^{\sim} | 0 \leq x \leq 1\}\}$ is an open cover of U, and it consists of uncountably many sets.

Thus, given a metric space (X, d), a subset A is "compact" (or "ball cover compact") if every cover for A by open balls with centers in A has a finite subcover. It is easily seen that this definition of compactness, which is based on the notions of a cover and a finite subcover, is semantically equivalent to the definition of compactness that is based on the notions of a convergent sequence and a convergent subsequence (in the former case, we think in terms of collections of open balls, while, in the latter, we think in terms of sequences of points and the limit of a sequence). Notice that the definition of convergence and the fact that open balls are open sets in a metric space imply that every convergent sequence in a metric space has an open cover consisting of open balls centered at the limit point.

In general, if a subset A of a metric space (X, d) is compact, then it is: (i) bounded and (ii) closed.

Proof: (i) For the sake of contradiction, let A be an unbounded subset of (X, d). Then, given an $x_0 \in A$, it holds that, for each natural number n, there exists an $x_n \in A$ such that $d(x_n, x_0) > n$. Hence, the sequence (x_n) in A is unbounded, and, therefore, every subsequence of (x_n) is unbounded. This fact implies that no subsequence of (x_n) can be convergent, and, therefore, A is not compact, which contradicts the hypothesis.

(ii) For the sake of contradiction, suppose that A is not bounded. Then there exists a cluster point x_0 of A in $X - A$, and, thus, there exists a sequence (x_n) in A that converges to x_0, which is a point of $X - A$. Moreover, all subsequences of (x_n) converge to $x_0 \in X - A$, and, therefore, no subsequence of (x_n) converges to a point of A, meaning that A is not compact, which contradicts the hypothesis.■

However, a closed and bounded subset of a metric space (X, d) need not be compact. For instance, consider a metric space $(\mathbb{R}, d)$ where d is the discrete metric, and let $A = [0,1]$. Then the subset $A = [0,1]$ of $(\mathbb{R}, d)$ is closed and bounded but not compact: I have already mentioned that, in a discrete metric space, every set is both open and closed. Moreover, I have already mentioned that, in a discrete metric space, every set is bounded. Then A is a closed and bounded subset of $\mathbb{R}$ with the discrete metric d. Now, we shall prove that A is not compact. Let $K = \{\{x\} | x \in A\}$ be a collection of open sets in the discrete metric space $(\mathbb{R}, d)$, where by the definition of a discrete metric space, we are allowed to call the sets $\{x\}$ open. Obviously, $A = \cup_{x \in A} \{x\}$, meaning that K is an open cover of A. Thus, we have to prove that K has no finite subcover for A, in order to prove that A is not compact. For the sake of contradiction, suppose that K has a finite subcover for A, say $K' = \{\{x_1\}, \{x_2\}, \dots, \{x_n\}\}$. Then this means that K' can also cover A, that is, $A = \cup_{i=1}^{n} \{x_i\} = \{x_1\} \cup \{x_2\} \cup \dots \cup \{x_n\} = \{x_1, x_2, \dots, x_n\}$. But, by hypothesis, $A = [0,1]$, which has infinitely many elements, whereas $\{x_1, x_2, \dots, x_n\}$ is a finite set, and, therefore, $A \neq \cup_{i=1}^{n} \{x_i\}$. This contradiction proves that K has no finite subcover for A, and, therefore, A is not compact.

Nevertheless, in real analysis, compact sets have a very simple characterization, which is known as the Heine–Borel theorem (named after the German mathematician Eduard Heine and the French mathematician Émile Borel).

Heine–Borel Theorem: Every open cover of a closed and bounded set in $\mathbb{R}$ equipped with the Euclidean metric d_E admits a finite subcover, and, therefore, by the definition of a compact set, such a set is compact. Moreover, every compact set in $\mathbb{R}$ equipped with the Euclidean metric d_E is closed and bounded. Hence,

$$(closed\ and\ bounded) \Leftrightarrow compact$$

in $(\mathbb{R}, d_E)$.

Proof: Let S be a closed and bounded set, and let $\mathcal{C} = \{U_\alpha | \alpha \in \mathcal{A}\}$ be an open cover of S, so that $S \subseteq \cup_{\alpha \in \mathcal{A}} U_\alpha$. Moreover, because S is bounded (by hypothesis), there exist two real numbers a and b such that $S \subseteq [a, b]$. For the sake of contradiction, assume that S does not have a finite subcover. Let us bisect $[a, b]$ at c, so that we obtain two subintervals $[a, c]$ and $[c, b]$. Then at least one of these subintervals contains a subset of S that does not have a finite subcover, and we rename this subinterval as $[a_1, b_1]$. The length of $[a_1, b_1]$ is $b_1 - a_1 = \frac{b-a}{2}$. Subsequently, we bisect $[a_1, b_1]$ at point c_1, and we select that subinterval as $[a_2, b_2]$ which contains a subset of S that does not have a finite subcover. Repeating this process of bisection and selection, we obtain nested closed intervals $[a_n, b_n]$, where $n = 1,2, \dots$, such that:

i. the length of $[a_n, b_n]$, which is equal to $\frac{b-a}{2^n}$, tends to 0 as $n \to \infty$, and

ii. each $[a_n, b_n]$ contains a subset of S that does not have a finite subcover.

Hence, applying Cantor's Intersection Theorem, we obtain $[a_n, b_n] \subset (\varepsilon - \delta, \varepsilon + \delta)$ for $\delta > 0$, and $\cap_{n\in\mathbb{N}} [a_n, b_n] = \{\varepsilon\}$, so that ε is an accumulation point of the set S. Because S is a closed set (by hypothesis), $\varepsilon \in S$. Moreover, $\mathcal{C}$ is an open cover of S, so that, for some n, $\varepsilon \in U_n$, and, since U_n is an open set, $\varepsilon \in (\varepsilon - \delta, \varepsilon + \delta) \subset U_n$. Hence, condition (i) implies that $[a_n, b_n] \subset U_n$ for some n, so that $[a_n, b_n]$ is covered by a single member U_n of $\mathcal{C}$, which contradicts condition (ii). Therefore, S has a finite subcover.

Regarding the converse (i.e., the statement that every compact set is closed and bounded), I have already proved that, if a subset A of any metric space (X, d) is compact, then it is: (i) bounded and (ii) closed (in our proof, we used the concepts of a convergent sequence and a convergent subsequence).■

Remark: The above theorem is due to Émile Borel (1871–1956), who gave its formal statement in 1895. The reason for attaching Heine's name is that Eduard Heine (1821–81) used the underlying idea in 1872 in order to prove that a real function which is continuous on a finite closed interval is uniformly continuous. As we did in the proof of the completeness of $(\mathbb{R}^n, d_E)$, we can generalize the Heine–Borel theorem in $(\mathbb{R}^n, d_E)$. Therefore, we obtain the following characterization of compact sets: a set S in $(\mathbb{R}^n, d_E)$ is closed and bounded if and only if it is compact, that is, every open cover of S admits a finite subcover (for S).

Chapter 8
Infinitesimal Calculus: Functions, Limits, Continuity, the Topology of $\mathbb{R}^n$, Differentiation, and Integration

“Infinitesimal calculus” is a branch of mathematics that concerns itself with the systematic study of the concept of an “infinitely small function,” a function of a variable x whose absolute value, $|f(x)|$, becomes and remains smaller than any given number as a result of variation of x. The method of the “infinitesimals” (“infinitely small” quantities), whose origin can be traced back to ancient Greek mathematicians, underpins the analytic way of thinking. The analytic way of thinking is based on the awareness that, when we treat geometric figures and the motions of physical bodies as “wholes,” we cannot demonstrate significant apparent similarities between them, but, when we analyze them into (sufficiently) “small” pieces, they display great similarities to each other. Hence, the major problem of seventeenth-century mathematics consisted of determining the proper processes for dividing the “whole” into “small” parts, which would be more easily and more rigorously studied than the “whole,” as well as of determining the proper processes for reassembling the behavior of the “whole” from the behavior of its “small” parts. In particular, the “small” parts into which an object of scientific research is divided are similar to the “small” parts into which another object of scientific research is divided, and, thus, we can formulate generalizations (scientific laws) as the dimensions of such “small” parts tend to zero (hence, we have to work with “infinitesimals”).

The ancient Greek mathematician and physicist Archimedes can be considered to be the most important ancient pioneer of infinitesimal calculus. Some other great pioneers of infinitesimal calculus are the Flemish Jesuit and mathematician Gregory of Saint Vincent (1584–1667), the Dutch-French philosopher and mathematician René Descartes (1596–1650), the Italian mathematician and Jesuate Bonaventura Francesco Cavalieri (1598–1647), the French lawyer and amateur mathematician Pierre de Fermat (1607–65), the English clergyman and mathematician John Wallis (1616–1703), the English Christian theologian and mathematician Isaac Barrow (1630–77), and the Scottish mathematician and astronomer James Gregory (1638–75).

Infinitesimal calculus is primarily aimed at solving problems concerning “change.” Thus, infinitesimal calculus is used in many scientific disciplines, including physics, engineering, biology, economics, statistics, mathematical psychology, neuroscience, and strategic studies (including warfare problems and arms races). In the seventeenth century, infinitesimal calculus was erected as a rigorous framework of science as a result of the revolutionary achievements that took place in the scientific discipline of celestial mechanics, whose protagonists were Nicolaus Copernicus, Galileo Galilei, Tycho Brahe, Johannes Kepler, and Isaac Newton. In its contemporary rigorous form, infinitesimal calculus was formulated independently in England by Isaac Newton and in Germany by Gottfried Wilhelm Leibniz in the last quarter of the seventeenth century, using the algebraic set-up and, especially, the Cartesian set-up, which had been introduced and developed by their predecessors.

Functions

Whenever, by a known value of one quantity, we can find the value of another quantity, we say that there is a "functional dependence" between these quantities. For instance, if the length x of the side of a square is known, then its area can be found by the formula $A = x^2$. In this way, we specify the functional dependence between the length of the side of a square and its area.

As already explained, the specification of a "numerical function" requires a set of numbers X and a rule f, according to which every number x that belongs to the set X is associated with a certain number (the value of the function). An independent variable taking on values from the set X is said to be the "argument" of the function. Given a member a of the set X, the value of the function f for the argument a is denoted by $f(a)$.

If a function f is specified on a set X, then the set X is said to be the "domain" of this function, and the set of all the values of the function is said to be its "range." As already mentioned, a function $f: X \rightarrow Y$ assigns to each element $x \in X$ exactly one element $y \in Y$.

We can read the expression $y = f(x)$ as follows: "y is a function of x," meaning that, as the variable x varies, the variable y also varies according to some rule f; in this case, y is the dependent variable, and x is the independent variable.

Analytic representation of a function: Assume that we are given a collection of operations that must be performed with the argument x in order to obtain a function value. Then the function is said to be represented by an "analytic expression." For instance, consider the following functions: $y = x^2 + x + 1$, $x \in [0,1]$; $y = x^2 + x + 1$, $x \in [-2,3]$; and $y = x^2 + x + 1$, $x \in (-\infty, +\infty)$. Even though the analytic expressions of these functions are the same in form, we have three different functions, because they are defined on three different sets (their domains are different).

Graphical representation of a function: Assume that a function f is given by an analytic expression $f(x)$, that is, $y = f(x)$ with $x \in X$, where X is the corresponding real interval, on which f is defined. The "graph" of the function f is a set of points of the coordinate plane that have coordinates $\big(x, f(x)\big)$, where $x \in X$. If a function is even, then its graph is symmetric with respect to the axis of ordinates (i.e., its graph remains unchanged after reflection about the y-axis). If a function is odd, then its graph is symmetric about the origin.

A function $y = f(x)$ is defined to be "increasing" on its domain if, for any two of its points x_1 and x_2 such that $x_1 < x_2$, the inequality $f(x_1) \leq f(x_2)$ is satisfied; in other words, if to a greater value of the argument there corresponds a greater value of the function. If $f(x_1) < f(x_2)$ whenever $x_1 < x_2$, then the fuction $f(x)$ is called "strictly increasing." A function $y = f(x)$ is defined to be "decreasing" on its domain if, for any two of its points x_1 and x_2 such that $x_1 < x_2$, the inequality $f(x_1) \geq f(x_2)$ is satisfied; in other words, if a smaller value of the function corresponds to a greater value of the argument. If $f(x_1) > f(x_2)$ whenever $x_1 < x_2$, then the fuction $f(x)$ is called "strictly decreasing."

A function f is said to have a "period" T if, for any value of x for which f is defined, the following equalities hold:

$f(x - T) = f(x) = f(x + T)$.

The aforementioned definition implies that, if a function f with period T is defined at the point x, it is also defined at the points $x + T$ and $x - T$. If a function f has a non-zero period T, then it is said to be "periodic." For instance, if time is measured in years, then the distance from the Earth to the Sun is given by a periodic function whose period is equal to 1. In general, the period of a periodic function represents the interval of x values on which one copy of the repeated pattern occurs. For instance, the functions $sinx$ and $cosx$ have period 2π, and the functions $tanx$ and $cotx$ have period π. "Frequency" is defined to be the reciprocal of period, that is, $frequency = \frac{1}{period} = number\ of\ events\ per\ unit\ time$.

One of the simplest functions is the "linear function" (or "linear equation"), where $y = mx + c$. In this, y and x are "variables" (that is, they can take on many values), while m and c are "constants" (that is, they have fixed values). As already explained, if we plot y against x on a diagram, the result will be a straight line, hence the name. A "nonlinear function" ("nonlinear equation") is any other sort of function (equation).

For instance, $y = x^2$ is a quadratic equation that is downward-sloping for negative values of x and upward-sloping for positive values of x. Functions come in many forms, and they are very useful as models of the real world when they are simple or can be satisfactorily approximated by, or manipulated into simple forms.

The Limit of a Function

The concept of a limit, or a limiting process, is central to all mathematical analysis. In fact, one can argue that, from the perspective of mathematical analysis, "analysis" means taking limits. In his book entitled *Cours d'analyse*, the French mathematician Augustin-Louis Cauchy (1789–1857), one of the founders of modern mathematical analysis, explained the concept of a limit of a function in a clear, formal, and arithmetic, rather than geometric, way by arguing as follows: "when the successive values attributed to a variable approach indefinitely a fixed value so as to end by differing from it by as little as one wishes, this last is called the limit of all the others" (quoted in: Boyer, *The History of Calculus and Its Conceptual Development*, p. 272).
Consider an arbitrary function $f(x)$ defined at all values in an open interval of the number line $\mathbb{R}$ containing a point x_0, with the possible exception of x_0 itself, and let L be a real number. The "limit of a function" $f(x)$ at a point x_0 is L if and only if the values of x (where $x \neq x_0$) approach the number x_0 (notice that $f(x_0)$ may not be defined, since, according to the definition of a limit, x tends to x_0, but x never becomes equal to x_0). In other words, as x gets closer to x_0, $f(x)$ gets closer and stays close to L; symbolically:
$lim_{x \to x_0} f(x) = L$.
Remark: Let a be a real number and c a constant. Then
$lim_{x \to a} x = a$, and
$lim_{x \to a} c = c$.
Let us recall that the distance between any two points a and b on the number line $\mathbb{R}$ is $|a - b|$. Therefore, the statement

$$|f(x) - L| < \varepsilon$$

means that the distance between $f(x)$ and L is less than ε, and, by the definition of an absolute value, the statement

$$0 < |x - a| < \delta$$

is equivalent to the statement

$$a - \delta < x < a + \delta, \text{ so that } x \neq a.$$

Thus, the *Cauchy epsilon-delta definition of a limit* is the following: assume that, for all $x \neq a$, an arbitrary function $f(x)$ is defined over an open interval containing a. Then

$$lim_{x \to a} f(x) = L$$

if and only if, for every $\varepsilon > 0$, there exists a $\delta > 0$ such that, if $0 < |x - a| < \delta$, then $|f(x) - L| < \varepsilon$. The statement (with the universal quantifier) "for every $\varepsilon > 0$" means "for every positive distance ε from L"; the statement (with the existential quantifier) "there exists a $\delta > 0$" means that there is a positive distance δ from a; and the conditional statement "if $0 < |x - a| < \delta$, then $|f(x) - L| < \varepsilon$" means that, if x is closer than δ to a, and $x \neq a$, then the value of $f(x)$ is closer than ε to L. Hence, in the 1-dimensional Euclidean metric space, the Cauchy epsilon-delta definition of a limit means that, if, for each $\varepsilon > 0$, there exists a sufficiently small $\delta > 0$ such that, for all points x that belong to an open 1-dimensional ball centered at a and of radius δ, except possibly for a itself (i.e., this open 1-dimensional ball is a deleted neighborhood of a), the vale of $f(x)$ belongs to an open 1-dimensional ball centered at L and of radius ε, then we say that the limit of $f(x)$ as x tends to a is L (recall that an open ball in $\mathbb{R}$ is an open interval).

Limit laws: If $lim_{x \to a} f(x) = L_1$ and $lim_{x \to a} g(x) = L_2$, then:

$lim_{x \to a}\big(f(x) \pm g(x)\big) = lim_{x \to a} f(x) \pm lim_{x \to a} g(x) = L_1 \pm L_2$;

$lim_{x \to a}\big(f(x)g(x)\big) = lim_{x \to a} f(x) lim_{x \to a} g(x) = L_1 L_2$;

$lim_{x \to a} \frac{f(x)}{g(x)} = \frac{lim_{x \to a} f(x)}{lim_{x \to a} g(x)} = \frac{L_1}{L_2}$, provided that $L_2 \neq 0$.

Squeeze Theorem: Suppose that, for all $x \in [p, q]$ (except possibly at $x = a$), it holds that $g(x) \leq f(x) \leq h(x)$. Moreover, suppose that $lim_{x \to a} g(x) = lim_{x \to a} h(x) = L$ for some $p \leq a \leq q$. Then $lim_{x \to a} f(x) = L$.

Proof: The Squeeze Theorem follows from the definition of the limit of a function as follows: By the definition of limits,

$lim_{x \to a} g(x) = L$ means that

$\forall \varepsilon > 0, \exists \delta_1 > 0 | |x - a| < \delta_1 \Rightarrow |g(x) - L| < \varepsilon$.

Hence, $|x - a| < \delta_1 \Rightarrow -\varepsilon < g(x) - L < \varepsilon$. (1)

Similarly, $lim_{x \to a} h(x) = L$ means that

$\forall \varepsilon > 0, \exists \delta_2 > 0 | |x - a| < \delta_2 \Rightarrow |h(x) - L| < \varepsilon$.

Hence, $|x - a| < \delta_2 \Rightarrow -\varepsilon < h(x) - L < \varepsilon$. (2)

By hypothesis, $g(x) \leq f(x) \leq h(x)$, and, thus,

$g(x) - L \leq f(x) - L \leq h(x) - L$.

Choosing $\delta = min\{\delta_1, \delta_2\}$, and using inequalities (1) and (2), we obtain the following results: whenever $|x - a| < \delta$,

$-\varepsilon < g(x) - L \leq f(x) - L \leq h(x) - L < \varepsilon \Rightarrow -\varepsilon < f(x) - L < \varepsilon \Rightarrow lim_{x\to a} f(x) = L$.■

The small angle approximation: A very important limit is the following:
$lim_{x\to 0}\frac{sinx}{x} = 1$,
which can be restated as follows: $sinx \approx x$ for small x, meaning that, for small values of angle x, the sine of x is approximately equal to x. Following Leonhard Euler's *Foundations of Differential Calculus*, we can prove this limit geometrically by thinking as follows: Consider the unit circle, centered at $(0,0)$ of radius equal to 1. Let x be the length of an arc along the unit circle, from the point $(1,0)$ in a counter-clockwise direction to some point $(cosx, sinx)$ on the circle. Then, obviously, $sinx$ is the height of this point above the x-axis. Now, let us imagine what happens if $x \to 0$. Then the arc is just an infinitely short vertical line, and the height of the endpoint above the x-axis is just the length of the arc. Hence, when $x \to 0$, $sinx \approx x$, meaning that $lim_{x\to 0}\frac{sinx}{x} = 1$.

Corollary: $lim_{x\to 0}\frac{cosx-1}{x} = 0$. *Proof:* Using the above result, we work as follows:
$lim_{x\to 0}\frac{cosx-1}{x} = lim_{x\to 0}\frac{(cosx-1)(cosx+1)}{x(cosx+1)} = lim_{x\to 0}\frac{cos^2x-1}{x(cosx+1)}$.
Recall that $cos^2x + sin^2x = 1 \Rightarrow cos^2x - 1 = -sin^2x$.
Hence, the last equation becomes
$lim_{x\to 0}\frac{cosx-1}{x} = lim_{x\to 0}\frac{-sin^2x}{x(cosx+1)} = lim_{x\to 0}\left[\left(\frac{sinx}{x}\right)\left(\frac{-sinx}{cosx+1}\right)\right] =$
$lim_{x\to 0}\frac{sinx}{x} lim_{x\to 0}\frac{-sinx}{cosx+1} = (1)(0) = 0$.

We can adapt the above definition of a limit to define a limit of a function in n-variables, that is, in the n-dimensional Euclidean metric space, as follows: if, for each $\varepsilon > 0$, there exists a sufficiently small $\delta > 0$ such that, for all points $(x_1, \dots, x_n)$ that belong to an open n-dimensional ball centered at $(a_1, \dots, a_n)$ and of radius δ, except possibly for $(a_1, \dots, a_n)$ itself (i.e., this open n-dimensional ball is a deleted neighborhood of $(a_1, \dots, a_n)$), the vale of $f(x_1, \dots, x_n)$ is less than ε away from L, then we say that the limit of $f(x_1, \dots, x_n)$ as $(x_1, \dots, x_n)$ approaches $(a_1, \dots, a_n)$ is L; symbolically:
$lim_{(x_1,\dots,x_n)\to(a_1,\dots,a_n)} f(x_1, \dots, x_n) = L$.
For instance, in $\mathbb{R}^2$, the limit of $f(x_1, x_2)$ as (x_1, x_2) approaches (a_1, a_2) is L, written $lim_{(x_1,x_2)\to(a_1,a_2)} f(x_1, x_2) = L$, if and only if, for each $\varepsilon > 0$, there exists a sufficiently small $\delta > 0$ such that, for all points (x_1, x_2) in an open 2-dimensional ball (i.e., in an open disc) centered at (a_1, a_2) and

of radius δ, except possibly for (a_1, a_2) itself, the value of $f(x_1, x_2)$ is less than ε away from L, that is, $|f(x_1, x_2) - L| < \varepsilon$ whenever $0 < \sqrt{(x_1 - a_1)^2 + (x_2 - a_2)^2} < \delta$.
The same limit laws hold for functions in n-variables.

Continuity, Topological Structures, and Homeomorphisms

In Chapter 7, I defined "continuity" and "uniform continuity" using the concept of distance (i.e., in the context of metric spaces). In this section, I shall revisit the concept of continuity in order to study some more details regarding the definition of this concept and its difference from the definition of a limit, as well as in order introduce the concept of a topological structure, which enables us to define continuity without depending on a metric. Moreover, I shall explain the meaning of a homeomorphism, which is an isomorphism in the category of topological spaces.

Consider a function f whose domain is D_f. Let a be an interior point of D_f. Then f is said to be "continuous at the point" a if and only if
$lim_{x \to a} f(x)$ exists finitely and
$lim_{x \to a} f(x) = f(a)$,
meaning: if and only if the limit of $f(x)$ as x tends to a is equal to the value of $f(x)$ at a. If a is a boundary point of D_f (i.e., in this case, an endpoint of a closed interval), then we distinguish the following two cases:

i. if $D_f = (x_1, a]$, then $f(x)$ is said to be "continuous from the left" at a if $lim_{x \to a^-} f(x) = f(a)$;
ii. if $D_f = [a, x_2)$, then $f(x)$ is said to be "continuous from the right" at a if $lim_{x \to a^+} f(x) = f(a)$.

The aforementioned definition of continuity (known as the limit definition of continuity) can also be given in the following equivalent forms:

(i) A function f is continuous at $a \in D_f$ if and only if, for every sequence (x_n) with $lim_{n \to \infty} x_n = a$, where $x_n \in D_f$, it holds that $lim_{n \to \infty} f(x_n) = f(a)$. As I explained in Chapter 2, an infinite sequence (x_n) of real numbers $x_1, x_2, \dots, x_n$ has a limit a if and only if the distance $|x_n - a|$ tends to zero as the indices of the terms of this sequence become greater than some value n_0. This means that, after a finite set of n_0 terms of this sequence, the remaining infinitely many terms of the given sequence, namely, $x_{n_0+1}, x_{n_0+2}, x_{n_0+3}, \dots$, converge indefinitely to the value a. The sequential definition of continuity was originally developed by the German mathematician Eduard Heine (1821–81).

(ii) A function f is continuous at $x = a \in D_f$ if and only if: $\forall \varepsilon > 0, \exists \delta > 0 | |x - a| < \delta \Rightarrow |f(x) - f(a)| < \varepsilon$.

A function f is said to be "continuous over (or on, or in) an open interval" (x_1, x_2) if f is continuous at every point in that interval (x_1 may be $-\infty$, and/or x_2 may be $+\infty$). A function f is said to be "continuous over (or on, or in) the closed interval" $[x_1, x_2]$ if the following conditions hold: firstly, f is continuous at every x in the open interval (x_1, x_2); secondly, $f(x_1)$ and $f(x_2)$ both exist; and, thirdly, $lim_{x \to x_1^+} f(x) = f(x_1)$, and $lim_{x \to x_2^-} f(x) = f(x_2)$.

If we compare the definition of the limit of a function with the definition of the continuity of a function, we realize that they have the same structure, but they also have the following differences:

i. In the case of the limit of a function (Cauchy epsilon-delta definition), we have $0 < |x - a| < \delta$, or $x \neq a$, whereas, in the case of continuity, we have only $|x - a| < \delta$, meaning that the definition of continuity holds also when $x = a$.

ii. Instead of the value L that is used in the definition of the limit of a function, the definition of the continuity of a function uses the value $f(a)$, meaning that, in the case of the continuity of a function, the function must be defined at the point a. Indeed, it is meaningless to talk about the continuity (or the discontinuity) of a function at a point that does not belong to its domain.

iii. In the definition of the limit of a function (Cauchy epsilon-delta definition), the point a must be an accumulation point of the domain D_f of the corresponding function. Therefore, it may not belong to D_f. However, in the definition of the continuity of a function, the point a must belong to the domain D_f of the corresponding function.

For instance, notice that every polynomial function is continuous everywhere, since: constant functions are continuous, x (the identity mapping) is continuous, multiplication is continuous, addition is continuous, and composition of continuous functions is continuous. Polynomials are precisely functions obtained by repeatedly composing addition, multiplication, constants, and x.

Let $A \subseteq \mathbb{R}^{\mathrm{n}}$. Then a function $f: A \to \mathbb{R}$ is said to be "continuous" at a point $P_0 \in A$ if and only if: for every $\varepsilon > 0$, there exists a $\delta > 0$ such that, for every $P \in A$ with $\|PP_0\| < \delta$, it holds that $\|f(P) - f(P_0)\| < \varepsilon$. Equivalently, we can say that a function $f: A \to \mathbb{R}$, where $A \subseteq \mathbb{R}^{\mathrm{n}}$, is

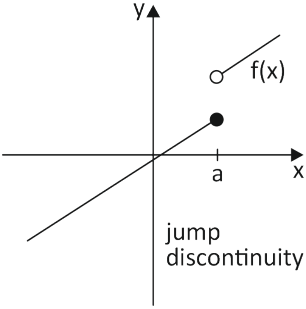
continuous at a point $P_0(x_1', x_2', \dots, x_n') \in A$ if and only if: for every $\varepsilon > 0$, there exist $\delta_i > 0$, $i = 1,2,\dots,n$, such that, for every point $P(x_1, x_2, \dots, x_n) \in A$ with $|x_i - x_i'| < \delta_i$, $i = 1,2,\dots,n$, it holds that $\|f(x_1, x_2, \dots, x_n) - f(x_1', x_2', \dots, x_n')\| < \varepsilon$.
A third equivalent definition of the continuity of a function in n real variables is the following: Let $A \subseteq \mathbb{R}^{\mathrm{n}}$, and let A' be the set of the accumulation points of A. Then a function $f\colon A \to \mathbb{R}$ is continuous at a point $P_0 \in A \cap A'$ if and only if $lim_{P \to P_0} f(P) = f(P_0)$.

Properties of continuous functions: If a function f is continuous at x_0, which belongs to the domain of f, and if $f(x_0\,) \neq 0$, then there exists a neighborhood of x_0 (specifically, there exists an open and bounded interval centered at x_0) wherein $f(x) \neq 0$. In other words, there exists a $\delta > 0$ such that $f(x) \neq 0$ for all $x \in N_\delta(x_0) \cap D_f$, where $N_\delta(x_0)$ denotes a δ-neighborhood of x_0, and D_f denotes the domain of f. In particular, if $f(x_0) > 0$ (resp. $f(x_0) < 0$), then $f(x) > 0$ (resp. $f(x) < 0$) for all $x \in N_\delta(x_0) \cap D_f$. *Proof:* Given that f is continuous at x_0, it holds that
$\forall \varepsilon > 0, \exists \delta > 0 \,||x - x_0| < \delta \Rightarrow |f(x) - f(x_0)| < \varepsilon$,
where x is any element of the domain of f. Since $f(x_0\,) \neq 0$, if we set $\varepsilon = \frac{1}{2}|f(x_0\,)|$, then we shall get
$|f(x) - f(x_0)| < \frac{1}{2}|f(x_0\,)|$, $\forall x \in D_f$ with $|x - x_0| < \delta$, so that
$f(x_0) - \frac{1}{2}|f(x_0\,)| < f(x) < f(x_0) + \frac{1}{2}|f(x_0\,)|$.
Hence, if $f(x_0) > 0$, then $f(x) > f(x_0) - \frac{1}{2}f(x_0\,) = \frac{1}{2}f(x_0\,) > 0$, and, if $f(x_0) < 0$, then $f(x) < f(x_0) - \frac{1}{2}f(x_0\,) = \frac{1}{2}f(x_0\,) < 0$, $\forall x \in D_f$ with $|x - x_0| < \delta$, that is, $\forall x \in N_\delta(x_0) \cap D_f$, *quod erat demonstrandum.*
Given two functions f and g that have the same domain, if they are continuous at x_0, which is an element of their common domain, then the functions
kf (for any constant k),
$f \pm g$,
$f \cdot g$,
$|f|$, and
$\frac{f}{g}$ (with $g \neq 0$)
are also continuous at x_0. These properties follow directly from the limit definition of continuity and the properties of limits. Moreover, given that

$min\{f,g\} = \frac{1}{2}(f + g - |f - g|)$ and $max\{f,g\} = \frac{1}{2}(f + g + |f - g|)$, the above properties of continuity imply that, if f and g are continuous at x_0, then $min\{f,g\}$ and $max\{f,g\}$ are also continuous at x_0.
If f and g are functions such that g is continuous at x_0 and f is continuous at $g(x_0)$, then the composition $f\big(g(x)\big)$ is continuous at x_0; since, given a convergent sequence x_n, $n \in \mathbb{N}$, with $x_n \to x_0$, the fact that $g(x)$ is continuous implies that $g(x_n) \to g(x_0)$ as $x_n \to x_0$, and the fact that $f(x)$ is continuous implies that $f\big(g(x_n)\big) \to f(g(x_0)\ )$ as $x_n \to x_0$, as required.

Discontinuities: In intuitive terms, a function is said to be continuous if it varies with no abrupt breaks or jumps. Hence, points of continuity are characterized by the fact that, for small changes in the argument, the value of the function changes but little, whereas points of discontinuity are characterized by the fact that, for small changes in the argument, the function can change considerably. For instance, consider a load that is suspended on a thread above a table. Due to this load (supposed to be a material particle), the thread extends, and the distance l from the load to the point of thread suspension is a function of the mass m of the load, symbolically, $l = f(m)$, where $m \geq 0$. For small changes in the mass of the load, the distance l will change but little. But, if the mass of the load approaches the tensile strength m_0 of the thread, then a small increase in the mass of the load may cause a break in the thread. Thus, the distance l will increase jump-wise and become equal to the distance L from the suspension point to the surface of the table. On the half-closed interval $[0, m_0)$, the graph of the function $l = f(m)$ is a continuous line, and, at the point m_0, it suffers a discontinuity. Consequently, we get a graph consisting of two branches: at all points except m_0, the function $l = f(m)$ is continuous, in the sense that it exhibits a smooth change. At the point m_0, however, it has a discontinuity, in the sense that it exhibits a jump-wise change. In Figure 8-1, we see an example of a “jump discontinuity.”

Figure 8-1: Jump discontinuity.



In case of a "jump discontinuity," the right-hand limit and the left-hand limit both exist, but they are not equal. In fact, the size of the jump is the difference between the right-hand limit and the left-hand limit. For instance, the piecewise function

$$f(x) = \begin{cases} 1\ if\ x < 0 \\ 2\ if x > 0 \end{cases}$$

has a jump discontinuity at $x = 0$, where the value of the function changes suddenly from 1 to 2.

In case of an "infinite discontinuity," the one-sided limits exist, and at least one of them is equal to $\pm\infty$. A common example of a function with an infinite discontinuity is the function $f(x) = \frac{1}{x}$, which has a vertical asymptote at $x = 0$. The function $f(x) = \frac{1}{x}$ is continuous on $(0, \infty)$ and on $(-\infty, 0)$, but it has a single point of discontinuity, namely, $x = 0$, and, in particular, it has an infinite discontinuity there.

Continuity on a Closed Interval: If a function $f: [a, b] \to \mathbb{R}$ is continuous on the closed interval $[a, b]$, then f is bounded in $[a, b]$. *Proof:* This theorem means that that, if $f: [a, b] \to \mathbb{R}$, is continuous on the closed interval $[a, b]$, then there exists an $M > 0$ such that, for all $x \in [a, b]$, it holds that $|f(x)| \leq M$. For the sake of contradiction, suppose that this does not hold, so that, for any $M > 0$, there exists some $x \in [a, b]$ such that $|f(x)| > M$. For $M = n$, in particular, let's assume that there exists a sequence $x_n \in [a, b]$ with $|f(x_n)| > n$. The sequence x_n, $n \in \mathbb{N}$, is bounded, since $a \leq x_n \leq b$ for all $n \in \mathbb{N}$. Therefore, by the Bolzano–Weierstrass Theorem (proved in Chapter 2), there exists a convergent subsequence, say x_{k_n}, $n \in \mathbb{N}$, with $lim_{n\to\infty} x_{k_n} = x_0$. Because $x_0 \in [a, b]$ and f is continuous at x_0, it must hold that $lim_{n\to\infty} f(x_{k_n}) = f(x_0) \in \mathbb{R}$.

Hence, $f(x_{k_n})$ is convergent and, therefore, bounded. But this result contradicts the assumed property that $|f(x_{k_n})| > k_n$. Consequently, f is bounded in $[a, b]$, *quod erat demonstrandum*.

Weierstrass's Extreme Value Theorem: If a function $f: [a, b] \to \mathbb{R}$ is continuous on the closed interval $[a, b]$, then f attains its supremum (least upper bound) and infimum (greatest lower bound) in $[a, b]$. *Proof:* According to the previous theorem, continuity of a function on a closed interval implies boundedness of the function. Therefore, $f: [a, b] \to \mathbb{R}$ is bounded in $[a, b]$, meaning that it has a supremum, say M, and an infimum, say m. We shall prove that there exist x_M and x_m in $[a, b]$ such that $f(x_M) = M$ and $f(x_m) = m$. If, for the sake of contradiction, we assume that there exists no $x_M \in [a, b]$ such that $f(x_M) = M$, then it should hold that $f(x) < M$ for all $x \in [a, b]$. Then $M - f(x)$ is positive and continuous on $[a, b]$. Moreover, the function

$g(x) = \frac{1}{M-f(x)}$, $x \in [a, b]$,

is continuous on $[a, b]$ and, therefore, bounded. Because it is positive, it has a positive supremum, say k, so that

$g(x) = \frac{1}{M-f(x)} \leq k$, $\forall x \in [a, b]$,

which implies that $f(x) \leq M - \frac{1}{k} < M$ for all $x \in [a, b]$. This means that $M - \frac{1}{k}$ is an upper bound of the range of f strictly smaller than M. But this is impossible, because $M = sup(f)$. Therefore, there exists an $x_M \in [a, b]$ such that $f(x_M) = M$. The proof for the infimum is similar; *quod erat demonstrandum*.

The Intermediate Value Theorem (due to Bolzano and Cauchy): Suppose that a function $f: [a, b] \to \mathbb{R}$ is continuous on the closed interval $[a, b]$ and $f(a) \neq f(b)$. If N is any value between $f(a)$ and $f(b)$, then there exists an $x_0 \in (a, b)$ such that $f(x_0) = N$. *Proof:* Without loss of generality, suppose that $f(a) < f(b)$ and $f(a) < N < f(b)$ (we can work analogously in case $f(a) > f(b)$). Let's consider the set $A = \{x \in [a, b] | f(x) \leq N\}$. Then $A \neq \emptyset$, since $a \in A$ and A is bounded from above by b. Hence, the supremum of A exists, and let $sup(A) = x_0$. We shall prove that $f(x_0) = N$. Indeed, if $f(x_0) > N$, then $x_0 > a$, and, since f is continuous at x_0, there exists some $\varepsilon > 0$ such that $f(x) > N$ over the interval $x_0 - \varepsilon < x \leq x_0$ (since I have already proved that, if a function f is continuous at x_0 and $f(x_0) \neq 0$, then there exists a neighborhood of x_0 wherein $f(x) \neq 0$, and, in particular, if $f(x_0) > 0$ (resp. $f(x_0) < 0$), then $f(x) > 0$ (resp. $f(x) < 0$) for all x in the intersection of this neighborhood of x and the domain of f). Therefore, $x_0 - \varepsilon$ is an upper bound for A, which contradicts the assumption that $sup(A) = x_0$. If $f(x_0) < N$, then $x_0 < b$, and, since f is continuous at x_0, there exists some $\varepsilon > 0$ such that $f(x) < N$ over the interval $x_0 \leq x < x_0 + \varepsilon$, meaning that there exist values of x that are greater than x_0 and belong to A for which it holds that $f(x) < N$ and, therefore, $x_0 \neq sup(A)$, thus contradicting our assumption that $sup(A) = x_0$. Consequently, $f(x_0) = N$, *quod erat demonstrandum*.

Corollary 1: If a function $f: [a, b] \to \mathbb{R}$ is continuous on the closed interval $[a, b]$ and $k \in \mathbb{R}$ such that
$inf\big(f([a, b])\big) \leq k \leq sup\big(f([a, b])\big)$,
then there exists an $x_0 \in [a, b]$ such that $f(x_0) = k$. In other words, every continuous function defined on a closed and bounded interval takes on all the values between its smallest value and its largest value in this interval. This corollary follows from the above Intermediate Value Theorem, given that, by Weierstrass's Extreme Value Theorem, there exist x_M and x_m in $[a, b]$ such that $inf\big(f([a, b])\big) = f(x_m) \leq k \leq f(x_M) = sup\big(f([a, b])\big)$.

Corollary 2: In case, $N = 0$, the above Intermediate Value Theorem reduces to the following corollary, known as Bolzano's theorem: If a function $f: [a, b] \to \mathbb{R}$ is continuous on the closed interval $[a, b]$ and $f(a) \cdot f(b) < 0$, then there exists some $x_0 \in (a, b)$ such that $f(x_0) = 0$.

Remark: The geometric significance of the above-mentioned Intermediate Value Theorem is the following: If the graph of a continuous function passes from one side of a horizontal line to the other, then it necessarily intersects that line somewhere. The geometric significance of the above-mentioned Bolzano's theorem (Corollary 2) is the following: If a continuous function on $[a, b]$ has values of opposite sign at the interval's endpoints, then it has at least one root in that interval.

Continuity and the Topology of $\mathbb{R}^n$*:* The epsilon-delta definition of a limit and the definition of continuity that is based on the epsilon-delta definition of a limit, as well as, generally, the study of limits and continuity in the context of metric spaces, depend on the concept of distance and assume that we have a clear rule for measuring distances. Moreover, these definitions are based on the concept of closeness. Thus, one may ask whether we can go up to such a high level of abstraction that we can rigorously define the continuity of a function in terms of closeness alone, without having to resort to distance measurement, that is, without having a metric. The answer to this question is positive and is one of the fundamental topics studied in topology.

Topology is a highly abstract kind of qualitative geometric knowledge, in the sense that it deals with the qualitative concept of nearness to spaces that might be conceptually close, without, however, using the quantitative concept of a distance function. Hence, intuitively, topology offers tools to model the concept of nearness in a set. In the context of topology, instead of using a ruler, we can think of two points x and y as being near each other if there are many open sets that contain both x and y, whereas, if there are no open sets containing two given points, then these two points are far apart (of course, the whole space is considered to be an open set containing every point under consideration). It is conventional to call the qualitative properties “topological properties.”

In order to understand what we mean by the qualitative properties of geometric figures, one can imagine a solid sphere to be a rubber ball that can be stretched and shrunk in any manner without being torn or gluing any two of its points together. Such transformations of a sphere are called homeomorphisms, and the different replicas that can be obtained as a result of homeomorphisms are said to be homeomorphic to each other. In other words, “homeomorphisms” are isomorphisms in the category of topological spaces. Hence, the qualitative properties of the sphere are those that it shares with all its homeomorphic replicas, that is, those which are preserved under homeomorphisms. For instance, one of the qualitative (“topological”) properties of the sphere is its integrity, namely, “connectedness.”

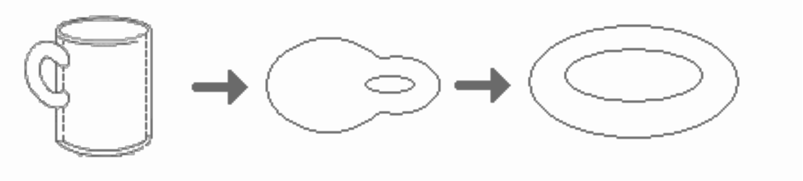
In few words, “topology” is the study of continuous shapes, and it is mainly preoccupied with properties that survive continuous deformation. In topology, we are allowed to deform objects, and, as long as we deform them *continuously*, we agree that they are topologically the same. From the topological point of view, it doesn’t matter if we bend, distort, or twist a geometric figure. To the topologist, homeomorphic spaces are indistinguishable, in the sense that they have the same topological properties (the term “homeomorphic” means being equal in the topological sense).

For instance, a topologist is not concerned with the differences between a circle and a square, since, from a topologist’s perspective, both a circle and a square are just simple closed curves (a curve is said to be “simple” if it does not cross itself, and a curve is said to be “closed” if it has no endpoints and, thus, forms a closed loop). A topologist is interested in those properties of a thing that, while they are in a sense geometric, are the most permanent, namely, the ones that will remain invariant after bending, distorting, or twisting a geometric figure. The roundness of a circle will not remain invariant, because we can tie or glue the ends of a bit of string together and make it into a circle, and, subsequently, without cutting or disconnecting it, we can make it into a square. But the facts that a circle has no endpoints and does not cross itself remain invariant (and, thus, every simple closed plane curve is homeomorphic to a circle). The straightness of a straight line is not a topological property (since, in topology, a straight line does not have to remain straight in the Euclidean sense, since it may be drawn on a globe and become a "geodesic"), but a straight line retains the quality of being continuously connected along itself, and it is this connectedness and this continuity that topology holds on to; and for this reason, in topology, deformations are only allowed if one does not disconnect what was connected, nor connect what was not. As I mentioned in Chapter 7, the concept of connectedness generalizes an intuitive concept of the wholeness or unseparatedness of a geometric figure, and the concept of a disconnected space generalizes the concept of the negation of wholeness, that is, separatedness.

According to the topological concept of a homeomorphism, we can take a doughnut-shaped (or, formally, a torus-shaped) lump of clay and make up a mug with a handle, and vice versa, without any tearing or gluing together, thus showing that a doughnut and a mug with a handle are topologically equivalent (the hole in the doughnut corresponds to the hole in the mug's handle), as shown in Figure 8-2. In view of Figure 8-2, the torus can be construed as a surface of revolution (revolving a small circle along a line made by a bigger circle) and, *equivalently*, as the solid sphere with one handle. By contrast, a round lump of clay without a hole (i.e., a solid sphere without a handle) and a mug with a handle are not topologically equivalent, because a round lump of clay cannot be transformed into such a mug (with a handle) without giving it a handle, and, since a round lump of clay (solid sphere) does not have a hole, it cannot be continuously deformed into a mug, which has a hole in the handle.

In topology, an object is said to be "simply connected" if and only if (like the lump of clay without a hole) it consists of one piece and does not have any "holes" that pass all the way through it. Notice that neither a doughnut (torus) nor a mug (with a handle) is simply connected, but a round lump of clay (solid sphere) is simply connected, in the sense that it can (continuously) contract to a point. In other words, connectedness can be defined as follows: a space is "connected" if and only if there are no two open sets that cover the entire space and have no points in common; whereas "simple connectedness" can be defined as follows: a space is said to be "simply connected" if and only if it is connected and every simple closed curve in the space can be continuously shrunk to a single point (i.e., in a simply connected domain D, every simple closed curve within it encloses only points of D).

Figure 8-2: A homeomorphism between a mug (with a handle) and a doughnut (source: Wikimedia Commons: Author: CHW; https://commons.wikimedia.org/wiki/Category:Homeomorphisms#/media/File:Homeo_tasse.png).



Some of the most important pioneers and founders of topology are the French mathematician, epistemologist, and theoretical physicist Henri Poincaré (1854–1912), the German mathematician Felix Hausdorff (1868–1942), and the Soviet mathematicians Pavel Sergeyevich Alexandrov (1896–1982) and Andrey Nikolayevich Tikhonov (1906–93).

Topology is the weakest structure (that is, the most "economical" structure in terms of assumptions) that can be established on a set and secure a good definition of continuity of mappings. By the term "topological space," we mean a set endowed with a topology defined on it. By the term "topology" (or "topological structure"), we mean a collection of subsets of the given set that are declared to be open. In fact, the intention of defining and using open sets in the context of topology is to give a meaning for "nearby," in the sense that two points are, in some sense, "nearby" if they are both in an open set. However, it does not suffice to declare a set open, since we want our open sets to have additional qualities, and we want to be able to perform set operations on them to preserve the given sets' qualities. In fact, in $\mathbb{R}^n$, the union of any collection of open sets is an open set, and the intersection of a finite collection of open sets is an open set. Thus, with these conditions and with the declarations that the empty set and the whole set are open sets, we come up with the "Euclidean topology" $\mathcal{T}_E$ of $\mathbb{R}^n$. In general, a topology endows a set with a structure based on the concept of a neighborhood, and, thus, a topology organizes a set into chunks of nearby points. The formal definition of a topology is the following: A "topology" $\mathcal{T}$ on a non-empty set X is a collection of subsets of X, called open sets, such that:

(T1) the empty set, $\emptyset$, and X are open, symbolically, $\emptyset, X \in \mathcal{T}$;

(T2) the union of any collection of open sets is open, symbolically, if $U_a \in \mathcal{T}$ for $a \in \mathcal{A}$, then $\cup_{a \in \mathcal{A}} U_a \in \mathcal{T}$;

(T3) the intersection of a finite collection of open sets is open, symbolically, if $U_i \in \mathcal{T}$ for $i = 1,2, \dots, n$, then $\cap_{i=1}^{n} U_i \in \mathcal{T}$.

Then the pair $(X, \mathcal{T})$ is called a "topological space." Whereas the concept of a metric space is based on the concept of a distance (or, more specifically, on the concept of a distance function), the concept of a topological space is based on the more abstract concept of closeness alone, or, more specifically, on the concept of a neighborhood. Notice that, if X is a topological space and U a subset, then U is said to be "open" in X if and only if, for each $p \in U$, U is a neighborhood of p; and a subset Y is said to be "closed" in X if and only if $X - Y$ is open.

For instance, given the set $X = \{1,2,3,4,5\}$, the family $F_1 = \{\emptyset, X, \{1\}, \{3,4\}, \{1,3,4\}, \{2,3,4\}\}$ is not a topology on X, because $\{1,3,4\}$ and $\{2,3,4\}$ belong to F_1, but $\{1,3,4\} \cup \{2,3,4\} = \{1,2,3,4\} \notin F_1$, whereas the family $F_2 = \{\emptyset, X, \{1\}, \{3,4\}, \{1,3,4\}, \{2,3,4,5\}\}$ is a topology on X (it satisfies conditions (T1), (T2), and (T3)).

In a metric space (X, d), open sets are defined in terms of the metric d as open balls as follows:

$(A \subseteq X, A\ open\ set) \Leftrightarrow (\forall x \in A, \exists r > 0 | B_r(x) \subseteq A)$,

where $B_r(x)$ is an open ball centered at x and of radius r in the metric space (X, d). Similarly, in a normed space $(E, \|\cdot\|)$, open sets are defined in terms of the norm $\|\cdot\|$ as open balls as follows:

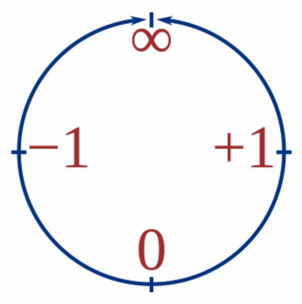
$(A \subseteq E, A\ open\ set) \Leftrightarrow (\forall x \in A, \exists r > 0 | B_r(x) \subseteq A)$,

where $B_r(x)$ is an open ball centered at x and of radius r in the normed space $(E, \|\cdot\|)$. The so defined open sets satisfy the conditions (T1), (T2), and (T3) mentioned in the above definition of a topology, and, therefore, they define a topology on the corresponding metric or normed space, respectively.

Notice that a metric space is a topological space with the properties required to define a metric (distance function), meaning that metric spaces have a richer structure than topological spaces. Hence, sameness of topology does not imply sameness of metric geometry.

The "usual" or "standard" topology on the real line is the topology whose open subsets are (unions of) open intervals, that is, they are sets of the form $\{x \in \mathbb{R} | a < x < b\}$, where $a, b \in \mathbb{R} \cup \{-\infty, +\infty\}$, and unions thereof (we have extended the order relation on $\mathbb{R}$ by declaring that $-\infty < +\infty$, $-\infty < x$, and $x < +\infty$, for any $x \in \mathbb{R}$). Notice that, in geometry, a "point at infinity," or an "ideal point," is an idealized limiting point that represents the "end" of each line, and, thus, a point at infinity completes a line into a topologically closed curve. The real line equipped with a point at infinity is called the "real projective line," extending from an original point 0 to an ideal point ∞, as shown in Figure 8-3.

Figure 8-3: The real projective line (source: Wikimedia Commons: Author: DerSpezialist; https://commons.wikimedia.org/wiki/File:Projective_Reals.svg).



Let f be an one-to-one mapping of the extended real line into $\mathbb{R}$ defined as follows:
$f(-\infty) = -1$,
$f(x) = \frac{x}{1+|x|}, x \in \mathbb{R}$,
$f(+\infty) = 1$.
Then the function

$$d(x,y) = |f(x) - f(y)| \; \forall x, y \in \mathbb{R} \cup \{-\infty, +\infty\}$$

is a metric on the extended real line, and the metric space of the extended real line is denoted by $\overline{\mathbb{R}}$. Notice that $\overline{\mathbb{R}}$ is isometric to the metric space that consists of the closed interval $[-1,1]$ with the Euclidean metric d_E (this metric space, which can be simply denoted by $[-1,1]$, is called a subspace of $\mathbb{R}$). Arguably, the n-dimensional sphere is the simplest non-Euclidean geometry. However, notice that Euclidean geometry is a local geometry on the sphere (in regions where the curvature of the sphere tends to zero), and the geometry on the sphere (Riemannian geometry) is a generalization of Euclidean geometry.

Given a metric space (X, d), the set of all open sets is a topology on X, and it is called the "metric topology" on X. The open sets of the Euclidean topology $\mathcal{T}_E$ on $\mathbb{R}^n$ are given by arbitrary unions of the open balls $B_r(p)$, defined as $B_r(p) = \{x \in \mathbb{R}^n | d_E(p,x) < r\}$, for all $r > 0$ and for all $p \in \mathbb{R}^n$, where d_E is the Euclidean metric. In fact, the circle S^1 is a topological space, in the sense that all the points that are on the circle lie in the set S^1, and, by analogy, the sphere S^2, which is embedded in $\mathbb{R}^3$ and inherits the topology $\mathcal{T}_E$ from the embedding topological space $(\mathbb{R}^3, \mathcal{T}_E)$, is a topological space, too (a 2-sphere is an ordinary 2-dimensional sphere in

3-dimensional Euclidean space, and it is the boundary of an ordinary 3-ball).

Notice that, given a non-empty set X, the collection $\{\emptyset, X\}$, consisting of the empty set and the whole set, is a topology on X, and it is known as the “trivial topology” on X. The power set $\wp(X)$ of X, consisting of all the subsets of X, is a topology on X, and it is called the “discrete topology” on X.

A topological space is called “Hausdorff” if and only if, for any two distinct points p and q, there exist neighborhoods U of p and V of q such that $U \cap V = \emptyset$ (i.e., distinct points are separated by disjoint neighborhoods). For instance, any Euclidean space is Hausdorff (the Euclidean topology is Hausdorff because, for any two distinct points in a Euclidean space, there exist disjoint open sets containing each point, and this property ensures that points can be separated).

Consider a topological space X and its subsets A and B. As I have already mentioned, A and B are separated if and only if $Cls(A) \cap \mathrm{B} = \emptyset$ and $A \cap Cls(B) = \emptyset$, where Cls denotes “closure.” A topological space X is “disconnected” if and only if it can be represented as the union of two non-empty separated sets, whereas a topological space not satisfying this condition is said to be “connected.” The simplest examples of connected topological spaces are a one-point space $X = \{*\}$ and an arbitrary set X equipped with the “trivial topology” on X. The simplest example of a disconnected topological space is a two-point set X equipped with the “discrete topology” on X.

If X and Y are topological spaces, then a mapping f from X to Y is said to be a “continuous mapping” if and only if $f^{-1}(A)$ is open in X (i.e., $f^{-1}(A)$ belongs to the topology of X) whenever A is open in Y (i.e., A belongs to the topology of Y). Notice that $f: X \to Y$ would be discontinuous if nearby points in the domain X were sent far apart in the codomain Y; and, reversing the direction of this statement, we require that all nearby points in X must be nearby in Y, thus securing the continuity of $f: X \to Y$.

Equivalently, we can say: If X and Y are topological spaces, then a mapping f from X to Y is said to be a “continuous mapping” if and only if, given $f(x) \in Y$ and a neighborhood $N_{f(x)}$ of $f(x)$, there exists a neighborhood N_x of x such that $f(N_x) \subseteq N_{f(x)}$. Therefore, the continuity of a mapping signifies the preservation of the nearness of points.

According to Pavel Sergeyevich Alexandrov, topology was born more in connection with clarifying the foundations of mathematical analysis, and it is in essence the most abstract theory of continuity. In topology, the concept of continuity is based on the existence of relations that are defined as local or neighborhood relations. Thus, according to Alexandrov, a topological space can be construed as a set in which certain subsets are defined and are associated to the points of the space as their neighborhoods.

Geometry is concerned with the study of such concepts as length, angle, area, and volume, whereas topology is concerned with the study of "closeness," or "connection," so that geometry will inform you about the length and the direction of a path between two points, but topology will tell you whether or not there is a path between two points. In other words, geometry is the branch of mathematics that you use in order to answer questions like "how far is it to get from point x to point y?" whereas topology is the branch of mathematics that you use in order to answer questions like "can I even get from point x to point y?" Thus, topology is frequently described as the study of shapes that can be stretched, squished, and otherwise distorted while keeping nearby points together (no tearing is allowed). Whereas geometry deals with specific kinds of spaces, topology deals with the most general kind of space possible. Topology transcends the particularities of different geometries, and it studies a common conception of "space," which amounts to considering one or more sets of objects (e.g., points, lines, etc.) endowed with a structure (namely, with a set of axioms describing the relations between these objects). Hence, a topological space can be intuitively construed as a geometric space in which "closeness" is defined in a rigorous way, but it cannot necessarily be measured in respect of a numeric distance.

If a topological space admits a metric, then it is called "metrizable." For instance, a metrizable topological space is $\mathbb{R}$ endowed with the discrete topology. The discrete topology is induced by the discrete metric. However, if $\mathbb{R}$ is endowed with the trivial topology $\mathcal{T} = \{X, \emptyset\}$, then this topological space is not metrizable (in this case, the only closed subsets of $\mathbb{R}$ are $\emptyset$ and the space $\mathbb{R}$, but we know that, in a metric space, singletons are closed sets; if $(\mathbb{R}, \mathcal{T})$ was metrizable, then its singletons should also be closed).

Given two topological spaces, $(X_1, \mathcal{T}_1)$ and $(X_2, \mathcal{T}_2)$, a function
$f: (X_1, \mathcal{T}_1) \to (X_2, \mathcal{T}_2)$
is said to be a "homeomorphism" if and only if f is a bijection (i.e., one-to-one and onto), f is continuous, and f^{-1} is continuous. If such a function exists, then the spaces X_1 and X_2 are said to be "homeomorphic" or "topologically equivalent" (they are actually, topologically speaking, the "same" space). Moreover, if the function $f^*: X_1 \to f(X_1)$, obtained by restricting the range of f, is a homeomorphism, then f is called an "embedding" of the space X_1 into X_2 (notice that $f(X_1)$ carries the subspace topology inherited from X_2).

Example 1: On the real (number) line with the usual (or standard) topology, the following open sets are homeomorphic (where $a, b, c, d \in \mathbb{R}$):

i. (a, b) and (c, d) : The function $f(x) = \frac{d-c}{b-a}(x-a) + c$, where $x \in (a, b)$, with $f^{-1}(y) = \frac{b-a}{d-c}(y-c) + a$, where $y \in (c, d)$, is a homeomorphism (notice that both f and f^{-1} are continuous, being linear functions).

ii. (a, b) and $\mathbb{R}$: The function $f(x) = tan\left[\frac{\pi}{b-a}\left(x - \frac{a+b}{2}\right)\right]$, where $x \in (a, b)$, with $f^{-1}(y) = \frac{b-a}{\pi} tan^{-1}y + \frac{a+b}{2}$, where $y \in \mathbb{R}$, is a homeomorphism (notice that both f and f^{-1} are continuous, because both $tanx$ and $tan^{-1}y$ are continuous).

iii. (a, b) and $(c, +\infty)$: The function $f(x) = \frac{1}{x-a} + c - \frac{1}{b-a}$, where $x \in (a, b)$, with $f^{-1}(y) = \frac{1}{y-c+\frac{1}{b-a}} + a$, where $y \in (c, +\infty)$, is a homeomorphism (notice that the continuity of both f and f^{-1} follows from the continuity of the function $\frac{1}{x}$).

iv. (a, b) and $(-\infty, c)$: The function $f(x) = -\frac{1}{x-a} + c + \frac{1}{b-a}$, where $x \in (a, b)$, with $f^{-1}(y) = \frac{1}{-y+c+\frac{1}{b-a}} + a$, where $y \in (-\infty, c)$, is a homeomorphism (notice that the continuity of both f and f^{-1} follows from the continuity of the function $\frac{1}{x}$).

Remark: The functions mentioned in the above four cases are not the only homeomorphisms between the corresponding sets, but they are simple in terms of operations, and this is the reason why I chose them in order to show that the corresponding sets are homeomorphic.

Example 2: The 2-sphere is locally homeomorphic to the Euclidean plane, in the sense that, for each $p \in S^2$, there is a neighborhood U of p such that U is homeomorphic to $\mathbb{R}^2$. In the XYZ-coordinate system (i.e., in $\mathbb{R}^3$), consider a unit 2-sphere with the origin as the center, namely, a subset of points of $\mathbb{R}^3$ that satisfy $|p| = \sqrt{p_1^2 + p_2^2 + p_3^2} = 1$. We denote the upper half of this sphere (i.e., $z > 0$) whose pole is (0,0,1) by S^+. It is easily noticed that each point $(q_1, q_2, q_3) \in S^+$ is projected to the point $(q_1, q_2, 0)$ of an open disc D of radius 1 in the XY-plane (see also Figure 6-16). The point $(q_1, q_2, 0)$ can be naturally identified with the point $(q_1, q_2) \in \mathbb{R}^2$, since D is a domain (an open disc) in $\mathbb{R}^2$ consisting of points (u, v) such that $u^2 + v^2 < 1$. Thus, we obtain a homeomorphism (and, in fact, an embedding)

$$f: D \to S^2 \subset \mathbb{R}^3$$

defined by

$$f(u, v) = \left(u, v, \sqrt{1 - u^2 - v^2}\right) \equiv (x, y, z)$$

(it is easily seen that (u, v) can be treated as coordinates of a point on the sphere). The pair $(f(D), f^{-1})$ constitutes a coordinate pair covering S^+. By analogy, we can construct five other coordinate pairs by taking $(0,0,-1)$, $(0,\pm 1,0)$, or $(\pm 1,0,0)$ as the poles. In fact, each hemisphere is mapped by a homeomorphism onto an open disc, and the coordinates of the points in the disc can be used in order to describe coordinates of points in the corresponding hemisphere. The 2-sphere is covered by a family of six coordinate neighborhoods each of which meets four other members of this family.

A topological space X each of whose points has a neighborhood homeomorphic to the open 2-disc (i.e., to the set of all points $(x, y) \in \mathbb{R}^2$ for which $x^2 + y^2 < r^2$ for some $r \in \mathbb{R}$) is a “two-dimensional manifold” (as we saw above, S^2 is a 2-dimensional submanifold of $\mathbb{R}^3$). In general, by a “topological n-dimensional manifold” M^n, we mean a connected Hausdorff space such that everyone of its points has a neighborhood homeomorphic to an open set in $\mathbb{R}^n$. If M^n is a topological n-dimensional manifold, then an indexed system $V = \{V_k\}$ of open sets is said to be a “covering” of M^n if each point of M^n belongs to at least one of these sets, and the union of these V_k’s equals M^n. Associated with each of these V_k’s are an open set U_k of $\mathbb{R}^n$ and a homeomorphism $\varphi_k: V_k \to U_k$.

Notice that, since a manifold is locally Euclidean while its global structure may be non-Euclidean, different geometries can be simultaneously valid, in the sense that they have different metrics, but they are logically isomorphic axiomatic systems (that is, their underlying manifolds are homeomorphic). In these cases, the choice of the appropriate geometry (e.g., plane geometry or spherical geometry) depends on our practical needs, the purpose of our work. For instance, locally, the Earth appears flat (ignoring hills, etc.), but long-distance observation leads us to the awareness that the Earth is roughly spherical.

Example 3: Any isometry is a homeomorphism. As mentioned in Chapter 7, given two arbitrary metric spaces (X, d_1) and (Y, d_2), an isometry is a bijective mapping $f: X \to Y$ such that $d_2\big(f(x), f(y)\big) = d_1(x, y)$ for all $x, y \in X$. In order to show that f is continuous, notice that, given $\varepsilon > 0$, if $d_1(x, y) < \varepsilon$, then $d_2\big(f(x), f(y)\big) = d_1(x, y) < \varepsilon$, meaning that f is (uniformly) continuous for $\delta = \varepsilon$. In order to show that f^{-1} is continuous,

simply notice that it is an isometry, and, therefore, by the first part, it is (uniformly) continuous as well.
Example 4: A torus is defined as the Cartesian product $S^1 \times S^1$, where S^1 is the 1-sphere, that is, a circle (roll a square so that two opposite edges meet to form a cylinder and then glue the cylinder's top and bottom edges to obtain a torus: this is a distorted square with all four vertices identified and with two pairs of opposite edges identified). A torus $T = S^1 \times S^1$ is not homeomorphic to S^2 (the 2-sphere), because a simple closed curve on S^2 can always be shrunk to a point, whereas this is not always the case on $S^1 \times S^1$ (some simple closed curves on $S^1 \times S^1$ can be shrunk to a point, but others cannot, since a torus has a hole). Therefore, the 2-sphere and the torus are not topologically the same, since simple closed curves on homeomorphic surfaces behave in the same way.
Another very important concept in topology is that of compactness. As I have already mentioned in Chapter 7, compactness can be intuitively construed as a sort of completed infinity: the concept of a compact space is a generalization of the concept of a closed and bounded subset of the real line. Let X be a set, and $A \subseteq X$. A collection $\mathcal{C}$ of subsets of X is called a"cover" for A if and only if
$A \subseteq \cup \{C | C \in \mathcal{C}\}$,
and, if this is the case, we say that $\mathcal{C}$ covers A. If a subcollection of $\mathcal{C}$ also covers A, then it is said to be a "subcover" of $\mathcal{C}$ for A. If X is a topological space, then an "open cover" is a cover each of whose members is open, and a "closed cover" is a cover each of whose members is closed. A topological space X is said to be "compact" if and only if every open cover for X contains a finite subcover (a subcover consisting of finitely many sets). Thus, a compact space has no "punctures" or "missing endpoints," so that it includes all its limit points. Since compactness is defined in terms of open sets, it is a topological property.
A compact Hausdorff space (i.e., a topological space that is both compact and Hausdorff) is a topological space in which every limit of a sequence that should exist does exist and does so uniquely. Hence, in topology, when we refer to a "compact space," we precisely mean a Hausdorff space with the property that every open cover contains a finite subcover.
If $f: X \to Y$ is a continuous mapping of a compact space X onto a Hausdorff space Y, then Y is compact: Given an open cover of Y, where the individual sets are denoted by U_i (with the i running over some set of indices), the sets $f^{-1}(U_i)$ form a cover of X that, due to the continuity of f , is open. Because X is compact, a finite collection, say $f^{-1}(U_1), \dots, f^{-1}(U_n)$, will cover it, and then $U_1, \dots, U_n$ form a finite

subcover of the given cover of Y, and, since we have assumed that Y is Hausdorff, it follows that Y is compact.

Curves and Surfaces in $\mathbb{R}^{n}$

Mathematically, dimensions are degrees of freedom. Consider a set of n real independent variables $x_1, x_2, \dots, x_n$. Such an n-tuple may be regarded as the coordinates of a current point in an n-dimensional space V_n, in the sense that each set of values of the variables defines a point of V_n. The totality of the points that correspond to values of the variables lying between certain specified limits constitutes a "region" of V_n.

The assemblage of points of V_n whose coordinates may be expressed as functions of a single parameter t is said to be a "curve" in the n-dimensional space V_n. Intuitively, a curve is an one-dimensional object, that is, an object that can be described by a single parameter. Hence, the equations

$x_i = x_i(t), i = 1,2, \dots, n,$

define a curve. However, points of V_n whose coordinates may be expressed as functions of two independent parameters u, v constitute a "surface" in the n-dimensional space V_n.

The totality of points whose coordinates may be expressed as functions of k independent parameters is said to be a k-dimensional "algebraic variety" or a k-dimensional "subspace" of V_n, and it may be denoted by V_k (i..e., an algebraic variety is the set of solutions of a system of polynomial equations over some field). Any such subspace is said to be "immersed" in V_n. If $k = n - 1$, then V_k is said to be a "hypersurface" of V_n.

Notice that, by a "constraint equation," we mean an equation of several variables that shows the relation between these variables. Usually, each constraint equation reduces the dimension by one, so that, usually, a set defined by m constraint equations in n variables is $(n - m)$-dimensional. Thus, an equation of the form

$$\varphi(x_1, x_2, \dots, x_n) = 0$$

is a constraint equation that determines a hypersurface in the n-dimensional space V_n, since such a relation reduces the number of independent variables to $n - 1$. Notice that, for instance, a line in $\mathbb{R}^2$ can be defined as the algebraic variety (the set of zeros) of the linear polynomial $x + y = 0$.

Moreover, if c is an arbitrary constant, then

$$\varphi(x_1, x_2, \dots, x_n) = c$$

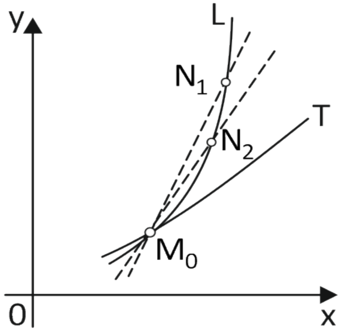
represents a family of hypersurfaces, each value of c determining a hypersurface. If the function φ is single-valued, then one hypersurface of the family passes through each point of V_n.
For instance, the constraint equation $x + 2y + 3z = 5$ (one constraint equation in three variables) determines a 2-dimensional set (specifically, a plane); and the constraint equations $x^2 + y^2 + z^2 = 100$ and $x + 2y + 3z = 5$ (two constraint equations in three variables) determine an 1-dimensional set (specifically, a circle, arising from the intersection of a sphere and a plane).

Differential Calculus in ℝ

Assume that a function $y = f(x)$ is defined at the points x and x_1. The difference $x_1 - x$ is called the "increment of the argument," and it is denoted by Δx (or sometimes simply by h). The difference $f(x_1) - f(x)$ is called the "increment of the function," and it is denoted by Δf or Δy. Therefore, $\Delta x = x_1 - x \Leftrightarrow x_1 = x + \Delta x$, and $\Delta f = f(x_1) - f(x) = f(x + \Delta x) - f(x)$. Using this formula, we can compute the value of Δf for any given x and Δx. Moreover, notice that a function $y = f(x)$ is continuous at a point $x = a$ if and only if $lim_{\Delta x \to 0} \Delta f = 0$, where $\Delta x = x - a$ and $\Delta f = f(x) - f(a)$. Indeed, the function $y = f(x)$ is continuous at the point $x = a$ if and only if $lim_{x \to a} f(x) = f(a)$ or, which is the same, if $lim_{x-a \to 0}\big(f(x) - f(a)\big) = 0$, that is, if $lim_{\Delta x \to 0} \Delta f = 0$.
Let $f(x)$ be a function defined on an interval $[a, b]$, and let $p \in (a, b)$. Assume that the limit
$lim_{x \to p} \frac{f(x)-f(p)}{x-p}$
exists. Then the function $f(x)$ is said to be "differentiable" at the point $p \in (a, b)$, and the limit $lim_{x \to p} \frac{f(x)-f(p)}{x-p}$ is called the "derivative" of f at p, and it is denoted by $f'(p)$, or $y'|_{x=p}$, or $\frac{df(p)}{dx}$. Symbolically:
$\frac{df(p)}{dx} \equiv f'(p) \equiv y'|_{x=p} = lim_{x \to p} \frac{f(x)-f(p)}{x-p}$.
The "right-hand derivative" of $f(x)$ at $x = p$ is defined as $f'_+(p) = lim_{x \to p^+} \frac{f(x)-f(p)}{x-p}$, provided that the limit exists. The "left-hand derivative" of $f(x)$ at $x = p$ is defined as $f'_-(p) = lim_{x \to p^-} \frac{f(x)-f(p)}{x-p}$, provided that the limit exists. Hence, $f'(p)$ exists if and only if $f'_+(p) = f'_-(p)$.
A function is said to be differentiable on a closed interval $[a, b]$ if it is differentiable at all points of (a, b) and has a right-hand derivativative at a and a left-hand derivative at b.

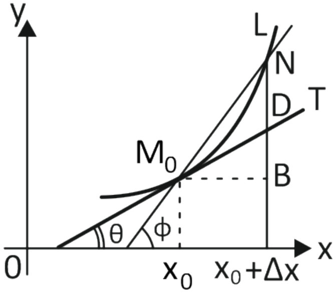
Let $\Delta x = dx$ be an increment given to x, and let the increment in $y = f(x)$ be $\Delta y = f(x + \Delta x) - f(x)$. If the function $f(x)$ is continuous and the derivative $f'(x)$ is also continuous on an interval, then the increment $\Delta y = f'(x)\Delta x + \varepsilon\Delta x = f'(x)dx + \varepsilon dx$, where $\varepsilon \to 0$ as $\Delta x \to 0$.
The first member of the right-hand side, that is, the term $f'(x)dx$, is called the "differential of y," and it is denoted by dy. Hence,
$dy = f'(x)dx$,
and dx is called the "differential of x." In general, $\Delta y \neq dy$, but, if $\Delta x = dx$, which is an infinitesimal, then the infinitesimal dy approximates Δy closely. Therefore, we write:
$\frac{dy}{dx} \equiv f'(x) = lim_{\Delta x \to 0} \frac{f(x+\Delta x)-f(x)}{\Delta x} = lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}$.
Notice that $\frac{dy}{dx} \equiv f'(x)$ is a new function defined at every such point x at which the indicated limit exists; this function is called the "derivative of the function $y = f(x)$," and it measures the rate of change of y with regard to x.
The geometric significance of the derivative of a function: Given a function $y = f(x)$, we realize that, in order to find the rate of change of y with regard to x at a particular point, we need to find the slope of the tangent line to the curve at that point. In differential calculus, a main objective is to try to understand tangents to curves, as shown in Figure 8-4. Hence, it is important to define a tangent line to an arbitrary plane curve in a rigorous way. A tangent line cannot be rigorously defined as a straight line having only one common point with the corresponding curve. In order to define a tangent line to an arbitrary plane curve in a rigorous way, we must use the concept of a limit. Let L be an arc of some curve, and M_0 be a point of this curve. We draw a secant M_0N through the point M_0. If the point N, moving in the curve, approaches the point M_0, then the secant M_0N turns about the point M_0. Thus, it may so happen that, as the point N approaches M_0, the secant tends to a certain limit position M_0T, so that M_0T is referred to as the "secant" to the curve L at the point M_0, as shown in Figure 8-4. Then the "tangent line" to the curve L at the point M_0 is defined as the limit position of the secant M_0N as $N \to M_0$.

Figure 8-4: A tangent line to a curve.



Let us try to compute the slope of the tangent line for the case when the curve L is the graph of a certain function $y = f(x)$. Let M_0 be a point of the graph with abscissa x_0 and ordinate $y_0 = f(x_0)$. Assuming that the tangent line to the curve L at the point M_0 does exist, we take one more point $N(x_0 + \Delta x, y_0 + \Delta y)$ on the curve, as shown in Figure 8-5, and we draw a straight line through the points M_0 and N. If φ is the slope of this secant to the positive direction of the x-axis, then
$|BN| = \Delta y$, $|M_0B| = \Delta x$, and $tan\varphi = \frac{|BN|}{|M_0B|} = \frac{\Delta y}{\Delta x}$,
so that $k_{tan} = lim_{N \to M_0} tan\varphi = lim_{\Delta x \to 0} tan\varphi$.
If we denote the slope of the tangent line to the axis of abscissas with θ, as shown in Figure 8-5, then the slope of the tangent line is
$k_{tan} = tan\theta = lim_{\Delta x \to 0} tan\varphi = lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}$.

Figure 8-5: The slope of a tangent line.



Consequently, in order to draw a non-vertical tangent line to the graph of the function $y = f(x)$ at a point with abscissa x_0, it is necessary and sufficient that, at this point, the limit $lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}$ exists (finitely); and this limit is equal to the slope of the tangent line. In other words, we create an infinite sequence of slopes, and then we say that the slope of the given tangent line is the infinite limit of this sequence. Hence, infinitesimal calculus provides us with abstract objects (such as a tangent to a curve) at which only infinite tasks can arrive through the concept of a limit. The concept of a limit has a deep philosophical significance, because it secures the theoretical convenience of being able to do an infinite number of tasks through a theoretical concept—namely, that of a limit—without actually doing each one of them, which would be practically impossible. This abstraction underpins the foundations of calculus as it was articulated by Newton and Leibniz in the seventeenth century. In view of the foregoing, the slope of the tangent line to the graph of a function $y = f(x)$ at the point x_0 is equal to the value of the derivative at the point of tangency; symbolically: $k_{tan} = f'(x)$. *This is the geometric significance of the derivative of a function.*

Theorem: If a function f is differentiable at x having a finite derivative, then f is continuous at x. However, the converse is not necessarily true.

Proof: Suppose that f is differentiable at $x = p$. Then the limit

$$lim_{x \to p} \frac{f(x) - f(p)}{x - p}$$

exists, it is finite, and, by definition, it is equal to $f'(p)$. Then notice that:

$$lim_{x \to p}\big(f(x) - f(p)\big) = lim_{x \to p}\left(\frac{f(x)-f(p)}{x-p} \cdot (x - p)\right) =$$

$$lim_{x \to p} \frac{f(x)-f(p)}{x-p} \cdot lim_{x \to p}(x - p) = f'(p) \cdot 0.$$

Hence,

$$lim_{x \to p}\big(f(x) - f(p)\big) = 0 \Leftrightarrow lim_{x \to p} f(x) = f(p),$$

and this proves that the function f is continuous at $x = p$. We have, thus, proved that, whenever a function has a finite derivative at a point, it is continuous there. In order to prove that the converse is not necessarily true, it suffices to give a counterexample. Indeed, consider, for instance, the function $f(x) = |x|$ for all $x \in \mathbb{R}$. Then, at $x = 0$, the function is continuous, because $lim_{x \to 0} f(x) = f(0)$, but the function is not differentiable at $x = 0$, because $f'_+(0) \neq f'_-(0)$; and, in fact, for $f(x) = |x|$, $f'_+(0) = lim_{h \to 0^+} \frac{f(h)-f(0)}{h} = lim_{h \to 0^+} \frac{|h|-0}{h} = lim_{h \to 0^+} \frac{|h|}{h}$, which is equal to $lim_{h \to 0^+} \frac{h}{h} = 1$, since, in this case, $h > 0$, whereas $f'_-(0) =$

$lim_{h\to 0^-}\frac{f(h)-f(0)}{h} = lim_{h\to 0^-}\frac{|h|}{h}$, which is equal to $lim_{h\to 0^-}\frac{-h}{h} = -1$, since, in this case, $h < 0$.■

Techniques and rules of differentiation: The formula for the derivative of x (for any $x \in \mathbb{R}$) is given by $\frac{dx}{dx} \equiv (x)' = 1$. Indeed, using the limit definition of the derivative, if $f(x) = x$, then we obtain:
$\frac{dx}{dx} = lim_{\Delta x\to 0}\frac{x+\Delta x-x}{\Delta x} = lim_{\Delta x\to 0}\frac{\Delta x}{\Delta x} = lim_{\Delta x\to 0}1 = 1.$
The intuition behind this result is that, given that the derivative of a function at a point represents the slope of the tangent drawn to the graph of that function at that particular point, and given that $f(x) = x$ represents a straight line, the derivative of x will be 1 at all points.
Obviously, $\frac{d}{dx}(c) = 0$ for any constant c (the slope, that is, the rate of change, of a constant function is zero; constant functions $f(x) = c$ are always horizontal lines parallel to the x-axis and cutting the y-axis at c).
If n is a positive integer, then $f(x) = x^n$ can be differentiated as follows: First of all, by definition, we shall have
$$f'(a) = lim_{x\to a}\frac{f(x)-f(a)}{x-a} = lim_{x\to a}\frac{x^n-a^n}{x-a}. \tag{1}$$
Moreover, it holds that
$$x^n - a^n = (x-a)(x^{n-1} + ax^{n-2} + a^2x^{n-3} + \cdots + a^{n-3}x^2 + a^{n-2}x + a^{n-1}), \tag{2}$$
and we notice that there are n terms in the second factor (we shall use this observation in the sequel). By substituting (2) into (1), we obtain:
$f'(a) = lim_{x\to a}\frac{(x-a)(x^{n-1}+ax^{n-2}+a^2x^{n-3}+\cdots+a^{n-3}x^2+a^{n-2}x+a^{n-1})}{x-a} =$
$lim_{x\to a}(x^{n-1} + ax^{n-2} + a^2x^{n-3} + \cdots + a^{n-3}x^2 + a^{n-2}x + a^{n-1}) =$
$a^{n-1} + aa^{n-2} + a^2a^{n-3} + \cdots + a^{n-3}a^2 + a^{n-2}a + a^{n-1} = na^{n-1}$. By replacing the a with an x, we obtain $(x^n)' = nx^{n-1}$, for any positive integer n. This result is known as the "power rule."
Let $X \subseteq \mathbb{R}$ be an interval, $a \in X$, and $f: X \to \mathbb{R}$ and $g: X \to \mathbb{R}$ be functions that are differentiable at a. Then, by the limit definition of the derivative, the following relations hold:
If $k \in \mathbb{R}$, then the function kf is differentiable at a, and
$$(kf)'(a) = kf'(a).$$
The function $f \pm g$ is differentiable at a, and
$$(f \pm g)'(a) = f'(a) \pm g'(a).$$
The function $f \cdot g$ is differentiable at a, and
$$(f \cdot g)'(a) = f'(a)g(a) + f(a)g'(a).$$
If $g(a) \neq 0$, then the function $\frac{f}{g}$ is differentiable at a, and

$\left(\frac{f}{g}\right)'(a) = \frac{f'(a)g(a)-f(a)g'(a)}{g(a)^2}$.

Differentiation of a composite function (the"chain rule"): $\left(f\left(g(x)\right)\right)' = f'\left(g(x)\right)\cdot g'(x)$. This result is known as the "chain rule."

In other words, if $y = y\left(u(x)\right)$, then $\frac{dy}{dx} = \frac{dy}{du}\cdot\frac{du}{dx}$.

For instance, to apply the chain rule to $f(x) = (x^2+1)^{10}$, the outside function is $h(\cdot) = (\cdot)^{10}$, and, by the "power rule," its derivative is $10(\cdot)^9$; while the inside function is $g(x) = x^2+1$, whose derivative is $2x$. Therefore, the chain rule implies that $f'(x) = 10(x^2+1)^9 2x$.

Implicit differentiation: If, in a function $f(x)$, one variable is not directly expressed in terms of the other variable, then this fuction is called "implicit." Implicit differentiation is illustrated in the following example: $x^4+y^3 = 7$. Since $y = f(x)$, $(x^4)' + (y^3)' = (7)' \Leftrightarrow 4x^3 + 3y^2\frac{dy}{dx} = 0 \Leftrightarrow 3y^2\frac{dy}{dx} = -4x^3 \Leftrightarrow \frac{dy}{dx} = -\frac{4x^3}{3y^2}$.

Higher order derivatives: It is evident that the first derivative $\frac{dy}{dx}$ expresses the rate of change of y with respect to x (e.g., velocity). Then $\frac{d}{dx}\left(\frac{dy}{dx}\right) \equiv \frac{d^2y}{dx^2} \equiv y''$ expresses the rate of change of the first derivative of y with respect to x (e.g., acceleration), and $\frac{d^3y}{dx^3} \equiv y''' \equiv y^{(3)}$ expresses the rate of change of the second derivative of y with respect to x (e.g., jerk). Of course, we can compute the nth derivative of $y = f(x)$, denoted by $\frac{d^ny}{dx^n} \equiv y^{(n)}$, where n is called the order of the derivative.

Basic differentiation formulae (following from the limit definition of the derivative):

i. $\frac{d}{dx}(a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0) = a_n\cdot nx^{n-1} + a_{n-1}\cdot(n-1)x^{n-2} + \cdots + a_1$, by the sum rule and the power rule.

ii. $\frac{d}{dx}(e^x) = e^x$; which can be proved as follows: If $f(x) = e^x$, so that $f(x+h) = e^{x+h}$, then the limit definition of the derivative implies that $f'(x) = lim_{h\to 0}\frac{e^{x+h}-e^x}{h} = lim_{h\to 0}\frac{e^x(e^h-1)}{h} = e^x lim_{h\to 0}\frac{e^h-1}{h}$. Set $e^h - 1 = n$, so that, as $h\to 0$, $n\to 0$.Then $e^h = n+1 \Rightarrow lne^h = ln(n+1) \Rightarrow h = ln(n+1)$. Therefore,

$f'(x) = e^{x} lim_{n\to 0} \frac{n}{ln(n+1)} = e^{x} lim_{n\to 0} \frac{n}{ln(n+1)} \times \frac{\frac{1}{n}}{\frac{1}{n}} =$ $e^{x} lim_{n\to 0} \frac{1}{\frac{1}{n}ln(n+1)} = e^{x} lim_{n\to 0} \frac{1}{ln(n+1)^{\frac{1}{n}}} = e^{x} \frac{1}{ln\left(lim_{n\to 0}(n+1)^{\frac{1}{n}}\right)}$

where $ln\left(lim_{n\to 0}(n+1)^{\frac{1}{n}}\right) = lne = 1$, since $lim_{n\to 0}(n+1)^{\frac{1}{n}} = e$, and, hence, we have proved that $\frac{d}{dx}(e^{x}) = e^{x}$.

iii. $\frac{d}{dx}(lnx) = \frac{1}{x}$; notice that the method of implicit differentiation implies that, since $y = lnx$, $e^{y} = x \Leftrightarrow e^{y}\frac{dy}{dx} = 1 \Leftrightarrow \frac{dy}{dx} = \frac{1}{e^{y}} = \frac{1}{x}$.
Using logarithmic differentiation and implicit differentiation, we can prove the power rule for any real number n as follows:
Let us define $y = x^{n}$, and then take the natural logarithm of both sides: $lny = lnx^{n} \Rightarrow lny = nlnx \Rightarrow \frac{y'}{y} = n\frac{1}{x} \Rightarrow y' = y\frac{n}{x} = x^{n}\frac{n}{x} = nx^{n-1}$.

iv. $\frac{d}{dx}(a^{x}) = a^{x}lna$, since we can set $y = a^{x} \Leftrightarrow lny = x \cdot lna$ and then differentiate both sides implicitly with respect to x, obtaining $\frac{1}{y}y' = lna \Rightarrow y' = ylna$, where $y = a^{x}$.

v. $\frac{d}{dx}(x^{x}) = x^{x}(1 + lnx)$; notice that we can set $y = x^{x} \Leftrightarrow lny = lnx^{x} = xlnx$ (and then we apply the product rule).

vi. $\frac{d}{dx}(log_{a}x) = \frac{1}{xlna}$, since, by the method of implicit differentiation, setting $y = log_{a}x$, we get $a^{y} = x \Leftrightarrow (lna) \cdot a^{y} \cdot \frac{dy}{dx} = 1 \Leftrightarrow \frac{dy}{dx} = \frac{1}{lna} \cdot \frac{1}{a^{y}} = \frac{1}{lna} \cdot \frac{1}{x}, x > 0$.

vii. $\frac{d}{dx}(sinx) = cosx, x \in \mathbb{R}$; and $\frac{darcsinx}{dx} = \frac{1}{\sqrt{1-x^{2}}}$ for $-1 < x < 1$.
Remark: We can prove $\frac{d}{dx}(sinx) = cosx$ by applying the limit definition of the derivative, some basic trigonometric formulae, and the small angle approximation (i.e., $lim_{x\to 0}\frac{sinx}{x} = 1$). If $f(x) = sinx$, then we have:
$f'(x) = lim_{h\to 0}\frac{f(x+h)-f(x)}{h} = lim_{h\to 0}\frac{sin(x+h)-sin(x)}{h}$, and, by using the sum and difference of angles in trigonometry (i.e., $sin(A + B) = sinAcosB + cosAsinB$), the above limit can be restated as follows: $f'(x) = lim_{h\to 0}\frac{sinxcosh+cosxsinh-sinx}{h} = lim_{h\to 0}\frac{[-sinx(1-cosh)+cosxsinh]}{h} = lim_{h\to 0}\frac{[-sinx(1-cosh)]}{h} +$

$lim_{h\to 0}\frac{cosxsinh}{h} = (-sinx)\left[lim_{h\to 0}\frac{(1-cosh)}{h}\right] + (cosx)lim_{h\to 0}\frac{sinh}{h}$.

Now, by using the half-angle formula $1 - cosh = 2sin^2\frac{h}{2}$, the above equation can be restated as follows:

$$f'(x) = (-sinx)lim_{h\to 0}\frac{2sin^2\frac{h}{2}}{h} + (cosx)lim_{h\to 0}\frac{sinh}{h}$$
$$= (-sinx)\left[lim_{h\to 0}\left(\frac{sin\frac{h}{2}}{\frac{h}{2}}\right)\cdot lim_{h\to 0}sin\frac{h}{2}\right]$$
$$+ (cosx)lim_{h\to 0}\frac{sinh}{h}$$

which, by the small angle approximation, gives

$f'(x) = (-sinx)\left(1\cdot sin\frac{0}{2}\right) + cosx(1) = (-sinx)(0) + cosx = cosx.$

In order to compute $\frac{darcsinx}{dx}$, we work as follows: Set $y = sin^{-1}x = arcsinx$ and $siny = x$, and then take the derivative of both sides of the equation and solve for y', namely: $siny = x \Rightarrow (cosy)\cdot y' = 1 \Rightarrow y' = \frac{1}{cosy}$. Recall that $cos^2y + sin^2y = 1 \Rightarrow cosy = \sqrt{1 - sin^2y}$, $cosy > 0$ on the range of $y = sin^{-1}x$. Plugging this in the above equation for y', we obtain

$y' = \frac{1}{cosy} = \frac{1}{\sqrt{1-sin^2y}} = \frac{1}{\sqrt{1-x^2}}$.

Following similar techniques, we can prove the derivatives of the other trigonometric functions.

viii. $\frac{d}{dx}(cosx) = -sinx$, $x \in \mathbb{R}$; and $\frac{darccosx}{dx} = \frac{-1}{\sqrt{1-x^2}}$ for $-1 < x < 1$.

ix. $\frac{d}{dx}(tanx) = \frac{1}{cos^2x} = sec^2x$; and $\frac{darctanx}{dx} = \frac{1}{1+x^2}$.

x. $\frac{d}{dx}(cotx) = -\frac{1}{sin^2x} = -csc^2x$; and $\frac{darccotx}{dx} = \frac{-1}{1+x^2}$.

xi. Hyperbolic functions: $\frac{d}{dx}(sinhx) = \frac{d}{dx}\left(\frac{e^x - e^{-x}}{2}\right) = \frac{1}{2}\left[\frac{d}{dx}(e^x) - \frac{d}{dx}(e^{-x})\right] = \frac{1}{2}(e^x + e^{-x}) = coshx$; and, similarly, we find $\frac{d}{dx}(coshx) = sinhx$, $\frac{d}{dx}(tanhx) = sech^2x$, and $\frac{d}{dx}(cothx) = -csch^2x$.

Investigation of the behavior of a function using differential calculus: If a function $y = f(x)$ is differentiable on an interval (a, b), then:

i. f is increasing on the interval (a, b) if and only if its derivative is non-negative in this interval; symbolically: $f'(x) \geq 0 \; \forall \, x \in (a, b)$;

ii. f is decreasing on the interval (a, b) if and only if its derivative is non-positive in this interval; symbolically: $f'(x) \leq 0 \; \forall \, x \in (a, b)$.

Geometric significance: A differentiable function increases where its graph has positive slopes, and decreases where its graph has negative slopes. If $f'(x) = 0$, then $f(x)$ is constant (in a sense, it increases and decreases *simultaneously*).

We often have to solve optimization problems—that is, to choose from various variants the best one for some reasons. For instance, builders must know how to select the dimensions of a square beam in order to ensure its best tensile strength, aircraft builders must know what orbit ensures minimum fuel consumption, agronomists must know what seeding rate will guarantee the richest harvest, logistics managers must know how to minimize the transportation cost, production managers must know how to minimize costs and maximize utility, artillery officers must know what inclination of a gun tube will result in the greatest range of fire, and so on. Most optimization problems reduce to finding the extreme values, meaning the greatest and the lowest values, of a function.

A point $x = c$ is called a "point of maximum" (resp. "minimum") for a function $y = f(x)$ if there is a neighborhood $(c - \delta, c + \delta)$ of this point in which the inequality $f(x) \leq f(c)$ (resp. $f(x) \geq f(c)$) holds. If a function $y = f(x)$ has an extremum (i.e., a maximum or a minimum) at a point x_0 of its domain, then the derivative of the given function either does not exist or is equal to zero at this point; because, at a point of extremum, the tangent line to the graph of the function is either horizontal or, in case the gaph has a cusp (i.e., a sharp bend or a corner), does not exist at all. In particular, at a "cusp" x_0, the right-hand derivative is not equal to the left-hand derivative, that is, $f'_+(x_0) \neq f'_-(x_0)$; and a characteristic example of a cusp is the point $\big(0, f(0)\big)$ where $f(x) = |x|$.

Assume that a function $y = f(x)$ is continuous at a point $x = c$, and that there exists a neighborhood $(c - \delta, c + \delta)$ of this point such that the inequality $f'(x) > 0$ holds in the interval $(c - \delta, c)$, and the inequality $f'(x) < 0$ holds in the interval $(c, c + \delta)$. Then $x = c$ is a "point of maximum" for $f(x)$. In other words, if $f(x)$ increases in the interval $(c - \delta, c)$ to the left of c, and decreases in the interval $(c, c + \delta)$ to the right of c, then $x = c$ is a "point of maximum" for $f(x)$.

On the other hand, assume that a function $y = f(x)$ is continuous at a point $x = c$, and that, for some $\delta > 0$, it holds that $f'(x) < 0$ in the interval $(c - \delta, c)$, and $f'(x) > 0$ in the interval $(c, c + \delta)$. Then $x = c$ is a "point of minimum" for $f(x)$. In other words, if $f(x)$ decreases in the interval $(c - \delta, c)$ to the left of c, and increases in the interval $(c, c + \delta)$ to the right of c, then $x = c$ is a "point of minimum" for $f(x)$.

Consequently, we obtain the following algorithm for investigating a function $y = f(x)$ for an extremum (maximum or minimum):

i. Find the derivative $f'(x)$.

ii. Find the critical points, that is, the points at which the function is continuous and the derivative $f'(x)$ is either equal to zero or does not exist.

iii. Consider the neighborhood of each critical point found that does not contain another critical point and investigate the sign of the derivative to the left and to the right of the critical point under consideration.

iv. Using the aforementioned sufficient conditions for a maximum and a minimum, draw relevant conclusions (when passing through a maximum, the derivative changes sign from plus to minus, whereas, when passing through a minimum, the derivative changes sign from minus to plus).

For instance, let us investigate the function $f(x) = x^3 - 9x^2 + 24x$ for an extremum. We work as follows:

i. We have $f'(x) = 3x^2 - 18x + 24$.

ii. Equating the derivative to zero, we find the two roots (solutions) of the equation $3x^2 - 18x + 24 = 0$, namely: $x_1 = 2$ and $x_2 = 4$ (the curve has horizontal tangents at these values). In this case, the derivative is defined everywhere, and, therefore, there are no other critical points.

iii. We study the behavior of the function in a neighborhood of the point $x_1 = 2$ and in a neighborhood of the point $x_2 = 4$. We see the following: when passing through the point $x_1 = 2$, the derivative changes sign from plus to minus, whereas, when passing through the point $x_2 = 4$, the derivative changes sign from minus to plus.

iv. At $x_1 = 2$, the function has a maximum $f_{max} = 20$. At $x_2 = 4$, the function has a minimum $f_{min} = 16$.

We have, thus, learnt that the first derivative of a function, f', provides important information about f. Now, we shall apply the same techniques to f' itself, and learn what this tells us about f. Therefore, we shall study f''.

A function $f(x)$ is said to be "concave up" on an interval X if all the tangents to $f(x)$ on X are below the graph of $f(x)$, as shown, for instance, in Figure 8-6 (i.e., it "opens" up). A function $f(x)$ is said to be "concave down" on an interval X if all the tangents to $f(x)$ on X are above the graph of $f(x)$, as shown, for instance, in Figure 8-7 (i.e., it "opens" down).

Figure 8-6: A concave-up function.

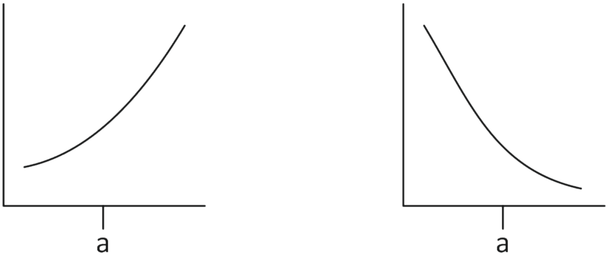
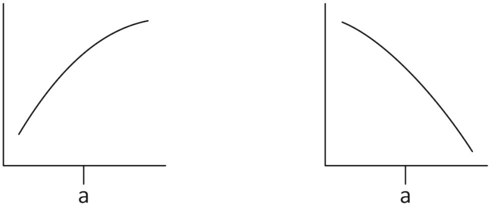


Figure 8-7: A concave-down function.



Let f be a function differentiable on (a, b). (i) If f' is increasing (namely, if $f''(x) > 0$ on (a, b)), then f is concave up on (a, b). (ii) If f' is decreasing (namely, if $f''(x) < 0$ on (a, b)), then f is concave down on (a, b). (iii) If f' is constant, then the graph of f has no concavity.

If $f: (a, b) \to \mathbb{R}$ changes its direction of concavity at x_0, then the point $(x_0, f(x_0))$ is said to be a "point of inflection." In other words, x_0 is a point of inflection if $x_0 \in (a, b)$ so that either f is concave down in (a, x_0) and concave up in (x_0, b), or f is concave up in (a, x_0) and concave down in (x_0, b).

Notice that, if x_0 is a critical point of $f(x)$ and the second derivative of $f(x_0)$ is positive (resp. negative), then x_0 is a "local minimum" (resp. a "local maximum") of $f(x)$. In other words, if the critical point has positive concavity (i.e., $f''(x_0) > 0$), then it is a local minimum; and, if the critical point has negative concavity (i.e., $f''(x_0) < 0$), then it is a local maximum.

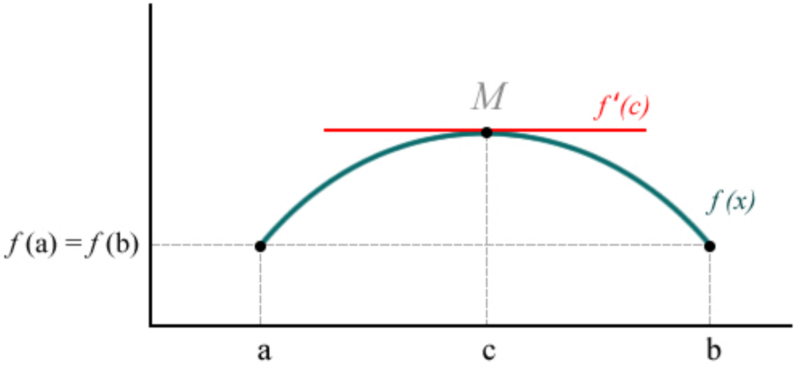
Rolle's Theorem: Let $f: [a, b] \to \mathbb{R}$ be a function satisfying the following conditions:

i. f is continuous on the closed interval $[a, b]$,
ii. f is differentiable on the open interval (a, b), and
iii. $f(a) = f(b)$.

Then there exists at least one point $c \in (a, b)$ such that $f'(c) = 0$.
Proof: Since f is continuous on the closed interval $[a, b]$, it is bounded and attains its supremum (least upper bound) and its infimum (greatest lower bound) in $[a, b]$. Let $\inf(f) = m$, $\sup(f) = M$, and $f(a) = f(b) = k$. Then it must hold that $m \leq k \leq M$.
First case: If $m = k = M$ (i.e., if f is a constant function), then $f(x) = k$, and, therefore, $f'(c) = 0\ \forall c \in (a, b)$.
Second case: If $m \neq M$, then $m < k$ or $k < M$. Suppose that $k < M$. There exists a $c \in (a, b)$ such that $f(c) = M$, since, if f is continuous on the closed interval $[a, b]$, then it attains its supremum and its infimum in $[a, b]$. Moreover, $f'(c)$ exists, because $a < c < b$. Notice that $f(x) \leq M\ \forall x \in [a, b]$. Therefore, if $a \leq x < c$, then $\frac{f(x)-f(c)}{x-c} = \frac{f(x)-M}{x-c} \geq 0$, so that $lim_{x \to c^-} \frac{f(x)-f(c)}{x-c} \geq 0 \Leftrightarrow f_-^{'}(\mathrm{c}) \geq 0$. If $c < x \leq b$, then $\frac{f(x)-f(c)}{x-c} = \frac{f(x)-M}{x-c} \leq 0$, so that $lim_{x \to c^+} \frac{f(x)-f(c)}{x-c} \leq 0 \Leftrightarrow f_+^{'}(c) \leq 0$. Consequently, $0 \leq f_-^{'}(c) = f^{'}(c) = f_+^{'}(c) \leq 0 \Rightarrow f'(c) = 0$. We can work similarly in order to prove the theorem for $m < k$.■
Geometric interpretation of Rolle's Theorem: Under the above conditions, there exists a point c at which the tangent line to the graph of $y = f(x)$ is parallel to the x-axis, as shown in Figure 8-8. In particular, conditions (i) and (ii) imply that the curve $y = f(x)$ is continuous from $x = a$ to $x = b$, and it has a definite tangent at each point between $x = a$ and $x = b$; and condition (iii) implies that the ordinates at the endpoints a and b are equal.
Algebraic interpretation of Rolle's Theorem: Since, according to condition (iii), $f(a) = f(b)$, let $f(a) = f(b) = 0$. Then Rolle's Theorem means that, if $f(x)$ is a polynomial in x, and if a and b are two roots of the equation $f(x) = 0$, then the equation $f'(x) = 0$ has at least one root between a and b. In fact, the French mathematician Michel Rolle, after whom the above theorem is named, proved the given theorem in 1691 only in the case of polynomial functions, and a general proof of this theorem was achieved and published by Augustin-Louis Cauchy in 1823. The name

"Rolle's Theorem" was first used by the German mathematician, logician, psychologist, and philosopher Moritz Wilhelm Drobisch in the 1830s.

Figure 8-8: Rolle's Theorem (source: Wikimedia Commons: Author: Benboyadjian; https://commons.wikimedia.org/wiki/File:Teorema_de_Rolle_(caso_2).jpg?uselang=it).



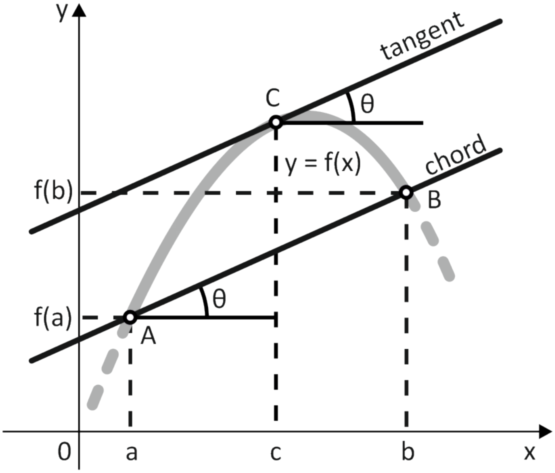
In mathematical analysis, the mean value theorems play a very important role, because they examine the relationship between the values of a function and the values of the derivative of the given function. The Italian-French mathematician and astronomer Joseph-Louis Lagrange (1736–1813) proved the following mean value theorem, which allows us to express the increment of a function on an interval through the value of the derivative at an intermediate point of the corresponding segment:

Lagrange's Mean Value Theorem: If $f:[a,b]\to\mathbb{R}$ is a function continuous on $[a,b]$ and differentiable on (a,b), then there exists a point $c\in(a,b)$ such that $f'(c)=\frac{f(b)-f(a)}{b-a}\Leftrightarrow f(b)-f(a)=f'(c)(b-a)$.
Proof: On $[a,b]$, we define another function $g(x)$ as follows:
$g(x)=f(x)-kx$ for all $x\in[a,b]$,
where k is a constant defined in such a way that
$g(a)=g(b)$.
Thus, $f(a)-ka=f(b)-kb\Rightarrow k=\frac{f(b)-f(a)}{b-a}$.
The assumptions that $f:[a,b]\to\mathbb{R}$ is a function continuous on $[a,b]$ and differentiable on (a,b) and the above value of k imply that the above function $g(x)$ satisfies every condition of Rolle's theorem. Therefore, by

Rolle's theorem, there exists a $c \in (a,b)$ such that $g'(c) = 0 \Rightarrow f'(c) - k = 0 \Rightarrow f'(c) = k = \frac{f(b)-f(a)}{b-a}$, *quod erat demonstrandum.*

Geometric interpretation of Lagrange's Mean Value Theorem: As shown in Figure 8-9, Lagrange's Mean Value Theorem implies that the slope of the chord passing through the points of the graph corresponding to the ends of the segment a and b is equal to $k = tan\theta = \frac{f(b)-f(a)}{b-a}$, and then there exists a point $x = c$ inside the closed interval $[a,b]$ such that the tangent to the graph at $x = c$ is parallel to the chord. In other words, if a function f is continuous on the closed interval $[a,b]$ and differentiable on the open interval (a,b), then there exists a point c in the interval (a,b) such that $f'(c)$ is equal to the function's average rate of change over $[a,b]$.

Figure 8-9: Lagrange's Mean Value Theorem.



Corollary 1: If $f'(x) = 0$ for all $x \in (a,b)$, then $f(x)$ is constant on (a,b).

Proof: Let x_1 and x_2 be two arbitrary elements of the interval (a,b). Then, since $f(x)$ is continuous and differentiable on (a,b), it must also be continuous and differentiable on $[x_1, x_2]$. Therefore, we can apply the Mean Value Theorem for x_1 and x_2. This means that $f(x_2) - f(x_1) = f'(c)(x_2 - x_1)$ where $x_1 < c < x_2$. By hypothesis, $f'(c) = 0$. Hence, $f(x_2) - f(x_1) = 0 \Rightarrow f(x_2) = f(x_1)$, and, since x_1 and x_2 are two arbitrary elements of the interval (a,b), the function f is a constant on (a,b).■

Corollary 2: If $f'(x) = g'(x)$ for all $x \in (a,b)$, then, in this interval, $f(x) = g(x) + c$, where c is a constant.

Proof: This is a direct result of Corollary 1.■

Example 1: Given $f(x) = x^2 + x + 1$, if we are asked to find the point c at which $f'(x)$ gets its mean value over [0,2], then we work as follows: we confirm that the hypotheses of Lagrange's Mean Value Theorem are satisfied, and, therefore, $\exists c \in (a,b) | \frac{f(b)-f(a)}{b-a} = f'(c) \Rightarrow \frac{f(2)-f(0)}{2-0} = 3 = f'(c) = 2c + 1 \Rightarrow c = 1$.

Example 2: Let $0 < a < b$. Then we can prove that

$$1 - \frac{a}{b} < ln\frac{b}{a} < \frac{b}{a} - 1$$

as follows: Set $f(x) = lnx$. By Lagrange's Mean Value Theorem,

$f'(c) = \frac{f(b)-f(a)}{b-a}, a < c < b,$

so that $\frac{1}{c} = \frac{lnb-lna}{b-a} = \frac{lnb/a}{b-a}$.

We have: $a < c < b \Rightarrow \frac{1}{b} < \frac{1}{c} < \frac{1}{a}$ because $0 < a < b$. Therefore,

$\frac{1}{b} < \frac{lnb/a}{b-a} < \frac{1}{a} \Rightarrow \frac{b-a}{b} < ln\frac{b}{a} < \frac{b-a}{a} \Rightarrow 1 - \frac{a}{b} < ln\frac{b}{a} < \frac{b}{a} - 1$, *quod erat demonstrandum.*

Cauchy's Mean Value Theorem: If functions $f(x)$ and $g(x)$ are continuous on a closed interval $[a,b]$ and differentiable on the open interval (a,b), then there exists some point $c \in (a,b)$ such that

$[f(b) - f(a)]g'(c) = [g(b) - g(a)]f'(c),$

which, for $g'(x) \neq 0$ for all $x \in (a,b)$, can be equivalently restated as

$\frac{f'(c)}{g'(c)} = \frac{f(b)-f(a)}{g(b)-g(a)}.$

Proof: Consider the function

$h(x) = [f(x) - f(a)][g(b) - g(a)] - [f(b) - f(a)][g(x) - g(a)].$

This function is continuous on $[a,b]$ and differentiable on (a,b), with

$h'(x) = f'(x)[g(b) - g(a)] - g'(x)[f(b) - f(a)].$

Moreover, $h(a) = 0 = h(b)$. Hence, by Rolle's theorem, there exists a point $c \in (a,b)$ such that $h'(c) = 0$, and then $[f(b) - f(a)]g'(c) = [g(b) - g(a)]f'(c)$.■

Remark: When $g(x) = x$, Cauchy's Mean Value Theorem reduces to Lagrange's Mean Value Theorem. Cauchy's Mean Value Theorem can be geometrically interpreted as follows: the functions $f(x)$ and $g(x)$ determine a plane curve with parametric equations $x = f(t)$ and $y =$

$g(t)$, where $t \in [a, b]$, and then Cauchy's Mean Value Theorem states that, for some $c \in (a, b)$, there exists a point $\big(f(c), g(c)\big)$ on this plane curve such that the slope $\frac{f'(c)}{g'(c)}$ of the tangent line to the curve at this point is equal to the slope of the chord that joins the endpoints of the curve.

Optimization problems:

i. *Maximum area enclosed by a fence:* Assume that a man has a farm that is adjacent to a river, and he wants to build a rectangular pen for his cows with $500ft.$ of fencing. Given that one side of the pen is the river, which functions as a natural fence (since cows will not swim away), the largest area of the pen that he can build can be calculated as follows: This rectangular field needs to be fenced on 3 sides, and two of these sides (which need to be fenced), say side s_1 and side s_2, are equal to each other. Let $s_1 = s_2 = x\,ft$, meaning that this man will use $2x\,ft$ of fencing for these two sides, and the remaining amount of fencing will be $500 - 2x\,ft$, corresponding to the third side of this rectangular farm. Then (given that $area = width \times length$) we have to maximize the area function $A(x) = x(500 - 2x\,)$. We have to differentiate and find the critical points: $A'(x) = 500 - 4x$. We want to know where $A'(x)$ is equal to zero and where $A'(x)$ is undefined. However, $A'(x) = 500 - 4x$ is defined for all values of x. Setting $A'(x) = 500 - 4x = 0$, we obtain $x = 125$, which is the critical point of $A(x)$. By checking the behavior of $A'(x)$ around $x = 125$, we realize that, when passing through the point $x = 125$, the derivative changes sign from plus to minus, and, therefore, $x = 125$ is a maximum for $A(x)$. Therefore, the area of the largest pen that this man can build is $A(125) = 125(500 - 2 \times 125) = 31{,}250ft^2$.

ii. *Closest points (minimum distance between a curve and a point):* We can determine the points on $y = 6 - x^2$ that are closest to the point $(0,3)$ as follows: we have to minimize the distance function $d^2 = (x - 0)^2 + (y - 3)^2$ subject to the constraint of $y = 6 - x^2$. Hence,

$$d = f(x) = \sqrt{(x - 0)^2 + (y - 3)^2} = \sqrt{x^2 + (6 - x^2 - 3)^2} \Rightarrow f(x) = \sqrt{x^2 + (3 - x^2)^2}.$$

Now, we have to minimize $f(x)$, and, therefore, we have to find its critical points, that is, the points at which the derivative $f'(x)$ is equal to zero or undefined. Firstly, we have to compute $f'(x)$, using the chain rule:

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

$$= \frac{-10x + 4x^3}{2\sqrt{x^2 + (3 - x^2)^2}}$$

(the derivative is equal to zero when the numerator is equal to zero, and the derivative is undefined when the denominator is equal to zero; but notice that, in the denominator of this fraction, the expression $\sqrt{x^2 + (3 - x^2)^2}$ is $d = f(x)$, and, therefore, it will never be equal to zero, since it represents the distance between a point on the y-axis, specifically, $(0,3)$, and the parabola $y = 6 - x^2$, meaning that the denominator is always different from zero, and, therefore, the derivative is always defined). The derivative $f'(x) = 0$ when $-10x + 4x^3 = 0 \Rightarrow x(4x^2 - 10) = 0 \Rightarrow 4x\left(x^2 - \frac{10}{4}\right) = 0 \Rightarrow 4x\left[x^2 - \left(\sqrt{\frac{5}{2}}\right)^2\right] = 4x\left(x - \sqrt{\frac{5}{2}}\right)\left(x + \sqrt{\frac{5}{2}}\right) = 0,$

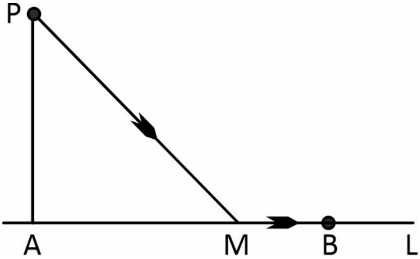
and, therefore, we obtain three solutions, which are the three critical values of $f(x)$, namely: 0, $\sqrt{\frac{5}{2}}$, and $-\sqrt{\frac{5}{2}}$; and we have to find out which of them is the minimum value. Investigating the manner in which the sign of the derivative $f'(x)$ changes when passing through each of these three points, we realize that $\sqrt{\frac{5}{2}}$ is a minimum, and $-\sqrt{\frac{5}{2}}$ is a minimum (indeed, given the parabola $y = 6 - x^2$, which is symmetric with respect to the y-axis, and the point $(0,3)$, which is on the y-axis, we expected to obtain two minimum values, and we expected them to be symmetric with respect to the y-axis). Consequently, $f(x) = d$ has a minimum when $x = \pm\sqrt{\frac{5}{2}}$, since $f'(x)$ changes from negative to positive at those x values.

iii. *Optimal route:* A team of archaeologists is camped on an archaeological site $9km$ to the north of a highway running from west to east. There is a town $15km$ to the east of the nearest point on the highway to the camp. The archaeologists send a messenger by bicycle to the town. What route should the messenger follow in order to reach the town in the shortest time if he can ride at $8\,km/h$ across the archaeological site and at $10\,km/h$ along the highway? In Figure 8-10, we can see a model of this problem: The point P denotes the location of this team of archaeologists, the straight line L denotes the highway, B denotes the town, $PA = 9km$, $AB = 15km$, PMB is the route of the messenger, and the position of the point M between A and B is not known yet. The quantity that we must optimize is the time t during which the messenger will move from P to B. Let $AM = x$. According to the problem, the point M may be anywhere between A and B including points A and B . Therefore, the real bounds within which x varies are $0 \leq x \leq 15$. We express time t in terms of x as follows: We have $PM = \sqrt{PA^2 + AM^2} = \sqrt{81 + x^2}$. The messenger, using his bicycle, covers this distance at a speed of $8\,km/h$, that is, $t_1 = \frac{\sqrt{81+x^2}}{8}$. Moreover, $MB = 15 - x$, and the messenger, using his bicycle, covers this distance at a speed of $10\,km/h$, so that $t_2 = \frac{15-x}{10}$. Hence, the total time t that the messenger spent in order to cover the entire distance is equal to $t_1 + t_2$, that is,

$$t = \frac{\sqrt{81+x^2}}{8} + \frac{15-x}{10}.$$

We have to minimize the function $t = \frac{\sqrt{81+x^2}}{8} + \frac{15-x}{10}$ in the closed interval $[0,15]$. The derivative $\frac{dt}{dx} = \frac{x}{8\sqrt{81+x^2}} - \frac{1}{10}$ exists for all x, and we have to find the points at which $\frac{dt}{dx} = 0$. We have $\frac{x}{8\sqrt{81+x^2}} - \frac{1}{10} = 0 \Rightarrow x = 12$, which belongs to the closed interval $[0,15]$. Let's examine the values of the function t at the endpoints of the closed interval $[0,15]$ and at $x = 12$ in order to find the least value of t. At $x = 0$, $t = 105/40$. At $x = 12$, $t = 87/40$. At $x = 15$, $t = 5\sqrt{306}/40$. Thus, the least value of t is $t = 87/40$, and it is reached for $x = 12$. This result implies that the messenger, using his bicycle, has to ride along a route PMB such that the distance between the points A and M on the highway is equal to $12km$. This type of problems is very common in logistics and transportation.

Figure 8-10: An optimization model.



iv. In view of what we discussed in Chapter 4, we have:
Total Cost: $C(x) = fixed\ cost + variable\ cost$ for producing x items, where fixed cost consists of all types of cost that do not change with the level of output (e.g., the rent of the premises, the insurance, taxes, etc.), and variable cost is the sum of all costs that are dependent on the level of production (e.g., labor cost, the cost of raw materials, the cost of energy, the cost of packaging, etc.).
Total Revenue: $R(x) = xp(x)$, where x denotes units sold, and p denotes price per unit (i.e., $R(x)$ is the revenue obtained from selling x items).
Demand Function (linear approximation): $Q_D = a + bp$, where a stands for the quantity-intercept (i.e., the x-intercept) of the demand (i.e., a is quantity demanded when price is zero, and it is known as the "autonomous demand"), b measures the change in quantity demanded resulting from a particular change in price (i.e., indicating the responsiveness of consumers to a particular increase in the price of a commodity, $b = \frac{\Delta Q_D}{\Delta P}$, which, given Figure 4-1, is the reciprocal of the slope of the demand curve), and p denotes price (notice that the independent variable p is graphically represented by the vertical axis, that is, by the y-axis, whereas the dependent variable Q_D is graphically represented

by the horizontal axis, that is, by the x-axis, as shown in Figure 4-1).

Supply Function (linear approximation): $Q_S = k + lp$, where k denotes the quantity-intercept (i.e., the x-intercept) of the supply (usually, it is negative, since, at a price of zero, no producers are generally willing or able to provide a commodity; but, in the case of some subsidies, the value of k may be positive), l denotes the price coefficient of supply (i.e., indicating the responsiveness of producers to a particular increase in the price of a commodity, l is given by the change in quantity supplied divided by the change in price, and, thus, $l = \frac{\Delta Q_S}{\Delta P}$, which, given Figure 4-1, is the reciprocal of the slope of the supply curve), and p denotes price (notice that the independent variable p is graphically represented by the vertical axis, that is, by the y-axis, whereas the dependent variable Q_S is graphically represented by the horizontal axis, that is, by the x-axis, as shown in Figure 4-1).

Profit: $P = R - C$, where R denotes revenue, and C denotes cost.

Break-Even Point: the point at which $R(x) = C(x)$, that is, the point at which the revenue function and the cost function cross.

Average Cost: $\bar{C} = \frac{C(x)}{x}$, that is, the cost per unit item.

Average Price: $\bar{p} = \frac{p(x)}{x}$, that is, the price per unit item.

Marginal Revenue: $R'(x) = \frac{dR(x)}{dx}$.

Marginal Cost: $C'(x) = \frac{dC(x)}{dx}$.

Minimization of Average Cost: In order to find the level of output for which the average cost is minimum, we define the function of the average cost, say $AC(x)$, in the case under consideration, and then we calculate the cost-minimizing level of output x by solving $AC'(x) = \frac{dAC(x)}{dx} = 0$, and, finally, we get the value of x for which $AC''(x) = \frac{d^2AC(x)}{dx^2} > 0$.

Maximization of Total Revenue: In order to find the level of output for which the total revenue is maximum, we define the function of the total revenue, say $R(x)$, in the case under consideration, and then we calculate the revenue-maximizing level of output x by solving $R'(x) = \frac{dR(x)}{dx} = 0$, and, finally, we get the value of x for which $R''(x) = \frac{d^2R(x)}{dx^2} < 0$.

Marginal Profit: $P'(x) = R'(x) - C'(x)$, that is, marginal profit is defined to be the difference between marginal revenue and marginal cost.

Maximization of Profit: In order to find the level of output for which the profit is maximum, we define the function of the profit, say $P(x)$, in the case under consideration, and then we calculate the profit-maximizing level of output x by solving $P'(x) = \frac{dP(x)}{dx} = 0$, and, finally, we get the value of x for which $P''(x) = \frac{d^2P(x)}{dx^2} < 0$.

Polynomial approximation of functions: the formulae of Taylor and MacLaurin: Polynomial functions are always continuous everywhere (i.e., at any real value), and they are also differentiable for all arguments. Moreover, polynomial functions being linear combinations of $1, x, x^2, x^3, \dots, x^n$, they are easier to differentiate and integrate than other functions, and algorithms have been devised to differentiate and integrate polynomial functions, whereas often there are no such algorithms for other functions, and this often compels us to use laborious graphing techniques to solve problems. Hence, it is very important to be able to approximate any function by means of polynomials.

Suppose that we have a function $f(x)$ and we want to express this function, or approximate this function, as a polynomial, symbolically, $f(x) \approx p_n(x)$, where n symbolizes that the highest power of x we are going to consider is x^n, and $f(x)$ can be any function you can think of as long as it is differentiable. The key idea that underpins the polynomial approximation of functions is the following: Firstly, we define $p_n(x)$ by expanding it around some general point a, namely:

$p_n(x) = a_0 + a_1(x - a) + a_2(x - a)^2 + \cdots + a_n(x - a)^n$,

and we stop at the nth power because we have fixed our polynomial $p_n(x)$ to be of maximum degree n. Then we have to decide how to define the coefficients a_k (where $k = 0,1,2, \dots, n$) of $p_n(x)$. In particular, we define the coefficients of $p_n(x)$ in such a way that the kth derivative of $p_n(x)$ at the point a is equal to the kth derivative of the function $f(x)$ at the point a, symbolically:

$p_n^{(k)}(a) = f^{(k)}(a)$.

For instance,

$p_n'(x) = a_1 + 2a_2(x-a) + \cdots + na_n(x-a)^{n-1}$,
$p''(x) = 2a_2 + \cdots + n(n-1)(x-a)^{n-2}$,
$\vdots$

Notice that

$$p_n^{(k)}(x) = a_k k! + a_{k+1}(k+1)k(k-1)\ldots 2(x-a) + \cdots \quad (1)$$

where, after the term $a_k k!$, every other term of $p_n^{(k)}(x)$ includes a power of $(x-a)$. But the rule that we use in order to determine the coefficients of $p_n(x)$ is $p_n^{(k)}(a) = f^{(k)}(a)$, and, if we set $x = a$ in the polynomial (1), the first term, that is, $a_k k!$, remains, because it is a constant, and every other term vanishes, because it includes a power of $(x-a)$. Hence,

$$p_n^{(k)}(a) = f^{(k)}(a) = a_k k! \Rightarrow a_k = \frac{f^{(k)}(a)}{k!}$$

(this is the formula for determining the coefficients of the approximating polynomial, which, in fact, leads us to Taylor's formula), and

$$p_n(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n.$$

Furthermore, due to Cauchy's Mean Value Theorem, we can argue that, because the functions $f(x)$ and $p_n(x)$ agree on these n derivatives at this point a, they are in fact almost the "same," in the sense that $p_n(x)$ tends to become exactly $f(x)$ as we continue this pattern forever; that is, as we increase the degree of the approximating polynomial, we get better approximations of the function $f(x)$.

Generalization: If $f: \mathbb{R} \to \mathbb{R}$ is a continuous function such that f has continuous derivatives of all orders at $x = a$, then $f(x)$ can be expanded in a power series as follows:

$$f(x) \approx f(a) + \frac{x-a}{1!}f'(a) + \frac{(x-a)^2}{2!}f''(a) + \cdots + \frac{(x-a)^n}{n!}f^{(n)}(a) + \cdots$$

(we get a more and more accurate approximation of $f(x)$ the more terms we take, that is, the more derivatives of $f(x)$ we calculate at a). This equation is known as Taylor's formula, and it approximates a function around a point.

For $a = 0$, in particular, we obtain:

$$f(x) \approx f(0) + \frac{x}{1!}f'(0) + \frac{x^2}{2!}f''(0) + \cdots + \frac{x^n}{n!}f^{(n)}(0) + \cdots$$

(this equation is a special case of Taylor's formula, and it approximates a function around the origin). This equation is known as MacLaurin's formula.

Examples: For any $x \in \mathbb{R}$, MacLaurin's formula implies that

$$e^x \approx 1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + \cdots \Rightarrow e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

since $f(x) = e^x \Rightarrow f^{(n)}(x) = e^x$, and then $f^{(n)}(0) = 1$;

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

since $f(x) = sinx \Rightarrow f^{(n)}(x) = sin\left(x + n\frac{\pi}{2}\right)$, and then $f^{(n)}(0) = sin\left(n\frac{\pi}{2}\right) = \begin{cases} 0 \; when \; n = 2k \\ (-1)^n \; when \; n = 2k+1 \end{cases}$;

$$cosx = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots \Rightarrow cosx = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!}$$

since $f(x) = cosx \Rightarrow f^{(n)}(x) = cosx\left(x + n\frac{\pi}{2}\right)$, and then $f^{(n)}(0) = cos\left(n\frac{\pi}{2}\right) = \begin{cases} 0 \; when \; n = 2k+1 \\ (-1)^n \; when \; n = 2k \end{cases}$.

The binomial series: Let $f(x) = (1+x)^m$, where m is an arbitrary rational number (positive or negative). Then
$f^{(n)}(x) = m(m-1)\ldots(m-n+1)(1+x)^{m-n}$;
and MacLaurin's formula implies that
$(1+x)^m \approx 1 + \binom{m}{1}x + \binom{m}{2}x^2 + \cdots$,
where the binomial coefficient is defined by
$\binom{m}{k} = \frac{m!}{k!(m-k)!}$, where $m! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot m$, m is a positive integer, and $m \geq k \geq 0$ (as I have already mentioned, $m = 0 \Rightarrow 0! = 1$, since the binomial coefficient gives the number of combinations of m elements taken k at a time).

Analytic functions: A function is said to be "analytic" in a neighborhood (open disc) centered at x_0 if and only if its Taylor series converges to the value of the function at each point of the neighborhood. For instance, the functions e^x, $sinx$, and $cosx$ are analytic for all $x \in \mathbb{R}$, whereas the function $(1+x)^m$ is analytic in the open interval $(-1,1)$.

L'Hôpital's rule and indeterminate forms: Consider the following limits:
$lim_{x\to a} f(x) = 0$ and $lim_{x\to a} g(x) = 0$.
Then the limit
$lim_{x\to a} \frac{f(x)}{g(x)}$ assumes the form $\frac{0}{0}$,
which is called an "indeterminate form." Other such forms are $\frac{\infty}{\infty}$, $\infty - \infty$, and $0 \cdot \infty$. For evaluating such forms of limits, we apply L'Hôpital's rules:

i. Let f and g be two functions that are both differentiable at each point of the neighborhood $N_\varepsilon(a)$ of a, and let $g'(x) \neq$

0 for all $x \in N_\varepsilon(a)$. If $lim_{x\to a}f(x) = 0$ and $lim_{x\to a}g(x) = 0$, then

$$lim_{x\to a}\frac{f(x)}{g(x)} = lim_{x\to a}\frac{f'(x)}{g'(x)},$$

provided that the limit in the right exists. Notice that, if $x = a$ is the point at which we are trying to take the limit, we can expand both the functions $f(x)$ and $g(x)$ into Taylor series about a, so that $f(x) \approx f(a) + f'(a)(x-a)$ and $g(x) \approx g(a) + g'(a)(x-a)$. But, since $lim_{x\to a}f(x) = 0 = lim_{x\to a}g(x)$, then (as we get infinitely close to a) we have $f(x) \approx f'(a)(x-a)$ and $g(x) \approx g'(a)(x-a)$. Hence, as $x \to a, \frac{f(x)}{g(x)} \approx \frac{f'(a)}{g'(a)}$.

ii. If f and g are two functions such that $lim_{x\to a}f(x) = \infty$ and $lim_{x\to a}g(x) = \infty$, then

$$lim_{x\to a}\frac{f(x)}{g(x)} = lim_{x\to a}\frac{f'(x)}{g'(x)},$$

provided that the limit in the right exists. Notice that $lim_{x\to a}\frac{f(x)}{g(x)}$ can be written as $lim_{x\to a}\frac{1/g(x)}{1/f(x)}$, which reduces to the above $0/0$ form.

The indeterminate forms $\infty - \infty$ and $0 \cdot \infty$ are reducible to the indeterminate form $\frac{0}{0}$. If $lim_{x\to a}f(x) = 0$ and $lim_{x\to a}g(x) = \infty$, then

$$lim_{x\to a}f(x) \cdot g(x) = lim_{x\to a}\frac{f(x)}{\frac{1}{g(x)}},$$

which is of the form $\frac{0}{0}$.

If $lim_{x\to a}f(x) = \infty$ and $lim_{x\to a}g(x) = \infty$, then

$$lim_{x\to a}[f(x) - g(x)] = lim_{x\to a}\left[\frac{\frac{1}{g(x)} - \frac{1}{f(x)}}{\frac{1}{g(x)f(x)}}\right],$$

which is of the form $\frac{0}{0}$.

Example: In case of $lim_{x\to 0}\frac{sinx}{x}$, we can use L'Hôpital's rule, since the limit reduces to the indeterminate form $\frac{0}{0}$, and, therefore,

$$lim_{x\to 0}\frac{sinx}{x} = lim_{x\to 0}\frac{\frac{d}{dx}sinx}{\frac{d}{dx}x} = lim_{x\to 0}\frac{cosx}{1} = \frac{1}{1} = 1.$$

The curvature of a curve: Intuitively, by the term "curvature," we refer to the measure of how sharply a curve bends, that is, of how much a curve deviates from being a straight line. Formally, we can say that a curved line is a line that gradually changes direction from one point to the next, and the rate of this change of direction, per unit length along the curve, is called the "curvature." The ancient Greek mathematician Apollonius calculated the curvature of conic sections, the French scholastic intellectual and mathematician Nicolas d'Oresme understood and studied "curvature" in an abstract way, Descartes studied curvature as a local measure of a curve's bending in the context of the Cartesian coordinate system and his "algebraic calculus," Kepler and Leibniz studied curvature in terms of the "closest" circle at a point (the "osculating circle"), the seventeenth-century Dutch mathematician Christiaan Huygens found a way to calculate the curvature of any curve, and Newton formulated the concept of curvature in its modern form, which will be studied here.

If $y = f(x)$ is a plane curve, then the curvature at any point $P(x, y)$ is expressed in terms of the first and the second derivatives of the function $f(x)$ by the formula

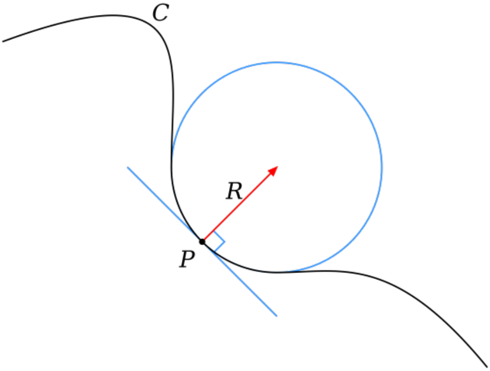
$$K = \frac{|f''(x)|}{[1 + (f'(x))^2]^{\frac{3}{2}}}$$

where K characterizes the speed of rotation of the tangent to the curve at the given point.

Proof: First of all, let us consider the following preliminary concepts and principles:

1. The radius of a circle drawn to a point of tangency between the circle and the tangent line is perpendicular to the tangent line.
2. If two separate lines are tangent to a circle at two different points, then the lines drawn perpendicular to the tangent lines at their points of tangency intersect each other at the circle's center; and each perpendicular line's segment from its point of tangency to the point of intersection is a radius.
3. Historically, the curvature of a (differentiable) curve was defined by means of the "osculating circle," that is, the circle that best approximates the curve under consideration (as shown in Figure 8-11). The curvature of a circle is defined as the reciprocal of its radius (i.e., the curvature of a circle of radius R is $1/R$). For any two nearby points on a curve, at least one circle minimizes the absolute area between the curve and the circle between the two points. This circle's radius can be viewed as approximating the curve's radius of curvature R, and, then, the curve's curvature is $K = \frac{1}{R}$ as the two points approach each other. See, for instance, Figure 8-11.

Figure 8-11: Radius of curvature (source: Wikimedia Commons: Author: Emperorhoney; https://commons.wikimedia.org/wiki/File:Osculating.svg).



4. A curve's curvature between two points approaches its curvature at a point as the two points approach each other (Figure 8-11).
5. The curvature of a circle that minimizes the absolute area between the curve and the circle between two nearby points on the curve approaches the curve's curvature K as the two points approach each other (Figure 8-11).
6. The radius of the absolute area-minimizinng circle approaches the curve's radius of curvature $R = \frac{1}{K}$ as the two points approach each other.
7. The function that represents the slope of a function f is the derivative of f (precisely, the slope of a tangent line at a point).

Consider a curve given by the twice differentiable function

$$y = f(x). \tag{1}$$

Let (x_0, y_0) and (x_1, y_1) be two points on the given curve. Using the point-slope form of a line, and denoting its slope by m_0, the tangent line to the curve at (x_0, y_0) is given by the following equation:

$$(y - y_0) = m_0(x - x_0). \tag{2}$$

Because the slope of a line perpendicular to another line is the negative inverse of that line's slope, the line perpendicular to the tangent line at (x_0, y_0) is given by the following equation:

$$(y - y_0) = -\frac{(x - x_0)}{m_0}. \tag{3}$$

Similarly, the corresponding tangent and perpendicular lines at (x_1, y_1) are, respectively, given by the following equations:

$$(y - y_1) = m_1(x - x_1) \tag{4}$$
and
$$(y - y_1) = -\frac{(x-x_1)}{m_1}. \tag{5}$$
The intersection of the two perpendicular lines approximates the curve's center of curvature. As the distance between x_0 and x_1 tends to zero, the aforementioned intersection becomes the center of the circle of curvature that matches exactly the curve's curvature at the point (x_0, y_0). This intersection is the solution of the simultaneous equations (3) and (5). In fact, substituting equation (3) into equation (5), we obtain
$$(y - y_0) = (y - y_1) + (y_1 - y_0) = -\frac{(x-x_0)}{m_0} \tag{6}$$
$$(y - y_1) = -\frac{(x-x_0)}{m_0} - (y_1 - y_0) \tag{7}$$
$$(x - x_1) = (x - x_0) - (x_1 - x_0) \tag{8}$$
$$-\frac{(x-x_0)}{m_0} - (y_1 - y_0) = -\frac{(x-x_0)}{m_1} + \frac{(x_1-x_0)}{m_1} \tag{9}$$
$$-\frac{(x-x_0)}{m_0} = -\frac{(x-x_0)}{m_1} + \frac{(x_1-x_0)}{m_1} + (y_1 - y_0). \tag{10}$$
Given that $f(x)$ is twice differentiable (by hypothesis), it holds that $(y_1 - y_0) = m(x_1 - x_0)$, where m denotes the curve's slope somewhere in the closed interval $[x_0, x_1]$. We have:
$$(x - x_0)\left(\frac{1}{m_1} - \frac{1}{m_0}\right) = \frac{(x_1-x_0)}{m_1} + m(x_1 - x_0) \tag{11}$$
$$(x - x_0)\left(\frac{m_0-m_1}{m_0 m_1}\right) = \left(\frac{1}{m_1} + m\right)(x_1 - x_0) \tag{12}$$

$$(x - x_0) = \frac{\left(\frac{1}{m_1}+m\right)}{\left(\frac{m_0-m_1}{m_0 m_1}\right)}(x_1 - x_0) \tag{13}$$

$$(x - x_0) = \frac{m_0(1+m m_1)}{(m_0-m_1)}(x_1 - x_0). \tag{14}$$
Substituting equation (14) into equation (3), we obtain:
$$(y - y_0) = -\frac{(x-x_0)}{m_0} = -\frac{(1+m m_1)}{(m_0-m_1)}(x_1 - x_0). \tag{15}$$
Since $y = f(x)$ is given to be twice differentiable, it has a slope at each point, and these slopes can be treated as another function (the slope function of the original function (curve)). Moreover, notice that, just as the tangent line to the original function (curve) at (x_0, y_0), namely, $(y - y_0) = m_0(x - x_0)$, is a good approximation to the original curve near (x_0, y_0), the tangent line to the slope function at (x_0, y_0), namely, $(m - m_0) = n_0(x - x_0)$, where n_0 denotes the slope (rate of change) of the slope function, is a good approximation to the slope function near (x_0, y_0).

Given that $f(x)$ is twice differentiable, it holds that

$$(m_1 - m_0) = n(x_1 - x_0) \tag{16}$$

where the term $(m_1 - m_0)$ is the difference between the original curve's slopes at the points (x_1, y_1) and (x_0, y_0), respectively, and n denotes the slope of the slope function somewhere in the closed interval $[x_0, x_1]$. Substituting equation (16) into equations (14) and (15), we obtain:

$$(x - x_0) = -\frac{m_0(1+mm_1)}{n} \tag{17}$$

and

$$(y - y_0) = \frac{(1+mm_1)}{n}. \tag{18}$$

As x_1 approaches x_0, all the slopes approach their values at (x_0, y_0), and, therefore, $(x - x_0)$ and $(y - y_0)$ approach:

$$(x - x_0) = -\frac{m_0(1+m_0^2)}{n_0} \tag{19}$$

and

$$(y - y_0) = \frac{(1+m_0^2)}{n_0}. \tag{20}$$

Equations (19) and (20) give the x-coordinate and the y-coordinate of the center of the circle that corresponds to the radius of curvature at (x_0, y_0). The "radius of curvature," R, is the distance of the point (x, y) given by equations (19) and (20) from the point (x_0, y_0). Hence,

$$R^2 = (x - x_0)^2 + (y - y_0)^2 = \frac{m_0^2(1 + m_0^2)^2}{n_0^2} + \frac{(1 + m_0^2)^2}{n_0^2} = \frac{(1 + m_0^2)^3}{n_0^2}$$

meaning that, finally,

$$R = \left|\frac{(1 + m_0^2)^{\frac{3}{2}}}{n_0}\right|$$

where, using the terminology of differential calculus, the term m_0 can be written as $\frac{dy}{dx} \equiv f'(x)$, and the term n_0 can be written as $\frac{d^2y}{dx^2} \equiv f''(x)$, so that

$$R = \frac{\left[1+(f'(x))^2\right]^{\frac{3}{2}}}{|f''(x)|}.$$

Given that curvature is the reciprocal of the radius of curvature,

$$K = \frac{1}{R} = \frac{|f''(x)|}{[1+(f'(x))^2]^{\frac{3}{2}}},$$

quod erat demonstrandum.

Curvature is one of the key concepts of differential geometry. Differential geometry is a combination of infinitesimal calculus and analytic geometry applied to curves and surfaces. The pioneers of differential geometry are C. Huygens, A. C. Clairaut, L. Euler, A.-L. Cauchy, and G. Monge. In the twentieth century, curvature played a very important role in the development of modern physics, since, according to the general theory of relativity, objects of great mass bend space-time. Geometrizing the theory of gravity, we could say, following Einstein, that a heavy body modifies the geometry around it in such a way that the geodesics in the corresponding geometry are the curved trajectories of the attracted particles.

A very simple way in which one can present Einstein's general theory of relativity is the following metaphor: imagine a big rubber sheet stretched nice and taut before your eyes. If you watch a little marble as it rolls across the surface of this rubber sheet, then you will realize that it follows a simple straight-line trajectory. But if you watch the movement of a heavy rock on this rubber sheet, then you will realize that now the rubber sheet is deformed, warped, curved. In contrast to the previous marble, this rock does not follow a straight-line trajectory, but it follows a curved trajectory along the curved surface of the rubber sheet. Einstein took this idea and applied it to the fabric of space. Originally, the fabric of space may look nice and flat, like the rubber sheet in the previous example. However, if the Sun appears, the fabric of space curves. Similarly, in the vicinity of the Earth, the fabric of space curves, and the Moon is kept in orbit around the Earth because it rolls along a valley in the curved environment that is created by the Earth's mass. This is the manner in which, according to Einstein, gravity is communicated from place to place: through warps and curves in the fabric of the space, more specifically through warps and curves in space-time. For instance, the Earth is kept in orbit around the Sun because it rolls along a valley in the curved environment that is created by the Sun's mas, and, similarly, as I mentioned before, the Moon is kept in orbit around the Earth because it rolls along a valley in the curved environment that is created by the Earth's mass. For this reason, the general theory of relativity is necessarily founded on Riemannian geometry (in Riemannian geometry, we do not talk about "straight lines," but about "straightest lines," that is, "geodesics" (the shortest path between two points on a curved surface) or "great circles" (the shortest path between two points along a spherical surface in particular)).

It is worth mentioning that the general theory of relativity makes the following predictions: rays of light passing close to a star should be bent towards it, and physical processes should take place more slowly in regions of low gravitational potential than in regions of high gravitational potential (thus, kinetic energy changes throughout an orbit, resulting in a higher speed when a planet is closer to the Sun).

According to the “Bing-Bang” cosmological model, gravity underpinned and, actually, determined the transition from the “Bing-Bang” cosmological “soup” to the galactic structure that we observe today: gravity started from the initial conditions of the Big Bang and made the universe much more complex because, even though the density of the universe was almost uniform, there were density quantum-mechanical fluctuations. Put slightly differently, there were small differences in the density of the universe from one region to another. Thus, a region of the universe with density slightly greater than the mean density of the universe acted upon itself by its own gravity, and, gradually, it made itself denser. Consequently, instead of expanding with the rest of the universe, it drew matter into the given region. Ultimately, this region collapsed upon itself and did not participate in the universal expansion. In this way, a physical object was made out of such a region. Gradually, the universe was filled with small density inhomogeneities resulting from inflation due to quantum-mechanical fluctuations, which ultimately merged into the structures of the universe that we observe today.

The physical significance of differentiation (basic applications in mechanics): By the term “energy,” we mean the impetus that underpins all motion and all activity—more specifically, the capacity for doing work. In physics, we typically look at the work that a constant force, F, does when moving an object over a distance of s. In these cases, the work is
$W = Fs$;
the force is parallel to the displacement.

Mechanics is the branch of physics that studies the relationships between the following three physical concepts:

i. *Force:* an agent that changes or tends to change the state of motion (i.e., the state of rest or of uniform motion) of an object. The “velocity” of an object is the rate of change of its position with respect to a frame of reference, and it is a function of time (i.e., velocity is the first derivative of displacement with respect to time). Notice that the “relative velocity” of a moving body A with respect to a moving body B is denoted by $\vec{v}_{A,B}$ and is defined as the vector sum of the velocity $\vec{v}_A$ of A and the negative of the velocity $\vec{v}_B$ of B; symbolically: $\vec{v}_{A,B} = \vec{v}_A + (-\vec{v}_B)$, which is the vector equation of relative motion, whereas the corresponding algebraic equation is $v_{A,B} = v_A - v_B$.

ii. *Mass:* the quantity of matter that is concentrated in an object. The product of the mass times the velocity of an object is the "momentum" of that object.

iii. *Motion:* a change in the position of an object with respect to time.

The part of mechanics that is concerned with the study of motion is called kinematics. Due to the rigorous study of classical mechanics by Isaac Newton, the SI (Système International) unit of force, newton (denoted by N), has been named in his honor. One newton is defined as the force needed in order to accelerate one kilogram (kg) of mass at the rate of one meter (m) per second (sec) squared in the direction of the applied force: $1N = 1kg\frac{m}{sec^2}$.

Regarding the measurement of time and physical distance, it should be mentioned that the German mathematician Hermann Minkowski depicted time as a length by proposing the following definition:

$distance = speed\ of\ light \times time = ct.$

Hence, if the speed of light in vacuum, commonly denoted by the letter c, is approximately $300{,}000{,}000\ meters/second$ (according to Rosa and Dorsey), then we say that $1/300{,}000{,}000$ of a second is one meter. In other words, one meter is the distance travelled by light in vacuum during a time interval of $1/300{,}000{,}000$ of a second.

First Law of Motion: An object will remain at rest or in a uniform state of motion unless that state is changed by an external force.

Second Law of Motion: The vector sum of the forces on an object is equal to the mass of that object multiplied by the acceleration of that object ("acceleration" is the rate of change of the velocity v of an object with respect to time, meaning that acceleration is the first derivative of velocity with respect to time or, equivalently, the second derivative of displacement s with respect to time); symbolically:

$F = ma \equiv m\frac{dv}{dt} \equiv m\frac{d^2s}{dt^2},$

where F denotes force, m denotes the mass of an object, and a denotes the acceleration of the given object (thus, for any force you put on an object, an object of small mass will accelerate a lot, and an object of large mass will accelerate just a little). In case of circular motion (i.e., a movement of an object along the circumference $C = 2\pi r$ of a circle of radius r), if the period for one rotation is T, then:

the angular velocity (i.e., the angular rate of rotation) is

$\omega = \frac{2\pi}{T} = \frac{d\varphi(t)}{dt},$

where $\varphi(t)$ denotes the angular displacement from the x-axis and is measured in radians, and t denotes time (measured in seconds);

the speed of the object travelling the circle is
$v = \frac{2\pi r}{T} = \omega r$;
the angular acceleration of the particle is
$\alpha = \frac{d\omega}{dt}$,
and, in case of uniform circular motion, $\alpha = 0$;
the acceleration due to change in the direction is
$\alpha_c = \frac{v^2}{r} = \omega^2 r$;
and the centripetal and centrifugal force can be computed using acceleration as follows (the centripetal force and the centrifugal force are actually the same force, depending upon the frame of reference):
$F_c = m\alpha_c = \frac{mv^2}{r}$.
For instance, a "satellite" is any object that is orbiting the Earth (or any other massive body). Once launched into orbit, a satellite is a projectile acted upon by a single force, specifically, by the force of gravity. In particular, a satellite is a projectile that is launched horizontally at such a high speed that, due to gravity, it falls *towards* the Earth, but it never falls *into* the Earth (thus, making a circular path *around* the Earth) because the curvature of the satellite's path matches the curvature of the Earth (approximately, for every 8 km horizontally, the Earth curves downward 5m, and, therefore, 5m is the distance that a projectile falls in one second, so that, if we shoot a projectile that travels 8 km horizontally per second, it will fall towards the Earth but never touch the Earth; in other words, since a satellite moves at 8 km/sec, it "falls" at the same rate as the Earth "curves downward").
Third Law of Motion: For every action in nature, there is an equal and opposite reaction. The "internal forces" of a system of objects are those forces which are exerted between the members of the given system. The "external forces" of a system of objects are those forces exerted by bodies not belonging to the given system on the members of the given system. Internal forces are interaction forces (that is, pairs of action and reaction), and, therefore, their resultant is equal to zero. Hence, the total momentum of an isolated system of objects remains constant. For instance, the operation of rockets and jet planes is based on the conservation of momentum: as the fuel burns, it gives off hot gas that shoots out from an opening at the back of the chamber, so that the force of the gas moving backward pushes the rocket/the jet plane forward.
Newton's Law of Universal Gravitation: An object attracts another object with a force that is directly proportional to the product of the masses of the

objects and inversely proportional to the square of the distance between them, symbolically:

$F_g = G\frac{m_1 m_2}{r^2}$,

where F_g is the magnitude of the gravitational force on either object, m_1 and m_2 are their masses, r is the distance between them, and G is the gravitational constant, whose value is found to be (in SI units) $6.673 \times 10^{-11} N \cdot m^2 \cdot kg^{-2}$ (thus, the "weight" of a body is the total gravitational force exerted on the body by all other bodies in the universe).

Total Mechanical Energy of a System: $E_m = K + U$,

where E_m denotes mechanical energy, K denotes kinetic energy, and U denotes potential energy.

By the term "potential energy," we mean the energy possessed by a body by virtue of its position relative to others, stresses within itself, its electric charge, or other factors. For instance, gravitational potential energy (e.g., in the case of a ball whose mass is m and is dropped from height h) can be computed using the following formula:

$U = mgh$,

where m denotes the mass of the object, g denotes the acceleration constant due to the Earth's gravity ($\approx 9.8\, m/sec^2$), and h denotes the height (displacement) of the object as a function of time (gravitational acceleration g differs from planet to planet; for instance, at the surface of the Earth, gravitational acceleration is approx. $9.8\, m/sec^2$, whereas, at the surface of Mars, gravitational acceleration is approx. $3.7\, m/sec^2$).

By the term "kinetic energy," we mean the energy possessed by a body by virtue of its motion. Let us consider a body of mass m moving along the x-axis under the action of a constant resultant force of magnitude F directed along the axis. The body's acceleration is constant, and, according to Newton's Second Law of Motion, it is given by $F = ma$. The kinetic energy of this body can be computed using the following formula:

$K = \frac{1}{2} m v^2$,

where v denotes the body's velocity (which is, by definition, a function of time), and m denotes the mass of the object. Thus, the work done by the resultant external force on a body is equal to the change in kinetic energy of the body.

The eighteenth-century French mathematician and natural philosopher Émilie du Châtelet proposed and tested the law of "conservation of energy," according to which the total energy of an "isolated system" (i.e., one that does not interact with other systems) remains constant.

In order to clarify the meaning of the principle of the conservation of energy, let us consider the following example: setting fire to coal. The chemical bonds of the coal molecules store great amounts of energy. If we set fire to coal, then fire causes a chain reaction between the coal and oxygen in the air. In this reaction, energy from the chemical bonds is converted into kinetic energy of air molecules. Hence, the air becomes warm, and, for this reason, it will rise. This rising air can be used in order to drive a turbine and, for instance, move a vehicle, or in order to create electricity (by feeding it into the grid). Alternatively, we can just burn coal without doing anything with the produced energy. This does not change the total energy in the system, because the total energy in the system is conserved. The chemical energy of the coal is converted into kinetic energy of air molecules, which are distributed in the atmosphere. Even though, in this case, the energy is useless, the total energy in the system remains the same. The difference between the aforementioned cases is entropy, or the measure of the molecular disorder, or randomness, of the system under consideration. Initially, the energy was packed into the coal, and the level of entropy was low. By setting fire to coal, the energy was distributed in the motion of air molecules, and the level of entropy became high. When a system has energy in a state of low entropy, its energy can be used in order to create macroscopic change (e.g., drive a turbine), and this useful energy is called "free energy." Free energy is a type of energy that does "work." But, if the energy in the system is in a state of high entropy, then the energy is useless, and it is called "heat." Heat is a type of energy that does not do "work." Even though *total* energy is conserved, *free* energy is not conserved.

Escape velocity: By "escape velocity," we mean the minimum velocity that a moving body, such as a rocket, must have to escape from the gravitational field of a celestial body, such as the Earth, and move outward into space. Suppose that a mass m is launched from the surface of the Earth with a velocity equal to v, and it travels to a height h. According to the law of conservation of energy,

$$(K+U)_{lowest\ point} = (K+U)_{highest\ point}$$

where K denotes kinetic energy, and U denotes potential energy. The potential energy of an object of mass m on the surface of the Earth is $\frac{-GMm}{R}$, where M denotes the mass of the Earth, and R denotes the radius of the Earth (this potential energy is due to the gravitational pull of the Earth, and it is negative because the work is done by the gravitational force of attraction). The above formula of the law of conservation of energy implies that, for some height h,

$$\frac{1}{2}mv_{initial}^2 + \frac{-GMm}{R} = \frac{1}{2}mv_{final}^2 + \frac{-GMm}{R+h}$$

where, we realize that, as $h \to \infty$, $R + h \to \infty$, and, thus, potential energy tends to zero, and, since we do not need any excess speed at the end (as we are looking for the absolute minimum velocity), $v \to 0$ (as $h \to \infty$), and, thus, kinetic energy tends to zero, too. Under these assumptions, the right-hand part of the last equation vanishes, and we obtain

$$\frac{1}{2} m v_{initial}^2 = \frac{GMm}{R}$$

(this $v_{initial}$ is what we call "escape velocity"). Hence, solving for $v_{initial}$, we get the escape velocity:

$$v_{escape} = \sqrt{\frac{2GM}{R}}$$

(where M and R, respectively, denote the mass and the radius of the planet from which the projectile is launched (in this case, the Earth); and if one launches a projectile with the speed v_{escape}, or higher, then the projectile will fly away and will not return).

Differentiation of Multivariable Functions

So far, we have studied exclusively functions of a single (independent) variable x, but we can also apply the concept of differentiation to functions of several variables $x, y, \ldots$ Suppose that $f(x, y)$ is a function of two variables x and y, and that the limits

$$lim_{\Delta x \to 0} \frac{f(x + \Delta x, y) - f(x, y)}{\Delta x}$$

and

$$lim_{\Delta y \to 0} \frac{f(x, y + \Delta y) - f(x, y)}{\Delta y}$$

exist for all values of x and y in question—that is, $f(x, y)$ possesses a derivative $\frac{df}{dx}$ with respect to x and a derivative $\frac{df}{dy}$ with respect to y. These derivatives are called the "partial derivatives" of f, and they are respectively denoted by

$\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}$ or

$f_x'(x, y), f_y'(x, y)$.

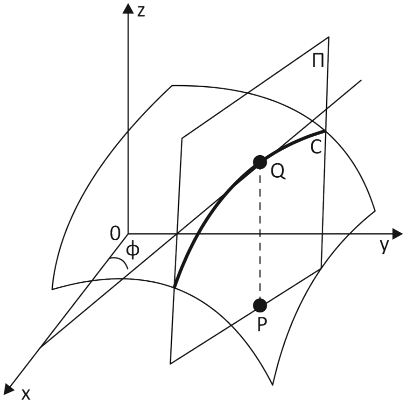
We use the following notation for second-order partial derivatives:

$\frac{\partial^2 f}{\partial x^2} \equiv f_{xx}''$ and $\frac{\partial^2 f}{\partial y^2} \equiv f_{yy}''$;

and, in case of second-order mixed derivatives, we write:

$\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) \equiv \frac{\partial^2 f}{\partial y \partial x} \equiv f''_{xy}$ and $\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right) \equiv \frac{\partial^2 f}{\partial x \partial y} \equiv f''_{yx}$.
Similarly, we can differentiate functions of three or more variables.
In general, when calculating partial derivatives, we treat all independent variables other than the variable with respect to which we differentiate as constants. For instance, if $f(x,y) = x^2 - 3xy - 5$, then
$\frac{\partial f}{\partial x} = \frac{\partial}{\partial x}(x^2 - 3xy - 5) = \frac{\partial}{\partial x}(x^2) - \frac{\partial}{\partial x}(3xy) - \frac{\partial}{\partial x}(5) = 2x - 3y$, and
$\frac{\partial f}{\partial y} = \frac{\partial}{\partial y}(x^2 - 3xy - 5) = \frac{\partial}{\partial y}(x^2) - \frac{\partial}{\partial y}(3xy) - \frac{\partial}{\partial y}(5) = -3x.$
The geometric significance of $\frac{\partial f}{\partial x}|_{(x_0,y_0)}$ and $\frac{\partial f}{\partial y}|_{(x_0,y_0)}$ is illustrated in Figure 8-12. Let us consider a function $z = f(x,y)$, whose graph in $\mathbb{R}^3$ is a surface. We suppose that $P(x_0, y_0)$ is an arbitrary point of the domain of f. Notice that, in $\mathbb{R}^3$, the equation $y = y_0$ represents a plane Π that is perpendicular to the y-axis. This plane intersects the surface $z = f(x,y)$ by a curve C whose equation is $z = f(x, y_0)$. If $Q(x_0, y_0, z_0)$ is a point belonging to C, so that its orthogonal projection to the plane xOy is the point P, then the slope of the tangent to the curve C at Q is equal to $\frac{\partial f}{\partial x}|_{(x_0,y_0)} = tan\varphi$, where φ is the angle formed by the x-axis and the tangent to the curve C at Q, as shown in Figure 8-12. Similarly, we can show that the slope of the tangent to the curve C at Q is equal to $\frac{\partial f}{\partial y}|_{(x_0,y_0)} = tan\theta$, where θ is the angle formed by the y-axis and the tangent to the curve C at Q.
Remark: If f, f'_x, f'_y, f''_{xy}, and f''_{yx} are all continuous at (x_0, y_0), then $f''_{xy}(x_0, y_0) = f''_{yx}(x_0, y_0)$.

Figure 8-12: The geometric significance of a partial derivative.



Generalization: If $f\colon \mathbb{R}^n \to \mathbb{R}$ is a function,
$\mathbb{R}^n \ni (x_1, x_2, \dots, x_n) \to f(x_1, x_2, \dots, x_n) \in \mathbb{R}$,
then

$$\frac{\partial f(x_1, x_2, \dots, x_i, \dots, x_n)}{\partial x_i} = lim_{\Delta x_i \to 0} \frac{f(x_1, x_2, \dots, x_i + \Delta x_i, \dots, x_n) - f(x_1, x_2, \dots, x_i, \dots, x_n)}{\Delta x_i}$$

is the partial derivative of $f(x_1, x_2, \dots, x_n)$ with respect to x_i, where $i = 1,2, \dots, n$ (the "round d," that is, the symbol ∂, was originally used in the 1770s by the French mathematician and philosopher Marquis de Condorcet and the Swiss mathematician Leonhard Euler for partial differentials; and this symbol was used for the first time in the modern combination $\partial f / \partial x$ in the 1780s by the French mathematician Adrien-Marie Legendre).

We take for granted the obvious generalizations of the theorems of differentiation to two or more variables.

Partial derivatives of composite functions (chain rule for multivariable functions): If a function $f(x, y)$ is defined on an open set A of $\mathbb{R}^2$, and if $x = x(r)$ and $y = y(r)$ with $r \in [a, b]$, then the derivative of the composite function f with respect to r is given by

$$\frac{\partial f}{\partial r} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial r} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial r}$$

provided that f has continuous partial derivatives over A and that $x = x(r)$ and $y = y(r)$ are differentiable over $[a, b]$. Since f is written as a function of the parameter r, we can also write $\frac{df}{dr}$, instead of $\frac{\partial f}{\partial r}$.

The above formula can be generalized for functions $f(x_1, x_2, \dots, x_i, \dots, x_n)$ of n variables with $x_i = x_i(r)$, $i = 1,2, \dots, n$, as follows:

$$\frac{\partial f}{\partial r} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial r} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial r} + \cdots + \frac{\partial f}{\partial x_n}\frac{\partial x_n}{\partial r}$$

under similar assumptions as previously.

Harmonic functions: A function $f(x, y)$ defined on a subset A of $\mathbb{R}^2$ is said to be a "harmonic function" if and only if it satisfies the following equation:

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = 0. \tag{1}$$

Equation (1) is called the "Laplace equation," and the operator $\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ is called the "Laplace operator." For instance, on their respective domains in $\mathbb{R}^2$, the functions $f(x, y) = x + y$, $f(x, y) = x^2 - y^2$, and $f(x, y) = ln(x^2 + y^2)$ are harmonic (the domain of $f(x, y) = ln(x^2 + y^2)$ may be any open subset of $\mathbb{R}^2$ that does not include 0). An interesting property of a harmonic function is that its value at a point is always equal to the average of its values over a ball centered at that point (i.e., harmonic functions are always equal to the average of their nearby values). Thus, harmonic functions are used in simplifying complex processes, since a first harmonic response gives us an indication of the linear approximation.

Homogeneous functions: A function $f(x, y)$ defined on a subset A of $\mathbb{R}^2$ is said to be a "homogeneous function of degree k" if it holds that

$$f(\lambda x, \lambda y) = \lambda^k f(x, y)$$

for all $(x, y) \in A$, where $\lambda > 0$, and k is a real number (i.e., a homogeneous function is scale-invariant, meaning that, if every variable is replaced with a scaled version of itself, this scale being the same for each variable, then the whole function is scaled by some power of that original scale). For instance, the function $f(x, y) = x^2 + y^2$ is a homogeneous function of degree 2, since $f(\lambda x, \lambda y) = (\lambda x)^2 + (\lambda y)^2 = \lambda^2(x^2 + y^2) = \lambda^2 f(x, y)$. Moreover, notice that a function $f(x, y)$ that can be expressed in the form of

$x^k g\left(\frac{y}{x}\right)$ or $y^k g\left(\frac{x}{y}\right)$

is a homogeneous function of degree k.

Euler's theorem for homogeneous functions: If $u(x, y)$ is a homogeneous function of degree k, then

$x\frac{\partial u}{\partial x}+y\frac{\partial u}{\partial y}=ku(x,y).$

The proof of this theorem follows directly from the definition of a homogeneous function: Since $u(x,y)$ is a homogeneous function of degree k, it can be expressed in the form of $u(x,y)=x^k f\left(\frac{y}{x}\right)$. Then (applying the product rule and the chain rule) we obtain $\frac{\partial u}{\partial x}=kx^{k-1}f\left(\frac{y}{x}\right)+x^k f'\left(\frac{y}{x}\right)\cdot y\cdot\left(-\frac{1}{x^2}\right)\Rightarrow\frac{\partial u}{\partial x}=kx^{k-1}f\left(\frac{y}{x}\right)-x^{k-2}yf'\left(\frac{y}{x}\right)$. Similarly (treating x^k as a constant), we obtain $\frac{\partial u}{\partial y}=x^k f'\left(\frac{y}{x}\right)\cdot\frac{1}{x}\Rightarrow\frac{\partial u}{\partial y}=x^{k-1}f'\left(\frac{y}{x}\right)$. Therefore, $x\frac{\partial u}{\partial x}+y\frac{\partial u}{\partial y}=x\left[kx^{k-1}f\left(\frac{y}{x}\right)-x^{k-2}yf'\left(\frac{y}{x}\right)\right]+y\left[x^{k-1}f'\left(\frac{y}{x}\right)\right]=kx^k f\left(\frac{y}{x}\right)=ku(x,y)$, *quod erat demonstrandum.*

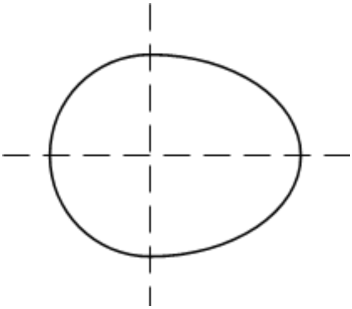
Partial derivatives of implicit functions: Variables x, y, and z are said to be “related implicitly” if they depend on each other by an equation of the form $F(x,y,z)=0$, where F is some function. For instance, the points on a sphere centered at the origin with radius 2 are related implicitly by the equation $x^2+y^2+z^2-2^2=0$. In such situations, we can compute the partial derivatives of one of the variables with respect to the other variables by using the method of implicit differentiation, in the context of which we treat the variables as independent in order to find the partial derivatives of the function F, while simultaneously keeping in mind the fact that the variables depend on each other due to the equation $F(x,y,z)=0$.

Suppoze that variables x, y, and z are related by the equation $F(x,y,z)=0$ and that we want to compute $\frac{\partial z}{\partial y}$. In order to do this, we have to think that the equation $F(x,y,z)=0$ determines z as a differentiable function of the independent (yet implicitly related to z) variables x and y, $\frac{\partial F}{\partial z}\neq 0$. In fact, the method of implicit differentiation is underpinned by the chain rule, in the sense that the original independent variables are x, y, and z, but we can reconsider them from a different persepective according to which the independent variables are x and y, and we treat z as a function of x and y.

Let y be implicitly related to x by the equation $F(x,y)=0$, and suppose that the locus of $F(x,y)=0$ is a closed curve, as shown, for instance, in Figure 8-13. In this situation, we observe the following: On the one hand, y is not a “function” of x, since to a particular value of x there correspond several values of y (applying the vertical line test, mentioned in Chapter 2, we realize that, in Figure 8-13, a line perpendicular to the x-axis intersects the given locus at more than one point, whereas, by definition, a function is single-valued). On the other hand, even though the equation $F(x,y)=0$

does not define y as a function of x, there are certain segments of this locus (e.g., the segment of the oval in Figure 8-13 that lies above the horizontal axis) that can be considered in isolation (i.e., separately from the totality of the given oval) in such a way that they can be considered to constitute a function of x, namely, $y = f(x)$. In other words, we can separate out a segment of a closed curve C that can successfully pass the vertical line test, thus defining a function $y = f(x)$. In fact, this is the reason why we say that the equation $F(x, y) = 0$ defines y *implicitly* as a function of x, so that y is an *implicit* function of x.

Figure 8-13: A closed curve (source: Wikimedia Commons: Author: Herbee; https://commons.wikimedia.org/wiki/File:Oval1.PNG).



In view of the foregoing, if we consider the locus of $F(x, y) = 0$ in Figure 8-13, we can meaningfully try to compute the derivative of the function $y = f(x)$ with respect to x at a particular point of the given locus (provided that we are working on a segment of the given locus that defines a function $y = f(x)$), and, in fact, this derivative is:

$$\frac{dy}{dx} = -\frac{\frac{\partial F(x,y)}{\partial x}}{\frac{\partial F(x,y)}{\partial y}}$$

where $\frac{\partial F(x,y)}{\partial y} \neq 0$. Thus, we have come up with the following theorem of implicit differentiation: If $F(x, y) = 0$, if x and y are restricted to those values which satisfy the equation $F(x, y) = 0$, and if $\frac{\partial F(x,y)}{\partial y} \neq 0$, then $\frac{dy}{dx} = -\frac{\frac{\partial F(x,y)}{\partial x}}{\frac{\partial F(x,y)}{\partial y}}$.

Proof: If we set $W = F(x, y)$, then the total differential of the function W is

$$dW = \frac{\partial W}{\partial x}dx + \frac{\partial W}{\partial y}dy. \qquad (1)$$

If x and y are restricted to those values which satisfy the equation $F(x, y) = 0$, then $W = F(x, y) = 0$, and $dW = 0$. Hence, if we set $dW = 0$ in equation (1) and solve for $\frac{dy}{dx}$, we obtain

$\frac{dy}{dx} = -\frac{\frac{\partial F(x,y)}{\partial x}}{\frac{\partial F(x,y)}{\partial y}}, \frac{\partial F(x,y)}{\partial y} \neq 0$, *quod erat demonstrandum.*

The aforementioned theorem of implicit differentiation can be formally stated as follows: Given that an equation of the form $F(x, y) = 0$ defines an implicit function $y = f(x)$ if and only if $F[x, f(x)] = 0$ for all $x \in \mathbb{R}$, let $F(x, y) = 0$ be defined on $\mathbb{R}^2$, and $(x_0, y_0) \in \mathbb{R}^2$. Suppose that F_x' and F_y' are continuous, $F(x_0, y_0) = 0$, and $F_y'(x_0, y_0) \neq 0$. Then there exists a neighborhood of x_0, say $U(x_0)$, wherein the equation $F(x, y) = 0$ defines a function $y = f(x)$ in a unique way, so that F_x' is continuous, $y_0 = f(x_0)$, and

$$\frac{dy}{dx} = -\frac{F_x'}{F_y'}$$

(this theorem delineates the method of implicit differentiation for multivariable functions).

Example: Let (2,1) be a point, and let $F(x, y) = x^2 - 2xy = 0$ be an equation. Then we can show that the given equation defines a function $y = f(x)$ on a neighborhood $x_0 = 2$ and find F_x' as follows: In order to show that an equation of the form $F(x, y) = 0$ defines a function $y = f(x)$ on a neighborhood of x_0, it is enough to show that:

i. $F_x' = 2x - 2y$ and $F_y' = -2x$ are continuous, which they are.

ii. $F(x_0, y_0) = 0 \Rightarrow F(2,1) = 0$, which is true.

iii. $F_y'(x_0, y_0) \neq 0 \Rightarrow F_y'(2,1) \neq 0$, which is true.

Therefore, $x^2 - 2xy = 0$ defines a function $y = f(x)$ on a neighborhood $U(x_0)$. Moreover, $\frac{dy}{dx} = -\frac{F_x'}{F_y'} = -\frac{y-x}{x}$.

We can work in $\mathbb{R}^3$ in a similar way: A function of the form $F(x, y, z)$ such that $z = z(x, y)$, that is, $F[x, y, z(x, y)]$, is called "implicit," while $z = z(x, y)$ is an "explicit" function. An equation of the form $F(x, y, z) = 0$ defines an implicit function $z = z(x, y)$ if and only if $F[x, y, z(x, y)] = 0$ for all $(x, y) \in \mathbb{R}^2$. Let $F(x, y, z) = 0$ be defined on $\mathbb{R}^3$, and $(x_0, y_0, z_0) \in \mathbb{R}^3$. We assume that F_x', F_y', and F_z' are continuous, $F(x_0, y_0, z_0) = 0$, and $F_z'(x_0, y_0, z_0) \neq 0$. Then there exists a neighborhood of (x_0, y_0), say $U(x_0, y_0)$, wherein the equation $F(x, y, z) =$

0 defines a function $z = z(x, y)$ in a unique way, so that F'_x and F'_y are continuous, $z_0 = z(x_0, y_0)$, and

$$\frac{\partial z}{\partial x} = -\frac{F'_x}{F'_z}$$

and

$$\frac{\partial z}{\partial y} = -\frac{F'_y}{F'_z}$$

(the proof is essentially the same as the above proof for the case $F(x, y) = 0$). Of course, the rule of implicit differentiation does not hold for the points of the surface defined by $F(x, y, z) = 0$ at which $F'_x = F'_y = F'_z = 0$, and such points are called "singular points" (that is, the singular points are those points at which all the partial derivatives simultaneously vanish; and, thus, at a singular point, we cannot solve for any variable in terms of the others, and an algebraic variety looks strange near such a point, or it may not look at all like the graph of a function). The total differential of the implicit function $z = z(x, y)$ defined by the equation $F(x, y, z) = 0$ is given by

$\frac{\partial F}{\partial x}dx + \frac{\partial F}{\partial y}dy + \frac{\partial F}{\partial z}dz = 0.$

The Jacobian determinant: The Jacobian determinant of n functions, $f_1, f_2, \dots, f_n$, in n real variables, $x_1, x_2, \dots, x_n$, with respect to $x_1, x_2, \dots, x_n$ is defined by

$$J = \frac{\partial(f_1, f_2, \dots, f_n)}{\partial(x_1, x_2, \dots, x_n)} = \begin{vmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{vmatrix}$$

(and it is named after the nineteenth-century German mathematician Carl Gustav Jacob Jacobi). Notice that, if $x_1 = x_1(r_1, r_2, \dots, r_n), \dots, x_n = x_n(r_1, r_2, \dots, r_n)$, then

$$\frac{\partial(f_1, f_2, \dots, f_n)}{\partial(r_1, r_2, \dots, r_n)} = \frac{\partial(f_1, f_2, \dots, f_n)}{\partial(x_1, x_2, \dots, x_n)} \cdot \frac{\partial(x_1, x_2, \dots, x_n)}{\partial(r_1, r_2, \dots, r_n)}$$

which is the Jacobian determinant of n functions whose variables $x_1, x_2, \dots, x_n$ are functions of n variables $r_1, r_2, \dots, r_n$.

Jacobian determinants are useful in order to compute the partial derivatives of implicit functions that are defined by a system of equations.

Case I: Let $\begin{Bmatrix} f(x, y, z) = 0 \\ g(x, y, z) = 0 \end{Bmatrix}$ be a system such that:

the functions f and g have continuous first-order partial derivatives, $f(x_0,y_0,z_0)=0=g(x_0,y_0,z_0)$, and the Jacobian determinant $\frac{\partial(f,g)}{\partial(y,z)}|_{(x_0,y_0,z_0)}\neq 0$. Every system of this form has a unique solution $y=y(x)$ and $z=z(x)$, where $y(x)$ and $z(x)$ are two functions whose derivatives (with respect to x) are continuous on a neighborhood of x_0, so that $y_0=y(x_0)$ and $z_0=z(x_0)$. Then:

$$\frac{dy}{dx}=-\frac{\frac{\partial(f,g)}{\partial(x,z)}}{\frac{\partial(f,g)}{\partial(y,z)}}=-\frac{\begin{vmatrix} f_x' & f_z' \\ g_x' & g_z' \end{vmatrix}}{\begin{vmatrix} f_y' & f_z' \\ g_y' & g_z' \end{vmatrix}}$$

and

$$\frac{dz}{dx}=-\frac{\frac{\partial(f,g)}{\partial(y,x)}}{\frac{\partial(f,g)}{\partial(y,z)}}=-\frac{\begin{vmatrix} f_y' & f_x' \\ g_y' & g_x' \end{vmatrix}}{\begin{vmatrix} f_y' & f_z' \\ g_y' & g_z' \end{vmatrix}}$$

(this result gives us the partial derivatives of implicit functions defined by a system of the form $\{f(x,y,z)=0,g(x,y,z)=0\}$).

Case II: Let $\begin{Bmatrix} f(x,y,z,t)=0 \\ g(x,y,z,t)=0 \end{Bmatrix}$ be a system such that:

the functions f and g have continuous first-order partial derivatives, $f(x_0,y_0,z_0,t_0)=0=g(x_0,y_0,z_0,t_0)$, and the Jacobian determinant $\frac{\partial(f,g)}{\partial(z,t)}|_{(x_0,y_0,z_0,t_0)}\neq 0$. Every system of this form has a unique solution $z=z(x,y)$ and $t=t(x,y)$, where $z(x,y)$ and $t(x,y)$ are two functions whose partial derivatives are continuous on a neighborhood of (x_0,y_0), so that $z_0=z(x_0,y_0)$ and $t_0=t(x_0,y_0)$. Then the derivatives of these functions are given by the following formulae:

$$\frac{\partial z}{\partial x}=-\frac{\frac{\partial(f,g)}{\partial(x,t)}}{\frac{\partial(f,g)}{\partial(z,t)}}$$

$$\frac{\partial z}{\partial y}=-\frac{\frac{\partial(f,g)}{\partial(y,t)}}{\frac{\partial(f,g)}{\partial(z,t)}}$$

$$\frac{\partial t}{\partial x}=-\frac{\frac{\partial(f,g)}{\partial(z,x)}}{\frac{\partial(f,g)}{\partial(z,t)}}$$

and

$$\frac{\partial t}{\partial y}=-\frac{\frac{\partial(f,g)}{\partial(z,y)}}{\frac{\partial(f,g)}{\partial(z,t)}}$$

(this result result gives us the partial derivatives of implicit functions defined by a system of the form $\{f(x,y,z,t)=0, g(x,y,z,t)=0\}$).

Functional dependence and functional independence: Consider m functions in n variables (each), namely,

$f_1(x_1,x_2,\ldots,x_n), f_2(x_1,x_2,\ldots,x_n),\ldots,f_m(x_1,x_2,\ldots,x_n)$.

Suppose that the domain of each of these functions is a subset A of $\mathbb{R}^n$ (i.e., they have the same domain). These functions are said to be "functionally dependent" in the set A if, for every n-tuple $(x_1,x_2,\ldots,x_n)\in A$, they satisfy at least one equation of the form $F(f_1,f_2,\ldots,f_m)=0$; otherwise, they are said to be "functionally independent."

For instance, the functions $f_1(x,y,z)=y-xz$, $f_2(x,y,z)=yz-x$, and $f_3(x,y,z)=(x-y)(z+1)$, which are defined on $\mathbb{R}^3$, are functionally dependent in $\mathbb{R}^3$, because $f_1+f_2+f_3=y-xz+yz-x+(x-y)(z+1)=0$.

In particular, n functions in n variables (each), namely,

$f_1(x_1,x_2,\ldots,x_n), f_2(x_1,x_2,\ldots,x_n),\ldots,f_n(x_1,x_2,\ldots,x_n)$

are functionally dependent in $\mathbb{R}^n$ if and only if the Jacobian determinant

$$J=\frac{\partial(f_1,f_2,\ldots,f_n)}{\partial(x_1,x_2,\ldots,x_n)}=0$$

(otherwise, i.e., when $J\neq 0$, they are functionally independent).

Mean Value Theorem: If a function $f\colon A\to\mathbb{R}$, where $A\subseteq\mathbb{R}^2$, is differentiable at the points of the straight line segment $\overline{ab}$, where $a=(a_1,a_2)$ and $b=(b_1,b_2)$, then there exists a number θ with $0<\theta<1$ such that

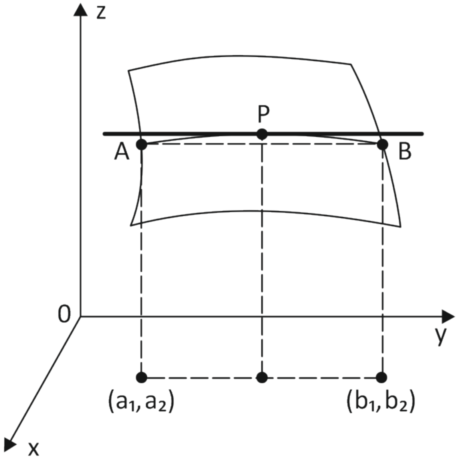
$$f'\big(a+\theta(b-a)\big)=\frac{f(b)-f(a)}{b-a}$$

where f' is the partial derivative of the function $f(x,y)$ with respect to x.

Geometric interpretation of the mean value theorem for a function in two variables: In Figure 8-14, we realize that, if $A\big(a_1,a_2,f(a_1,a_2)\big)$ and $B\big(b_1,b_2,f(b_1,b_2)\big)$ are two points of the surface $z=f(x,y)$ that correspond to the points a and b, then there exists a point P on the curve

$$\begin{Bmatrix} x = a_1 + t(b_1 - a_1) \\ y = a_2 + t(b_2 - a_2) \\ z = f\big(a + t(b - a)\big) \end{Bmatrix}, \text{ where } t \in [0,1],$$
of the surface $z = f(x, y)$, such that the tangent at P is parallel to the chord AB.

Figure 8-14: The geometric significance of the mean value theorem for a function in two variables.



Remark: An equivalent formula for the mean value theorem for a function $f(x, y)$ and the points (x_0, y_0) and $(x_0 + h, y_0 + \lambda)$ is the following:

$$f(x_0 + h, y_0 + \lambda) - f(x_0, y_0) \\ = h\frac{\partial f(x_0 + \theta h, y_0 + \theta\lambda)}{\partial x} + \lambda\frac{\partial f(x_0 + \theta h, y_0 + \theta\lambda)}{\partial y}$$

where $0 < \theta < 1$.
Recall that a set A is called convex if, for every $a, b \in A$, $ka + (1 - k)b \in A$. It is easily seen that, if a function $f(x, y)$ is differentiable on a convex subset A of $\mathbb{R}^2$, and if, for every $(x, y) \in \mathbb{R}^2$, it holds that $\frac{\partial f(x,y)}{\partial x} = 0 = \frac{\partial f(x,y)}{\partial y}$, then $f(x, y)$ is a constant function.

Extreme values: If $z = f(x, y)$ is a continuous function defined on a set R, then the extreme values of f may occur only at:

i. the boundary points of R in case the set R is compact (i.e., closed and bounded);

ii. the interior points of R where $f'_x = f'_y = 0$ ("critical points");

iii. the points where f'_x or f'_y fail to exist ("critical points").

Hence, if there are no boundary points (e.g., when $z = f(x, y)$ is defined on an open disc or on a quadrant minus the axes, or on the entire plane, etc.), the function may have no extrema on R, but, if it does, they must occur at its critical points in R.

Maximum-Minimum Tests: If f has continuos first-order and second-order partial derivatives on some open disc containing the point (a, b), and if $f'_x(a, b) = f'_y(a, b) = 0$, then:

- if $f''_{xx} < 0$ and $f''_{xx}f''_{yy} - \left(f''_{xy}\right)^2 > 0$ at the point (a, b), then the point (a, b) is a "local maximum";
- if $f''_{xx} > 0$ and $f''_{xx}f''_{yy} - \left(f''_{xy}\right)^2 > 0$ at the point (a, b), then the point (a, b) is a "local minimum";
- if $f''_{xx}f''_{yy} - \left(f''_{xy}\right)^2 < 0$ at the point (a, b), then the point (a, b) is a "saddle point" (a "saddle point," also known as a "minimax point," is a point on the surface of the graph of a function where the slopes, i.e., the derivatives, in orthogonal directions are all zero (and, thus, it is a "critical point"), but it is not a local extremum of the function: a saddle point is a stable point where the function has a local maximum in one direction, but a local minimum in another direction, as shown, for instance, in Figure 8-15);
- if $f''_{xx}f''_{yy} - \left(f''_{xy}\right)^2 = 0$ at the point (a, b), then the test is inconclusive.

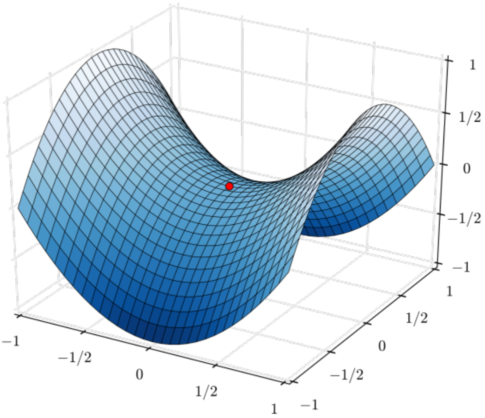
Note: The expression $f''_{xx}f''_{yy} - \left(f''_{xy}\right)^2$ is called the (Hessian) "discriminant" of f, and it is sometimes easier to remember in the determinant form

$$f''_{xx}f''_{yy} - \left(f''_{xy}\right)^2 = \begin{vmatrix} f''_{xx} & f''_{xy} \\ f''_{yx} & f''_{yy} \end{vmatrix}$$

(this method was developed in the nineteenth century by the German mathematician Ludwig Otto Hesse).

Because f''_{xy} is continuous, the order of mixed partial derivatives f''_{xy} and f''_{yx} does not matter (notice that, if the function is continuous, then it is continuous in every variable, and, therefore, if we measure the mixed rates of change of x and y, the result should be the same as measuring firstly y and then x, since everything is happening "smoothly" with respect to both variables; however, if the function is not continuously second differentiable, then, at the points where the derivative is discontinuous, the partial derivatives may not commute).

Figure 8-15: Saddle point with the coordinates $z = x^2 - y^2$ *(source: Wikimedia Commons: Author: Nicoguaro; https://commons.wikimedia.org/wiki/File:Saddle_point.svg).*



If the local extrema refer to the entire domain of the function under consideration, then they are "global extrema," and, as we know, a continuous function takes on an absolute maximum value and an absolute minimum value on any closed and bounded set on which it is defined.

Example: In order to find the extreme values of the function $f(x, y) = xy$, we work as follows: Since the function is differentiable everywhere and its domain has no boundary points, the function can assume extreme values only where

$f_x' = y = 0$ and $f_y' = x = 0$.

Hence, the origin is the only point where f might have an extreme value. In order to examine what happens there, we calculate

$f_{xx}'' = 0$, $f_{yy}'' = 0$, and $f_{xy}'' = 1$.

Then the discriminant of f is

$f_{xx}'' f_{yy}'' - \left(f_{xy}''\right)^2 = -1$,

and, since it is negative, we realize that $f(x, y) = xy$ has a saddle point at $(0,0)$, meaning that this function assumes no extreme values at all. Notice that, if we restrict the domain of $f(x, y) = xy$ to the closed disc $x^2 + y^2 \leq 1$, then the maximum value of f is $+\frac{1}{2}$, and the minimum value of f

is $-\frac{1}{2}$ (as shown by changing to polar coordinates: $xy = r^2 sin\theta cos\theta = \frac{1}{2}r^2 sin2\theta$).

In order to calculate the local extrema of a function $f(x, y, z)$ that is defined on an open subset A of $\mathbb{R}^3$ and has continuous first-order and second-order partial derivatives, we work as follows: Firstly, we solve the system of equations

$\{f_x' = 0, f_y' = 0, f_z' = 0\}$,

since the solutions to this system represent the possible locations of the local extrema of the given function. Then we compute the following expressions:

f_{xx}'',

$D_2 = \begin{vmatrix} f_{xx}'' & f_{xy}'' \\ f_{yx}'' & f_{yy}'' \end{vmatrix}$, and

$D_1 = \begin{vmatrix} f_{xx}'' & f_{xy}'' & f_{xz}'' \\ f_{yx}'' & f_{yy}'' & f_{yz}'' \\ f_{zx}'' & f_{zy}'' & f_{zz}'' \end{vmatrix}$.

We find the value of each of these three expressions at the critical points of $f(x, y, z)$, that is, at the solutions to the system $\{f_x' = 0, f_y' = 0, f_z' = 0\}$. If (x_0, y_0, z_0) is a critical point of f, then:

- if $f_{xx}''(x_0, y_0, z_0) > 0$, $D_2(x_0, y_0, z_0) > 0$, and $D_1(x_0, y_0, z_0) > 0$, then (x_0, y_0, z_0) is a "local minimum";
- if $f_{xx}''(x_0, y_0, z_0) < 0$, $D_2(x_0, y_0, z_0) > 0$, and $D_1(x_0, y_0, z_0) < 0$, then (x_0, y_0, z_0) is a "local maximum."

Integral Calculus in $\mathbb{R}$

In infinitesimal calculus, we start with two general questions about functions. Firstly, how steep is a function at a point? Secondly, what is the area underneath a graph over some region? The first question is answered using a tool called the "derivative." In other words, the derivative measures the rate of change of a function at a point. The second question is answered using a tool called the "integral."

Indefinite Integrals in $\mathbb{R}$

Let $f: I \to \mathbb{R}$ be a function, where I is an interval; in fact, I may have one of the following forms:

$[a, b], [a, b), (a, b], (a, b), [a, +\infty), (a, +\infty), (-\infty, b], (-\infty, b), (-\infty, +\infty)$

where $a, b \in \mathbb{R}$. When the interval I is closed, for instance, $[a, b]$, the expression $F'(x) = f(x)\ \forall x \in I$ implies that the following functions (derivatives) exist: $F'(x)\ \forall x \in (a, b)$, $F'_+(a)$, and $F'_-(b)$. Then the primitive function F is called the "antiderivative" of f in I, and it is denoted by

$F(x) = \int f(x)dx$, where $x \in I$,

according to Leibniz's notation. Hence, $\int f(x)dx = F(x) + c$ if and only if $[F(x) + c]' = f(x)$, where c is an arbitrary constant. Notice that the definition of the primitive function F of a function f includes an arbitrary constant c, and, therefore, the expression $\int f(x)dx$ is not uniquely determined. If F is a primitive function of a function f in an interval I, then the function $F + c, \forall c \in \mathbb{R}$, is called the "indefinite integral" of f in I. The aforementioned definition implies that the "indefinite integral" of a given function with respect to x is a new function plus a constant (known as the "constant of integration") if and only if the derivative of the new function and of the constant equals the given function, and it is based on a principle of differential calculus, where the assumption that I is an interval is substantial. Thus, differentiation can be used in order to verify the result of an indefinite integral: given that integration is the reverse process of differentiation, if the indefinite integral of a function $f(x)$ is $F(x)$, then differentiating $F(x)$ gives $f(x)$ back.

Integrals of elementary functions:

i. $\int adx = ax + c$, where a is an arbitrary constant.

ii. $\int x^n dx = \frac{x^{n+1}}{n+1} + c$ over the following intervals: (i) $n \neq -1, x > 0$; (ii) $n \neq -1, x < 0$; and (iii) $n \geq 0, x \in \mathbb{R}$. For instance, $\int \sqrt{x}dx = \int x^{1/2}dx = \frac{x^{3/2}}{\frac{3}{2}} + c = \frac{2}{3}x^{3/2} + c$, and $\int xdx = \frac{x^2}{2} + c$.

iii. $\int \frac{dx}{x} = ln|x| + c$.

iv. $\int a^x dx = \frac{a^x}{lna} + c$.

v. $\int sinxdx = -cosx + c$.

vi. $\int cosxdx = sinx + c$.

vii. $\int \frac{dx}{cos^2x} = tanx + c$.

viii. $\int \frac{dx}{sin^2x} = -cotx + c$.

ix. $\int \frac{dx}{\sqrt{1-x^2}} = arcsinx + c = \frac{\pi}{2} - arccosx + c$.

x. $\int \frac{dx}{1+x^2} = arctanx + c$.

xi. $\int sinhxdx = coshx + c$.

xii. $\int coshxdx = sinhx + c.$

Let $f: I \to \mathbb{R}$ and $g: I \to \mathbb{R}$ be two functions. If their indefinite integrals exist over I, then there exists the indefinite integral of $af + bg$, where a and b are constants, and
$\int[af(x) + bg(x)]dx = a\int f(x)dx + b\int g(x)dx.$
Proof: Given the definition of the indefinite integral, if $F(x) = \int f(x)dx$ and $G(x) = \int g(x)dx$, then $F'(x) = f(x)$ and $G'(x) = g(x)$ for all $x \in I$, and, therefore,
$\frac{d}{dx}[aF(x) + bG(x)] = aF'(x) + bG'(x) = af(x) + bg(x)$, *quod erat demonstrandum*.
Corollary: $\int \sum_{i=1}^{n} c_i f_i(x)dx = \sum_{i=1}^{n} c_i \int f_i(x)dx.$

Examples:

i. $\int \frac{x}{2x+1}dx = \frac{1}{2}\int \frac{2x}{2x+1}dx = \frac{1}{2}\int \frac{2x+1-1}{2x+1}dx = \frac{1}{2}\int \frac{2x+1}{2x+1}dx - \frac{1}{2}\int \frac{dx}{2x+1} = \frac{1}{2}\int dx - \frac{1}{4}\int \frac{d(2x+1)}{2x+1} = \frac{1}{2}x - \frac{1}{4}ln|2x+1| + c$, where $x > -\frac{1}{2}$ or $x < -\frac{1}{2}$.

ii. $\int \frac{dx}{sin^2x \cdot cos^2x} = \int \frac{sin^2x + cos^2x}{sin^2x \cdot cos^2x}dx = \int \frac{dx}{cos^2x} + \int \frac{dx}{sin^2x} = tanx - cotx + c$, where $k\pi < x < k\pi + \frac{\pi}{2}$ or $k\pi + \frac{\pi}{2} < x < k\pi + \pi$.

iii. $\int \frac{dx}{sinx \cdot cosx} = \int \frac{\frac{dx}{cos^2x}}{tanx} = \int \frac{dtanx}{tanx} = ln|tanx| + c$, where $x \in \mathbb{R} - \left\{k\frac{\pi}{2} | k \in \mathbb{Z}\right\}$.

iv. $\int \frac{dx}{sinx} = \int \frac{dx}{2sin\frac{x}{2}cos\frac{x}{2}} = \int \frac{d\frac{x}{2}}{sin\frac{x}{2}cos\frac{x}{2}} = ln\left|tan\frac{x}{2}\right| + c$, where $k\pi < x < k\pi + \pi$.

v. $\int cos^2xdx = \int \frac{1}{2}(1 + cos2x)\,dx = \frac{1}{2}\int(1 + cos2x)dx = \frac{1}{2}\left(x + \frac{sin2x}{2}\right) + c = \frac{1}{2}x + \frac{1}{4}sin2x + c$, and, because $sin2x = 2sinxcosx$, we can write $\int cos^2xdx = \frac{1}{2}x + \frac{1}{2}sinxcosx + c$.

Integration by substitution: The method of integration by substitution (or change of variable) is based on the following theorem: Let A and B be two real intervals, and let $f: A \to \mathbb{R}$ be a continuous function. If $g: B \to \mathbb{R}$ is a differentiable function such that $g'(t) \neq 0$ for all $t \in B$ and the range of g is a subset of A, then

$$\int f(x)\,dx = \int f\big(g(t)\big)\,g'(t)dt$$

(where, ultimately, after the computation of the last integral, we shall return to the original variable via the substitution $t = g^{-1}(x)$). The proof of this theorem follows directly from the definition of the indefinite integral by applying the chain rule for the differentiation of composite functions: Let $F(x) = \int f(x)\,dx$, so that $F'(x) = f(x)\ \forall x \in A$. By setting $G(t) = F\big(g(t)\big)$ and applying the chain rule for the differentiation of composite functions, we obtain $G'(t) = F'\big(g(t)\big)g'(t) = f\big(g(t)\big)g'(t)$, *quod erat demonstrandum*.

For instance, given the integral $\int \sqrt{1-x^2}\,dx$, $-1 < x < 1$, we set $x = sint$, $-\frac{\pi}{2} < t < \frac{\pi}{2}$, so that the range of $x = sint$ is the interval $-1 < x < 1$, and $x' = cost > 0\ \forall t \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. Hence,
$\int \sqrt{1-x^2}\,dx = \int \sqrt{1-sin^2t} \cdot (sint)'dt = \int cost \cdot costdt =$
$\int cos^2tdt = \frac{1}{2}sint \cdot cost + \frac{1}{2}t = \frac{1}{2}x\sqrt{1-x^2} + \frac{1}{2}arcsinx + c$, $-1 < x < 1$.

Notice that, sometimes, we may need to make the linear substitution $t = ax + b$, where $a, b \in \mathbb{R}$ and $a \neq 0$, so that $\int f(ax+b)\,dx = \frac{1}{a}\int f(t)dt$. For instance, for $x \in \mathbb{R}$, and setting $t = ax + b$, $\int sin(ax+b)dx = \int \frac{1}{a}sintdt = -\frac{1}{a}cost + c = -\frac{1}{a}cos(ax+b) + c$.

Furthermore, notice that, if the integrand is the quotient of two functions such that the numerator is the derivative of the denominator, then the indefinite integral is equal to the logarithm of the denominator. In other words, given the indefinite integral $\int \frac{f'(x)}{f(x)}dx$, where$f(x) \neq 0$, we set $f(x) = t$ to obtain $\int \frac{f'(x)}{f(x)}dx = \int \frac{dt}{t} = ln|t| + c = ln|f(x)| + c$ in the intervals where $f(x) \neq 0$. For instance: (i) $\int tanxdx = \int \frac{sinx}{cosx}dx = \int \frac{(-cosx)'}{cosx}dx = -ln|cosx| + c = ln|(cosx)^{-1}| + c = ln|secx| + c$; (ii) similarly, $\int cotxdx = \int \frac{cosx}{sinx}dx = \int \frac{(sinx)'}{sinx}dx = ln|sinx| + c$; and (iii) $\int \frac{dx}{xln|x|} = \int \frac{(ln|x|)'}{ln|x|}dx = ln\big|ln|x|\big| + c$.

In general, the choice of a suitable substitution depends on the integral that we have to compute. However, we can highlight the following cases:

Case 1: In integrals containing a term of the form $(ax+b)^n$, set $ax + b = t$.

Case 2: If the integral includes the expression $\sqrt{a^2-x^2}$, then we set $x = |a|sin\theta$ or $x = |a|cos\theta$, so that: (i) if $x = |a|sin\theta$, then $dx = |a|cos\theta d\theta$ and $\sqrt{a^2-x^2} = |a|cos\theta$; (ii) if $x = |a|cos\theta$, then $dx = -|a|sin\theta d\theta$ and $\sqrt{a^2-x^2} = |a|sin\theta$.

Case 3: If the integral includes the expression $\sqrt{a^2+x^2}$, then we set $x = |a|tan\theta$ (or $x = |a|cot\theta$). If $x = |a|tan\theta$, then $dx = \frac{|a|}{cos^2\theta}d\theta$ and $\sqrt{a^2+x^2} = \frac{|a|}{cos\theta}$.

Case 4: If the integral includes the expression $\sqrt{x^2-a^2}$, then we set $x = |a|\frac{1}{cos\theta}$, so that $dx = |a|\frac{sin\theta}{cos^2\theta}d\theta$ and $\sqrt{x^2-a^2} = |a|\frac{sin\theta}{cos\theta} = |a|tan\theta$.

Case 5: If the integral includes the expression $\sqrt{ax+b}$, then we set $\sqrt{ax+b} = t$. For instance, in order to compute the integral $\int\frac{sin\sqrt{x}}{\sqrt{x}}dx$, we work as follows: Setting $\sqrt{x} = t \Rightarrow x = t^2$, and $dx = 2tdt$. Hence, $\int\frac{sin\sqrt{x}}{\sqrt{x}}dx = \int\frac{sint}{t}2tdt = 2\int sintdt = -2cost + c = -2cos\sqrt{x} + c$.

Case 6: If the integral includes the expression $\sqrt{2ax-x^2}$, $a > 0$, then we set $x = a(1-cos\theta)$, so that $dx = asin\theta d\theta$ and $\sqrt{2ax-x^2} = asin\theta$.

Case 7: Integrals of the form $\int\frac{dx}{\sqrt{ax^2+bx+c}}$ can always be reduced to one or other of the three standard forms: $\int\frac{dx}{\sqrt{a^2-x^2}}$, $\int\frac{dx}{\sqrt{a^2+x^2}}$, $\int\frac{dx}{\sqrt{x^2-a^2}}$ (and we work as above).

Case 8: In case of integrals of the form $\int\frac{dx}{a+bcosx}$ or $\int\frac{dx}{a+bsinx}$, set $tan\frac{x}{2} = t$.

Case 9: In case of integrals of the form $\int\frac{dx}{(px+q)\sqrt{ax^2+bx+c}}$, set $px + q = \frac{1}{t}$.

Integration by parts: The method of integration by parts is based on the following theorem: If two functions f and g are differentiable on a real interval I, and if the indefinite integral of the function $f'g$ exists in I, then the idefinite integral of the function fg' also exists in I, and it holds that

$$\int f(x)\, g'(x)dx = f(x)g(x) - \int f'\,(x)g(x)dx, x \in I$$

(this theorem can be immediately verified by the definition of the indefinite integral, since $\frac{d}{dx}[f(x)g(x) - \int f'\,(x)g(x)dx] = \frac{d}{dx}[f(x)g(x)] - \frac{d}{dx}\int f'\,(x)g(x)dx = f'(x)g(x) + f(x)g'(x) - f'(x)g(x) = f(x)g'(x)$).

Remark: For convenience, the formula of integration by parts is often stated as follows:

$$\int udv = uv - \int vdu$$

(in order to use this formula, the integrand must be expressed as the product of a function and the differential of another function; and, in particular, we often follow the LIATE rule, which tells us to choose u to be the function that appears first in the following list: Logarithmic functions, Inverse trigonometric functions, Algebraic functions, Trigonometric functions, and Exponential functions).

For instance, given the integral $\int xe^x dx$, we set $u = x$ and $dv = e^x dx$. Then $du = dx$ and $v = \int e^x dx = e^x$. Hence, $\int xe^x dx = \int xde^x = xe^x - \int e^x dx = xe^x - e^x + c$.

Similarly, we can compute the integral $\int xcosxdx$ using the method of integration by parts as follows:

$\int xcosxdx = \int xd(sinx) = xsinx - \int sinxdx = xsinx - (-cosx) + c = xsinx + cosx + c$.

Integration by reduction formulae: A reduction formula is applied to a given integral in order to express it in terms of a much simpler integral, using the rule for integration by parts, namely, $\int udv = uv - \int vdu$. When we have to compute integrals of higher order, we usually have to look for a reduction formula, such as the following (by successive applications of the corresponding reduction formula, the integral of any power of the integrand can be obtained):

Case 1: $\int x^n\, e^x dx$. We shall apply the rule for integration by parts, setting $u = x^n$ and then $du = nx^{n-1}dx$. Similarly, we can set $dv = e^x dx$, so that $\int dv = \int e^x dx$, and, differentiating both sides, we get $v = e^x$. Hence,

$$\int x^n\, e^x dx = x^n e^x - n\int x^{n-1}\, e^x dx$$

(and, thus, we have obtained a reduction formula). Notice that

$\int x^n\, e^{mx} dx = \frac{1}{m} x^n e^{mx} - \frac{n}{m}\int x^{n-1}\, e^{mx} dx$.

Case 2: $\int ln^n\, xdx$. We shall apply the rule for integration by parts, setting $u = ln^n x$ and then $du = nln^{n-1}x \cdot \frac{1}{x} dx = \frac{n}{x} ln^{n-1}xdx$. Similarly, we can set $dv = dx$, so that $v = x$. Hence,

$$\int ln^n\, xdx = ln^n x \cdot x - \int x\frac{n}{x} ln^{n-1}xdx \Rightarrow \int ln^n\, xdx$$
$$= xln^n x - n\int ln^{n-1}xdx$$

(and, thus, we have obtained a reduction formula).

Case 3: $\int x^m \, ln^n x dx$. We shall apply the rule for integration by parts, setting $u = ln^n x$ and then $du = \frac{nln^{n-1}x}{x}dx$. Similarly, we can set $dv = x^m$, so that $v = \frac{x^{m+1}}{m+1}$. Hence,

$$\int x^m \, ln^n x dx = \frac{x^{m+1} ln^n x}{m+1} - \frac{n}{m+1}\int x^m \, ln^{n-1} x dx$$

(and, thus, we have obtained a reduction formula).

Case 4: $\int sin^n x \, dx$. In this case, in order to apply the rule for integration by parts, we shall separate the integrand into two parts by writing $sin^n x = sin^{n-1}x \cdot sinx$, so that $\int sin^n x \, dx = \int sin^{n-1}x \cdot sinxdx$. Now, we are ready to apply the rule for integration by parts, setting $u = sin^{n-1}x$, $dv = sinxdx$, and then $du = (n-1)sin^{n-2}x \cdot cosxdx$, and $v = -cosx$. Hence, we obtain

$$\begin{aligned}\int sin^n x \, dx = \int sin^{n-1}x \cdot sinxdx \\ &= -sin^{n-1}xcosx + (n-1)\int cos^2 xsin^{n-2}xdx \\ &= -sin^{n-1}xcosx + (n-1)\int (1 - sin^2 x) \, sin^{n-2}xdx \\ &= -sin^{n-1}xcosx + (n-1)\int sin^{n-2} xdx \\ &- (n-1)\int sin^n xdx \Leftrightarrow n\int sin^n x \, dx \\ &= -sin^{n-1}xcosx + (n-1)\int sin^{n-2} xdx \Leftrightarrow \int sin^n x \\ &= -\frac{1}{n} sin^{n-1}xcosx + \frac{n-1}{n}\int sin^{n-2} xdx\end{aligned}$$

(this is the reduction formula for this type of integrals; so that, writing $I_n = \int sin^n x \, dx$ and $I_{n-2} = \int sin^{n-2} xdx$, the reduction formula can be written as follows: $nI_n = -sin^{n-1}xcosx + (n-1)I_{n-2} \Leftrightarrow I_n = -\frac{1}{n} sin^{n-1}xcosx + \frac{n-1}{n} I_{n-2}$).

Case 5: $\int cos^n xdx$. By rewriting the given integral as $\int cos^{n-1}x \cdot cosxdx$, we can apply the rule for integration by parts by setting $u = cos^{n-1}x$, $dv = cosxdx$, and then $du = (n-1)cos^{n-2}x \cdot (-sinx)dx$, and $v = sinx$. Hence, we ultimately obtain

$$\int cos^n xdx = \frac{1}{n} cos^{n-1}xsinx + \frac{n-1}{n}\int cos^{n-2} xdx$$

(this is the reduction formula for this type of integrals).

Case 6: $\int sin^m xcos^n xdx$, where m and n are natural numbers. Let us call this integral $I_{m,n}$, where m represents the power of the sine term in the integrand, and n represents the power of the cosine term in the integrand. Then the following reduction formulae hold:

$$I_{m,n} = \frac{sin^{m+1}xcos^{n-1}x}{m+n} + \frac{n-1}{m+n}I_{m,n-2}$$

for $n \geq 2$; and

$$I_{m,n} = -\frac{sin^{m+1}xcos^{n+1}x}{m+n} + \frac{m-1}{m+n}I_{m-2,n}$$

for $m \geq 2$.

Let us prove the first reduction formula (applying the rule for integration by parts):

$$\begin{aligned} I_{m,n} = \int sin^m xcos^{n-1}\, xd(sinx) \\ &= sin^{m+1}xcos^{n-1}x \\ &- \int sinx[msin^{m-1}xcos^n x \\ &- (n-1)sin^{m+1}xcos^{n-2}x]\, dx \\ &= sin^{m+1}xcos^{n-1}x - m\int sin^m\, xcos^n xdx \\ &+ (n-1)\int sin^{m+2}xcos^{n-2}\, xdx \\ &= sin^{m+1}xcos^{n-1}x - mI_{m,n} \\ &+ (n-1)\int sin^m x(1-cos^2x)\, cos^{n-2}xdx \\ &= sin^{m+1}xcos^{n-1}x - mI_{m,n} + (n-1)I_{m,n-2} \\ &- (n-1)I_{m,n} \Leftrightarrow I_{m,n} \\ &= \frac{sin^{m+1}xcos^{n-1}x}{m+n} + \frac{n-1}{m+n}I_{m,n-2} \end{aligned}$$

(this is the required reduction formula). In case $m = 0$, this reduction formula implies that

$$\int cos^n\, xdx = \frac{sinxcos^{n-1}x}{n} + \frac{n-1}{n}\int cos^{n-2}\, xdx$$

(which we obtained in Case 5).

If we set $I_{m,n} = \int sin^m\, xcos^n xdx = \int cos^n\, xsin^{m-1}xsinxdx$, and if we work in a way analogous to the way we proved the first reduction formula, then we obtain the second reduction formula, namely: $I_{m,n} = -\frac{sin^{m+1}xcos^{n+1}x}{m+n} + \frac{m-1}{m+n}I_{m-2,n}$. In case $n = 0$, this reduction formula implies that

$$\int sin^m\, xdx = -\frac{cosxsin^{m-1}x}{m} + \frac{m-1}{m}\int sin^{m-2}\, xdx$$

(which we obtained in Case 4).

Case 7: $\int tan^n x\, dx$. We work as follows (based on the formula of integration by parts):
$I_n = \int tan^n x\, dx = \int tan^{n-2} x tan^2 x dx = \int tan^{n-2} x\, (sec^2 x - 1) dx = \int (tan^{n-2} x sec^2 x - tan^{n-2} x) dx = \int tan^{n-2} x sec^2 x dx - \int tan^{n-2} x dx$.
In case of the integral $\int tan^{n-2} x sec^2 x dx$, let us apply the rule for integration by parts with $u = tan^{n-2} x$ and $dv = sec^2 x$, so that $du = (n-2) sec^2 x tan^{n-3} x$ and $v = tanx$. Then, by the rule for integration by parts, $\int tan^{n-2} x sec^2 x dx = tan^{n-1} x - (n-2) \int sec^2 x tan^{n-2} x dx = tan^{n-1} x - (n-2) \int (1 + tan^2 x)\, tan^{n-2} x dx = tan^{n-1} x - (n-2) \int (tan^{n-2} x + tan^n x)\ dx$.
Therefore, returning to $I_n = \int tan^n x\, dx$, we have found that

$$\int tan^n x\, dx = tan^{n-1} x - (n-2) \int tan^{n-2} x dx - (n-2) \int tan^n x dx - \int tan^{n-2} x dx \Leftrightarrow \int tan^n x\, dx + (n-2) \int tan^n x\, dx = (n-1) \int tan^n x\, dx = tan^{n-1} x - (n-1) \int tan^{n-2} x dx \Leftrightarrow \int tan^n x\, dx = \frac{tan^{n-1} x}{n-1} - \int tan^{n-2} x dx$$

(this is the required reduction formula; so that, writing $I_n = \int tan^n x\, dx$ and $I_{n-2} = \int tan^{n-2} x dx$, the reduction formula can be written as follows: $I_n = \frac{tan^{n-1} x}{n-1} - I_{n-2}$).
Case 8: $\int \frac{dx}{(a^2+x^2)^n}$, where n is a natural number. If $n = 1$, then we have $\int \frac{dx}{a^2+x^2}$, and we can compute it by making the substitution $x = atan\theta$ and $dx = asec^2\theta d\theta$, so that $\int \frac{dx}{a^2+x^2} = \int \frac{1}{a^2+(atan\theta)^2} asec^2\theta d\theta = \int \frac{1}{a^2(1+tan^2\theta)} asec^2\theta d\theta = \int \frac{1}{a^2 sec^2\theta} asec^2\theta d\theta = \frac{1}{a} \int d\theta = \frac{1}{\alpha}\theta + c$, and, because $x = atan\theta \Leftrightarrow \theta = tan^{-1}\frac{x}{a}$; and, therefore, $\int \frac{dx}{a^2+x^2} = \frac{1}{a} arctan \frac{x}{a} + c$.
If $n \geq 2$, then we find a reduction formula as follows: Let us call this integral I_n, where n represents the power of the denominator. Then

$$
\begin{aligned}
I_n = \int \frac{dx}{(a^2+x^2)^n} &= \frac{1}{a^2}\int \frac{a^2+x^2-x^2}{(a^2+x^2)^n}dx \\
&= \frac{1}{a^2}\int \frac{dx}{(a^2+x^2)^{n-1}} - \frac{1}{a^2}\int \frac{x^2}{(a^2+x^2)^n}dx \\
&= \frac{1}{a^2}I_{n-1} - \frac{1}{a^2}\int x\frac{x}{(a^2+x^2)^n}dx \\
&= \frac{1}{a^2}I_{n-1} - \frac{1}{2}\cdot\frac{1}{a^2}\int x\frac{d(a^2+x^2)}{(a^2+x^2)^n} \\
&= \frac{1}{a^2}I_{n-1} - \frac{1}{2a^2}\int xd\frac{(a^2+x^2)^{1-n}}{1-n} \\
&= \frac{1}{a^2}I_{n-1} - \frac{1}{2a^2}\cdot\frac{x}{(1-n)(a^2+x^2)^{n-1}} + \frac{1}{2a^2} \\
&\cdot\frac{1}{1-n}\int\frac{dx}{(a^2+x^2)^{n-1}} \\
&= \frac{1}{a^2}I_{n-1} - \frac{1}{2a^2}\cdot\frac{x}{(1-n)(a^2+x^2)^{n-1}} + \frac{1}{2a^2} \\
&\cdot\frac{1}{1-n}I_{n-1} \Leftrightarrow I_n \\
&= \frac{1}{2(n-1)a^2}\cdot\frac{x}{(a^2+x^2)^{n-1}} + \frac{2n-3}{2(n-1)}\cdot\frac{1}{a^2}I_{n-1}
\end{aligned}
$$

(this is the required reduction formula for $n \geq 2$).

Integration of rational functions: In general, expressions of the form $\frac{f(x)}{g(x)}$ where $f(x)$ and $g(x)$ are rational integral algebraic functions of x can be resolved into partial fractions, provided that the degree of $f(x)$ is less than the degree of $g(x)$, and $g(x)$ itself can be expressed in terms of linear and quadratic factors. Hence, the aforementioned expression $\frac{f(x)}{g(x)}$ can be integrated if each of the corresponding separate partial fractions can be integrated. In fact, the following types of partial fractions will arise: $\frac{A}{ax+b}$, $\frac{A}{(ax+b)^n}$, $\frac{Ax+B}{ax^2+bx+c}$, and $\frac{Ax+B}{(ax^2+bx+c)^n}$.

Table 8:1: Partial fractions decomposition.

Form of the rational function	Form of the partial fraction
$\boldsymbol{\frac{px+q}{(x-a)(x-b)}, a \neq b}$	$\frac{A}{x-a}+\frac{B}{x-b}$
$\boldsymbol{\frac{px+q}{(x-a)^2}}$	$\frac{A}{x-a}+\frac{B}{(x-a)^2}$
$\boldsymbol{\frac{px^2+qx+r}{(x-a)(x-b)(x-c)}}$	$\frac{A}{x-a}+\frac{B}{x-b}+\frac{C}{x-c}$
$\boldsymbol{\frac{px^2+qx+r}{(x-a)^2(x-b)}}$	$\frac{A}{x-a}+\frac{B}{(x-a)^2}+\frac{C}{x-b}$
$\boldsymbol{\frac{px^2+qx+r}{(x-a)(x^2+bx+c)}}$	$\frac{A}{x-a}+\frac{Bx+C}{x^2+bx+c}$

Example 1: $\int \frac{1}{x^2-4}dx$. Factoring the denominator, we obtain two distinct linear factors: $\int \frac{1}{x^2-4}dx = \int \frac{1}{(x+2)(x-2)}dx = \int \frac{A}{x+2}dx + \int \frac{B}{x-2}dx$, and we have to determine the constants A and B. Thus, $\frac{1}{(x+2)(x-2)} = \frac{A}{x+2} + \frac{B}{x-2} \Rightarrow$ $(x+2)(x-2)\frac{1}{(x+2)(x-2)} = (x+2)(x-2)\left(\frac{A}{x+2}+\frac{B}{x-2}\right) \Rightarrow 1 =$ $A(x-2) + B(x+2)$. Given the linear factors, $x-2$ and $x+2$, we shall find the values of A and B as follows: Plugging in the value $x = 2$ (derived from $x-2=0 \Leftrightarrow x=2$), $1 = A(x-2)+B(x+2) \Rightarrow 1 =$ $0 + B(4) \Rightarrow B = \frac{1}{4}$. Plugging in the value $x = -2$ (derived from $x+2 =$ $0 \Leftrightarrow x = -2$), $1 = A(x-2)+B(x+2) \Rightarrow 1 = A(-2-2)+0 \Rightarrow A =$ $-\frac{1}{4}$. Now, the given indefinite integral becomes $\int \frac{1}{x^2-4}dx = \int \frac{-\frac{1}{4}}{x+2}dx +$ $\int \frac{\frac{1}{4}}{x-2}dx = -\frac{1}{4}\int \frac{dx}{x+2} + \frac{1}{4}\int \frac{dx}{x-2} = -\frac{1}{4}\ln|x+2| + \frac{1}{4}\ln|x-2| + c =$ $\frac{1}{4}(\ln|x-2| - \ln|x+2|) + c = \frac{1}{4}\ln\left|\frac{x-2}{x+2}\right| + c.$

Example 2: $\int \frac{x-4}{x^2+2x-15}dx$. We need to factor the denominator, and, therefore, in this trinomial, the leading coefficient being 1, we need to find two numbers whose product is equal to the constant term, that is, -15, and whose sum is equal to the linear coefficient, that is, 2. These two numbers are 5 and -3, so that $x^2+2x-15 = (x+5)(x-3)$. Thus, $\int \frac{x-4}{x^2+2x-15}dx = \int \frac{A}{x+5}dx + \int \frac{B}{x-3}dx$, and we can determine A and B as follows: $\frac{x-4}{(x+5)(x-3)} = \frac{A}{x+5} + \frac{B}{x-3} \Rightarrow (x+5)(x-3)\frac{x-4}{(x+5)(x-3)} =$

$(x+5)(x-3)\left(\frac{A}{x+5}+\frac{B}{x-3}\right) \Rightarrow x-4 = A(x-3)+B(x+5)$. As in the previous example, we shall plug in values of x to calculate A and B; and, specifically, given the linear terms $x-3$ and $x+5$, let us firstly plug in $x=3$ (derived from $x-3=0 \Leftrightarrow x=3$), so that $x-4=A(x-3)+B(x+5) \Rightarrow 3-4=0+B(3+5) \Rightarrow B=-\frac{1}{8}$. Plugging in $x=-5$ (derived from $x+5=0 \Leftrightarrow x=-5$), $x-4=A(x-3)+B(x+5) \Rightarrow -5-4=A(-5-3)+0 \Rightarrow A=\frac{9}{8}$. Now, the given indefinite integral becomes $\int \frac{x-4}{x^2+2x-15}dx = \int \frac{\frac{9}{8}}{x+5}dx + \int \frac{-\frac{1}{8}}{x-3}dx = \frac{9}{8}\int \frac{dx}{x+5} - \frac{1}{8}\int \frac{dx}{x-3} = \frac{9}{8}ln|x+5| - \frac{1}{8}ln|x-3| + c$.

Example 3: $\int \frac{x}{(x-1)(x-2)^2}dx = \int \frac{A}{x-1}dx + \int \frac{B}{x-2}dx + \int \frac{C}{(x-2)^2}dx$, so that $\frac{x}{(x-1)(x-2)^2} = \frac{A}{x-1} + \frac{B}{x-2} + \frac{C}{(x-2)^2} \Rightarrow (x-1)(x-2)^2\frac{x}{(x-1)(x-2)^2} = (x-1)(x-2)^2\left[\frac{A}{x-1}+\frac{B}{x-2}+\frac{C}{(x-2)^2}\right] \Rightarrow x = A(x-2)^2 + B(x-1)(x-2) + C(x-1)$. In order to determine A, B, and C, we must plug in some values of x. Let's focus on $x-2=0 \Leftrightarrow x=2$, and $x-1=0 \Leftrightarrow x=1$. Plugging in $x=2$, we get: $x=A(x-2)^2+B(x-1)(x-2)+C(x-1) \Rightarrow 2=0+0+C \Rightarrow C=2$. Plugging in $x=1$, we get: $x=A(x-2)^2+B(x-1)(x-2)+C(x-1) \Rightarrow 1=A(1-2)^2+0+0 \Rightarrow A=1$. In order to determine B, let's plug in $x=3$, so that $x=A(x-2)^2+B(x-1)(x-2)+C(x-1) \Rightarrow 3=A(1)+B(2)(1)+C(2)=1+2B+4 \Rightarrow 3=5+2B \Rightarrow B=-1$. Now, the given indefinite integral becomes $\int \frac{x}{(x-1)(x-2)^2}dx = \int \frac{dx}{x-1} - \int \frac{dx}{x-2} + \int \frac{2}{(x-2)^2}dx$, where $\int \frac{dx}{x-1} = ln|x-1| + c_1$, $\int \frac{dx}{x-2} = ln|x-2| + c_2$, and, in order to compute $\int \frac{2}{(x-2)^2}dx = 2\int \frac{dx}{(x-2)^2}$, we set $u=x-2$ and $du=dx$, so that we obtain $2\int \frac{dx}{(x-2)^2} = 2\int \frac{du}{u^2} = 2\int u^{-2}\,du = 2\frac{u^{-1}}{-1} + c_3 = -\frac{2}{u} + c_3 = -\frac{2}{x-2} + c_3$. Hence, $\int \frac{x}{(x-1)(x-2)^2}dx = \int \frac{dx}{x-1} - \int \frac{dx}{x-2} + \int \frac{2}{(x-2)^2}dx = ln|x-1| - ln|x-2| - \frac{2}{x-2} + c$.

Example 4: $\int \frac{x^2+9}{(x^2-1)(x^2+4)}dx$. We cannot factor x^2+4, but we can factor x^2-1 as $(x+1)(x-1)$, so that $\frac{x^2+9}{(x^2-1)(x^2+4)} = \frac{x^2+9}{(x+1)(x-1)(x^2+4)}$, and, thus, in the denominator of the integrand, we have two linear factors and one quadratic factor. We work as follows: $\frac{x^2+9}{(x+1)(x-1)(x^2+4)} = \frac{A}{x+1} + \frac{B}{x-1} +$

$\frac{Cx+D}{x^2+4} \Rightarrow \int \frac{x^2+9}{(x+1)(x-1)(x^2+4)} dx = \int \frac{A}{x+1} dx + \int \frac{B}{x-1} dx + \int \frac{Cx+D}{x^2+4} dx$. Then, working as before, $\frac{x^2+9}{(x+1)(x-1)(x^2+4)} = \frac{A}{x+1} + \frac{B}{x-1} + \frac{Cx+D}{x^2+4} \Rightarrow (x+1)(x-1)(x^2+4)\frac{x^2+9}{(x+1)(x-1)(x^2+4)} = (x+1)(x-1)(x^2+4)\left(\frac{A}{x+1} + \frac{B}{x-1} + \frac{Cx+D}{x^2+4}\right) \Rightarrow x^2+9 = A(x-1)(x^2+4) + B(x+1)(x^2+4) + (Cx+D)(x+1)(x-1)$. In order to determine A, B, C, and D, we must plug in some values of x. Let's focus on $x-1=0 \Leftrightarrow x=1$, and $x+1=0 \Leftrightarrow x=-1$. Plugging in $x=1$, we get: $x^2+9 = A(x-1)(x^2+4) + B(x+1)(x^2+4) + (Cx+D)(x+1)(x-1) \Rightarrow 1+9 = 0 + B(2)(5) + 0 \Rightarrow B=1$. Plugging in $x=-1$, we get: $x^2+9 = A(x-1)(x^2+4) + B(x+1)(x^2+4) + (Cx+D)(x+1)(x-1) \Rightarrow 10 = A(-2)(5) + 0 + 0 \Rightarrow A = -1$. In order to determine C and D, let's plug in $x=0$ (since, for $x=0$, C disappears in $Cx+D$, and we can solve for D). Indeed, for $x=0$, and given that we have found that $A=-1$ and $B=1$, we have: $x^2+9 = A(x-1)(x^2+4) + B(x+1)(x^2+4) + (Cx+D)(x+1)(x-1) \Rightarrow 0^2+9 = -1(-1)(4) + (1)(1)(4) + (C\cdot 0 + D)(1)(-1) \Rightarrow D = -1$. Finally, we need to determine C, and, for this reason, let's plug in $x=2$, so that, for $x=2$, and given that we have found that $A=-1$, $B=1$, and $D=-1$, we have: $x^2+9 = A(x-1)(x^2+4) + B(x+1)(x^2+4) + (Cx+D)(x+1)(x-1) \Rightarrow C=0$. Now, the given indefinite integral becomes $\int \frac{x^2+9}{(x^2-1)(x^2+4)} dx = \int \frac{-1}{x+1} dx + \int \frac{1}{x-1} dx + \int \frac{-1}{x^2+4} dx$, where $\int \frac{-1}{x+1} dx = -ln|x+1| + c_1$, $\int \frac{1}{x-1} dx = ln|x-1| + c_2$, and, in order to compute $\int \frac{-1}{x^2+4} dx$, we shall use the formula $\int \frac{dx}{a^2+x^2} = \frac{1}{a} arctan\frac{x}{a} + c$ (which was proved earlier), so that $\int \frac{-1}{x^2+4} dx = -\frac{1}{2} arctan\frac{x}{2} + c_3$. Hence, $\int \frac{x^2+9}{(x^2-1)(x^2+4)} dx = -ln|x+1| + ln|x-1| - \frac{1}{2} arctan\frac{x}{2} + c = ln\left|\frac{x-1}{x+1}\right| - \frac{1}{2} arctan\frac{x}{2} + c.$

Remarks:

i. In case of $\int \frac{Ax+B}{x^2+px+q} dx$ with $p^2 - 4q < 0$, we set $x + \frac{p}{2} = t$.

ii. In case of $\int R(e^{ax})\, dx$, where R is a rational function, we set $e^{ax} = t$.

Integration of irrational functions: Some types of integrals of irrational algebraic functions are reducible to integrals of rational functions via suitable substitutions.

Case 1: $\int R\left(x, \sqrt[n]{\frac{ax+b}{cx+d}}\right) dx$, where R is a rational function, n is a positive integer $\neq 1$, $ad \neq bc$, and $\frac{ax+b}{cx+d} > 0$ if n is even. In this case, we set $\frac{ax+b}{cx+d} = t^n$, so that $x = \frac{b-dt^n}{ct^n-a}$, and $dx = \frac{n(ad-bc)t^{n-1}}{(a-ct^n)^2} dt$. Hence, the original integral becomes $\int R\left(x, \sqrt[n]{\frac{ax+b}{cx+d}}\right) dx = n(ad - bc) \int R\left(\frac{b-dt^n}{ct^n-a}, t\right) \frac{t^{n-1}}{(a-ct^n)^2} dt.$

For instance, given the integral

$$\int \sqrt[3]{\frac{x+1}{x-1}} dx$$

(which is of the above form), we set $\frac{x+1}{x-1} = t^3$, so that $x = \frac{1+t^3}{t^3-1}$, and $dx = -\frac{6t^2}{(t^3-1)^2} dt$, and then the original integral becomes

$\int \sqrt[3]{\frac{x+1}{x-1}} dx = \int \frac{-6t^3}{(t^3-1)^2} dt = 2\int t d \frac{1}{t^3-1} = \frac{2t}{t^3-1} - 2\int \frac{dt}{t^3-1} = \frac{2t}{t^3-1} - \frac{2}{3}\int \frac{dt}{t-1} + \frac{2}{3}\int \frac{t+2}{t^2+t+1} dt = \frac{2t}{t^3-1} + \frac{1}{3} ln \frac{t^3-1}{(t-1)^2} + \frac{2}{\sqrt{3}} arctan \frac{2t+1}{\sqrt{3}} + c,$

and, finally, we make the substitution $t = \sqrt[3]{\frac{x+1}{x-1}}$ to get the result as a function of x.

Case 2: $\int R\left(x, \sqrt[m]{\frac{ax+b}{cx+d}}, \sqrt[n]{\frac{ax+b}{cx+d}}, \ldots\right) dx$, where R is a rational function, $m, n, \ldots$ are natural numbers, and $ad \neq bc$. In this case, we set $\frac{ax+b}{cx+d} = t^p$, where p is the least common multiple of $m, n, \ldots$

For instance, given the integral

$$\int \frac{x}{\sqrt{x+1} - \sqrt[3]{x+1}} dx$$

(which is of the above form), we set $x + 1 = t^6$, so that we obtain $\int \frac{x}{\sqrt{x+1} - \sqrt[3]{x+1}} dx = \int \frac{(t^6-1)6t^5}{t^3-t^2} dt = 6\int (t^8 + t^7 + t^6 + t^5 + t^4 + t^3)\, dt$, which can be computed very easily, and, finally, we make the substitution $t = \sqrt[6]{x+1}$ to get the result as a function of x.

Case 3: $\int R\left(x, \sqrt{ax^2 + bx + c}\right) dx$, where R is a rational function, a, b, and c are real numbers, and $a \neq 0$. In this case, we apply the so-called “Euler’s substitutions,” namely:

i. If $a > 0$, then we set $\sqrt{ax^2 + bx + c} = t - \sqrt{a}x$ or $\sqrt{ax^2 + bx + c} = t + \sqrt{a}x$. Notice that, if $\sqrt{ax^2 + bx + c} = t - \sqrt{a}x$, then, by raising both sides to the square, we see that $x = \frac{t^2 - c}{2\sqrt{a}t + b}$, and $dx = 2\frac{\sqrt{a}t^2 + bt + c\sqrt{a}}{\left(2\sqrt{a}t + b\right)^2}dt$.

ii. If $a < 0$ and $b^2 - 4ac > 0$, then we set $\sqrt{ax^2 + bx + c} = t|x - r_1|$, where r_1 is a root of $ax^2 + bx + c = 0$.

iii. If $a < 0$ and $c > 0$, then we set $\sqrt{ax^2 + bx + c} = tx + \sqrt{c}$ or $\sqrt{ax^2 + bx + c} = tx - \sqrt{c}$. Moreover, by setting $x = \frac{1}{t}$, the situation reduces to case (i).

Binomial integrals: $\int x^m (a + bx^n)^p dx$, where m, n, and p are rational numbers, and a and b are non-zero real numbers. Integrals of this type can be computed only if at least one of the numbers p, $\frac{m+1}{n}$, $\frac{m+1}{n} + p$ is an integer (Chebyshev conditions). If p is an integer, then we set $t^k = x$, where k is the least common multiple of the denominators of the numbers m and n. If $\frac{m+1}{n}$ is an integer, then we set $a + bx^n = t^\lambda$, where λ is the denominator of p. If $\frac{m+1}{n} + p$ is an integer, then we set $ax^{-n} + b = t^\lambda$, where λ is the denominator of p.
For instance, the integral

$$\int \frac{dx}{\sqrt[3]{x^2}\left(1 + \sqrt[3]{x}\right)^3}$$

can be written as $\int x^{-2/3}\left(1 + x^{1/3}\right)^{-3}dx$, where $m = -\frac{2}{3}$, $n = \frac{1}{3}$, and $p = -3$. Because p is an integer, the above methodological rules imply that we should set $t^3 = x$, and then we obtain: $\int x^{-2/3}\left(1 + x^{1/3}\right)^{-3}dx = \int t^{-2}(1 + t)^{-3}3t^2dt = 3\int \frac{dt}{(1+t)^3} = -\frac{3}{2(1+t)^2} + c = -\frac{3}{2}\frac{1}{\left(1+\sqrt[3]{x}\right)^2} + c$.

Definite Integrals in $\mathbb{R}$

The “definite integral” is written as

$$\int_a^b f(x)dx$$

and represents the area bounded by the curve $y = f(x)$, the x-axis, and the ordinates $x = a$ and $x = b$ if $f(x) \geq 0$. If $f(x)$ is sometimes positive and sometimes negative, then the definite integral represents the algebraic sum of the areas above and below the x-axis. In particular, the areas that are above the x-axis are considered to be positive, whereas the areas that are below the x-axis are considered to be negative.

The development of analytic geometry gave rise to a new method for the calculation of the area of a curvilinear figure. The old method for the calculation of the area of a curvilinear figure consisted of a series of approximating polygons. The new method for the calculation of the area of a curvilinear figure consisted of a sequence of sums of approximating rectangles, as illustrated in Figure 8-16. The area of each of these rectangles can be represented by the product $f(x_i)\Delta x_i$ (corresponding to the product $height \times width$), and the sum of these rectangles is given by $S_n = \sum_{i=1}^{n} f(x_i)\Delta x_i$. Then the area of the figure can be defined as the limit of the infinite sequence of sums S_n as the number of subdivisions n increases indefinitely and, thus, as the intervals Δx_i approach zero. Moreover, in this context, the use of infinitesimal rectangles is intimately related to the explanation and treatment of an arbitrary curve as the limit of or a sum of infinitesimal bits (infinitesimal straight line segments), so that an arbitrary curve is locally (i.e., at the infinitesimal level) straight (in fact, this is the underlying idea of Riemann's theory of integration). In this way, the analytic representation of the curve set the stage for the development of the "definite integral" on the basis of the ordinary operations of arithmetic and on the basis of the concept of the limit of an infinite sequence of terms (S_n).

As shown in Figure 8-16, the definite integral $\int_a^b f(x)dx$ can be defined as follows:

We subdivide the closed interval $[a,b]$ into n subintervals
$[a,x_1],[x_1,x_2],\dots,[x_{i-1},x_i],[x_i,x_{i+1}],\dots,[x_{n-1},b]$
by means of the points $x_1,x_2,\dots,x_i,\dots,x_{n-1}$, which have been chosen arbitrarily (and, obviously, $x_0 < x_1 < \cdots < x_{n-1} < x_n$). Hence, the set of points
$P = \{a = x_0, x_1, x_2, \dots, x_i, \dots, x_{n-1}, x_n = b\}$
is a "partition" of $[a,b]$. Let Δx_i be the length of the ith subinterval, that is, $\Delta x_i = x_i - x_{i-1}$. Then the "norm" of the partition P is denoted by $||P||$, and it is equal to $max\{\Delta x_i | i = 1,2,\dots,n\}$. Now, in each of the n subintervals mentioned in the aforementioned partition, we choose points $\xi_1,\xi_2,\dots,\xi_n$ in an arbitrary way, and we form the sum

$$S(P,f,\xi_i) = \\ f(\xi_1)\Delta x_1 + f(\xi_2)\Delta x_2 + \cdots + f(\xi_i)\Delta x_i + \cdots + f(\xi_n)\Delta x_n \\ = \sum_{i=1}^{n} f(\xi_i)\Delta x_i$$

where $\Delta x_i = x_i - x_{i-1}$. In other words, an arbitrary domain value, ξ_i, is chosen in each subinterval, and the corresponding function value, $f(\xi_i)$, is determined, so that we can define the product of each function value times the corresponding subinterval's length (Δx_i) and then add these n products to determine their sum. This sum is called a "Riemann sum," and it may be positive, negative, or zero, depending on the behavior of the function on the given closed interval. Notice that the subintervals of the partition can be taken to be of equal length $\Delta x = \frac{b-a}{n}$ (in this case, $\|P\| = \Delta x$).

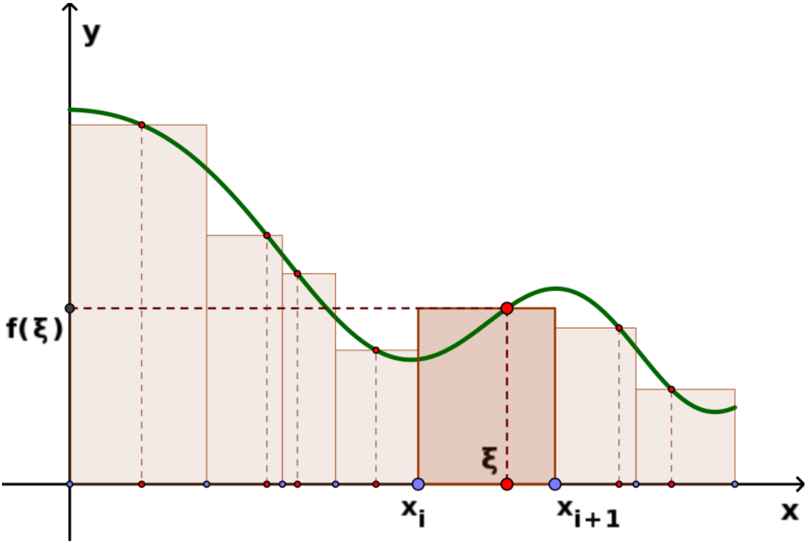
In general, as the number of subdivisions n increases, $\|P\|$ vanishes—that is, $\|P\| \to 0$ as $n \to \infty$ (and, obviously, if the subintervals of the partition have been taken to be of equal length $\Delta x = \frac{b-a}{n}$, then $\Delta x \to 0$ as $n \to \infty$).

Hence, if $lim_{\|P\| \to 0} S(P, f, \xi_i)$ exists and is independent of the mode of subdivision of $[a, b]$, then this limit is said to be the integral of f on $[a, b]$; symbolically:

$$lim_{\|P\| \to 0} S(P, f, \xi_i) = \int_a^b f(x)dx$$

where $f(x)dx$ is called the "integrand," $[a, b]$ is called the "range of integration," and a and b are respectively called the lower and the upper "limit of integration." Leibniz symbolized the definite integral of a function $f(x)$ on $[a, b]$ as $\int_a^b f(x)dx$, because the sign $\int$ is an elongated S standing for the word "sum," since Leibniz defined $\int_a^b f(x)dx$ as the summation of infinitely many rectangles of height $f(x)$ and infinitesimally small width dx (see Figure 8-16).

Figure 8-16: The integral as the limit of a sum (Wikimedia Commons: Author: Helder, Marcos Antônio Nunes de Moura; https://commons.wikimedia.org/wiki/File:Integral_de_Riemann.svg).



The above definition of the definite integral can be, equivalently, restated as follows (the epsilon-delta definition of the definite integral): A function f is called “integrable” on the interval $[a,b]$ if and only if there exists a number A such that: for every $\varepsilon > 0$, there exists a $\delta > 0$ such that every Riemann sum of f that corresponds to a partition $P = \{a = x_0, x_1, x_2, \dots, x_i, \dots, x_{n-1}, x_n = b\}$ of $[a,b]$ with $\|P\| < \delta$ satisfies the inequality

$$|S(P,f,\xi) - A| < \varepsilon$$

for any choice of sample points ξ of P. Then the number A is called the definite (or the Riemann) integral of f on $[a,b]$, and it is written as

$$\int_a^b f(x)dx$$

(that is, $A \equiv \int_a^b f(x)dx$).

Remark: Intuitively, the epsilon-delta definition of the definite (or Riemann) integral means that, as the partition becomes finer and finer, the Riemann sums converge to a limit, which is the definite (or Riemann) integral of f on $[a,b]$ (given a partition P of the interval $[a,b]$, another

partition Q of $[a,b]$ is said to be "finer" than P, or a "refinement" of P, if Q contains all the points of P and possibly others).
Hence, if $f:[a,b]\to\mathbb{R}$ is integrable on $[a,b]$ according to the aforementioned definition, then f is bounded on $[a,b]$. This theorem can be easily proved by reasoning as follows: As outlined above, the definite integral is calculated by partitioning $[a,b]$ into smaller intervals, and then, in each such subinterval, we choose a value of f, we multiply it by the length of the subinterval, and, finally, we sum all of these products. If, for the sake of contradiction, we assume that f is unbounded, then, in one of these subintervals (of the partition), f will still be unbounded, and, therefore, in this subinterval, we can choose a value for f so large that the resultant sum will also become analogously large, meaning that, among all the approximations of the integral, there will be sums of arbitrarily large size, thus rendering the function f unintegrable, which contradicts the assumption that f is integrable. Therefore, "integrable" on $[a,b]$ implies "bounded" on $[a,b]$.
For a given continuous function $f(x)$, of a real variable x, defined on the interval $[a,b]$, the definite integral is

$$\int_a^b f(x)dx = F(x)|_a^b = F(b) - F(a)$$

where $F(x)$ is the antiderivative (i.e., the indefinite integral) $F(x) = \int f(x)dx$, so that we calculate a definite integral as follows: (i) we calculate the antiderivative $F(x)$, (ii) we calculate the values $F(b)$ and $F(a)$, and (iii) we calculate $F(b) - F(a)$. For instance, we calculate the value of $\int_2^3 x^2dx$ as follows:

$$\int_2^3 x^2dx = \frac{x^3}{3}|_2^3 =$$

$$\left(value\ of \frac{x^3}{3} when\ x = 3\right) - \left(value\ of \frac{x^3}{3} when\ x = 2\right) = \frac{3^3}{3} - \frac{2^3}{3} = \frac{19}{3}.$$

Properties of the definite integral: The study of the definite integral was placed in a rigorous mathematical setting in the nineteenth century by Bernhard Riemann, Thomas Joannes Stieltjes, and Jean-Gaston Darboux, and their work underpins the following theorems (the properties of the definite integral).

Property 1: $\int_a^b f(x)dx = -\int_b^a f(x)dx$.
Proof: By the definition of the definite integral,
$\int_a^b f(x)dx = lim_{n\to\infty}\sum_{i=1}^n f(\xi_i)\,\Delta x$ where $\Delta x = \frac{b-a}{n}$,

and similarly
$\int_b^a f(x)dx = lim_{n\to\infty} \sum_{i=1}^n f(\xi_i)\, \Delta x$ where $\Delta x = \frac{a-b}{n}$.
Hence,
$\int_a^b f(x)dx = lim_{n\to\infty} \sum_{i=1}^n f(\xi_i) \frac{b-a}{n} = lim_{n\to\infty} \sum_{i=1}^n f(\xi_i) \left[\frac{-(a-b)}{n}\right] =$
$lim_{n\to\infty} \left(-\sum_{i=1}^n f(\xi_i) \frac{a-b}{n}\right) = -lim_{n\to\infty} \sum_{i=1}^n f(\xi_i) \frac{a-b}{n} = -\int_b^a f(x)dx.$■

Property 2: $\int_a^a f(x)dx = 0.$
Proof: By the definition of the definite integral,
$\int_a^a f(x)dx = lim_{n\to\infty} \sum_{i=1}^n f(\xi_i)\, \Delta x$ where $\Delta x = \frac{a-a}{n} = 0$,
and, therefore, $\int_a^a f(x)dx = lim_{n\to\infty} \sum_{i=1}^n f(\xi_i)\, (0) = 0.$■

Property 3: $\int_a^b cf(x)dx = c \int_a^b f(x)\, dx.$
Proof: By the definition of the definite integral (and the properties of summations and limits), we have:
$\int_a^b cf(x)dx = lim_{n\to\infty} \sum_{i=1}^n cf(\xi_i)\, \Delta x = lim_{n\to\infty} c \sum_{i=1}^n f(\xi_i)\, \Delta x =$
$c lim_{n\to\infty} \sum_{i=1}^n f(\xi_i)\, \Delta x = c \int_a^b f(x)dx$, where $\Delta x = \frac{b-a}{n}$.■

Property 4: $\int_a^b [f(x) \pm g(x)]\, dx = \int_a^b f(x)dx \pm \int_a^b g(x)dx.$
Proof: Firstly, we shall prove the formula for "+" (using the definition). Indeed, by the definition of the definite integral, using $\Delta x = \frac{b-a}{n}$, we have:
$\int_a^b [f(x) + g(x)]\, dx = lim_{n\to\infty} \sum_{i=1}^n [f(\xi_i) + g(\xi_i)]\, \Delta x =$
$lim_{n\to\infty} [\sum_{i=1}^n f(\xi_i)\, \Delta x + \sum_{i=1}^n g(\xi_i)\, \Delta x\ ] = \int_a^b f(x)dx + \int_a^b g(x)dx.$
The formula can be proved for "−" by repeating the above work with a minus sign.■

Property 5: $\int_a^b cdx = c(b-a)$, where c is a real number (constant).
Proof: If we define $f(x) = c$ (a constant function), then, by the definition of the definite integral, using $\Delta x = \frac{b-a}{n}$, we have:
$\int_a^b cdx = \int_a^b f(x)dx = lim_{n\to\infty} \sum_{i=1}^n f(\xi_i)\, \Delta x = lim_{n\to\infty} (\sum_{i=1}^n c) \frac{b-a}{n} =$
$lim_{n\to\infty} (cn) \frac{b-a}{n} = lim_{n\to\infty} c(b-a) = c(b-a).$■

Property 6: If $f(x) \geq 0$ for $x \in [a,b]$, then $\int_a^b f(x)dx \geq 0$.
Proof: By the definition of the definite integral, using $\Delta x = \frac{b-a}{n}$, we have:

$\int_a^b f(x)dx = lim_{n\to\infty} \sum_{i=1}^n f(\xi_i)\, \Delta x$,
and, because $f(x) \geq 0$ and $\Delta x \geq 0$, it holds that $\sum_{i=1}^n f(\xi_i)\, \Delta x \geq 0$. Thus, by the properties of limits, $lim_{n\to\infty} \sum_{i=1}^n f(\xi_i)\, \Delta x \geq lim_{n\to\infty} 0 \Leftrightarrow \int_a^b f(x)dx \geq 0$.■

Property 7: If $f(x) \geq g(x)$ for $x \in [a,b]$, then $\int_a^b f(x)dx \geq \int_a^b g(x)dx$.
Proof: Because $f(x) \geq g(x)$, it holds that $f(x) - g(x) \geq 0$ for $x \in [a,b]$, and, therefore, by Property 6, $\int_a^b [f(x) - g(x)]\, dx \geq 0$. Moreover, by Property 4, $\int_a^b [f(x) - g(x)]dx = \int_a^b f(x)dx - \int_a^b g(x)dx$. Therefore, $\int_a^b f(x)dx - \int_a^b g(x)dx \geq 0 \Rightarrow \int_a^b f(x)dx \geq \int_a^b g(x)dx$.■

Property 8 (Extreme Value Theorem for Definite Integrals): If $m \leq f(x) \leq M$ for $x \in [a,b]$, then $m(b-a) \leq \int_a^b f(x)dx \leq M(b-a)$.
Proof: Given that $m \leq f(x) \leq M$, we can use Property 7 on each inequality to obtain $\int_a^b mdx \leq \int_a^b f(x)dx \leq \int_a^b Mdx$. Then, by Property 5, we obtain $m(b-a) \leq \int_a^b f(x)dx \leq M(b-a)$.■

Property 9: $\left|\int_a^b f(x)dx\right| \leq \int_a^b |f(x)|\, dx$.
Proof: By the definition of the absolute value,
$-|f(x)| \leq f(x) \leq |f(x)|$.
Therefore, using Property 7, we obtain
$\int_a^b -|f(x)|dx \leq \int_a^b f(x)dx \leq \int_a^b |f(x)|dx \Rightarrow -\int_a^b |f(x)|dx \leq \int_a^b f(x)dx \leq \int_a^b |f(x)|dx$.
Hence, given that, in general, $|u| \leq v \Leftrightarrow -v \leq u \leq v$, we obtain the required result: $\left|\int_a^b f(x)dx\right| \leq \int_a^b |f(x)|\, dx$.■
Remark: The Cauchy–Schwarz–Bunyakovsky inequality for definite integrals: If f and g are continuous real-valued functions on $[a,b]$, then

$$\int_a^b |f(x)||g(x)|dx \leq \left[\int_a^b f^2(x)dx\right]^{\frac{1}{2}} \left[\int_a^b g^2(x)dx\right]^{\frac{1}{2}}$$

(this is a very useful result for proving oher inequalities in real analysis). The proof of this inequality can be obtained as follows: For a variable λ, let's define the function $p(\lambda) = \int_a^b [\lambda f(x) + g(x)]^2\, dx$. Hence,

$\int_a^b [\lambda f(x) + g(x)]^2\, dx = \int_a^b [\lambda^2 f^2(x) + 2\lambda f(x)g(x) + g^2(x)]\, dx =$ $\lambda^2 \int_a^b f^2(x)dx + 2\lambda \int_a^b f(x)g(x)dx + \int_a^b g^2(x)dx = I.$

Notice that I is a quadratic polynomial (in λ), and $I \geq 0$ if the discriminant is less than or equal to zero. Recall that: if the discriminant is positive, then we have two distinct real roots; if the discriminant is equal to zero, then we have a double root; and, if the discriminant is negative, then we do not have any real roots. The discriminant of I is $D = \left[2\int_a^b f(x)g(x)dx\right]^2 - 4\left[\int_a^b f^2(x)dx\right]\left[\int_a^b g^2(x)dx\right] \leq 0 \Rightarrow$ $\left[\int_a^b f(x)g(x)dx\right]^2 \leq \left[\int_a^b f^2(x)dx\right]\left[\int_a^b g^2(x)dx\right] \Rightarrow \left|\int_a^b f(x)g(x)dx\right| \leq$ $\left[\int_a^b f^2(x)dx\right]^{\frac{1}{2}}\left[\int_a^b g^2(x)dx\right]^{\frac{1}{2}}$. Then, by Property 9, we receive the required result: $\int_a^b |f(x)||g(x)|\, dx \leq \left[\int_a^b f^2(x)dx\right]^{\frac{1}{2}}\left[\int_a^b g^2(x)dx\right]^{\frac{1}{2}}$.■

Property 10: If a function $f: [a, b] \to \mathbb{R}$ is integrable on $[a, b]$, and if $a \leq a_1 \leq b_1 \leq b$, then f is integrable on $[a_1, b_1]$.

Proof: This property follows directly from the epsilon-delta definition of the definite integral (and it can be demonstrated by contradiction, since the lack of integrability over a subinterval results in the lack of integrability over the whole interval).■

Property 11 (additivity of domain for definite integrals): If $f: [a, b] \to \mathbb{R}$ is a continuous function, and if a, b, and c are real numbers such that $a < c < b$, then $f(x)$ is integrable on $[a, b]$ if and only if $f(x)$ is integrable on both $[a, c]$ and $[c, b]$, and then it holds that

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

Proof: Geometrically, this property means that, if we consider the signed area bounded by the graph of $y = f(x)$ and the x-axis from $x = a$ to $x = b$, then this signed area is equal to the sum of the signed area from $x = a$ to $x = c$ plus the signed area from $x = c$ to $x = b$. Algebraically, this property means that, if $F(x)$ is an antiderivative of $f(x)$, then $F(b) - F(a) = [F(c) - F(a)] + [F(b) - F(c)]$.

Let $P = \{a = x_0, x_1, x_2, \dots, x_i, \dots, x_{n-1}, x_n = b\}$ be a partition of $[a, b]$ such that c coincides with some point belonging to P, say $x_r = c$. Then P can be divided into the following two partitions:

$P_1 = \{a = x_0, x_1, .., x_r\}$ with norm $||P_1||$ and

$P_2 = \{x_r, x_{r+1}, .., x_n = b\}$ with norm $||P_2||$.

For any choice of sample points ξ of P, we shall have the following Riemann sum:
$S(P,f,\xi)=\sum_{i=1}^{n}f(\xi_i)(x_i-x_{i-1})=\sum_{i=1}^{r}f(\xi_i)(x_i-x_{i-1})+\sum_{i=r+1}^{n}f(\xi_i)(x_i-x_{i-1}).$ (1)
If we assume that f is integrable on $[a,b]$, then, by Property 10, it is also integrable on $[a,c]$ and $[c,b]$. If $\|P\|<\delta$, then, obviously, $\|P_1\|<\delta$ and $\|P_2\|<\delta$, and, therefore, by the epsilon-delta definition of the definite integral, we have:
$\left|S(P,f,\xi)-\int_a^b f(x)dx\right|<\frac{\varepsilon}{3}$,
$\left|S(P_1,f,\xi)-\int_a^c f(x)dx\right|<\frac{\varepsilon}{3}$, and
$\left|S(P_2,f,\xi)-\int_c^b f(x)dx\right|<\frac{\varepsilon}{3}$.
Due to relation (1), the above three inequalities imply that
$\left|\int_a^b f(x)\,dx-\int_a^c f(x)dx-\int_c^b f(x)\,dx\right|<\varepsilon$,
and, because ε is arbitrary, we obtain the required result:
$\int_a^b f(x)\,dx=\int_a^c f(x)dx+\int_c^b f(x)\,dx.$
The converse, starting from the assumption that f is integrable on $[a,c]$ and $[c,b]$, can be easily established from relation (1).■

First Fundamental Theorem of Calculus: If a function $f(x)$ is continuous on $[a,b]$, then the function

$$g(x)=\int_a^x f(t)dt$$

is continuous on $[a,b]$ and differentiable on (a,b), and it holds that

$$g'(x)=f(x)$$

(the first formulations of this theorem are due to Isaac Barrow, Isaac Newton, Gottfried Leibniz, and James Gregory, independently of each other; and this theorem establishes the relationship between differentiation and integration).
Proof: Suppose that x and $x+h$ are elements of the open interval (a,b). Then
$g(x+h)-g(x)=\int_a^{x+h}f(t)dt-\int_a^x f(t)\,dt.$ (1)
Using Property 11, we can rewrite relation (1) as follows:

$$\begin{aligned}g(x+h)-g(x)&=\left(\int_a^x f(t)\,dt+\int_x^{x+h}f(t)dt\right)-\int_a^x f(t)\,dt\\&=\int_x^{x+h}f(t)dt\end{aligned}$$

and, assuming that $h\neq 0$, we obtain:

$\frac{g(x+h)-g(x)}{h} = \frac{1}{h}\int_x^{x+h} f(t)dt.$ (2)

If we assume that $h > 0$, and given that x and $x + h$ are elements of the open interval (a, b), then $f(x)$ is continuous on $[x, x + h]$. Therefore, by Weierstrass's Extreme Value Theorem, there exist numbers c and d in $[x, x + h]$ such that $f(c) = m$ is the minimum of $f(x)$ in $[x, x + h]$, and $f(d) = M$ is the maximum of $f(x)$ in $[x, x + h]$. Then, by Property 8, it holds that

$$mh \le \int_x^{x+h} f(t)dt \le Mh \Rightarrow f(c)h \le \int_x^{x+h} f(t)dt \le f(d)h \Rightarrow f(c) \le \frac{1}{h}\int_x^{x+h} f(t)dt \le f(d)$$

and, by relation (2), we obtain:

$f(c) \le \frac{g(x+h)-g(x)}{h} \le f(d).$ (3)

If we assume that $h < 0$, then we can follow the same reasoning, except we shall be working on $[x + h, x]$ in order to obtain the same inequality as above. Consequently, we have proved that inequality (3) is true provided that $h \neq 0$.

Now, consider the case in which $h \to 0$. In this case, $c \to x$ and $d \to x$, since c and d are between x and $x + h$. Therefore,

$lim_{h\to 0} f(c) = lim_{c\to x} f(c) = f(x)$ and

$lim_{h\to 0} f(d) = lim_{d\to x} f(d) = f(x)$.

Then, by the Squeeze Theorem,

$lim_{h\to 0} \frac{g(x+h)-g(x)}{h} = f(x).$ (4)

The left-hand side of relation (4) is the definition of the derivative of $g(x)$, and, therefore,

$g'(x) = f(x).$ (5)

In other words, we have proved that $g(x)$ is differentiable on (a, b). Moreover, in the section on differential calculus, we proved that, if a function f is differentiable at x (having a finite derivative), then f is continuous at x. For this reason, relation (5) implies that $g(x)$ is also continuous on (a, b). Finally, if we set $x = a$ or $x = b$, we can follow a type of reasoning similar to the one we followed in order to obtain relation (4) using one-sided limits in order to obtain the same result, and, thus, the fact that "differentiability" implies "continuity" will lead us to the conclusion that $g(x)$ is continuous at $x = a$ or $x = b$, so that it will be ultimately established that $g(x)$ is continuous on $[a, b]$.■

Second Fundamental Theorem of Calculus: If a function $f(x)$ is continuous on $[a, b]$, and if $F(x)$ is any antiderivative of $f(x)$, then

$$\int_a^b f(x)dx = F(x)\,|_a^b = F(b) - F(a)$$

(the first formulations of this theorem are due to Isaac Barrow, Isaac Newton, Gottfried Leibniz, and James Gregory, independently of each other; and this theorem complements the First Fundamental Theorem of Calculus).

Proof: Let $g(x) = \int_a^x f(x)dt$. Then, by the First Fundamental Theorem of Calculus, $g'(x) = f(x)$, meaning that $g(x)$ is an antiderivative of $f(x)$ on $[a, b]$. Additionally, suppose that $F(x)$ is any antiderivative of $f(x)$ on $[a, b]$ that we want to choose. Thus, $g'(x) = F'(x)$. Then, by Corollary 2 of Lagrange's Mean Value Theorem, we know that $g(x)$ and $F(x)$ can differ by no more than an additive constant on (a, b). In other words, for $x \in (a, b)$, it holds that $F(x) = g(x) + c$. Because $g(x)$ and $F(x)$ are continuous on $[a, b]$, if we compute the corresponding limits as $x \to a^+$ and as $x \to b^-$, we realize that the last conclusion is also true at $x = a$ and $x = b$. Therefore, for all $x \in [a, b]$, $F(x) = g(x) + c$. This conclusion and the definition of $g(x)$ imply that

$$\begin{aligned} F(b) - F(a) = [g(b) + c] - [g(a) + c] &= g(b) - g(a) \\ &= \int_a^b f(t)dt + \int_a^a f(t)dt \\ &= \int_a^b f(t)dt + 0 = \int_a^b f(x)\,dx \end{aligned}$$

(in the last step, the change of t's into x's is legitimate, because the name of the variable used in the integral does not matter) .■

The Average Value of a Function on a Compact Interval: The average value of a function $f(x)$ over the compact interval $[a, b]$ is given by

$$\bar{f} = \frac{1}{b-a}\int_a^b f(x)dx$$

(in its modern form, this theorem is due to A.-L. Cauchy).

Proof: First of all, recall that the average value of n numbers is the sum of all these numbers divided by n. Now, let's divide the interval $[a, b]$ into n subintervals each of length

$\Delta x = \frac{b-a}{n}$.

From each of these subintervals, we choose the points $\xi_1, \xi_2, \ldots, \xi_n$, and the manner in which we choose these points does not matter as long as they come from the appropriate interval. The average of the function values $f(\xi_1), f(\xi_2), \ldots, f(\xi_n)$ is

$\frac{f(\xi_1)+f(\xi_2)+\cdots+f(\xi_n)}{n}$. (1)

The above definition of Δx implies that $n=\frac{b-a}{\Delta x}$. Hence, the fraction (1) becomes

$$\frac{f(\xi_1)+f(\xi_2)+\cdots+f(\xi_n)}{\frac{b-a}{\Delta x}}$$
$$=\frac{1}{b-a}[f(\xi_1)\Delta x+f(\xi_2)\Delta x+\cdots+f(\xi_n)\Delta x]$$
$$=\frac{1}{b-a}\sum_{i=1}^{n}f(\xi_i)\Delta x$$

where, by increasing n, we can compute the average of more and more function values in the interval $[a,b]$, and, in fact, the larger we choose n the better approximation of the average value of the function we shall obtain. If we take the limit as n tends to infinity, we shall actually obtain the average function value $\bar{f}$; symbolically:

$\bar{f}=lim_{n\to\infty}\frac{1}{b-a}\sum_{i=1}^{n}f(\xi_i)\Delta x=\frac{1}{b-a}lim_{n\to\infty}\sum_{i=1}^{n}f(\xi_i)\Delta x$,

where $lim_{n\to\infty}\sum_{i=1}^{n}f(\xi_i)\Delta x$ is the standard definition of the definite integral $\int_a^b f(x)dx$. Therefore, $\bar{f}=\frac{1}{b-a}\int_a^b f(x)dx$.■

Example: The average value of the function $f(x)=8-2x$ over the interval [0,4] is $\frac{1}{4-0}\int_0^4(8-2x)dx=4$. The point x_0 at which $f(x_0)$ is equal to the average value of f over [0,4] can be found as follows: $8-2x_0=4\Rightarrow x_0=2$.

The Mean Value Theorem for Integrals: If a function $f(x)$ is continuous on $[a,b]$, then there exists a number c in $[a,b]$ such that

$$\int_a^b f(x)\,dx=f(c)(b-a)$$

(in its modern form, this theorem is due to A.-L. Cauchy). This means that $f(c)=\bar{f}$, that is, $f(c)$ is equal to the average value of $f(x)$ over the interval $[a,b]$.

Proof: Let $F(x)=\int_a^x f(t)dt$. Because $f(x)$ is given to be continuous on $[a,b]$, the First Fundamental Theorem of Calculus implies that $F(x)$ is continuous on $[a,b]$ and differentiable on (a,b), as well as that $F'(x)=f(x)$. From Lagrange's Mean Value Theorem, we know that there exists a number c such that $a<c<b$ and $F(b)-F(a)=F'(c)(b-a)$. Additionally, we know that $F'(c)=f(c)$, $F(b)=\int_a^b f(t)dt=$

$\int_a^b f(x)dx$, and $F(a) = \int_a^a f(t)dt = 0$. Hence, we obtain $\int_a^b f(x)dx = f(c)(b-a)$, as required.■

Applications of the definite integral: In this section, we shall study a few applications of the definite integral.

1. *The area of a region in* $\mathbb{R}^2$

Suppose that f is a non-negative continuous function defined on the interval $[a, b]$. Let R be the set of all points (x, y) such that
$0 \leq y \leq f(x)$ and $a \leq x \leq b$,
meaning that R is a plane region bounded by the straight lines $x = a$ and $x = b$, the x-axis ($y = 0$), and the curve of the function $y = f(x)$. As stated previously, the area of R (being approximated by Riemann sums) is ultimately equal to

$$A = \int_a^b f(x)dx. \qquad (1)$$

For instance, let us consider the area of a rectangle whose sides are parallel to the axes of a Cartesian (rectangular) co-ordinate system. If the height of the rectangle is $h > 0$, and its width is $b - a$, then we set $f(x) = h$ over the interval $a \leq x \leq b$, and, therefore, the area of this rectangle is $\int_a^b f(x)dx = \int_a^b hdx = (b-a)\,h$. The area of a square, in particular, can be calculated as follows: if a is the length of the side of the square, where a is the distance from the origin of the coordinate system to $x = a$, then the area of the square is given by $A = \int_0^a adx = ax|_0^a = a^2$. Similarly, if we are given a right-angled triangle with height h and base b, where the base of this triangle is equal to the distance from the origin of the coordinate system to $x = b$, then, in order to calculate its area, we think as follows: in this case, our function is a straight line (the hypotenuse), and the general equation of a straight line is $f(x) = mx + c$, where c is a constant, but, in this case, $c = 0$ because y passes through the origin of the coordinate system, and the slope $m = \frac{h}{b}$ (since $m = \frac{"rise"}{"run"}$, as we explained in Chapter 6), so that the area of this triangle is $\int_0^b f(x)\,dx = \int_0^b \frac{h}{b}xdx = \frac{bh}{2}$ (by analogy, we can compute the area of any triangle; and, using analytic geometry and infinitesimal calculus, we can prove that the area of any triangle is given by $\frac{base \times height}{2}$).

The definition of the area of a region in $\mathbb{R}^2$ that is expressed by formula (1) is applicable for any function $f(x)$ that is non-negative and integrable over the interval under consideration (the function need not be continuous, although it is usually the case in applications that the function is continuous). The restriction that $f(x)$ is non-negative is not essential. Indeed, if $f(x) \leq 0$ and $a \leq x \leq b$, then, because the regions $R_1 = \{a \leq x \leq b, f(x) \leq y \leq 0\}$ and $R_2 = \{a \leq x \leq b, 0 \leq -y \leq -f(x)\}$ have the same area, it is sufficient to work with the function $-f(x)$, which is positive. In the general case where $f(x)$ does not have a constant sign over $[a, b]$, we divide $[a, b]$ into subintervals in which $f(x)$ has a constant sign, and we calculate the corresponding areas, so that, in this case, we have the following formula for the calculation of the area of a region in $\mathbb{R}^2$:

$A = \int_a^b |f(x)|\, dx.$ (2)

For instance, in order to calculate the area of the region bounded by the curve of $f(x) = x^2 - 4x + 3$, the x-axis, and the straight lines $x = -2$ and $x = 4$, we work as follows: because $f(x) \geq 0$ when $x \in [-2,1] \cup [3,4]$, and $f(x) \leq 0$ when $x \in [1,3]$, we have $A = \int_{-2}^{4} |x^2 - 4x + 3| dx = \int_{-2}^{1} (x^2 - 4x + 3) dx - \int_1^3 (x^2 - 4x + 3)\, dx + \int_3^4 (x^2 - 4x + 3) dx = 18$.

The area between two arbitrary curves can be calculated as follows: In the first case, we want to determine the area A between the equations $y = f(x)$ and $y = g(x)$ over the interval $[a, b]$ under the assumption that $f(x) \geq g(x)$, meaning that the graph of $f(x)$ is above the graph of $g(x)$. Then

$$A = \int_a^b [(upper\, function) - (lower\, function)] dx$$
$$= \int_a^b [f(x) - g(x)]\, dx$$

where $a \leq x \leq b$.

In the second case, we want to determine the area A between the equations $x = f(y)$ and $x = g(y)$ over the interval $[c, d]$ under the assumption that $f(y) \geq g(y)$, namely, $x = f(y)$ is on the right-hand side of $x = g(y)$. Then

$$A = \int_c^d [(rightfunction) - (leftfunction)] dy$$
$$= \int_c^d [f(y) - g(y)]\, dy$$

where $c \leq y \leq d$.

For instance, in order to calculate the area of the region bounded by the parabola $f(x) = x^2 - 3x$ and the straight line $g(x) = x$, as shown in

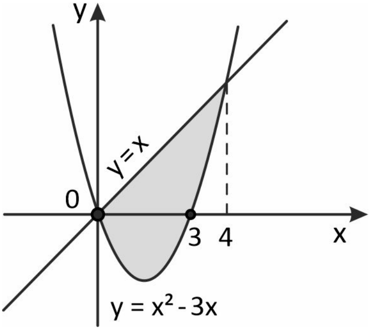
Figure 8-17, we work as follows: Firstly, we consider the system of equations

$$\begin{Bmatrix} f(x) = x^2 - 3x \\ g(x) = x \end{Bmatrix},$$

which gives the abscissas of the common points of the two curves. Thus, here, we have: $x^2 - 3x = x \Leftrightarrow (x_1 = 0, x_2 = 4)$, and $g(x) \geq f(x) \Rightarrow x \geq x^2 - 3x \Leftrightarrow x \in [0,4]$. Therefore, the required area is

$A = \int_0^4 [x - (x^2 - 3x)]dx = \int_0^4 (4x - x^2)dx = \frac{32}{3}$.

Figure 8-17: The area between two curves.



Now, as another example, let us use integral calculus in order to calculate the area of a triangle ABC whose vertices are $A(2,5)$, $B(4,7)$, and $C(6,2)$. Recall that, if (x_1, y_1) and (x_2, y_2) are any two points, then the equation of a straight line y passing through these two points is given by $\frac{y-y_1}{y_2-y_1} = \frac{x-x_1}{x_2-x_1}$. Thus, firstly, we must find the equations of the three sides of the triangle using the formula $\frac{y-y_1}{y_2-y_1} = \frac{x-x_1}{x_2-x_1}$.

The equation of the line AB is

$\frac{y-5}{7-5} = \frac{x-2}{4-2} \Rightarrow y = x + 3.$

The equation of the line BC is

$\frac{y-7}{2-7} = \frac{x-4}{6-4} \Rightarrow y = -\frac{5}{2}x + 17.$

The equation of the line AC is

$\frac{y-5}{2-5} = \frac{x-2}{6-2} \Rightarrow y = -\frac{3}{4}x + \frac{13}{2}.$

Thus,

$$Area\ of\ ABC$$
$$= (area\ under\ AB) + (area\ under\ BC) - (area\ under\ AC)$$

$$= \int_2^4 (x+3)dx + \int_4^6 \left(-\frac{5}{2}x + 17\right)dx - \int_2^6 \left(-\frac{3}{4}x + \frac{13}{2}\right)dx$$
$= 7 square\ units.$

If the function of a region is given in a parametric form, then we work as follows in order to calculate the area of that region using integral calculus: Let us consider the equations $x = g(t)$ and $y = f(t)$ where $t \in [t_1, t_2]$. If $g'(t) \neq 0$ for all $t \in (t_1, t_2)$, then the equations $x = g(t)$ and $y = f(t)$ define y as a function of x, and, if this is the case, then we apply the following rule: If y is a continuous function of x over the interval $[a, b]$ where $x = g(t)$ and $y = f(t)$, then the area of the region defined by y and the x-axis (along $b - a$) is

$$A = \int_a^b y dx = \int_{t_1}^{t_2} f(t)g'(t)dt$$

under the conditions that $g(t_1) = a$, $g(t_2) = b$, and the functions g' and f are continuous on $[t_1, t_2]$.

When a region is defined in polar coordinates, its area can be calculated by using integral calculus as follows: If a function f is continuous and non-negative over the interval $[a, b]$ with $0 \leq b - a \leq 2\pi$, then the area of the region bounded by $r = f(\varphi)$, $\varphi = a$, and $\varphi = b$ is given by the formula

$$A = \frac{1}{2}\int_a^b [f(\varphi)]^2 d\varphi$$

(regarding polar coordinates, see Chapter 6). In other words, if a curve's radius function can be expressed as a function $r(\varphi)$ of its angle with the positive side of the x-axis, then the area of the curve between two half-lines $\varphi = \alpha$ and $\varphi = \beta$ is $A = \frac{1}{2}\int_\alpha^\beta r^2 (\varphi)d\varphi$, because it is the summation of infinitely many infinitesimally small triangular pie wedges (sectors) such that: the arc length of the base of each triangular pie wedge is $rd\varphi$, the height of each triangular pie wedge is $r(\varphi) = r$, the apex angle of each pie wedge is $d\varphi$, and (using the formula of the area of a triangle: $\frac{1}{2} base \times height$) the area of each triangular pie wedge is (approximately) $A(d\varphi) = r^2 \frac{d\varphi}{2}$. In fact, the area of each triangular pie wedge is $\frac{1}{2} r^2 sind\varphi$, but, since $d\varphi$ is infinitesimally small, $sind\varphi \approx d\varphi$, and, thus, the area of each such small triangle is approximately $r^2 \frac{d\varphi}{2}$. Hence, in polar coordinates, the area of a circle of radius r ($r(\varphi) = r$) is

$\int_0^{2\pi} \frac{1}{2} r^2 d\varphi = \frac{1}{2} r^2 \int_0^{2\pi} d\varphi = r^2\pi$ (in essence, this is ancient Greek mathematicians' method of exhaustion formulated in modern mathematical language).

Moreover, if the functions f and g are continuous over the interval $[a, b]$, and if $0 \leq g(\varphi) \leq f(\varphi)$ for all $\varphi \in [a, b]$, then the area of the region bounded by $r = f(\varphi)$, $r = g(\varphi)$, $\varphi = a$, and $\varphi = b$ is given by the formula

$$A = \frac{1}{2} \int_a^b \{[f(\varphi)]^2 - [g(\varphi)]^2\} \, d\varphi$$

(regarding polar coordinates, see Chapter 6).

Finally, it is worth mentioning that integrals can be thought of as inner products on infinite-dimensional spaces. In fact, if $C[a, b]$ denotes the vector space of continuous functions on the interval $[a, b]$, then we obtain an inner product on $C[a, b]$ by defining, for all $f, g \in C[a, b]$,

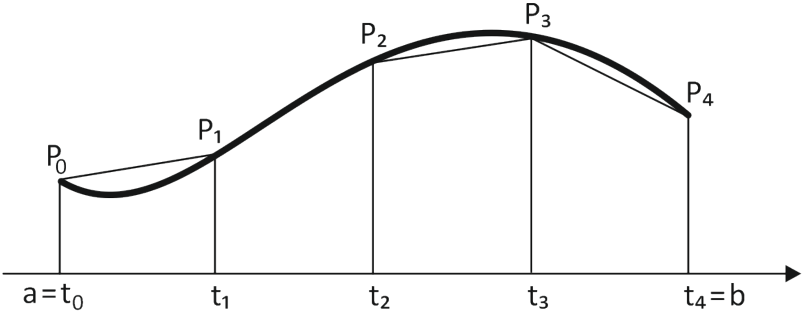
$$\langle f, g \rangle = \int_a^b f(x) g(x) dx$$

(notice that $\langle f, f \rangle = \int_a^b f(x) f(x) dx = \int_a^b [f(x)]^2 \, dx$, which gives the (signed) area between the graph of $y = [f(x)]^2$ and the x-axis from $x = a$ to $x = b$). An integral is a linear operator that takes one thing (specifically, a function) and returns a number; and an inner product is a bilinear operator that takes two things and returns a number, so that here we see that it is the integral of the product of two inputs.

2. *The Arc Length of a Curve*

Let us consider a curve γ defined by the parametric equations
$x = g(t)$ and $y = f(t)$ where $t \in [a, b]$,
as shown, for instance, in Figure 8-18, and let $P = \{t_0, t_1, \dots, t_n\}$ be a partition of $[a, b]$. Intuitively, if we regard parameter t as the time variable, then the curve may be thought of as the path of a moving point whose position vector at time t is $\gamma(t) = \big(g(t), f(t)\big)$.

Figure 8-18: The arc length of a curve.



Let $P_k = [g(t_k), f(t_k)]$ be the corresponding points of γ, as shown in Figure 8-18. Then these points define a polygonal line. The sum

$$L_P = \sum_{i=1}^{n} \sqrt{[g(t_i) - g(t_{i-1})]^2 + [f(t_i) - f(t_{i-1})]^2}$$

is the length of the polygonal line that is defined by the points P_k (corresponding to a partition P); and the finer the partition P, the more the corresponding polygonal line tends to be identified with the curve γ. Now, let us consider the set L of all the numbers L_P, which correspond to all possible partitions P of $[a, b]$, symbolically:
$L = \{L_P | P \ is \ a \ partition \ of \ [a, b]\}$.
If this set L is bounded, then the curve is said to be “rectifiable,” and the supremum $S = L(\gamma)$ of this set is said to be the length of the curve γ. Moreover, we write $S = L_a^b(\gamma)$ in order to denote the arc length of the curve that is defined on the interval $[a, b]$.
Notice that, if γ is a rectifiable curve on $[a, b]$, and if $a < c < b$, then
$L_a^b(\gamma) = L_a^c(\gamma) + L_c^b(\gamma)$.
Given a curve γ defined by the parametric equations
$x = g(t)$ and $y = f(t)$ where $t \in [a, b]$,
if the derivatives g' and f' are continuous on $[a, b]$, then the curve γ is rectifiable on $[a, b]$, and its length is given by

$$S = L(\gamma) = \int_a^b \sqrt{[g'(t)]^2 + [f'(t)]^2}\, dt$$

where $t \in [a, b]$.
If a curve γ is defined by $y = f(x)$, where $x \in [a, b]$, and if the derivative $f'(x)$ exists and is continuous on $[a, b]$, then, setting $x = t$ and $y = f(t)$ in the aforementioned formula, we obtain the following formula:

$$S = \int_a^b \sqrt{1 + [f'(x)]^2}\, dx$$

where $x \in [a, b]$.
If a curve γ is defined in polar coordinates $r = r(\varphi)$, $\varphi \in [\varphi_1, \varphi_2]$, then

$$S = \int_{\varphi_1}^{\varphi_2} \sqrt{r^2(\varphi) + [r'(\varphi)]^2}\, d\varphi$$

where $\varphi \in [\varphi_1, \varphi_2]$.

3. The volume of a solid of revolution

As shown in Figure 8-19, in order to obtain a solid of revolution, we start out with a curve $y = f(x)$ on an interval $[a, b]$, and then we rotate this curve (360°) about a given axis, so that a volume is generated. In order to determine the volume of a solid of revolution on the interval $[a, b]$, we work as follows: we divide the interval $[a, b]$ into n subintervals, each of which has width $\Delta x = \frac{b-a}{n}$, and then we choose a point ξ_k (where $k = 1, 2, \dots, n$) from each subinterval. When we want to determine the area between two curves, we approximate the area by using rectangles on each subinterval. Understandably, when we want to calculate the volume of a solid of revolution, we use discs on each subinterval to approximate the area. The area of the face of each disc is given by $A(\xi_k)$, and the volume of each disc is given by $V_k = A(\xi_k)\Delta x$, where Δx is the thickness of the disc. Hence, the volume of the corresponding solid of revolution on the interval $[a, b]$ can be approximated by $V \approx \sum_{k=1}^{n} A(\xi_k)\Delta x$. Then, its exact volume is

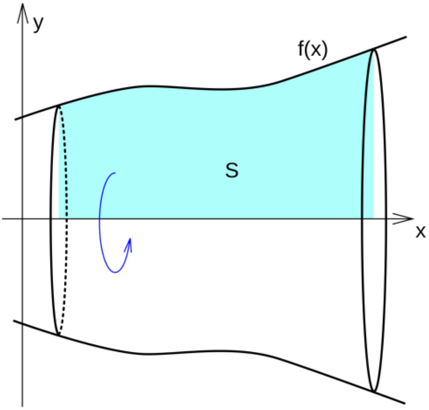
$$V = lim_{n \to \infty} \sum_{k=1}^{n} A(\xi_k)\Delta x = \int_a^b A(x)dx$$

where $a \le x \le b$.
In other words, in this case, the volume is the integral of the cross-sectional area$A(x)$ at any x, and $x \in [a, b]$. Given that $A = \pi r^2$, $r = f(x)$, and $f(x)$ is a non-negative continuous function from $[a, b]$ to $\mathbb{R}$, the volume of the solid generated by a region under $y = f(x)$ bounded by the x-axis and the vertical lines $x = a$ and $x = b$ via revolution about the x-axis is

$$V = \pi \int_a^b [f(x)]^2\, dx$$

(we take discs with respect to x, and $r = y = f(x)$; dx indicates that the area is rotated about the x-axis).

Figure 8-19: A solid of revolution (source: Wikimedia Commons: Author: Pajs; https://commons.wikimedia.org/wiki/File:Integral_apl_rot_objem1.svg).



If we rotate a curve about the y-axis, thus obtaining a cross-sectional area that is a function of y instead of x, then the aforementioned formula becomes

$$V = \int_c^d A(y)dy$$

where $c \leq y \leq d$. Given that, in this case, $A = \pi r^2$, and $r = f(y)$, the volume of the solid generated by a region under $x = f(y)$ bounded by the y-axis and the horizontal lines $y = c$ and $y = d$ via revolution about the y-axis is

$$V = \pi \int_c^d [f(y)]^2 \, dy$$

(we take discs with respect to y, and $r = x = f(y)$; dy indicates that the area is rotated about the y-axis).

If we have two curves y_1 and y_2 that enclose some area, and we rotate that area about the x-axis, then the volume of the solid formed is given by

$$V = \pi \int_a^b [(y_2)^2 - (y_1)^2] \, dx$$

where $y_1 = f(x)$, $y_2 = g(x)$, $x \in [a, b]$, and we assume that y_1 and y_2 are continuous on $[a, b]$, and $y_2 \geq y_1$ over $[a, b]$.

For instance, a sphere of radius r centered at the origin (0,0,0) can be generated by revolving the upper semicircular disc enclosed between the

x-axis and $x^2 + y^2 = r^2$ about the x-axis. If we revolve the semi-circle given by

$$y = f(x) = \sqrt{r^2 - x^2}$$

about the x-axis, we obtain a sphere of radius r. A cross-section of the sphere is a circle with radius $f(x)$ and area $\pi[f(x)]^2$. If we slice the sphere vertically into discs, then each disc has infinitesimal thickness dx, and the volume of each disc is approximately $\pi[f(x)]^2 dx$. If we add up the volumes of the discs, then we obtain the volume of the sphere—namely:

$$V = \pi \int_a^b [f(x)]^2\, dx = \pi \int_{-r}^{r} (r^2 - x^2) dx = \pi \left(r^2 x - \frac{x^3}{3}\right) |_{-r}^{r} = \frac{4}{3}\pi r^3.$$

Similarly, the volume of a cone can be calculated as follows: A cone with base radius r and height h can be formed by rotating a straight line through the origin $(0,0,0)$ about the x-axis. The slope of the straight line is $tan\theta = \frac{r}{h}$, so that the equation of the line is $y = \frac{r}{h}x$, and the limits of integration are $x = 0$ and $x = h$. Therefore, the volume of the corresponding cone is

$$V = \pi \int_0^h \left(\frac{r}{h}x\right)^2 dx = \frac{\pi r^2}{h^2}\left(\frac{x^3}{3}\right) |_0^h = \frac{1}{3}\pi r^2 h.$$

Similarly, the volume of a cylinder with base radius r and height h (assuming that the plane xOy is the cylinder's base plane) is $V = \pi \int_0^h r^2 dx = \pi r^2 h$,

since the volume of an infinitesimal circular strip of a cylinder having radius r and infinitesimally small height dx is $dv = area \times height = \pi r^2 dx$.

Following the same reasoning, the volume of a pyramid of height h with a $b \times b$ square base can be calculated using integration as follows: If y is the vertical distance from the top of the pyramid (placed at the origin of the rectangular coordinate system), then the square cross-sectional area $A(y)$ is given by $A(y) = \left(\frac{b}{h}y\right)^2 = \frac{b^2}{h^2}y^2$, and, hence, the volume of this pyramid is given by $\int_0^h A(y)dy = \frac{b^2}{h^2}\int_0^h y^2 dy = \frac{1}{3}b^2 h$.

4. *The physical significance of the definite integral and basic applications of integral calculus in mechanics*

The development of infinitesimal calculus by Newton and Leibniz is intimately related to the study of celestial mechanics (and physics in general) by them. Infinitesimal calculus, also known as the differentiation-integration method, is concerned with the limits of applicability of physical laws. The content of a physical law is not absolute, and the validity of a law is restricted to the framework of the applicability limits (i.e., certain conditions). However, a physical law can be expanded by changing its form beyond the limits of applicability by means of infinitesimal calculus. This method is based on the following two principles: (i) the principle that a law can be represented in a differential form, and (ii) the superposition principle, according to which the quantities that enter into the law are additive.

Suppose that a physical law has the form

$$X = YZ, \qquad (*)$$

where X, Y, and Z are physical quantities, and, in particular, Y is a constant representing the given law's limits of applicability. We can generalize the given law to the case where Y is not a constant but a function of Z, that is, $Y = Y(Z)$, as follows: As shown in Figure 8-20, we isolate an interval dZ so small that the variation of Z over this interval can be ignored. Hence, in the interval ("infinitesimal") dZ, we can approximately assume that Y is constant, and that the law $(*)$ is valid in this interval. Therefore, as shown in Figure 8-20,

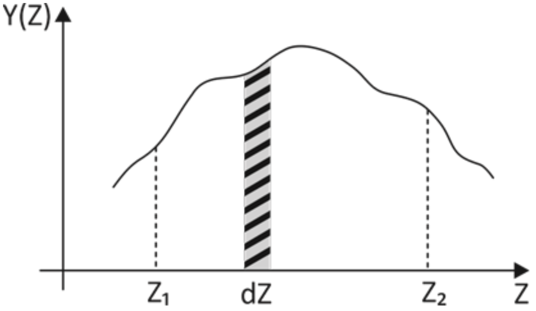
$$dX = Y(Z)dZ, \qquad (**)$$

where dX is the variation of X over dZ. Due to the superposition principle, that is, by summing the quantities $(**)$ over all the intervals of variation of Z, we obtain an expression for X in the form

$$X = \int_{Z_1}^{Z_2} Y(Z)dZ,$$

where Z_1 and Z_2 are the initial and the final values of Z, respectively, as shown in Figure 8-20.

Figure 8-20: The method of infinitesimal calculus.



As a conclusion, the method of infinitesimal calculus consists of two parts: in the first part of the method, we find the differential $(**)$ of the quantity under investigation; in the second part of the method, we sum, or "integrate," having properly determined the integration variable and the limits of integration (in order to determine the integration variable, we must analyze the quantities on which the differential of the investigated quantity depends and choose the most important variable; and the limits of integration are the lower and the upper values of the integration variable).

Eaxample 1: The work done by any force $F(x)$, assuming that $F(x)$ is continuous, over the range $a \leq x \leq b$ is

$$W = \int_a^b F(x)dx$$

(the force is parallel to the displacement). This formula can be proved, using the method infinitesimal calculus, as follows: We divide the range $[a, b]$ into n subintervals of width Δx, and, from each of these intervals, we choose the points $\xi_1, \xi_2, \ldots, \xi_n$. If n is large enough, and given that $F(x)$ is continuous, the variation of $F(x)$ over the ith interval ($i = 1,2, \ldots, n$) can be ignored, and we can assume that, over such an interval, the force is approximately constant, so that $F(x) \approx F(\xi_i)$. Thus, the work on each interval is approximately $W_i \approx F(\xi_i)\Delta x$, and then the total work over $[a, b]$ is approximately $W \approx \sum_{i=1}^{n} W_i = \sum_{i=1}^{n} F(\xi_i)\Delta x$. If we compute the limit of this summation as $n \to \infty$, then we shall get the exact work done, namely: $W = lim_{n\to\infty} \sum_{i=1}^{n} F(\xi_i)\Delta x$, which is the definition of the definite integral, and, hence, $W = \int_a^b F(x)dx$.

Example 2: Using the method of infinitesimal calculus, we can compute velocity from displacement, acceleration from velocity, displacement from velocity, and velocity from acceleration: Since, as I have already mentioned, the time derivative of the velocity function $v(t)$ is acceleration $a(t)$, that is,

$\frac{dv(t)}{dt} = a(t),$

we can integrate both sides to obtain

$\int \frac{dv(t)}{dt} dt = \int a(t)dt + c_1,$

where c_1 is a constant of integration. Since $\int \frac{dv(t)}{dt} dt = v(t)$, velocity is given by

$v(t) = \int a(t)dt + c_1.$

Additionally, as I have already mentioned, the time derivative of the position function $s(t)$ is the velocity function,

$\frac{ds(t)}{dt} = v(t),$

and, similarly, by integrating both sides, we obtain the displacement function

$s(t) = \int v(t)\,dt + c_2$,

where c_2 is another constant of integration.

Using these integrals, we can derive the three fundamental kinematic equations for a constant acceleration $a(t) = a$ as follows: Since

$a = \frac{dv}{dt} \Leftrightarrow dv = adt$,

integrating both sides with proper limits, we obtain

$\int_{v_0}^{v} dv = \int_0^t adt \Rightarrow v|_{v_0}^{v} = at|_0^t \Rightarrow v - v_0 = a(t-0) \Rightarrow v = v_0 + at$, (1)

where v_0 denotes initial velocity, v denotes final velocity, $t = 0$ denotes initial time, and t denotes final time. Moreover,

$v = \frac{ds}{dt} \Leftrightarrow ds = vdt$,

and, similarly, integrating both sides with proper limits, and using equation (1), we obtain

$\int_{s_0}^{s} ds = \int_0^t vdt \Rightarrow s|_{s_0}^{s} = \int_0^t (v_0 + at)dt \Rightarrow s|_{s_0}^{s} = \int_0^t v_0 dt +$

$\int_0^t atdt \Rightarrow s|_{s_0}^{s} = v_0 t|_0^t + a\frac{t^2}{2}|_0^t \Rightarrow s - s_0 = v_0(t-0) + \frac{a}{2}(t^2 - 0)$,

which ultimately yields

$s = s_0 + v_0 t + \frac{a}{2}t^2$, (2)

where s_0 denotes initial position, and s denotes final position. We can also write

$a = \frac{dv}{dt} = \frac{dv}{ds}\frac{ds}{dt} \Rightarrow a = v\frac{dv}{ds} \Leftrightarrow vdv = ads$,

and, similarly, integrating both sides with proper limits, we obtain

$\int_{v_0}^{v} vdv = \int_{s_0}^{s} ads \Rightarrow \frac{v^2}{2}|_{v_0}^{v} = as|_{s_0}^{s}$,

which ultimately yields

$v^2 = v_0^2 + 2a(s - s_0)$. (3)

The aforementioned kinematic equations refer to an object moving horizontally. Notice that, in case of strictly vertical motion, the only difference is that the acceleration will be the acceleration due to the Earth's gravity (i.e., $\approx -9.8\,m/sec^2$). In case of projectile motion, however, we deal with objects moving in both directions (i.e., moving along a curved path under the influence of gravity). For instance, consider a cannonball that is fired at some angle from the horizontal: it will travel some distance up into the air before eventually falling back down and hitting the ground, a distance away from the cannon. The path of this object (projectile) can be represented by a parabola. In projectile motion, the horizontal motion and the vertical motion are independent of each other, and, therefore, we use separate equations in order to study motion in each direction (one equation that corresponds to the x-coordinate of the object, and one equation that corresponds to the y-coordinate of the object). The time a projectile spends in the air relates only to its y-direction behavior, whereas the distance it travels from its initial position to its final position on the ground depends only on its x-direction behavior (horizontal velocity). Thus, in projectile motion, we use the above equations (equations (1), (2), and (3)), but we split up the velocity vector $\vec{v}$ into x and y components, i.e., $\vec{v}_x$ and $\vec{v}_y$ (the initial launch angle can be anywhere between 0 and 90 degrees). Notice that, in projectile motion, the horizontal velocity will be the same at every moment in the corresponding trajectory (as long as we ignore wind resistance), whereas the vertical velocity will be the greatest at the moment the projectile is launched, and then it will be decreasing until it reaches zero at the zenith, after which it will become increasingly negative until it hits the ground (since there is a constant acceleration in the negative direction due to gravity); and, of course, the angle at which the object is launched affects the range, the height, and the time of flight it will experience while in projectile motion.

Example 3: We shall use the method of infinitesimal calculus in order to find the "center of mass" or "centroid" of a thin plate with uniform density ρ. Given a homogeneous region, its center of mass is the average position of all the parts of the given system weighted according to their masses. An object with mass m and volume V has density $\rho = \frac{m}{V}$. Hence, given an object of constant cross-sectional area whose mass is distributed along a signle axis according to the function $\rho(x)$ (whose units are units of mass per unit of length), the total mass, M, of the given object between $x = a$ and $x = b$ is given by $M = \int_a^b \rho(x)dx$.

Assume that the plate under consideration is a region bounded by the curves $f(x)$ and $g(x)$ on the interval $[a, b]$. In order to find its center of mass, we work as follows: Firstly, we find the "total mass" of the plate, using the following formula:

$$M = \rho \times area\ of\ the\ plate = \rho \int_a^b [f(x) - g(x)]dx$$

(without loss of generality, we assume that the curve $f(x)$ is above the curve $g(x)$). Secondly, we find the two "moments" of the region, namely, M_x and M_y, which measure the tendency of the region to rotate about the x-axis and the y-axis, respectively. The two moments are given by:

$$M_x = \rho \int_a^b \frac{1}{2}\{[f(x)]^2 - [g(x)]^2\}dx$$

and

$$M_y = \rho \int_a^b x[f(x) - g(x)]\, dx$$

(*note:* in its most basic form, $Moment = Force \times Distance$, meaning that the magnitude of the moment of a force acting about a point or an axis is directly proportional to the distance of the force from the point or the axis, and it measures the tendency of a force to cause a body to rotate about a specific point or axis). Thirdly, we find the coordinates of the "center of mass" $(\bar{x}, \bar{y})$ by using the following formulae:

$$\bar{x} = \frac{M_y}{M}$$

and

$$\bar{y} = \frac{M_x}{M}$$

M, M_x, and M_y are defined as bove, that is: $M = \rho \int_a^b [f(x) - g(x)]dx$, $M_x = \rho \int_a^b \frac{1}{2}\{[f(x)]^2 - [g(x)]^2\}dx$, and $M_y = \rho \int_a^b x[f(x) - g(x)]\, dx$. Hence, we can write:

$$\bar{x} = \frac{1}{Q}\int_a^b x[f(x) - g(x)]dx$$

and

$$\bar{y} = \frac{1}{Q}\int_a^b \frac{1}{2}\{[f(x)]^2 - [g(x)]^2\}dx$$

where

$$Q = \int_a^b [f(x) - g(x)]dx$$

(the center of mass helps us to analyze how objects move and interact; for instance: when we are going to lift something with a crane, we have to center the lifting cable over the center of mass, or the center of mass of the load plus the counterweights must be within the crane's stabilizing struts, since otherwise the object will shift and tumble when we lift it, and it might fall, or the crane itself will tip over; the center of mass plays a critical role in designing cars in order to make sure that they have control and stability in dangerous conditions; and an airplane must be balanced around its its designed center of mass, since otherwise it may not fly correctly or may not fly at all).

5. *Basic applications of integral calculus in the social sciences*

In economics, the integral shows how to find total revenue, TR, from marginal revenue, MR, and how to find total cost, TC, from marginal cost,

MC. Since, MR can be defined as the derivative of TR with respect to quanty Q sold, that is,
$MR = \frac{dTR}{dQ}$,
it follows that, if we know the marginal revenue function $MR(Q)$, the total revenue is
$TR(Q) = \int MR(Q)dQ$.
Similarly, since MC can be defined as the derivative of TC with respect to quanty Q produced, that is,
$MC = \frac{dTC}{dQ}$,
it follows that, if we know the marginal cost function $MC(Q)$, the total cost is
$TC(Q) = \int MC(Q)\, dQ$.
Moreover, if $I(t)$ denotes the rate of investment (where t denotes time), then the total accumulation of capital K during the time interval $[t_1, t_2]$ is given by the formula

$$K = \int_{t_1}^{t_2} I(t)dt$$

(the business investment rate is defined as gross investment (gross fixed capital formation) divided by gross value added of non-financial corporations, and, thus, this ratio relates the investment of non-financial businesses in fixed assets (e.g., buildings, machinery, etc.) to the value added created during the production process).

Approximate integration: When the integrand $f(x)$ is known only at certain points (e.g., those obtained by sampling), or when a formula for the integrand is known but it is difficult or impossible to find an antiderivative that is an elementary function, we may use numerical methods of integration—that is, approximate formulae for definite integrals. The simplest approximate formula for definite integrals is
$\int_a^b f(x)dx \approx \frac{1}{2}(b-a)[f(a)+f(b)]$,
which is exact when $f(x)$ is linear. However, a much better approximate formula for definite integrals is
$\int_a^b f(x)dx \approx \frac{1}{6}(b-a)\left[f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right]$,
which is known as "Simpson's Rule," named after the eighteenth-century British mathematician Thomas Simpson, who formulated it. Before him, however, Johannes Kepler had already used similar formulae. For this reason, "Simpson's Rule" is sometimes called "Kepler's Rule." Simpson's Rule derives from the observation that, if $p(x) = Ax^2 + Bx + C$, then

$\int_a^b p(x)\,dx = \frac{b-a}{6}\left[p(a) + 4p\left(\frac{a+b}{2}\right) + p(b)\right]$, and it is used in order to approximate any integral $\int_a^b f(x)dx$, where f is an arbitrary function, and not necessarily a quadratic polynomial (i.e., a parabola).

Generalized integrals: A "generalized integral" (also known as an "improper integral") is an integral with one or more infinite limits of integration and/or discontinuous integrands (specifically, integrands with vertical asymptotes).
First Case:f is discontinuous at some points or at one point in the closed interval of integration $[a, b] \subset \mathbb{R}$.

i. If $f(x)$ is discontinuous at $x = x_0$, that is, if $f(x_0) \to \infty$ (and, thus, f has a vertical asymptote at this point), and $a < x_0 < b$, then
$\int_a^b f(x)dx = lim_{\varepsilon\to 0}\int_a^{x_0-\varepsilon} f(x)dx + lim_{\varepsilon\to 0}\int_{x_0+\varepsilon}^b f(x)dx.$

ii. If $f(x)$ is discontinuous at $x = x_0 = b$, that is, if $f(b) \to \infty$, then
$\int_a^b f(x)dx = lim_{\varepsilon\to 0}\int_a^{b-\varepsilon} f(x)dx$, or, equivalently,
$\int_a^b f(x)dx = lim_{k\to b^-}\int_a^k f(x)dx.$

iii. If $f(x)$ is discontinuous at $x = x_0 = a$, that is, if $f(a) \to \infty$, then
$\int_a^b f(x)dx = lim_{\varepsilon\to 0}\int_{a+\varepsilon}^b f(x)dx$, or, equivalently,
$\int_a^b f(x)dx = lim_{k\to a^+}\int_k^b f(x)dx.$

For instance, in $\int_0^n \frac{dx}{\sqrt{n^2-x^2}}$, the integrand is discontinuous at $x = n$, and, therefore,
$\int_0^n \frac{dx}{\sqrt{n^2-x^2}} = lim_{\varepsilon\to 0}\int_0^{n-\varepsilon}\frac{dx}{\sqrt{n^2-x^2}} = lim_{\varepsilon\to 0} arcsin\frac{x}{n}\Big|_0^{n-\varepsilon} =$
$lim_{\varepsilon\to 0}\left(arcsin\frac{n-\varepsilon}{n} - arcsin\frac{0}{n}\right) = lim_{\varepsilon\to 0} arcsin\frac{n-\varepsilon}{n} = arcsin 1 = \frac{\pi}{2}.$
Second Case: the interval of integration is infinite, that is, $(-\infty, b]$, $[a, +\infty)$, or $(-\infty, -\infty)$.

i. If $f(x)$ is continuous on (a, b) where $b = +\infty$, then
$\int_a^\infty f(x)dx = lim_{k\to\infty}\int_a^k f(x)dx.$

ii. If $f(x)$ is continuous on (a, b) where $a = -\infty$, then
$\int_{-\infty}^b f(x)dx = lim_{k\to -\infty}\int_k^b f(x)dx.$

iii. If $f(x)$ is continuous on (a, b) where $a = -\infty$ and $b = +\infty$, then
$\int_{-\infty}^\infty f(x)dx = lim_{k_1\to -\infty}\int_{k_1}^p f(x)dx + lim_{k_2\to\infty}\int_p^{k_2} f(x)dx$, where p is any number; in other words, $\int_{-\infty}^\infty f(x)dx = \int_{-\infty}^p f(x)\,dx + \int_p^\infty f(x)dx.$

For instance,
$\int_1^\infty \frac{1}{x^2}dx = lim_{k\to\infty}\int_1^k \frac{1}{x^2}dx = lim_{k\to\infty}\left(-\frac{1}{x}\right)|_1^k = lim_{k\to\infty}\left(1-\frac{1}{k}\right) = 1.$

Integration of Multivariable Functions

We can integrate functions of several variables as follows: suppose that the domain of a bivariate function is the Cartesian product of two closed intervals—that is, a rectangle—say
$R = [a,b]\times[c,d] = \{(x,y)\in\mathbb{R}^2 | a\le x\le b, c\le y\le d\}$.
If $R = [a,b]\times[c,d]$, whenever the integrand is $f(x,y)$, we have to integrate over two variables, x and y, so that, for each variable, we have an integration sign. In order to indicate the variables involved, we have dx and dy, symbolically:
$\iint_R f(x,y)\,dxdy \equiv \int_c^d\int_a^b f(x,y)dxdy,$
where $f(x,y)$ is an integrable function of two real variables. In this case, we compute the innermost integral first, and then we work our way outward. In particular, we compute the dx integral inside first, while treating y as a constant, and then we integrate the result over y as we would do with any variable. One interpretation of the double integral of $f(x,y)$ over the rectangle R is the volume under the function (surface) $f(x,y)$ and above the xy-plane.
For instance, $\int_0^2\int_0^1 x^2y^2dxdy$ can be calculated as follows: We focus on the inner integral first: $\int_0^2\left[\int_0^1 x^2y^2dx\right]dy$; and, treating y as a constant, we integrate normally for x^2dx, thus obtaining $\int_0^2\left[\frac{x^3y^2}{3}|_0^1\right]dy = \int_0^2\left[\frac{1^3y^2}{3}-\frac{0^3y^2}{3}\right]dy = \int_0^2\left[\frac{y^2}{3}\right]dy$. Now, we are left with an ordinary definite integral: $\int_0^2\frac{y^2}{3}dy = \frac{y^3}{3\cdot 3}|_0^2 = \frac{y^3}{9}|_0^2 = \frac{2^3}{9}-\frac{0^3}{9} = \frac{8}{9}$. Therefore, $\int_0^2\int_0^1 x^2y^2dxdy = \frac{8}{9}$.
Recall that ordinary integration, such as $\int_a^b f(x)dx$, gives us the area under the curve $y = f(x)$, above the x-axis, and between the lines $x = a$ and $x = b$; that's when f is a positive function (when f also takes negative values, we get a signed area). Double integration, such as $\int_c^d\int_a^b f(x,y)dxdy$, gives us the volume under the surface $z = f(x,y)$, above the xy-plane, and above the region described by the limits of integration (thus, we refer to this volume as the "volume under the surface"). The limits of integration in case of

$$\int_c^d \int_a^b f(x,y)dxdy$$

indicate that the corresponding region is the rectangle consisting of the points (x,y) such that $a \le x \le b$ and $c \le y \le d$; and the fact that dx is written before dy means that the function $f(x,y)$ is firstly integrated with respect to x (using the "inner" limits of integration a and b) and then the resulting function is integrated with respect to y (using the "outer" limits of integration c and d).

Double integrals can be used in order to compute areas, too. Recall that, if a region R is bounded from below by the curve $y = h_1(x)$ and bounded from above by the curve $y = h_2(x)$, and if $a \le x \le b$, then the area of R is given by

$A = \int_a^b [h_2(x) - h_1(x)]\,dx.$

However, we can obtain the same result using double integrals as follows:

$$A = \int_a^b \int_{h_1(x)}^{h_2(x)} dydx$$

(which gives the area of the same region R), since $\int_a^b \int_{h_1(x)}^{h_2(x)} dydx = \int_a^b \left(y|_{h_1(x)}^{h_2(x)}\right) dx = \int_a^b [h_2(x) - h_1(x)]\,dx.$

Therefore, the area A of a plane region $R = \{(x,y) \in \mathbb{R}^2 | a \le x \le b, h_1(x) \le y \le h_2(x)\}$ is given by

$$A = \iint_R dydx = \int_a^b \int_{h_1(x)}^{h_2(x)} dy\,dx = \int_a^b [h_2(x) - h_1(x)]\,dx$$

(i.e., R lies between two vertical lines and the graphs of two continuous functions $h_1(x)$ and $h_2(x)$).

Example 1: We can use double integrals in order to calculate the area between the curves $y = \frac{1}{2}x^2$ (which is a parabola that opens upward) and $y = 3x - x^2$ (which is a parabola that opens downward) as follows: Firstly, we have to find where these two curves meet by solving $\frac{1}{2}x^2 = 3x - x^2 \Rightarrow \frac{3}{2}x^2 = 3x \Rightarrow \frac{x^2}{2} = x \Rightarrow x^2 - 2x = 0 \Rightarrow x(x-2) = 0 \Rightarrow$ $(x = 0 \; or \; x = 2)$. Therefore, these two curves meet at $x = 0$ and at $x = 2$; and the given region (which is enclosed by these two curves) is bounded from above by $y = 3x - x^2$ and bounded from below by $y = \frac{1}{2}x^2$. Then the area of this region is given by

$A = \int_0^2 \int_{y=\frac{1}{2}x^2}^{y=3x-x^2} dydx = \int_0^2 \left(y|_{y=\frac{1}{2}x^2}^{y=3x-x^2}\right) dx = \int_0^2 \left(3x - x^2 - \frac{1}{2}x^2\right) dx = \int_0^2 \left(3x - \frac{3}{2}x^2\right) dx = \left(\frac{3x^2}{2} - \frac{3}{2}\frac{x^3}{3}\right)|_0^2 = 2 \; square \; units.$

Example 2: Now, we shall use double integrals in order to find the volume between the xy-plane and $z = 6 - 3x - 2y$ (i.e., z is the height function) above the unit square $R = \{(x, y)|0 \leq \mathrm{x} \leq 1, 0 \leq \mathrm{y} \leq 1\}$: $V = \int_0^1 \int_0^1 (6 - 3x - 2y)\, dydx = \int_0^1 [6y - 3xy - y^2]\, |_0^1 dx = \frac{7}{2}\ cubic\ units.$

In summary, it is important to understand and keep in mind the following:

- $\int_a^b dx$ represents length, specifically, $b - a$ (one could say that $\int_a^b dx$ is the area under the curve $f(x) = 1$ over the region of integration $[a, b]$).
- $\int_a^b f(x)\, dx$ represents area, specifically, it is the area of a curvilinear trapezoid bounded by the straight lines $y = 0$, $x = a$, and $x = b$ and by the graph of the function $y = f(x)$, assuming that $f(x)$ is continuous and non-negative on the interval $[a, b]$.
- $\int_c^d \int_a^b dxdy$ represents area associated with the region of integration $[a, b] \times [c, d]$.
- $\int_c^d \int_a^b f(x, y) dxdy$ represents the (three-dimensional) volume under the surface $z = f(x, y)$, above the xy-plane, and above the region described by the limits of integration, in the three-dimensional space (we assume that $f(x, y)$ is continuous and non-negative on the region of integration). *Remark:* For a function $f(x, y)$ that is continuous over a region of the type $R = \{(x, y) \in \mathbb{R}^2 | a \leq x \leq b, h_1(x) \leq y \leq h_2(x)\}$, we have $\iint_R f(x, y) dydx = \int_a^b \int_{h_1(x)}^{h_2(x)} f(x, y) dy\, dx$. For a function $f(x, y)$ that is continuous over a region of the type $R = \{(x, y) \in \mathbb{R}^2 | c \leq y \leq d, h_1(y) \leq x \leq h_2(y)\}$, we have $\iint_R f(x, y) dxdy = \int_c^d \int_{h_1(y)}^{h_2(y)} f(x, y)\, dxdy.$
- $\int_k^l \int_c^d \int_a^b dxdydz$ represents (three-dimensional) volume associated with the region of integration $[a, b] \times [c, d] \times [k, l]$, in the three dimensions (x, y, z). For instance, the volume of the tetrahedron bounded by the planes $x = 0$, $y = 0$, and $z = 0$, and by the equation $x + y + z = 1$ can be calculated using triple integrals as follows: In this case, the limits of integration can be determined as follows: (i) limits for z: $x + y + z = 1 \Rightarrow z = 1 - x - y$, and, therefore, z varies from 0 to $1 - x - y$; (ii) limits for y: $x + y = 1 \Rightarrow y = 1 - x$, and, therefore, y varies from 0 to $1 - x$; (iii) limits for x: x varies from 0 to 1. Hence, the required volume is given by

$$V = \int\int\int_R dzdydx$$
$$= \int_0^1 \int_0^{1-x} \int_0^{1-x-y} dzdydx$$
$$= \int_0^1 \int_0^{1-x} \left(z|_0^{1-x-y}\right) dydx$$
$$= \int_0^1 \int_0^{1-x} (1-x-y)\, dydx$$
$$= \int_0^1 \left[(1-x)(y|_0^{1-x}) - \left(\frac{y^2}{2}|_0^{1-x}\right)\right] dx$$
$$= \int_0^1 \left[(1-x)^2 - \frac{(1-x)^2}{2}\right] dx$$
$$= \int_0^1 \frac{(1-x)^2}{2} dx = \frac{1}{6}$$

(in cubic units).

- $\int_k^l \int_c^d \int_a^b f(x,y,z)dxdydz$ represents the four-dimensional hypervolume under the hypersurface $t = f(x,y,z)$, above the xyz-space, and above the region described by the limits of integration, in the four dimensions (x,y,z,t) (we assume that $f(x,y,z)$ is continuous and non-negative on the region of integration). For instance, in order to understand the physical significance of this triple integral, consider the following: $\int\int\int_R \rho(x,y,z)dxdydz$ is the total mass of a region R in space, where ρ is the density (i.e., mass per unit volume), which may vary from one point to another (R is the region occupied by the solid under consideration).

Of course, the area, the volume, and the hypervolume are usually taken to be signed, so that parts below the axis, or the plane, or the space, respectively, are negative, and those above are positive (however, integrating the absolute value of the function gives the unsigned corresponding quantity).

The order in which we do the integrations does not matter, provided that we keep track of the limits of integration of each variable. For instance, in the double integral $\int_c^d \int_a^b f(x,y)dxdy$, dx is associated with the x integrand, which runs from a to b, while dy is associated with the y integrand, which runs from c to d, and, therefore,

$$\int_c^d \int_a^b f(x,y)dxdy = \int_a^b \int_c^d f(x,y)dydx$$

(meaning that the limits of integration of each integrand remain the same). This result is known as Fubini's Theorem: given that a definite double integral can be thought of as a process of adding up all the infinitesimal elements of a Cartesian area $dxdy$ (imagine little rectangles) over the required region, thus obtaining the area of that region, the equality between the aforementioned two iterated integrals (i.e., Fubini's Theorem) can be thought of as an infinite version of the idea that addition is commutative and associative. By analogy, Fubini's Theorem applies to triple integrals, etc.

Increasing the number of integrals in the context of multiple integration is the same as increasing the number of dimensions, so that a single integral gives a two-dimensional area, a double integral gives a three-dimensional volume, a triple-integral gives a four-dimensional hypervolume, etc. In general, the multiple integral of a function $f(x_1, \dots, x_n)$ in n variables over a domain U is represented by n nested integral signs in the reverse order of computation (in the sense that the leftmost integral is computed last), followed by the function and the integrand arguments in such an order that indicates that the integral with respect to the rightmost argument is computed last; and the domain of integration is either represented symbolically for every argument over each integral sign or it is indicated by a characteristic letter (variable) at the rightmost integral sign:

$$\int \dots \int_U f(x_1, \dots, x_n) d\, x_1 \dots dx_n$$

$(x_1, \dots, x_n \in U)$. We take for granted the obvious generalizations of the theorems of integration to two or more variables.

Line integrals: Let C be a continuous curve. Then C is said to be "piecewise smooth" if it is a finite union of smooth curves. Curves are said to be "smooth" if they have no corners, or cusps, associated with them. In general, the "smoothness" of a function is a property measured by the number of continuous derivatives that a function has over its domain (e.g., a function is said to be of "differentiability class" C^k if it has a kth derivative that is continuous over its domain, and then it is also said to be of smoothness at least k).

A "line integral" (or "path integral," or "curve integral") helps us to calculate the area of a fence that lies above a piecewise smooth curve C and under the graph of a continuous non-negative function $f(x, y)$. In general, by a "line integral," we mean an integral where the function to be integrated is evaluated along a curve.

For instance, suppose that it has snowed, and there are snowbanks; at some spots, the snow is higher, and, at some other spots, the snow is less high.

You want to walk out into the snow shoveling as you move. The question is the following: how much snow do you have to actually shovel when you go out and walk? The answer to this question depends on the path you take (in particular, it depends on the length of the route you take and on the concentration of snow at various points along the route). The concept of a line integral captures this notion of snow accumulation along a path (simply put, we have some path and a function that gives us the "height" of snow above every point along that curve).

When the function to be integrated is a scalar field, the value of the line integral is the sum of values of the field at all points on the curve, weighted by some scalar function on the curve, commonly arc length. Thus, line integrals generalize definite integrals.

The line integral with respect to the arc length of a continuous function $f(x,y)$ along a piecewise smooth curve whose parametric expression is $c(t) = \big(x(t), y(t)\big)$, where $a \leq t \leq b$, is defined to be

$$\int_c f(x,y)ds = \int_a^b f\big(x(t), y(t)\big)\sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt$$

where ds represents an element of arc length along the curve (notice that, if we set $f(x,y) = 1$, then we obtain the formula for the calculation of the length of the curve c). Line integrals do not depend on the parametrization, as long as the curve is traversed once (counter-clockwise).

For instance, in order to compute the line integral $\int_c xy^2 ds$ where c is the right half of a circle with radius 2, that is, of the circle $x^2 + y^2 = 4$ (traversed once counter-clockwise), we work as follows: Firstly, we parametrize c, and, thus, we set

$c(t) = (2cost, 2sint)$, that is,

$x = 2cost$ and $y = sint$,

so that

$$ds = \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt = \sqrt{(-2sint)^2 + (2cost)^2}dt = 2dt.$$

Since t here represents the angle, and we have the right-hand side of the circle, our t will go from $-\pi/2$ to $+\pi/2$, that is, $-\pi/2 \leq t \leq \pi/2$.

Now, we are ready to make the corresponding substitutions in the line integral:

$$\int_c xy^2 ds = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} 2cost(2sint)^2 2dt = 8\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} sin^2 tcostdt$$

which can be calculated by setting $u = sint$ and $du = costdt$, thus obtaining

$8\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} u^2 du = 8\frac{u^3}{3}|_{-\frac{\pi}{2}}^{\frac{\pi}{2}} = \frac{8}{3}sin^3t|_{-\frac{\pi}{2}}^{\frac{\pi}{2}} = \frac{8}{3}\left[sin^3\left(\frac{\pi}{2}\right) - sin^3\left(-\frac{\pi}{2}\right)\right] =$
$\frac{8}{3}\left[sin^3\left(\frac{\pi}{2}\right) + sin^3\left(\frac{\pi}{2}\right)\right] = \frac{8}{3}(2) = \frac{16}{3}.$
By analogy, the line integral with respect to the arc length of a continuous function $f(x,y,z)$ along a piecewise smooth curve whose parametric expression is $c(t) = \big(x(t), y(t), z(t)\big)$, where $a \leq t \leq b$, is defined to be

$$\int_c f(x,y,z)ds = \int_a^b f\big(x(t), y(t), z(t)\big)\sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2}\,dt$$

etc.

Surface integrals: Recall that, whereas a curve in $\mathbb{R}^3$ is a set of points having an one-dimensional character, a surface is a set of points such that each point has two degrees of freedom. Moreover, a surface S can be represented in the following ways:
i. *Explicit representation:* S is the set of points $\{(x,y,z)\}$ such that
$z = f(x,y)$ for a smooth function f with domain U_{xy} in $\mathbb{R}^2$
or $y = g(x,z)$ for a smooth function g with domain V_{xz} in $\mathbb{R}^2$
or $x = h(y,z)$ for a smooth function h with domain W_{yz} in $\mathbb{R}^2$.
ii. *Implicit representation:* S is the set of points $\{(x,y,z)\ such\ that\ F(x,y,z) = 0\}$ where F is a smooth function on a domain D in $\mathbb{R}^3$.
iii. *Parametric representation:* S is the set of points $\{(x,y,z)\}$ such that

$$x = x(s,t)$$
$$y = y(s,t)$$
$$z = z(s,t)$$

where $a \leq s \leq b$, $c \leq t \leq d$, and the terms x, y, and z are smooth functions on the rectangle $[a,b] \times [c,d]$.
Surface integrals generalize double integrals to integrating over a surface that lies in an n-dimensional space. The double integral of a function of two real variables over a region D in $\mathbb{R}^2$ is written as $\iint_D f(x,y)\,dA$ or $\iint_D f(x,y)dxdy$, and these integrals can be evaluated as iterated single integrals, but we need a generalization similar to how line integrals generalize definite integrals. This need is satisfied by the concept of a surface integral. Whereas double integrals work when the region of integration is on a plane and, therefore, flat, surface integrals also work when the region of integration is not flat and, therefore, does not sit on a plane (in case of surface integrals, the region over which we integrate is an arbitrary smooth surface).

Recall that the concept of a line integral means that we integrate over a curve that has a range of movement in, for example, two dimensions, and, thus, our input curve being in two dimensions (i.e., parametrically defined by $\big(x(t), y(t)\big)$), we compute the surface area of something that looks like a fence or a curtain as it moves through three dimensions. However, in case of a surface integral, our surface is already a three-dimensional shape, and, thus, if we want to represent the function evaluated at some point on this shape, which exists in three dimensions, we require a fourth dimension in order to represent the corresponding "height." The "surface integral" of a scalar field (in this case, a function of three real variables) is written as follows:

$$\iint_S F(x,y,z)\,dS$$

where S is the surface over which the integral is evaluated, and dS is an element of S. This surface integral (i.e., the integral of a smooth scalar field $F(x,y,z)$ over a smooth surface S) can be calculated as follows:

$$\iint_S F(x,y,z)\,dS = \iint_{U_{xy}} F\big(x,y,f(x,y)\big)\sqrt{\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2 + 1dxdy}$$

in case the surface S is given by $z = f(x,y)$;

$$\iint_S F(x,y,z)\,dS = \iint_{V_{xz}} F(x,g(x,z),z)\sqrt{\left(\frac{\partial g}{\partial x}\right)^2 + \left(\frac{\partial g}{\partial z}\right)^2 + 1dxdz}$$

in case the surface S is given by $y = g(x,z)$;

$$\iint_S F(x,y,z)\,dS = \iint_{W_{yz}} F(h(y,z),y,z)\sqrt{\left(\frac{\partial h}{\partial y}\right)^2 + \left(\frac{\partial h}{\partial z}\right)^2 + 1dydz}$$

in case the surface S is given by $x = h(y,z)$. Notice that, if we set $F(x,y,z) = 1$, then the surface integral yields the exact surface area of S, that is: $\int\int_{U_{xy}}\sqrt{\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2 + 1dxdy}$ is the surface area of the surface $z = f(x,y)$ over the region U_{xy}, $\int\int_{V_{xz}}\sqrt{\left(\frac{\partial g}{\partial x}\right)^2 + \left(\frac{\partial g}{\partial z}\right)^2 + 1dxdz}$ is the surface area of the surface $y = g(x,z)$ over the region V_{xz}, and $\int\int_{W_{yz}}\sqrt{\left(\frac{\partial h}{\partial y}\right)^2 + \left(\frac{\partial h}{\partial z}\right)^2 + 1dydz}$ is the surface area of the surface $x = h(y,z)$ over the region W_{yz}.

Numerical approximations of multiple integrals: Frequently, when we have to integrate multivariable functions, we cannot calculate multiple integrals exactly, and, therefore, we approximate their values by applying numerical methods, which are based on the concept of the average value of a function. By analogy with single variable calculus, the "average value" of a bivariate function $f(x, y)$ over a region R is

$$\bar{f} = \frac{1}{A(R)} \int \int_R f(x, y)\, dA$$

where $A(R)$ is the area of the region R; and, thus,

$$\int \int_R f(x, y)\, dA = \bar{f} A(R)$$

where $\bar{f}$ over the region R represents the sum of all the values of $f(x, y)$ divided by the number of points in R, and, because there are infinitely many points in every region, we need an approximation method that will be based on determining a very large number N of random points in the region R (which can be generated by a computer), calculating the average value of f for those points, and using that average value as the value of $\bar{f}$ in the above formula. This is the so-called Monte Carlo method. Hence, we obtain the following approximation formula:

$$\int \int_R f(x, y)\, dA \approx A(R)\bar{f} \pm A(R) \sqrt{\frac{\bar{f^2} - \left(\bar{f}\right)^2}{N}}$$

where

$$\bar{f} = \frac{\sum_{i=1}^{N} f(x_i, y_i)}{N}$$

and

$$\bar{f^2} = \frac{\sum_{i=1}^{N} \left(f(x_i, y_i)\right)^2}{N}$$

(the sums are taken over N random points $(x_1, y_1), \dots, (x_N, y_N)$, and the $\pm$ "error term" in the above approximation formula represents a single standard deviation from the expected value of the integral).

Similarly, the average value of a trivariate function $f(x, y, z)$ over a solid S is

$$\bar{f} = \frac{1}{V(S)} \int \int \int_S f(x, y, z) dV$$

where dV is the volume of the solid S; etc.

Differentiation and Integration of Vector-Valued Functions

When a function takes a real number and sends it to a vector, then it is said to be a vector-valued function. In the real plane, or in the xy-plane, the general form of a vector-valued function is the following:

$$\vec{r}(t) = f(t)\hat{\imath} + g(t)\hat{\jmath}; \tag{1}$$

and, in the real 3-dimensional space, or in the xyz-space, the general form of a vector-valued function is the following:

$$\vec{r}(t) = f(t)\hat{\imath} + g(t)\hat{\jmath} + h(t)\hat{k}; \tag{2}$$

where the component functions f, g, and h are real-valued functions of the parameter t, and $\hat{\imath}$, $\hat{\jmath}$, and $\hat{k}$ are the corresponding unit vectors on the x-axis, the y-axis, and the z-axis, respectively. The standard unit vectors in the direction of the x, the y, and the z axes of a 3-dimensional Cartesian coordinate system are

$$\hat{\imath} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \hat{\jmath} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \text{ and } \hat{k} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

The “limit” of a vector-valued function $\vec{r}(t)$ is $\vec{L}$ as t tends to a, symbolically:

$$lim_{t\to a}\vec{r}(t) = \vec{L}$$

if and only if

$$lim_{t\to a}\left\|\vec{r}(t) - \vec{L}\right\| = 0.$$

Therefore, (1) implies that

$$lim_{t\to a}\vec{r}(t) = [lim_{t\to a}f(t)]\hat{\imath} + [lim_{t\to a}g(t)]\hat{\jmath},$$

and (2) implies that

$$lim_{t\to a}\vec{r}(t) = [lim_{t\to a}f(t)]\hat{\imath} + [lim_{t\to a}g(t)]\hat{\jmath} + [lim_{t\to a}h(t)]\hat{k},$$

provided that the limits of the component functions f, g, and h as $t \to a$ exist. Similarly, we can define the limit of a vector-valued function of n component functions for $n > 3$.

A vector-valued function $\vec{r}(t)$, where $t \in [a, b]$, is said to be “continuous” at a point $t_0 \in [a, b]$ if and only if $lim_{t\to t_0}\vec{r}(t) = \vec{r}(t_0)$; and $\vec{r}(t)$ is said to be continuous on $[a, b]$ if and only if it is continuous at every point of $[a, b]$.

The derivative of a vector-valued function $\vec{r}(t)$, where $t \in [a, b]$, is defined as follows:

$$\overrightarrow{r'}(t) \equiv \frac{d\vec{r}(t)}{dt} = lim_{\Delta t\to 0}\frac{\vec{r}(t + \Delta t) - \vec{r}(t)}{\Delta t}$$

provided that the limit exists. If $\overrightarrow{r'}(t)$ exists, then $\vec{r}(t)$ is said to be differentiable at t. If $\overrightarrow{r'}(t)$ exists $\forall t \in (a, b)$, then $\vec{r}(t)$ is said to be differentiable on the interval (a, b). In order for $\vec{r}(t)$ to be differentiable on $[a, b]$, $\vec{r}(t)$ must be differentiable on the interval (a, b), and the following two limits must exist as well:

$\overrightarrow{r'}(a) = lim_{\Delta t \to 0^+} \frac{\vec{r}(a+\Delta t)-\vec{r}(a)}{\Delta t}$ and
$\overrightarrow{r'}(b) = lim_{\Delta t \to 0^-} \frac{\vec{r}(b+\Delta t)-\vec{r}(b)}{\Delta t}$.
Consequently, (1) implies that
$\overrightarrow{r'}(t) = f'(t)\hat{\imath} + g'(t)\hat{\jmath}$,
and (2) implies that
$\overrightarrow{r'}(t) = f'(t)\hat{\imath} + g'(t)\hat{\jmath} + h'(t)\hat{k}$,
where $\hat{\imath}$, $\hat{\jmath}$, and $\hat{k}$ are the unit vectors for the x-axis, the y-axis, and the z-axis, respectively.
The properties of the derivative of a vector-valued function are analogous to those of the derivative of a scalar-valued function.
If C is a smooth curve represented by the vector-valued function $\vec{r}(t)$ on some interval I, then the "unit tangent vector" $\vec{T}$(t) at t can be found by taking the derivative of $\vec{r}(t)$ and then normalizing it (i.e., we divide it by its magnitude in order to become a *unit* vector); symbolically:

$$\vec{T}(t) = \frac{\overrightarrow{r'}(t)}{\left\|\overrightarrow{r'}(t)\right\|}$$

(i.e., the normalized derivative; and, in physical terms, this is the normalized velocity vector). Notice that the smoothness of the curve C guarantees that $\overrightarrow{r'}(t) \neq \vec{0}$.
In general, the equation of the tangent line to a curve C at the point $\vec{r}(t_0)$ is given by the formula
$\vec{R} = \vec{r}(t_0) + k\overrightarrow{r'}(t_0)$, where $k \in \mathbb{R}$.
The angle of intersection between two curves is the angle of intersection between their tangent vectors. Hence, given two curves C_1 and C_2 represented by the vector-valued functions $\vec{r}(u)$ and $\vec{s}(v)$, respectively, the angle of intersection between them at the point that corresponds to the parameter values u_0 and v_0 is given by the formula

$$cos\omega = \frac{\overrightarrow{r'}(u_0) \cdot \overrightarrow{s'}(v_0)}{\left|\overrightarrow{r'}(u_0)\right|\left|\overrightarrow{s'}(v_0)\right|}$$

(ω is the required angle).
A vector-valued function $\vec{r}(t)$ is perpendicular to its derivative $\overrightarrow{r'}(t)$ if and only if the magnitude of $\vec{r}(t)$ is constant; symbolically:

$$\|\vec{r}(t)\| = constant$$

(intuitively, the magnitude of a vector in space with tail at the origin and head moving changes if and only if the direction of its motion is not orthogonal to itself). Indeed, notice that a vector of constant magnitude is orthogonal to its derivative because $\vec{r}(t) \cdot \vec{r}(t) = \|\vec{r}(t)\|^2 = \vec{c}$ for some constant vector $\vec{c}$ independent of t, and, therefore, differentiating with respect to t, we obtain $2\vec{r}(t)\cdot\overrightarrow{r'}(t)=\vec{0}$.

A vector-valued function $\vec{r}(t)\neq\vec{0}$ has a "constant direction" if and only if the cross product

$$\vec{r}(t)\times\frac{d\vec{r}(t)}{dt}=\vec{0}$$

(i.e., if and only if $\vec{r}(t)$ is parallel to its derivative). We can prove this theorem as follows: Let $\hat{r}$ be the unit vector in the direction of vector $\vec{r}$. Then, by the definition of a unit vector, $\vec{r}=\|\vec{r}\|\hat{r}$, where $\vec{r}$ has a constant direction, and so $\hat{r}$ has also a constant direction. Thus,

$$\vec{r}=\|\vec{r}\|\hat{r}\Rightarrow\frac{d\vec{r}}{dt}=\|\vec{r}\|\frac{d\hat{r}}{dt}+\frac{d\|\vec{r}\|}{dt}\hat{r}\Rightarrow\vec{r}\times\frac{d\vec{r}}{dt}=\vec{r}\times\left(\|\vec{r}\|\frac{d\hat{r}}{dt}+\frac{d\|\vec{r}\|}{dt}\hat{r}\right)$$

$$=\|\vec{r}\|\hat{r}\times\left(\|\vec{r}\|\frac{d\hat{r}}{dt}+\frac{d\|\vec{r}\|}{dt}\hat{r}\right)$$

$$=\|\vec{r}\|^2\hat{r}\times\frac{d\hat{r}}{dt}+\|\vec{r}\|\frac{d\|\vec{r}\|}{dt}\hat{r}\times\hat{r}=\|\vec{r}\|^2\hat{r}\times\frac{d\hat{r}}{dt}+\vec{0}$$

since $\hat{r}\times\hat{r}=\vec{0}$. This means that

$$\vec{r}\times\frac{d\vec{r}}{dt}=\|\vec{r}\|^2\hat{r}\times\frac{d\hat{r}}{dt}$$

and, since $\hat{r}$ also has a constant direction,

$$\frac{d\hat{r}}{dt}=\vec{0}$$

we obtain

$$\vec{r}\times\frac{d\vec{r}}{dt}=\vec{0}$$

as required (therefore, the condition is necessary). Conversely, suppose:

$$\vec{r}(t)\times\frac{d\vec{r}(t)}{dt}=\vec{0}\Rightarrow\|\vec{r}\|\hat{r}\times\frac{d}{dt}(\|\vec{r}\|\hat{r})=\vec{0}\Rightarrow\|\vec{r}\|\hat{r}\times\|\vec{r}\|\frac{d\hat{r}}{dt}=\vec{0}$$

$$\Rightarrow\|\vec{r}\|^2\left(\hat{r}\times\frac{d\hat{r}}{dt}\right)=\vec{0}\Rightarrow\hat{r}\times\frac{d\hat{r}}{dt}=\vec{0}$$

since $\|\vec{r}\|^2\neq\vec{0}$. Moreover, since $\hat{r}$ is a unit vector of constant length,

$$\hat{r}\cdot\frac{d\hat{r}}{dt}=\vec{0}$$

(i.e., this dot product is equal to zero, as we explained in the previous theorem). Hence, we have:

$$\hat{r}\times\frac{d\hat{r}}{dt}=\vec{0}=\hat{r}\cdot\frac{d\hat{r}}{dt}\Rightarrow\frac{d\hat{r}}{dt}=\vec{0}$$

which implies that the unit vector $\hat{r}$ is of constant direction, and, therefore, the vector $\vec{r}$ (where $\vec{r}=\|\vec{r}\|\hat{r}$) is also of constant direction, as required (therefore, the condition is sufficient); *quod erat demonstrandum*.

Recall that

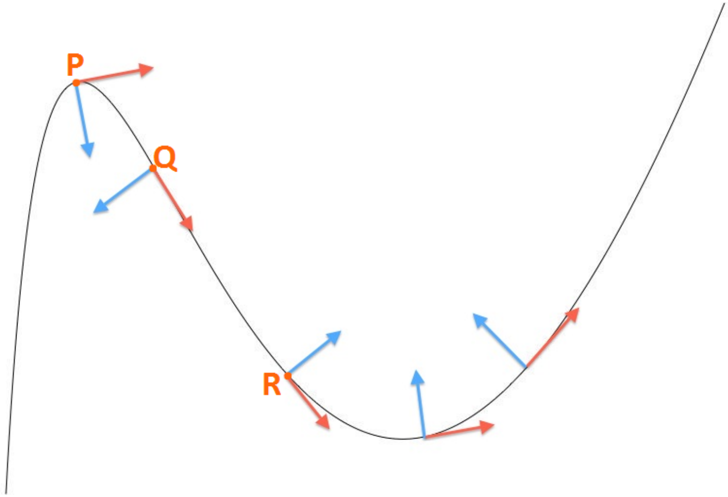
$$\vec{T}(t) \cdot \vec{T}(t) = \left\|\vec{T}(t)\right\|^2$$

(here $\vec{T}(t)$ is a unit vector, namely, the unit tangent vector, as above, but this formula holds in general), and, since $\vec{T}(t)$ is a unit vector, its magnitude, namely, $\left\|\vec{T}(t)\right\|$, is equal to 1, and, therefore, $\left\|\vec{T}(t)\right\|^2 = 1$. Therefore, $\vec{T}(t) \cdot \vec{T}(t) = constant$, and then, as I have already proved, $\vec{T}(t) \cdot \overrightarrow{T'}(t) = \vec{0}$. As I have already mentioned, when the dot product of any two vectors is equal to zero, these vectors are perpendicular (or orthogonal) to each other (Chapter 7). In this case, $\vec{T}(t) \perp \overrightarrow{T'}(t)$ (where the symbol $\perp$ means "perpendicular"). If we normalize this $\overrightarrow{T'}(t)$, then we obtain the "principal unit normal vector":

$$\vec{N}(t) = \frac{\overrightarrow{T'}(t)}{\left\|\overrightarrow{T'}(t)\right\|}$$

(where $\vec{T}(\mathrm{t})$ is the unit tangent vector $\vec{T}(\mathrm{t})$ at t on the smooth curve C represented by the vector-valued function $\vec{r}(t)$ on some interval I). Notice that the principal unit normal vector points in the direction in which the curve is curving, as shown, for instance, in Figure 8-21. Once you know a tangent vector (a, b), there are two obvious vectors that are normal (i.e., perpendicular) to (a, b), namely, $(b, -a)$ and $(-b, a)$; so that, if you pick the one that points in the direction in which the curve is curving and you divide it by its magnitude (norm), then you have the principal unit normal vector.

Figure 8-21: The unit tangent vector and the principal unit normal vector (source: Wikimedia Commons: Author: https://math.libretexts.org; https://commons.wikimedia.org/wiki/File:Curvatura_PQR.png).



Let f, g, and h be integrable real-valued functions on $[a, b]$. Then (1) implies that the indefinite integral of a vector-valued function $\vec{r}(t) = f(t)\hat{\imath} + g(t)\hat{\jmath}$ is

$$\int [f(t)\hat{\imath} + g(t)\hat{\jmath}]\, dt = [\int f(t)dt]\hat{\imath} + [\int g(t)dt]\hat{\jmath},$$

and the definite integral of a vector-valued function $\vec{r}(t) = f(t)\hat{\imath} + g(t)\hat{\jmath}$ is

$$\int_a^b [f(t)\hat{\imath} + g(t)\hat{\jmath}]dt = \left[\int_a^b f(t)dt\right]\hat{\imath} + \left[\int_a^b g(t)dt\right]\hat{\jmath}.$$

By analogy, (2) implies that

$$\int \left[f(t)\hat{\imath} + g(t)\hat{\jmath} + h(t)\hat{k}\right] dt = [\int f(t)dt]\hat{\imath} + [\int g(t)dt]\hat{\jmath} + [\int h(t)dt]\hat{k},$$

and

$$\int_a^b \left[f(t)\hat{\imath} + g(t)\hat{\jmath} + h(t)\hat{k}\right]dt = \left[\int_a^b f(t)dt\right]\hat{\imath} + \left[\int_a^b g(t)dt\right]\hat{\jmath} + \left[\int_a^b h(t)dt\right]\hat{k}.$$

The properties of the integral of a vector-valued function are analogous to those of the integral of a scalar-valued function.

Differential operators and their applications in physics: Let us consider a function $f(x, y)$; f depends on both x and y, and its graph is a surface in space. Then, in order to interpret and compute the rate of change of $f(x, y)$, we find the rate of change of $f(x, y)$ in a specific direction independently. If we want the rate of change in the x-direction, then we differentiate $f(x, y)$ with respect to x while treating y as a constant. In other words, we compute the partial derivative $\frac{\partial f(x,y)}{\partial x}$. Similarly, if we want the rate of change in the y-direction, then we differentiate $f(x, y)$ with respect to y while treating x as a constant. In other words, we compute the partial derivative $\frac{\partial f(x,y)}{\partial y}$. The "gradient" of $f(x, y)$ is denoted by ∇f, and it is a concept that combines the aforementioned two partial derivatives; specifically, the gradient of $f(x, y)$ is a vector consisting of both partial derivatives of f in their associated positions, symbolically:

$$gradf \equiv \nabla f = \frac{\partial f(x,y)}{\partial x}\hat{\imath} + \frac{\partial f(x,y)}{\partial y}\hat{\jmath}$$

(where $\hat{\imath}$ is the unit vector in the x-direction, and $\hat{\jmath}$ is the unit vector in the y-direction). By analogy, we can define the gradient of a function $f(x, y, z)$, etc. If you draw a little disc on the surface at the point you want to find the gradient, then the axis of the disc is normal (i.e., perpendicular) to the plane of the disc, and, therefore, it is also normal to the corresponding surface; and, in fact, the axis of the disc is said to be the gradient vector of the corresponding surface at the given point. Therefore, the "outward unit normal vector" to a given surface defined by a function f at a given point P (on this surface) is

$$\frac{\nabla f(P)}{\|\nabla f(P)\|}$$

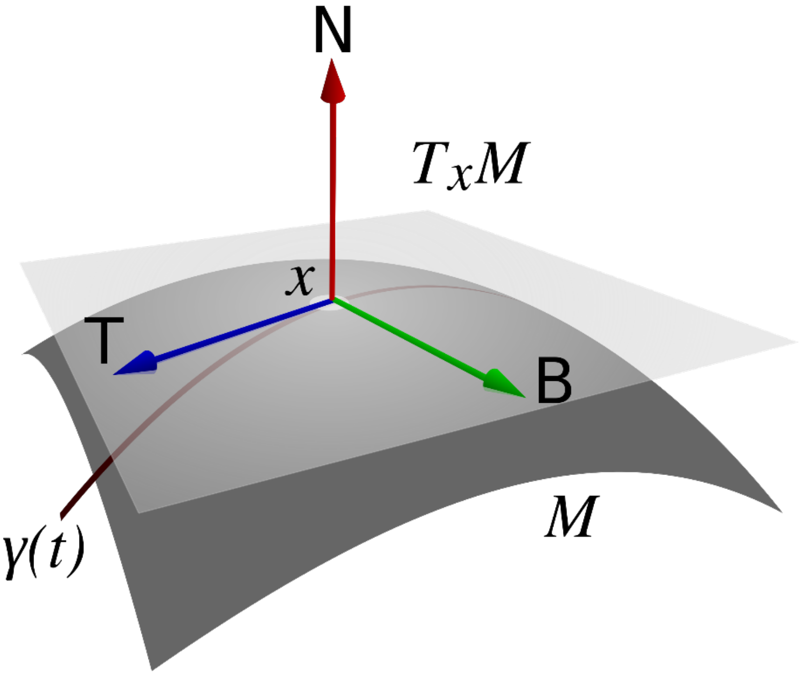
(i.e., we normalize the gradient at P in order to turn it into a unit vector). In general, a normal vector is a vector that points directly away from the corresponding plane, and, thus, if we know the normal vector, we know the orientation of the corresponding plane. If we have a normal vector $\vec{n}$ emanating from a fixed point $P_0(x_0, y_0, z_0)$ on a plane, then there exists a vector that emanates from the same point $P_0(x_0, y_0, z_0)$ and terminates at another point $P(x, y, z)$ lying in this plane. Obviously, the normal vector $\vec{n}$ is orthogonal to the vector $\overrightarrow{P_0P}$ (for any terminal point $P(x, y, z)$, since $\vec{n}$ is orthogonal to every vector that lives in this plane), and, therefore, their dot product is equal to zero. Consequently, in vector notation, the formula of a plane, in general, can be written as follows:

$\vec{n} \cdot \overrightarrow{P_0P} = 0,$

that is, by setting the dot product between the normal vector $\vec{n}$ and the generic vector $\overrightarrow{P_0P}$ that lies in the plane (and emanates from the point $P_0(x_0, y_0, z_0)$ and terminates at the point $P(x, y, z)$) equal to zero (see Figure 8-22). Furthermore, normal vectors help us to find the equations of tangent planes to surfaces at given points, as shown in Figure 8-22.

If the components of the normal vector are a, b, and c, that is, if $\vec{n} = \langle a, b, c \rangle$, then we can expand the aforementioned formula of a plane as follows:
$a(x - x_0) + b(y - y_0) + c(z - z_0) = 0$,
and we make the simplifying assumption that $c = -1$, so that we obtain
$a(x - x_0) + b(y - y_0) + z_0 = z$,
which is a linear function $z = z(x, y)$, where z represents "height" and depends on x and y in a linear way.

Figure 8-22: Tangent plane to a surface at a point (source: Wikimedia Commons: Author: A2569875; https://commons.wikimedia.org/wiki/File:Vertex_tangent,_bitangent_and_normal_vector.svg).



We can use the gradient vector for a function in order to find the tangent plane equation for the function at a particular point. If we have a function in two variables and the gradient vector is $\nabla f(x, y) = \langle \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \rangle$, where

(x,y) denotes the point in which we are interested, and if the result of evaluating the gradient vector at the point (x,y) is
$\nabla f(x,y) = \langle a,b\rangle$,
then a and b represent the slope of the original function f in the x and the y directions, respectively. Hence, the equation of the tangent plane at a particular point $P(x_0,y_0)$ can be found by substituting the corresponding values a and b together with the point $P(x_0,y_0)$ into the equation of the tangent plane

$$a(x-x_0)+b(y-y_0)-(z-z_0)=0$$

where a and b come from $\nabla f(x,y)=\langle a,b\rangle$, x_0 and y_0 are the coordinates of the given point P, and z_0 is obtained by substituting $P(x_0,y_0)$ into $f(x,y)$. In other words, the equation of the tangent plane to the surface $z=f(x,y)$ at the point $\big(x_0,y_0,f(x_0,y_0)\big)$ is

$$\frac{\partial f(x_0,y_0)}{\partial x}(x-x_0)+\frac{\partial f(x_0,y_0)}{\partial y}(y-y_0)-z+f(x_0,y_0)=0$$

(as above).

For instance, let us use the gradient vector in order to find the equation of the tangent plane to the surface $x^4-5x^3y-y^2+3y^4=6$ at the point $P(3{,}4)$. The corresponding function is
$f(x,y)=x^4-5x^3y-y^2+3y^4-6$,
and, thus, $\frac{\partial f}{\partial x}=4x^3-15x^2y$, and $\frac{\partial f}{\partial y}=-5x^3-2y+12y^3$. Then the gradient vector is
$\nabla f(x,y)=\langle 4x^3-15x^2y,-5x^3-2y+12y^3\rangle$,
and its value at the point $P(3{,}4)$ is
$\nabla f(3{,}4)=\langle -432,625\rangle$,
which is the vector that is normal to the curve at the point $P(3{,}4)$; and, in order to find the equation of the tangent plane to $f(x,y)$ at this point, we work as follows: The equation of the tangent plane is
$a(x-x_0)+b(y-y_0)-(z-z_0)=0$,
and, in this case, $a=-432$, $b=625$, $x_0=3$, and $y_0=4$, so that we obtain
$-432(x-3)+625(y-4)-(z-z_0)=0 \Rightarrow z-z_0=-432x+625y-1{,}204$.
In order to find z_0, we have to substitute $P(3{,}4)$ into $f(x,y)$, and, thus, we obtain $f(3{,}4)=287$. Therefore, setting $z_0=287$ in $z-z_0=-432x+625y-1{,}204$, we obtain the equation of the tangent plane to the surface $x^4-5x^3y-y^2+3y^4=6$ at the point $P(3{,}4)$, namely:
$z=-432x+625y-917$.

By analogy, if a surface is defined implicitly by an equation of the form $F(x, y, z) = 0$, then the tangent plane to the surface at a point (x_0, y_0, z_0) is given by the following equation:

$$\frac{\partial F(x_0, y_0, z_0)}{\partial x}(x - x_0) + \frac{\partial F(x_0, y_0, z_0)}{\partial y}(y - y_0) + \frac{\partial F(x_0, y_0, z_0)}{\partial z}(z - z_0) = 0$$

(recall that the equation of a plane that contains the point (x_0, y_0, z_0) with normal vector $\vec{n} = \langle a, b, c\rangle$ is $a(x - x_0) + b(y - y_0) + c(z - z_0) = 0$). Thus, we observe that the equation $\frac{\partial f(x_0, y_0)}{\partial x}(x - x_0) + \frac{\partial f(x_0, y_0)}{\partial y}(y - y_0) - z + f(x_0, y_0) = 0$ (which we formulated previously) is the special case of the equation $\frac{\partial F(x_0, y_0, z_0)}{\partial x}(x - x_0) + \frac{\partial F(x_0, y_0, z_0)}{\partial y}(y - y_0) + \frac{\partial F(x_0, y_0, z_0)}{\partial z}(z - z_0) = 0$ where $F(x, y, z) = f(x, y) - z$, and $z_0 = f(x_0, y_0)$.

Since the gradient of a function is given by the vector field whose components are the partial derivatives of the function (and, thus, the gradient attaches a vector to each point of the domain of the corresponding function), we can, more precisely, use the term “gradient vector field” rather than simply “gradient” or “gradient vector.”

In physics, the term “field” refers to an area in which forces are exerted on things in its midst. The modern concept of a physical field was originally formulated in the nineteenth century by the English physicist Michael Faraday. An electric charge creates an “electric field” in the region of space surrounding it, in the sense that the properties of space are modified by the presence of an electric charge. “Electric field” (sometimes called “electric intensity”) is defined as the electric force per unit charge. In particular, the “electric field” is a vector field that associates to each point in space the force per unit of charge exerted on an infinitesimal test charge at rest at the given point. Therefore (in SI unites), the unit of electric field magnitude is one newton per coulomb (i.e., $1N \cdot C^{-1}$). According to Coulomb’s Law, the magnitude of the force of interaction between two point charges (i.e., electric charges) is directly proportional to the product of the charges and inversely proportional to the square of the distance between them, symbolically:

$$F = k\frac{|q_1 q_2|}{r^2}$$

where F denotes the magnitude of the force that each of two point charges q_1 and q_2 a distance r apart exerts on the other, and k is a proportionality constant, whose value is (in SI units) approximately $8.988 \times 10^9 N \cdot m^2 \cdot$

C^{-2}. Due to the rigorous description of the electrostatic force of attraction and repulsion by the French military engineer and physicist Charles-Augustin de Coulomb (1736–1806), the SI unit of electric charge, the coulomb (denoted by C), has been named in his honor; it is approximately equivalent to 6.24×10^{18} electrons. "Charge" is a property of matter (just like mass, volume, or density), and it can come in two types: positive $(+)$ or negative $(-)$. In particular, a positive charge occurs when the number of protons exceeds the number of electrons, and a negative charge occurs when the number of electrons exceeds the number of protons. The fundamental building blocks of ordinary matter are the negatively charged "electron," the positively charged "proton," and the uncharged "neutron."[4] In a neutral atom, the number of electrons equals the number of protons that exist in the nucleus, and the net electric charge is zero. If one or more electrons are removed (resp. added), then the remaining positively (resp. negatively) charged structure is called a "positive ion" (resp. a "negative ion").

For instance, consider static electricity (triboelectric effect): Friction can give loosely bound electrons enough energy to leave their atoms and get attached to others, migrating between different surfaces. When this happens, the first object is left with more protons than electrons and, thus, becomes positively charged, whereas the object with more electrons accumulates a negative charge. This situation is called "net charge separation." However, when one of these newly charged bodies comes into

[4] In simple terms, to construct an atom, one needs some protons and neutrons for the construction of the nucleus, and then one has to put some electrons around the nucleus until the whole system is electrically neutral (in fact, once you have a positively charged nucleus, it attracts electrons, which automatically form shells around the nucleus). In 1911, the New Zealand physicist Ernest Rutherford discovered the basic structure of the atom: it consists of a small and dense core of positive electric charge called the nucleus, surrounded by a "cloud" (probability distribution) of negatively charged electrons; and, in particular, electrons move in orbitals around the nucleus in an energy level (precisely, an electron has a probability of being in various locations based on its energy). However, it should be mentioned that the construction of an atomic nucleus is a complex process, because protons, being positively charged, repel each other. As a result, they have to come very close to each other in order for the nuclear force to start operating and, thus, keep them together, given that there exist sufficiently many neutrons. This process requires extremely high temperatures (hundreds of millions of degrees Kelvin). Such high temperatures existed briefly after the Big Bang. The "atomic number," which defines the identity of an element, is the number of protons in the nucleus of an atom, and, since atoms are electrically neutral, the atomic number also indicates the number of electrons in an uncharged atom.

contact with another material, the mobile electrons will take the first chance they get to go where they are most needed, thus relocating from the negatively charged object to a positively charged one, restoring the neutral charge equilibrium. This quick movement of electrons is called "static discharge," and it is recognized as a sudden spark. This process happens only with specific objects. In particular, "conductors," such as metals and salt water, tend to have loosely bound outer electrons, which can easily flow between molecules, whereas "insulators," such as plastic, rubber, and glass, have tightly bound electrons, which do not regularly jump to other atoms. "Static buildup" is the phenomenon wherein electric charges are exchanged between the surfaces of two objects that come into contact with each other; and it is most likely to occur when one of the materials involved is an insulator. For instance, when you shuffle your feet across a rug, you are creating many surface contacts between your feet and the rug, and, thus, electrons relocate from your body to the rug (due to friction), whereas the rug's insulating wool will resist loosing its own electrons. Your body and the rug together constitute a system that is electrically neutral, but there is a charge polarization between your body and the rug (your body representing the positive pole, and the rug representing the negative pole), so that, when you reach to touch the metal door knob, you will experience an electric shock, since the metal door knob's loosely bound electrons will relocate to your hand in order to replace the electrons that your body has lost. Similarly, when you rub a plastic comb on your head, it causes opposite static charges to build up both on your hair and the plastic comb, and, therefore, when you pull the plastic comb slowly away from your head, you can see these two opposite static charges attracting each other and making your hair stand up. Charge separation may happen in clouds, and, in this case, it is neutralized by being released towards another body, such as a building, the earth, or another cloud, in a giant spark that we know as a lightning.

"Eelectricity" is the flow of electric charge along a path provided by a conductor (conductors are materials with high electron mobility). If you have two charges, one positive and one negative, then they have an electric field between them.

The amount of work needed in order to move a unit of electric charge from a reference point to a specific point in an electric field without producing acceleration is called an "electric potential." In terms of SI units, it is represented by

$$V = \frac{potential\ energy}{charge} = \frac{joule}{coulomb},$$

where joule is the unit for work done, and $1\,joule = (1\,newton)(1\,meter)$; coulomb is the unit for the charge; and V denotes "volt," the derived unit for electric potential (electromotive force), and it is named after the Italian physicist Alessandro Volta (1745–1827). The motion across the field is supposed to proceed with negligible acceleration in order to avoid the test charge acquiring kinetic energy or producing radiation. When we move a charge at constant speed, it becomes a current, and it generates a magnetic field (actually consisting of attraction and repulsion of electric fields) that is perpendicular to the motion of the charge; whereas, if we accelerate the charge, then the charge produces a squeezed electric field, which is no longer spherical, but is shaped like an hour glass.

The key to the flow of electricity is making a continuous electric circuit: connecting a wire between a source of electrons and an attractor of electrons (for which reason, for instance, a battery has two poles: a source (a negative), and an attractor (a positive); and, similarly, an electric plug has at least two tongs, one for incoming electrons and one for outgoing electrons). Electrons do not cease to exist. Rather, being carriers of charge, they move from the negative (source) to the positive (attractor), and they are useful as they follow the path to their destination in the context of a continuous electric circuit. By contrast, connecting two poles of a power source directly can actually be very dangerous: this is what is called a "short circuit," because there is no electric device between the source and the destination of electrons to power, such as a PC or a TV set. In case of a short circuit, the electron flow does not encounter any resistance, therefore the release of energy is instant, often paired with the involved wire heating dangerously.

The electric field at a point can be calculated by using Coulomb's law in order to find the total force F on a test charge q' placed at the point, and then we divide F by q' to obtain the electric field E. If q' is positive, then the direction of E is the direction of F. The force on a negative charge, such as an electron, is opposite to the direction of E.

In order to analyze the motion of a particle with charge q in an electric field, we need to use Newton's Second Law of Motion, $F = ma$, with F caused by the electric field E, so that the magnitude of the electric force F is given by

$$F = qE$$

(in vector notation, $\vec{F} = q\vec{E}$). If the field is uniform, then the acceleration is constant.

In simple terms, electric interactions can be described as follows: a charge distribution sets up an electric field E, and the field exerts a force $F = qE$ on any charge q that is present. The same pattern can be followed in order to describe magnetic interactions (phenomena of attraction or repulsion that arise between electrically charged particles because of their motion). A moving charge, or a current, sets up a magnetic field in the space around it, and this field exerts a force F on a moving charge. Like electric field, magnetic field is a vector field (a vector quantity associated with each point in space). The symbol for magnetic field is B.

Whereas the electric-field force is the same whether the charge is moving or not, the magnetic force is proportional to the particle's speed. Thus, a particle at rest experiences no magnetic force at all. Furthermore, the magnetic force F acting on a charge q moving with velocity v does not have the same direction as the corresponding magnetic field B, but it is perpendicular to both the magnetic field B and v. Hence, the magnitude of the magnetic force F is given by

$F = |q|vBsin\varphi,$

where $|q|$ is the magnitude of the charge, and φ is the angle measured from the direction of v to the direction of B. The SI unit of B is $1N \cdot sec \cdot C^{-1} \cdot m^{-1}$, where N stands for newton, sec stands for second, C stands for coulomb, and m stands for meter. This unit is called 1 tesla ($1T$), in honor of the prominent Serbian-American scientist and inventor Nikola Tesla (1857–1943).

Using vector notation, the force that a magnetic field $\vec{B}$ exerts on a charge q with velocity $\vec{v}$ is given by

$\vec{F} = q\vec{v} \times \vec{B},$

where $\vec{v} \times \vec{B}$ denotes the cross product of the velocity and the magnetic field.

In 1831, the English scientist Michael Faraday discovered electromagnetic induction: he placed a stationary magnet inside or outside a coil, and he observed no deflection in the galvanometer. However, at the moment that he moved the magnet towards (into/above/below) the coil, he saw the pointer deflecting in one direction, and, at the moment that he moved the magnet way from the coil, he saw the pointer deflecting in the opposite direction. Using the aforementioned notation, the entire electromagnetic force F on the charged particle is called the Lorentz force (after the Dutch physicist H. A. Lorentz), and its magnitude is given by

$$F = F_{electric} + F_{magnetic}$$

(and, as I have already mentioned, $\vec{F}_{electric} = q\vec{E}$, and $\vec{F}_{magnetic} = q\vec{v} \times \vec{B}$). Faraday's discovery was really amazing, because one could make something move without ever touching it, only by using the field. Indeed, we can affect things far away and develop telecommunications using electromagnetic fields. Notice that the operation of antennas is based on electromagnetism (by an "antenna," we mean anything that transfers electricity from the air to a wire, or from a wire to the air). In other words, antennas are a way of transmitting and receiving information through changes in the electromagnetic fields that surround them. Moreover, Faraday was the first to understand that waves of the electromagnetic field are what we call light.

By the term "wave," we mean a disturbance or oscillation that travels through space-time accompanied by a transfer of energy. The basic properties of a wave are its amplitude (i.e., the distance from the center line, that is, the still position, to the top of a crest or the bottom of a trough), its frequency (i.e., the number of cycles occurring per second; specifically, it can be measured by counting the number of crests of waves that pass a fixed point in one second), and its length (i.e., the distance over which the wave's shape repeats; for instance, the distance between two adjacent crests). According to the theory of wave mechanics, which was formulated in the 1920s by the Austrian-Irish physicist Erwin Schrödinger, a wave itself does not have units of matter or energy, but it is just form, specifically, a pattern of information.

In simple terms, electromagnetic radiation consists of electric and magnetic fields oscillating around each other, creating a freely propagating wave that can travel from one place to another. This event explains light, the operation of radio stations, the operation of microwave ovens, etc. These are electromagnetic phenomena, and they differ from each other only with respect to the wavelength of the corresponding oscillation, so that we use different names for electromagnetic radiation depending on the corresponding wavelength; for instance, if we can see electromagnetic radiation, then we call it light, light with large wavelengths is red, light with larger wavelengths that is invisible is called infrared, while, at even larger wavelengths, electromagnetic radiations are called microwaves, and, if the wavelengths are even larger, then electromagnetic radiations are called radio-waves.

By the term "radiation," we generally mean energy transferred by waves or particles. For instance, radiation may take the form of electromagnetic waves—which, however, are made of particles, photons specifically. A photon is a type of elementary particle that serves as the quantum of the electromagnetic field and the force carrier for the electromagnetic force.[5]

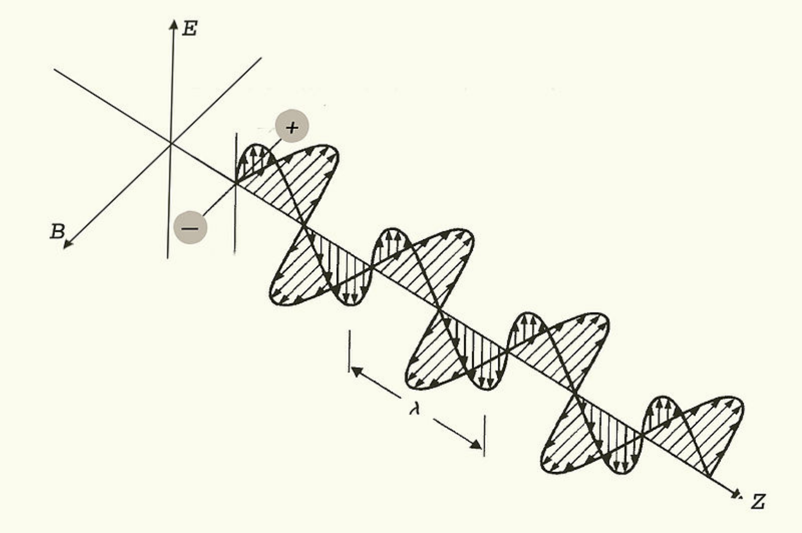
[5] The term "quantum" derives from the Latin language, and it means an amount of something. In the context of quantum mechanics, the term "quantum" means the smallest amount of energy that can be measured. In fact, light is made up of photons, which we can think of as small packets ("quanta") of energy. For instance, when I point a flashlight at an object, I direct photons (which make up the given beam of light) to hit the object. In opaque solids, when photons hit the

In particular, quantum electrodynamics describes the manner in which electrically charged particles interact by shooting photons back and forth between each other. Electrons, being zero-dimensional, lack spatial extension (that is, they have practically zero volume). Therefore, they interact with each other by exchanging photons. As two electrons move towards each other, a photon is passed from one to another, and it changes the momentum of both of them, thus pushing them off.

In Figure 8-23, we see the graph of a linearly polarized electromagnetic wave going in the z-axis with E denoting the electric field (corresponding to the x-axis) and B denoting the magnetic field (corresponding to the y-axis). In electrodynamics, by the term "linear polarization," we refer to a confinement of the electric field vector or the magnetic field vector to a given plane along the direction of propagation, and this term was coined by the French civil engineer and physicist Augustin-Jean Fresnel (1788–1827).

surface of the material, the energy of the photons is absorbed by the electrons to excite themselves to the next atomic orbital, and, when this happens, the photons lose all of their energy (and there are not any photons any more; this is the reason why an opaque solid is opaque). Due to the structure of opaque solids, the electron orbitals are not far enough from each other (in terms of energy), and the photon has enough energy to push the electron up to a higher energy orbital (and, thus, the electron absorbs the photon). However, due to the amorphous structure of glass (silicon dioxide), the energy gap between the atomic orbitals is too large, and, therefore, when a photon hits the glass, the electron does not absorb it, because the photon does not have enough energy to push the electron up to a higher energy orbital. Hence, in this case, the photon retains its energy, and the electron lets it pass; and this is the reason why the glass is transparent.

Figure 8-23: A linearly polarized electromagnetic wave (source: Wikimedia Commons: Author: Витольд Муратов; https://commons.wikimedia.org/wiki/File:%D0%93%D0%B5%D0%BD%D0%B5%D1%80%D0%B0%D1%86%D0%B8%D1%8F_%D1%8D%D0%BB%D0%B5%D0%BA%D1%82%D1%80%D0%BE%D0%BC%D0%B0%D0%B3%D0%BD%D0%B8%D1%82%D0%BD%D0%BE%D0%B9_%D0%B2%D0%BE%D0%BB%D0%BD%D1%8B.jpg).



Now, let us consider a vector-valued function (vector field) $\vec{r}(x,y,z) = f(x,y,z)\hat{\imath} + g(x,y,z)\hat{\jmath} + h(x,y,z)\hat{k}$ such that the partial derivatives $\frac{\partial f}{\partial x}$, $\frac{\partial g}{\partial y}$, and $\frac{\partial h}{\partial z}$ exist and are continuous on $U \subseteq \mathbb{R}^3$. Then the "divergence" of $\vec{r}(x,y,z)$ is a vector operator that operates on a vector field, producing a scalar field that gives the quantity of the vector field's source at each point; and it is defined as follows:

$$div\vec{r} \equiv \vec{\nabla} \cdot \vec{r} = \frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} + \frac{\partial h}{\partial z}$$

(where, in particular, we have: $\vec{\nabla} \cdot \vec{r} = \left(\frac{\partial}{\partial x}\hat{\imath} + \frac{\partial}{\partial y}\hat{\jmath} + \frac{\partial}{\partial z}\hat{k}\right)\left(f(x,y,z)\hat{\imath} + g(x,y,z)\hat{\jmath} + h(x,y,z)\hat{k}\right)$).

In other words, the divergence of a function tells us how the corresponding vector field behaves towards or away from a point: the divergence of a vector field represents the tendency of the field to either converge or diverge at a given point (e.g., in the context of mechanical systems, electromagnetism, and fluid dynamics). In particular, in physics, the divergence of a vector field is the extent to which the vector field flux behaves like a source at a given point, and, specifically, it is a local measure of the extent to which, in a vector field, there are more vectors exiting from an infinitesimal region of space than entering it (the term "flux" refers to any effect that appears to pass or travel through a surface or substance). A point at which the flux is outgoing has positive divergence, and it is said to be a "source" of the field, whereas a point at which the flux is directed inward has negative divergence, and it is said to be a "sink" of the field. Obviously, the greater the vector field flux through a small surface enclosing a point, the greater the value of divergence at that point.

The divergence of an electrostatic field $\vec{E}$ is

$$div\vec{E} = \frac{\rho}{\varepsilon_0}$$

where ρ is the electric charge density (i.e., charge per unit volume), and ε_0 is the permittivity of free space (i.e., a physical constant that reflects the ability of electric fields to pass through a classical vacuum; $\varepsilon_0 \approx 8.85 \times 10^{-12}$ farads per meter). The divergence of an electrostatic field provides important information: a region with zero divergence is either a constant field or contains no charges; whereas non-zero divergence regions indicate the presence of charges, and the sign of divergence determines whether the charges are positive or negative.

Given a vector-valued function (vector field) $\vec{r}(x,y,z) = f(x,y,z)\hat{\imath} + g(x,y,z)\hat{\jmath} + h(x,y,z)\hat{k}$ such that the partial derivatives $\frac{\partial f}{\partial x}$, $\frac{\partial g}{\partial y}$, and $\frac{\partial h}{\partial z}$ exist and are continuous on $U \subseteq \mathbb{R}^3$, the "curl" (also known as "rotor") of $\vec{r}(x,y,z)$ is the vector-valued function (vector field)

$$curl\vec{r} \equiv \vec{\nabla} \times \vec{r} = \left(\frac{\partial h}{\partial y} - \frac{\partial g}{\partial z}\right)\hat{\imath} + \left(\frac{\partial f}{\partial z} - \frac{\partial h}{\partial x}\right)\hat{\jmath} + \left(\frac{\partial g}{\partial x} - \frac{\partial f}{\partial y}\right)\hat{k}$$

$$= \begin{vmatrix} \hat{\imath} & \hat{\jmath} & \hat{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ f & g & h \end{vmatrix}$$

(notice that, in this case, we expand the determinant only across the first row, and it is used as a mnemonic rule). The curl of a vector field represents the rotation (or "circulation") of the field around a point (it is used, for instance, in order to analyze the rotation of a fluid flow, magnetic fields, and stress distributions). In particular, the curl at a point in the vector field is a vector whose length and direction denote, respectively, the magnitude and the axis of the maximum rotation of the field. A zero curl implies that a vector field is irrotational (circulation is zero), and, in electrostatics, it is indicative of electrostatic fields with no magnetic field present (*note:* a moving electric charge generates a magnetic field, and a magnetic field induces electric charge movement, producing an electric current, but electric fields and magnetic fields may also exist independently of each other: an electric field with no magnetic field present exists in charges at rest; just as a magnetic field with no electric field present exists in permanent magnets).

Line integrals of vector fields: Consider a vector field

$$\vec{F}(x,y,z) = P(x,y,z)\hat{\imath} + Q(x,y,z)\hat{\jmath} + R(x,y,z)\hat{k}$$

and the three-dimensional smooth curve C given by

$$\vec{r}(t) = x(t)\hat{\imath} + y(t)\hat{\jmath} + z(t)\hat{k}$$

where $a \leq t \leq b$. The "line integral of $\vec{F}$ along C" is

$$\int_C \vec{F} \cdot d\vec{r} = \int_a^b \vec{F}\big(\vec{r}(t)\big) \cdot \vec{r}'(t)dt$$

where, in the integral on the left side, the dot denotes a dot product of the vector field, and the differential is a vector. Moreover, notice that $\vec{F}\big(\vec{r}(t)\big) = \vec{F}\big(x(t), y(t), z(t)\big)$. The line integral can be written with respect to the arc length as follows:

$$\int_C \vec{F} \cdot d\vec{r} = \int_C \vec{F} \cdot \vec{T} ds$$

where $\vec{T}(t)$ is the unit normal vector, that is,

$$\vec{T}(t) = \frac{\overrightarrow{r'}(t)}{\left\|\overrightarrow{r'}(t)\right\|}$$

(as we have previously explained). Line integrals of vector fields are useful in physics for computing the work done by a force on a moving object along a curve.

For instance, let us compute the line integral $\int_C \vec{F} \cdot d\vec{r}$ where $\vec{F}(x,y,z) = 8x^2yz\hat{\imath} + 5z\hat{\jmath} - 4xy\hat{k}$, and the curve C is defined by $\vec{r}(t) = t\hat{\imath} + t^2\hat{\jmath} + t^3\hat{k}$ with $0 \leq t \leq 1$. Firstly, the given vector field along the given curve is $\vec{F}\big(\vec{r}(t)\big) = 8t^2(t^2)(t^3)\hat{\imath} + 5t^3\hat{\jmath} - 4t(t^2)\hat{k} = 8t^7\hat{\imath} + 5t^3\hat{\jmath} - 4t^3\hat{k}$.

Secondly, the derivative of the parametric expression of the curve is $\overrightarrow{r'}(t) = \hat{\imath} + 2t\hat{\jmath} + 3t^2\hat{k}$.

Thirdly, the corresponding dot product is $\vec{F}\big(\vec{r}(t)\big) \cdot \overrightarrow{r'}(t) = 8t^7 + 10t^4 - 12t^5$.

Hence, the given line integral is

$\int_C \vec{F} \cdot d\vec{r} = \int_0^1 (8t^7 + 10t^4 - 12t^5) dt = 1.$

In general, as we can easily see, another (equivalent) way of computing line integrals of vectors fields is the following: Given a vector field

$$\vec{F}(x,y,z) = P(x,y,z)\hat{\imath} + Q(x,y,z)\hat{\jmath} + R(x,y,z)\hat{k}$$

and the three-dimensional smooth curve C defined by

$$\vec{r}(t) = x(t)\hat{\imath} + y(t)\hat{\jmath} + z(t)\hat{k}$$

where $a \leq t \leq b$, the "line integral of $\vec{F}$ along C" is

$$\int_C \vec{F} \cdot d\vec{r} = \int_C (Pdx + Qdy + Rdz)$$

since

$\int_C \vec{F} \cdot d\vec{r} = \int_a^b (P\hat{\imath} + Q\hat{\jmath} + R\hat{k}) \cdot (x'\hat{\imath} + y'\hat{\jmath} + z'\hat{k}) = \int_a^b (Px' + Qy' + Rz')\, dt = \int_a^b Px'dt + \int_a^b Qy'dt + \int_a^b Rz'dt = \int_C Pdx + \int_C Qdy + \int_C Rdz.$

Green's and Stokes's Formulae: Let us consider in the plane $\mathbb{R}^2$ a smooth closed curve K without self-intersections which bounds an open domain D in $\mathbb{R}^2$, as shown in Figure 8-24. Suppose that, on K, a parameter t is valid and defines the circulation direction and, therefore, the orientation of K as an one-dimensional manifold. Then the closure $Cls(D) \equiv \overline{D}$ is an oriented two-dimensional manifold with the boundary $\partial\overline{D} = K$. If the orientation of the domain $\overline{D}$ is defined by a linear coordinate system (x, y), then the orientation on the boundary K will be compatible with the orientation of the entire $\overline{D}$, provided of course that the domain $\overline{D}$ lies on the left of K when K is traversed in the direction of increasing parameter t. Let $\vec{F} = (P, Q)$ be a C^1 vector field on $\mathbb{R}^2$. In the coordinate system (x, y), the vector field $\vec{F}$ can be expressed as $\vec{F} = P(x,y)dx + Q(x,y)dy$. Then the integral of $\vec{F}$ along the curve K is given by

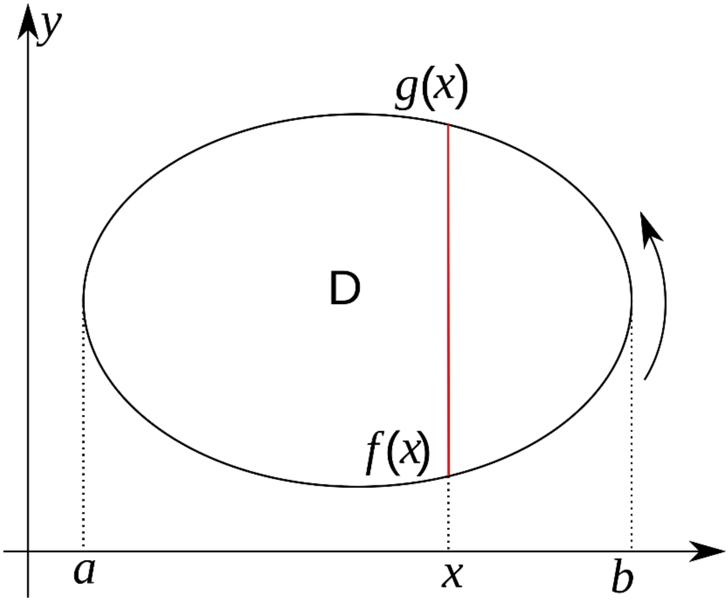
$$\int_K (Pdx + Qdy) = \int_{t_0}^{t_1} \left(P\big(x(t), y(t)\big) \frac{dx}{dt} + Q\big(x(t), y(t)\big) \frac{dy}{dt} \right) dt$$
$$= \int \int_D \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dxdy$$

(this result is "Green's formula," named after the British mathematical physicist George Green, who published an initial version of this formula in 1828). Hence, Green's formula relates a line integral around a simple closed curve K to a double integral over the plane region D bounded by K. In other words, Green's formula tells us the following: If U is a region in $\mathbb{R}^2$ whose boundary ∂U consists of a finite union of curves of class C^1, if we orient ∂U so that, whenever we traverse the boundary in the direction of orientation, U remains on the left, and if $\vec{F} = (P, Q)$ is a C^1 vector field on U, then

$$\int_{\partial U} (Pdx + Qdy) = \int\int_U \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dxdy$$

(where $P(x, y)$ and $Q(x, y)$ are the components of the given vector field $\vec{F}$).

Figure 8-24: Green's formula (source: Wikimedia Commons: Author: Theon; https://commons.wikimedia.org/wiki/File:Th%C3%A9or%C3%A8me_de_Green-Riemann.svg?uselang=eo).



Similarly, let K be a smooth closed curve without self-intersections in the space $\mathbb{R}^3$, and let this curve be the boundary of a two-dimensional surface D. Given a C^1 vector field on $\mathbb{R}^3$, Stokes's formula (named after the Irish physicist and mathematician George Stokes) relates the surface integral of the curl of the vector field over the given surface to the line integral of the vector field around the boundary of the given surface, as follows:

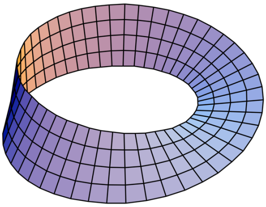
$$\int_K (Pdx + Qdy + Rdz)$$
$$= \int\int_D \left[\left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dxdy + \left(\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z}\right) dydz + \left(\frac{\partial P}{\partial z} - \frac{\partial R}{\partial x}\right) dzdx\right]$$

(i.e., $\int_K \vec{F}\, ds = \int\int_D curl\vec{F} \cdot d\vec{S}$ for $\vec{F} = (P, Q, R)$).

Surface integrals of vector fields: Let us think of a two-sided[6] smooth (or at least piecewise smooth) surface immersed in a vector field, so that the vector field under consideration contains this surface. If a vector field $\vec{F}$ contains such a surface S, then $\vec{F}$ describes the velocity of the flow at any point across the surface. The rate of flow (i.e., the amount or the volume of flow across the surface) is called the "flux," and this concept is the key to understanding surface integrals of vector fields. For instance, if the vector field $\vec{F}$ represents the flow of a fluid, then the surface integral of $\vec{F}$ represents the flux, that is, the amount or the volume of fluid flowing through the surface (per unit time). This is the reason why the surface integral of a vector field is frequently called a "flux integral." If dS is an element (infinitesimal) of the surface (i.e., a really small surface area), and

[6] It should be mentioned that a surface may be one-sided. For instance, a "Möbius strip" can be constructed by gluing together the edges of a sheet of paper with a twist. Thus, we obtain an one-sided, non-orientable surface (within it, one cannot consistently distinguish clockwise from counterclockwise turns), as shown in Figure 8-25. A surface is "orientable" if and only if, during any smooth shifting across the surface, any sufficiently small circle on the surface with a fixed direction of the journey along its boundary preserves the original direction of the journey along its boundary (we assume that the circle does not intersect the edge of the surface).

Figure 8-25: Möbius strip (source: Wikimedia Commons: Author: Fropuff; https://commons.wikimedia.org/wiki/File:MobiusStrip-01.svg).



if $\vec{n}$ is the unit normal vector to dS at one of its points, then the flux of the vector field $\vec{F}$ through the elementary region dS is given by

$$\vec{F} \cdot \vec{n} dS$$

where $\vec{F}$ is taken at the same point as $\vec{n}$, and $\vec{F} \cdot \vec{n}$ denotes the dot product between $\vec{F}$ and $\vec{n}$. For instance, if a fluid is flowing perpendicular to the surface, a lot of that fluid will flow through the surface, and the flux will be large, whereas, if a fluid is flowing parallel to the surface, that fluid will not flow through the surface, and the flux will be zero. In order to calculate the total amount of a fluid flowing through the surface, we must add up (i.e., integrate) the component of the vector field $\vec{F}$ that is normal (i.e., perpendicular) to the surface. Notice that, if $\vec{n}$ is the unit normal vector to dS, the dot product $\vec{F} \cdot \vec{n}$ will be positive if $\vec{F}$ and $\vec{n}$ are pointing in the same direction, and it will be negative if $\vec{F}$ and $\vec{n}$ are pointing in opposite directions (in general, the choice of a normal vector orients the surface and determines the sign of the flux).

Hence, by analogy, the flux of the vector field $\vec{F}$ through the *whole* surface S is defined to be the integral

$$\iint_S \vec{F} \cdot \vec{n} dS = \iint_S \vec{F} \cdot d\vec{S}$$

(this is the "flux integral," that is, the surface integral of the vector field under consideration).

Thus, the total flux of fluid flow through a surface S is denoted by $\iint_S \vec{F} \cdot d\vec{S}$, since it is the integral of the vector field $\vec{F}$ over the surface S. Similarly, we can calculate the "electric flux" through a surface as follows:

$$\Phi_E = \iint_S \vec{E} \cdot d\vec{S}$$

where $\vec{E}$ is the electric field (having units of volt per meter), and $d\vec{S}$ is a differential area on the given surface S with an outward facing normal vector defining its direction. The "electric flux" through a surface is proportional to the number of field lines crossing that surface. In other words, its magnitude is proportional to the portion of the field perpendicular to the surface area:

$Electric\ Flux = (Electric\ Field) \cdot (Surface\ Area) \cdot (cos\theta)$,

where $cos\theta$ denotes the cosine of the angle θ between the electric field and the vector that is perpendicular to the area. A "field line" is an imaginary line drawn through a region of space in such a way that, at every point, it is tangent to the direction of the electric-field vector at that point. In particular, in an "electrostatic field," every field line is a continuous curve with a positive charge at one end and a negative charge at the other. In mathematics, we use the term “trajectory of a vector field” in order to refer to a curve whose tangent at every point has the same direction as the corresponding vector field.

Suppose that a surface is defined by a function $z = g(x, y)$. In order to work with with surface integrals of vector fields, we have to define a new function, namely,

$$f(x, y, z) = z - g(x, y)$$

(and, in this way, the surface is defined by the equation $f(x, y, z) = 0$). The unit normal vector to the surface defined by the equation $f(x, y, z) = 0$ is

$$\vec{n} = \frac{\nabla f}{\|\nabla f\|}$$

and, in this case,

$$\vec{n} = \frac{-g'_x\hat{\imath} - g'_y\hat{\jmath} + \hat{k}}{\sqrt{(g'_x)^2 + (g'_y) + 1}}$$

(where, of course, $\hat{\imath}$, $\hat{\jmath}$, and $\hat{k}$ are the corresponding unit vectors on the x-axis, the y-axis, and the z-axis, respectively). Notice that the component of this normal vector in the z direction (that is, $\hat{k}$ in the aforementioned formula of $\vec{n}$) is positive, meaning that the normal vector general points upward, specifically, it has an upward component to it. However, in general, “positive orientation” points out of the region under consideration, and, sometimes, this may mean downward. Thus, if we need the downward orientation, we can take the negative of $\vec{n}$ to obtain the required result. Hence, if a surface S is defined by $z = g(x, y)$, if the corresponding vector field is defined by $\vec{F}(x, y, z) = P(x, y, z)\hat{\imath} + Q(x, y, z)\hat{\jmath} + R(x, y, z)\hat{k}$, and if the orientation in which we are interested is the upward orientation, then the surface integral of $\vec{F}$ over S is computed according to the following formula:

$$\iint_S \vec{F}\cdot\vec{n}dS = \iint_S \vec{F}\cdot d\vec{S}$$

$$= \int\int_D \left(P\hat{\imath}+Q\hat{\jmath}+R\hat{k}\right)$$

$$\cdot\left(\frac{-g'_x\hat{\imath}-g'_y\hat{\jmath}+\hat{k}}{\sqrt{(g'_x)^2+\left(g'_y\right)+1}}\right)\sqrt{(g'_x)^2+\left(g'_y\right)+1}dA$$

$$= \int\int_D \left(P\hat{\imath}+Q\hat{\jmath}+R\hat{k}\right)\cdot\left(-g'_x\hat{\imath}-g'_y\hat{\jmath}+\hat{k}\right)dA$$

$$= \int\int_D \left(-Pg'_x-Qg'_y+R\right)dA$$

(notice that this computation holds in case the surface is given in the form $z=f(x,y)$, but we can obviously think and work in a similar fashion when the surface is given in the form $y=g(x,z)$, in which case we we define $f(x,y,z)=y-g(x,z)$, as well as when the surface is given in the form $x=g(y,z)$, in which case we define $f(x,y,z)=x-g(y,z)$). Given that we can consider two different orientations, there are six possible surface integrals, that is, two for each form of the surface: $z=f(x,y)$, $y=g(x,z)$, and $x=g(y,z)$, so that, given each form of the surface, there will be two possible unit normal vectors, and we have to choose the one that matches the given orientation of the surface (but the derivation of the corresponding formula of the surface integral is always similar to that given above).

Now, suppose that the surface S is defined parametrically, as follows:

$$\vec{r}(u,v)=x(u,v)\hat{\imath}+y(u,v)\hat{\jmath}+z(u,v)\hat{k}$$

(as always, $\hat{\imath}$, $\hat{\jmath}$, and $\hat{k}$ denote the corresponding unit vectors on the x-axis, the y-axis, and the z-axis, respectively). In this case, the vector $\overrightarrow{r'_u}\times\overrightarrow{r'_v}$ is normal to the tangent plane to the curve at a particular point of the curve (and, therefore, to that point of the curve itself), and the corresponding unit normal vector is

$$\vec{n}=\frac{\overrightarrow{r'_u}\times\overrightarrow{r'_v}}{\left\|\overrightarrow{r'_u}\times\overrightarrow{r'_v}\right\|}$$

(and, as previously, we have to consider the adequate direction). Therefore, if the surface S is given parametrically by $\vec{r}(u,v)$ with parameter domain D, the surface integral of $\vec{F}$ over S is computed according to the following formula:

$$\begin{aligned}\iint_S \vec{F}\cdot\vec{n}dS = \iint_S \vec{F}\cdot d\vec{S} \\ &= \int\int_D \vec{F}\cdot\left(\frac{\overrightarrow{r_u'}\times\overrightarrow{r_v'}}{\left\|\overrightarrow{r_u'}\times\overrightarrow{r_v'}\right\|}\right)\left\|\overrightarrow{r_u'}\times\overrightarrow{r_v'}\right\|dA \\ &= \int\int_D \vec{F}\cdot\left(\overrightarrow{r_u'}\times\overrightarrow{r_v'}\right)dA\end{aligned}$$

(as previously, we may have to change the sign of $\overrightarrow{r_u'}\times\overrightarrow{r_v'}$ in order to match the orientation of the surface).

Chapter 9
Complex Numbers and Complex Analysis

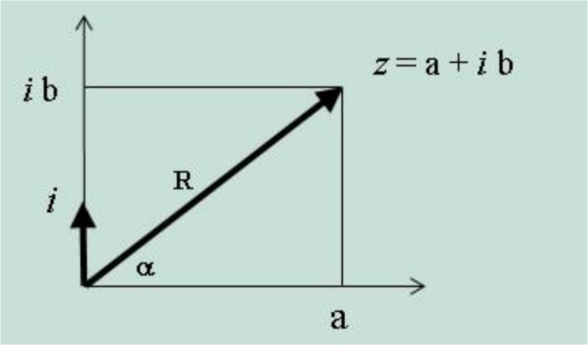
As I have already mentioned, the concept of a number has been extended from natural to real numbers, both because of human practice and because of the needs of mathematics itself. In particular, the concept of a number grew out of the counting of objects. Counting gave rise to the numbers 1, 2, 3, and so on, which are called natural numbers. Then the necessity of performing the operation of division led to the concept of positive fractional numbers; furthermore, the necessity of performing the operation of subtraction led to the concepts of zero and negative numbers; finally, the necessity of taking roots of positive numbers led to the concept of irrational numbers. The aforementioned operations are feasible in the set of real numbers. However, there are still impracticable operations—for instance, taking a square root of a negative number. Hence, there is a need to extend the concept of a number even further, specifically, to invent new numbers different from the real numbers.

Indeed, if we adjoin to the real system $\mathbb{R}$ a root i of the polynomial $x^2 + 1 = 0$, which is irreducible to $\mathbb{R}$, we obtain the system of complex numbers $\mathbb{C} \equiv \mathbb{R}(i)$. The symbol

$z = a + bi$, where $a, b \in \mathbb{R}$ and $i = \sqrt{-1}$,

is called a "complex number"; the number a is called the "real part" of $z = a + bi$, and it is denoted by $Re(z)$; the number b is called the "imaginary part" of $z = a + bi$, and it is denoted by $Im(z)$; and $i = \sqrt{-1}$ is called the "imaginary unit." As we can see in Figure 9-1, a complex number is a two-dimensional number. The number $i = \sqrt{-1}$ signifies a 90^o rotation about the real axis, turning 1 into -1. Hence, $i = \sqrt{-1}$ done twice, or squared, is equal to -1. Two complex numbers $z = a + bi$ and $w = c + di$ are "equal" if and only if $a = c$ (that is, $Re(z) = Re(w)$) and $b = d$ (that is, $Im(z) = Im(w)$).

Figure 9-1: A complex number (source: Wikimedia Commons: Author: Zgyorfi; https://commons.wikimedia.org/wiki/File:Depicting_complex_numbers.JPG).



Any polynomial equation with coefficients can be solved in the system of complex numbers, and the system of complex numbers is the fundamental connection between geometry and algebra. The development of the system of complex numbers is originally due to the sixteenth-century Italian mathematicians Gerolamo Cardano and Rafael Bombelli; and, in the nineteenth century, the system of complex numbers was put in a more rigorous and conceptually richer mathematical setting by Cauchy and Riemann.

As shown in Figure 9-1, we picture the complex number $z = a + bi$ by putting a on the x-axis and b (or rather bi) on the y-axis.

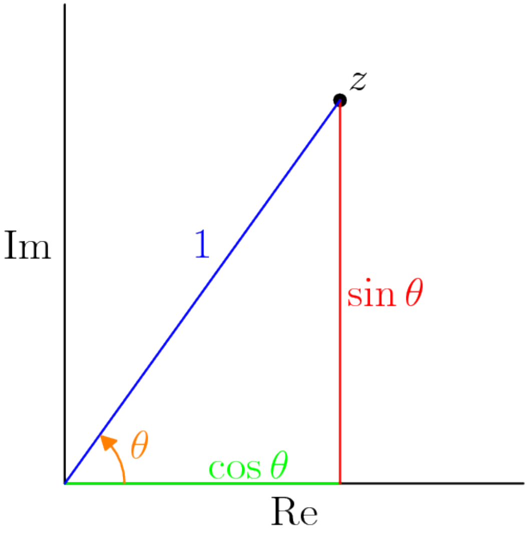
The "modulus" or "absolute value" of $a + bi$ is $\sqrt{a^2 + b^2}$, and it is denoted by $mod(a + bi)$ or $|a + bi|$. The square of the modulus of a complex number $z = a + bi$ is called its "norm," and it is denoted by $Nm(z)$; so that, if $z = a + bi$, then $Nm(z) = a^2 + b^2$.

Now, let us consider Figure 9-2. The "argument" of $z = a + bi$, denoted by $arg(z)$, is a quantity θ such that $cos\theta = \frac{a}{|z|}$ and $sin\theta = \frac{b}{|z|}$. It is many-valued and determined only up to multiples of 2π. In other words, the argument of a complex number is the inclined angle developed in between the real axis and the complex number in the direction of the complex number; and, given a complex number $z = a + bi$, its argument is $\theta = tan^{-1}\left(\frac{b}{a}\right)$.

As shown in Figure 9-2, the angle (in radians) that θ intercepts forms an arc of length s, so that $s = r\theta$ (where r denotes the radius of the corresponding circle), and, if $r = 1$, that is, for the unit circle, $s = \theta$. The study of the unit circle implies that the sine of an angle θ equals the y-

value of the endpoint on the unit circle of an arc of length θ, and the cosine of an angle θ equals the x-value of the endpoint. Therefore, using the unit circle, and given Figure 9-2, we obtain the following trigonometric form of a complex number:
$z = a + bi = |z|cos\theta + |z|sin\theta \cdot i = |z|(cos\theta + isin\theta) = |z|e^{i\theta}$,
where $|z|$ is the modulus of z, $|z|$ is equal to the radius vector of the point z (in the case of the unit circle, $|z| = 1$), and e is the base of the natural logarithm. Hence, we obtain Euler's formula:
$e^{ix} = cosx + isinx$ for any real or complex number x. It is noteworthy that, when $x = \pi$, Euler's formula yields $e^{i\pi} + 1 = 0 \Leftrightarrow e^{i\pi} = -1$. Moreover, in polar coordinates, for some r and θ depending on x, Euler's formula can be written as follows:
$e^{ix} = r(cos\theta + isin\theta)$.

Figure 9-2: The complex plane (source: Wikimedia Commons: Author: Lickyvi; https://commons.wikimedia.org/wiki/File:Nthrootofunity.png).



In general, a circle of radius r (where r is a positive real number) centered at a point $a \in \mathbb{C}$ is given by the equation

$$|z - a| = r$$

(where $|\cdot|$ denotes the complex modulus).

Notice that the periods of hyperbolic functions are complex numbers: the functions $sinhx$ and $coshx$ have period $2\pi i$, and $tanhx$ has period πi, where $i = \sqrt{-1}$.

In 1833, at the Royal Irish Academy, the Irish mathematician and astronomer Sir William Rowan Hamilton presented the complex numbers as ordered pairs of real numbers, thus denoting a complex number by an ordered pair (a, b), and denoting the imaginary unit by $i = \sqrt{-1}$, so that $i^2 = (0,1) \cdot (0,1) = (-1,0) = -1$.

The zero of $\mathbb{C}$ is $(0,0)$, and the unit of $\mathbb{C}$ is $(1,0)$.

The (complex) "conjugate" of $a + bi$ is $a - bi$, and the conjugate of a complex number z is denoted by $\bar{z}$ or by z^*; so that, if $z = a + bi$, then: $z + \bar{z} = 2a$, $z - \bar{z} = 2ib$, and $z \cdot \bar{z} = a^2 + b^2 = |z|^2$.

As shown by Hamilton, the complex number system $\mathbb{C}$ is the set $\mathbb{R} \times \mathbb{R}$ with operations defined as follows:

$$(a + bi) + (c + di) = (a + c) + (b + d)i,$$
$$(a + bi) - (c + di) = (a - c) + (b - d)i,$$
$$(a + bi)(c + di) = (ac - bd) + (ad + bc)i, \text{ and}$$
$$\frac{(a + bi)}{(c + di)} = \frac{(a + bi)(c - di)}{(c + di)(c - di)} = \frac{(ac + bd) + (bc - ad)i}{c^2 + d^2}$$

where $a, b \in \mathbb{R}$, and $i = \sqrt{-1}$.

The algebraic form of a complex number enables us to easily carry out such arithmetic operations on complex numbers as addition, subtraction, multiplication, and division, but raising a complex number to a natural power is more convenient in trigonometric form. For this purpose, we usually use "De Moivre's formula" (named after the French mathematician Abraham de Moivre):

$$(cosx + isinx)^n = cosnx + isinnx$$

(where $i = \sqrt{-1}$). De Moivre's formula can be easily proved using mathematical induction and the angle sum and difference trigonometric identities. Moreover, we can obviously derive De Moivre's formula from Euler's formula and the exponential law for integral powers: $(e^{ix})^n = e^{inx}$, where, by Euler's formula, $(e^{ix})^n = (cosx + isinx)^n$, and $e^{inx} = cosnx + isinnx$.

Complex Vector Spaces: The set of complex numbers $\mathbb{C}$ with addition and multiplication as defined above is a field with additive and multiplicative identities $(0,0)$ and $(1,0)$, respectively (the notion of a "field" was discussed in Chapter 7). Thus, we can define complex vector spaces. A "complex vector space" is a vector space whose scalar field is the complex numbers. The set $\mathbb{C}^n$ is the set of column vectors

$$v = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix}$$

where $z_i \in \mathbb{C}$, $i = 1,2, \dots, n$. Such vectors can be added componentwise, and any such vector can be multiplied by a complex scalar. Hence, this is the fundamental example of a complex vector space; and a complex vector space $V \subseteq \mathbb{C}^n$ is a subset such that: for any $v, w \in V$, it holds that $v + w \in V$, and, for any $v \in V$ and any $z \in \mathbb{C}$, it holds that $zv \in V$. The concepts of vector basis and of linearly independent and linearly dependent vectors can be defined in the case of complex vector spaces in the same way as they are defined in the case of real vector spaces (see Chapter 7). The algebra we have done with matrices over the real numbers works perfectly for matrices over $\mathbb{C}$, without any change (see Chapter 3). The dot product of complex vectors is defined as follows:
$\vec{v} \cdot \vec{w} = \sum_i v_i \overline{w_i}$, where $\overline{w_i}$ is the complex conjugate of w_i.
Let $z_1 = a_1 + b_1 i$ and $z_2 = a_2 + b_2 i$. The "cross product" of z_1 and z_2 is defined as follows:
$z_1 \times z_2 = a_1 b_2 - b_1 a_2 = |z_1||z_2| sin\theta$,
where $|\cdot|$ denotes the complex modulus, and θ denotes the angle from z_1 to z_2 measured in the positive direction.

The nth Roots of Unity: The solutions to the equation $z^n = 1$, where $z \in \mathbb{C}$ and n is a positive integer, are said to be the "nth roots of unity," and each root of unity is given by

$$z = cos\frac{2k\pi}{n} + isin\frac{2k\pi}{n} = e^{\frac{2k\pi i}{n}}$$

where $k = 0,1,2, \dots, n-1$. If we set $\omega = cos\frac{2\pi}{n} + isin\frac{2\pi}{n} = e^{\frac{2\pi i}{n}}$, then the n roots are $1, \omega^1, \omega^2, \dots, \omega^{n-1}$, and, geometrically, they represent the n vertices of a regular polygon of n sides inscribed in a circle of radius 1 centered at the origin (the circle is given by the equation $|z| = 1$, and it is the "unit circle" in the complex plane).

Differentiation and Integration of Complex-Valued Functions: If a function f takes real inputs and gives complex outputs, then the "derivative" with respect to its real input is computed by taking the derivatives of the real and the imaginary parts separately, namely:

$$\frac{df}{dx} = \frac{dRe(f)}{dx} + i\frac{dIm(f)}{dx}$$

where $i = \sqrt{-1}$, $Re(f)$ is the real part of f, and $Im(f)$ is the imaginary part of f. In other words, if $f = u + iv$ is a complex-valued function of a *real* variable x, then the derivative of f at the point x_0 is defined by

$$f'(x_0) = u'(x_0) + iv'(x_0)$$

where u' and v' are the derivatives of u and v, respectively, and $i = \sqrt{-1}$. However, the situation becomes more complicated when we consider functions that take *complex* inputs and give complex outputs. Let us consider a complex-valued function $f = u + iv$ of a *complex* variable $z = x + iy$. As in real analysis, we can define the "derivative" of a complex-valued function $f = u + iv$ of a complex variable $z = x + iy$ as follows:

$$f'(z_0) = lim_{\Delta z \to 0} \frac{f(z_0 + \Delta z) - f(z_0)}{\Delta z}$$

where Δz, being a complex number, can approach zero in more than one way; specifically, if we write $\Delta z = \Delta x + i\Delta y$, then we observe that we can approach zero along the real axis $\Delta y = 0$, or along the imaginary axis $\Delta x = 0$, or indeed along any direction. Therefore, this derivative exists if and only if its value does not depend on how Δz approaches zero, and, in particular, if and only if the following equations are satisfied at the point (x_0, y_0):

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}$$

and

$$\frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}$$

(where $f = u + iv$). These equations are called the "Cauchy–Riemann equations" (the first formulation of these conditions appeared in an essay on fluid mechanics that was published by Jean-Baptiste le Rond d'Alembert in 1752). A complex-valued function $f = u + iv$ of a complex variable $z = x + iy$ is "differentiable" at z_0 if and only if $f'(z_0)$ is well-defined at z_0, and, in view of the foregoing,

$$f'(z) = \frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x} = \frac{\partial v}{\partial y} - i\frac{\partial u}{\partial y}$$

(according to the Cauchy–Riemann equations). The function f is said to be "analytic" (or "holomorphic") in a neighborhood U of z_0 if it is differentiable everywhere in U (i.e., a function can be differentiable *at a point*, but analyticity of complex functions only makes sense in an open set). If a function is analytic in the whole complex plane, then it is called "entire."

Having discussed differentiation of complex-valued functions, let us now discuss integration of complex-valued functions. Suppose that $f(x) =$

$g(x) + ih(x)$ is a complex-valued function of a *real* variable x. Then the "integral" of $f(x)$ between the limits a and b is defined by

$$\int_a^b f(x)dx = \int_a^b [g(x) + ih(x)]dx = \int_a^b g(x)dx + i\int_a^b h(x)dx$$

where $i = \sqrt{-1}$, and x is a real variable. Obviously, the properties of such integrals may be deduced from the properties of the real integrals.

Now, let us consider a complex-valued function $f = g + ih$ of a *complex* variable $z = x + iy$. Let z_0 and z_1 be two points in the complex plane. A curve joining z_0 and z_1 can be defined as follows: imagine a point-particle moving in the complex plane, starting at some time t_0 at the point z_0, and ending at some later time t_1 at the point z_1, so that, at any given instant in time $t_0 \leq t \leq t_1$, this point-particle is at the point $z(t)$ in the complex plane. Thus, a curve joining z_0 and z_1 can be defined by a function $z(t)$ that takes points $t \in [t_0, t_1]$ to points $z(t)$ in the complex plane in such a way that $z(t_0) = z_0$ and $z(t_1) = z_1$. In other words, a "(parametrized) curve" joining z_0 and z_1 is a continuous function $z: [t_0, t_1] \to \mathbb{C}$ such that $z(t_0) = z_0$ and $z(t_1) = z_1$; and, obviously, $z(t)$ can be decomposed into its real part, which is the continuous real-valued function $x(t)$, and into its imaginary part, which is the continuous real-valued function $y(t)$. The curve $z(t)$ is "smooth" if and only if its velocity (first derivative with respect to t), that is, $z'(t)$, is a non-zero continuous function $[t_0, t_1] \to \mathbb{C}$.

Suppose that γ is a smooth curve joining z_0 and z_1, and let $f(z)$ be a complex-valued function that is continuous on the curve γ ($z \in \mathbb{C}$). Then the "integral of $f(z)$ along γ" is defined as follows:

$$\int_\gamma f(z)dz = \int_{t_0}^{t_1} f(z(t))z'(t)dt$$

(as I have already mentioned, $t \in [t_0, t_1]$, $z = x + iy$, $x(t_0) = x_0$, $x(t_1) = x_1$, $y(t_0) = y_0$, $y(t_1) = y_1$, $z_0 = x_0 + iy_0$, and $z_1 = x_1 + iy_1$).

The Fundamental Theorem of Algebra (originally due to Carl Friedrich Gauss)

The Fundamental Theorem of Algebra (also known as the D'Alembert–Gauss theorem) is the statement that every univariate polynomial of positive degree with complex (possibly real) coefficients has at least one complex (possibly real) zero. Therefore, any non-zero polynomial $p(z)$ over $\mathbb{C}$ can be written uniquely (except for order) as a product
$p(z) = k(z - \lambda_1)(z - \lambda_2) \dots (z - \lambda_n)$, where $k, \lambda_i \in \mathbb{C}$, and $n = deg(p)$, meaning that a polynomial in a single variable of degree $n > 0$ with complex (possibly real) coefficients has exactly n complex (possibly real)

zeros, counting multiplicity (i.e., zeros of multiplicity k are counted k times).

In G. H. Hardy's book *A Course of Pure Mathematics*, Appendix II (entitled "The proof that every equation has a root"), we can find an explanation of the proof of this theorem that is based on the concept of continuity and can be summarized as follows: In essence, we want to prove that every algebraic equation must have a root. First of all, we have to realize that a polynomial equation in the complex variable $z = x + yi$ is equivalent to a pair of real equations in the variables x and y, whose loci are curves in the plane. The proof of the Fundamental Theorem of Algebra involes the idea of curves winding around the origin. In the case of real functions, we can visualize the action of functions by sketching graphs in the plane. But, since the space of complex numbers has two real dimensions, the visualization of the action of functions of complex variables requires a four-dimensional space in which we have to construct the graph of a complex function of a complex variable. Alternatively, we can consider two complex planes, one for the domain of the function (the "input plane") and the other for the range of the function (the "output plane"). In other words, if $f(z)$ is a function of the complex variable z, and if $w = f(z)$, then, for each z in the complex plane of the domain (i.e., in the "input plane"), we plot the corresponding point $w = f(z)$ in the complex plane of the range (i.e., in the "output plane"), and we write $z = x + yi$ and $w = u + vi$. However, in order to understand the behavior of the function f, we must envisage z as a moving point in the "input plane" (z-plane) and consider how the image point $f(z)$ moves correspondingly through the "output plane" (w-plane). In fact, as z traces out a certain curve in the "input plane," $f(z)$ traces out a curve in the "output plane," and, by examining the images of special curves, we can analyze the behavior of the function f. In particular, if a curve in the "input plane" (z-plane) passes through a zero of $f(z)$, then its image in the "output plane" (w-plane) must pass through the origin. Hence, the problem of showing that f has a zero reduces to the problem of showing that some image curve must pass through the origin.

Now, let us become more specific: We shall continue thinking in terms of a complex plane called the "input plane," on which we locate the input values of the polynomial, and these inputs are mapped to outputs on another complex plane, the "output plane." For, instance, given a polynomial $p(z) = a_0 z^n + a_1 z^{n-1} + \cdots + a_n$, if $z = 0$, then $p(z) = a_n$. Hence, we know that the zero point on the input plane goes to the point a_n on the output plane (a_n is a complex constant). This is not really helpful, because we wanted to get a point to go to zero, not to a_n. However, we know that, if we take z to have an enormously large magnitude, then z^n will have that magnitude to the nth power, and it will be very much bigger than z^{n-1}. Therefore, for sufficiently large z's (i.e., $|z| \gg 0$), z ranges around a big circle centered at the origin of the the input plane (we get a circle because we can have any argument, any angle), and z^n would be even bigger. Let the constant term a_n of $p(z)$ be non-zero; otherwise, $z = 0$ is a root. Consider the circuit created by $p(z)$ as z ranges around a very small circle centered at the origin. If we make the circle sufficiently small, then all the terms involving powers of z are insignificant compared to the constant a_n, and, therefore, we realize that the image of the circle is contained in a circle winding around a_n that cannot wind around the origin (this follows from the (epsilon-delta)-definition of continuity, setting, for instance, $\varepsilon = a_n/2$). However, for a very large circle (i.e., when $|z| \gg 0$), the highest power of z dominates, and, therefore, the image of the circle will wind around the origin (n times, where n is the degree of $p(z)$). Due to the continuity of $p(z)$, as the radius of the circle grows, there must be some point in between where the image passes through the origin, namely, there must exist a zero of the polynomial (as required).

Regarding the exact number of zeros, we can think as follows: If λ is complex (possibly real) zero of the polynomial $p(z)$, where $deg(p(z)) = n \geq 1$, with complex (possibly real) coefficients, then, by dividing this polynomial by $z - \lambda$, we obtain $p(z) = (z - \lambda)q(z) + r$, where $q(z)$ is a polynomial of degree $n - 1$, and r is a constant. But $p(\lambda) = r = 0$, and, therefore, $p(z) = (z - \lambda)q(z)$. Continuing by induction, we conclude that $p(z)$, which is an arbitrary polynomial of degree $n \geq 1$, has exactly n complex (possibly real) zeros (although some might be repeated), *quod erat demonstrandum*.

The Applications of Complex Numbers in Quantum Physics

Everything that we can definitively say about the physical world, and about the past of the physical world, is based on the classical worldview, which is founded on two major theoretical pillars, depending on the scale of our analysis: Newtonian mechanics and the general theory of relativity. In fact, the general theory of relativity is a geometric theory of gravitation and of space-time, explaining the behavior of the universe on the large scale. On the other hand, the quantum world is not directly observable, and it can be used only for calculating probabilities. Hence, quantum mechanics, pioneered by Niels Bohr, Werner Heisenberg, Wolfgang Pauli, and Erwin Schrödinger, is a theory of physical probability.

In quantum physics, everything is described in terms of wave functions, a wave function is a vector in a complex Hilbert space, and the vector coefficients are complex numbers. According to Paul Dirac's notation, in quantum physics, vectors are symbolized in the following way, known as the bra-ket notation:

$|\Psi\rangle = a_1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + a_2 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + a_3 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$, where $a_1, a_2, a_3 \in \mathbb{C}$.

The aforementioned type of brackets helps us to keep track of whether a vector is a row vector or a column vector: $|\Psi\rangle$ is a column vector, whereas $\langle\Psi|$ is a row vector. In quantum mechanics, if we convert a row vector to a column vector, then we have to take the complex conjugate of each coefficient. In other words, for instance,

$|\Psi\rangle = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}$ and $\langle\Psi| = (a_1^*, a_2^*, a_3^*)$, where a_1^*, a_2^*, a_3^* are, respectively, the complex conjugates of a_1, a_2, a_3.

In quantum mechanics, all vectors describe probabilities. Usually, we choose the basis of the space under consideration in such a way that the basis vectors correspond to possible measurement outcomes; for instance:

$|\Psi\rangle = a_1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + a_2 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + a_3 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ corresponds to

$|\Psi\rangle = a_1|X\rangle + a_2|Y\rangle + a_3|Z\rangle$.

Hence, the probability of a particular measurement outcome is the absolute square of the scalar product with the basis vector that corresponds to the outcome; so that, for instance, the probability of measuring X is

$|\langle X|\Psi\rangle|^2 = a_1 a_1^*$,

and this is known as Born's Rule. In other words, the probability density of finding a particle at a given point, when measured, is proportional to the square of the amplitude of the particle's wave function at that point. In quantum physics, the gradient of a wave function is denoted as follows:

$\nabla|\Psi\rangle = \frac{\partial}{\partial x}|\Psi\rangle\hat{\imath} + \frac{\partial}{\partial y}|\Psi\rangle\hat{\jmath} + \frac{\partial}{\partial z}|\Psi\rangle\hat{k}$.

In order to understand quantum physics, we must understand the difference between the potential mode of being and the actual mode of being. In the context of quantum mechanics, a molecule can be thought of like a mountain range (described by a wave function) filled with infinitely many energy steps, where each energy step, representing a quantum of energy, is a quantum state. A molecule stands on one of these quantum states, and all the other infinitely many quantum states are empty, they are virtual states. Moreover, each quantum state is characterized by a wave form. When a system stands on one of these states, the other states also exist, but potentially. This means that they cannot be observed, and they actually look empty. Those virtual states are potential modes of being, by virtue of which a molecule can jump into other quantum states. Due to Heisenberg's uncertainty principle, we know that molecules can make "quantum jumps," because they have empty states into which they can jump. Of course, understanding the difference between "actuality" and "potentiality," we must never confuse the realm of potentiality with the realm of actuality—that is, we must never attribute actuality to probability; and quantum physics is a physical probability theory.

A successful scientific theory (such as the general theory of relativity, quantum mechanics, etc.) is a mathematical framework—that is, an abstract system—from which we can derive predictions that agree with observation. Therefore, physical objects, such as time, black holes, quarks, bosons, etc., which are said to "exist actually" in the physical world are names that physicists give to mathematical structures or concepts that are necessary parts of successful hypothetico-deductive systems. In physics, a hypothetico-deductive system is said to be successful if the predictions, or the generalizations, that derive from it agree with observations and logic, and then the physical objects that constitute *necessary* parts of such a hypothetico-deductive system (which is consistent with *both* observation and mathematics) are said to exist actually in the physical universe.

As already mentioned, in quantum physics, every system is described by a wave function, usually denoted by the Greek letter Ψ, from which physicists calculate the probability of obtaining a specific measurement outcome. In other words, this wave function is a way of studying the realm of potentiality in a scientifically rigorous way. For instance, from this wave function, one can calculate that a particle that enters a beam-splitter has a 50% chance of going left and a 50% chance of going right. This is a way of analyzing that particle's *potential* mode of being. On the other hand, we can analyze that particle's *actual* mode of being by measuring the given particle.

After measuring the particle, we know with 100% probability where it is. Therefore, we must update our probabilistic study of the particle under consideration accordingly and with it the wave function. This update is known as the "wave function collapse," and it is an observational requirement that stems from the fact that, by measuring the particle, we have achieved a transition from potentiality to actuality. At the level of potentiality, or when we study the potential mode of being of a particle, that particle may be 50% at point A and 50% at point B; but, at the level of actuality, or when we study the actual mode of being of a particle by managing to measure it, that particle is 100% in a particular position, and we never observe a particle that is 50% at point A and 50% at point B. If we observe a particle at all, then we find that it is either in a particular position or not.

Chapter 10
Basic Principles of Ordinary Differential Equations

The Fundamental Theorem of Infinitesimal Calculus is a rigorous explanation of the dialectical relationship between integration and differentiation, and, thus, it is a major underpinning of the theory of differential equations. Moreover, in Chapter 8, I explained the significance of the method of infinitesimal calculus in general.

By the term “ordinary differential equation,” we refer to any equation that contains an unknown function, some of its derivatives, and an independent variable. The “order” of a differential equation is the order of the highest-ordered derivative occurring in the given differential equation. The fundamental problem of the theory of differential equations is to find all of the functions $y = f(x)$ that satisfy some differential equation. Every function $y = f(x)$ that satisfies some differential equation is said to be a “solution” of the given differential equation.

A family of functions

$$y = f(x, c) \qquad (*)$$

where c is a constant belonging to $A \subseteq \mathbb{R}$, is said to be a “general solution” of a differential equation

$$y' \equiv \frac{dy}{dx} = F(x, y) \qquad (**)$$

if, for every $c \in A$, $(*)$ is a solution to $(**)$. The solution that we obtain for each particular value of c is said to be a “partial solution” of the differential equation $(**)$.

The theory of differential equations is a branch of mathematics in which the study of theoretical problems can hardly be distinguished from the study of practical problems, and dynamics, which is a characteristic aspect of modern mathematics, is clearly manifested. Moreover, the theory of differential equations has played an important role in the transition from the eighteenth-century infinitesimal calculus to advanced mathematical analysis and modern geometry. One of the major advantages of differential equations is that they constitute one of the most important underpinnings and instruments of the “mathematization” (i.e., of the “mathematical modeling”) of many problems both in the context of the natural sciences and in the context of the social sciences.

The systematic study of differential equations began in the 1670s by Leibniz. The methods that I present in this chapter are based on the scientific works of Leibniz, Newton, the Bernoulli brothers, Euler, Riccati, Lagrange, and Cauchy.

General Methods for the Solution of Differential Equations

In this section, we shall study different types of ordinary differential equations, and we shall present general methods for finding their general solutions.

The Method of Separation of Variables

This method was originally developed by Leibniz. If a differential equation may be written in the form

$$\frac{dy}{dx} = f(x)g(y),$$

or, similarly, in the form $f(x) + g(y)\frac{dy}{dx} = 0$, then it is said to be solvable by "separation of variables" as follows:

$$\int \frac{dy}{g(y)} = \int f(x)\,dx + c.$$

Remark: In case we have a differential equation of the form

$$y^{(n)} = f(x) \Leftrightarrow \frac{d^n y}{dx^n} = f(x), \qquad (1)$$

then, by integrating (1), we obtain

$$\frac{d^{n-1}y}{dx^{n-1}} = \int f(x)dx + c_1. \qquad (2)$$

By setting $\int f(x)dx = f_1\,(x)$ and then integrating (2), we obtain

$$\frac{d^{n-2}y}{dx^{n-2}} = \int f_1(x)dx + c_1\,x + c_2.$$

Repeating the same process, we obtain the general solution to (1), which is of the form

$$y = w(x) + \frac{c_1}{(n-1)!}x^{n-1} + \frac{c_2}{(n-2)!}x^{n-2} + \cdots + c_n,$$

meaning that the general solution to $y^{(n)} = f(x)$ can be obtained through n successive integrations.

For instance, let us find the general solution to the differential equation

$$x^2dy - ydx = 0,$$

and then let us find its partial solution that satisfies the condition $y(2) = 4$ (i.e., the integral curve that passes through the point $P(2{,}4)$). We shall apply the method of separation of variables:

$x^2dy - ydx = 0 \Rightarrow \frac{dy}{y} = \frac{dx}{x^2} \Rightarrow \frac{dy}{y} = x^{-2}dx \Rightarrow \int \frac{dy}{y} = \int x^{-2}dx \Rightarrow lny = \frac{x^{-1}}{-1} + c \Rightarrow lny = -\frac{1}{x} + c \Rightarrow y = e^{-\frac{1}{x}+c} \Rightarrow y = e^c e^{-\frac{1}{x}} \Rightarrow y = ke^{-\frac{1}{x}}$, which is the general solution to the given differential equation. In order to find the partial solution for which $x = 2 \Rightarrow y = 4$ (i.e., the integral curve that passes through the point $P(2,4)$), we must determine the constant k. If we substitute $x = 2$ and $y = 4$ into the general solution, then we obtain $4 = ke^{-\frac{1}{2}} \Rightarrow k = 4e^{\frac{1}{2}} = 4\sqrt{e}$. Hence, if we substitute this value of k into the general solution, then we shall obtain the required partial solution, namely, $y = 4\sqrt{e}e^{-\frac{1}{x}}$.

Homogeneous Differential Equations

The systematic study of homogeneous differential equations is originally due to Johann Bernoulli, who first applied the term "homogeneous" to differential equations in his research paper "On the Integration of Differential Equations" (1726). A differential equation is said to be "homogeneous" if it may be written in the form

$$f(x,y)dx + g(x,y)dy = 0, \qquad (1)$$

where the functions $f(x,y)$ and $g(x,y)$ are homogeneous with respect to x and y of the same degree of homogeneity, meaning that

$f(x,y)$ may be written in the form $x^m A\left(\frac{y}{x}\right)$ and (2)

$g(x,y)$ may be written in the form $x^m B\left(\frac{y}{x}\right)$. (3)

Thus, due to (2) and (3), (1) becomes (for $x^m \neq 0$):

$$A\left(\frac{y}{x}\right)dx + B\left(\frac{y}{x}\right)dy = 0 \Rightarrow \frac{dy}{dx} = -\frac{A\left(\frac{y}{x}\right)}{B\left(\frac{y}{x}\right)},$$

which ultimately reduces to the form

$$\frac{dy}{dx} = f\left(\frac{y}{x}\right) \Leftrightarrow y' = f\left(\frac{y}{x}\right), \qquad (4)$$

where $f\left(\frac{y}{x}\right)$ is a homogeneous function whose degree of homogeneity is equal to zero. In order to find the general solution to (4), we set

$$\frac{y}{x} = w \Leftrightarrow y = xw \qquad (5)$$

where w is a function of the independent variable x, that is, $w = w(x)$. By differentiating (5), we obtain

$$dy = wdx + xdw,$$

and, after dividing by dx, we obtain

$$\frac{dy}{dx} = w + x\frac{dw}{dx}. \qquad (6)$$

Therefore, due to (5) and (6), the differential equation (4) becomes

$$w + x\frac{dw}{dx} = f(w) \Rightarrow x\frac{dw}{dx} = f(w) - w \Rightarrow \frac{dw}{f(w)-w} = \frac{dx}{x}. \qquad (7)$$

The differential equation (7), which is equivalent to (4), and, therefore, equivalent to (1), can be solved by the method of separation of variables. In particular, (7) yields:

$$\int \frac{dw}{f(w)-w} = lnx + lnc \Rightarrow \int \frac{dw}{f(w)-w} = lncx \Rightarrow cx = e^{\int \frac{dw}{f(w)-w}}. \qquad (8)$$

In (8), we have to compute the integral $\int \frac{dw}{f(w)-w}$ and then to make the substitution $w = \frac{y}{x}$ in order to ultimately find the general solution to (1).

For instance, let us solve the differential equation
$(x^2 - y^2)dx + 2xydy = 0.$

This differential equation is homogeneous, because the expressions $f(x,y) = x^2 - y^2$ and $g(x,y) = 2xy$ are homogeneous with respect to x and y, and, in particular, their degree of homogeneity is 2. We set

$$\frac{y}{x} = w \Leftrightarrow y = xw \qquad (*)$$

where $w = w(x)$. By differentiating $(*)$ with respect to x, we obtain

$$y' = w + x\frac{dw}{dx}. \qquad (**)$$

Due to $(*)$ and $(**)$, the given differential equation becomes
$(x^2 - x^2w^2) + 2x^2w\left(w + x\frac{dw}{dx}\right) = 0 \Rightarrow x^2(1 - w^2) + 2x^2w\left(w + x\frac{dw}{dx}\right) = 0$, and, because, by $(*)$, $x \neq 0$, we divide the last expression by x^2 to obtain

$$(1 - w^2) + 2w\left(w + x\frac{dw}{dx}\right) = 0 \Rightarrow 1 - w^2 + 2w^2 + 2xw\frac{dw}{dx} = 0 \Rightarrow$$
$$1 + w^2 + 2xw\frac{dw}{dx} = 0 \Rightarrow \frac{2wdw}{w^2+1} = -\frac{dx}{x} \Rightarrow \int \frac{2wdw}{w^2+1} =$$
$$-\int \frac{dx}{x} \Rightarrow ln(w^2 + 1) = -lnx + lnc \Rightarrow ln(w^2 + 1) = ln\left(\frac{c}{x}\right) \Rightarrow$$
$$w^2 + 1 = \frac{c}{x}.$$

By the substitution $w = \frac{y}{x}$, we find that the general solution to the given differential equation is $y^2 + x^2 = cx$.

It is worth pointing out that homogeneous equations have important applications in electromagnetism, communication technology, and optics. For instance, homogeneous equations formulated by Maxwell (specifically, $\vec{\nabla} \cdot \vec{E} = \rho/\varepsilon_0$ and $\vec{\nabla} \cdot \vec{B} = 0$, where $\vec{E}$ is the electric field, $\vec{B}$ is the magnetic field, ρ is the electric charge density, and ε_0 is the vacuum permittivity, and, of course, $\vec{\nabla} \cdot$ is the divergence operator) predict the existence of electromagnetic waves, which are fundamental in communication technology, and they are essential in optics, too, because they explain the behavior of light, including reflection (change in the direction of waves when they bounce off a barrier), refraction (change in the direction of waves as they pass from one medium to another), and diffraction (change in the direction of waves as they pass through an opening or around a barrier in their path), and optics is fundamental in various technologies (e.g., lenses, microscopes, telescopes, lasers, and fiber optics communication). Moreover, homogeneous Maxwell's equations underpin our understanding of the electromagnetic spectrum, and this knowledge is important in astronomy, medical imaging (X-rays and MRI), and spectroscopy (the study of absorption and emission of light and other radiation by matter). The equation $\vec{\nabla} \cdot \vec{E} = \rho/\varepsilon_0$ implies that the electric field produced by electric charge diverges from positive charge and converges upon negative charge; and the equation $\vec{\nabla} \cdot \vec{B} = 0$ implies that the divergence of the magnetic field at any point is zero, as well as the assumption that there no magnetic monopoles (a magnetic flux that is generated from magnetic materials is a closed loop, specifically, the direction of the flux lines is from the north pole to the south pole in the atmosphere, so that, in the absence of any poles, these flux lines are unthinkable).

Differential Equations Reducible to Homogeneous Differential Equations

The following methodology is originally due to Johann Bernoulli. The differential equations of the form

$$\frac{dy}{dx} = f\left(\frac{a_1x+b_1y+c_1}{a_2x+b_2y+c_2}\right), \tag{$*$}$$

where $a_1, b_1, c_1, a_2, b_2, c_2$ are real constants, are reducible to homogeneous differential equations. In order to solve $(*)$ by reducing it to a homogeneous differential equation, we distinguish the following two cases:

Case I: If $\frac{a_1}{a_2} \neq \frac{b_1}{b_2} \Leftrightarrow a_1b_2 - a_2b_1 \neq 0$, then we can find the general solution to $(*)$ as follows: We solve the system of equations

$$\begin{Bmatrix} a_1x + b_1y + c_1 = 0 \\ a_2x + b_2y + c_2 = 0 \end{Bmatrix}. \tag{1}$$

Let $(x, y) = (x_0, y_0)$ be the solution to (1). Then we set

$$\begin{Bmatrix} x = x_0 + w \\ y = y_0 + v \end{Bmatrix}, \tag{2}$$

where $w = w(x)$ and $v = v(x)$, and, by differentiating (2), we obtain

$$\begin{Bmatrix} dx = dw \\ dy = dv \end{Bmatrix}, \tag{3}$$

so that, by (2) and (3), the differential equation $(*)$ becomes

$$\frac{dv}{dw}=f\left(\frac{a_1(x_0+w)+b_1(y_0+v)+c_1}{a_2(x_0+w)+b_2(y_0+v)+c_2}\right)\Rightarrow\frac{dv}{dw}=f\left(\frac{a_1x_0+b_1y_0+c_1+a_1w+b_1v}{a_2x_0+b_2y_0+c_2+a_2w+b_2v}\right).$$

But $a_1x_0+b_1y_0+c_1=0$ and $a_2x_0+b_2y_0+c_2=0$, because (x_0,y_0) is the solution to (1), and, therefore,

$$\frac{dv}{dw}=f\left(\frac{a_1w+b_1v}{a_2w+b_2v}\right). \tag{4}$$

The differential equation (4) is homogeneous with respect to v and w, and, in order to find its general solution, we set $\frac{v}{w}=z\Leftrightarrow v=wz$, where $z=z(w)$, and we work according to the method of solving homogeneous differential equations, which I have already explained. When we find the general solution to (4), we set $z=\frac{v}{w}$, and then, by (2), we set $w=x-x_0$ and $v=y-y_0$ in order to ultimately find the general solution to $(*)$.

Case II: If $\frac{a_1}{a_2}=\frac{b_1}{b_2}=\lambda\Rightarrow a_1b_2-a_2b_1=0$, then we can find the general solution to $(*)$ as follows: Because $a_1=\lambda a_2$ and $b_1=\lambda b_2$, $(*)$ becomes

$$\frac{dy}{dx}=f\left(\frac{\lambda(a_2x+b_2y)+c_1}{a_2x+b_2y+c_2}\right). \tag{5}$$

We set $a_2x+b_2y=w$, where $w=w(x)$, and, by differentiating with respect to x, we obtain $a_2+b_2y'=w'\Leftrightarrow y'=\frac{1}{b_2}(w'-a_2)$, so that (5) becomes

$$\frac{1}{b_2}\left(\frac{dw}{dx}-a_2\right)=f\left(\frac{\lambda w+c_1}{w+c_2}\right). \tag{6}$$

The differential equation (6) can be solved by the method of separation of variables, which I have already explained. When we find the general solution to (6), we set $w=a_2x+b_2y$ in order to ultimately find the general solution to $(*)$.

First-Order Linear Differential Equations

The following methodology is originally due to L. Euler and Leibniz. The general form of these equations is

$$\frac{dy}{dx}+Ay=B \tag{$*$}$$

where A and B are functions of x, that is, $A=A(x)$ and $B=B(x)$. In other words, the dependent variable and all of its derivatives appear in a linear fashion (recall that "linearity" is a property of functions, meaning that a function $f(x)$ is linear if and only if $f(x+y)=f(x)+f(y)$ and $f(kx)=kf(x)$ for any constant k). The general solution to $(*)$ is:

$$y=e^{-\int Adx}\left(c+\int Be^{\int Adx}\,dx\right)$$

where c is an arbitrary constant.

Proof: If $B(x) = 0$, then $(*)$ becomes $\frac{dy}{dx} + Ay = 0$, and it is said to be a homogeneous linear differential equation, which can be solved by separation of variables: $\frac{dy}{y} = -Adx \Rightarrow \int \frac{dy}{y} = -\int Adx + c \Rightarrow lny = -\int Adx + c \Rightarrow y = e^{-\int Adx + c} = e^{c}e^{-\int Adx} = ce^{-\int Adx}$, which is the general solution to the aforementioned homogeneous linear differential equation. If $c = 1$, then we obtain its partial solution $y_1 = e^{-\int Adx}$.

If $A(x)$ and $B(x)$ are constant functions, then $(*)$ is solvable by separation of variables.

In order to find the general solution to $(*)$, we consider a new unknown function z of x such that

$$y = y_1 z, \tag{1}$$

where, as I have already mentioned, y_1 is a partial solution to $\frac{dy}{dx} + Ay = 0$.

By differentiating (1) with respect to x, we obtain

$$y' = y_1' z + y_1 z'. \tag{2}$$

Hence, by (1) and (2), the differential equation $(*)$ becomes

$y_1' z + y_1 z' + Ay_1 z = B \Leftrightarrow (y_1' + Ay_1)z + y_1 z' = B.$

But $y_1' + Ay_1 = 0$, since y_1 is a partial solution to $\frac{dy}{dx} + Ay = 0$, and, therefore, since $y_1 = e^{-\int Adx}$,

$y_1 z' = B \Rightarrow e^{-\int Adx} z' = B \Rightarrow z' = Be^{\int Adx} \Rightarrow z = \int Be^{\int Adx} dx + c.$

Because $y_1 = e^{-\int Adx}$ and $z = \int Be^{\int Adx} dx + c$, equation (1) gives the general solution to $(*)$, which is

$y = e^{-\int Adx}\left(c + \int Be^{\int Adx}\, dx\right).$■

A simple example is the following: The differential equation $y' - 2xy = x - x^3$ is a linear differential equation, whose general solution is given by the above formula where $A = -2x$ and $B = x - x^3$, so that

$$y = e^{\int 2xdx}\left(c + \int (x - x^3)\, e^{-\int 2xdx} dx\right) = ce^{x^2} + \frac{1}{2}x^2$$

(the above formula is a general method for solving linear differential equations).

Remark: If y is the general solution to $(*)$, and if y_1 and y_2 are two partial solutions to $(*)$, then the ratio

$$\frac{y - y_1}{y_2 - y_1}$$

is constant. The difference between two solutions of the linear differential equation (i.e., of $(*)$) yields a solution of the corresponding homogeneous linear differential equation (i.e., of $y' + Ay = 0$). If y_1 is a partial solution to $(*)$, then the general solution to $(*)$ is given by

$$y = ce^{-\int A dx} + y_1$$

(that is, the general solution of the linear differential equation is the sum of the general solution of the corresponding homogeneous linear differential equation and a partial solution of the linear differential equation). For instance, given the linear differential equation

$y' + \frac{x-1}{x} y = -x$,

we can find its general solution, by working as follows: Firstly, we shall find the general solution to the corresponding homogeneous linear differential equation, that is, to $y' + \frac{x-1}{x} y = 0$, and, in particular, we obtain: $y' + \frac{x-1}{x} y = 0 \Rightarrow \frac{dy}{y} = \frac{1-x}{x} dx \Rightarrow \int \frac{dy}{y} = \int \frac{dx}{x} - \int dx + c_1 \Rightarrow y = \frac{cx}{e^x}$.

Now, we shall find a partial solution to the original linear differential equation, and it will be of the form

$y = c(x) \frac{x}{e^x}$.

Thus, $y = c(x) \frac{x}{e^x} \Rightarrow y' = c'(x)xe^{-x} + c(x)e^{-x} - c(x)xe^{-x}$.

Then, by substituting $y = \frac{cx}{e^x}$ and $y' = c'(x)xe^{-x} + c(x)e^{-x} - c(x)xe^{-x}$ into the original linear differential equation, we obtain

$c'(x)xe^{-x} = -x \Rightarrow c'(x) = -e^x \Rightarrow c(x) = -\int e^x dx \Rightarrow c(x) = -e^x$,

so that $y = c(x) \frac{x}{e^x}$ yields $y = -e^x \frac{x}{e^x} \Rightarrow y = -x$.

Hence, the general solution to the original linear differential equation is the sum of $y = \frac{cx}{e^x}$ and $y = -x$, that is, $y = \frac{cx}{e^x} - x$.

Linear Systems (LS), Nonlinear Systems (NLS), and Linearization of Nonlinear Differential Equations

By a “nonlinear system,” we mean a phenomenon whose behavior can be described by a model that is a nonlinear differential equation. The characteristic properties of linear and nonlinear systems can be studied in relation to the following concepts and issues:

- *The principle of superposition:* This principle, originally stated by Daniel Bernoulli (1775), consists of two properties: (i) the sum of any number of linearly independent partial solutions of a differential equation is also a solution of the given differential equation; and (ii) any constant multiple of a solution is also a solution. This principle holds in linear systems (LS), but, generally, it does not hold in nonlinear systems (NLS).

- *The global property:* This property characterizes the LS, but not the NLS; that is, in LS, the local behavior of the solutions yields their global behavior, whereas, in general, the global behavior of a NLS cannot be deduced from its local behavior. The solutions of NLS may not be extendible beyond a certain time or may not be defined for all values of time.
- *Limit cycles:* Periodic phenomena of LS and NLS correspond to closed trajectories, called "cycles," whose period is a finite number. If the cycles are isolated, in the sense that, in a neighborhood of them, no other cycles exist, then they are called "limit cycles." Limit cycles may exist in a NLS but never in a LS. If a LS has a periodic solution y, then, due to the principle of superposition, ky is also a periodic solution for any constant k; therefore, no limit cycles exist in LS. On the other hand, the special nature of the nonlinearities of some nonlinear differential equations (NLDE) may lead to limit cycles. For several examples, see: E. C. Zeeman, "Stability of Dynamical Systems," *Nonlinearity*, vol. 1, 1988, pp. 115–155.
- *Self-excited oscillations:* Self-excited oscillations are special periodic phenomena corresponding to limit cycles, and, therefore, they may be produced in NLS, but never in LS. In particular, they may be produced in NLS where the nonlinearities appear in damping forces (i.e., forces that act to "damp," reduce, attenuate the amplitude of an oscillation) and no external force is exerted.
- *Subharmonic phenomena:* Subharmonic phenomena may be produced in NLS, but, generally, not in LS. They occur when the systems undergo external periodic forces. If the frequency of the external force is ω, then the system on which the force is exerted may undergo periodic motions with frequency $\frac{\omega}{n}$, where $n = 2,3, \ldots$, and such motions are called "subharmonic oscillations" of order $\frac{1}{n}$, where $n = 2,3, \ldots$ For instance, an aerodynamical model of subharmonic oscillations may be due to the fact that certain parts of an airplane may incur violent oscillations by an engine running with frequency much larger than the frequency of the oscillating parts.
- *Amplitude and frequency of periodic solutions of free linear and nonlinear systems:* The amplitude of the periodic solutions of free (unforced) LS is independent of the frequency, and the frequency

is the same for all trajectories. By contrast, in NLS, the amplitude depends on the frequency, and the frequency changes from one trajectory to another.

- *Resonance phenomena:* The resonance phenomena may occur in forced LS and NLS when the free frequency of the system becomes (almost) equal to the frequency of the external force. For instance, when a group of soldiers marches in step over a suspension bridge, the feet of the group exert a periodic force on the bridge; and, if the period of marching equals the natural period of the bridge, then resonance occurs, and the sustained bridge oscillations may even bring on the collapse of the bridge. However, the nonlinearity of a NLS can prevent resonance, even in the absence of damping, because, as frequency is changed, resonance ceases. For a detailed study of resonance phenomena, see: J. K. Hale and J. P. LaSalle, "Differential Equations: Linearity vs. Nonlinearity," *SIAM Review*, vol. 5, 1963, pp. 249–272.
- *Hysteresis phenomena:* Jump discontinuities, or hysteresis phenomena, may occur in damped forced NLS, but not in LS. In particular, there are regions where the amplitude of the oscillations jumps discontinuously, and, in these regions, the oscillations are unstable. In general, "hysteresis" means a lag between input and output in a system upon a change in direction. In engineering, the problem of vibration is of great importance, since it refers to the oscillation or movement of objects or systems around an equilibrium point, and, in certain scenarios, excessive vibrations can give rise to several issues, ranging from discomfort and noise to catastrophic system failure. In particular, vibration problems occur when different variables, such as mass, elasticity, and damping, interact within a system, and, when these variables create a disruptive back-and-forth movement, this is a key sign that there is a vibration problem, specifically, the system is not perfectly balanced. Mathematically, the vibration problem can be formulated using Newton's Second Law of Motion as follows:

  $$ma = F - bv$$

  where m denotes the object's mass, a denotes acceleration, F denotes the net force acting on the object, and bv accounts for the damping force, where b is the damping coefficient, and v denotes velocity.

- *Combination frequencies:* Hermann von Helmholtz and H. Poincaré were the first scientists to observe that, except for certain fundamental frequencies ω_1 and ω_2 in a NLS, there exist solutions of the same differential equation with frequencies $\omega = a\omega_1 + b\omega_2$, where a and b are integers; and these are called "combination frequencies" of the system, and they are phenomena of NLS (difficulties emanating from combination frequencies may be circumvented by using viscous damping in the system; notice that, when an oscillating body is subjected to viscous drag, the kinetic energy of the body is dissipated at a much faster rate than if the oscillating body was placed in the air, and viscous drag often causes the oscillating system to be overdamped, which results in the oscillations quickly dying down to zero). In fact, Hermann von Helmholtz observed that, just as seemingly pure, white light actually contains all the colors of the rainbow, clearly defined musical notes are composed of many different tones, and, thus, for instance, if you play the A above the middle C on a musical instrument, then the sound that you hear has a clearly defined "fundamental" pitch of $440Hz$, but the sound does not only contain a simple "fundamental" vibration at $440Hz$, but also a "harmonic series" of integral multiples of this frequency, called "overtones" (e.g., $880Hz$, $1320Hz$, $1760Hz$, etc.).

Problems of nonlinear analysis started to exist ever since the creation of the universe. Some of them were solved by ancient Greek mathematicians, but many new nonlinear problems were created, both in pure mathematics and in other sciences, such as biology, physics, astronomy, economics, etc. The distinction between linear and nonlinear analysis is not quite clear, because a considerable part of information about a nonlinear system can be extracted from a linear approximation of the corresponding nonlinear problem. Moreover, it is often possible to extract information about the solution to a linear system from a relevant nonlinear one, and this fact was explicitly studied by the Russian-American mathematician Victor Lomonosov in his research paper "Invariant Subspaces for the Family of Operators which Commute with a Completely Continuous Operator" (published in *Functional Analysis and Its Applications*, vol. 7, 1973, pp. 213–214). The term "linearization" of a nonlinear differential equation refers to the reduction of a nonlinear differential equation to a linear differential equation that is either equivalent or almost equivalent to the given nonlinear differential equation, that is, the solution to the linear differential equation may give the solution to the nonlinear differential equation either exactly or approximately within an acceptable error.

Exact methods of linearization: Two well-known examples of exact linearization of nonlinear differential equations are the Bernoulli equation and the Riccati equation. The Bernoulli equation and the Riccati equation are special because they are nonlinear differential equations with known exact solutions, which are obtainable through linearization.

i. *The Bernoulli equation:*

$$\frac{dy}{dx} + Ay = By^n, \tag{1}$$

where A and B are functions of x, and $n \in \mathbb{R} - \{0,1\}$ (if $n = 0$, then the equation is linear; if $n = 1$, then the equation can be solved by separation of variables). Multiplying both sides of (1) by y^{-n}, we obtain

$$y^{-n}\frac{dy}{dx} + Ay^{1-n} = B. \tag{2}$$

Let $y^{1-n} = w,$ (3)

where $w = w(x)$. By differentiating (3) with respect to x, we obtain

$$(1-n)y^{-n}\frac{dy}{dx} = \frac{dw}{dx} \Leftrightarrow y^{-n}y' = w'/(1-n). \tag{4}$$

Hence, (2), due to (3) and (4), yields

$$\frac{w'}{1-n} + Aw = B \Rightarrow \frac{dw}{dx} + (1-n)Aw = (1-n)B, \tag{5}$$

which is a first-order linear differential equation (whose dependent variable is w), and it can be solved according to the aforementioned method of solving first-order linear differential equations. When we find the general solution to (5), we set $w = y^{1-n}$, according to (3), and, thus, we obtain the general solution to (1).

Remark: The Bernoulli equation was originally discussed in a work of 1695 by Jacob Bernoulli, after whom it is named, but the earliest solution to this equation was obtained by Gottfried Leibniz, who published it in 1696.

ii. *The Riccati equation:*

$$\frac{dy}{dx} + A + By + Cy^2 = 0, \tag{1}$$

where A, B, and C are functions of x. We can find the general solution to the Riccati equation only if we know one of its partial solutions. Suppose that $y = y_1$ is a partial solution to (1), so that

$$\frac{dy_1}{dx} + A + By_1 + Cy_1^2 = 0. \tag{2}$$

Then, by setting

$$y = y_1 + w, \tag{3}$$

where $w = w(x)$, and differentiating (3) with respect to x, we obtain

$$\frac{dy}{dx} = \frac{dy_1}{dx} + \frac{dw}{dx}. \tag{4}$$

Hence, (1), due to (3) and (4), yields

$$\frac{dy_1}{dx} + \frac{dw}{dx} + A + B(y_1 + w) + C(y_1 + w)^2 = 0,$$

which ultimately becomes

$$\frac{dy_1}{dx} + A + By_1 + Cy_1^2 + \frac{dw}{dx} + (B + 2Cy_1)w + Cw^2 = 0.$$

But, due to (2), $\frac{dy_1}{dx} + A + By_1 + Cy_1^2 = 0$ (since y_1 is a partial solution), so that we obtain

$$\frac{dw}{dx} + (B + 2Cy_1)w + Cw^2 = 0,$$

and, hence,

$$\frac{dw}{dx} + (B + 2Cy_1)w = -Cw^2, \tag{5}$$

which is a Bernoulli equation (where w is the dependent variable), and it can be solved according to the aforementioned general method of solving the Bernoulli equation (in this case, we begin by multiplying both sides of (5) by w^{-2}, etc.). By substituting the value of w that we receive from the general solution to (5) into equation (3), that is, into $y = y_1 + w$, we obtain the general solution to (1).

Notice that, alternatively, we can solve the Riccati equation by setting $y = y_1 + \frac{1}{w} \Leftrightarrow y' = y_1' - \frac{w'}{w^2}$ (where, as above, y_1 is a partial solution to (1), and $w = w(x)$), but then the Riccati equation will reduce to a linear differential equation, which is solvable according to the aforementioned general method of solving linear differential equations.

Remark: This equation is named after the Italian mathematician Jacopo Francesco (Count) Riccati (1676–1754), who wrote on philosophy, physics, and differential equations.

Approximate methods of linearization: Usually, the nonlinear differential equations that come from applied mathematics cannot be linearized by exact methods, and, therefore, we search for approximate methods of linearization, which give approximations of particular solutions. Let us consider a general nonlinear differential equation in its normal form:

$x_i' = f_i(t, x_1, \dots, x_n),$ (∗)
where $i = 1,2, \dots, n$, and the functions f_i are such that there exists a unique solution of (∗) through any point x_0 in the region of the validity of (∗). We frequently write (∗) as follows:
$x_i' = Ax + X,$ (∗∗)
where x_i', x, and X are n-column matrices, and A is an $n \times n$-matrix (either constant or time-dependent). In (∗∗), X is the set of the nonlinearities of (∗), and $X(t, 0) \equiv 0$. Then the system
$x_i' = Ax$ (∗∗∗)
is the linear part of (∗); and (∗∗∗) is said to be the "first approximation" of (∗). It should be mentioned that not always the results of approximate linearization are acceptable, depending mainly on the nature of the problem under consideration.

Differential Equations of the Form $f(y') = 0$ where $f(y')$ Is an Integral Polynomial in y'

Suppose that the degree of the polynomial $f(y')$ is n. If the roots of the polynomial are $r_1, r_2, \dots, r_n$, then we obtain the relations $y' = r_1, y' = r_2, \dots, y' = r_n$, which, by integration, yield $y = r_1x + c, y = r_2x + c, \dots, y = r_nx + c$, and, therefore, $y - r_1x - c = 0, y - r_2x - c = 0, \dots, y - r_nx - c = 0$ .Then the general solution to the differential equation is
$(y - r_1x - c)(y - r_2x - c) \dots (y - r_nx - c) = 0,$
which defines a family of n parallel straight lines on the xy-plane.
For instance, the differential equation $(y')^3 - 2(y')^2 - y' + 2 = 0$ is of the first order and the third degree. The roots of the corresponding algebraic equation in y' are $y' = -1$, $y' = 1$, and $y' = 2$, or $\frac{dy}{dx} = -1$, $\frac{dy}{dx} = 1$, and $\frac{dy}{dx} = 2$, and, therefore, we have $dy = -dx$, $dy = dx$, and $dy = 2dx$, which, by integration, yield $y = -x + c$, $y = x + c$, and $y = 2x + c$. Then the general solution to the given differential equation is $(y + x - c)(y - x - c)(y - 2x - c) = 0.$

Differential Equations that Do Not Include the Unknown Function

These differential equations are of the form $F(x, y') = 0$. There are two ways in which we can find the general solution to such a differential equation:

First way: We solve for y', thus obtaining $y' = g(x)$, which is solvable by separation of variables, and its general solution is $y = \int g(x)dx + c$.
Second way: We solve for x, in which case we obtain $x = f(y')$, and we set
$y' = p \Leftrightarrow dy = pdx.$ (1)
Then
$x = f(p),$ (2)
and $dx = f'(p)dp$. Substituting the value of dx into (1), we obtain $dy = pf'(p)dp$, which, by integration, yields
$y = \int pf'(p)dp + c.$ (3)
The above relations (2) and (3) imply that the general solution to the given differential equation is obtained in terms of the parameter p and in the following parametric form:
$x = f(p)$ and $y = \int pf'(p)dp + c.$
For instance, let us consider the differential equation $x(y')^2 - 1 = 0$, which can be solved both with respect to y' and with respect to x.
If we solve this differential equation with respect to y', then we have: $y' = \pm\frac{1}{\sqrt{x}}$, and, by integration, we obtain $y = \pm\int\frac{1}{\sqrt{x}}dx + c = \pm\int x^{-\frac{1}{2}}dx + c = \pm\frac{x^{\frac{1}{2}}}{\frac{1}{2}} + c = \pm 2\sqrt{x} + c$, and, hence,

$$(y-c)^2 = 4x$$

(this is the general solution to the given differential equation).
If we solve this differential equation with respect to x, then we have: $x = \frac{1}{(y')^2}$, and we set $y' = p \Leftrightarrow dy = pdx$, so that $x = \frac{1}{p^2}$, and $dx = -\frac{2}{p^3}dp$. Therefore, $dy = pdx \Rightarrow dy = p\left(-\frac{2}{p^3}\right)dp = -\frac{2}{p^2}dp$, which, by integration, yields $y = -\int\frac{2}{p^2}dp + c = \frac{2}{p} + c$. Hence, the parametric form of the general solution to the given differential equation is:

$$x = \frac{1}{p^2}$$

and

$$y = \frac{2}{p} + c$$

(as I have already explained, p is the parameter, since $x = x(p)$).
In a similar fashion, we can solve differential equations of the form $F(y, y') = 0$, that is, differential equations that do not include the independent variable x.

Second-Order and Higher-Order Linear Differential Equations

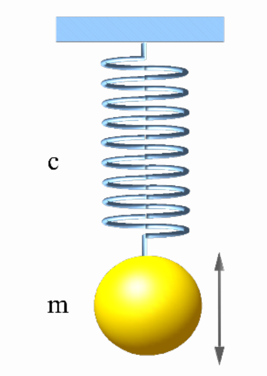
The generic second-order linear ordinary differential equation with constant coefficients has the form

$$ay'' + by' + cy = F(x)$$

where a, b, and c are real constants, $F(x)$ is a given function of the independent variable x, and, obviously, the differentiation of the uknown function is symbolized as follows: $y'' \equiv \frac{d^2y}{dx^2}$ and $y' = \frac{dy}{dx}$.

The systematic study of second-order linear ordinary differential equations with constant coefficients has been significantly motivated by physics. For instance, one physical system whose behavior is governed by a second-order linear ordinary differential equation with constant coefficients is the linear mechanical oscillator shown in Figure 10-1, where we see a linear mechanical oscillator consisting of a mass m attached to a rigid wall by a linear spring of spring stiffness c and a damper of damping constant k, a time-dependent force $F(t)$ is applied to the mass m, and the displacemt of the mass from its rest position is represented by $x(t)$.

Figure 10-1: A linear mechanical oscillator (source: Wikimedia Commons: Author: Lokilech; https://commons.wikimedia.org/wiki/File:Feder_Masse_Schwinger.svg).



The behavior of the mass–spring system shown in Figure 10-1 (with an applied external force) is governed by the second-order linear ordinary differential equation with constant coefficients

$$mx''(t) + kx'(t) + cx(t) = F(t)$$

where m denotes the mass of the particle attached to the spring, k is a measure of the strength of the damper, c represents the spring stiffness, $F(t)$ is the applied external force, t denotes time, $x(t)$ is the displacement of the mass from its rest position, the term $kx'(t)$ represents the force exerted by the damper on the mass (and, in case of a linear damper, this force is proportional to the velocity x', and it resists the motion), the term $cx(t)$ represents the force exerted by the spring on the mass (and, in case of a linear spring, this force is proportional to the displacement x, and it acts in the direction opposite to the displacement, that is, it is a "restoring force"), and this second-order linear ordinary differential equation represents Newton's Second Law of Motion, according to which force equals mass times acceleration, that is, $mx''(t)$.

The Homogeneous Equation: If $F(x) = 0$, then the generic second-order linear ordinary differential equation with constant coefficients reduces to its "homogeneous" form:

$$ay'' + by' + cy = 0. \tag{1}$$

A typical solution of the homogeneous equation (1) is in the form

$$y = e^{rx}$$

where r is a constant to be determined; and, thus, $y' = re^{rx}$, and $y'' = r^2e^{rx}$. Hence, substituting these values of y, y', and y'' into the homogeneous equation (1), we obtain the corresponding "characteristic" (or "auxiliary") equation:

$$ar^2e^{rx} + bre^{rx} + ce^{rx} = 0 \Rightarrow e^{rx}(ar^2 + br + c) = 0 \Rightarrow ar^2 + br + c = 0.$$

In this way, we have transformed the given ordinary differential equation into the "characteristic polynomial"

$$r^2 + br + c = 0,$$

which is a quadratic equation in the unknown r, and then we have to solve for r, using the quadratic formula

$$r = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Therefore, we have to consider three cases:

Firstly, if the discriminant $b^2 - 4ac > 0$, then the quadratic equation has two distinct real roots, say r_1 and r_2, and, in this case, by the principle of superposition, the general solution to the homogeneous equation (1) is

$$y = c_1e^{r_1x} + c_2e^{r_2x}$$

for any value of the two constants c_1 and c_2 (and all we have to do is to substitute r_1 and r_2 into this formula; and, when we have to solve an initial-value problem or a boundary-value problem, we have to solve for c_1 and c_2).

Secondly, if the discriminant $b^2 - 4ac = 0$, then the quadratic equation has a repeated real root, say r, and, in this case, the general solution to the homogeneous equation (1) is

$$y = c_1e^{rx} + c_2xe^{rx}$$

for any value of the two constants c_1 and c_2 (the second exponential is multiplied by x because in this way it represents a second independent solution).

Thirdly, if the discriminant $b^2 - 4ac < 0$, then the quadratic equation has two complex (conjugate) roots, say $r_1 = a + bi$ and $r_2 = a - bi$, and then the general solution to the homogeneous equation (1) is

$$y = e^{ax}[c_1cos(bx) + c_2sin(bx)]$$

for any value of the two constants c_1 and c_2. Notice that, due to the special nature of r_1 and r_2, and due to the fact that

$$e^{a+bi} = e^a(cosb + isinb)$$

(by the definition of the complex exponential and Euler's formula), in this case, we have:

$$y = c_1e^{r_1x} + c_2e^{r_2x} \Rightarrow y = e^{ax}[c_1cos(bx) + c_2sin(bx)]$$

(for (different) arbitrary constants c_1 and c_2).

For instance, let us consider "simple harmonic motion": Let us consider a spring whose upper end is securely fastened (of natural, that is, unstretched, length l_0), and suppose that we attach to it an object of mass m, so that the addition of the mass m stretches the spring to length l. This static elongation of the spring is the result of two forces, namely: the force of gravity (i.e., $F_g = mg$), acting downward, and the spring force (F_s), acting upward. The model of simple harmonic motion is based on the following assumptions: (i) All motion is along a vertical line through the center of gravity of the object, and the object is treated as a point mass. (ii) There is no damping force due to the medium in which the mass is moving (e.g., we ignore air resistance). (iii) No other forces (except for the ones already mentioned) are applied to the mass. According to Hooke's Law, if a spring is stretched (or compressed) x units from its natural (equilibrium) length, then it exerts a force that is proportional to x, so that $restoring\ force = -kx$ (where k is a positive constant, and $x = x(t)$ is displacement); and, therefore, by substituting this equation into Newton's Second Law of Motion (force equals mass times acceleration: $F_{net} = m\frac{d^2x}{dt^2}$), and, ignoring any external resisting forces, we obtain the following second-order linear differential equation (which describes "simple harmonic motion"): $-kx = m\frac{d^2x}{dt^2} \Leftrightarrow \frac{d^2x}{dt^2} + \frac{k}{m}x = 0$, whose typical solution is $x = e^{rt}$, and, hence, we have:

$$r^2e^{rt}+\frac{k}{m}e^{rt}=0\Rightarrow e^{rt}\left(r^2+\frac{k}{m}\right)=0\Rightarrow r^2+\frac{k}{m}=0\Rightarrow r=\pm\sqrt{-\frac{k}{m}}$$
$$=\pm i\sqrt{\frac{k}{m}}$$

and, for simplicity, we set $\lambda=\sqrt{\frac{k}{m}}$. Hence, the two values of r are $r=\pm i\lambda$, which means that the two solutions (for $x=x(t)$) are $e^{+i\lambda t}$ and $e^{-i\lambda t}$, which are two functions that satisfy the differential equation $\frac{d^2x}{dt^2}+\frac{k}{m}x=0$. However, according to the principle of superposition, the most general solution, which describes all the solutions to this differential equation is a linear combination of these two solutions, namely: $x(t)=c_1e^{+i\lambda t}+c_2e^{-i\lambda t}$, for any value of the two constants c_1 and c_2. By Euler's formula, we obtain

$$e^{+i\lambda t}=cos(\lambda t)+isin(\lambda t)$$

and

$$e^{-i\lambda t}=cos(-\lambda t)+isin(-\lambda t)=cos(\lambda t)-isin(\lambda t)$$

(given that $cosine$ is even, and $sine$ is odd). Therefore, our solution $x(t)$ can be written as follows:

$$x(t)=c_1[cos(\lambda t)+isin(\lambda t)]+c_2[cos(\lambda t)-isin(\lambda t)]\Rightarrow x(t)$$
$$=(c_1+c_2)cos(\lambda t)+i(c_1-c_2)sin(\lambda t)$$

where, since c_1 and c_2 are constants, we set $c_1+c_2=A$ and $i(c_1-c_2)=B$ for some new constants A and B, thus obtaining

$$x(t)=Acos(\lambda t)+Bsin(\lambda t)$$

(and this is the general solution in terms of real functions). Notice that, in SI units, displacements are measured in meters (m), and forces are measured in neutons (N), and, therefore, the spring constant k is measured in newotns per meter.

In the aforementioned model of simple harmonic motion, we assumed that the only forces involved were gravity and the spring force. However, in order to account for such things as friction in the spring and air restinance, we must assume that there is a damping force (i.e., a force that tends to slow the motion of the given object, which we still treat as a point mass), and this damping force can be thought of as the resultant of all other external forces acting on the given object (the magnitude of the damping force is proportional to the velocity of the particle). Therefore, we have to add a term $\delta\frac{dx}{dt}$, where δ is the damping constant, so that we come up with a new homogeneous equation, which describes (free) damped vibrations, namely:

$$\frac{d^2x}{dt^2}+\frac{\delta}{m}\frac{dx}{dt}+\frac{k}{m}x=0$$

(and we can solve this homogeneous equation according to the aforementioned method). Given that the motion of this point mass is determined by the inherent forces of the spring–mass system and the natural forces acting on the system, the vibrations are called “free vibrations.” Things change dramatically if we assume that the point mass is also subject to an external periodic force $F_0 sin\varphi t$, which is due to the motion of the object to which the upper end of the spring is attached, so that, in this case, the mass undergoes “forced vibrations,” which are described by the differential equation

$$\frac{d^2x}{dt^2}+\frac{\delta}{m}\frac{dx}{dt}+\frac{k}{m}x=\frac{F_0 sin\varphi t}{m}$$

(where $F_0 sin\varphi t \neq 0$), and, obviously, this differential equation is not homogeneous; this is a non-homogeneous second-order linear differential equation.

The Non-homogeneous Equation (forced motions): When we are dealing with non-homogeneous second-order linear differential equations, that is, with differential equations of the form

$y''+a_1y'+a_2y=F(x)$ where $F(x)\neq 0$, (2)

the general solution to (2) is given by the formula

$$y(x)=y_p(x)+y_c(x)$$

where $y_p(x)$ is a partial solution to the non-homogeneous differential equation (2), and $y_c(x)$ is the general solution to the corresponding homogeneous differential equation, that is, to $y''+a_1y'+a_2y=0$ (notice that, in physics, $F(x)$ in (2) represents the forcing function). Obviously, $y_c(x)$ can be found by applying the aforementioned method of solving homogeneous second-order linear ordinary differential equations. However, $y_p(x)$ can be found by considering $F(x)$ and making the proper algebraic tasks, since the choice for the partial solution should match the structure of the right-hand side of the non-homogeneous differential equation. In particular, in order to determine the form of the partial solution $y_p(x)$, we distinguish the following cases:

Case 1: If $F(x)$ has the form be^{ax}, then $y_p(x)=Ae^{ax}$. Notice that $y_p(x)$ contains the same exponential as $F(x)$, but A is unknown and must be determined such that the given ordinary differential equation is fulfilled.

Case 2: If $F(x)$ has the form $ax^n + (lower\ order\ powers\ of\ x)$, then $y_p(x) = c_nx^n + c_{n-1}x^{n-1} + \cdots + c_0$. Notice that $y_p(x)$ is a complete integral polynomial in x whose degree is equal to the degree of $F(x)$, and its coefficients must be determined such that the given ordinary differential equation is fulfilled.
Case 3: If $F(x)$ has the form $pcosax$ or $qsinax$, then $y_p(x) = Acosax + Bsinax$; and the coefficients A and B must be determined such that the given ordinary differential equation is fulfilled. Notice that the most important form of forcing in engineering is the harmonic forcing, where $F(x)$ has the form $pcosax$ or $qsinax$.
The aforementioned methods can be generalized to solve corresponding nth-order linear differential equations.

Example: Let us find the general solution to the differential equation
$y'' - y = x^2 + x + 1.$ (1)
The general solution to this differential equation is of the form $y(x) = y_p(x) + y_c(x)$ where $y_c(x)$ is the general solution to the corresponding homogeneous differential equation, that is, to $y'' - y = 0$. According to the above method of solving homogeneous second-order linear differential equations, in order to find $y_c(x)$, we solve the corresponding characteristic equation, which can be found by setting $y = e^{rx}$ and $y'' = r^2e^{rx}$, so that $y'' - y = 0 \Rightarrow r^2e^{rx} - e^{rx} = 0 \Rightarrow e^{rx}(r^2 - 1) = 0$. The roots of $r^2 - 1 = 0$ are $r_1 = 1$ and $r_2 = -1$, and, therefore,
$y_c(x) = c_1e^x + c_2e^{-x}.$ (2)
In order to find a partial solution $y_p(x)$ to (1), we set
$y = Ax^2 + Bx + C,$ (3)
that is, a complete integral polynomial in x whose degree is equal to the degree of $F(x) = x^2 + x + 1$. Now, we must determine the coefficients A, B, and C in order for the given differential equation to be fulfilled, and we can do this as follows:
By differentiating (3) twice with respect to x, we obtain
$y' = 2Ax + B$ and $y'' = 2A.$ (4)
Due to (3) and (4), the differential equation (1) can be written as follows:

$$2A - Ax^2 - Bx - C = x^2 + x + 1$$

$\Leftrightarrow -Ax^2 - Bx + 2A - C = x^2 + x + 1.$ (5)
We want both sides of the equation (5) to be identically equal to each other, and, therefore, the coefficients of the corresponding x's must be equal to each other, namely: $-A = 1 \Leftrightarrow A = -1$, $-B = 1 \Leftrightarrow B = -1$, and $2A - C = 1 \Leftrightarrow C = -3$. Substituting these values of A, B, and C into (3), we obtain a partial solution to (1), specifically:

$y_p(x) = -x^2 - x - 3.$

Hence, the general solution to the differential equation (1) is

$y_p(x) + y_c(x) = -x^2 - x - 3 + c_1e^x + c_2e^{-x}.$

Systems of Differential Equations

By the term "system of differential equations," we refer to two or more simultaneous differential equations that contain an independent variable x, the dependent variables $y, z, \dots$, as well as the derivatives of the dependent variables $y, z, \dots$ with respect to the independent variable x. The order of a system of differential equations is the sum of the orders of the highest-order derivatives that appear in the equations of the given system. For instance, the order of the system

$$\begin{Bmatrix} y''' + z' + x = 0 \\ z'' + y' + 3x = 0 \end{Bmatrix},$$

where the order of the highest-order derivative in the first equation is 3, and the order of the highest-order derivative in the second equation is 2, is $3 + 2 = 5$.

The systems of differential equations with constant coefficients where the number of equations is equal to the number of dependent variables are usually solved by means of differentiation and term deletions, so that we ultimately come up with one differential equation. For instance, consider the following system:

$$\begin{Bmatrix} \frac{dy}{dx} - 3y + z = 0 \\ \frac{dz}{dx} - 4y + z = 0 \end{Bmatrix}. \tag{1}$$

We differentiate the first equation of the system (1) with respect to x, so that we obtain

$$\frac{d^2y}{dx^2} - 3\frac{dy}{dx} + \frac{dz}{dx} = 0. \tag{2}$$

Combining (1) with (2), so that we obtain a system of three equations in three unknowns, we can use standard algebraic techniques for solving systems of equations to delete z, and, therefore, we obtain

$$\frac{d^2y}{dx^2} - 2\frac{dy}{dx} + y = 0. \tag{3}$$

The differential equation (3) is homogeneous, and its general solution is

$$y = e^x(c_1 + c_2x). \tag{4}$$

Finally, the value of y that we found in (4) must be substituted into the first equation of the system (1) to obtain

$[e^x(c_1 + c_2x)]' - 3[e^x(c_1 + c_2x)] + z = 0 \Rightarrow z = e^x(2c_1 - c_2 + 2c_2x).$

Therefore, the solution to the system (1) is

$y = e^x(c_1 + c_2x)$ and $z = e^x(2c_1 - c_2 + 2c_2x)$.

Some Applications of Differential Equations in Mathematical Modeling

In this section, we shall consider a few examples of applications of differential equations in physics, biology, neuroscience, cognitive psychology, strategic studies, and economics.

Mechanics

As aforementioned, Newton's Second Law of Motion, which is the "backbone" of mechanics, states that, if an object of mass m is moving with acceleration a and being acted upon with force F, then
$F = ma.$
This is actually a differential equation, because

$$a = \frac{dv}{dt} = \frac{d^2s}{dt^2}$$

where $v = v(t)$ denotes the velocity of the object under consideration, and $s = s(t)$ denotes the position function of the given object, at any time t. Hence, Newton's Second Law of Motion can be written as a differential equation in terms of either the velocity, v, or the position, s, of the object under consideration as follows:

$$m\frac{dv}{dt} = F$$

and

$$m\frac{d^2s}{dt^2} = F$$

where F, the force acting on the particle, need not be constant, but it may vary with the position s or the velocity $\frac{ds}{dt}$ of the particle.

Electricity

Let us consider a simple series electric circuit, that is, an RLC circuit, which has the following components: a resistor R (implementing electrical resistance, thus reducing current flow, adjusting signal levels, dividing voltages, etc.), an inductor L (slowing down surges or spikes by temporarily storing energy in an electro-magnetic field and then releasing it back into the circuit), and a Capacitor C (storing and releasing electricity into a circuit by distributing charged particles on (generally two) plates to create a potential difference). Moreover, this circuit has a source of voltage (something like a battery) V, as shown, for instance, in Figure 10-2. Here, R, L, and C are constants (independent of time), and we are interested in the current $I(t)$ across the circuit, which is a function of time t. In the corresponding differential-equation model, time t will be the independent variable. We can also have $Q(t)$, which is the charge on the capacitor, and then $I = \frac{dQ}{dt}$. By Kirchhoff's Law, the total voltage around the circuit is equal to zero (since a circuit loop is a closed conducting path, and, therefore, no energy is lost), so that the voltage $V(t)$ from the battery is equal to the voltage V_R across the resistor plus the voltage V_L across the inductor plus the voltage V_C on the capacitor:

$$V(t) = V_R + V_L + V_C$$

where:

by Ohm's Law, the voltage V_R across the resistor is given by

$$V_R = R \cdot I(t)$$

(this is the relationship between voltage, current, and resistance); the voltage V_C on the capacitor is given by

$$V_C = \frac{1}{C} Q(t)$$

(where C is the capacitance of the capacitor, that is, the ability of the capacitor to store charge in it); and, by Faraday's Law, the voltage V_L across the inductor is given by

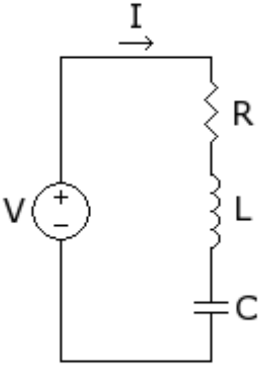
$$V_L = L \frac{dI}{dt}$$

(Faraday's Law says that a changing magnetic flux through a circuit will induce an electromagnetic flux in the circuit, and the induced electromagnetic flux can act like a battery and affect the flow of charge, that is, current, in the circuit). Furthermore, since $I = \frac{dQ}{dt}$, the aforementioned equations yield the following differential equation in terms of $Q = Q(t)$ (i.e., charge):

$$V(t) = L \frac{d^2Q}{dt^2} + R \frac{dQ}{dt} + \frac{1}{C} Q$$

(which is a non-homogeneous second-order linear differential equation with constant coefficients, which appears in electric circuits, $V(t) \neq 0$).

Figure 10-2: RLC series circuit (soure: Wikimedia Commons; Author: Omegatron; https://commons.wikimedia.org/wiki/File:RLC_series_circuit.png).



Demography

In the 1790s, the English economist Thomas Malthus assumed that the rate at which the population of a country grows at a certain time is proportional to the total population of the country at that time. Hence, on the basis of this assumption, we can develop the following model of population growth: If $p(t)$ denotes the total population at time t, then Malthus's assumption can be mathematically expressed in terms of the following differential equation:

$$\frac{dp(t)}{dt} = kp(t)$$

where k is the growth constant or the decay constant, as appropriate, and $p(t_0) = p_0$ is the initial condition (initial population). If $k > 0$, then the population grows, and, if $k < 0$, then the population will shrink. This differential equation is linear, and its solution is

$$p(t) = p_0 e^{kt}$$

where p_0 denotes the initial population. If we modify this model in order to allow the growth rate to vary linearly with time, then the model becomes

$$\frac{dp(t)}{dt} = k(t)p(t)$$

where $k(t) = at + b$ (for constants a and b), $p(t_0) = p_0$ is the initial condition (initial population), and this linear differential equation can be solved separation of variables.

Epidemiology

Let us consider the spread of a disease through a population. Suppose that we have a number of people, say N, who are infected with a disease. We want to know how N will change in time. Hence, N is a function of t, which denotes time. Each of the N people has a certain probability to spread the disease to other people during some period of time. Let us quantify infectiousness by using a constant k, so that the rate of change of the number of infected people with respect to time equals this constant k times the number of people who are already infected. In general, the rate of change of a function with respect to time is the derivative of that function with respect to time. Therefore, we obtain the following differential equation:

$$\frac{dN(t)}{dt} = k \cdot N(t) \Leftrightarrow \frac{dN(t)}{dt} - k \cdot N(t) = 0$$

which yields

$$N(t) = N_0 e^{kt}$$

where N_0 is the number of the infected people at the initial time ($t = 0$), and the probability of infecting someone appears in the exponent (kt). Thus, we understand why infectious diseases begin by speading exponentially (since the rate of growth of the infected population is proportional to the number of people who are already infected). When a disease begins to spread, the constant k in the aforementioned exponent is

$$k = \frac{R_0 - 1}{\tau}$$

where τ is the time an infected person remains infectious, and R_0 denotes the average number of people someone infects.

Interspecific Competition: The Lotka–Volterra Equations

The problem of the growth of two species competing for the same resources has signigant applications in biology, ecology, and economics. Consider two mixed populations of species that are mutually interdependent and compete for the same resources. Let N_1 and N_2 denote the number of individuals of species one and of species two, respectively. Both N_1 and N_2 are functions of time t. Then we obtain the following "logistic equations" for these two species:

$$Population\ growth\ of\ species\ 1\text{:}\ \frac{dN_1}{dt} = a_1 N_1 \left(1 - \frac{N_1}{M_1}\right)$$

$$Population\ growth\ of\ species\ 2\text{:}\ \frac{dN_2}{dt} = a_2 N_2 \left(1 - \frac{N_2}{M_2}\right)$$

which are uncoupled equations (i.e., we study the population growth of each species without accounting for the presence of another species), and $N_1 \to M_1$ and $N_2 \to M_2$, where the factor M denotes the corresponding "carrying capacity," or largest sustainable population (the value of $M$is determined by available resources and by each individual's resource demand, so that the logistic equation has *intra*-specific competition built into it, since there is also competition between the members of the same species).

However, we have to model the competition between these two populations (i.e., *inter*-specific competition). If N_1 is much smaller than M_1, and if N_2 is much smaller than M_2, then resources are plentiful, and these two populations, N_1 and N_2, grow exponentially with growth rates a_1 and a_2, respectively. If species one and species two compete, then the growth of species one reduces resources available to species two, and vice versa. Because we do not know the exact impact species one and species two have on each other, we introduce two additional parameters in order to model interspecific competition. In particular, let q_{12} and q_{21} be dimensionless parameters (constants) that respectively model the consumption of species one's resources by species two, and vice versa (for instance, if both species eat exactly the same food, but species two consumes twice as much as species one, then $q_{12} = 2$ and $q_{21} = 0.5$); that is, q_{12} represents the effect of species two on species one, and q_{21} represents the effect of species one on species two. Then we can modify and couple the two aforementioned logistic equations as follows:

$$Population\ growth\ of\ species\ 1\ in\ the\ presence\ of\ species\ 2\text{:}$$
$$\frac{dN_1}{dt} = a_1 N_1 \left(1 - \frac{N_1 + q_{12} N_2}{M_1}\right)$$

$$Population\ growth\ of\ species\ 2\ in\ the\ presence\ of\ species\ 1\text{:}$$
$$\frac{dN_2}{dt} = a_2 N_2 \left(1 - \frac{N_2 + q_{21} N_1}{M_2}\right)$$

(the outcome of competition, according to the Lotka–Volterra model, is ultimately determined by carrying capacity, that is, the M parameter, and by the competition coefficient, that is, the q parameter). As time increases, the solution to this model (system of differential equations), which starts at (N_1^*, N_2^*), approaches a point (N_1^T, N_2^T), so that one of the following cases holds: (i) the point (N_1^T, N_2^T) lies in the fully positive quadrant of the x, y-plane, so that both N_1^T and N_2^T are positive, which means that the species co-exist; (ii) the point $(N_1^T, N_2^T) = (0,0)$, which indicates extinction of both species; or (iii) one of N_1^T and N_2^T may be zero and the other positive, indicating a situation of competitive exclusion.

Differentiation and Growth of Cells

The minimal constituent matter elements of organic matter (such as DNA) are subject to differentiations, which underpin the actualization and the manifestation of the structural program of an organic being. In fact, due to their differentiation, the cells of an organic being underpin its organic constitution, which determines the corresponding organic being's unity and cohesion (namely, the attraction of molecules for other molecules of the same kind). Furthermore, it is important to mention that eukaryotes (that is, organisms whose cells have a nucleus enclosed within a nuclear envelope), such as the human being, have two types of DNA: the DNA of the cells (namely, the agent of the genetic information of the cells) and the mitochondrial DNA (namely, the DNA located in mitochondria, which are double membrane-bound organelles supplying cellular energy and controlling the cell cycle and the cell growth; mitochondrial proteins—that is, proteins transcribed from mitochondrial DNA—vary depending on the tissue and the species).

For a cell of mass m, its growth rate may be proportional to m, and then the model of the growth of a simple cell is given by the following differential equation:

$$\frac{dm}{dt} = km \Rightarrow m = m_0 e^{kt}$$

where m_0 denotes the initial condition (initial mass), and, usually, some restriction, like $m < m^*$, is assumed (that is, it is usually assumed that the cell undergoes division once mass m^* is reached rather than continuing to grow).

Moreover, we can assume that the growth rate of a cell is proportional to the rate at which it can absorb nutrient and, thus, proportional to its surface area, specifically, to the two-third power of its mass, thus obtaining the differential equation

$$\frac{dm}{dt} = km^{\frac{2}{3}}$$

where the $2/3$-scaling surface law was proposed in 1919 by the American biologists James Arthur Harris and Francis Gano Benedict, who conducted biometric studies of basal metabolism. According to the $2/3$-scaling surface law, the basal metabolism of animals differing in size is nearly proportional to their respective body surfaces, and, as organisms increase in size, their volume and, therefore, their mass increase at a much faster rate than their surface area. In particular, the $2/3$-scaling surface law is based on the assumption that metabolic rates scale to avoid heat exhaustion (bodies lose heat passively via their surface, but they produce heat metabolically throughout their mass). In the 1930s, the Swiss biologist Max Kleiber argued that a $3/4$-power scaling (instead of the Harris–Benedict surface law's $2/3$ -power scaling) describes more accurately the relationship between an animal's metabolic rate and its mass (symbolically, if B is an animal's metabolic rate, and if M is this animal's mass, then, according to Kleiber's law, $B \approx M^{3/4}$).

Neuroscience: The Standard Leaky Integrate-and-Fire (LIF) Model

The brain (the central nervous system) contains nerve cells that are highly specialized in transmitting messages. Each nerve cell, called a neuron, consists of the soma (i.e., the central body of the cell), the (neur)axon, and the dendrites. At the end of the neural tube, there is a special structure called a synapse, through which the neurons communicate with each other. When a message created in one neuron is about to be transmitted to the next, the first neuron releases specialized chemicals called neurotransmitters. The released neurotransmitters are taken up by specially shaped regions, called receptors, on the cell membrane of the next neuron involved in the particular synapse.

Neurons send signals along an axon to a dendrite through junctions called synapses. The standard Leaky Integrate-and-Fire (LIF) model is a point neuron model that helps us to represent and study the dynamics of the neuron, and it is given by the following differential equation:

$$V'(t) \equiv \frac{dV(t)}{dt} = \frac{1}{C}\left[I_e(t) - \frac{1}{R}(V(t) - E_L)\right]$$
$$\text{with } V(t) \leftarrow V_r, if\ V(t) > \Theta$$

where:

$V(t)$ denotes membrane potential (i.e., the difference in electric potential between the interior and the exterior of a biological cell; in other words, the difference in the energy required for electric charges to move from the internal to the exterior cellular environments and vice versa, so that, for instance, the resting membrane potential of a neuron is approximately $-70\ millivolts$, meaning that the inside of the neuron is approximately $70\ millivolts$ less than the outside);

C denotes membrane capacitance (parameter) and is proportional to the cell surface area;
R denotes membrane resistance (parameter) and is a function of open ion channels (the greater the number of open channels, the lower the membrane resistance);
E_L denotes resting (membrane) potential (parameter) and is the electric potential difference across the cell membrane when the cell is in a non-excited state;
I_e denotes trans-membrane current (an excitatory synaptic input initiates a current flow across the membrane and into the neuron, and this current consists of an ionic flow of positive ions (e.g., sodium ions Na^+) in addition to capacitive currents, and it is by convention a negative trans-membrane current; this changes the membrane potential at the location of the synaptic input, initiating axial currents, that is, currents inside the neuron);
V_r denotes reset membrane potential (transmission of a signal within a neuron (from dendrite to axon terminal) is carried by a brief reversal of the resting membrane potential called an "action potential," and, when neurotransmitter molecules bind to receptors located on a neuron's dendrites, ion channels open, so that, at excitatory synapses, this opening allows positive ions to enter the neuron and results in "depolarization" of the membrane, that is, a decrease in the difference in voltage between the inside and the outside of the neuron; a stimulus from a sensory cell or another neuron depolarizes the target neuron to its threshold potential (e.g., $-55mV$), and Na^+ channels in the axon hillock open, allowing positive ions to enter the cell, and, once depolarization is complete, the cell must now "reset" its membrane voltage back to the resting potential by closing the Na^+ channels);
Θ denotes firing threshold (i.e., the level that a depolarization must reach for an action potential to occur, and, in most neurons, the threshold is around $-55mV$ to $-65mV$; if the neuron does not reach this critical threshold level, then no action potential will fire);
t denotes time.
It is worth mentioning that the combination of differential equations with neural networks (computer systems modeled on the human brain and nervous system) gives rise to Neural Differential Equations (NDE), which empower Artificial Intelligence systems to synthesize time-evolving data in an effective way. By a "neural differential equation," we mean a differential equation with neural network vector field, and, thus, its generic form is the following:

$$\frac{dy(t)}{dt} = f_\theta\big(t, y(t)\big)$$
$$with\, y(0) = y_0$$

where the subscript θ represents some vector of learnt parameters, so that $f_\theta: \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ represents a standard neural network (for a systematic study of these issues, see: R. Rico-Martínez *et al.*, "Discrete-vs. Continuous-Time Nonlinear Signal Processing of Cu Electrodissolution Data," *Chemical Engineering Communications*, vol. 118, 1992, pp. 25–48).

A Mathematical Approach to Cognitive Psychology: The Weber–Fechner Law

In general, "mathematical psychology" is a branch of psychology that is based on mathematical modeling of perceptual, thought, cognitive, and motor processes, and it tries to formulate scientific laws that relate quantifiable strimulus characteristics with quantifiable behavior.

Psychological experiments conducted by the German physicist, philosopher, and experimental psychologist Gustav Theodor Fechner and the German physician Ernst Heinrich Weber (who is considered one of the founders of experimental psychology) suggest that the minimum change that we can detect in a stimulus' magnitude (the just perceptible difference) varies in such a way that the fractional change is a constant, and, in particular, the intensity of our sensation increases as the logarithm of an increase in the physical magnitude of the stimulus. Weber demonstrated that, if S is the physical magnitude of the stimulus, then we shall just perceive the change to $s + \Delta s$ where $\frac{\Delta s}{s} = k$, a constant.

According to Fechner, this constant k represents a standard increase in the psychological magnitude of the stimulus, I. Therefore,

$$\Delta I = \frac{\Delta S}{S}$$

or

$$\frac{\Delta S}{\Delta I} = S^*$$

(in a discrimination experiement, we are interested in measuring ΔI as a function of I, that is, we want to find the discrimination threshold ΔI such that a stimulus with intensity $I + \Delta I$ is just discriminable from a stimulus of intensity I). Treating ΔI and ΔS as infinitesimals, we realize that S is related to I through the differential equation

$$\frac{dS}{dI} = kS$$

where k is a constant of proportionality. The aforementioned differential equation (by separation of variables) yields

$$I = \frac{1}{k} lnS + c$$

where c is a constant (this result means that the subjective sensation is proportional to the logarithm of the intensity of the corresponding stimulus; for instance, perceived loudness or brightness is proportional to the logarithm of the actual intensity measured by means of an accurate technical instrument). In fact, this is the reason why the intensity of sounds (decibels), the brightness of stars (magnitudes), and many other similar quantities are measured on logarithmic scales.

Strategic Studies and Warfare Problems: The Lanchester–Deitchman Models

In the scholarly discipline of International Relations, the term “strategy” is used in order to relate military means and political ends, in both war and piece. When we study ancient history, “strategy” means a commander’s battle plan and, generally, the “art of war” (a term usually associated with the Chinese strategist and intellectual Sun Tzu); in the eighteenth and the nineteenth centuries, “strategy” evolved into a country’s whole disposition for war, both in peacetime and during periods of conflict; in the second half of the twentieth century, “strategy” and “foreign policy” were usually treated as two concepts and two practical activities inseparable from each other, if not synonymous (at least among the industrial nations); and, by the beginning of the twenty-first century, “strategy” explicitly included an international actor’s disposition for economic and technological war.

In warfare problems, the calculation of a force ratio may be achieved by simple rules or may include complex assumptions and subjective judgments. For the quantitative study of a force ratio, the following three variables are difficult to handle: (i) the disparity in number and lethality of weapons between similar organizations; (ii) the variations in concepts of combat support; and (iii) the concentration of forces.

In this section, we shall study the warfare modeling approach of the English polymath and engineer Frederick W. Lanchester and the guerilla and the mixed conventional-guerilla combat models developed by S. J. Deitchman, who followed the methodology of F. W. Lanchester.

Let $x(t)$ and $y(t)$ denote respective strengths of the forces at time t, where t is measured in days from the start of the combat. We shall identify the strengths with the numbers of combatants. We shall consider the ideal case where $x(t)$ and $y(t)$ are differentiable functions of time. Even though we may not have a specific formula for $x(t)$ as a function of time, we may have sufficient information about the operational loss rate (OLR) of the x-force (i.e., the loss rate due to inevitable diseases, desertions, and other non-combat mishaps), the combat loss rate (CLR), due to encounters with the y-force, and the reinforcement rate (RR). Hence,

$$\frac{dx(t)}{dt} = OLR + CLR + RR,$$

and a similar equation applies to the y-force.

Lanchester assumed that the loss rate of a force is directly proportional to the enemy force strength. The following three Lanchester-type models are of great significance; $x(t)$ and $y(t)$ denote the strengths of the opposing forces at time t, and t denotes time from the start of the combat (you may add reinforcement rates $P(t)$ and $Q(t)$ per day if relevant).

Model I: Conventional Combat (CONCOM; "aimed fire"):

$$\begin{cases} x' \equiv \dfrac{dx(t)}{dt} = -Ay(t), x(0) = x_0 \\ y' \equiv \dfrac{dy(t)}{dt} = -Bx(t), y(0) = y_0 \end{cases}$$

where the coefficients are non-negative loss rate constants: A denotes the fighting effectiveness of y, and B denotes the fighting effectiveness of x. In general, we assume that the "fighting effectiveness" is proportional to a power of

$$\frac{T - x(0)}{x(0)}$$

where T denotes the total number of troops at time $t = 0$, $x(0)$ denotes the total number of fighting troops at time $t = 0$, and, therefore, $T - x(0)$ denotes the total number of support troops at time $t = 0$ (a similar formula applies to the y-force). Solving the above system of differential equations gives

$$\frac{y'}{x'} = \frac{-Bx}{-Ay} = \frac{B}{A}\frac{x}{y}$$

which, by the chain rule, yields

$$\frac{y'}{x'} = \frac{\frac{dy}{dt}}{\frac{dx}{dt}} = \frac{dy}{dx} \Rightarrow \frac{dy}{dx} = \frac{B}{A}\frac{x}{y}$$

so that (given that we have a separable equation, meaning that we can separate variables and integrate each side) we obtain Lanchester's Square Law:

$$Aydy = Bxdx \Rightarrow A\int ydy = B\int xdx \Rightarrow A\frac{y^2}{2} + C_1 = B\frac{x^2}{2} + C_2$$
$$\Rightarrow Ay^2 - Bx^2 \equiv C$$

where C is a constant, Ay^2 is the fighting strength of y, and Bx^2 is the fighting strength of x. Then x wins if $C < 0$, y wins if $C > 0$, and a stalemate (equilibrium) occurs if $C = 0 \Leftrightarrow Ay^2 = Bx^2 \Leftrightarrow \left(\frac{y}{x}\right)^2 = \frac{B}{A}$ (notice that we can use definite integrals, too, in which case we integrate y over $[y_0, y]$, and we integrate x over $[x_0, x]$).

Model II: Guerilla Combat (GUERCOM):

$$\begin{cases} x' \equiv \dfrac{dx(t)}{dt} = -Ax(t)y(t) \\ y' \equiv \dfrac{dy(t)}{dt} = -Bx(t)y(t) \end{cases}$$

and we work in the same way as above to obtain the analogue of Lanchester's Square Law for a guerilla combat. Hence, we have:

$$\frac{y'}{x'} = \frac{-Bx(t)y(t)}{-Ax(t)y(t)} \Rightarrow \frac{y'}{x'} = \frac{B}{A}$$

so that

$$\frac{y'}{x'} = \frac{\frac{dy}{dt}}{\frac{dx}{dt}} = \frac{dy}{dx} \Rightarrow \frac{dy}{dx} = \frac{B}{A} \Rightarrow Ady = Bdx$$
$$\Rightarrow A\int dy$$
$$= B\int dx \Rightarrow Ay + C_1 = Bx + C_2 \Rightarrow Ay - Bx \equiv C$$

where C is a constant. Thus, a stalemate (equilibrium) occurs if $C = 0 \Leftrightarrow Ay = Bx \Leftrightarrow \frac{y}{x} = \frac{B}{A}$.

When each side is visible to the other, and every fighter on each side can fire on any opponent, the loss rate on one side is proportional to the number of opponents firing, that is, $x' = -Ay(t)$ and $y' = -Bx(t)$, and this leads to the quadratic "square law" for "equality of fighting strength" (i.e., the condition under which neither side wins), namely, $Ay^2 = Bx^2$. However, when each side is invisible to the other (since guerillas, or "insurgents," strike at a time and place of their own choosing and then disappear), and each fires into the area that is *believed* to be occupied by the other, the loss rate on one side is proportional to the number of fighters on the other *and* to the number of fighters occupying the area under fire,

so that $x' = -Ax(t)y(t)$ and $y' = -Bx(t)y(t)$, and this leads to the "linear law" for equilibrium (i.e., equality of fighting strength), namely, $Ay = Bx$.
Moreover, Lanchester's differential equations are the basis for the application of the slightly more complex Deitchman's law of mixed combat, which enables the simulation of the combat dynamics of qualtitatively different opponents x and y, such as the warfare of two adversaries in guerilla and conventional combat: this problem can be solved by a combination of quadratic and linear laws.

Model III: Mixed Guerilla-Conventional Combat (e.g., Vietnam War):

$$\begin{cases} x' \equiv \dfrac{dx(t)}{dt} = -Ax(t)y(t) \\ y' \equiv \dfrac{dy(t)}{dt} = -Bx(t) \end{cases}$$

where y is the conventional force (out in the open), and x is the guerilla force (hard to find); and we work in the same way as above to obtain the analogue of Lanchester's Square Law for a mixed guerilla-conventional combat. Hence, we have:

$$\frac{y'}{x'} = \frac{-Bx(t)}{-Ax(t)y(t)} \Rightarrow \frac{y'}{x'} = \frac{B}{Ay(t)}$$

so that

$$\frac{y'}{x'} = \frac{\frac{dy}{dt}}{\frac{dx}{dt}} = \frac{dy}{dx} \Rightarrow \frac{dy}{dx} = \frac{B}{Ay(t)} \Rightarrow A\int ydy = B\int dx \Rightarrow A\frac{y^2}{2} + C_1$$

$$= Bx + C_2 \Rightarrow Ay^2 - 2Bx = C$$

where C is a constant. Therefore, y wins if $C > 0$, x wins if $C < 0$, and a stalemate (equilibrium) occurs if $C = 0 \Leftrightarrow Ay^2 = 2Bx$. In the history of war, this type of combat is also known under the terms "counterinsurgency" and "counter-revolutionary" operations, and its history has been thoroughly studied by Joseph MacKay in his book *The Counterinsurgent Imagination* (Cambridge: Cambridge University Press, 2023), and, in this type of war, information asymmetry is of paramount importance in determining the outcome of war.

Arms Race Models: Richardson's Methodology

One manifestation of tension between nations is the existence of an arms race. In the context of an arms race, each nation responds, in some fashion, to any increase in the military capabilities of the opposing nation. The pioneer in mathematical theorizing about arms races was the English mathematician, physicist, and psychologist Lewis Fry Richardson (1881–1953). In developing his model of arms races, Richardson relied on three basic assumptions: (i) In an armaments race between two countries, each country attempts to increase its armaments proportionately to the size of the armaments of the other. (ii) The burden of armaments upon the economy of the country imposes a restraint upon further expenditures, and this restraint is proportional to the size of the existing force. (iii) A nation would procure arms, guided either by ambition, grievances, hostility, or the need to maintain internal order, even if no other nation posed a threat. The aforementioned three assumptions yield the following pair of linear differential equations, which summarize a two-nation arms race:

$$M_A' = kM_B - aM_A + g \tag{1}$$

$$M_B' = lM_A - bM_B + h \tag{2}$$

where: M_A' and M_B' denote the rates of change in the arms stocks (or military budgets) of nations A and B, represented by M_A and M_B, respectively; k and l are "response" coefficients; k (resp. l) indicates the influence of B's (resp. A's) total arms stock on the increase in A's (resp. B's) arms, and this influence is assumed to be positive, meaning that the higher the level of B's (resp. A's) weapons stocks the greater the increase in A's (resp. B's) weapons stocks will be; the coefficients a and b are "fatigue" factors indicating the damping effects on the arms race of the economic burden of maintaining the present level of armaments, and this effect is expressed as a proportion (a, b) of the present arms stocks (M_A, M_B); and, finally, the constants g and h denote "grievance" and "minimal security," summarizing the contribution to increased armaments of basic hostility between the opposing nations as well as the influence of the need to maintain internal order.

Given the above definitions, the differential equation (1) means that the change in A's level of armaments (i.e., M_A') is equal to a certain proportion (k) of B's arms stocks (i.e., M_B) minus a certain amount due to economic constraints (aM_A) plus an amount reflecting grievances and hostility; and a similar interpretation holds for the differential equation (2).

Richardson wanted to determine whether the arms race would reach an equilibrium, and, if so, if this equilibrium would be stable. In an arms race model, an equilibrium is a point at which neither side has any reason to further increase or decrease its stock of arms. In terms of the differential equations (1) and (2), an equilibrium occurs when both derivatives are equal to zero, that is, when

$$M_A' = M_B' = 0, \tag{3}$$

and the simultaneous equations (1) and (2) can then be solved to find the equilibrium point. An equilibrium point is said to be stable if and only if any movements away from the equilibrium (for whatever reason) are followed by movements back to the equilibrium. Therefore, an arms race reaches a stable equilibrium if neither side has an incentive to icrease its arms stock, and, should there be a movement away from the equilibrium, the equilibrium is soon recovered. Mathematically, in this arms race model, the equilibrium point is stable if and only if

$$kl < ab, \tag{4}$$

that is, if and only if the product of the response coefficients is less than the product of the fatigue coefficients (the political interpretation of condition (4) is that the nations' collective fatigue must overwhelm their joint defense competition).

Notice that, since the differential equations (1) and (2) are linear, they can be represented by a pair of straight lines whose intersection will be the equilibrium point, satisfying condition (3). The stability condition (4) will be satisfied whenever the slope of $M_B' = 0$ is less than the slope of $M_A' = 0$.

Elementary Ballistics

Ballistics (i.e., the field of mechanics concerned with the launching, flight behavior, and impact effects of projectiles) was put in a rigorous mathematical context by Isaac Newton, Johann Bernoulli, and Euler. The main problem of exterior ballistics is to determine the trajectory of a projectile launched from a cannon with a given angle and a given velocity. The differential equation of motion involves the gravity g, the velocity v of the projectile, the tangent inclination θ of the projectile, and the air resistance $F(v)$, which is an unknown function of v; namely:

$$gd(vcos\theta) = vF(v)d\theta$$

(if $F(v) = 0$, that is, if we ignore air resistance, then we obtain a parabolic trajectory; but the actual trajectory is calculated for a given non-zero $F(v)$).

If we ignore air resistance, then the distance travelled by a bullet is given by the formula

$$x = v_0\sqrt{\frac{2h}{g}}$$

where v_0 is the initial velocity of the bullet, h is the height from which the bullet is fired, and g is the acceleration due to gravity. If this formula incorporates drag (or resistance), then it becomes

$$x = v_0 t - \frac{C\rho A v^2 t^2}{2m}$$

where C is the drag coefficient of the bullet (a dimensionless quantity that quantifies the drag of the bullet in its environment, and, understandably, the shape of an object has a very significant effect on the amount of drag), ρ is the air density, A is the area of the bullet, t is the time of flight, and m is the mass of the bullet.

Remark: The "drag force" of an object as it moves through a fluid is given by

$$F_d = \frac{1}{2}\rho v^2 CA \Leftrightarrow C = \frac{2F_d}{\rho v^2 A}$$

where F_d is the drag force (measured in $newtons$), ρ denotes density (kg/m^3), v denotes velocity (m/sec), C denotes the drag coefficient, and A denotes the frontal cross-sectional area (m^2).

Business Cycles and Economic Growth

By the term "Gross Domestic Product," we mean the total monetary value of all final goods and services produced (and sold on the market) within a country during a period of time (typically one year). The formula for calculating Gross Domestic Product (GDP) is the following:

$$\begin{aligned} GDP = {} & private\ consumption + gross\ private\ investment \\ & + government\ investment \\ & + government\ spending + (exports - imports) \end{aligned}$$

and the term "gross" indicates that products are counted regardless of their subsequent use (a product can be used for consumption, for investment, or to replace an asset). Nominal GDP uses current prices in its measure. Real GDP is an inflation-adjusted measure of the total monetary value of all final goods and services produced (and sold on the market) within a country during a period of time (typically one year):

$$Real\ GDP = \frac{Nominal\ GDP}{GDP\ Deflator}$$

(for instance, if an economy's prices have increased by 1% since the base year that is used in order to calculate the Real GDP, then the GDP Deflator is equal to 1.01). If $Y(t)$ is the current state of GDP, then $\frac{dY(t)}{dt}$ is the rate of change of $Y(t)$ with respect to time t (i.e., the "growth rate").

The Harrod–Domar model was developed independently by the English economist Sir Roy Harrod and the Russian-American economist Evsey Domar in order to analyze business cycles, and it was used in order to explain an economy's growth rate through savings and capital productivity. Growth is measured in terms of GDP, and, according to the Harrod–Domar model, savings (denoted by S) lead to investment (denoted by I), so that $S = I$, investment leads to changes in capital stock (denoted by $\varDelta K$), so that $I = \varDelta K$, and the capital-output ratio is constant, so that the ratio $\frac{K}{Y} = c$ is constant, and then $\frac{dK}{dY} = c$ (meaning that the marginal product of capital is constant and equal to the average product of capital). Hence, this model postulates that the output growth rate is given by the differential equation

$$\frac{1}{Y}\frac{dY(t)}{dt} = sc - \delta$$

where s denotes the savings rate, δ denotes the rate of depreciation of capital stock, and c is the aforementioned constant (marginal product of capital). The solution

$$Y(t) = Y_0 e^{(sc-\delta)t}$$

demonstrates that increasing investment through savings and productivity boosts economic growth.

Bibliography

Altland, Alexander, and J. von Delft. *Mathematics for Physics: Introductory Concepts and Methods*. Cambridge: Cambridge University Press, 2019.

Anton, Howard. *Elementary Linear Algebra*, tenth edition. New Jersey: John Wiley & Sons, 2010.

Anton, Howard, I. C. Bivens, and S. Davis. *Calculus*, tenth edition. New Jersey: John Wiley & Sons, 2012.

Armstrong, M. A. *Basic Topology*. New York: Springer-Verlag, 1983.

Arrow, Kenneth. *Social Choice and Individual Values*, second edition. New York: John Wiley and Sons, 1963.

Athans, Michael, and P. L. Falb. *Optimal Control*. New York: Dover, 2007.

Ayres, Frank. *Differential Equations*, Schaum's Outline Series. New York: McGraw-Hill, 1967.

Bachelard, Gaston. *The New Scientific Spirit*, translated by Arthur Goldhammer. Boston: Beacon, 1986.

Bachelard, Gaston. *The Formation of the Scientific Mind*, translated by Mary McAllester Jones. Manchester: Clinamen, 2002.

Bachelard, Gaston. *The Poetics of Space*, translated by Maria Jolas. Boston: Beacon, 1994.

Balakrishnan, V. K. *Introductory Discrete Mathematics*. New York: Dover, 2010.

Balasko, Yves. *Foundations of the Theory of General Equilibrium*, second edition. Singapore: World Scientific, 2016.

Beer, Stafford. "Knowing Norbert," *Kybernetes*, vol. 23, 1994, pp. 17–20.

Bell, Eric Temple. *The Search for Truth*. New York: Reynal and Hitchcock, 1934.

Bellman, Richard. *Perturbation Techniques in Mathematics, Engineering & Physics*. New York: Dover, 2003.

Bertalanffy, Ludwig von. *General System Theory*, third printing. New York: George Braziller, 1972.

Blei, Ira, and G. Odian. *An Introduction to General Chemistry*, second edition. New York: W. H. Freeman, 2006.

Boyd, Richard, P. Gasper, and J. T. Trout, eds. *The Philosophy of Science*, seventh edition. Cambridge, Mass.: MIT Press, 1991.

Boyer, Carl B. *The History of Calculus and Its Conceptual Development*. New York: Dover, 1959.

Cambridge Companions to Philosophy. Cambridge: Cambridge University Press.

Carathéodory, Constantin. *Theory of Functions*, translated by F. Steinhardt, 2 vols. New York: Chelsea, 1958 and 1960.

Clagett, Marshall. *Greek Science in Antiquity*. New York: Abelard-Schuman, 1955 and Dover, 2002.

Derrick, William R., and S. I. Grossman. *Introduction to Differential Equations with Boundary Value Problems*, third edition. Minnesota: West Publishing Co., 1999.

Di Bernardo, Giuliano. *Introduzione alla logica dei sistemi normativi*. Bologna: Il Mulino, 1972.

Di Bernardo, Giuliano. *Le regole dell'azione sociale*. Milano: Il Saggiatore, 1983.

Di Bernardo, Giuliano, ed. *Normative Structures of the Social World*. Amsterdam: Rodopi, 1988.

Di Bernardo, Giuliano. *The Epistemological Foundation of Sociology*. Amazon, 2021.

Di Bernardo, Giuliano. *The Future of Homo Sapiens*. Amazon, 2021.

Di Bernardo, Giuliano. *Liberalismo contro Totalitarismo*. Amazon, 2023.

Dierker, E. *Topological Methods in Walrasian Economics*. New York: Springer-Verlag, 1974.

Dieudonné, J. *Foundations of Modern Analysis*. New York: Academic Press, 1960.

Dummett, Michael. *Truth and Other Enigmas*. Cambridge, Mass.: Harvard University Press, 1978.

Dunbar, Robin. *The Human Story*. London: Faber & Faber, 2004.

Dunbar, Robin. *How Religion Evolved and Why It Endures*. Oxford: Oxford University Press, 2022.

Edwards, Paul, ed. *The Encyclopedia of Philosophy*, 8 vols. New York: Macmillan, 1972.

Euclid's Elements, all thirteen books complete in one volume, edited by Dana Densmore. Santa Fe, New Mexico: Green Lion Press, 2007.

Euler, *Foundations of Differential Calculus*, tanslated by J. D. Blanton. New York: Springer, 2000.

Fechner, Gustav Theodor. *Elemente der Psychophysik*. Leipzig: Druck und Verlag von Breitkopf & Härtel, 1889.

Fitzpatrick, Patrick. *Advanced Calculus*, second edition. California: Thomson Brooks/Cole, American Mathematical Society, 2006.

Fraleigh, John B. *Calculus with Analytic Geometry*, third edition. Boston: Addison Wesley, 1990.

Galbraith, John Kenneth. *The New Industrial State*, with a new foreword by James K. Galbraith. Princeton, N.J.: Princeton University Press, 2007.

Gowers, Timothy, ed. *The Princeton Companion to Mathematics*. Princeton, N.J.: Princeton University Press, 2008.

Haaser, Norman B., and J. A. Sullivan. *Real Analysis*. New York: Dover, 1991.

Hadot, Pierre. *Philosophy As a Way of Life*, edited by Arnold I. Davidson, and translated by Michael Chase. Oxford: Blackwell, 1995.

Hale, J. K., and J. P. LaSalle. "Differential Equations: Linearity vs. Nonlinearity," *SIAM Review*, vol. 5, 1963, 249–272.

Halmos, Paul R. *Naive Set Theory*. New York: Springer, 1998.

Halperin, Steve. *Introduction to Proof in Analysis*, 2020 Edition; online: http://www2.math.umd.edu/~shalper/text.pdf

Hardy, G. H. *A Course of Pure Mathematics*, tenth edition. Cambridge: Cambridge University Press, 1993.

Heath, Sir Thomas L. *A Manual of Greek Mathematics*. New York: Dover, 2003.

Hirsch, Morris W., and S. Smale. *Differential Equations, Dynamical Systems, and Linear Algebra*. San Diego: Academic Press, 1974.

Hritonenko, Natali, and Y. Yatsenko. *Mathematical Modeling in Economics, Ecology and the Environment*. New York: Springer, 2016.

Kaplansky, Irving. *Set Theory and Metric Spaces*. New York: Chelsea, 1977.

Knopp, Konrad. *Theory and Applications of Infinite Series*, translated by R. C. H. Young. New York: Dover, 1990.

Landau, Edmund. *Elementary Number Theory*, translated by Paul T. Bateman. New York: Chelsea, 1958.

Lang, Serge. *Basic Mathematics*. New York: Springer, 1988.

Laos, Nicolas. *Topics in Mathematical Analysis and Differential Geometry*, Series in Pure Mathematics, vol. 24. Singapore: World Scientific, 1998.

Lavrentiev, M. M. *Some Improperly Posed Problems of Mathematical Physics*, translated by R. J. Sacker. New York: Springer, 1967.

Libet, Benjamin. *Mind Time: The Temporal Factor in Consciousness*. Cambridge, Mass.: Harvard University Press, 2004.

Lloyd, G. E. R. *Early Greek Science: Thales to Aristotle*. New York: W. W. Norton and Company, 1974.

Lomonosov, V. I. "Invariant Subspaces for the Family of Operators which Commute with a Completely Continuous Operator," *Functional Analysis and Its Applications*, vol. 7, 1973, 213–214.

Maurer, Stephen B., and A. Ralston. *Discrete Algorithmic Mathematics*, third edition. Florida: Taylor & Francis Group, 2005.

Mavroudeas, Stavros. “Periodising Capitalism: Problems and Method – The Case of the Regulation Approach,” *Research in Political Economy*, vol. 17, 1999, 27–61.

McConnell, Campbell R., S. Brue, and S. Flynn. *Economics*, 22nd edition. New York: McGraw-Hill, 2021.

McGraw-Hill Encyclopedia of Science and Technology, 20 vols., eleventh edition, 2012.

McLellan, David. *The Thought of Karl Marx*, second edition. London: Macmillan, 1981.

Milnor, John W. *Topology from the Differentiable Viewpoint*, revised edition. Princeton, N.J.: Princeton University Press, 1997.

Murray, J. D. *Mathematical Biology: An Introduction*, third edition. New York: Springer, 2002.

Newman, M. H. A. *Elements of the Topology of Plane Sets of Points*. New York: Dover, 1992.

Nicholson, Michael. *Formal Theories in International Relations*. Cambridge: Cambridge University Press, 1989.

Nicholson, Michael. *Rationality and the Analysis of International Conflict*. Cambridge: Cambridge University Press, 1992.

Nove, Alex, and D. M. Nuti, eds. *Socialist Economics: Selected Readings*. Baltimore: Penguin, 1972.

Parsons, Talcott. *The Structure of Social Action*. Glencoe, Ill.: The Free Press, 1949.

Pedoe, Dan. *Geometry: A Comprehensive Course*, new edition. New York: Dover, 1988.

Polya, G. *How to Solve It: A New Aspect of Mathematical Method*, second edition. Princeton, N.J.: Princeton University Press, 1988.

Pugh, Charles C. *Real Mathematical Analysis*, second edition. New York: Springer, 2016.

Rassias, G. M., and Th. M. Rassias, eds. *Selected Studies: Physics-Astrophysics, Mathematics, History of Science, A Volume Dedicated to the Memory of Albert Einstein* (with a Foreword by S. M. Ulam). Amsterdam: North-Holland Publ. Co., 1982.

Rassias, Themistocles M. *Foundations of Global Nonlinear Analysis*. Leipzig: Teubner-Texte zür Mathematik, Band 86, 1986.

Rassias, Themistocles M., ed. *Topics in Mathematical Analysis: A Volume Dedicated to the Memory of A.-L. Cauchy*. Singapore: World Scientific, 1989.

Rassias, Themistocles M., and P. M. Pardalos, eds. *Mathematical Analysis and Applications*. New York: Springer, 2019.

Richardson, Lewis F. *Arms and Insecurity: A Mathematical Study of the Causes and Origins of War*. Pittsburgh, Penn.: The Boxwood Press, 1960.

Rico-Martínez, R., *et al.* “Discrete-vs. Continuous-Time Nonlinear Signal Processing of Cu Electrodissolution Data,” *Chemical Engineering Communications*, vol. 118, 1992, 25–48.

Robbins, Lionel. A *History of Economic Thought: The LSE Lectures*, edited by S. G. Medema and W. J. Samuels. Princeton, N.J.: Princeton University Press, 1998.

Rosenlicht, Maxwell. *Introduction to Analysis*. New York: Dover, 1968.

Ryden, Barbara. *Introduction to Cosmology*, second edition. Cambridge: Cambridge University Press, 2016.

Samuelson, Paul A. *Foundations of Economic Analysis*, second edition. Cambridge, Mass.: Harvard University Press, 1983.

Sears, Francis W., M. W. Zemansky, and H. D. Young. *College Physics*, seventh edition. Reading, Mass.: Addison Wesley, 1991.

Shanker, S. G., ed. *Gödel’s Theorem in Focus*. London: Routledge, 1991.

Shapiro, Stewart, ed. *The Oxford Handbook of Philosophy of Mathematics and Logic*. Oxford: Oxford University Press, 2007.

Simmons, George F. *Introduction to Topology and Modern Analysis*. Tokyo: McGraw-Hill Kogakusha, 1963.

Smith, David E., ed. *A Source Book in Mathematics*. New York: Dover, 1959.

Spiegel, Murray R. *Vector Analysis and an Introduction to Tensor Analysis*, Schaum’s Outlines. New York: McGraw-Hill, 1959.

Spiegelhalter, David. *The Art of Statistics: How to Learn from Data*. New York: Basic Books, 2021.

Srivastava, S. P. *Discrete Mathematics*. New Delhi: Shree Publishers, 2014.

Stoker, J. J. *Nonlinear Vibrations in Mechanical and Electrical Systems*. New York: Interscience Publishers, 1950.

Struik, Dirk J. *A Concise History of Mathematics*, fourth revised edition. New York: Dover, 1987.

Swokowski, Earl W., and J. A. Cole. *Algebra and Trigonometry with Analytic Geometry*, thirteenth edition. Boston: Cengage Learning, 2011.

Theon of Smyrna. *Mathematics Useful for Understanding Plato*, translated by Robert and Deborah Lawlor. San Diego: Wizards Bookshelf, 1979.

Thom, René. *Mathématiques essentielles*. École d’automne de biologie théorique, Abbaye de Solignac, 24 septembre – 4 octobre 1981.

Thom, René. “Les intuitions topologiques primordiales de l’aristotélisme,” *Revue Thomiste*, tome 88, no. 3, 1988, 393–409.

Thom, René. *Structural Stability and Morphogenesis: An Outline of a General Theory of Models*, translated by D. H. Fowler. Reading, Mass.: W. A. Benjamin, 1975.

Toeplitz, Otto. *The Calculus: A Genetic Approach*, reprint edition. Chicago: University of Chicago Press, 2007.

Villani, Cédric. *Birth of a Theorem: A Mathematical Adventure*, translated by Malcolm DeBevoise. New York: Farrar, Straus and Giroux, 2016.

Warren, Bill. *Imperialism: Pioneer of Capitalism*. London: Verso and NLB, 1980.

Weiss, Neil A. *Elementary Statistics*, eighth edition. London: Pearson, 2011.

Wintner, Aurel. *The Analytical Foundations of Celestial Mechanics*, reprint edition. New York: Dover, 2014.

Wittgenstein, Ludwig. *Philosophical Investigations*, fourth edition, translated by G. E. M. Anscombe, P. M. S. Hacker, and J. Schulte. New Jersey: Wiley-Blackwell, 2009.

Zeeman, E. C. “Stability of Dynamical Systems,” *Nonlinearity*, vol. 1, 1988, 115–155.

Nicolas Laos's Photographs with Other Scientists and Philosophers

With my mathematics mentor Professor Themistocles M. Rassias (former Chairman of the Department of Mathematics at the University of La Verne's Athens Campus and Professor at the National Technical University of Athens). Professor Th. M. Rassias has received several awards, and he is an active member of an array of journals in mathematical analysis and optimization. His work extends over several fields of mathematical analysis, and he has published numerous research papers and research books on nonlinear functional analysis, functional equations, approximation theory, analysis on manifolds, calculus of variations, inequalities, and metric geometry. Professor Th. M. Rassias's research work is known in the field of mathematical analysis with the terms "Hyers–Ulam–Rassias stability (of functional equations)" and "Cauchy–Rassias stability," and in geometry with the term "Aleksandrov–Rassias problem (for isometric mappings)." Moreover, Professor Th. M. Rassias has conducted pioneering research in the Morse theory of critical points and in the study of Plateau's problem (i.e., the problem of determining the surfaces of minimum area spanned in a given curve or subject to other boundary conditions), modifying Marston Morse's critical point theory in order to solve Plateau's problem.

With Professor Svetoslav Jordanov Bilchev (former Chairman of the Department of Algebra and Geometry at the “Angel Kanchev” University of Ruse), at whose invitation, I addressed the Fifth International Conference on Differential Equations and Applications held in Ruse, Bulgaria, 24–29 August 1995. Professor S. J. Bilchev (1946–2010) received several awards, and his research interests included geometry/differential geometry, differential equations, inequalities, game theory, and mathematical models in economics. I cooperated with Professor S. J. Bilchev in the fields of differential equations and topology, and some results of our joint work have been published by the Union of Bulgarian Mathematicians and have been presented at mathematical conferences in Ruse and Kazanlak.

With Professor Stepan Tersian (faculty member of the Department of Mathematics at the “Angel Kanchev” University of Ruse and member of the Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences). Professor S. Tersian and Professor S. J. Bilchev were the editors of the *Proceedings of the Fifth International Conference on Differential Equations and Applications* held in Ruse, Bulgaria, 24–29 August 1995, published by the “Angel Kanchev” University of Ruse and the Union of Bulgarian Mathematicians. The aforementioned volume of proceedings includes my research paper “A Comparative Study of Linear and Nonlinear Differential Equations with Applications.”

With Professor Stojan Chernev (faculty member of the Center of Applied Mathematics and Informatics at the "Angel Kanchev" University of Ruse). At the University of Ruse, I had the opportunity to investigate various problems by means of dynamical systems, that is, expressions of the form

$$\frac{d}{dt}\vec{x} = \vec{f}\left(\vec{x}, t, \vec{u}, \vec{\beta}\right)$$

where: t denotes time; the vector $\vec{x}$ represents the state of the system, and it consists of the minimal set of variables $(x_1, x_2, x_3, ...)$ that are needed in order to describe the system under consideration; the vector field $\vec{f}$, called the "dynamics," is a set of functions $(f_1, f_2, f_3, ...)$ that describe the dynamics of the corresponding states of the system (so that the time derivative of x_1 will be given by the first-row equation f_1, the time derivative of x_2 will be given by the second-row equation f_2, etc.); the vector $\vec{u}$ denotes all the variables over which we have active control (i.e., variables that we can manipulate in order to change the behavior of the system); and the vector $\vec{\beta}$ denotes the system's parameters over which we do not explicitly have control, but which are important in order to understand the corresponding dynamical system (and big changes in the $\vec{\beta}$ parameters may give rise to big changes in the system's behavior, called "bifurcations," meaning that curves may branch, or "bifurcate," at a critical point of the corresponding function, so that two or more values of y may be possible for a single value of x).

From right to left, with: Professor Ruslan Mitkov (Research Professor at the Institute of Mathematics of the Bulgarian Academy of Sciences), Professor Myron Grammatikopoulos (Professor of Mathematics at the University of Ioannina, Greece, and Visiting Professor of Mathematics at the “Angel Kanchev” University of Ruse, Bulgaria), Professor Emiliya Velikova (Professor of Mathematics at the “Angel Kanchev” University of Ruse), and Ms R. Gatzova (Secretary, Center of Applied Mathematics and Informatics at the “Angel Kanchev” University of Ruse).

With the President of the University of La Verne, California, Dr. Stephen C. Morgan (first on the left) and other senior board members of the University of La Verne, 1996. During my studies at the University of La Verne, while majoring in Mathematics, I also conducted interdisciplinary studies, which underpinned my subsequent further studies and work in the fields of interdisciplinary mathematics and epistemology. During my studies at the University of La Verne, while majoring in Mathematics, I also conducted interdisciplinary studies, which underpinned my subsequent further studies and work in the fields of interdisciplinary mathematics and epistemology.

With my philosophy mentor Professor Giuliano Di Bernardo, who held the Chair of Philosophy of Science and Logic at the Faculty of Sociology of the Università degli Studi di Trento from 1979 until 2010, and he is a member of the Académie Internationale de Philosophie des Sciences. Apart from the field of academic philosophy, my cooperation with Professor Giuliano Di Bernardo includes work in the context of the Dignity Order, an international private exclusive membership association of which Professor Giuliano Di Bernardo is the Founder and Grand Master, with the goal of promoting the dignity of humanity. Prof. Di Bernardo personally inducted me into the Dignity Order and bestowed upon me the titles of a Knight and a Grand Prior of the Dignity Order.

Sardinia, Italy, 2023; a convocation of the Knights and the Dames of Dignity in Orosei: Based on, and in line with, the teachings and honors that I have received from Professor Giuliano Di Bernardo in the context of the Dignity Order (on the left), and using Freemasonry as an instrument and a symbolic technology for accomplishing my projects (totally detached from the profound degeneration that mainstream contemporary Freemasonries have suffered), I manage a unique and exclusive, autonomous Masonic association of literati (on the right) in order to operate as a guild of rigorously educated men and women who share and serve concrete epistemological, moral, aesthetic, and ideological values, principles, and visions, as well as in order to preserve and promote the concept and the value of the *Homo universalis* and to operate as a custodian of sophisticated and complex knowledge. My initiative to create a new, genuinely philosophically informed, intellectually significant, and historically relevant Freemasonry *from literati for literati* is based on my argument that contemplation must be rigorous and combined with action and on my attempt to articulate a creative synthesis between various aspects of Plato's political theory, modern philosophy, and cybernetics.

With Dr. Spyros Kiartzis, electrical engineer and business economist, Director of Alternative Energy Sources and New Technologies for the Hellenic Petroleum Group (HELLENiQ ENERGY Holdings S.A.), at an event that I organized in December 2023 in order to present some results of my scholarly endeavors.

Presenting some results of my research work regarding epistemology and mathematical modeling in the social sciences in the Ceremonial Hall of the Rectorate of the National and Kapodistrian University of Athens ("Ioannis Drakopoulos" amphitheater), in May 2022; with Dr. Stavros Mavroudeas (Ph.D./Birkbeck College, University of London), Professor of Political Economy at the Department of Social Policy of Panteion University in Athens, Greece (on my right). Professor Stavros Mavroudeas's areas of expertise include Marxist political economy, macroeconomics, and growth theory.

I present part of my research work and social action as well as my initiative to create a new Freemasonry that is scholarly rigorous, historically responsible, and politically aware and active, according to my training and work in the Dignity Order, at a press conference in Thessaloniki, Greece, 2023: https://www.thessnews.gr/thessaloniki/mathimata-tektonikis-filosofias-o-nikolaos-laos-ypografei-ena-endiaferon-vivlio/

Presenting my work in the scholarly discipline of epistemology based on Giuliano Di Bernardo’s thought and publications, in the Ceremonial Hall of the Greek Society of Writers, in Athens, in 2023; with Dr. Ioannis Katselidis, Lecturer of History of Economic Theory at the National and Kapodistrian University of Athens and at the Athens University of Economics and Business (on my right). The major focus of my philosophy is structuralism. Structuralism does not imply that a structural argument (e.g., a theorem) should dictate anything to reality, but it implies that, due to a valid structural argument, we have to expect that the empirical morphology will take a particular form, and, whenever reality does not comply with a structural argument, it simply makes the situation more thought-provoking and intellectually challenging.

In the Ceremonial Hall of the Greek Society of Writers, in Athens, in 2023, having on my left the lawyer and criminologist Mrs. Christina Ch. Florou (member of the Athens Bar Association), the biologist Dr. Vasileios Balis (Academic Director of the Aegean College's Thessaloniki Campus and associate of the Center for Regenerative Medicine of the Aristotle University of Thessaloniki), and the business consultant and biochemist Dr. Stamatis Tournis (Managing Partner and Principal Consultant of the Sigma Business Network, expert in the full spectrum of operations research and risk management). An epistemology roundtable focused on the Greek edition of Professor Giuliano Di Bernardo's book *The Epistemological Foundation of Sociology*.

Between the distinguished Greek political journalist and analyst Mr. Spyros Sourmelidis (on the left) and the Greek publisher and philologist Mr. Agisilaos Kalamaras (on the right), Athens 2023. In that event, I explained that my methodology and my mindset, in general, are inspired by cybernetics, which cultivates an inter-disciplinary approach to knowledge. As the renowned British polymath, cybernetics expert, management consultant, and university professor Stafford Beer (1926–2002) has aptly pointed out, all worthwhile thinking is underpinned by syntheses of different fields of study. Cybernetics is an integral part of my thought and of my arguments in favor of modernity. Whereas informatics is a branch of engineering and applied mathematics that deals with the study of computing and computational systems, cybernetics is a strongly interdisciplinary field that deals with the study of systems and control, communication, as well as information processing in living organisms and machines. Thus, cybernetics is focused on the application of principles from mathematics, engineering, biology, neuroscience, and social sciences in order to understand structures and explain the behavior of complex systems and in order to develop models for the control and regulation of the systems under study, whereas computer science is focused on the design, the development, and the use of software and hardware systems, including their underlying principles, technologies, and methodologies.