Undergraduate Lecture Notes in Physics

Lev Kantorovich

Mathematics for Natural Scientists

Fundamentals and Basics

Springer

Undergraduate Lecture Notes in Physics

More information about this series at http://www.springer.com/series/8917

Undergraduate Lecture Notes in Physics (ULNP) publishes authoritative texts covering topics throughout pure and applied physics. Each title in the series is suitable as a basis for undergraduate instruction, typically containing practice problems, worked examples, chapter summaries, and suggestions for further reading.

ULNP titles must provide at least one of the following:

- An exceptionally clear and concise treatment of a standard undergraduate subject.
- A solid undergraduate-level introduction to a graduate, advanced, or non-standard subject.
- A novel perspective or an unusual approach to teaching a subject.

ULNP especially encourages new, original, and idiosyncratic approaches to physics teaching at the undergraduate level.

The purpose of ULNP is to provide intriguing, absorbing books that will continue to be the reader's preferred reference throughout their academic career.

Lev Kantorovich

Mathematics for Natural Scientists

Fundamentals and Basics

Springer

Lev Kantorovich
Physics
King's College London
London, UK

ISSN 2192-4791 ISSN 2192-4805 (electronic)
Undergraduate Lecture Notes in Physics
ISBN 978-1-4939-2784-5 ISBN 978-1-4939-2785-2 (eBook)
DOI 10.1007/978-1-4939-2785-2

Library of Congress Control Number: 2015943266

Springer New York Heidelberg Dordrecht London

Printed on acid-free paper

Springer Science+Business Media LLC New York is part of Springer Science+Business Media (www.springer.com)

Preface

The idea to write a mathematics textbook for physics undergraduate students presented itself only after more than 10 years of teaching a mathematics course for the second year students at King's College London. I discovered that students starting their studies at university level are lacking elementary knowledge of the fundamentals of mathematics; yet in their first year, this problem is normally left untreated. Instead, the assumption is made that a baseline understanding of numbers, functions, numerical sequences, differentiation and integration was adequately covered in the school curriculum and thus they remain poorly represented at university. Consequently, the material presented to first years is almost exclusively new and leaves the students wondering where it had all come from and struggling to follow the logic behind it. Students, for instance, have no understanding of "a proof". They are simply not familiar with this concept as everything they have learned so far was presented as "given", which is a product of the spoon-feeding teaching methods implemented in so many schools. In fact, the whole approach to teaching mathematics (and physics) at school is largely based on memorising facts. How can one learn, appreciate and apply mathematics in unknown situations if the foundations of what we now call mathematics, formed from lines of thought by some of mankind's best minds over the centuries, remain completely untouched? How can we nurture independent thinkers, thoughtful engineers, physicists, bankers and so on if they can only apply what is known with little ability to create new things and develop new ideas? We are raising a generation of dependents. This can only be avoided if students are taught how to think, and part of this learning must come from understanding the foundations of the most beautiful of all sciences, mathematics, which is based on a magnificent edifice of logic and rigour.

One may hope, naturally, that students can always find a book and read. Unfortunately, very little exists in the offering today for physics students. Indeed, there are some good books currently being used as textbooks for physics and engineering students [1–5], which give a broad account of various mathematical facts and notions with a lot of examples (including those specifically from physics) and problems for students to practice on. However, these books in their presentation lack rigour; most of the important statements (or theorems) are not even mentioned while others are discussed rather briefly and far too "intuitively". Most worryingly, all these books assume previous knowledge by the student (e.g. of limits, derivatives and integrals) and start immediately from the advanced material leaving essentially an insurmountable vacuum in students' knowledge of fundamental principles.

I myself studied physics and mathematics between 1974 and 1979 at the Latvian University in Riga (the capital of Latvia, formerly USSR) where a very different approach to teaching students was adopted. We were taught by active research mathematicians (rather than physicists) and at a very different level to the one which is being widely used today at western universities. The whole first year of studies was almost completely devoted to mathematics and general physics; the latter was an introductory course with little mathematics, but covering most of physics from mechanics to optics. Concerning mathematics, we started from the very beginning: from logic, manifolds and numbers, then moving on to functions, limits, derivatives and so on, proving every essential concept as we went along. There were several excellent textbooks available to us at the time, e.g. [6, 7], which I believe are still among the best available today, especially for a student yearning for something thought provoking. Unfortunately, the majority of these books are inaccessible for most western students as they are only available in Russian, or are out of print. In this project, an attempt has been made to build mathematics from numbers up to functions of a single and then of many variables, including linear algebra and theory of functions of complex variables, step-by-step. The idea is that the order of introduction of material would go gradually from basic concepts to main theorems to examples and then problems. Practically everything is accompanied by proofs with sufficient rigour. More intuitive proofs or no proofs at all are given only in a small number of exceptional cases, which are felt to be less important for the general understanding of the concept. At the same time, the material must be written in such a way that it is not intimidating and is easy to read by an average physics student and thus a decision was made not to overload the text with notations mathematicians would find commonplace in their literature. For instance, symbols such as $\in$, $\ni$, $\subset$, $\supset$, $\cup$, etc. of mathematical logic are not used although they would have enabled in many places (e.g. in formulating theorems and during proofs) to shorten the wording considerably. It was felt that this would make the presentation very formal and would repel some physics students who search for something straightforward and palpable behind the formal notation. As a result, the material is presented by a language that is more accessible and hence should be easier to understand but without compromising the clarity and rigour of the subject.

The plan was also to add many examples and problems for each topic. Though adding examples is not that difficult, adding many problems turned out to be a formidable task requiring considerable time. In the end, a compromise was reached to have problems just sufficient for a student reading the book to understand the material and check him/herself. In addition, some of the problems were designed to illustrate the theoretical material, appearing at the corresponding locations throughout the text. This approach permitted having a concise text of reasonable size and a correspondingly sensible weight for the hardcopy book. Nevertheless, there are up to a hundred problems in each chapter, and almost all of them are accompanied by answers. I do not believe that not having hundreds of problems for each section (and therefore several hundreds in a chapter) is a significant deficiency of this book as it serves its own specific purpose of filling the particular gap in the whole concept of teaching mathematics today for physics students. There are many other books available now to students, which contain a large number of problems, e.g. [1–5], and hence students should be able to use them much more easily as an accompaniment to the explanations given here.

As work progressed on this project, it became clear that there was too much material to form a single book and so it was split into two. This one you are holding now is the first book. It presents the basics of mathematics, such as numbers, operations, algebra, some geometry (all in Chap. 1), and moves on to functions and limits (Chap. 2), differentiation (Chaps. 3 and 5), integration (Chaps. 4 and 6), numerical sequences (Chap. 7) and ordinary differential equations (Chap. 8). As we build up our mathematical tools, examples from physics start illustrating the maths being developed whenever possible. I really believe that this approach would enable students to appreciate mathematics more and more as the only language which physics speaks. This is a win-win situation as on the other hand, a more in-depth knowledge of physics may follow by looking more closely at the mathematics side of a known physics problem. Ultimately, this approach may help the students to better apply mathematics in their future research or work.

In the second book, more advanced material will be presented such as: linear algebra, Fourier series, integral transforms (Fourier and Laplace), functions of complex variable, special functions including general theory of orthogonal polynomials, general theory of curvilinear coordinates including their applications in differential calculus, partial differential equations of mathematical physics, and finally calculus of variation. Many more examples from physics will be given there as well. Both volumes taken together should comprise a comprehensive collection of mathematical wisdom hopefully presented in a clear, gradual and convincing manner, with real illustrations from physics. I sincerely hope that these volumes will be found sufficient to nurture future physicists, engineers and applied mathematicians, all of which are desperately lacking in our society, particularly those with a solid background in mathematics.

I am also convinced that the book should serve well as a reach reference material for lecturers. In spite of a relatively small number of pages in this volume, there really is a lot of in-depth material for the lecturers to choose from.

In preparing the book, I have consulted extensively a number of excellent existing books [6–9], most of them written originally in Russian. Some of these are the same books I used to learn from when I was a student myself, and this is probably the main reason for using them. To make the reading easier, I do not cite the specific source I used in the text, so I'd like to acknowledge the above-mentioned books in general here. A diligent student may want to get hold of those books for further reading to continue his/her education. In addition, I would also like to acknowledge the invaluable help I have been having along the way from Wikipedia, and therefore my gratitude goes to all those scholars across the globe who contributed to this fantastic online resource.

I would also like to thank my teachers and lecturers from the Latvian University who taught me mathematics there and whose respect, patience and love for this subject planted the same feeling in my heart. These are, first of all, Era Lepina who taught me the foundations of analysis during my first year at University (I still keep the notes from her excellent lectures and frequently consulted them during the work on this book!), Michael Belov and Teodors Cirulis (theory of functions of complex variables and some other advanced topics such as asymptotic series). I would also like to mention some of my physics lecturers such as Boris Zapol (quantum mechanics), Vladimir Kuzovkov (solid state physics) and Vladimir Ivin (group theory) who made me appreciate mathematics even more as the universal language of physics.

Finally, I would like to apologise to the reader for possible misprints and errors in the book which are inevitable for the text of this size in spite of all the efforts to avoid these. Please send your comments, corrections and any criticism related to this book either to the publisher or directly to myself (lev.kantorovitch@kcl.ac.uk). Happy reading!

London, UK Lev Kantorovich

References

1. McQuarrie D (2003) Mathematical methods for scientists and engineers. University Science Books, Sausalito. ISBN 1-891389-29-7
2. Riley, KF, Hobson, MP, Bence SJ (2006) Mathematical methods for physics and engineering. Cambridge University Press, Cambridge. ISBN 0521679710
3. Boas M (1983) Mathematical methods in the physical sciences, 2nd edn. Wiley, New York. ISBN 0-471-04409-1
4. Stroud K (2001) Engineering mathematics, 5th edn. Palgrave, New York. ISBN 0-333-919394
5. Arfken GB, Weber, HJ Harris, FE (2013) Mathematical methods for physicists. A comprehensive guide, 7th edn. Academic Press, New York. ISBN: 978-0-12-384654-9
6. Смирнов ВИ (1974) Курс высшей математики, т. 1–5, Москва, Наука (in Russian). Apparently, there is a rather old English translation: Smirnov VI A course of higher mathematics: Adiwes international series in mathematics, vol 1. ISBN 1483123944, 2. ISBN 1483120171, 3-Part-1. ISBN B0007ILX1K and 3-Part-2. ISBN B00GWQPPMO. Pergamon, New York 1964
7. Фихтенголц ГМ (2001) Курс дифференциального и интегрального исчисления, Москва, Физматлит (in Russian). [Fihtengolc GM (2001) A course of differentiation and integration, vol 1–3, Fizmatgiz,] I was not able to find an English translation of this book
8. Шипачев ВС (1998) Высшая математика, Москва, Высшая Школа (in Russian). [Shipachev VS (1998) Higher mathematics Vishaja Shkola] I was not able to find an English translation of this book
9. Widder DV (1989) Advanced calculus, 2nd edn. Dover, New York ISBN-13 978-0-486-66103-2

Famous Scientists Mentioned in the Book

Throughout the book various people, both mathematicians and physicists, who are remembered for their outstanding contribution in developing science, will be mentioned. For reader's convenience their names (together with some information borrowed from their Wikipedia pages) are listed here in the order they first appear in the text:

René Descartes (Latinized: **Renatus Cartesius**) (1596–1650) was a French philosopher, mathematician and writer.

Pythagoras of Samos (c. 570 BC – c. 495 BC) was an Ionian Greek philosopher, mathematician and founder of the religious movement called Pythagoreanism.

Tullio Levi-Civita (1873–1941) was an Italian mathematician.

Leopold Kronecker (1823–1891) was a German mathematician who worked in the areas of number theory and algebra.

Gerolamo (or **Girolamo**, or **Geronimo**) **Cardano** (1501–1576) was an Italian mathematician, physician, astrologer and gambler.

Rafael Bombelli (1526–1572) was an Italian mathematician.

William Rowan Hamilton (1805–1865) was an Irish mathematician, physicist and astronomer.

Blaise Pascal (1623–1662) who was a French mathematician, physicist, inventor, writer and Christian philosopher.

Abu Bakr ibn Muhammad ibn al Husayn al-Karaji (or **al-Karkhi**) was a Persian mathematician and engineer.

Baron **Augustin-Louis Cauchy** (1789–1857) was a famous French mathematician who contributed in many areas of mathematics.

Viktor Yakovych Bunyakovsky (1804–1889) was a Ukrainian mathematician.

Karl Hermann Amandus Schwarz (1843–1921) was a famous German mathematician.

Guido Grandi (1671–1742) was an Italian mathematician.

Oliver Heaviside (1850–1925) was a self-taught English electrical engineer, mathematician and physicist.

Lodovico Ferrari (1522–1565) was an Italian mathematician.

Niccolo Fontana Tartaglia (1499/1500–1557) was an Italian mathematician.

Gerolamo Cardano (1501–1576) was an Italian mathematician.

Paolo Ruffini (1765–1822) was an Italian mathematician.

Niels Henrik Abel (1802–1829) was a Norwegian mathematician. He is famously known for proving that the solution of the quintic algebraic equation cannot be represented by radicals.

Abraham de Moivre (1667–1754) was a French mathematician.

Due to Heinrich Eduard Heine (1821–1881) was a German mathematician.

Bernhard Placidus Johann Nepomuk Bolzano (1781–1848) was a Czech mathematician, philosopher and theologist.

Sir **Isaac Newton** PRS MP (1642–1726/7) was an outstanding English scientist who contributed in many areas of physics (especially, mechanics) and mathematics (calculus).

Sir **William Rowan Hamilton** (1805–1865) was an Irish physicist, astronomer and mathematician.

Gottfried Wilhelm von Leibniz (1646–1716) was a German mathematician and philosopher.

Brook Taylor (1685–1731) was an English mathematician who is best known for Taylor's theorem and Taylor's series.

Michel Rolle (1652–1719) was a French mathematician.

Colin Maclaurin (1698–1746) was a Scottish mathematician.

Joseph-Louis Lagrange (1736–1813) was an Italian Enlightenment Era mathematician and astronomer. He made significant contributions to the fields of analysis, number theory, and both classical and celestial mechanics.

Guillaume François Antoine, Marquis de l'Hôpital (1661–1704) was a French mathematician.

Johannes Diderik van der Waals (1837–1923) was a Dutch theoretical physicist and thermodynamicist.

Lev Davidovich Landau (1908–1968) was a prominent Soviet physicist who made fundamental contributions to many areas of theoretical physics, 1962 Nobel laureate. He is a co-author of a famous ten volume course of theoretical physics.

Vitaly Lazarevich Ginzburg (1916–2009) was a Soviet and Russian theoretical physicist, astrophysicist, 2003 Nobel laureate.

Georg Friedrich Bernhard Riemann (1826–1866) was a German mathematician.

Thomas Joannes Stieltjes (1856–1894) was a Dutch mathematician.

Henri Léon Lebesgue (1875–1941) was a French mathematician.

Jean-Gaston Darboux (1842–1917) was a French mathematician.

Jean Baptiste Joseph Fourier (1768–1830) was a French mathematician and physicist, best known for Fourier series and integral.

Leonhard Euler (1707–1783) was a famous Swiss mathematician and physicist. He made important discoveries in various fields of mathematics and introduced much of the modern mathematical terminology and notation. He is also renowned for his work in mechanics, fluid dynamics, optics, astronomy and music theory.

Paul Langevin (1872–1946) was a prominent French physicist.

Ludwig Eduard Boltzmann (1844–1906) was an Austrian physicist and philosopher famously known for his fundamental work in statistical mechanics and kinetics.

Lorenzo Romano Amedeo Carlo Avogadro di Quaregna e di Cerreto (1776–1856) was an Italian scientist.

Jean-Baptiste Biot (1774–1862) was a French physicist, astronomer and mathematician.

Félix Savart (1791–1841) was a French physicist.

Pierre-Simon, marquis de Laplace (1749–1827) who was a French mathematician and astronomer.

Hermann Ludwig Ferdinand von Helmholtz (1821–1894) was a German physician and physicist.

Josiah Willard Gibbs (1839–1903) was an American physicist, chemist and mathematician.

James Clerk Maxwell (1831–1879) was a Scottish mathematical physicist best known for his unification of electricity and magnetism into a single theory of electromagnetism.

Karl Theodor Wilhelm Weierstrass (1815–1897) was an outstanding German mathematician contributed immensely into modern analysis.

Carl Gustav Jacob Jacobi (1804–1851) who was a German mathematician.

Andrey (Andrei) Andreyevich Markov (1856–1922) was a Russian mathematician.

Albert Einstein (1879–1955) was a German-born theoretical physicist famously known for his relativity and gravity theories.

Marian Smoluchowski (1872–1917) was an Austro-Hungarian Empire scientist of a Polish origin.

Andrey Nikolaevich Kolmogorov (1903–1987) was a Soviet mathematician, one of the founders of modern probability theory.

Sydney Chapman (1888–1970) was a British mathematician and geophysicist.

George Green (1793–1841) was a British mathematical physicist.

August Ferdinand Möbius (1790–1868) was a German mathematician and theoretical astronomer.

Johann Benedict Listing (1808–1882) was a German mathematician.

George Gabriel Stokes (1819–1903) was an Irish and British mathematician, physicist, politician and theologian.

Mikhail Vasilyevich Ostrogradsky (1801–1862) was a Russian - Ukrainian mathematician, mechanician and physicist.

Johann Carl Friedrich Gauss (1777–1855) was a German mathematician and physicist.

Archimedes of Syracuse (c. 287 BC–c. 212 BC) was a Greek mathematician, physicist, engineer, inventor and astronomer.

Siméon Denis Poisson (1781–1840) was a French mathematician, geometer and physicist.

Hendrik Antoon Lorentz (1853–928) was a Dutch physicist.

André-Marie Ampère (1775–1836) was a French physicist and mathematician.

Michael Faraday (1791–1867) was a famous English physicist.

Adolf Eugen Fick (1829–1901) was a German-born physician and physiologist.

Jean-Baptiste le Rond d'Alembert (1717–1783) was a French mathematician, physicist, philosopher.

Lorenzo Mascheroni (1750–1800) was an Italian mathematician.

Douglas Rayner Hartree (1897–1958) was an English mathematician and physicist.

Max Karl Ernst Ludwig Planck (1858–1947) was a German theoretical physicist.

Satyendra Nath Bose (1894–1974) was an Indian Bengali physicist specialising in mathematical physics.

Adrien-Marie Legendre (1752–1833) was a French mathematician.

Charles Hermite (1822–1901) was a French mathematician.

Edmond Nicolas Laguerre (1834–1886) was a French mathematician.

Pafnuty Lvovich Chebyshev (1821–1894) was a Russian mathematician.

Augustin-Jean Fresnel (1788–1827) was a French engineer known for his significant contribution to the theory of wave optics.

Józef Maria Hoene-Wroński (1776–1853) was a Polish Messianist philosopher who worked in many fields of knowledge, including mathematics and physics.

Erwin Rudolf Josef Alexander Schrödinger (1887–1961) was an Austrian physicist, one of the founders of quantum theory, he is famously known for Schrödinger (wave) equation of quantum mechanics. He also contributed in other fields fo physics such as statistical mechanics and thermodynamics, physics of dielectrics, colour theory, electrodynamics, general relativity and cosmology.

Ferdinand Georg Frobenius (1849–1917) was a German mathematician.

Friedrich Wilhelm Bessel (1784–1846) was a German mathematician and astronomer.

Gustav Robert Kirchhoff (1824–1887) was a German physicist who contributed to the fundamental understanding of electrical circuits, spectroscopy and the emission of black-body radiation by heated objects.

Konstantin Eduardovich Tsiolkovsky (1857–1935) was a Russian and Soviet rocket scientist and pioneer of the astronautic theory.

Svante August Arrhenius (1859–1927) was a Swedish physicist, one of the founders of physical chemistry.

Contents

Part I
Fundamentals

Chapter 1
Basic Knowledge

This is an introductory chapter which sets up some basic definitions we shall need throughout the book. We shall also introduce some elementary concepts (some of which will be refined and generalised later on in forthcoming chapters) and derive simple formulae which the reader may well be familiar with; however, it is convenient to have everything we will need under "one roof". Hence the content of this chapter will enable the reader to work on the forthcoming chapters without the need of consulting other texts or web pages, and should make the book more or less self-contained.

1.1 Logic of Mathematics

In mathematics (and in other natural sciences, especially in physics) we would like to establish relations between various properties A, B, C, etc. In particular, we might like to know if property A is equivalent to B, or, less strictly, whether B follows from A (i.e. B is true if A is, i.e. A⇒B). This kind of argument is what we will be referring to as a "proof". It must be based on known facts. For instance, if we would like to prove that A⇒B, then we use known properties of A, and via a logical argument, the property B should follow.

What needs to be done in order to prove that some property A is *equivalent* to B? In this case it is not sufficient to show that B follows from A, i.e. A⇒B; it is also necessary to show that, conversely, A follows from B, i.e. B⇒A. If it was possible to provide a logical path in both directions, then one can say that B is true "if and only if" A is true. It is also said that there is *one-to-one correspondence* between A and B, or that for B to be true it is a *necessary and sufficient* for A to be true. For instance, A could be the statement that "a number consists of three digits", while B is that the

L. Kantorovich, *Mathematics for Natural Scientists*, Undergraduate Lecture Notes in Physics, DOI 10.1007/978-1-4939-2785-2_1

number lies inclusively between 100 and 999. Both these statements are equivalent. Indeed, from A (three digits) obviously B follows (between and including 100 and 999), and, conversely, from B immediately follows A.

If, however, it is only possible to provide a proof in one direction, e.g. A⇒B, but not in the other (e.g. by providing a contradicting example), then it is said that for B to happen it is *sufficient* that A is true, i.e. A implies B. If for B to be true A is necessarily required (but there are also some additional conditions required for B to be true as well), then it is said that A is a *necessary* condition for B. If A is necessarily required for B to be true, then B⇒A. Therefore, if both statements are proven, i.e. that A is both necessary and sufficient for B to be true, then obviously A and B are equivalent.

For instance, we know that any even number is divisible by 2; conversely, any number divisible by 2 can be written as $2n$ with n being an integer, and so it is even. Hence, these two conditions are indeed equivalent. However, not every even number is divisible by 4, i.e. it is a necessary condition for the number to be even in order to be divisible by 4, but not sufficient. Similarly, all numbers ending with 2 are divisible by 2, this is a sufficient condition; however, there are also other even numbers divisible by 2. Therefore, it is not possible to show that any number divisible by 2 ends with the digit 2, i.e. having the digit 2 at the end of the number is sufficient for the number to be even, but not necessary.

In some cases several conditions are equivalent. In such cases the logic of a proof goes in a cycle. If, e.g., we would like to prove the equivalence of three conditions A, B and C, then we prove that: A⇒B, then B⇒C and, finally, C⇒A.

How is the proof of say A⇒B is done in practice? Normally, one constructs, using algebra or geometry, or maybe some other argument, a logical path to B based on accepting A, and this process may contain several steps. If the reverse is to be proven, the process is repeated, but this time assuming that B is true, and, on the acceptance of that, a logical path is built towards A. Alternatively, an example may be proposed (just one such example would be enough) that corresponds to B, but contradicts A, in which case the reverse passage B⇒A is not possible, and hence only the sufficient condition is proven. Other methods of proof exist in mathematical logic as well but will not be used in this book.

If one makes a statement, it can be proven wrong by suggesting a single example which contradicts it.

In some cases a statement can be proven right by assuming an opposite and then proving that the assumption was wrong—this is called *proving by contradiction.*[1]

There is one rather famous logical method of a proof called *mathematical induction* (or simply *induction*), and since we are going to use it quite often, let us discuss it here. Suppose, there is a property which depends on the natural number $n = 1, 2, 3, \ldots$. We would like to establish a formula for that property which would be valid for any $n = 1, 2, 3, \ldots$ and up to infinity. We perform detailed calculations for some small values of n, say $n = 1, 2, 3, 4$, and see a pattern or a rule, and devise

[1] The corresponding Latin expression "reductio ad absurdum" is also frequently used.

such a formula for a general n. But would it be valid for any n? Obviously no one can repeat the calculation for all possible integer values of n as there are an infinite number of them. Nevertheless, there is a beautiful logical construction which allows one to prove (or disprove) our formula. The method consists of three steps and runs as follows:

- check that the formula is valid for the smallest values of n of interest, e.g. for $n = 1$;
- *assume* that the formula is valid for some integer value n;
- prove then that it is also valid for $n + 1$ by performing the necessary calculation assuming the formula for the previous integer value (i.e. for n); if the result for $n + 1$ looks exactly the same as for n but with n replaced by $n + 1$, the formula is correct; if this is not the case, the formula is wrong.

If the last step is successfully proven, then that means the formula is indeed valid for *any* value of n. Indeed, let us assume our formula is trivially verified for $n = 1$. Then, because of the last two steps of the induction, it must also be valid for the next value of $n = 2$. Once it is valid for $n = 2$, by virtue of the same argument, it must also be valid for the next value of $n = 3$, and so on till infinity.

As an example, consider a trivial case of even natural numbers. Let us prove that any even natural number can be written using the formula $N_k = 2k$, where k is an integer. Indeed, for $k = 1$ we get the number $N_1 = 2 \cdot 1 = 2$, which is obviously even. Now we *assume* that the number $N_k = 2k$ is even for any natural number k. Let us prove that the number associated with the next value of k is also even. We add 2 to the even number N_k to get the next even number $N_k + 2 = 2k + 2 = 2(k + 1)$ which is exactly of the required N_{k+1} form. **Q.E.D.**[2]

Problem 1.1. *Demonstrate that any odd number can be written as* $N_k = 2k - 1$*, where* $k = 1, 2, 3, \ldots$.

There will be more examples, less trivial than this one, later on in this and other chapters.

1.2 Real Numbers

In applications we often talk about various quantities having a certain value. For this we use *numbers*.

The simplest numbers arise from countable quantities which may only take one of the positive integer numbers $1, 2, 3, \ldots$ called natural numbers. These numbers

[2] Which means (from Latin "quod erat demonstrandum") "which was to be demonstrated".

form a sequence (or a set) containing an infinite number of numbers (members). By *infinite* here we mean that the set never ends: after each, however big, number there is always standing the next one bigger by one. We shall denote this set with the symbol $\mathbb{N}at$, it is formed only by positive integer numbers. We also define the simplest *operations* on the numbers such as addition, subtraction, multiplication and division. Addition and multiplication of any two numbers from $\mathbb{N}at$ is a number from the same set; subtraction and division of two numbers may be either the number of the same set or not.

Using the subtraction operation, negative integer numbers and the zero number are created forming a wider set $\mathbb{Z} = \{0, \pm 1, \pm 2, \pm 3, \ldots\}$ of positive and negative integers (including zero). Specifically, the number zero, 0, has a property that when added to (or subtracted from) any other number belonging to $\mathbb{Z}$, the number does not change.[3] The integer numbers have a property that summation, subtraction or multiplication of any two members of the set $\mathbb{Z}$ is a member of the same set, i.e. it is an integer number. An integer power operation n^k (where k is from $\mathbb{N}at$) is then defined for any $n \neq 0$ as a multiplication of n with itself k times: $n^1 = 1$, $n^2 = n \cdot n$, $n^3 = n \cdot n \cdot n$, and so on. We also define that zero power gives always one: $n^0 = 1$.

Any integer number can be represented as a sum containing the base ten:

$$\pm\, a_n a_{n-1} \cdots a_1 a_0 = \pm \left(a_n 10^n + a_{n-1} 10^{n-1} + \cdots + a_1 10^1 + a_0 \right), \tag{1.1}$$

where on the left we wrote an integer consisting of $n + 1$ digits (with the corresponding plus or minus sign), and on the right the decimal representation of the whole number. For instance, $375 = 300 + 70 + 5 = 3 \cdot 10^2 + 7 \cdot 10^1 + 5$ and $-375 = -\left(3 \cdot 10^2 + 7 \cdot 10^1 + 5\right)$. The representation above allows proving familiar divisibility rules. For instance, it follows from the above that any number is divisible by 2 or 5 if the last digit a_0 is divisible or is 0 since any base 10 is divisible by 2 or 5.

Problem 1.2. *Prove that a number is divisible by 3 or 9 if the sum of all its digits is divisible by 3 or 9, respectively.* [*Hint*: note that the number $10^k - 1$ is divisible by 3 or 9 for any integer k.]

Problem 1.3. *Prove the following divisibility rule for* 7*: multiply the first, the second, the third, etc. digits by* 1, 3, 2, 6, 4 *and* 5*, respectively (cut or continue the same sequence as necessary depending on the number of digits in the given number), then sum up all the numbers thus obtained. If the sum is divisible by* 7*, so is the whole number.*

[3] Often 0 is considered belonging to natural numbers $\mathbb{N}at$, but this is not so important for us here.

Rational numbers form a set $\mathbb{Q}$. These are formed by dividing all possible numbers m and n from $\mathbb{Z}$, i.e. via m/n with $n \neq 0$. Integer numbers are a particular case of the rational numbers formed by taking the denominator $n = 1$. Any sum, difference, product (including power) or division of two rational numbers is a rational number, i.e. it belongs to the same set over all four operations. For instance, in the case of the summation:

$$z_1 + z_2 = \frac{m_1}{n_1} + \frac{m_2}{n_2} = \frac{m_1 n_2 + m_2 n_1}{n_1 n_2} = \frac{m_3}{n_3} = z_3 ,$$

where the result, z_3, is obviously a rational number as well since $m_3 = m_1 n_2 + m_2 n_1$ and $n_3 = n_1 n_2$ are each integers.

Problem 1.4. *Prove that a product or division of two rational numbers is a rational number.*

Problem 1.5. *Prove that there exist an infinite number of rational numbers between any two rational numbers.* [*Hint*: construct, e.g., a set of n equidistant numbers lying along a linear interpolation between the two numbers and show that they are all rational.]

Any rational number $z = m/n$ can be represented as an integer i plus a rational number r which is between -1 and 1. For instance, consider $z > 0$ with $0 < m < n$. In this case $0 < m/n < 1$ and hence $i = 0$ and $r = z$. If $m > n$, then it can always be written as $m = i \cdot n + m_1$, where m_1 is the remainder between 0 and n excluding n (i.e. $0 \leq m_1 < n$). Then

$$z = \frac{i \cdot n + m_1}{n} = i + \frac{m_1}{n} = i + r .$$

Similarly for negative rational numbers. For instance, $13/6 = (2 \cdot 6 + 1)/6 = 2 + 1/6$; here, $i = 2$ and $r = 1/6$, while $-13/6 = -2 - 1/6$.

In the *decimal representation* positive rational numbers are represented as a number with the dot: the integer part is written on the left of the dot, while the rest of it (which is between 0 and 1) on the right. This is done in the following way: consider a number $m/n = i + m_1/n$ with $0 \leq m_1 < n$. Let us *assume* that there exists the smallest positive integer k such that $n \cdot k = 1\underbrace{0 \cdots 0}_{l} = 10^l$, i.e. it is represented as 1 following by l zeroes. For instance, if $n = 2$, then the smallest possible k is 5 as $2 \cdot 5 = 10 = 10^1$; if $n = 8$, then the smallest possible $k = 125$ yielding $8 \cdot 125 = 1000 = 10^3$. Then,

$$\frac{m_1}{n} = \frac{m_1 \cdot k}{n \cdot k} = \frac{m_1 \cdot k}{10^l} .$$

Since $m_1 < n$, then $m_1 \cdot k < 10^l$ and hence the number m_1/n is represented by *no more* than l digits. These are the digits which are written on the right of the dot in the decimal number. For instance, $3/8 = (3 \cdot 125) / (8 \cdot 125) = 375/1000 = 0.375$, $12/5 = 2 + 2/5 = 2 + 4/10 = 2.4$. For negative numbers the same procedure is applied first to its positive counterpart, and then the minus sign is attached: $-3/8 = -0.375$, $-12/5 = -2.4$.

It is easy to see that if a positive rational number has an integer part i and the rest $r = m_1/n$ consists of digits $d_1 d_2 \cdots d_l$, i.e. $r = 0.d_1 d_2 \cdots d_l$, then one can always write:

$$z = \frac{m}{n} = i + \frac{m_1}{n} = i + \frac{d_1}{10^1} + \frac{d_2}{10^2} + \frac{d_3}{10^3} + \cdots + \frac{d_l}{10^l} . \tag{1.2}$$

For instance,

$$\frac{3}{8} = 0.375 = 0 + \frac{3}{10^1} + \frac{7}{10^2} + \frac{5}{10^3} \quad \text{and} \quad \frac{12}{5} = 2.4 = 2 + \frac{4}{10^1} .$$

If a rational number has a finite number of digits after the dot (e.g. $0.415 = 415/1000$), then its representation (1.2) via a sum of inverse powers of 10 will contain a finite number of such terms. However, in some cases this is not the case: some rational numbers may have an *indefinitely repeating* sequence of digits (called *periods*), e.g. $1/3 = 0.33333\ldots = 0.(3)$ (which means it is of the period of 3), $1/6 = 0.16666\ldots = 0.1(6)$, and hence an infinite number of terms with inverse powers of 10 are needed:

$$\frac{1}{3} = 0.(3) = 0 + \frac{3}{10^1} + \frac{3}{10^2} + \frac{3}{10^3} + \frac{3}{10^4} + \cdots \quad \text{or} \quad \frac{1}{6} = 0.1(6) = 0 + \frac{1}{10^1} + \frac{6}{10^2} + \frac{6}{10^3} + \frac{6}{10^4} + \cdots .$$

These are examples of infinite numerical series which we shall be studying in other chapters of this course. The numbers m/n with periods in the decimal representation arise if it is impossible to find such integer k that $n \cdot k$ is represented as 10^l with some positive integer l.

One can also introduce numbers which cannot be represented as a ratio of two integers; in other words, there are also numbers r, for which it is impossible to find such m and n from $\mathbb{Z}$ such that $z = m/n$. These are called *irrational* numbers. For instance, the rational number $1/4$ can be also written as a square of $1/2$, i.e. $(1/2)^2 = 1/4$, or, inversely, $1/2 = \sqrt{1/4}$ is expressed via the square root operation $\sqrt{\ldots}$. However, not every square root of a rational number is also a rational number, e.g. $\sqrt{1/3}$ is not; this is because the integer number 3 cannot be represented as a square of another integer number. In the decimal representation irrational numbers are constructed from a sequence of an infinite number of digits which is never repeated; $\sqrt{2} = 1.414213562373095\ldots$ is an example of an irrational number.

All positive and negative rational and irrational numbers form a complete set of *real* numbers $\mathbb{R}$. Addition, subtraction, multiplication (and power) and division of any two real numbers (avoiding division by zero) always results in a real number. Operations of summation and multiplication are *commutative,* i.e. they do not depend on the order of terms ($x + y = y + x$ and $xy = yx$), and *associative* ($(x + y) + z = x + (y + z)$ and $(xy)z = x(yz)$); they are also *distributive*: $x(y + z) = xy + xz$. The subtraction operation $x - y$ is understood as summation of x and $-y$; division of x and y is understood as a multiplication of x and $1/y$.

Real numbers can be represented as points on the number axis as shown in Fig. 1.1. The distance between a point on the axis and zero represents an absolute value $|x|$ of the number x; its sign is plus if the point is on the right of 0, while the sign is minus if it is on the left. Between any two real numbers x_1 and x_2, no matter how close they are to each other, there is always an infinite number of real numbers x between them, $x_1 < x < x_2$. Numbers run in both directions from zero indefinitely, i.e. for any number $x > 0$ there is always another number $x' > x$ on the right of it; similarly, for any number $x < 0$ on the left of the zero there can always be found a number $x' < x$ on the left of it. It is convenient to introduce symbols on the right and the left edges of the set of real numbers, called plus and minus infinities, ∞ and $-\infty$, which designate the largest possible positive and negative numbers, respectively. These are not actual numbers (since for any real number there is always a real number either larger or smaller), but special symbols indicating unbounded limits of the sequence of real numbers at both ends. Although the symbols $\pm\infty$ are shown at

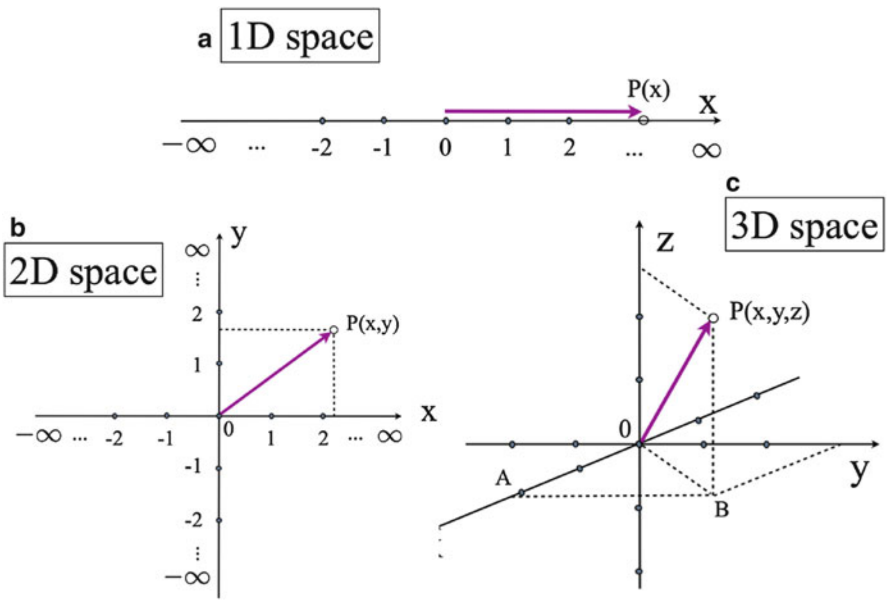


Fig. 1.1 Graphical representation of (**a**) one-, (**b**) two- and (**c**) three-dimensional spaces of real numbers

the right and left edges of the number axis, they are actually located indefinitely far away from zero, and any number x satisfies $-\infty < x < \infty$. Division of a positive real number by 0 is defined to be $+\infty$, while division of a negative number by 0 is defined to be $-\infty$.

We shall frequently be dealing with continuous subsets of numbers enclosed between two real numbers a and b. These subsets are called *intervals* and could be either: (i) unbounded, when $a < x < b$ (both ends are open); (ii) semi-bounded, if either $a \leq x < b$ or $a < x \leq b$ (one end is open and one closed); or (iii) bounded, if $a \leq x \leq b$ (both ends are closed). Intervals symmetric with respect to zero can be written using the absolute value symbol which is defined as follows:

$$|x| = \begin{cases} x\,, & \text{if } x \geq 0 \\ -x\,, & \text{if } x < 0 \end{cases} \tag{1.3}$$

Then, $|x| < a$ corresponds to the unbounded interval between $-a$ and a, and $|x| \leq a$ to the same interval with the addition of the upper and lower bounds $-a$ and a (assuming here that $a > 0$). A set of real numbers which are on the left of $-a$ or on the right of a can shortly be written as $|x| > a$. All real numbers can be represented as an interval between $\pm\infty$, i.e. $-\infty < x < \infty$. The interval $|x - x_0| < a$ corresponds to $-a + x_0 < x < a + x_0$, while $|x - x_0| > a$ to $x < -a + x_0$ or $x > a + x_0$.

Frequently other notations for the intervals are also used: $[a, b]$ is equivalent to $a \leq x \leq b$, $[a, b)$ is equivalent to $a \leq x < b$, $(a, b]$ is equivalent to $a < x \leq b$, and, finally, (a, b) is equivalent to $a < x < b$. In the mathematics literature the fact that an object "belongs" to a particular set is expressed with a special symbol $\in$ (the membership relation); e.g. the fact that x is somewhere between a and b, $a < x < b$, is expressed compactly as $x \in (a, b)$; however, we shall not use these notations here in the book.

1.3 Cartesian Coordinates in 2D and 3D Spaces

The number axis in Fig. 1.1(a) corresponds to a one-dimensional (often denoted as 1D) space $\mathbb{R}$. Any operation between two numbers from $\mathbb{R}$ such as summation, subtraction, multiplication (and, hence, power) and division results in another number from the same set $\mathbb{R}$.

By taking an ordered pair (x, y) of two real numbers x and y, a two-dimensional (or 2D) space is formed, see Fig. 1.1(b). In this case one draws two perpendicular number axes, one for x and another for y. A point on thus constructed plane, P, is said to have Cartesian[4] coordinates x and y, if two projections (perpendiculars)

[4]Due to René Descartes.

drawn from the point onto the two axes will cross them at the real numbers x and y, respectively. This is denoted $P(x, y)$. The 2D space formed by pairs of real numbers is denoted $\mathbb{R}^2$.

Similarly, a set (x, y, z) of three real numbers, x, y and z, corresponds to a point in a three-dimensional space $\mathbb{R}^3$ (or 3D). It is built by three mutually orthogonal number axes for the x, y and z, see Fig. 1.1(c), and a point P is said to have Cartesian coordinates x, y and z if the corresponding projections onto the three axes cross them at points x, y and z, denoted $P(x, y, z)$. Obviously, the 2D Cartesian system is a particular case of the 3D space corresponding to the $z = 0$ plane in it.

1.4 Elementary Geometry

Here we shall revisit the main ideas of the elementary geometry. We shall start from a circle on the $x - y$ plane. The simplest circle of radius R is the one centred at the origin. Every point (x, y) on the circle satisfies the equation $x^2 + y^2 = R^2$. Such a circle of unit radius is shown in Fig. 1.2(a).

Next, we shall consider angles. We have already mentioned above that two lines can be “perpendicular”, meaning they make 90 degrees (denoted 90°) with each other. This angle corresponds to the most symmetrical position of two lines crossing each other when they make four such angles with each other. Here we shall introduce general angles properly.

If we draw a circle of unit radius, see Fig. 1.2(a), then the positive angle $\theta = \angle AOB$ formed by two lines AO and BO crossing at point O is defined by the length L_{AB} of the arc AB relative to the length L_{circle} of the whole circle. Here the point A is obtained from the point B by moving around the circle in the anticlockwise direction. The whole circle corresponds to the angle of 360°, half of the circle to 180° and a quarter to 90°, as shown in Fig. 1.2(b). Alongside degrees, *radians* are also frequently used: 360° are equivalent to 2π radians, where $\pi = 3.1415\ldots$ is an irrational number. Therefore, generally the angle θ in Fig. 1.2(a) is given by the formula:

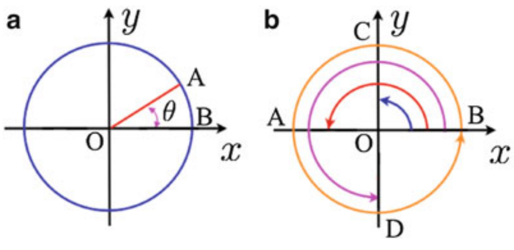


Fig. 1.2 The angle $\theta = \angle AOB$ in (**a**) is formed by lines AO and BO crossing at the point O in the centre of the unit radius circle. Specific angles of $90° = \pi/2$ (*blue*), $180° = \pi$ (*red*), $270° = 3\pi/2$ (*magenta*) and $360° = 2\pi$ (*orange*) are shown in (**b**)

$$\theta = \angle AOB = 360 \frac{L_{AB}}{L_{\text{circle}}} \text{ (degrees)} \quad \text{or} \quad 2\pi \frac{L_{AB}}{L_{\text{circle}}} \text{ (radians)}.$$

The angle of $180° = \pi$ corresponds to the half of the circle, while quarter of the circle gives $90° = \pi/2$. The latter angle is sometimes called the right angle. One also introduces *negative angles* which correspond to going around the unit circle in the clockwise direction starting from the positive part of the x axis. This is a fully equivalent representation. Then, for instance, $\angle BOD$ in Fig. 1.2(b) is $-90°$ (or $-\pi/2$). Hence the angles $-90°$ and $270°$ correspond to the same point D on the unit circle and hence are equivalent.

Consider now a line L_1 in Fig. 1.3(a). If we cross it with another line T at some angle, then at the crossing there will be four angles, but at most two will be different. This is because the opposite angles are always the same. Moreover, it can also be seen that the supplementary angles (defined as two angles on the same side of a straight line) on any side sum up to $180°$ (or π): $\alpha + \beta = \pi$. Exactly the same angles will be formed at the crossing with another line L_2 parallel to the first one: If we slide L_1 along T keeping it parallel all the way (i.e. keeping the same angles α and β), then we shall arrive at L_2 and the crossing will be perfectly identical; this follows from the fact that the lines are parallel: we call lines parallel if they never cross. This discussion gives us a useful relationship between various angles formed at the crossing of two parallel lines with a third (called transversal) line, as shown in Fig. 1.3(a).

We shall be using a lot of various 2D figures, and amongst them the *triangle* and *parallelogram* play an essential role. Consider first a general triangle $\triangle ABC$ shown in Fig. 1.3(b).

Problem 1.6. *Prove using Fig. 1.3(b) that the sum of three (internal) angles of a general triangle is* $180°$, *i.e.* $\alpha + \beta + \gamma = \pi$.

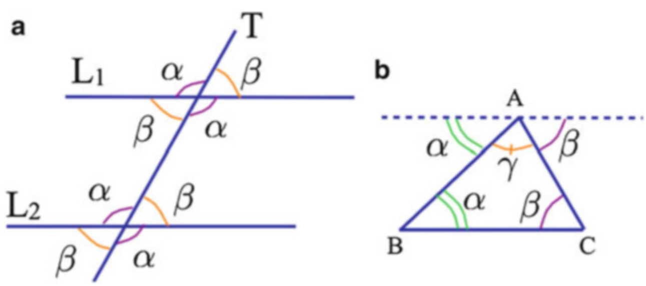


Fig. 1.3 (**a**) When a transversal line T crosses two parallel lines L_1 and L_2, eight angles are formed, four at each crossing. (**b**) The sum of the angles of any triangle is $180° = \pi$

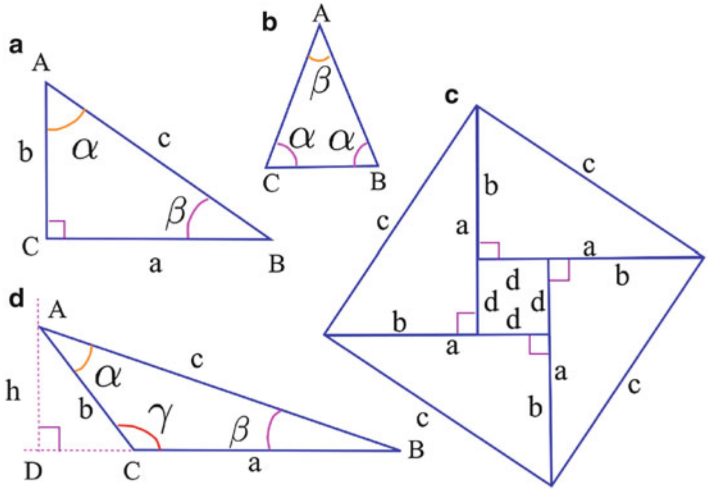


Fig. 1.4 (**a**) A right triangle ABC with sides a, b and c and two acute angles α and β. (**b**) An isosceles triangle with two equal sides $AC = AB$ and two equal angles $\angle ACB = \angle ABC = \alpha$. (**c**) To the proof of the Pythagoras theorem. (**d**) To the proof of the cosine theorem for a general triangle ABC

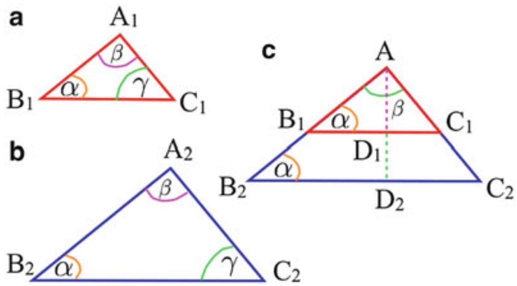


Fig. 1.5 Triangles in (**a**) and (**b**) are *similar*; the heights of the triangles in (**c**), AD_1 and AD_2, also differ by the same factor as the sides

The triangle in Fig. 1.4(a) is called a right triangle as one of its angles is 90° (or $\pi/2$). Then, the sum of its other two angles is $\pi/2$ as well. A general triangle would have its three angles α, β and γ all different, and its three sides AB, BC and AC different as well. If a triangle has all three sides equal to each other, $AB = BC = AC$, the three angles are equal as well: $\alpha = \gamma = \beta$. It is called equilateral. If only two sides are the same, e.g. $AB = AC$ as in Fig. 1.3(b), then the angles which are opposite to these sides are the same. This is an isosceles triangle. If one draws a perpendicular from its unique angle to the opposite side, it will divide the angle in equal halves and the opposite side into two equal lines.

In practice very often it is necessary to consider *similar triangles*, shown in Fig. 1.5(a) and (b); these are the ones which are scaled up or down with respect to each other; these have exactly the same angles, and their sides are scaled by the same factor: $A_1B_1/A_2B_2 = A_1C_1/A_2C_2 = B_1C_1/B_2C_2$.

Problem 1.7. *Show that the heights of the two triangles, AD_1 and AD_2, see Fig. 1.5(c), obtained by drawing perpendiculars from A to the opposite side, are scaled by exactly the same factor as their sides.*

1.5 Introduction to Elementary Functions and Trigonometry

A more detailed description of functions will be given in the next chapter. However, for the purposes of this chapter, some elementary knowledge is necessary. Although the reader should be familiar with what is to follow, we shall concisely summarise it here.

We shall start by reminding the reader the main ideas of *algebra*. In mathematics it is often useful (and convenient) to use letters instead of numbers as many operations on numbers such as multiplication, addition, division, etc. result in a single answer in a form of another number and no general rules are seen through this arithmetic calculation. Replacing numbers by letters allows one to perform calculations in a general form obtaining the answer (we call it *formula*) in the form of a combination of letters. Replacing the letters by numbers in the answer (i.e. in the resulting formula) gives the required numerical value. However, the convenience of this "algebraic way", as opposed to manipulating directly with numbers, is in the fact that the calculation from the original point (initial input) to the final result only needs to be done once and then the resulting expression (formula) will be valid for any numerical values of the input numbers, and hence can be used many times. Moreover, using algebraic manipulations it is often discovered that many terms may cancel out in the final result yielding a simple final expression. Therefore, using algebra the same final formula can be applied for different initial values and this may save time and effort.

However, we need to work out rules for algebraic manipulations. These follow from the rules which are valid for numbers. Since multiplication of two numbers does not depend on their order, e.g. $5 \cdot 3 = 3 \cdot 5$, we establish a general rule that multiplication of two letters, say a and b, is *commutative*: $a \cdot b = b \cdot a$. The dot here indicates the multiplication operation explicitly, it will be omitted from now on as only a single letter is used for a single number, so two letters one after another imply two numbers multiplied with each other. Similarly summation is also commutative, e.g. $3 + 5 = 5 + 3$, and so is summation of letters in algebra: $a + b = b + a$. A combination of addition and multiplication, e.g. $3 \cdot (5 + 6)$, may also be performed as $3 \cdot 5 + 3 \cdot 6$; we say that this binary operation is *distributive*. Therefore, the same can be then stated for letters: using distributivity, one can write $a(b + c)$ as $ab + ac$, i.e. "open the brackets". We can also use powers to indicate multiplication of a number with itself, e.g. $a^0 = 1, a^2 = aa$, etc. Both summation and multiplication are also *associative*, i.e. they do not depend on the order in which they are performed: $(a + b) + c = a + (b + c)$ and $(ab)\,c = a\,(bc)$.

As an example, let us consider an expression $(a+b)^2$. By definition of the power, this is $(a+b)(a+b)$. Then we use distributivity, treating one bracket as another number $c = a + b$:

$$(a+b)^2 = (a+b)\,c = ac + bc = a\,(a+b) + b\,(a+b) \ .$$

Next, we use the distributivity again:

$$(a+b)^2 = a\,(a+b) + b\,(a+b) = a^2 + ab + ba + b^2 \ .$$

Because a and b commute, $ab = ba$, we can sum up ab and ba together as $ab+ba = ab + ab = 2ab$, which finally allows us to arrive at the familiar result:

$$(a+b)^2 = a^2 + 2ab + b^2 \ .$$

Problem 1.8. *Prove the following algebraic identities:*

$$a^2 - b^2 = (a-b)\,(a+b) \ ; \quad (a-b)^2 = a^2 - 2ab + b^2 \ ;$$

$$(a \pm b)^3 = a^3 \pm 3a^2b + 3ab^2 \pm b^3 \ ; \quad a^3 \pm b^3 = (a \pm b)\left(a^2 \mp ab + b^2\right) \ .$$

In the last two cases either upper or lower signs should be consistently used when $\pm$ or $\mp$ signs are given.

We must also introduce *inequalities* as these are frequently used in mathematics and will be met very often in this book: these are relations between numbers (or corresponding letters) stating their relation to each other. If a number a is strictly *larger* than b, it is written as $a > b$; by writing $a \geq b$, we mean that a could also be equal to b, i.e. a is *not smaller* than b. Similarly, $a < b$ and $a \leq b$ state that either a is strictly *less* than b or it is *not larger* than it. Obviously, $a > b$ and $b < a$ are equivalent, as are $a \geq b$ and $b \leq a$.

One can also perform some algebraic manipulations on inequalities. Firstly, one may add to both sides of an inequality the same number, i.e. if $a < b$, then $a + c < b+c$. One can take a number from one side of the inequality to the other by changing the sign of the number, i.e. $a < b$ is equivalent to $a - b < 0$ (here b was taken from the right to the left side) or $0 < b - a$ (conversely, here a was taken from the left to the right side). The latter inequality can also be written as $b-a > 0$. For instance, let us prove that from $a < b$ follows $a - b < 0$: Adding to both sides of $a < b$ the number $-b$ yields $a - b < b - b$ or $a - b < 0$, as required.

It is possible to multiply both sides of the given inequality by some number $c \neq 0$: If $c > 0$, the sign of the inequality does not change; however, if $c < 0$, it changes to the opposite one. For instance, if $a < b$, then $5a < 5b$ as the number 5 is positive.

However, $-5a > -5b$. Indeed, the latter inequality can be easily manipulated into the first one:

$$-5a > -5b \quad \Longrightarrow \quad 0 > 5a - 5b \quad \Longrightarrow \quad 5b > 5a \ ,$$

which we know is correct.

Problem 1.9. *Prove, that if $a < b$ and $0 < c < d$ (i.e. both c and d are positive numbers), then $ac < bd$, i.e. one can "multiply both inequalities" side-by-side.*

Problem 1.10. *Show that if $0 < a < b$, then $\sqrt{a} < \sqrt{b}$ and $a^2 < b^2$.*

Two inequalities of the same type (i.e. both are either "less than" or "larger than" types) can be "summed up", i.e. from $a < b$ and $c < d$ follows $a+c < b+d$. Indeed, if we add to both sides of the inequality $a < b$ the *same* number c, the inequality will not change: $a + c < b + c$. However, since $c < d$, then we can similarly state that $b + c < b + d$, which proves the above made statement:

$$a + c < b + c \quad \text{and} \quad b + c < b + d \quad \Longrightarrow \quad a + c < b + d \ .$$

Similarly, from $a \geq b$ and $c \geq d$ follows $a + c \geq b + d$. Two inequalities of the opposite types can be subtracted: e.g., if $a < b$ and $c > d$, then $a - c < b - d$. Indeed, $d < c$, so that summing up this one with $a < b$, we get $a + d < c + b$ or $a - c < b - d$, as required.

Next we define a *function*. A function $y = f(x)$ establishes a correspondence between two real numbers x and y, both from $\mathbb{R}$. There must be a single value of y for each value of the x. The absolute value (or modulus) function introduced above by Eq. (1.3) is one example of a function: each value of the x is related to its absolute value $|x|$. The other simplest class of functions are *polynomials*:

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 = \sum_{i=0}^{n} a_i x^i. \tag{1.4}$$

This is a general way of writing a polynomial of degree n, where a_0, a_1, etc. are the $(n + 1)$ real coefficients; also, it is assumed that at least $a_n \neq 0$. The symbol for summation, $\sum$ (called sigma), we used above is very useful and we shall be employing it a lot. It is simply a short-cut notation for the sum written explicitly in Eq. (1.4); the index i is called a summation index and it runs here from 0 (shown underneath the sum symbol) to the value of n (shown on top of it) and it is a matter of simple exercise to check that the full expression just after the $f(x)$ is exactly reproduced by allowing i to run between its two limiting values (note that $x^0 = 1$ and $x^1 = x$).

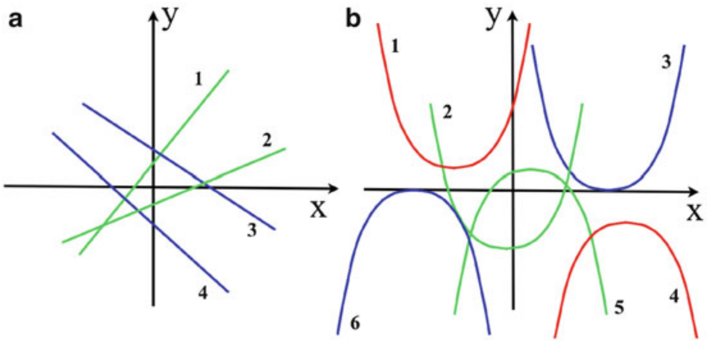


Fig. 1.6 Graphs of linear (**a**) and square (**b**) polynomials

If we divide one polynomial $P_n(x)$ of degree n on another one $Q_m(x)$ of degree m, we arrive at the class of *rational functions*.

The simplest polynomials are the first and the second order ones. Their graphs are shown in Fig. 1.6. When plotted, a linear polynomial $P_1(x) = a_0 + a_1 x$ appears as a straight line with $a_0 = P_1(0)$ being the point where its graph crosses the y axis, while a_1 shows its slope: if $a_1 > 0$, then the curve goes up from left to right (curves 1 and 2 in Fig. 1.6(a)), while when $a_1 < 0$ the line goes down (curves 3 and 4). The straight line crosses the x axis at the point $x = -a_0/a_1$. It happens always as long as $a_1 \neq 0$. If $a_1 = 0$, then the line is horizontal and hence never crosses the x axis.

Graphs of various second order polynomials (called *parabolas*), $P_2(x) = a_0 + a_1 x + a_2 x^2$, are symmetrical curves with either a minimum (if $a_2 > 0$) or maximum ($a_2 < 0$), see Fig. 1.6(b). Depending on the values of the constants a_0, a_1 and a_2 the parabola may either cross the x axis in 1 (curves 3 and 6) or 2 (curves 2 and 5) points or do not cross at all (curves 1 and 4). This property can be established by completing the square:

$$\begin{aligned} P_2(x) = a_2x^2 + a_1x + a_0 &= a_2\left[x^2 + 2\left(\frac{a_1}{2a_2}\right)x\right] + a_0 \\ &= a_2\left[\left(x^2 + 2\left(\frac{a_1}{2a_2}\right)x + \left(\frac{a_1}{2a_2}\right)^2\right) - \left(\frac{a_1}{2a_2}\right)^2\right] + a_0 \\ = a_2\left[\left(x + \frac{a_1}{2a_2}\right)^2 - \left(\frac{a_1}{2a_2}\right)^2\right] + a_0 &= a_2\left(x + \frac{a_1}{2a_2}\right)^2 + \left(-\frac{a_1^2}{4a_2} + a_0\right) . \end{aligned} \tag{1.5}$$

If the free term $a_0 - a_1^2/4a_2 = 0$, then the parabola just touches the x axis and hence has only a single *root*. If $a_0 - a_1^2/4a_2 > 0$ and $a_2 > 0$, then the entire graph of this function is above the x axis and there will be no roots (0 crossings of the x axis); if, however, $a_2 < 0$, then there will be 2 roots. Similarly, in the case of $a_0 - a_1^2/4a_2 < 0$

there will be 2 roots if $a_2 > 0$ and no roots if $a_2 < 0$. The roots can be established by setting to zero the complete square expression obtained above[5]:

$$a_2\left(x+\frac{a_1}{2a_2}\right)^2+\left(-\frac{a_1^2}{4a_2}+a_0\right)=0 \quad \Rightarrow \quad \left(x+\frac{a_1}{2a_2}\right)^2=\frac{a_1^2-4a_0a_2}{4a_2^2}$$

$$\Rightarrow x+\frac{a_1}{2a_2}=\pm\sqrt{\frac{a_1^2-4a_0a_2}{4a_2^2}} \quad \Rightarrow \quad x=-\frac{a_1}{2a_2}\pm\sqrt{\frac{a_1^2-4a_0a_2}{4a_2^2}} \quad \Rightarrow$$

$$x=\frac{-a_1\pm\sqrt{a_1^2-4a_0a_2}}{2a_2}\ . \tag{1.6}$$

No roots are obtained if the *determinant* $\mathcal{D}=a_1^2-4a_0a_2<0$, one root is obtained if $\mathcal{D}=0$, and, finally, two roots exist if $\mathcal{D}>0$. These are the same results as previously worked out by inspecting the graphs of the parabolas.

The other important class of functions are *trigonometric* ones. But first, we have to briefly return to the right triangles, i.e. the ones which have one right angle (so that the sum of the other two is exactly 90°), see Fig. 1.4(a). The sides a, b and c of the triangle $\triangle$ABC are related by the famous Pythagorean theorem stating that $a^2+b^2=c^2$. There are many proofs of this theorem, one of the simplest is based on the construction of four identical triangles depicted in Fig. 1.4(c). They form two squares: the outer of side c, and the inner of side $d=a-b$.

> **Problem 1.11.** *Prove the Pythagorean theorem using the construction in Fig. 1.4(c) made of four identical right triangles of Fig. 1.4(a).* [*Hint*: establish first, using properties of the triangles, that the outer and inner figures are exact squares.]

We are now in the position to define the necessary trigonometric functions. The *sine* function of angle α is defined from the right triangle of Fig. 1.4(a) as the ratio of the opposite side of the triangle a to its hypotenuse c:

$$\sin\alpha=\frac{a}{c}=\frac{a}{\sqrt{a^2+b^2}}\ . \tag{1.7}$$

Similarly, $\sin\beta$ for the other acute angle is defined as b/c, where b is the length of the side opposite to the angle β. The *cosine* function is defined as the ratio of the adjacent (with respect to the angle) side of the triangle to its hypotenuse:

[5] Here we wrote both roots at the same time using the combined symbol $\pm$. One has to understand this as two separate equations: one root is obtained with the upper sign (which is "+"), and another one with the lower sign "−".

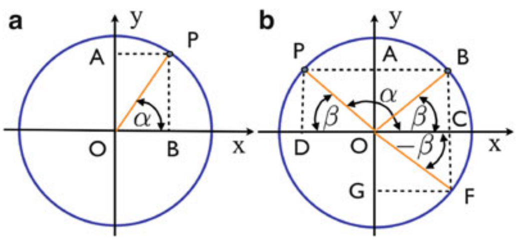


Fig. 1.7 To the definition of the trigonometric functions for arbitrary angles

$$\cos\alpha = \frac{b}{c} = \frac{b}{\sqrt{a^2+b^2}} \ . \tag{1.8}$$

From these two definitions we immediately have our first identity:

$$\sin^2\alpha + \cos^2\alpha = \frac{a^2}{a^2+b^2} + \frac{b^2}{a^2+b^2} = 1 \ , \tag{1.9}$$

which follows from the Pythagorean theorem. By looking at our triangle, we can also establish the following identities:

$$\cos\alpha = \frac{b}{c} = \sin\beta = \sin\left(\frac{\pi}{2}-\alpha\right) \quad \text{and} \quad \sin\alpha = \frac{a}{c} = \cos\beta = \cos\left(\frac{\pi}{2}-\alpha\right) \ .$$

Angles in a right triangle are limited to 90° (or $\pi/2$), and hence our definitions of the sine and cosine functions. However, these are generalised to any angle by means of the unit radius circle in the $x-y$ plane depicted in Fig. 1.7. If we consider a radius OP at an angle α with the x axis, see Fig. 1.7(a), then for angles $0 \le \alpha \le \pi/2$ the value of $\sin\alpha$ corresponds to the projection OA of the radius on the y axis; correspondingly, $\cos\alpha$ will be the projection on the x axis, OB. The same holds for any other angle. For instance, the radius OP in Fig. 1.7(b) corresponds to the angle $\pi/2 < \alpha < \pi$. In this case $\cos\alpha$ is defined as −OD and is negative (OD itself is a positive length) since the point D is on the negative part of the x axis; at the same time, $\sin\alpha$ =OA and is still positive. By rotating the radius around the circle the two projections change sign, and so do the two functions, sine and cosine. In fact, sine and cosine functions for arbitrary angle α can be defined as the projection on the y and x axes of the line drawn from the origin to the circle of unit radius and making the angle α with the x axis.

The sine and cosine functions of different angles are related to each other. Firstly, by rotating the radius by 2π we arrive at the same point on the circle, so that both functions remain the same for angles α and $\alpha + 2\pi$; we say that the functions are *periodic* with the period 2π:

$$\sin(\alpha+2\pi) = \sin\alpha \quad \text{and} \quad \cos(\alpha+2\pi) = \cos\alpha \ . \tag{1.10}$$

Secondly, consider points P and B in Fig. 1.7(b). We have:

$$\sin\alpha = \sin(\pi - \beta) = PD = BC = \sin\beta \ ,$$
$$\cos\alpha = \cos(\pi - \beta) = -OD = -OC = -\cos\beta \ .$$

If *negative angles* are used, the same definitions for the sine and cosine functions apply, see Fig. 1.7(b):

$$\sin(-\beta) = -OG = -OA = -\sin\beta \tag{1.11}$$

and

$$\cos(-\beta) = OC = \cos\beta \ . \tag{1.12}$$

A function is called *even* with respect to its argument if $f(-x) = f(x)$ for every value of the x, while it is called *odd* if $f(-x) = -f(x)$. It follows from these definitions that the sine is an odd function, while cosine is an even one.

Problem 1.12. *Now we shall consider a general triangle to derive the cosine theorem which generalises the Pythagoras theorem to arbitrary triangles. To this end, consider triangle $\triangle ABC$ in Fig. 1.4(d). Prove using the additional construction shown there the cosine theorem, stating that*

$$c^2 = a^2 + b^2 - 2ab\cos\gamma \ . \tag{1.13}$$

Problem 1.13. *Using the same triangle in Fig. 1.4(d), prove the* sine theorem *stating that the ratio of the sine of an angle in a triangle to the length of the opposite side is the same for all three angles (sides):*

$$\frac{\sin\beta}{b} = \frac{\sin\gamma}{c} = \frac{\sin\alpha}{a} \ . \tag{1.14}$$

[*Hint*: expressing the height h using either β or γ will immediately give the first part of the above identity; the other part is proven similarly by introducing the appropriate height h'.]

Two other frequently used functions are $\tan\alpha = \sin\alpha/\cos\alpha$ and $\cot\alpha = \cos\alpha/\sin\alpha = 1/\tan\alpha$. All their properties follow from their definition and the properties of the sine and cosine functions.

By looking at the right triangle in Fig. 1.4(a), we see that

$$\tan\alpha = \frac{\sin\alpha}{\cos\alpha} = \frac{a/c}{b/c} = \frac{a}{b} \ , \tag{1.15}$$

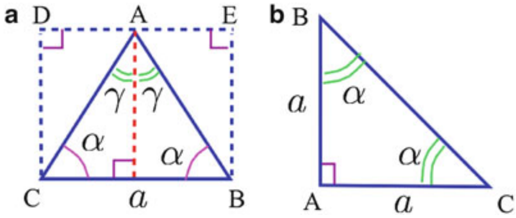


Fig. 1.8 To the proof of (**a**) $\sin(\pi/6)$ and (**b**) $\sin(\pi/4)$

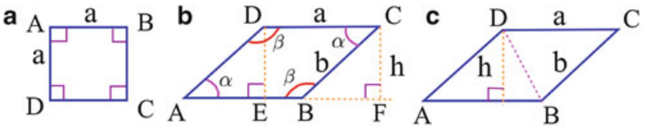


Fig. 1.9 (**a**) Square *ABCD*; (**b**) parallelogram *ABCD*; (**c**) to illustrate that the parallelogram *ABCD* consists of two identical triangles

i.e. it is equal to the ratio of the opposite and adjacent sides of the triangle. For $\cot\alpha$ this is the other way round.

Sine and cosine of some angles, such as $30° = \pi/6$, $60° = \pi/3$, $45° = \pi/4$ and some others can be calculated using special geometric considerations. Consider an equilateral $\triangle ABC$ shown in Fig. 1.8(a); it has all three angles equal: $\alpha = 60°$. Next, we draw a horizontal line DE parallel to the side CB of the triangle, and two perpendicular lines CD and BE. Since the lines DE and CB are parallel, the angles $\angle ADC$ and $\angle AEB$ are the right angles. Hence, $\angle DCA = \angle ABE = \gamma = 30°$. The two triangles $\triangle DAC$ and $\triangle AEB$ are equal.

Problem 1.14. *Show using Fig. 1.8(a), where an equilateral triangle $\triangle ABC$ is drawn, that* $\sin 30° = \cos 60° = 1/2$. *Correspondingly, show that* $\sin 60° = \cos 30° = \sqrt{3}/2$.

Problem 1.15. *Show using Fig. 1.8(b), where an isosceles right triangle $\triangle ABC$ is depicted, that* $\sin 45° = \cos 45° = 1/\sqrt{2}$.

In applications it is frequently needed to calculate areas of various objects, and we shall come across this type of problem very often in this course. Let us consider some elementary planar objects here. A square has four sides of the same length a and all four of its angles are the right angles, Fig. 1.9(a). Its area is $S_{\text{square}} = a^2$. The square is generalised by a rectangle in which case the two adjacent sides are not the same and equal to a and b, while the four angles are still of 90°; its area $S_{\text{rectang}} = ab$. Next, we bend the rectangle by pressing on one of its sides and obtain a parallelogram, Fig. 1.9(b). Its opposite sides are of the same length, but the four angles are in general different from 90°, although $\alpha + \beta = 180°$.

Problem 1.16. *Show using Fig. 1.9(b) that the area of the parallelogram* $S_{parallelogram} = ah$*, where* $h = b \sin\alpha$ *is the height shown in the figure.*

Problem 1.17. *Show using Fig. 1.9(c) that the area of a triangle* $\triangle ABD$ *is equal to* $S_{\triangle} = ah/2$*, where* h *is the height perpendicular to the side* a*.*

To finish this section, we shall consider equations for a general circle and of an ellipse. Coordinates (x, y) of points on a circle of radius R centred in the $x-y$ plane at the point (x_0, y_0) are given by

$$(x-x_0)^2 + (y-y_0)^2 = R^2 \ .$$

The simplest generalisation of the circle is an *ellipse*. It is obtained by stretching (squeezing) the circle along either x or y directions and the points lying on it satisfy the following equation:

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1 \ ,$$

where a and b are two positive constants and (x, y) are coordinates of an arbitrary points on the curve. This ellipse has the centre of the 2D coordinates system O as its symmetry point, see Fig. 1.10. There are two axes: the one drawn between points with coordinates $(-a, 0)$ and $(a, 0)$ along the x axis and the other between points $(0, -b)$ and $(0, b)$ along the y axis. The longer of the two axes is called the *major axis* of the ellipse, while the smaller one the *minor*. If $a = b$, the ellipse coincides with a circle of radius $R = a = b$. The *vertex* points $(\pm a, 0)$ and $(0, \pm b)$ correspond to the points on the curve of the ellipse which have, respectively, the largest and the smallest distance from the centre for the ellipse drawn in Fig. 1.10. The distance to the centre O from any other point on the curve of the ellipse lies between b and a values ($a > b$ for the ellipse in the figure). One also defines *eccentricity* of an ellipse,

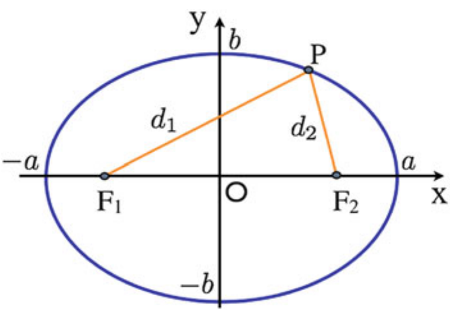


Fig. 1.10 An ellipse with the centre of symmetry at the centre of the coordinate system

$$e = \sqrt{1 - \left(\frac{b}{a}\right)^2} \ ,$$

where it is assumed that b and a are halfs of the minor and major axes, respectively ($a > b$).

An ellipse has two special points called *focal points* (or *foci*) shown in the figure as F_1 and F_2. A peculiar property of the ellipse is that the sum of distances $F_1P + F_2P = d_1 + d_2$ from the focal points to any point P on the ellipse is a constant quantity equal to $2a$ (obtained by taking the point P to be one of the vertices on the major axis).

Problem 1.18. *Let the foci have coordinates* $(\pm f, 0)$. *Using the equation of the ellipse, show that for any point* (x, y) *on the ellipse the distance* $D = d_1 + d_2 = 2a$ *is constant if and only if* $f^2 = a^2 - b^2$.

1.6 Simple Determinants

In many instances it appears to be extremely convenient to use special objects called *determinants*.

They are properly defined in linear algebra using matrices, however, it is quite useful to give a formal definition of the 2×2 and 3×3 determinants here; we shall do it quite formally without referring to matrices, so that determinants can be used in this and the following chapters as a convenient notation.

A 2×2 determinant is a scalar (a number) composed from four numbers a_{11}, a_{12}, a_{21} and a_{22} (which we distinguished by using unique double indices for convenience) written on a 2×2 grid between two vertical lines and its value is calculated as follows:

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21} \ . \tag{1.16}$$

A 3×3 determinant is composed of nine numbers arranged in three rows and three columns, and it is calculated as shown below:

$$\begin{aligned} \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} &= a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} \\ &= a_{11} \left(a_{22}a_{33} - a_{23}a_{32}\right) - a_{12} \left(a_{21}a_{33} - a_{23}a_{31}\right) \\ &\quad + a_{13} \left(a_{21}a_{32} - a_{22}a_{31}\right) \ . \end{aligned} \tag{1.17}$$

It is seen that the 3×3 determinant is expressed via three 2×2 determinants as shown. Note how the elements of the determinants are denoted: they have a double index, e.g. a_{13} has the left index 1 and the right index 3 in the 3×3 determinant. The left index indicates the row in which the given element is located, while the right index indicates the column. Indeed, the element a_{13} can be found at the end of the first row in the 3×3 determinant. We note that these notations come from matrices.

When defining the value of the determinant in (1.17), we "expanded" it along the first row as the elements of the first row (a_{11}, a_{12} and a_{13}) appear before the 2×2 determinants in the expansion. Notice that the 2×2 determinants can be obtained by removing all elements of the row and column which intersect at the particular element which appears as a pre-factor to them. Indeed, the determinant $\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix}$ is obtained by crossing out the first row and the first column in the original 3×3 determinant as these intersect exactly at the a_{11} element. The same is true for the other two as well as can easily be checked. It can be shown that the determinant can be expanded in a similar manner with respect to any row or any column although care is needed in choosing appropriate signs for each term.

An important property of the determinants that we shall need in the following is that a determinant is equal to zero if any one of its rows (or columns) is a linear combination of the other rows (columns). Let us consider in more detail what this means. Starting from the 2×2 case, if the second row is linearly dependent on the first, that means that the elements of the second row are related to the corresponding elements of the first row via the same scaling factor λ, i.e. $a_{21} = \lambda a_{11}$ and $a_{22} = \lambda a_{12}$. In this case, using the definition (1.16) of the determinant, we have:

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = \begin{vmatrix} a_{11} & a_{12} \\ \lambda a_{11} & \lambda a_{12} \end{vmatrix} = \lambda a_{11} a_{12} - \lambda a_{12} a_{11} = 0 \ ,$$

as required. In the 3×3 case the linear dependence is defined as follows: each element of, say, the first row is expressed via the corresponding elements of the other two rows using the same linear factors λ and μ:

$$a_{11} = \lambda a_{21} + \mu a_{31} \ , \quad a_{12} = \lambda a_{22} + \mu a_{32} \quad \text{and} \quad a_{13} = \lambda a_{23} + \mu a_{33} \ . \tag{1.18}$$

You can see that elements of the first row are indeed expressed as a linear combination of the elements from the other two rows taken from *the same column*. Jumping a little bit ahead, what we have just formulated can also be recast in a simple form if each row is associated with a vector; then, the linear combination expressed above would simply mean that the vector given by the first row is a linear combination of the vectors (with the coefficients λ and μ) corresponding to the second and the third rows. Now, let us see explicitly that if we have such a linear combination (1.18), then the determinant is equal to zero. This can be easily established from the definition of the determinant.

Problem 1.19. *Show using the definition of the* 3×3 *determinant that if the elements of the first row are related to those of the second and the third rows as given in Eq. (1.18), then the determinant is zero:*

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \begin{vmatrix} \lambda a_{21} + \mu a_{31} & \lambda a_{22} + \mu a_{32} & \lambda a_{23} + \mu a_{33} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = 0$$

for any values of λ *and* μ.

The other important property which we shall also need rather soon is that the determinant changes sign if two rows (or columns) are swapped around (permuted). Again, this is most easily checked by a direct calculation: use the definition (1.18) where say elements a_{1i} of the first row are replaced by the elements a_{2i} of the second, and vice versa:

$$\begin{vmatrix} a_{21} & a_{22} & a_{23} \\ a_{11} & a_{12} & a_{13} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{21}\left(a_{12}a_{33} - a_{13}a_{32}\right) - a_{22}\left(a_{11}a_{33} - a_{13}a_{31}\right) + a_{23}\left(a_{11}a_{32} - a_{12}a_{31}\right) .$$

Rearranging the terms in the right-hand side, we obtain:

$$\begin{vmatrix} a_{21} & a_{22} & a_{23} \\ a_{11} & a_{12} & a_{13} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = -a_{11}\left(a_{22}a_{33} - a_{23}a_{32}\right) + a_{12}\left(a_{21}a_{33} - a_{23}a_{31}\right)$$
$$- a_{13}\left(a_{21}a_{32} - a_{22}a_{31}\right) = - \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} ,$$

i.e. the minus of the original determinant. The same happens if any two rows or columns are interchanged. If two interchanges take place, the determinant does not change its sign.

Problem 1.20. *Show that*

$$|A| = \begin{vmatrix} 1 & 0 & 2 \\ 3 & -1 & 0 \\ 0 & 5 & 1 \end{vmatrix} = 29 , \quad |B| = \begin{vmatrix} 1 & 1 & 0 \\ 0 & 2 & 1 \\ 3 & -1 & 0 \end{vmatrix} = 4 .$$

1.7 Vectors

1.7.1 Three-Dimensional Space

As the 2D space is a particular case of the 3D space, it is sufficient to discuss the 3D space, and this is what we shall do here.

It is convenient to define objects, called *vectors*, belonging to the space denoted $\mathbb{R}^3$, such that they can undergo summation and subtraction operations resulting in the same type of objects (i.e. vectors as well) belonging to the same space $\mathbb{R}^3$. Vectors are defined by two points in space, A and B, and a direction between them, and are sometimes denoted by a directed arrow above the letters AB as $\overrightarrow{AB}$. This notation we shall be rarely using, however; instead, another notation will be frequently employed whereby a vector is shown by a bold letter, e.g. $\mathbf{a}$. If Cartesian coordinates of points A and B are (x_A, y_A, z_A) and (x_B, y_B, z_B), respectively, and the vector $\mathbf{a} = \overrightarrow{AB}$ is directed from A to B, then it is said that it has Cartesian coordinates $(x_B - x_A, y_B - y_A, z_B - z_A)$. It is seen that a vector is specified by three real numbers (its coordinates), hence the notation $\mathbb{R}^3$ for the space. The length of the vector (or its magnitude) is the distance between two points, A and B, and it is denoted $\left|\overrightarrow{AB}\right|$ or $|\mathbf{a}|$. Sometimes, the length of a vector is written using the same non-bold letter, e.g. $|\mathbf{a}| = a$. Two vectors $\mathbf{a}$ and $\mathbf{b}$ are said to be identical, $\mathbf{a} = \mathbf{b}$, if they have the same length and direction. This means that a vector can be translated in space to any position and remain the same. In particular, any vector can be translated to the position in which its starting point is at the centre of the coordinate system with zero coordinates as shown in Fig. 1.1(c). In this case the ending point of the vector P has coordinates $x = x_B - x_A$, $y = y_B - y_A$, $z = z_B - z_A$. It is easy to see from the same figure, that the length of the vector,

$$\left|\overrightarrow{OP}\right| = |\mathbf{a}| = \sqrt{x^2 + y^2 + z^2} = \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2 + (z_B - z_A)^2} \,, \tag{1.19}$$

corresponds to a diagonal of the cuboid with sides x, y and z. The formula above is obtained by using the Pythagoras theorem twice: first, we find the diagonal $OB = \sqrt{x^2 + y^2}$ of the right triangle $\triangle OAB$, then we obtain $OP = \sqrt{OB^2 + z^2} = \sqrt{x^2 + y^2 + z^2}$ as the diagonal of the other right triangle $\triangle OPB$, see Fig. 1.1(c). A vector of unit length is called *unit vector*. A null vector, $\mathbf{0}$, has all its coordinates equal to zero and the zero length. As was mentioned, formula (1.19) also naturally gives the distance OP between points O and P.

Two vectors are *collinear* if they have the same or opposite directions; in other words, if translated such that they begin at the centre of the coordinate system, they lie on the same line. Two vectors specified explicitly via their start and end points are *coplanar*, if there is a plane such that both vectors lie in that plane (no translation of the vectors is to be applied).

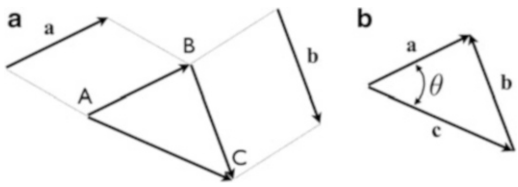


Fig. 1.11 Sum of two vectors **a** and **b**

Now we are in a position to define operations on vectors. A vector **a** if multiplied by a scalar (a real number) c results in another vector $\mathbf{b} = c\mathbf{a}$ which is collinear with **a** and has the length $|\mathbf{b}| = c\,|\mathbf{a}|$. Obviously, coordinates of **b** are obtained by multiplying the coordinates of **a** by the same factor c. The vector **b** has the same direction as **a** if $c > 0$, opposite to that of **a** if $c < 0$, and is the null vector if $c = 0$. Dividing a vector **a** by its length $|\mathbf{a}|$ results in a unit vector: $\mathbf{e} = \mathbf{a}/\,|\mathbf{a}|$ which has the same direction as **a**.

The sum of two vectors **a** and **b** is defined as follows: translate the vector **a** into the vector $\overrightarrow{AB}$, vector **b** into $\overrightarrow{BC}$, then the sum $\mathbf{c} = \mathbf{a} + \mathbf{b}$ is defined as the vector $\overrightarrow{AC}$, see Fig. 1.11(a). It is easy to see using the geometrical representation of the vectors in the 3D Cartesian system, that Cartesian coordinates of the vector **c** are obtained by summing up the corresponding coordinates of **a** and **b**. Similarly, one defines a difference $\mathbf{b} = \mathbf{c} - \mathbf{a}$ of two vectors **c** and **a** as a vector **b** for which $\mathbf{c} = \mathbf{a} + \mathbf{b}$, also see Fig. 1.11. Correspondingly, one can also define a general linear operation on two or any number n of vectors:

$$\mathbf{b} = c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + \cdots + c_n\mathbf{a}_n\;,$$

where c_1, c_2, etc. are real numbers. Note that any such *linear combination* of vectors results in a vector in the same space $\mathbb{R}^3$. The summation operation of vectors is both commutative and associative, similarly to real numbers; this is because by summing vectors we perform summation of their coordinates which are real numbers.

A *scalar* or *dot product* of two vectors $\mathbf{a} = (a_1, a_2, a_3)$ and $\mathbf{b} = (b_1, b_2, b_3)$ is defined as a scalar (a number)

$$\mathbf{a}\cdot\mathbf{b} = (\mathbf{a}, \mathbf{b}) = a_1b_1 + a_2b_2 + a_3b_3\;. \tag{1.20}$$

Both notations for the dot product (as $\mathbf{a}\cdot\mathbf{b}$ or $(\mathbf{a}, \mathbf{b})$) will be frequently used. The length of a vector **a** can be expressed via its dot product with itself: $|\mathbf{a}| = \sqrt{\mathbf{a}\cdot\mathbf{a}} = \sqrt{(\mathbf{a}, \mathbf{b})}$, which directly follows from the definitions of the two.

The dot product is *commutative*, $(\mathbf{a}, \mathbf{b}) = (\mathbf{b}, \mathbf{a})$, and *distributive*, $(\mathbf{c}, \mathbf{a} + \mathbf{b}) = (\mathbf{c}, \mathbf{a}) + (\mathbf{c}, \mathbf{b})$, which can easily be checked from their definitions; e.g. for the latter:

$$\begin{aligned}(\mathbf{c}, \mathbf{a} + \mathbf{b}) &= c_1(a_1 + b_1) + c_2(a_2 + b_2) + c_3(a_3 + b_3)\\ &= (c_1a_1 + c_2a_2 + c_3a_3) + (c_1b_1 + c_2b_2 + c_3b_3) = (\mathbf{c}, \mathbf{a}) + (\mathbf{c}, \mathbf{b})\;.\end{aligned}$$

Two vectors $\mathbf{a}$ and $\mathbf{b}$ are said to be *orthogonal*, if their dot product is zero. It is easy to see that this makes perfect sense since the orthogonal vectors make an angle of 90° with each other. Indeed, consider two vectors $\mathbf{a}$ and $\mathbf{c}$ with the angle θ between them, Fig. 1.11(b). Using distributivity and commutativity of the dot product, we can write:

$$(\mathbf{a}-\mathbf{c})^2 = (\mathbf{a}-\mathbf{c},\mathbf{a}-\mathbf{c}) = (\mathbf{a},\mathbf{a})-2(\mathbf{a},\mathbf{c})+(\mathbf{c},\mathbf{c}) = |\mathbf{a}|^2-2(\mathbf{a},\mathbf{c})+|\mathbf{c}|^2 = a^2+c^2-2(\mathbf{a},\mathbf{c}),$$

where $a = |\mathbf{a}|$ and $c = |\mathbf{c}|$. On the other hand, $\mathbf{a} - \mathbf{c} = \mathbf{b}$, and $|\mathbf{a} - \mathbf{c}| = |\mathbf{b}| = b$ is the length of the side of the triangle in Fig. 1.11(b) which is opposite the angle θ. Therefore, from the well-known cosine theorem of Eq. (1.13), $b^2 = |\mathbf{a} - \mathbf{c}|^2 = a^2 + c^2 - 2ac\cos\theta$. Comparing this with our previous result, we obtain another definition of the dot product in the 3D space as

$$(\mathbf{a},\mathbf{c}) = |\mathbf{a}|\,|\mathbf{c}|\cos\theta \; . \tag{1.21}$$

Therefore, if the dot product of two vectors of non-zero lengths is equal to zero, the two vectors make the right angle with each other, i.e. they are orthogonal.

It is convenient to introduce three special unit vectors, frequently called *unit base vectors* of the Cartesian system,

$$\mathbf{i} = (1,0,0)\;,\quad \mathbf{j} = (0,1,0)\;,\quad \mathbf{k} = (0,0,1)\;, \tag{1.22}$$

which run along the Cartesian x, y and z axes. These unit vectors are orthogonal, $(\mathbf{i},\mathbf{j}) = (\mathbf{i},\mathbf{k}) = (\mathbf{j},\mathbf{k}) = 0$, and all have the unit length. These are said to be *orthonormal*. Then, a vector $\mathbf{a} = (a_1,a_2,a_3)$ can be written as a linear combination of these unit base vectors:

$$\mathbf{a} = a_1(1,0,0) + a_2(0,1,0) + a_3(0,0,1) = a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k} \; . \tag{1.23}$$

There also exists another useful multiplication operation between two vectors which is called a *vector* (or *cross*) *product*. If $\mathbf{a} = (a_1,a_2,a_3)$ and $\mathbf{b} = (b_1,b_2,b_3)$, then one defines the third vector $\mathbf{c} = \mathbf{a}\times\mathbf{b}$ called their vector product as a vector of length

$$|\mathbf{c}| = |\mathbf{a}\times\mathbf{b}| = |\mathbf{a}|\,|\mathbf{b}|\sin\theta \tag{1.24}$$

and directed as shown in Fig. 1.12. Note that the vector $\mathbf{c}$ is perpendicular to both vectors $\mathbf{a}$ and $\mathbf{b}$.

In order to derive an explicit expression for the components of the vector product, $\mathbf{c} = \mathbf{a}\times\mathbf{b} = (c_1,c_2,c_3)$, let us first make use of the conditions that the vector $\mathbf{c}$ is perpendicular to both $\mathbf{a}$ and $\mathbf{b}$:

$$\mathbf{c}\perp\mathbf{a} \quad\Rightarrow\quad (\mathbf{c},\mathbf{a}) = 0 \quad\Rightarrow\quad a_1c_1 + a_2c_2 + a_3c_3 = 0\;, \tag{1.25}$$

$$\mathbf{c}\perp\mathbf{b} \quad\Rightarrow\quad (\mathbf{c},\mathbf{b}) = 0 \quad\Rightarrow\quad b_1c_1 + b_2c_2 + b_3c_3 = 0\;. \tag{1.26}$$

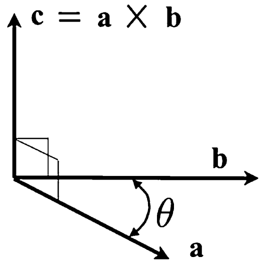


Fig. 1.12 Definition of the vector product of two vectors **a** and **b**

Multiplying the first equation by b_1 and the second by a_1 and subtracting from each other, one can relate c_2 to c_3:

$$c_2 = \frac{b_1 a_3 - b_3 a_1}{b_2 a_1 - b_1 a_2} c_3 \; . \tag{1.27}$$

Here we assumed that $b_2 a_1 - a_2 b_1 \neq 0$. Substituting the above expression back into any of the two equations (1.25) or (1.26), we also obtain a formula relating c_1 to c_3:

$$c_1 = \frac{b_3 a_2 - b_2 a_3}{b_2 a_1 - b_1 a_2} c_3 \; . \tag{1.28}$$

The obtained components of $\mathbf{c} = (c_1, c_2, c_3)$ make it perpendicular to both **a** and **b** for any value of c_3; choosing the latter fixes the length of the vector **c**. So, we must choose it in such a way as to comply with the definition of Eq. (1.24). Let us then consider the square of the length and make use of the fact that the dot product of **a** and **b** gives us access to the cosine of the angle θ between the two vectors, Eq. (1.21):

$$\begin{aligned} c^2 &= a^2 b^2 \sin^2\theta = a^2 b^2 \left(1 - \cos^2\theta\right) = a^2 b^2 \left[1 - \left(\frac{\mathbf{a}\cdot\mathbf{b}}{ab}\right)^2\right] = a^2 b^2 - (\mathbf{a}\cdot\mathbf{b})^2 \\ &= \left(a_1^2 + a_2^2 + a_3^2\right)\left(b_1^2 + b_2^2 + b_3^2\right) - (a_1 b_1 + a_2 b_2 + a_3 b_3)^2 \\ &= (a_2 b_1 - a_1 b_2)^2 + (a_3 b_1 - a_1 b_3)^2 + (a_2 b_3 - a_3 b_2)^2 \; . \end{aligned} \tag{1.29}$$

This should be equal to $c^2 = c_1^2 + c_2^2 + c_3^2$. Comparing (1.29) with Eqs. (1.27) and (1.28) for c_1 and c_2, we see that choosing $c_3 = \pm(b_2 a_1 - b_1 a_2)$ should give the desired expression for the c^2. To fix the sign here, consider a particular case when **a** is a unit vector directed along the x axis, $\mathbf{a} = \mathbf{i} = (1, 0, 0)$, and $\mathbf{b} = \mathbf{j} = (0, 1, 0)$ is a unit vector along the y axis. According to Fig. 1.12, the vector product of the two should be directed along the z axis, i.e. c_3 must be positive. Therefore, in $c_3 = \pm(1 \cdot 1 - 0 \cdot 0) = \pm 1$ the plus sign must be chosen. We finally have:

$$\mathbf{a} \times \mathbf{b} = (a_2 b_3 - a_3 b_2)\mathbf{i} + (a_3 b_1 - a_1 b_3)\mathbf{j} + (a_1 b_2 - a_2 b_1)\mathbf{k} \; . \tag{1.30}$$

The vector product of two vectors can most easily be remembered if written via a *determinant* which we introduced in Sect. 1.6; specifically, compare Eqs. (1.17) and (1.30). Hence, one can write:

$$\mathbf{a} \times \mathbf{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} . \tag{1.31}$$

Problem 1.21. *Prove that formula (1.30) has perfect sense also in the case of $b_2a_1 - a_2b_1 = 0$, i.e. show explicitly that in this case the vector $\mathbf{a} \times \mathbf{b}$ is perpendicular to both vectors $\mathbf{a}$ and $\mathbf{b}$, and that its length corresponds to the definition of Eq. (1.24).*

Problem 1.22. *If $\mathbf{a} = (1, 0, 1)$ and $\mathbf{b} = (-1, 1, 0)$, calculate: $(\mathbf{a}, \mathbf{b})$ and $\mathbf{a} \times \mathbf{b}$.*

Problem 1.23. *For $\mathbf{a} = (x, x^2, x^3)$ and $\mathbf{b} = (x^{-1}, x^{-2}, x^{-3})$, calculate: $x^2\mathbf{a}$, $(\mathbf{a}, \mathbf{b})$, $\mathbf{a} \times \mathbf{b}$ and $|\mathbf{a}|$ and $|\mathbf{b}|$.* [*Answers*: (x^3, x^4, x^5); 3; $(x^{-1} - x, -x^{-2} + x^2, x^{-1} - x)$; $x\sqrt{1 + x^2 + x^4}$ and $x^{-3}\sqrt{1 + x^2 + x^4}$.]

Another form of the vector product, very useful in practical analytical work, is based on the so-called Levi-Civita symbol ϵ_{ijk} defined as follows: if its indices i, j, k are all different and form a cyclic order (123, 231 or 312), then it is equal to $+1$, if the order is non-cyclic, then it is equal to -1; finally, it is equal to 0 if at least two indices are the same. With this definition, as can easily be checked, the component i of the vector product of $\mathbf{a}$ and $\mathbf{b}$ can be written as a double sum over all their components:

$$c_i = [\mathbf{a} \times \mathbf{b}]_i = \sum_{j=1}^{3}\sum_{k=1}^{3} \epsilon_{ijk} a_j b_k = \sum_{j,k=1}^{3} \epsilon_{ijk} a_j b_k \ . \tag{1.32}$$

In the last passage here a single sum symbol was used to simplify the notations; this is what we shall be frequently doing.

If two vectors are collinear (parallel), then the angle between them $\theta = 0$ and thus their vector product is zero. In particular, $\mathbf{a} \times \mathbf{a} = 0$.

It also follows from its definition that the vector product is *anti-commutative*,

$$\mathbf{a} \times \mathbf{b} = -\mathbf{a} \times \mathbf{b}$$

(the same length, but opposite direction), and non-associative,

$$(\mathbf{a} \times \mathbf{b}) \times \mathbf{c} \neq \mathbf{a} \times (\mathbf{b} \times \mathbf{c}) \ .$$

However, the vector product is *distributive*:

$$\mathbf{a} \times (\mathbf{b} + \mathbf{c}) = \mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c} \, .$$

Considering all vector products between the three unit base vectors (either directly using the definition based on Fig. 1.12 or by means of the coordinate representation (1.30) or (1.31)), we can establish the following simple identities: $\mathbf{i} \times \mathbf{j} = \mathbf{k}$, $\mathbf{i} \times \mathbf{k} = -\mathbf{j}$ and $\mathbf{j} \times \mathbf{k} = \mathbf{i}$. Of course, vector multiplying either of the three vectors with themselves gives zero.

Problem 1.24. *Using Eq. (1.30), prove that the vector product is anti-commutative.*

Problem 1.25. *Using Eq. (1.30), prove that the vector product is distributive.* [*Hint*: the algebra could be somewhat simplified if vector $\mathbf{a}$ is chosen along the x axis and vector $\mathbf{b}$ in the (x, y) plane; of course, the result should not depend on the orientation of the Cartesian axes!]

Problem 1.26. *Using the distributivity of the vector product, derive formula (1.30) by multiplying vectors* $\mathbf{a} = a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}$ *and* $\mathbf{b} = b_1\mathbf{i} + b_2\mathbf{j} + b_3\mathbf{k}$ *directly.*

Problem 1.27. *Using the explicit expression for the vector product given above, prove the following identities:*

$$[\mathbf{a} \times \mathbf{b}] \cdot [\mathbf{c} \times \mathbf{d}] = (\mathbf{a} \cdot \mathbf{c})\,(\mathbf{b} \cdot \mathbf{d}) - (\mathbf{a} \cdot \mathbf{d})\,(\mathbf{b} \cdot \mathbf{c}) \ , \tag{1.33}$$

$$\mathbf{a} \times [\mathbf{b} \times \mathbf{c}] = (\mathbf{a} \cdot \mathbf{c})\,\mathbf{b} - (\mathbf{a} \cdot \mathbf{b})\,\mathbf{c} \, , \tag{1.34}$$

$$[\mathbf{a} \times \mathbf{b}] \times \mathbf{c} = (\mathbf{a} \cdot \mathbf{c})\,\mathbf{b} - (\mathbf{c} \cdot \mathbf{b})\,\mathbf{a} \, , \tag{1.35}$$

and hence, correspondingly, that

$$\mathbf{a} \times [\mathbf{b} \times \mathbf{c}] + \mathbf{b} \times [\mathbf{c} \times \mathbf{a}] + \mathbf{c} \times [\mathbf{a} \times \mathbf{b}] = 0 \, . \tag{1.36}$$

Problem 1.28. *Show that the area of a parallelogram defined by two vectors* $\mathbf{a}$ *and* $\mathbf{b}$*, as shown in Fig. 1.13(a), is given by the absolute value of their vector product:*

$$S_{paralellogram} = |\mathbf{a} \times \mathbf{b}| \ . \tag{1.37}$$

For instance, if $\mathbf{u} = (0, 1, 3)$, $\mathbf{v} = (-1, 1, 2)$, then $\mathbf{u} + \mathbf{v} = (-1, 2, 5)$, $\mathbf{u} - \mathbf{v} = (1, 0, 1)$, $(\mathbf{u}, \mathbf{v}) = 0 \cdot (-1) + 1 \cdot 1 + 3 \cdot 2 = 7$ and

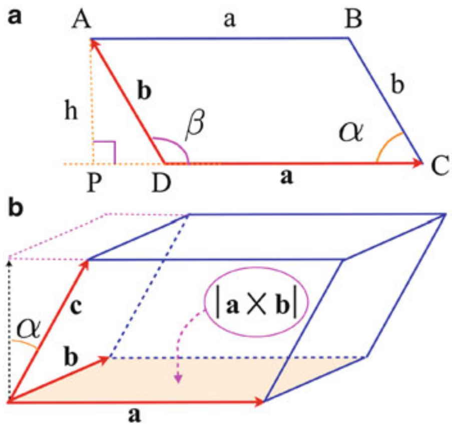


Fig. 1.13 (**a**) Illustration for the calculation of the area of a parallelogram defined by two vectors **a** and **b**, and (**b**) of the volume of a parallelepiped defined by three non-planar vectors **a**, **b** and **c**

$$\mathbf{u} \times \mathbf{v} = (1 \cdot 2 - 3 \cdot 1)\mathbf{i} + (3 \cdot (-1) - 0 \cdot 2)\mathbf{j} + (0 \cdot 1 - 1 \cdot (-1))\mathbf{k} = (-1, -3, 1) \ .$$

It is also possible to mix the dot and vector products into the so-called *mixed* or *triple product*, denoted by putting the three vectors within square brackets, e.g. $[\mathbf{c}, \mathbf{a}, \mathbf{b}]$. Indeed, a vector product of two vectors is a vector which can be dot-multiplied with a third vector giving a scalar:

$$\begin{aligned}
[\mathbf{c}, \mathbf{a}, \mathbf{b}] &= \mathbf{c} \cdot [\mathbf{a} \times \mathbf{b}] \\
&= (c_1\mathbf{i} + c_2\mathbf{j} + c_3\mathbf{k}) \cdot [(a_2b_3 - a_3b_2)\mathbf{i} + (a_3b_1 - a_1b_3)\mathbf{j} + (a_1b_2 - a_2b_1)\mathbf{k}] \\
&= (a_2b_3 - a_3b_2)c_1 + (a_3b_1 - a_1b_3)c_2 + (a_1b_2 - a_2b_1)c_3 \\
&= \begin{vmatrix} c_1 & c_2 & c_3 \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} \ .
\end{aligned} \tag{1.38}$$

The last equality trivially follows from formula (1.31) as we effectively replace **i** with c_1, **j** with c_2 and **k** with c_3 in Eq. (1.30); this is exactly what we have done in Eq. (1.38). The triple product has some important symmetry properties listed below which follow from its determinant representation and the fact, discussed earlier in Sect. 1.6, that a determinant changes sign if two of its rows are exchanged:

$$[\mathbf{c}, \mathbf{a}, \mathbf{b}] = [\mathbf{a}, \mathbf{b}, \mathbf{c}] = [\mathbf{b}, \mathbf{c}, \mathbf{a}] = -[\mathbf{c}, \mathbf{b}, \mathbf{a}] = -[\mathbf{b}, \mathbf{a}, \mathbf{c}] = -[\mathbf{a}, \mathbf{c}, \mathbf{b}] \ , \tag{1.39}$$

which show that the vectors in the triple product can be "rotated" in a cyclic order **abc** → **bca** → **cab** → **abc** → ⋯ and this does not affect the sign of the product; if the order is destroyed, a minus sign appears. If two permutations are performed (this corresponds to the cyclic permutation, e.g. **abc** → **bac** → **bca**), then no change of the sign occurs.

Problem 1.29. *Show that*

$$[\mathbf{a}, \mathbf{a}, \mathbf{b}] = [\mathbf{a}, \mathbf{b}, \mathbf{a}] = [\mathbf{b}, \mathbf{a}, \mathbf{a}] = 0 \ . \tag{1.40}$$

[*Hint*: use the explicit expression (1.38) for the triple product preceding the determinant.[6]]

Problem 1.30. *Using Fig. 1.13(b), show that the volume of a parallelepiped formed by three non-planar vectors* **a**, **b** *and* **c** *is given by the absolute value of their triple product:*

$$V_{parallelepiped} = |[\mathbf{a}, \mathbf{b}, \mathbf{c}]| = |[\mathbf{a} \times \mathbf{b}] \cdot \mathbf{c}| \ . \tag{1.41}$$

The immediate consequence of the definition of the triple product is that it is equal to zero if the three vectors lie in the same plane, i.e. are *coplanar*. Indeed, in this case one of the vectors must be a linear combination of the other two and hence the volume of the parallelepiped constructed from these three vectors will obviously be zero: in the corresponding determinant of the triple product one of its rows will be a linear combination of the other two.

The Cartesian basis introduced above, $\{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$, is not the only one possible; one can define any three non-collinear and non-planar vectors $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$ to serve as a new basis and then any vector $\mathbf{x}$ from $\mathbb{R}^3$ can then be expanded with respect to it:

$$\mathbf{x} = g_1\mathbf{a}_1 + g_2\mathbf{a}_2 + g_3\mathbf{a}_3 \ . \tag{1.42}$$

This is proven in the following way. Let us write the new basis via their Cartesian coordinates: $\mathbf{a}_1 = (a_{11}, a_{12}, a_{13})$ and similarly for $\mathbf{a}_2$ and $\mathbf{a}_3$. Here a_{ij} corresponds to the j-th Cartesian component ($j = 1, 2, 3$) of the vector $\mathbf{a}_i$ ($i = 1, 2, 3$). Then the Cartesian components $\{x_j\}$ of $\mathbf{x}$ can be expressed via $\{a_{ij}\}$ as follows:

$$\begin{aligned} &x_1 = g_1a_{11} + g_2a_{21} + g_3a_{31} \ , \quad x_2 = g_1a_{12} + g_2a_{22} + g_3a_{32} \ , \\ &x_3 = g_1a_{13} + g_2a_{23} + g_3a_{33} \ . \end{aligned} \tag{1.43}$$

One can see that the "coordinates" $\{g_1, g_2, g_3\}$ in the new basis define uniquely the Cartesian components $\{x_i\}$ of the vector $\mathbf{x}$; inversely, given the Cartesian components of $\mathbf{x} = (x_1, x_2, x_3)$, one determines uniquely the coordinates $\{g_1, g_2, g_3\}$ in the new basis by solving the system of three algebraic equations (1.43). It is

[6] The required statement follows immediately from the determinant formula for the triple product since the determinant is equal to zero if any two of its rows are identical.

shown in linear algebra that for the unique solution of these equations to exist it is necessary that the following condition be satisfied:

$$\begin{vmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \\ a_{13} & a_{23} & a_{33} \end{vmatrix} \neq 0 \ , \tag{1.44}$$

i.e. the determinant composed of Cartesian components of the three basis vectors be non-zero. It must be clear that this condition is satisfied as long as the three vectors are non-collinear and non-planar (in the latter case one must be a linear combination of the other two).

Example 1.1. ► Expand vector $\mathbf{x} = (2, 1, -1)$ in terms of three vectors $\mathbf{a}_1 = (1, 0, 0)$, $\mathbf{a}_2 = (1, 1, 0)$ and $\mathbf{a}_3 = (1, 1, 1)$.

Solution. solving the system of linear equations (1.43),

$$\begin{cases} g_1 + g_2 + g_3 = 2 \\ \quad g_2 + g_3 = 1 \\ \qquad g_3 = -1 \end{cases} ,$$

we obtain $g_3 = -1$, then $g_2 = 2$ and $g_1 = 1$, i.e. the expansion sought for is $\mathbf{x} = \mathbf{a}_1 + 2\mathbf{a}_2 - \mathbf{a}_3$, which can now be checked by direct calculation. ◄

Alongside a direct space basis $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$, it is sometimes useful to introduce another special basis, $\{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\}$, called *reciprocal* or *biorthogonal* basis. These basis vectors are defined by the following relations:

$$\mathbf{b}_1 = \frac{1}{v_c}\left[\mathbf{a}_2 \times \mathbf{a}_3\right] \ , \ \mathbf{b}_2 = \frac{1}{v_c}\left[\mathbf{a}_3 \times \mathbf{a}_1\right] \ , \ \mathbf{b}_3 = \frac{1}{v_c}\left[\mathbf{a}_1 \times \mathbf{a}_2\right] \ , \tag{1.45}$$

where $v_c = [\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3]$ is the volume of the parallelepiped (assumed to be positive) formed by the three original basis vectors $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$. Notice that the index of each reciprocal vector and the two old vectors it is related to form an ordered triple in each case. It is easy to see using properties (1.39) and (1.40) of the triple product that the new basis is related to the old one via:

$$\mathbf{a}_i \cdot \mathbf{b}_j = \begin{cases} 0 \ , & \text{if } i \neq j \\ 1 \ , & \text{if } i = j \end{cases} = \delta_{ij} \ , \tag{1.46}$$

where we introduced the *Kronecker symbol* which is equal to zero if its two indices are different, otherwise it is equal to one.[7]

The volumes of the parallelepipeds $v_c = [\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3]$ and $v_g = [\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3]$, formed by the two sets of vectors are inversely related. Indeed,

[7] In solid state physics and crystallography the reciprocal lattice is defined with an additional factor of 2π, so that instead of (1.46) we have $\mathbf{a}_i \cdot \mathbf{b}_j = 2\pi\delta_{ij}$.

$$v_g = \mathbf{b}_1 \cdot [\mathbf{b}_2 \times \mathbf{b}_3] = \frac{1}{v_c}[\mathbf{a}_2 \times \mathbf{a}_3] \cdot [\mathbf{b}_2 \times \mathbf{b}_3]$$
$$= \frac{1}{v_c}\left\{(\mathbf{a}_2 \cdot \mathbf{b}_2)(\mathbf{a}_3 \cdot \mathbf{b}_3) - (\mathbf{a}_2 \cdot \mathbf{b}_3)(\mathbf{a}_3 \cdot \mathbf{b}_2)\right\} = \frac{1}{v_c} , \tag{1.47}$$

where we made use of (1.33) and then (1.46) for the four dot products of direct and reciprocal vectors. Moreover, it is easy to see that direct basis vectors are reciprocal to the reciprocal ones; in other words, the two sets of vectors are mutually reciprocal. This is proven in the following problem.

Problem 1.31. *Show that direct basis vectors can be written via the reciprocal vectors similarly to Eq. (1.45), i.e. that*

$$\mathbf{a}_1 = \frac{1}{v_g}[\mathbf{b}_2 \times \mathbf{b}_3] \ , \ \mathbf{a}_2 = \frac{1}{v_g}[\mathbf{b}_3 \times \mathbf{b}_1] \ , \ \mathbf{a}_3 = \frac{1}{v_g}[\mathbf{b}_1 \times \mathbf{b}_2] \ .$$

[*Hint*: in calculating vector products of reciprocal vectors replace only one of them via direct ones using Eq. (1.45), and then make use of either Eq. (1.34) or (1.35) for the double vector product.]

Problem 1.32. *The crystal lattice is characterised by three lattice vectors* $\mathbf{a}_1$*,* $\mathbf{a}_2$ *and* $\mathbf{a}_3$*, which form its primitive unit cell. By repeating this cell in all three directions the whole periodic crystal lattice is constructed. Let* a*,* b *and* c *be the lengths of the three vectors, and* α*,* β *and* γ *the angles between the first and the third, the second and the third, and the first and the second vectors, respectively. Assuming that* $\mathbf{a}_1$ *lies along the* x *axes,* $\mathbf{a}_2$ *in the* $x-y$ *plane, work out the Cartesian components of all three lattice vectors.*

Direct and reciprocal vectors are frequently used in crystallography and solid state physics to build the crystal lattices in both spaces.

1.7.2 N-dimensional Space

Most of the material we considered above is generalised to abstract spaces of arbitrary N dimensions ($N \geq 1$). In this space any point X is specified by N coordinates x_1, x_2, ..., x_N, i.e. $X = (x_1, \ldots, x_N)$. The "distance" d_{XY} between two points $X = (x_1, \ldots, x_N)$ and $Y = (y_1, \ldots, y_N)$ is defined as a direct generalisation of the result (1.19) for the 3D space:

$$d_{XY} = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} \ . \tag{1.48}$$

We can also define an N-dimensional vector $\overrightarrow{XY} = \mathbf{a} = (y_1 - x_1, y_2 - x_2, \ldots, y_N - x_N)$ connecting the points X and Y, so that d_{XY} would serve as its length. Similarly to the 3D case, vectors can be added or subtracted from each other or multiplied by a number; each such an operation results in a vector in the same space: if there are two vectors $\mathbf{a} = (a_1, \ldots, a_N)$ and $\mathbf{b} = (b_1, \ldots, b_N)$, then we define

$$\mathbf{a} \pm \mathbf{b} = \mathbf{c} = (c_1, \ldots, c_N) \quad \text{with} \quad c_i = a_i \pm b_i \ , \ i = 1, \ldots, N \ ,$$

and

$$\alpha \mathbf{a} = \mathbf{g} \quad \text{with} \quad g_i = \alpha a_i \ , \ i = 1, \ldots, N \ ,$$

where α is a number from $\mathbb{R}$. Correspondingly, we can also define a dot or scalar product of two vectors $\mathbf{a} = (a_1, \ldots, a_N)$ and $\mathbf{b} = (b_1, \ldots, b_N)$ as a direct generalisation of Eq. (1.20):

$$\mathbf{a} \cdot \mathbf{b} = (\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{N} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_N b_N \ . \tag{1.49}$$

As was the case for the 3D space, the square of a vector length is given by the dot product of the vector with itself: $|\mathbf{a}|^2 = \mathbf{a} \cdot \mathbf{a}$. All properties of the dot product are directly transferred to the N-dimensional space. The distance between two points X and Y is then given by

$$d_{XY} = \sqrt{(\mathbf{x} - \mathbf{y})^2} = \sqrt{(\mathbf{x} - \mathbf{y}) \cdot (\mathbf{x} - \mathbf{y})} = |\mathbf{x} - \mathbf{y}| \ .$$

Next, we introduce N basis vectors (or unit base vectors) of the space as $\mathbf{e}_1 = (1, 0, \ldots, 0)$, $\mathbf{e}_2 = (0, 1, 0, \ldots, 0)$, etc., $\mathbf{e}_N = (0, 0, \ldots, 0, 1)$. Each of these vectors consists of $N-1$ zeros and a single number one; the latter is positioned exactly at the i-th place in $\mathbf{e}_i$. Each such vector is obviously of unit length and they are orthogonal, i.e. $\mathbf{e}_i \cdot \mathbf{e}_j = \delta_{ij}$. Then, any vector $\mathbf{a} = (a_1, \ldots, a_N)$ in such a space can be expanded via the basis vectors as

$$\mathbf{a} = a_1 \mathbf{e}_1 + a_2 \mathbf{e}_2 + \cdots + a_N \mathbf{e}_N = \sum_{i=1}^{N} a_i \mathbf{e}_i \ .$$

So far we have been using only vectors-rows, e.g. $X = (x_1, \ldots, x_N)$, when the coordinates of the vector are listed along a line (a row). In some cases vector-columns

$$X = \begin{pmatrix} x_1 \\ \cdots \\ x_N \end{pmatrix}$$

may be more useful.

1.8 Introduction to Complex Numbers

We know that a square of a real number is always positive. Therefore, a square root of a negative number does not make sense within the manifold of real numbers. This fact causes some problems in solving algebraic equations of power more than one. An extension of real numbers to a form which allows representation of a square root of negative numbers was first introduced by Gerolamo Cardano and are called *complex numbers*.[8]

These are defined by a pair of two real numbers x (which is called a real part) and y (a complex part) and a special object i as follows: $z = x + iy$. The i is defined in such a way that its square is equal to -1, i.e. $i^2 = -1$ or $i = \sqrt{-1}$. It is also a complex number with zero real part and the complex part equal to one. Two complex numbers are identical if and only if their corresponding real and imaginary parts coincide.

Complex numbers are conveniently shown as points (x, y) on a 2D plane, called the *complex plane*, with its real and imaginary parts x and y being the point coordinates, Fig. 1.14. The distance $r = \sqrt{x^2 + y^2}$ from the point to the origin of the coordinate system is called the *absolute value* of the complex number z, while the angle ϕ (see the figure) the *argument* or *phase*. Of course, $x = r \cos\phi$ and $y = r \sin\phi$, i.e. the complex number can be also written as

$$z = x + iy = r (\cos\phi + i \sin\phi) \ . \tag{1.50}$$

To the phase of a complex number a multiple of 2π may be added without any change of the number itself, i.e. the phases ϕ and $\phi + 2\pi k$ with $k = \pm 1, \pm 2, \ldots$

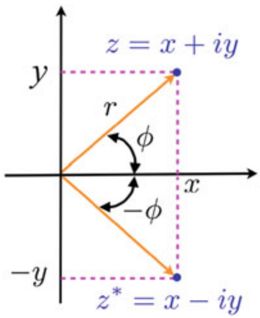


Fig. 1.14 A complex number $z = x+iy$ is depicted as a point (x, y) with coordinates x and y on the 2D plane, the so-called complex plane. Alternatively the vector from the origin of the coordinate system to the point is also frequently used. The complex conjugate number $(z)^* = z^* = x - iy$ is depicted by the point reflected upon the x axis. It has the same r but its phase is $-\phi$

[8]These were developed further by Rafael Bombelli and William Rowan Hamilton.

result in the same complex number. The phase can be obtained from the values of $y = r \sin\phi$ and $x = r \cos\phi$ via

$$\frac{y}{x} = \frac{r \sin\phi}{r \cos\phi} = \tan\phi \ , \tag{1.51}$$

i.e. by solving the above equation with respect to the angle ϕ. We shall learn in Sect. 2.3.9 that there is a function which does the job, called arctangent, but for now we shall just limit ourselves with this result.

Problem 1.33. *Determine the absolute value r and the phase ϕ of the following complex numbers:* $1 + i$; $1 - i$; $\sqrt{3} + i$. [*Answers*: $r = \sqrt{2}$, $\phi = \pi/4$; $r = \sqrt{2}$, $\phi = -\pi/4$; $r = 2$, $\phi = \pi/6$.]

The vector from the coordinate origin to the point (x, y) on the complex plane can also be used to represent the complex number $z = x + iy$ on the plane, Fig. 1.14. In that case r is the vector length and ϕ is the angle the vector makes with the x axis. Two identical complex numbers correspond to the same point on the complex plane.

If $y = 0$, the numbers are positioned on the x axis and are purely real, and the phase is $\phi = \pi n$ with any positive and negative integer n, including zero: even n corresponds to positive, while odd n corresponds to the negative direction of the x axis. If the complex number is purely imaginary, then the phase is given generally as $\phi = \pi/2 + \pi n$. Any arithmetic operation with any two real numbers would result in a number on the same axis, i.e. within the manifold of real numbers as we discussed at length above. However, taking a square root of a negative number takes the result immediately out of that axis, e.g. $\sqrt{-9} = \sqrt{9(-1)} = \sqrt{9}\sqrt{-1} = 3i$.

The addition and subtraction (two of the four elementary operations) on complex numbers are defined in the following way:

$$(x_1 + iy_1) + (x_2 + iy_2) = x_3 + iy_3 = (x_1 + x_2) + i(y_1 + y_2) \quad \text{(addition)}, \tag{1.52}$$

$$(x_1 + iy_1) - (x_2 + iy_2) = x_3 + iy_3 = (x_1 - x_2) + i(y_1 - y_2) \quad \text{(subtraction)}. \tag{1.53}$$

These definitions ensure that the sum (difference) of two numbers are equivalent to a sum (difference) of two vectors corresponding to them. It is also seen that the sum (difference) is commutative ($z_1 + z_2 = z_2 + z_1$ and $z_1 - z_2 = -z_2 + z_1$) and associative, e.g. $(z_1 + z_2) + z_3 = z_1 + (z_2 + z_3)$.

When defining the product of two numbers, $z_3 = z_1 z_2$, a number of conditions are to be satisfied: (i) the product must be equivalent to the product of real numbers if the imaginary parts of z_1 and z_2 are both zeros, i.e. $x_3 = x_1 x_2$ and $y_3 = 0$; (ii) the product should be a complex number, i.e. it must have the form $x_3 + iy_3$; (iii) algebraic rules must be respected. These conditions are satisfied if we treat a product of two complex numbers using usual algebraic rules; in fact, the form $x + iy$ already contains such a product (with i being multiplied by y, i.e. i is a number in its own right). Hence we *define* the product as follows:

$$\begin{aligned}(x_1 + iy_1)(x_2 + iy_2) &= x_1x_2 + iy_1x_2 + ix_1y_2 + i^2y_1y_2 \\ &= \left(x_1x_2 + i^2y_1y_2\right) + i(x_1y_2 + y_1x_2) \ .\end{aligned}$$

Since $i^2 = -1$ by definition, we finally obtain:

$$(x_1 + iy_1)(x_2 + iy_2) = (x_1x_2 - y_1y_2) + i(x_1y_2 + y_1x_2) \quad \text{(product)}. \tag{1.54}$$

It is easy to see from (1.54) that if $y_1 = y_2 = 0$, then the imaginary part of the product is also equal to zero, i.e. we remain within the real numbers and that indeed we get x_1x_2. Also, the product clearly has the form of a complex number. Finally, the product is commutative, i.e. $z_1z_2 = z_2z_1$, and associative, $(z_1z_2)z_3 = z_1(z_2z_3)$, which guarantee (together with the analogous properties of the sum and difference) that the algebraic rules, however complex, will be indeed respected.

Next, we consider division of complex numbers. This is defined as an operation inverse to the product, i.e. $z_1/z_2 = z_3$ is understood in the sense that $z_1 = z_2z_3$.

Problem 1.34. *Using this definition of the division and explicit definition of the product (1.54), prove the following formula for the division:*

$$\frac{z_1}{z_2} = \frac{x_1 + iy_1}{x_2 + iy_2} = \frac{x_1x_2 + y_1y_2}{x_2^2 + y_2^2} + i\frac{y_1x_2 - x_1y_2}{x_2^2 + y_2^2} \quad \textit{(division)}. \tag{1.55}$$

[*Hint*: write $z_1 = z_2z = z_2(x + iy)$ and solve for x and y using the fact that two complex numbers are equal if and only if their respective real and imaginary parts are equal.]

We shall see later on in Sect. 2.3.8 that when we multiply two complex numbers, their absolute values multiply and the phases sum up; when we divide two numbers, their absolute values are divided and the phases are subtracted.

Finally, a special operation characteristic only to complex numbers is also introduced whereby the signs to all i in the expression are inverted. This is called *complex conjugation* and is denoted by the superscript *, i.e. $(5 - i6)^* = 5 + i6$. The complex conjugate number z^* corresponds to the reflection of the position of the number z on the complex plane with respect to the x axis, Fig. 1.14.

Problem 1.35. *Prove the following identities:*

$$z + z^* = 2x \quad \text{and} \quad z - z^* = 2iy \; .$$

Problem 1.36. *Prove the following inequality:*

$$|z| = |x + iy| \leq |x| + |y| \; . \tag{1.56}$$

Problem 1.37. *The square of the absolute value of a complex number r^2 is very often denoted $|z|^2$ and called the square of the module of the complex number z. Prove:*

$$|z|^2 = x^2 + y^2 = zz^* \; . \tag{1.57}$$

Another (much simpler) way of proving the result for the division (1.55) can be done using complex conjugation: multiply and divide the number z_1/z_2 by z_2^* and make use of Eq. (1.57):

$$\frac{x_1 + iy_1}{x_2 + iy_2} = \frac{x_1 + iy_1}{x_2 + iy_2} \cdot \frac{x_2 - iy_2}{x_2 - iy_2} = \frac{(x_1 + iy_1)(x_2 - iy_2)}{x_2^2 + y_2^2}$$
$$= \frac{x_1 x_2 + y_1 y_2}{x_2^2 + y_2^2} + i\frac{y_1 x_2 - x_1 y_2}{x_2^2 + y_2^2} \; .$$

Problem 1.38. *Calculate the following complex numbers (i.e. present them in the form $a + ib$):*

$$(1 + i)(3 - 2i) \; , \quad (1 + i)(3 - 2i)^* \; , \quad \frac{1 + i}{3 - 2i} \; , \quad (1 + i) - (3 - 2i) \; .$$

[*Answers*: $5 + i$; $1 + 5i$; $(1 + 5i)/13$; $-2 + 3i$.]

Problem 1.39. *Work out the absolute values and the phases of the numbers:* $1 + i$, $1 - i$, i, 1 *and* -1. [*Answers*: $\sqrt{2}$, $\pi/4$; $\sqrt{2}$, $-\pi/4$; 1, $\pi/2$; 1, 0; 1, π.]

1.9 Summation of Finite Series

It is often necessary to sum a finite set of terms

$$S_n = a_0 + a_1 + a_2 + \cdots + a_n = \sum_{i=0}^{n} a_i \ , \tag{1.58}$$

containing $n+1$ terms constructed via a certain rule, i.e. each term a_i is constructed from its number i (the index). Here we shall consider several examples of such finite sums.

In a *geometrical progression* $a_i = a_0 q^i$, i.e. the numbers a_i form a finite *sequence* $a_0, a_0 q, a_0 q^2, \ldots, a_0 q^n$. This means that each next term is obtained from the previous one via a multiplication with q:

$$a_{i+1} = a_i q \ , \ \text{ where } i = 0, 1, 2, \ldots \ . \tag{1.59}$$

This is called a *recurrence relation*. We would like to calculate the sum

$$S_n = \sum_{i=0}^{n} a_i = a_0 + a_0 q + a_0 q^2 + \cdots + a_0 q^n \ . \tag{1.60}$$

The trick here is to notice that if we multiply S_n by q, we can manipulate the result into an expression containing S_n again:

$$\begin{aligned} S_n q &= a_0 q + a_0 q^2 + a_0 q^3 + \cdots + a_0 q^{n+1} = \left(a_0 + a_0 q + a_0 q^2 + \cdots + a_0 q^n\right) \\ &\quad - a_0 + a_0 q^{n+1} = S_n + a_0 \left(q^{n+1} - 1\right) \end{aligned}$$

Solving this equation with respect to S_n, we finally obtain:

$$S_n = a_0 \frac{q^{n+1} - 1}{q - 1}, \tag{1.61}$$

which is the desired result. Note that this formula is valid for both $q > 1$ and $q < 1$, but not for $q = 1$. In the latter case, however, all terms in the sum (1.60) are identical and $S_n = (n+1)a_0$ which is a trivial result. We shall see later on in Sect. 3.9 that the formula (1.61), in spite of the singularity, can in fact be used for the case of $q = 1$ as well, i.e. it is actually quite general, but we have to learn limits before we can establish that and hence have to postpone the necessary discussion.

Although we have derived the sum of the geometrical progression from its definition, it is also instructive to illustrate using this example how the induction principle we discussed in Sect. 1.1 works. So, let us suppose that we *are given* the result (1.61) for the sum of the progression (1.60), and now we would like to *prove* that it is correct for any n. First of all, we check that the formula (1.61) works for the lowest value of $n = 0$, and indeed, we get

$$S_0 = a_0 \frac{q^1 - 1}{q - 1} = a_0,$$

as expected. Then we *assume* that (1.61) is valid for some general n, and hence we consider

$$S_{n+1} = S_n + a_0 q^{n+1} = a_0 \frac{q^{n+1}-1}{q-1} + a_0 q^{n+1} = a_0 \frac{q^{n+1}-1+q^{n+2}-q^{n+1}}{q-1} = a_0 \frac{q^{n+2}-1}{q-1} ,$$

which does have the correct form as it can be obtained from (1.61) by the substitution $n \to n+1$, as required.

Example 1.2. ► Sum up the finite series

$$3^2 + 3 + 1 + \frac{1}{3} + \frac{1}{3^2} + \cdots + \frac{1}{3^{10}} .$$

Solution. This is a geometrical progression with $a_0 = 3^2 = 9$ and $q = 1/3$, containing $n + 1 = 13$ terms. Using our general result (1.61), we thus get

$$S_{12} = 9 \frac{(1/3)^{13} - 1}{1/3 - 1} = \frac{27}{2}\left(1 - \frac{1}{3^{13}}\right) . ◄$$

Problem 1.40. *Sum up the finite numerical series:*

$$(a)\ 1 + 2 + 4 + 8 + \cdots + 1024 ; \quad (b)\ 3^2 - 3 + 1 - \frac{1}{3} + \frac{1}{3^2} - \cdots + \frac{1}{3^{10}} .$$

[*Answers*: $2^{11} - 1$; $\left(3^3 + 3^{-10}\right)/4$.]

Problem 1.41. *Show that*

$$\sum_{n=1}^{N} \left(a^n + \frac{1}{a^n}\right)^2 = 2(N-1) + \frac{\left(a^{2N+2}-1\right)\left(a^{2N}+1\right)}{a^{2N}\left(a^2-1\right)} .$$

In an *arithmetic progression* $a_i = a_0 + ir$, i.e. each term is obtained from the previous one by adding a constant r:

$$a_{i+1} = a_i + r , \text{ where } i = 0, 1, 2, \ldots , \tag{1.62}$$

and we would like to calculate the sum:

$$S_n = \sum_{i=0}^{n} a_i = a_0 + [a_0 + r] + [a_0 + 2r] + \cdots + [a_0 + (n-1)r] + [a_0 + nr] . \tag{1.63}$$

The trick here which will enable us to derive a simple expression for the sum S_n is to notice that each term in the series can also be written via the last term: $a_k = a_0 + kr = a_n - (n-k)r$. Therefore, the sum (1.63) can alternatively be written as:

$$S_n = [a_n - nr] + [a_n - (n-1)r] + [a_n - (n-2)r] + \cdots + [a_n - r] + a_n \ . \tag{1.64}$$

Here we have exactly the same terms containing r as in (1.63), but in the reverse order and with the minus sign each. Therefore, all these terms cancel out if we sum the two equalities (1.63) and (1.64):

$$2S_n = (n+1)a_0 + (n+1)a_n \ \Rightarrow \ S_n = \frac{a_0 + a_n}{2}(n+1) \ , \tag{1.65}$$

where the factor of $(n+1)$ appeared since we have both a_0 and a_n that number of times in the two sums above. The final result is very simple: you take the average of the first and the last terms which is to be multiplied by the total number of terms in the finite series (which is $n+1$).

Problem 1.42. *Calculate the sum of all integers from 10 to 1000.* [*Answer*: 500455.]

Many other finite series are known, some of them are given in the problems below.

Problem 1.43. *Use the mathematical induction to prove the formulae given below:*

$$\sum_{i=1}^{n} i^2 = 1 + 2^2 + 3^2 + \cdots + n^2 = \frac{1}{6}n(n+1)(2n+1) \ ; \tag{1.66}$$

$$\sum_{i=1}^{n} i^3 = 1 + 2^3 + 3^3 + \cdots + n^3 = \frac{1}{4}n^2(n+1)^2 \ ; \tag{1.67}$$

$$\sum_{i=2}^{n} \frac{1}{i^2-1} = \frac{3}{4} - \frac{2n+1}{2n(n+1)} \ ; \tag{1.68}$$

$$\sum_{i=0}^{n} (a_0 + ir)q^i = a_0\frac{1-q^{n+1}}{1-q} + \frac{rq}{(1-q)^2}\left[1-(n+1)q^n + nq^{n+1}\right] \ . \tag{1.69}$$

The latter series is a combination of a geometric and arithmetic progressions.

Problem 1.44. *Prove that*

$$\sum_{i=1}^{n} \frac{1}{i(i+1)} = 1 - \frac{1}{n+1} \ . \tag{1.70}$$

[*Hint*: make use of the fact that $1/i(i+1) = 1/i - 1/(i+1)$.]

1.10 Binomial Formula

There are numerous examples when one needs to expand an expression like $(a+b)^n$ with integer $n \geq 1$. We can easily multiply the brackets for small values of the n to get:

$$(a+b)^1 = a + b = a^1b^0 + a^0b^1 \ ,$$

$$(a+b)^2 = a^2 + 2ab + b^2 = a^2b^0 + 2a^1b^1 + a^0b^2 \ ,$$

$$(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3 = a^3b^0 + 3a^2b^1 + 3a^1b^2 + a^0b^3 \ ,$$

$$(a+b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4 = a^4b^0 + 4a^3b^1 + 6a^2b^2 + 4a^1b^3 + a^0b^4 \ .$$

This can be continued, but the expressions get much longer. We can notice, however, some general rules in expanding the $(a+b)^n$ from the particular cases considered above. Indeed, the expansions contain all possible products a^ib^{n-i}, starting from $i = 0$ and ending with $i = n$, such that the sum of powers is always $i + (n - i) = n$. The coefficients to these products form sequences: $(1, 1)$ for $n = 1$, $(1, 2, 1)$ for $n = 2$, $(1, 3, 3, 1)$ for $n = 3$, $(1, 4, 6, 4, 1)$ for $n = 4$, and so on. One may notice that a set of coefficients for the power $n + 1$ can actually be obtained from the coefficients corresponding to the previous power n. This elegant rule known as Pascal's triangle[9] is illustrated in Fig. 1.15. Consider, for instance, the next to $n = 4$ set of coefficients, i.e. for $n = 5$. There will be six terms in the expansion with coefficients $(1, 5, 10, 10, 5, 1)$, and they are constructed following two simple rules: (i) the first and the last coefficients (for the terms $a^5b^0 = a^5$ and $a^0b^5 = b^5$) are 1; (ii) any other coefficient is obtained as a sum of the two coefficients directly above it from the previous line (corresponding to $n = 4$ in our case).

Although this construction is elegant, it can only be useful for small values of n; moreover, it is not useful at all if one has to manipulate the binomial expressions

[9] After Blaise Pascal.

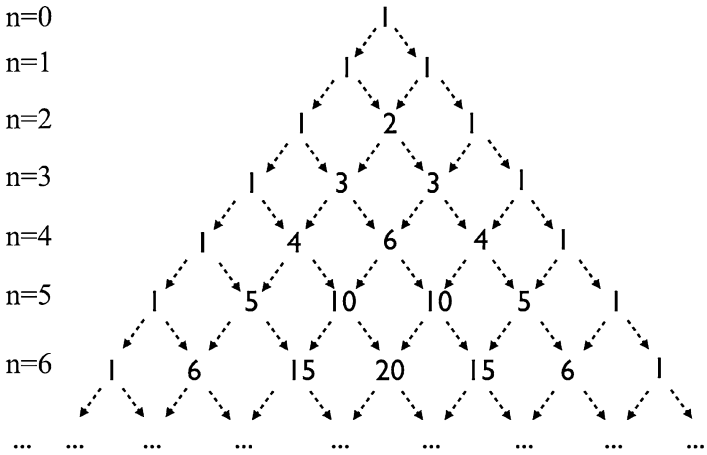


Fig. 1.15 Pascal's triangle

$(a+b)^n$ analytically for a general integer n. This problem is solved by the *binomial formula* (or theorem)[10]:

$$(a+b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i} , \tag{1.71}$$

where

$$\binom{n}{i} = \frac{n!}{i!(n-i)!} \tag{1.72}$$

are called *binomial coefficients*. Here $n! = 1 \cdot 2 \cdot 3 \ldots \cdot n$ is a short-hand notation for a product of all integers between 1 and n, called *n-factorial.* The formula (1.71) can also formally be written as follows:

$$(a+b)^n = \sum_{k_1+k_2=n} \binom{n}{k_1, k_2} a^{k_1} b^{k_2} , \tag{1.73}$$

[10]The formula and the triangle were described by Blaise Pascal, but apparently were known before him; in particular, a Persian mathematician and engineer Abu Bakr ibn Muhammad ibn al Husayn al-Karaji (or al-Karkhi) wrote about them in the tenth century as well; he has also proved the formula by introducing the principle of mathematical induction.

where $\begin{pmatrix} n \\ k_1, k_2 \end{pmatrix} = \begin{pmatrix} n \\ k_1 \end{pmatrix}$. The sum in this formula is taken over all possible indices k_1 and k_2 between 0 and n such that $k_1 + k_2 = n$. It is easy to see that this formula is equivalent to the previous one, Eq. (1.71), if we take $k_1 = i$ as then $k_2 = n - i$ has a single value and the sum in (1.73) becomes a single sum over i between 0 and n.

Let us establish some essential properties of these coefficients:

$$\begin{pmatrix} n \\ 0 \end{pmatrix} = \frac{n!}{0!n!} = 1 \ , \ \begin{pmatrix} n \\ n \end{pmatrix} = \frac{n!}{n!0!} = 1 \ , \tag{1.74}$$

i.e. the first and the last coefficients (which are by the highest powers of a and b) are, for any n, equal to 1. This is the first property of Pascal's triangle. The second property of Pascal's triangle is established by the following identity:

$$\begin{aligned} \begin{pmatrix} n \\ i-1 \end{pmatrix} + \begin{pmatrix} n \\ i \end{pmatrix} &= \frac{n!}{(i-1)!(n-i+1)!} + \frac{n!}{i!(n-i)!} \\ &= \frac{n!}{(i-1)!(n-i)!}\left[\frac{1}{n-i+1} + \frac{1}{i}\right] \\ &= \frac{n!}{(i-1)!(n-i)!}\frac{n+1}{i(n-i+1)} = \frac{(n+1)!}{i!(n+1-i)!} = \begin{pmatrix} n+1 \\ i \end{pmatrix} , \end{aligned} \tag{1.75}$$

where we have made use of the fact that $m! = (m-1)!m$ for any positive integer m. We can now indeed see that two binomial coefficients by the $(i-1)$-th and i-th terms of the expansion of $(a+b)^n$ give rise to the i-th coefficient of the $(a+b)^{n+1}$ expansion.

Problem 1.45. *Prove the symmetry property:*

$$\begin{pmatrix} n \\ k \end{pmatrix} = \begin{pmatrix} n \\ n-k \end{pmatrix} \ , \tag{1.76}$$

which basically means that the binomial coefficients are the same in the binomial formula whether one counts from the beginning to the end or vice versa. Equation (1.74) is a particular case of this symmetry written for the first and the last coefficients.

Problem 1.46. *Prove an identity:*

$$\begin{pmatrix} n \\ k \end{pmatrix}\begin{pmatrix} n-k \\ m-k \end{pmatrix} = \begin{pmatrix} m \\ k \end{pmatrix}\begin{pmatrix} n \\ m \end{pmatrix} .$$

These properties do not yet prove the binomial formula. However, before proving it, let us first see that it actually works. Consider the case of $n = 5$. Since the sum runs from $i = 0$ up to $i = 5$, there will be six terms with the coefficients:

$$\binom{5}{0} = 1\,, \quad \binom{5}{1} = \frac{5!}{1!4!} = 5\,, \quad \binom{5}{2} = \frac{5!}{2!3!} = 10\,,$$

$$\binom{5}{3} = \frac{5!}{3!2!} = 10\,, \quad \binom{5}{4} = \frac{5!}{4!1!} = 5\,, \quad \binom{5}{0} = 1\,,$$

i.e. the same results as we have got with the Pascal's triangle.

Now we shall prove the formula using induction. First of all, we check that it works for $n = 1$. Indeed, since $\binom{1}{0} = \binom{1}{1} = 1$, we immediately see that $(a+b)^1 = 1 \cdot a^1 b^0 + 1 \cdot a^0 b^1 = a + b$, as required. Next, we *assume* that the formula (1.71) is valid for a general n, and we then consider the expansion for the next value of $(n + 1)$:

$$(a+b)^{n+1} = (a+b)(a+b)^n = (a+b)\sum_{i=0}^{n}\binom{n}{i} a^i b^{n-i}$$
$$= \sum_{i=0}^{n}\binom{n}{i} a^{i+1} b^{n-i} + \sum_{i=0}^{n}\binom{n}{i} a^i b^{n+1-i}\,.$$

In the first sum we separate out the term with $i = n$, while in the second the term with $i = 0$:

$$(a+b)^{n+1} = \binom{n}{n} a^{n+1} b^0 + \sum_{i=0}^{n-1}\binom{n}{i} a^{i+1} b^{n-i} + \sum_{i=1}^{n}\binom{n}{i} a^i b^{n+1-i} + \binom{n}{0} a^0 b^{n+1}$$

Note that now the summation index i ends at $(n - 1)$ in the first sum and starts from $i = 1$ in the second. Also note that, according to the properties (1.74), the first and the last binomial coefficients in the last expression are both equal to one. We shall now change the summation index in the first sum to make it look similar to the second: instead of i running between 0 and $(n-1)$ we introduce $k = i+1$ changing between 1 and n:

$$\sum_{i=0}^{n-1}\binom{n}{i} a^{i+1} b^{n-i} = \sum_{k=1}^{n}\binom{n}{k-1} a^k b^{n-k+1} = \sum_{i=1}^{n}\binom{n}{i-1} a^i b^{n-i+1}\,,$$

where in the last passage we replaced k back with i for convenience.[11] Now the two sums look really similar and we can combine all our manipulations into:

$$(a+b)^{n+1} = a^{n+1}b^0 + \sum_{i=1}^{n} \left[\binom{n}{i-1} + \binom{n}{i} \right] a^i b^{n-i+1} + a^0 b^{n+1} \,. \tag{1.77}$$

Since $\binom{n+1}{0} = \binom{n+1}{n+1} = 1$ according to (1.74) and the sum of two binomial coefficients in the square brackets is $\binom{n+1}{i}$ because of (1.75), we see that we managed to manipulate $(a+b)^{n+1}$ in (1.77) exactly into the right-hand side of (1.71) with $(n+1)$ instead of n. This completes the proof by induction of the binomial formula.

Example 1.3. ► Calculate the following sum of the binomial coefficients:

$$\sum_{i=0}^{n} \binom{n}{i} = 1 + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n-1} + 1 \,.$$

Solution. consider the binomial expansion of order n for $a = b = 1$:

$$(1+1)^n = 2^n = 1 + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n-1} + 1 \,, \tag{1.78}$$

which shows that the sum is equal to 2^n.◄

Problem 1.47. *Prove the formulae:*

$$\binom{n}{2} + \binom{n}{4} + \binom{n}{6} + \cdots + \binom{n}{n-1} = 2^{n-1} - 1 \,, \quad \textit{if } n = \textit{odd}, \tag{1.79}$$

$$\binom{n}{2} + \binom{n}{4} + \binom{n}{6} + \cdots + \binom{n}{n-2} = 2^{n-1} - 2 \,, \quad \textit{if } n = \textit{even}. \tag{1.80}$$

[*Hint*: consider the expansion for $(1+1)^{n-1}$ and make use of Eq. (1.75).]

[11] After all, the summation index is a dummy index which is only used to indicate how the summation is made; any letter can be used for it.

Problem 1.48. *Similarly, prove:*

$$\binom{n}{1}+\binom{n}{3}+\binom{n}{5}+\cdots+\binom{n}{n-2}=2^{n-1}-1\,,\quad \text{if } n=\text{odd}, \tag{1.81}$$

$$\binom{n}{1}+\binom{n}{3}+\binom{n}{5}+\cdots+\binom{n}{n-1}=2^{n-1}\,,\quad \text{if } n=\text{even}. \tag{1.82}$$

Problem 1.49. *Prove that*

$$\sum_{k=0}^{n}\binom{n}{k}^2=\binom{2n}{n}\,. \tag{1.83}$$

[*Hint*: expand $(1+x)^n(1+x)^n$ in two different ways and compare the coefficients by x^n. Note the symmetry property (1.76).

Many other identities involving the binomial coefficients can be established in a similar fashion by giving a and b special values.

1.11 Combinatorics and Multinomial Theorem

The binomial coefficients we derived in the previous section have another significance related to the number of *combinations* in which different objects can be arranged with respect to each other when placed on a single line. This may seem only to deserve some attention of a mathematician as a purely abstract problem, which is entirely irrelevant to a physicist or engineer. Apparently, this is not the case: combinatorics, known as the science of counting different combinations of objects, has important applications in many sciences and hence deserves our full attention here.

There are several questions one may ask, the simplest of them being as follows: in how many combinations can one place n different objects? For instance, consider different numbers: there will be exactly two combinations, (ab) and (ba), in arranging two different numbers a and b; if there are three numbers, a, b and c, however, the number of combinations is six, namely: (abc), (acb), (bac), (bca), (cab) and (cba). It is easy to relate these combinations of three objects to the combinations of the two: in order to obtain all combinations of the three objects, we fix one object at the first "place" and perform all combinations of the other two objects; we know there are exactly two such combinations possible. For instance, choosing c at the first position, a and b are left giving us two possibilities: $c\,(ab)$

and c (ba). Since there are three different objects, then each of them is to be placed in the first position, and each time we shall have only two combinations, which brings the total number of them to $3 \times 2 = 6$.

Problem 1.50. *Use the method of induction to prove that n distinct objects can be arranged exactly in $n!$ different combinations.*

What if now some of the objects are the same? For instance, consider 2 white balls and 3 black ones, and let us try to calculate the total number of combinations in which we can arrange them. If the balls were all different, then there would have been $5! = 120$ such combinations. However, in our case the number of combinations must be much smaller since we have to remove all combinations (or arrangements) in which we permute balls of the same colour as these would obviously not bring any new arrangements of black and white balls. Within the $5! = 120$ combinations, for *any* given arrangement of white balls in 5 places, there will exactly be $3! = 6$ equivalent combinations of the black ones permuted between themselves; similarly, for any given arrangement of 3 black balls in 5 positions, there will be $2! = 2$ identical combinations associated with the permutation of the two white balls between themselves. Hence, this simple argument gives us $5!/(2! \cdot 3!) = 120/(2 \cdot 6) = 10$ different combinations altogether, all shown in Fig. 1.16. The same argument can be extended to n black and m white balls, in which case the total number of combinations will be $(n+m)!/(n! \cdot m!)$. But this is nothing but the binomial coefficient $\binom{n+m}{n} = \binom{n+m}{m}$, check out Eq. (1.72).

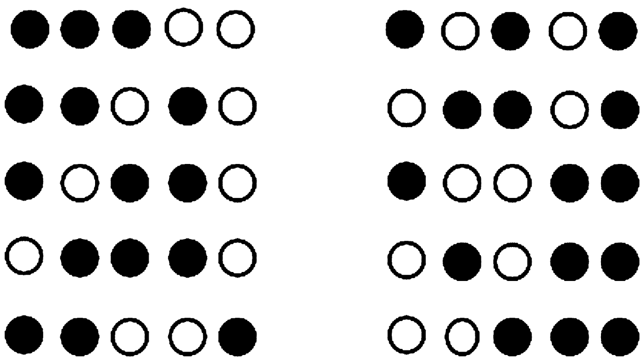


Fig. 1.16 There are ten combinations in arranging three black and two white balls

This is not accidental: if we look closely at the formula (1.71), each term in the sum contains i-th power of a and $(n-i)$-th power of b; if we multiply i numbers a and $n-i$ numbers b, like

$$\underbrace{aaa\ldots}_{i}\underbrace{bbbb\ldots}_{n-i} ,$$

then obviously the result does not depend on the order in which we multiply the numbers, it will be still $a^i b^{n-i}$. However, the total number of combinations in which we arrange the numbers in the product is equal to the total number of arrangements of i white and $n-i$ black balls, i.e. it is $\begin{pmatrix} n \\ i \end{pmatrix}$, the binomial coefficient in Eq. (1.71).

Finally, we are now ready to generalise our result to a problem of arranging m balls, when there are k_1 balls of one colour, k_2 of another, and so on, n colours altogether.[12] Since the total number of balls is $k_1 + k_2 + \cdots + k_n = m$, the number of different (unique) combinations will be

$$\begin{pmatrix} m \\ k_1, k_2, \ldots, k_n \end{pmatrix} = \frac{(k_1 + k_2 + \cdots + k_n)!}{k_1!k_2!\ldots k_n!} = \frac{\left(\sum_{i=1}^{n} k_i\right)!}{\prod_{i=1}^{n} k_i!} , \tag{1.84}$$

where the symbol $\prod$ means "a product of", i.e. in our case above this is simply the product of $k_1!$, $k_2!$, etc. The bracket notation is not accidental as will be clear in a moment; also, in the case of $n = 2$ the expression above is exactly the binomial coefficient in (1.73). This equation is readily understood: the total number of permutations of all m balls is $m!$. However, this number is to be reduced by the numbers $k_1!$, $k_2!$, and so on of permutations of all identically coloured balls, so we divide $m!$ by the product of $k_1!$, $k_2!$, etc.

If we now multiply the sum of n numbers by itself m times,

$$\underbrace{(a_1 + a_2 + \cdots + a_n)(a_1 + a_2 + \cdots + a_n)\cdots(a_1 + a_2 + \cdots + a_n)}_{m \quad \text{times}}$$
$$= (a_1 + a_2 + \cdots + a_n)^m ,$$

then upon opening the brackets and multiplying all terms, we shall have all possible combinations of the products of all individual numbers; there will be exactly m numbers in each such term, and some of the numbers may be the same or appear in different order.

[12] Of course, many other interpretations of the same result exist. For instance, in how many ways one can distribute m bins in n boxes so that there will be k_1 bins in the first box, k_2 in the second, and so on, $k_1 + k_2 + \cdots + k_n = m$ altogether.

Problem 1.51. *As an example, consider* $(a_1 + a_2 + a_3)^3$. *First, expand the cube without combining identical terms* $a_i a_j a_k$ *(where* i, j, k *are* 1, 2 *or* 3*); verify that there will be all possible combinations of the indices* (i, j, k). *Then, combine identical terms having the same powers for* a_1, a_2 *and* a_3. *Check that the number of identical terms (and hence the numerical pre-factor to each term) for the given indices* (i, j, k) *corresponds to the number of ways the indices can be permuted giving different sequences. Verify that this number is given by Eq. (1.84).*

Returning to the general case, we can combine together all combinations which contain a_1 exactly k_1 times, a_2 exactly k_2 times, and so on, as they give an identical contribution of $a_1^{k_1} a_2^{k_2} \dots a_n^{k_n}$. Of course, $k_1 + k_2 + \dots + k_n = m$ as there are exactly m numbers in each elementary term. How many such identical terms there will be? The answer is given by the same Eq. (1.84)! Therefore, combining all identical terms, we arrive at an expansion in which a numerical pre-factor (1.84) is assigned to each different term, i.e. we finally obtain:

$$(a_1 + a_2 + \dots + a_n)^m = \sum_{k_1 + k_2 + \dots + k_n = m} \binom{m}{k_1, k_2, \dots, k_n} a_1^{k_1} a_2^{k_2} \dots a_n^{k_n} , \qquad (1.85)$$

which is called the *multinomial expansion*. The sum here is run over all possible indices k_1, k_2, etc. between 0 and m in such a way that their sum is equal to m.

Problem 1.52. *Prove Eq. (1.85) using the method of induction.* [*Hint*: when considering the induction hypothesis for the case of $n + 1$, combine the last two numbers in the expression $(a_1 + a_2 + \dots + a_n + a_{n+1})^m$ into a single number $b_n = a_n + a_{n+1}$, apply the hypothesis and use the binomial theorem (1.73).]

Problem 1.53. *Consider* $2n$ *objects. Show that the number of different pairs of objects which can be selected out of all objects is*

$$\mathcal{N}_{pairs} = \frac{(2n)!}{2^n n!} . \qquad (1.86)$$

1.12 Some Important Inequalities

There is one famous result in mathematics known as Cauchy–Bunyakovsky–Schwarz inequality. It states that for any real set of numbers x_i and y_i (where $i = 1, \dots, n$)

$$\left(\sum_{i=1}^{n} x_i y_i\right)^2 \leq \left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right) \quad \text{or} \quad \sum_{i=1}^{n} x_i y_i \leq \sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}\,. \tag{1.87}$$

To prove this, consider a function of some real variable t defined in the following way:

$$f(t) = \sum_{i=1}^{n} (tx_i + y_i)^2 = t^2 \underbrace{\left(\sum_{i=1}^{n} x_i^2\right)}_{a} + 2t \underbrace{\left(\sum_{i=1}^{n} x_i y_i\right)}_{b} + \underbrace{\left(\sum_{i=1}^{n} y_i^2\right)}_{c} \geq 0\,.$$

This is a parabola with respect to t; it is non-negative, and hence should have either one or no roots. This means that the determinant of this square polynomial $\mathcal{D} = 4b^2 - 4ac$ is either equal to zero (one root: the parabola just touches the t axis at a single point) or is negative (no roots: the parabola is lifted above the t axis), therefore

$$\mathcal{D} = 4\left(b^2 - ac\right) = 4\left[\left(\sum_{i=1}^{n} x_i y_i\right)^2 - \left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right)\right] \leq 0\,,$$

which is exactly what we wanted to prove.

Using this rather famous inequality, we shall now prove another one related to lengths of vectors. Indeed, consider two vectors $\mathbf{x}=(x_{1,}, x_2, x_3)$ and $\mathbf{y}=(y_1, y_2, y_3)$. Then, the length of their sum will be shown to be less than or equal to the sum of their individual lengths:

$$|\mathbf{x} + \mathbf{y}| \leq |\mathbf{x}| + |\mathbf{y}|\,. \tag{1.88}$$

Indeed,

$$|\mathbf{x} + \mathbf{y}| = \sqrt{\sum_{i=1}^{3} (x_i + y_i)^2} = \sqrt{\sum_i x_i^2 + \sum_i y_i^2 + 2\sum_i x_i y_i}\,.$$

Now we can apply the inequality (1.87) for the third term inside the square root, which gives:

$$|\mathbf{x} + \mathbf{y}| = \sqrt{\sum_i x_i^2 + \sum_i y_i^2 + 2\sum_i x_i y_i} \leq \sqrt{\sum_i x_i^2 + \sum_i y_i^2 + 2\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$$

$$= \sqrt{\left(\sqrt{\sum_i x_i^2} + \sqrt{\sum_i y_i^2}\right)^2} = \sqrt{\sum_i x_i^2} + \sqrt{\sum_i y_i^2} = |\mathbf{x}| + |\mathbf{y}|\,,$$

which is the desired result.[13]

[13]The same inequality is of course valid for vectors in a space of any dimension as we did not really use the fact that our vectors only have three components.

Problem 1.54. *Prove similarly that for any two complex numbers z_1 and z_2, we have:*

$$|z_1 + z_2| \leq |z_1| + |z_2| \ . \tag{1.89}$$

Problem 1.55. *Using induction, generalise this result for a sum of n complex numbers:*

$$\left|\sum_i z_i\right| \leq \sum_i |z_i| \ . \tag{1.90}$$

Problem 1.56. *Prove that for any two complex numbers z_1 and z_2:*

$$|z_1 - z_2| \geq |z_1| - |z_2| \ . \tag{1.91}$$

[*Hint*: make a substitution in Eq. (1.89).]

Problem 1.57. *Prove that one side of a triangle is always shorter than the sum of two other sides.*

Of course, the same inequalities are valid for real numbers as well, as these are particular cases of their complex counterparts with zero imaginary parts:

$$\left|\sum_{i=1}^{n} x_i\right| \leq \sum_{i=1}^{n} |x_i| \ , \tag{1.92}$$

i.e. the absolute value of a sum of real numbers is always less than or equal to the sum of their absolute values.

However, it is instructive to prove the above inequality independently as well. Let us first consider the case of $n = 2$. Then, we can write:

$$|x_1 + x_2| = \sqrt{(x_1 + x_2)^2} = \sqrt{x_1^2 + x_2^2 + 2x_1x_2} = \sqrt{(|x_1| + |x_2|)^2 - 2|x_1x_2| + 2x_1x_2}$$
$$= \sqrt{(|x_1| + |x_2|)^2 + 2(x_1x_2 - |x_1x_2|)} \ .$$

Two cases are possible: if $x_1x_2 \geq 0$, then $x_1x_2 - |x_1x_2| = x_1x_2 - x_1x_2 = 0$; if, however, $x_1x_2 < 0$, then $x_1x_2 - |x_1x_2| = x_1x_2 + x_1x_2 = 2x_1x_2 < 0$. Hence, $2(x_1x_2 - |x_1x_2|) \leq 0$, and, therefore,

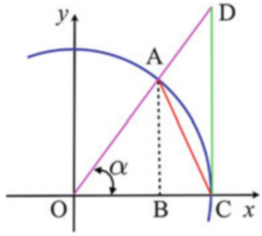


Fig. 1.17 To the proof of inequality (1.93): the area of the △OAC is less then the area of the sector OAC which in turn is less than the area of the △ODC

$$|x_1 + x_2| = \sqrt{(|x_1| + |x_2|)^2 + 2(x_1x_2 - |x_1x_2|)} \leq \sqrt{(|x_1| + |x_2|)^2} = |x_1| + |x_2| \ ,$$

as required. Now we can consider the case of three numbers, for which we can use the inequality for two numbers proven above:

$$|x_1 + x_2 + x_3| = |x_1 + (x_2 + x_3)| \leq |x_1| + |x_2 + x_3| \leq |x_1| + |x_2| + |x_3| \ ,$$

and so on.

Finally, let us consider another inequality which we shall also need. It is used in various estimates and proofs:

$$\sin\alpha < \alpha < \tan\alpha \ . \tag{1.93}$$

To show that, consider Fig. 1.17 where we have a circle of unit radius (blue), two triangles, △OAC and △ODC, as well as a sector OAC. Obviously, the area of the △OAC is strictly smaller than the area of the sector, which in turn is smaller than the area of the △ODC. Both triangles have the same unit base $OC = 1$, the height of △OAC is $AB = OA\sin\alpha = \sin\alpha$, while the height of the △ODC is $DC = OC\tan\alpha = \tan\alpha$. The areas of the two triangles are easy to calculate as $S_{\triangle OAC} = \frac{1}{2}OC \cdot AB = \frac{1}{2}\sin\alpha$ and $S_{\triangle ODC} = \frac{1}{2}OC \cdot DC = \frac{1}{2}\tan\alpha$. The area of the sector OAC of radius $R = 1$ can be calculated as a fraction $\alpha/2\pi$ of the area $\pi R^2 = \pi$ of the whole circle[14]: $\left(\frac{\alpha}{2\pi}\right)\pi R^2 = \frac{1}{2}\alpha$. Comparing the three and cancelling out on the factor of $1/2$, we arrive at (1.93).

Problem 1.58. *Prove the same result (1.93) by considering the lengths of AB and DC as well as the circle arc length AC.*[15]

[14]The area of a circle of radius R is πR^2, this will be established later on in Sect. 4.6.2.

[15]The length of the circle circumference is $2\pi R$, where R is the circle radius (see Sect. 4.6.1).

1.13 Lines and Planes

We shall be often dealing with lines and surfaces in the course, and it is convenient at this stage to summarise a number of useful results here. We shall start from lines, then turn to surfaces, after which we shall consider various geometrical problems related to both of them.

1.13.1 Straight Line

A straight line in the 3D space can be specified by any two points $A(x_A, y_A, z_A)$ and $B(x_B, y_B, z_B)$ on the line, or alternatively, by vectors $\mathbf{r}_A = (x_A, y_A, z_A)$ and $\mathbf{r}_B = (x_B, y_B, z_B)$. If the vector $\mathbf{r} = (x, y, z)$ corresponds to an arbitrary point P on the line, see Fig. 1.18(a), then we can construct a vector $\mathbf{a} = \overrightarrow{AB} = \mathbf{r}_B - \mathbf{r}_A$ along the line, so that any point P on the line characterised by the vector $\mathbf{r}$ can be obtained by adding to $\mathbf{r}_A$ the scaled vector $\mathbf{a}$:

$$\mathbf{r} = \mathbf{r}_A + t\mathbf{a} = \mathbf{r}_A + t\left(\mathbf{r}_B - \mathbf{r}_A\right) \ , \tag{1.94}$$

where $-\infty < t < \infty$ is a scalar parameter. This is a parametric equation of the line, specified either via two points A and B or via a vector $\mathbf{a}$ (it is frequently convenient for it to be a unit vector) and a point A.

Problem 1.59. *The vector equation (1.94) is equivalent to three scalar ones for the three Cartesian components of* $\mathbf{r}$. *Show that the equation of a line can also alternatively be written as follows:*

$$\frac{x - x_A}{x_B - x_A} = \frac{y - y_A}{y_B - y_A} = \frac{z - z_A}{z_B - z_A} \ . \tag{1.95}$$

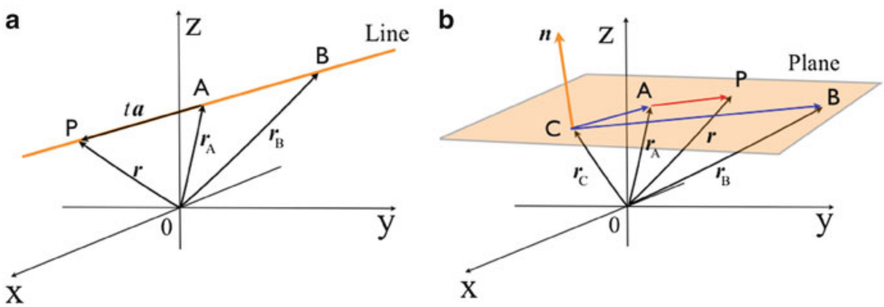


Fig. 1.18 An illustration to the derivation of the equations for a line (**a**) and a plane (**b**)

Problem 1.60. *Show from the result above that in the $x-y$ plane ($z = 0$, a 2D space), the equation of a line can be formally written as*

$$Ax + By = C \ . \tag{1.96}$$

Note that this form is more general than the familiar $y = kx + c$ (with k and c being constants) since it contains also lines parallel to the y axis ($Ax = C$) which are not contained in the other form.

Two lines are parallel (collinear) if the vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ in the corresponding parametric equations are parallel ($\mathbf{a}_1 \parallel \mathbf{a}_2$) , i.e. if $\mathbf{a}_1 \times \mathbf{a}_2 = 0$; two lines are perpendicular if these vectors are orthogonal, $\mathbf{a}_1 \cdot \mathbf{a}_2 = 0$ (or $\mathbf{a}_1 \perp \mathbf{a}_2$). Note, however, that in the second case the lines may not necessarily cross. A simple conditions for the two lines to cross (even if they are not perpendicular) is considered later on in Sect. 1.13.5.

1.13.2 Polar and Spherical Coordinates

In 2D space *polar coordinates* are frequently used instead of the Cartesian ones. In polar coordinates the length $r = |\mathbf{r}|$ of the vector $\mathbf{r}$ and its angle ϕ with the x axis ($0 \leq \phi \leq 2\pi$) are used instead of the Cartesian coordinates (x, y) as shown in Fig. 1.19(a). As follows from the figure, the relationship between the two types of coordinates is provided by the following *connection relations*:

$$x = r\cos\phi \ , \quad y = r\sin\phi \ . \tag{1.97}$$

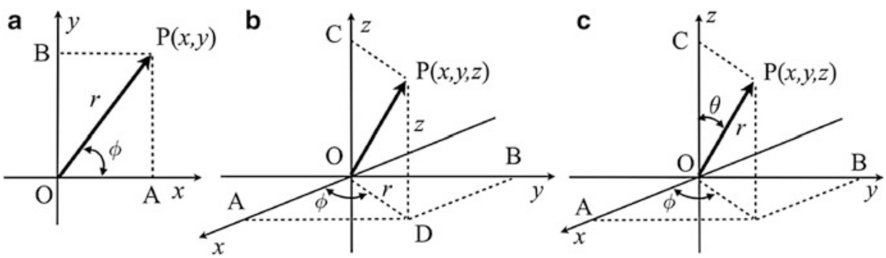


Fig. 1.19 Definitions of polar (**a**), cylindrical (**b**) and spherical (**c**) coordinate systems. The components of the vector $\mathbf{r}$ are its projections $x = OA$, $y = OB$ and $z = OC$ on the Cartesian axes

In 3D space polar coordinates are generalised to *cylindrical* coordinates by adding the z coordinate as shown in Fig. 1.19(b); in this case, (x, y, z) are replaced by (r, ϕ, z) as follows:

$$x = r\cos\phi\ , \quad y = r\sin\phi\ , \quad z = z\ . \tag{1.98}$$

Note that r here is the length of the projection OD of the vector $\mathbf{r}$ on the $x-y$ plane.

Finally, we shall also introduce another very useful coordinate system—*spherical*. It is defined in Fig. 1.19(c). Here the coordinates used are (r, θ, ϕ), where this time r is the length OP of the vector $\mathbf{r}$, θ is the angle it makes with the z axis ($0 \leq \theta \leq \pi$) and ϕ is the same angle as in the two previous systems ($0 \leq \phi \leq 2\pi$). The connection relations in this case are:

$$x = r\sin\theta\cos\phi, \quad y = r\sin\theta\sin\phi, \quad z = r\cos\theta, \tag{1.99}$$

since $CP = r\sin\theta$ and hence $x = OA = CP\cos\phi$, $y = OB = CP\sin\phi$ and $z = OC = r\cos\theta$.

The special coordinate systems introduced above are frequently used in the calculus to simplify calculations, and we shall come across many such examples throughout the course. Polar coordinates are useful in 2D calculations related to objects having circular symmetry; cylindrical coordinates—for 3D objects with cylindrical symmetry (with the axis of the cylinder along the z axis), while the spherical coordinate system is convenient for the 3D calculations with objects having spherical symmetry. Many other so-called *curvilinear coordinates* also exist.

1.13.3 Curved Lines

Here we shall briefly mention an application of polar coordinates for specifying curved lines in the 2D space. The idea is based on writing r as a function of ϕ, e.g. $r = r(\phi)$. It is essential to understand that this is simply a prescription for writing the Cartesian coordinates,

$$x(\phi) = r(\phi)\cos\phi \quad \text{and} \quad y(\phi) = r(\phi)\sin\phi\ , \tag{1.100}$$

as parametric equations of curved lines via the angle ϕ serving as a single parameter. Because of that, $r = r(\phi)$ may be even negative, i.e. only its absolute value corresponds to the distance of the (x, y) point to the centre O of the coordinate system.

As an example, consider a circle of radius R centred at the origin. Its equation is given simply by $r = R$, i.e. r does not depend on the angle ϕ. In that case from (1.100) we get (by taking the square of both sides of the two equations and summing them up) that the equation of the circle is the familiar $x^2 + y^2 = R^2$.

Problem 1.61. *Show that the curve specified via* $r = 2R\sin\phi$ *corresponds to a circle of radius* R *centred at* $(0, R)$. [*Hint*: solve for ϕ from one of the equations for x or y, and derive an equation $f(x, y) = 0$ for the curve.]

Problem 1.62. *Various famous curves are shown in Fig. 1.20. Convince yourself that these geometrical figures correspond to the equations given. In the case of the three-petal rose (a) verify the range of the angle* ϕ *necessary to plot a single petal.*[16]

Problem 1.63. *Consider a family of curves specified by* $r = a\cos(k\phi)$ *and sketch them. How does the number of petals depend on the value of* k*?*

1.13.4 Planes

A plane can be specified by either three points A, B, C or by a normal vector $\mathbf{n}$ and a single point A, see Fig. 1.18(b). The two ways are closely related as the normal can be constructed from the three points by performing a vector product $\mathbf{n} = \overrightarrow{CB} \times \overrightarrow{CA} = (\mathbf{r}_B - \mathbf{r}_C) \times (\mathbf{r}_A - \mathbf{r}_C)$. So, it is sufficient to consider the specification of the plane via a normal and a point, say the point A. If we take an arbitrary point P on the plane corresponding to the vector $\mathbf{r}$, then the vector $\overrightarrow{AP} = \mathbf{r} - \mathbf{r}_A$ lies within the plane, and hence is orthogonal to the normal:

$$\mathbf{n} \cdot (\mathbf{r} - \mathbf{r}_A) = 0 \quad \Longrightarrow \quad \mathbf{r} \cdot \mathbf{n} = \mathbf{r}_A \cdot \mathbf{n} \ , \tag{1.101}$$

which is the desired equation of the plane. One can see the point $\mathbf{r}_A$ lies on the plane (as $\mathbf{r} = \mathbf{r}_A$ also satisfies the equation), and that the normal does not need to be a unit vector. In addition, either direction of the normal can be taken, up or down.

Another form of the equation for the plane follows immediately from (1.101) if we write explicitly the dot product in the left-hand side and notice that the right-hand side is simply a number:

$$ax + by + cz = h \ . \tag{1.102}$$

It follows from this discussion that the coefficients (a, b, c) in this equation form components of the normal to the surface. Note that if $h \neq 0$, one may divide both sides of the equation by it, resulting in a simpler form $a'x + b'y + c'z = 1$, however, for some surfaces $h = 0$ (e.g. the $x - y$ plane has the equation $z = 0$), so that the form above is the most general. Let us mention some particular cases which might

[16] This and other related curves were studied by an Italian mathematician Guido Grandi.

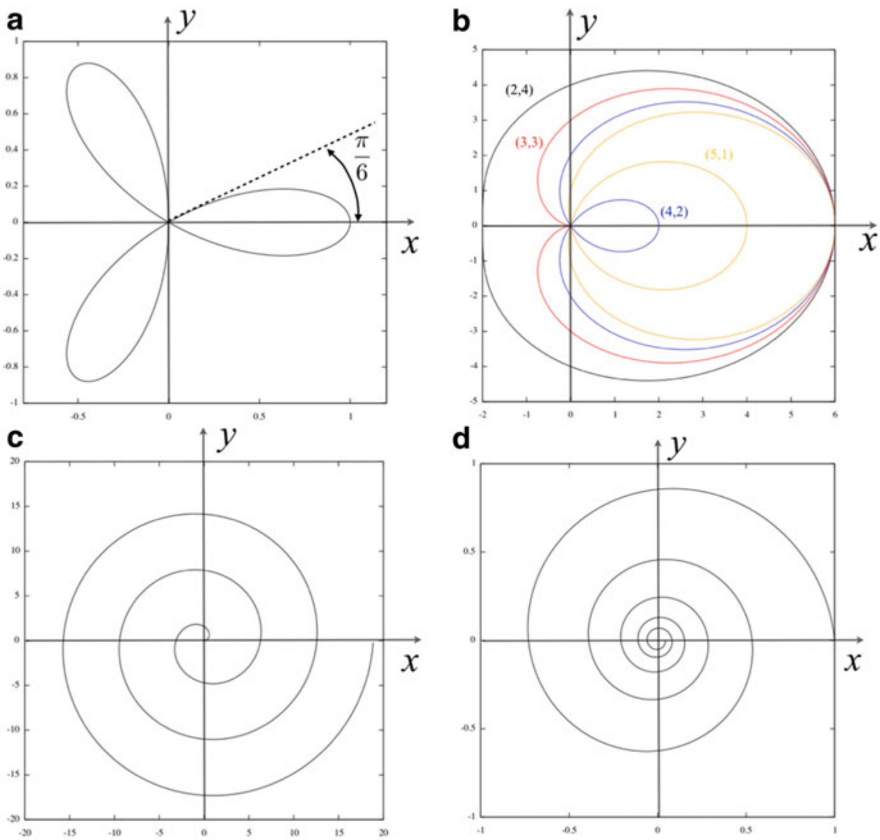


Fig. 1.20 Examples of various famous curves specified in polar coordinates: (**a**) a three-leaf rose (or rhodonea curve) specified as $r = a\cos(3\phi)$ (plotted for $a = 1$); (**b**) snail curves (limacons of Pascal) specified via $r = b + a\cos\phi$ and shown for various pairs of (a, b); note how the curves continuously change their shape when going from $a < b$ to $a > b$ across $a = b$; (**c**) Archimedean spiral $r = a\phi$ (plotted for $a = 1$); (**d**) spiral $r = ae^{-\alpha\phi}$ (plotted for $a = 1$ and $\alpha = 0.1$). In the latter two cases only a finite number of revolutions are depicted

be of interest: (i) $h = 0$ corresponds to a plane passing through the centre of the coordinate system; (ii) $a = 0$ corresponds to a plane perpendicular to the x axis; (iii) $a = b = 0$ corresponds to a plane $cz = h$ which is parallel to the $x - y$ plane.

We shall now derive a parametric equation for a plane. In Eq. (1.94) for a line we only had one parameter t; this is because a line is equivalent to a 1D space. In the case of a plane there must be two such parameters as the plane is equivalent to a 2D space. One may say that a line has one "degree of freedom", while a plane has two. Now, consider two vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ lying in the plane (it is usually convenient to choose them of unit length), and the vector $\mathbf{r}_A$ pointing to a point A on the plane.

Then, evidently, any point P on the plane would have the vector

$$\mathbf{r} = \mathbf{r}_A + t_1\mathbf{a}_1 + t_2\mathbf{a}_2 \tag{1.103}$$

pointing to it; here, t_1 and t_2 are the two parameters, each of which may take any values from $\mathbb{R}$.

Two planes are parallel if their normal vectors $\mathbf{n}_1$ and $\mathbf{n}_2$ are parallel, i.e. if $\mathbf{n}_1 \times \mathbf{n}_2 = 0$; two planes are perpendicular to each other if their normals are orthogonal, i.e. if $\mathbf{n}_1 \cdot \mathbf{n}_2 = 0$.

Unless two planes are parallel, they intersect at a line. Let us derive the equation of that line. This is most easily accomplished by noticing that a vector $\mathbf{a}$ along the line in Eq. (1.94) is perpendicular to both normal vectors of the planes $\mathbf{n}_1$ and $\mathbf{n}_2$, i.e. we can choose a vector $\mathbf{a}$ along the line as $\mathbf{a} = \mathbf{n}_1 \times \mathbf{n}_2$. The only other unknown to find is a point A with vector $\mathbf{r}_A$ lying on that crossing, as after that the parametric equation of the line is completely determined. To find a point on the line, we should solve for a point $\mathbf{r}$ that simultaneously satisfies equations for both planes taken in either of the forms discussed above; however, the simplest are the first two, i.e. either (1.101) or (1.102). This step will result in two algebraic equations for three unknowns x, y and z, so that one can only express any two of them via the third; we choose that third coordinate arbitrarily (our particular choice will result in a certain position of the point A along the line) and then would be able to determine the other two. Alternatively, the chosen coordinate may serve as a parameter t for the line equation.

Example 1.4. ► Determine the parametric equation of the line at which two surfaces $x + y + z = 1$ and $x = 0$ cross.

Solution. the normal of the first surface $\mathbf{n}_1 = (1, 1, 1)$, while $\mathbf{n}_2 = (1, 0, 0)$ for the second. Therefore, the vector of the line

$$\mathbf{a} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 1 & 1 & 1 \\ 1 & 0 & 0 \end{vmatrix} = \mathbf{j} - \mathbf{k} = (0, 1, -1) .$$

Considering the two equations simultaneously, we get $y + z = 1$. Taking $y = 0$, we get $z = 1$ and hence we now have the point $(0, 0, 1)$ on the crossing line. The parametric equation, written using vector-columns, is thus

$$\mathbf{r} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + t \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ t \\ 1 - t \end{pmatrix} .$$

It is easy to see that any point on this line (i.e. for any value of the t) satisfies the equations for both planes simultaneously.

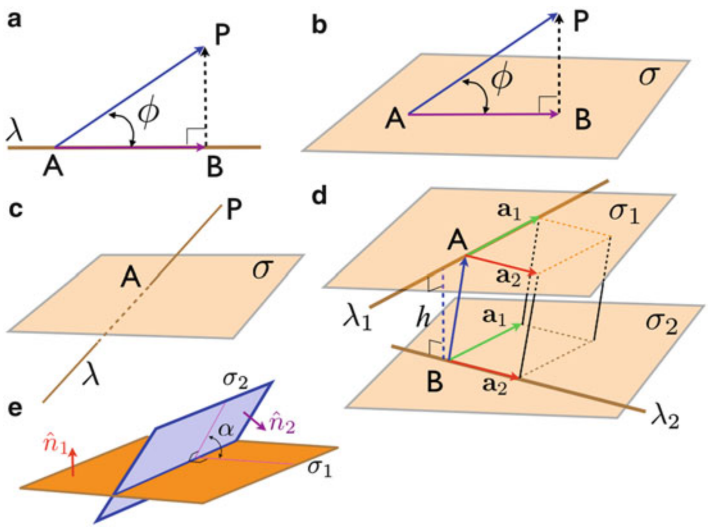


Fig. 1.21 Typical problems related to lines and planes. (**a**) To the calculation of the distance PB between the point P and the line λ; (**b**) to the calculation of the distance PB between the point P and the plane σ; (**c**) finding the points A of intersection of the line λ with the plane σ; (**d**) finding the distance h between two non-parallel and non-crossing lines λ_1 and λ_2; (**e**) finding angle α between two planes

Alternatively, consider y as the parameter t. Then, we have $x = 0$, $y = t$ and $z = 1 - y = 1 - t$. We see that exactly the same result is obtained.◀

Problem 1.64. *Prove that this method will always give the correct solution, i.e. points on the line constructed in this manner will satisfy Eq. (1.101) for both planes.*

Problem 1.65. *Determine the equation of the line at which two planes* $2x - y + z = 2$ *and* $-x + y + z = 1$ *cross.* [*Answer*: $\mathbf{r} = (3, 4, 0) + t(-2, -3, 1)$.]

1.13.5 Typical Problems for Lines and Planes

Let us now consider a number of typical problems which one comes across in applications related to lines and planes.

Distance from a Point to a Line or Plane. Let us take a point P specified by the vector $\mathbf{r}_P$ and a line $\mathbf{r} = \mathbf{r}_A + t\mathbf{a}$ specified via the point A and the vector $\mathbf{a}$. What we would like to calculate is the distance between them; this will be the length of the perpendicular drawn from the point P to the line as shown in Fig. 1.21(a). Here

$\overrightarrow{AB} \parallel \mathbf{a}$, and hence the angle ϕ can be obtained from the dot product of $\mathbf{a}$ and $\overrightarrow{AP}$, i.e. $\overrightarrow{AP} \cdot \mathbf{a} = |\mathbf{a}| \left|\overrightarrow{AP}\right| \cos\phi$. Indeed, by calculating the dot product using components of the vectors $\mathbf{a}$ and $\overrightarrow{AP}$, the $\cos\phi$ is obtained, and hence the angle ϕ. Then, the required distance $BP = \left|\overrightarrow{AP}\right| \sin\phi$.

Specifically, let us consider the 2D case where the line is given by the equation $y = kx + b$. Let point A be $(0, b)$ (i.e. it is the point at which the line crosses the y axis). The point P has coordinates (x_P, y_P). Then, by considering two points $A = (0, b)$ and $(x_P, kx_P + b)$ on the line, we can calculate its direction vector as

$$\mathbf{a} = (x_P - 0)\,\mathbf{i} + [(kx_P + b) - b]\,\mathbf{j} = x_P\mathbf{i} + x_Pk\mathbf{j} \ .$$

The vector $\overrightarrow{AP} = (x_P, y_P - b)$. Therefore,

$$\cos\phi = \frac{\overrightarrow{AP} \cdot \mathbf{a}}{\left|\overrightarrow{AP}\right| |\mathbf{a}|} = \frac{x_P^2 + kx_P\,(y_P - b)}{\sqrt{x_P^2 + (y_P - b)^2}\,x_P\sqrt{1 + k^2}} = \frac{x_P + k\,(y_P - b)}{\sqrt{x_P^2 + (y_P - b)^2}\sqrt{1 + k^2}}$$

and hence

$$\begin{aligned}\sin^2\phi = 1 - \cos^2\phi &= 1 - \frac{x_P^2 + 2kx_P\,(y_P - b) + k^2\,(y_P - b)^2}{\left[x_P^2 + (y_P - b)^2\right](1 + k^2)} \\ &= \frac{(kx_P - y_P + b)^2}{\left[x_P^2 + (y_P - b)^2\right](1 + k^2)} = \frac{(kx_P - y_P + b)^2}{(1 + k^2)\left|\overrightarrow{AP}\right|^2} \ ,\end{aligned}$$

so that the required distance is

$$d = \left|\overrightarrow{AP}\right| \sin\phi = \frac{|kx_P - y_P + b|}{\sqrt{1 + k^2}} \ . \tag{1.104}$$

Problem 1.66. *Show that the distance between a point $P(x_P, y_P, z_P)$ and a plane $ax + by + cz = h$ is (see Fig. 1.21(b)):*

$$d = BP = \mathbf{n} \cdot \overrightarrow{AP} \ , \tag{1.105}$$

where $\mathbf{n}$ is the unit normal to the plane. Here A is an arbitrary point on the plane, the result does not depend on its choice. Indeed, manipulate this expression into

$$d = \frac{ax_P + by_P + cz_P - h}{\sqrt{a^2 + b^2 + c^2}} \ . \tag{1.106}$$

Crossing of Two Lines. The necessary conditions for crossing of two lines can be established as follows. Suppose, we are given two lines $\mathbf{r} = \mathbf{r}_1 + t_1\mathbf{a}_1$ and $\mathbf{r} = \mathbf{r}_2 + t_2\mathbf{a}_2$, where t_1 and t_2 are two parameters. The lines will cross if and only if there exists a plane to which they both belong. This means that the vectors $\mathbf{a}_1$, $\mathbf{a}_2$ and $\mathbf{r}_1 - \mathbf{r}_2$ should lie in this plane. This in turn means that the vector $\mathbf{r}_1 - \mathbf{r}_2$ can be represented as a linear combination $s_1\mathbf{a}_1 + s_2\mathbf{a}_2$ of the vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ with some numbers s_1 and s_2 since any vector within the plane can be represented like this unless the two vectors $\mathbf{a}_1$, $\mathbf{a}_2$ are parallel which we assume is not the case.[17] There is also an algebraic way to show precisely that: since two lines cross, the point $\mathbf{r}$ of crossing will be common to both of them, therefore $\mathbf{r}_1 + t_1\mathbf{a}_1 = \mathbf{r}_2 + t_2\mathbf{a}_2$ and hence $\mathbf{r}_1 - \mathbf{r}_2 = -t_1\mathbf{a}_1 + t_2\mathbf{a}_2$, i.e. the desired linear combination. The simplest way of writing this condition down is by using the triple product:

$$[\mathbf{r}_1 - \mathbf{r}_2, \mathbf{a}_1, \mathbf{a}_2] = \begin{vmatrix} x_1 - x_2 & y_1 - y_2 & z_1 - z_2 \\ (\mathbf{a}_1)_x & (\mathbf{a}_1)_y & (\mathbf{a}_1)_z \\ (\mathbf{a}_2)_x & (\mathbf{a}_2)_y & (\mathbf{a}_2)_z \end{vmatrix} = 0\ , \tag{1.107}$$

i.e. the determinant composed of the components of the three vectors is zero. Indeed, as we know from the properties of the determinants mentioned above, it is equal to zero if any of its rows (or columns) is a linear combination of any other rows (columns). Saying it differently, if the three vectors lie in the same plane, their triple product is zero (see Sect. 1.7), which immediately brings us to the above condition, cf. Eq. (1.38).

Example 1.5. ► A line is given by the equation $\mathbf{r} = (1, 0, 0) + t_1(1, 0, 0)$. There is also a family of lines specified via $\mathbf{r} = (0, 1, 1) + t_2(0, \cos\varphi, \sin\varphi)$; each line in the family corresponds to a particular value of the angle φ. Determine which line in the family (i.e. at which value of the angle φ) crosses the first line, and find the crossing point.

Solution. we have to solve Eq. (1.107) with $\mathbf{r}_1 - \mathbf{r}_2 = (1, -1, -1)$, $\mathbf{a}_1 = (1, 0, 0)$ and $\mathbf{a}_2 = (0, \cos\varphi, \sin\varphi)$; hence, building up the determinant and calculating it, we get:

$$\begin{bmatrix} 1 & -1 & -1 \\ 1 & 0 & 0 \\ 0 & \cos\varphi & \sin\varphi \end{bmatrix} = 1 \cdot (0 \cdot \sin\varphi - 0 \cdot \cos\varphi) - (-1) \cdot (\sin\varphi - 0 \cdot 0) - 1 \cdot (\cos\varphi - 0 \cdot 0)$$

$$= \sin\varphi - \cos\varphi = 0\ ,$$

i.e. $\tan\varphi = 1$ and $\varphi = \pi/4$, so the equation of the second line is $\mathbf{r} = (0, 1, 1) + t_2\left(0, 1/\sqrt{2}, 1/\sqrt{2}\right)$. The point of crossing is then obtained by solving the vector equation

[17] Note that the case of $\mathbf{a}_1 \parallel \mathbf{a}_2$ is not interesting here as in this case the lines do not cross—unless they are the same, of course, which is of even less interest!

$$\mathbf{r} = (1,0,0) + t_1(1,0,0) = (0,1,1) + t_2\left(0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$$

with respect to t_1 and t_2 in components. There are three equations to solve (for x, y and z components) for only two variables; however, because of the condition of crossing (1.107) we just used, one of the equations should be equivalent to the other two, so we just have to consider two independent equations. In our case these are: $1 + t_1 = 0$ for the x component and $0 = 1 + t_2/\sqrt{2}$ for the y (the z component equation, $0 = 1+t_2/\sqrt{2}$, is in this case the same as the second), which give $t_1 = -1$ and $t_2 = -1/\sqrt{2}$, and hence the point of crossing $\mathbf{r} = (1,0,0) + t_1(1,0,0) = (0,0,0)$ is the centre of the coordinate system (of course, the same result is obtained by using the equation of the second line and the value of t_2). This problem and its solution can easily be verified by plotting the two lines in the Cartesian frame.◄

Crossing Point Between a Line and Plane. Consider a line λ, specified by the equation $\mathbf{r} = \mathbf{r}_B + t\mathbf{a}$, which crosses the plane σ given by the equation $\mathbf{n}\cdot\mathbf{r} = \mathbf{n}\cdot\mathbf{r}_C$. The line will definitely cross the plane at some point A, Fig. 1.21(c), if the vector $\mathbf{a}$ is not perpendicular to the line normal: $\mathbf{a}\cdot\mathbf{n} \neq 0$. To determine the crossing point $\mathbf{r}$, we substitute the $\mathbf{r}$ from the line equation into the plane equation to get the value of t, and hence determine the crossing point:

$$t = \frac{\mathbf{n}\cdot(\mathbf{r}_C - \mathbf{r}_B)}{\mathbf{n}\cdot\mathbf{a}}, \text{ giving } \quad \mathbf{r} = \mathbf{r}_B + \frac{\mathbf{n}\cdot(\mathbf{r}_C - \mathbf{r}_B)}{\mathbf{n}\cdot\mathbf{a}}\mathbf{a}, \tag{1.108}$$

where $\mathbf{r}_B$ is a point on the line, while $\mathbf{r}_C$ is a point on the plane.

Distance Between Two Non-parallel and Non-planar Lines. Consider two lines, λ_1 and λ_2, which are not parallel (their vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ are not parallel, $\mathbf{a}_1 \times \mathbf{a}_2 \neq 0$) and do not cross, Fig. 1.21(d). We would like to calculate the distance h between the two lines, i.e. to find a point on one line and also one on the other such that the distance between them is the smallest possible. It is easy to see that the line connecting these two points will be perpendicular to both lines. Indeed, let the equations of the two lines be $\mathbf{r} = \mathbf{r}_A + t_1\mathbf{a}_1$ and $\mathbf{r} = \mathbf{r}_B + t_2\mathbf{a}_2$. Let us first draw the vector $\mathbf{a}_2$ from the point A on the first line, and the vector $\mathbf{a}_1$ from the point B on the second, see Fig. 1.21(d). Two vectors, $\mathbf{a}_1$ and $\mathbf{a}_2$, drawn from the point A form a plane σ_1 containing the line λ_1 with the normal $\mathbf{n}_1 = \mathbf{a}_1 \times \mathbf{a}_2$, whereas the same pair of vectors drawn from the point B form another plane σ_2 containing the other line λ_2 and with the same normal $\mathbf{n}_2 = \mathbf{n}_1$. Since the two normal vectors are collinear, the two planes are parallel. Hence, the two lines lie in parallel planes. The distance between these planes, which is given by the perpendicular line drawn from one plane to the other, is exactly what is needed—the minimal distance between the points of the two lines, shown as a blue dotted line in Fig. 1.21(d).

One way of obtaining the distance h is to calculate the distance between two arbitrary points on the two lines (specified by the parameters t_1 and t_2) and then minimise it with respect to them. For this task to be accomplished, we, however, require a derivative which we have not yet considered. So, we shall use another purely geometrical method. The vectors $\overrightarrow{BA} = \mathbf{r}_A - \mathbf{r}_B$, $\mathbf{a}_1$ and $\mathbf{a}_2$ form a parallelepiped shown in the figure. The volume of it is determined by the absolute value of the triple product of these three vectors. On the other hand, its volume is also given by the area of the parallelogram at the base of the parallelepiped and formed by $\mathbf{a}_1$ and $\mathbf{a}_2$ (which is equal to the absolute value of the vector product $|\mathbf{a}_1 \times \mathbf{a}_2|$), multiplied by its height which is h. Hence,

$$h = \frac{|[\mathbf{r}_B - \mathbf{r}_A, \mathbf{a}_1, \mathbf{a}_2]|}{|\mathbf{a}_1 \times \mathbf{a}_2|} = \frac{|(\mathbf{r}_B - \mathbf{r}_A) \cdot [\mathbf{a}_1 \times \mathbf{a}_2]|}{|\mathbf{a}_1 \times \mathbf{a}_2|} , \tag{1.109}$$

which is the desired result.

Distance Between Two Parallel Planes. We know that two planes are parallel if their normal vectors are collinear. Consider two planes σ_1 and σ_2 with collinear normal vectors $\mathbf{n}_1$ and $\mathbf{n}_2$, given by the equations $\mathbf{n}_1 \cdot \mathbf{r} = \mathbf{n}_1 \cdot \mathbf{r}_A$ and $\mathbf{n}_2 \cdot \mathbf{r} = \mathbf{n}_2 \cdot \mathbf{r}_B$.

Problem 1.67. *Show, using the same method as we already used above for calculating the distance between a point and a line (or a plane), cf. Fig. 1.21(a, b), that the distance between two parallel planes as specified above can be calculated via*

$$h = \frac{(\mathbf{r}_B - \mathbf{r}_A) \cdot \mathbf{n}_1}{|\mathbf{n}_1|} . \tag{1.110}$$

Crossing of Two Planes. Orientation of a plane is determined by its normal vector. Two planes are parallel if their corresponding normal vectors are collinear, i.e. their vector product is equal to zero. Two planes are perpendicular, if the dot product of their normal vectors is equal to zero. In general, one can define an angle α between two planes as shown in Fig. 1.21(e); if $\alpha = 0$, the two planes are parallel, while if $\alpha = \pi/2$, the two planes are perpendicular.

Problem 1.68. *Show that the angle α between two crossing planes, as shown in Fig. 1.21(e), is the same as the angle between their normal vectors. Correspondingly, demonstrate that the angle α between two planes given by equations $a_1x + b_1y + c_1z = d_1$ and $a_2x + b_2y + c_2z = d_2$ is given by:*

$$\cos\alpha = \frac{a_1a_2 + b_1b_2 + c_1c_2}{\sqrt{a_1^2 + b_1^2 + c_1^2}\sqrt{a_2^2 + b_2^2 + c_2^2}} \tag{1.111}$$

(either α or $\pi - \alpha$ can be chosen). In particular, verify the above criteria for the two planes to be parallel or perpendicular to each other.

Chapter 2
Functions

2.1 Definition and Main Types of Functions

As was mentioned already in Sect. 1.5, a function $y = f(x)$ establishes a correspondence between values of x taken from $\mathbb{R}$ and values of y also from $\mathbb{R}$. It is not an essential condition for the function to exist that the values of x and/or y form continuous intervals. For instance, consider a factorial function $y = n! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot \ldots \cdot n$. It makes a correspondence between all integer numbers (serving as the values of x) and some specific set of integer numbers (1, 2, 6, 24, etc.) serving as y. It is essential, however, that to a single value of x there is always one and only one value of y. Note that this condition may not necessarily go in the reverse direction as equal values of y may indeed correspond to different values of $x_1 \neq x_2$. Indeed, consider, e.g., the function $y = x^2$ for which two values of $x = \pm 1$ result in the same value of $y = 1$.

Four arithmetic operations defined between real numbers (addition, subtraction, multiplication and division) can be applied to any two functions $f(x)$ and $g(x)$ to define new functions: $f(x) \pm g(x)$, $f(x)g(x)$ and $f(x)/g(x)$ (in the latter case the resulting function is not defined for the values of x where $g(x) = 0$). In addition, we can define a *composition* of two functions $f(x)$ and $g(x)$ via the following chain of operations: $y = g(w)$ and $w = f(x)$, i.e. $y = g(f(x))$. For instance, $y = \sin^2(x)$ is a composition of the functions $y = x^2$ and $y = \sin(x)$. Obviously, one can also define a composition of more than two functions, e.g. $y = g\left(f\left(d(x)\right)\right)$.

Functions can be specified in various ways, the most useful one of which is analytic using either a single expression (e.g. $y = x^3 + 1$) or several as is the case for the *Heaviside* or *unit step function*:

$$H(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases}, \tag{2.1}$$

L. Kantorovich, *Mathematics for Natural Scientists*, Undergraduate Lecture Notes in Physics, DOI 10.1007/978-1-4939-2785-2_2

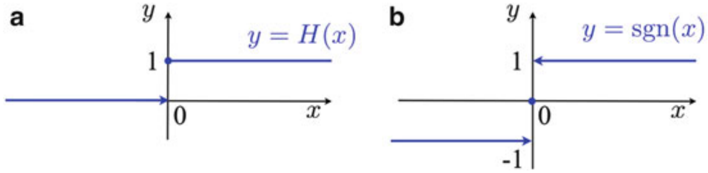


Fig. 2.1 Heaviside (**a**) and sign (**b**) functions, see Eqs. (2.1) and (2.2). The *arrows* at $x = 0$ at the end of some lines indicate that this point is excluded, while the *fat dot*—included

where we used $H(0) = 1$. It is shown in Fig. 2.1(a). Other definitions of $H(x)$ at $x = 0$ are also possible. Another useful example of such a definition is the *sign function* which is defined similarly as:

$$\mathrm{sgn}(x) = \begin{cases} -1, & \text{if } x < 0 \\ 0, & \text{if } x = 0 \\ 1, & \text{if } x > 0 \end{cases}, \tag{2.2}$$

shown in Fig. 2.1(b). Moreover, the two functions are related for all $x \neq 0$ as $\mathrm{sgn}(x) = 2H(x) - 1$. However, if $H(x)$ was defined at $x = 0$ as $H(0) = 1/2$, then the mentioned relationship would hold even for $x = 0$. This redefinition of $H(x)$ at $x = 0$ is sometimes useful.

Problem 2.1. *Prove that*

$$sgn(x) = 2H(x) - 1 = H(x) - H(-x).$$

We shall frequently be using graphical representations of functions where the points x and $y = f(x)$ are displayed on the $x - y$ plane. This could be either a collection of points (as is the case e.g. for the factorial function) or a continuous line (as, e.g., for the parabola $y = x^2$). However, when discussing functions, we shall mostly be dealing with continuous functions which are defined on a continuous interval of the values of x in $\mathbb{R}$.

Now we shall talk about the classifications of functions. A function can be *limited from above* on the given interval of x if for all values of the x within it there exists a real M such that $f(x) \leq M$. For instance, $y = -x^2$ is limited from above, $y < M = 0$, for all x in $\mathbb{R}$, i.e. for any $-\infty < x < \infty$. Similarly, if there exists such real m that for all values of x we have $f(x) \geq m$, then the function $y = f(x)$ is *limited from below*, as is the case, e.g., for $y = x^2$ ($m = 0$ in this case). A function can also be limited both from above and below if there exists such positive M that $|f(x)| \leq M$. An obvious example is $y = \sin(x)$ since $-1 \leq \sin(x) \leq 1$, i.e. for the sine function $M = 1$. On the other hand, the function $y = 1/x^2$ is not limited in the interval $0 < x < \infty$

since it is impossible to find a single constant $M > 0$ such that $1/x^2 \leq M$ for all positive x: whatever value of M we take, there always be values of x such that $1/x^2 > M$ in the interval $0 < x < 1/\sqrt{M}$.

A function $y = f(x)$ is *strictly increasing monotonically* within an interval, if for any $x_1 < x_2$ follows $f(x_1) < f(x_2)$, while the function is *strictly decreasing monotonically* if $f(x_1) > f(x_2)$ follows instead. For instance, $y = x^2$ is an increasing function for $x \geq 0$, while it is decreasing for $x \leq 0$. Strictly decreasing or increasing functions are called *monotonic*.

A function is *periodic* if there exists such number T' that for any x (within the function definition) $f(x + T') = f(x)$. The smallest such number T is called the *period*. For instance, $y = \sin(x)$ is a periodic function since $\sin(x + 2\pi n) = \sin(x)$ for any integer $n = 1, 2, \ldots$, i.e. the values of T' form a sequence 2π, 4π, etc.; however, only the smallest their value is the period, $T = 2\pi$.

Above we have defined a *direct* function whereby each value of x is put in the direct correspondence with a single value of y via $y = f(x)$. One can, however, solve this equation with respect to x and write it as a function of y. This will be a function different from $f(x)$ and it is called (if exists, see below) the *inverse function* $x = f^{-1}(y)$. Conventionally, however, we would like the x to be the argument and y to be the function. Hence, we swap the letters x and y and write the inverse function as $y = f^{-1}(x)$.

Example 2.1. ► Obtain the inverse function to $y = 5x$.

Solution. Solving $y = 5x$ for x, we get $x = \frac{1}{5}y$; interchanging x and y for the inverse function to be written in the conventional form, we get $y = \frac{1}{5}x$, i.e. the inverse function to $f(x) = 5x$ is $f^{-1}(x) = \frac{1}{5}x$. ◄

Example 2.2. ► Obtain the inverse function to $y = x^3$.

Solution. Solving for x and interchanging x and y, we obtain $y = f^{-1}(x) = x^{1/3}$ which is the *cube root* function (see Sect. 2.3.3). ◄

The inverse function exists if and only if there is one-to-one correspondence between x and $y = f(x)$. This means that for any $x_1 \neq x_2$ it follows that $y_1 = f(x_1) \neq y_2 = f(x_2)$. This is because for each value of the x there must be a unique value of $y = f(x)$ for the direct function to be well defined; at the same time, for the inverse function to be defined, it is necessary that for each value of y there will be one and only one value of $x = f^{-1}(y)$.

It is easy to see that if the direct function is increasing monotonically, then this is sufficient to establish the existence of the inverse function since the requirement for that condition is satisfied: indeed, if for any $x_1 < x_2$ (and hence $x_1 \neq x_2$) follows $y_1 = f(x_1) < y_2 = f(x_2)$, then obviously $y_1 \neq y_2$, as required. Similarly, the inverse function $f^{-1}(x)$ exists if the direct function $f(x)$ is monotonically decreasing.

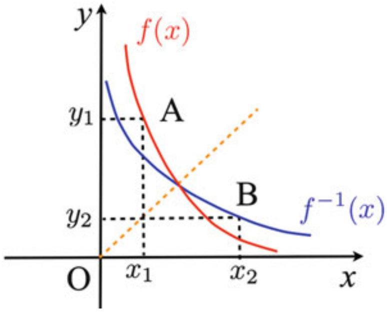


Fig. 2.2 To the proof of the symmetry of $f(x)$ and $f^{-1}(x)$ with respect to the diagonal (shown with the *brown dashed line*) of the quadrants I and III

Problem 2.2. *Prove that if $f(x)$ is an increasing (decreasing) function, then so is its inverse (if it exists).*

Problem 2.3. *Let us consider a function $f(x)$ whose inverse is $f^{-1}(x)$. Prove that the inverse to $f^{-1}(x)$ is $f(x)$, i.e. the two functions are mutually inverse to each other.*

It is easy to see that the graphs of the two functions, $y = f(x)$ and $y = f^{-1}(x)$, are symmetric with respect to the diagonal line drawn in the first ($x, y > 0$) and the third ($x, y < 0$) quadrants, see Fig. 2.2. Indeed, consider a point A on the curve of the function $f(x)$ with the coordinates $(x_1, y_1 = f(x_1))$. Consequently, $x_1 = f^{-1}(y_1)$ by the definition of the inverse function. However, conventionally, we write this as $y_2 = f^{-1}(x_2)$, i.e. $y_2 = x_1$ and $x_2 = y_1$. It is seen that the point A on $f(x)$ corresponds to point B on the curve of $f^{-1}(x)$ with the coordinates (x_2, y_2). It is now obvious from the drawing in Fig. 2.2 that the two points A and B are located symmetrically with respect to the dashed brown diagonal line. Since the point A was chosen arbitrarily on $f(x)$, the reasoning given would correspond to any point on $f(x)$, i.e. the whole two curves, $f(x)$ and $f^{-1}(x)$, are symmetric with respect to the diagonal line of the first and third quadrants.

It follows from that simple discussion that the graph of $f^{-1}(x)$ can be obtained from the graph of $f(x)$ by reflecting the latter across the diagonal of the quadrants I and III.

As should by now be clear, not every function has an inverse. Consider, for instance, a parabola $y = x^2$, shown in Fig. 2.3 with the red line. If we reflect $y = x^2$ in the first quadrant diagonal (shown as dashed brown, as before), we obtain the blue curve containing two parts: one shown with the solid blue line which is $y = \sqrt{x} > 0$, and another shown as dotted blue line and corresponding to $y = -\sqrt{x} < 0$. In other words, in this case the two curves (the red and blue) are rotated by 90° with respect to each other. It is easy to see, however, that to a single value of $x > 0$ there will be *two* values of $y = \pm\sqrt{x}$ which contradicts the definition of the function given above. This is of course because while solving the equation $y = x^2$ with respect

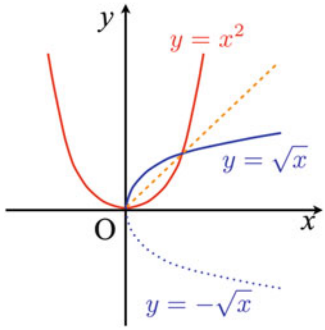


Fig. 2.3 Parabola and its inverse

to x, we obtain two values of the square root. It follows from this discussion that, strictly speaking, the function $y = x^2$ does not have an inverse. As we still would like to define an inverse to the parabola, we shall take only one value of the square root, and it is conventional to take the positive root. This way we obtain the function $y = \sqrt{x} > 0$ drawn in Fig. 2.3 with the solid blue line. This is a function since to any value of x there is only one value of $y = \sqrt{x}$ corresponding to it.

The reduction operation considered above for the parabola was necessary to define properly the inverse function to it, and the same operation is applied in some other cases as will be demonstrated in the next section.

2.2 Infinite Numerical Sequences

2.2.1 Definitions

To continue our gradual approach to functions, we have to consider now numerical sequences and their *limit*, otherwise it would be difficult for us to define power and exponential functions which are the matter of the next section.

Consider an infinite set of numbers $x_0, x_1, x_2, x_3, \ldots, x_n, \ldots$ which are calculated via a certain rule from the integer numbers $0, 1, 2, 3, \ldots, n, \ldots$ taken from $\mathbb{N}at$. In other words, the numbers $x_n = f(n)$ are obtained as a function depending on an integer number n. For instance, the set $1, -1, 1, -1, 1, \ldots$ can be obtained by calculating $(-1)^n$ sequentially for $n = 0, 1, 2, 3, \ldots$. Another set

$$1,\ \frac{1}{2},\ \frac{1}{3},\ \frac{1}{4},\ \frac{1}{5},\ \frac{1}{6},\ \frac{1}{7},\ \ldots \tag{2.3}$$

is obtained from $1/n$ with $n = 1, 2, 3, \ldots$. If the former set $(-1)^n$ results in an alternating sequence of numbers, the latter sequence $1/n$ obviously tends to smaller and smaller numbers and one may expect that the numbers in the sequence approach zero for large values of n, being every time slightly above it. One may say that if the former sequence does not have a definite limit when n becomes indefinitely large (we say "n tends to infinity", which is denoted as ∞, and write it as $n \to \infty$), the latter sequence does: $1/n \to 0$ as $n \to \infty$. We use a special symbol "lim" to say that:

$$\lim_{n\to\infty} \frac{1}{n} = 0.$$

Although intuitively our argumentation may be sound and clear, proper analysis requires a rigorous definition for the limit of an infinite numerical sequence to exist. To formulate the required definition, let us assume that a sequence $x_0, x_1, x_2, x_3, \ldots, x_n, \ldots$ approaches a. Intuitively that means that as n gets bigger and bigger, the values x_n of the sequence get closer and closer to a, i.e. the absolute value of the difference $|x_n - a|$ gets smaller and smaller (the absolute value is needed since the numbers in the sequence may appear on both sides of a, consider, e.g., the sequence $(-1)^n/n$). So, a rigorous formulation sounds like this: for any $\Delta > 0$ there always exists such positive integer N, that for all members of the sequence staying *after* it (i.e. with numbers greater than it, $n > N$) we shall have $|x_n - a| < \triangle$. This is written as follows:

$$\lim_{n\to\infty} x_n = a. \tag{2.4}$$

For example, consider our sequence $x_n = 1/n$ for which we suspect that it approaches $a = 0$. Let us take a small positive number $\triangle = 0.01$. Then, we have to show that starting from some number N all members of the sequence are smaller than $\triangle$, i.e. they satisfy $|x_n| < \triangle = 0.01$. Indeed, from the inequality $1/n < \triangle$ we get $n > 1/\triangle = 100$; in other words, there exist such number $N = 100$ so that any member of the sequence $x_{101} = \frac{1}{101} \simeq 0.0099$, $x_{102} = \frac{1}{102} \simeq 0.0098$, $x_{103} = \frac{1}{103} \simeq 0.0097$, etc. satisfies the condition $|x_n| < 0.01$, as required. If we take even smaller value of $\triangle = 0.001$, then solving for n in $1/n < \triangle$ will result in $N = 1/\triangle = 1000$, i.e. for all $n > 1000$ all members of the sequence x_n are smaller than this smaller value of $\triangle$. It is clear that whatever value of Δ we take, however small, one can always find such number N that all members x_n of the sequence with $n > N$ will be smaller than Δ. This means that indeed, according to the given above definition, the sequence $x_n = 1/n$ tends to zero.

This makes perfect sense: the smaller the value of the *deviation* from a (the value of $\triangle$), i.e. the closer we want the members of the sequence to be to that value, the larger the value of N needs to be taken, i.e. we have to consider more and more distant members of the sequence (with larger n); x_n with small values of n may deviate from the limit a significantly, but as n increases, the elements of the sequence x_n become closer and closer to it.

Problem 2.4. *Prove that the infinite numerical sequence $x_n = 1/n^2$ tends as well to zero, i.e. its limit is equal to zero.*

Problem 2.5. *Prove, using the definition, that the infinite numerical sequence* $x_n = (1 + 2n) / n$ *tends to* 2.

Problem 2.6. *Prove that the sequence* $x_n = q^n$ *does not have a limit if* $|q| > 1$.

Using the definition of the limit of a sequence is very tedious in practice, and it is much easier to perform algebraic manipulations on the terms of the sequence and this way find the limit of it. Consider again Problem 2.5. Each element of the sequence can alternatively be written as

$$x_n = \frac{1}{n} + 2,$$

Then, considering the original sequence $x_n = y_n + z_n$ as a sum of two sequences, $y_n = 1/n$ and $z_n = 2$, we may apply the limit to each sequence separately to get $y_n = 1/n \to 0$, $z_n \to 2$, and hence $x_n \to 2$ as $n \to \infty$.

Example 2.3. ► Consider the limit of the sequence

$$x_n = \frac{2n^2 + 3n + (-1)^n}{3n^2 - 4n + 5}.$$

Solution. Divide both the numerator and denominator by n^2:

$$x_n = \frac{2 + 3/n + (-1)^n / n^2}{3 - 4/n + 5/n^2} \to \frac{2}{3},$$

as all other terms tend to zero as $n \to \infty$. ◄

2.2.2 Main Theorems

The validity of the operations we have performed in the examples above requires a rigorous justification. This is accomplished by the following simple theorems which we shall now discuss.

Theorem 2.1 (Addition Theorem). *If two sequences* x_n *and* y_n *have certain limits* X *and* Y*, then their sum sequence* $z_n = x_n + y_n$ *has a limit and it is* $X + Y$.

Proof. Since the first sequence x_n has the limit, then for any $\Delta_x = \Delta/2$ there exists such number N_x that starting from it (i.e. for any $n > N_x$) all members of the sequence would not differ from the limit by more than Δ_x, i.e. we would have $|x_n - X| < \Delta_x$. Similarly, for the second sequence: for any $\Delta_y = \Delta/2$ there would exist such N_y, so that for any $n > N_y$ we have $|y_n - Y| < \Delta_y$. Then, we can write:

$$|z_n - (X + Y)| = |(x_n - X) + (y_n - Y)| \le |x_n - X| + |y_n - Y| < \Delta_x + \Delta_y = \Delta. \tag{2.5}$$

Here we used the inequality (1.92) stating that an absolute value of a sum of numbers is always less than or equal to the sum of their absolute values. Therefore, it appears that whatever value of $\Delta > 0$ we choose, one can always find an N such that the members of the sequence $z_n = x_n + y_n$ will be closer than Δ to the limiting value $X + Y$ as stated by inequality (2.5), e.g. $N = \max\left\{N_x, N_y\right\}$. This means that the limit of the sequence z_n indeed exists and the sequence converges to the sum of the individual limits, $X + Y$. **Q.E.D.**

Theorem 2.2 (Product Theorem). *If two sequences x_n and y_n have certain limits X and Y, then their product sequence $z_n = x_n y_n$ has a limit and it is XY.*

Proof. The logic is similar to that of the previous theorem; we only need to estimate the absolute value of the difference to derive the necessary value of Δ in this case:

$$\begin{aligned}|x_n y_n - XY| &= |(x_n - X)(y_n - Y) + X(y_n - Y) + Y(x_n - X)| \\ &\le |x_n - X|\,|y_n - Y| + |X|\,|y_n - Y| + |Y|\,|x_n - X| \\ &< \Delta_x \Delta_y + |X|\,\Delta_y + |Y|\,\Delta_x = \Delta,\end{aligned}$$

i.e. by taking this value of Δ and $N = \max\left(N_x, N_y\right)$ any member of the product sequence z_n with $n > N$ will be different from the value of XY by no more than Δ, i.e. it is the limit of the product sequence. **Q.E.D.**

Theorem 2.3 (Division Theorem). *If two sequences x_n and $y_n \neq 0$ have certain limits X and $Y \neq 0$, then their ratio sequence $z_n = x_n/y_n$ has a limit and it is X/Y.*

Proof. Again, this is proven by the following manipulation:

$$\left|\frac{x_n}{y_n} - \frac{X}{Y}\right| = \left|\frac{x_n Y - y_n X}{Y y_n}\right| = \left|\frac{(x_n - X)\,Y - (y_n - Y)\,X}{Y y_n}\right|$$
$$\leq \frac{|x_n - X|\,|Y| + |y_n - Y|\,|X|}{|Y|\,|y_n|} < \frac{\Delta_x\,|Y| + \Delta_y\,|X|}{|Y|\,|y_n|}.$$

Since $y_n \neq 0$ for any n, then it should be possible to find such $\delta > 0$ that $|y_n| \geq \delta$ for any n. Hence, $1/\,|y_n| \leq 1/\delta$, and we can then continue:

$$\left|\frac{x_n}{y_n} - \frac{X}{Y}\right| < \frac{\Delta_x\,|Y| + \Delta_y\,|X|}{|Y|\,|y_n|} \leq \frac{\Delta_x\,|Y| + \Delta_y\,|X|}{|Y|\,\delta}.$$

Choosing the expression in the right-hand side of the inequality above as Δ finally proves the theorem. **Q.E.D.**

The three theorems enable us to manipulate the limits using algebraic means, and this is exactly what we did in the previous example:

$$\lim_{n\to\infty} x_n = \lim_{n\to\infty} \frac{2n^2 + 3n + (-1)^n}{3n^2 - 4n + 5} = \lim_{n\to\infty} \frac{2 + 3/n + (-1)^n\,/n^2}{3 - 4/n + 5/n^2}$$
$$= \frac{2 + \lim_{n\to\infty}(3/n) + \lim_{n\to\infty}\left[(-1)^n\,/n^2\right]}{3 - \lim_{n\to\infty}(4/n) + \lim_{n\to\infty}(5/n^2)} = \frac{2+0+0}{3-0+0} \to \frac{2}{3}.$$

If we take a *subsequence* of the members of some original sequence (e.g. only terms with even n), then we shall get another infinite numerical sequence, e.g.

$$1,\ \frac{1}{2},\ \frac{1}{4},\ \frac{1}{6},\ \ldots$$

is a subsequence of the sequence (2.3). It might appear to be obvious that if the original sequence has a certain limit, then its arbitrary subsequence would also converge to the same limit. This is stated by the following

Theorem 2.4 (Subsequence Theorem). *If a sequence x_n (with indices n taking all integer values $\mathbb{N}at$) tends to a limit X, then its arbitrary subsequence y_k (where indices k take on only a subset of values from the original set $\mathbb{N}at$ and for these indices $y_k = x_k$) tends to the same limit.*

Proof. This follows directly from the definition of the limit. Since the original sequence x_n has a limit, then for any Δ there is such N so that for any $n > N$ we have $|x_n - X| < \Delta$. Take now a subset of indices, let us call them k. On the subset using the same value of Δ, we have that the same inequality, $|y_k - X| < \Delta$, is satisfied for all values of $k > N$ as these indices form a subset of the original indices $n > N$ for which it was satisfied. **Q.E.D.**

Therefore, if a sequence $x_n = 1/n$ (equal to all inverse integers) has the limit 0, then the same limit would have its subsequence $y_n = 1/(2n)$ containing only inverse even integers. However, if one can find at least two subsequences of the given sequence which converge to different limits individually, then we have to state that the original sequence does *not* have a limit.

Problem 2.7. *Prove that the sequence $x_n = (-1)^n$ does not have a limit.*

If a sequence has a limit, it must be unique, it is impossible for a sequence to have two different limits. This is stated by the following:

Theorem 2.5 (Uniqueness). *If a sequence x_n has a limit, it is unique.*

Proof. We shall prove this theorem by contradiction. Assume that the sequence x_n has two different limits, X_1 and $X_2 > X_1$. To prove this assumption wrong, we take a specific value of $\Delta = (X_2 - X_1)/2 > 0$. Since X_1 is the limit, than starting form some N_1 (i.e. for all $n > N_1$) we have $|x_n - X_1| < \Delta$. Similarly, since X_2 is also a limit, there must be such N_2 so that for any $n > N_2$ we have $|x_n - X_2| < \Delta$ with the same Δ. Then we write:

$$X_2 - X_1 = |X_2 - X_1| = |(X_2 - x_n) + (x_n - X_1)| \leq |X_2 - x_n| + |X_1 - x_n| < 2\Delta = X_2 - X_1,$$

which is obviously incorrect. Note that we have the strict "lesser than" sign here! Hence, our assumption is wrong. The theorem is proven. **Q.E.D.**

There are several theorems related to inequalities which are sometimes used to find particular limits. We shall consider them next.

Theorem 2.6. *If all elements of the sequence, starting from some particular element x_{n_0} (i.e. for all $n \geq n_0$), satisfy $x_n \geq A$ and the sequence has the limit X, then $X \geq A$.*

Proof. We prove this theorem by contradiction. Assume that $X < A$. Then, since the sequence has the limit, there is such $N > n_0$ that for any $n > N$ we have $|x_n - X| < \Delta$. This means that $-\Delta < x_n - X < \Delta$. In particular, $x_n - X < \Delta$ or

$x_n < \Delta + X$. Take $\Delta = A - X > 0$, then $x_n < \Delta + X = A$, which contradicts the fact that $x_n \geq A$. Our initial assumption is therefore wrong. **Q.E.D.**

Theorem 2.7. *If all elements of the sequence, starting from some particular number n_0, satisfy $x_n \leq A$ and the sequence has the limit X, then $X \leq A$.*

Problem 2.8. *Prove this theorem.*

Theorem 2.8. *If elements of two sequences x_n and y_n, starting from some number n_0, satisfy $x_n \leq y_n$ and their limits exist and equal to X and Y, respectively, then $X \leq Y$.*

Proof. Consider a sequence $z_n = y_n - x_n$. All elements of the new sequence, starting from some number n_0, satisfy $z_n \geq 0$ and its limit is $Z = Y - X$. Therefore, according to Theorem 2.6, $Z = Y - X \geq 0$, or $Y \geq X$, as required. **Q.E.D.**

This theorem has a simple corollary that if a sequence, y_n, is bracketed between two other sequences, $x_n \leq y_n \leq z_n$, then its limit is also bracketed between the two limits, i.e. $X \leq Y \leq Z$. In particular, if $Z = X$, then $Y = X = Z$.

2.2.3 Sum of an Infinite Numerical Series

Having defined a numerical sequence, x_n, one can consider a finite sum (cf. Sect. 1.9)

$$S_N = \sum_{n=0}^{N} x_n$$

of its first N members. These so-called *partial sums* S_N for different values of N form a numerical sequence in their own right, and the limit S of that sequence, if exists, should give us the sum of an infinite series built from the original sequence x_n. It is instructive to explicitly state here the definition of the convergence of the infinite series which is nothing but a rephrased definition (for S_N in this case) of the limit given above in Sect. 2.2.1 for the numerical sequence: the series $\sum_{n=0}^{\infty} x_n$ converges to the value S, if for any positive ϵ one can always find such number $\mathcal{N}$ that for any $N > \mathcal{N}$ the partial sum S_N would be different from S by no more than ϵ, i.e. $|S_N - S| < \epsilon$.

As an example of an infinite numerical series, consider an infinite geometrical progression [cf. Eq. (1.60)]:

$$S_{\text{geom}} = a_0 q^0 + a_0 q^1 + a_0 q^2 + a_0 q^3 + \cdots = \lim_{N\to\infty} S_N = \lim_{N\to\infty} \frac{a_0\left(1 - q^{N+1}\right)}{1-q}, \tag{2.6}$$

where we used the formula (1.61) for the sum S_N of the first N terms of the geometrical series. Now, if $|q| > 1$, then $q^{N+1} \to \infty$ as $N \to \infty$ (obviously, q^N increases with N), and hence the series does not have a finite limit. If $q = 1$, then the series converges to infinity as well since it consists of identical terms a_0. If $q = -1$, then the series contains alternating a_0 and $-a_0$ terms, and hence (similarly to the case of $x_n = (-1)^n$ discussed above) does not have a limit at all. Therefore, it only makes sense to consider the case of $|q| < 1$. In this case $q^{N+1} \to 0$ as $N \to \infty$ since q^N decreases when N increases, and we obtain after taking the limit in (2.6) an important result:

$$S_{\text{geom}} = \lim_{N\to\infty} \frac{a_0\left(1 - q^{N+1}\right)}{1-q} = \frac{a_0}{1-q}. \tag{2.7}$$

Example 2.4. ► Determine the fractional form $r = p/q$ of the rational number, $r = 0.9\,(123)$, which contains a periodic sequence of digits in its decimal representation.

Solution. we write:

$$r = 0.9 + 0.0123 + 0.0000123 + \cdots = 0.9 + 0.0123\left(1 + 10^{-3} + 10^{-6} + \cdots\right).$$

The expression in the brackets is nothing but the infinite geometric progression. Using (2.7) for its sum, we finally obtain:

$$r = 0.9 + \frac{0.0123}{1 - 10^{-3}} = \frac{9}{10} + \frac{123}{10000}\frac{1}{0.999} = \frac{9}{10} + \frac{123}{9990} = \frac{9114}{9990}. \blacktriangleleft$$

Problem 2.9. *Prove that any number $r < 1$ specified via a periodic sequence of digits in the decimal representation is a rational number, i.e. it can always be written as a ratio of two integers.*

2.3 Elementary Functions

Here we shall introduce all elementary functions and their main properties. As we are sure the reader is familiar with all of them, our consideration may be rather brief; at the same time, it is worthwhile to define all functions systematically gradually making emphasis on some specific points.

2.3.1 Polynomials

A constant is the simplest function: $y = C$ with C being some real number. Its graph is represented by a horizontal line crossing the y axis exactly at the value of C. It is also the simplest polynomial: a polynomial of degree zero.

A linear combination of x in integer powers, from zero to n, with real coefficients, as was already mentioned in Sect. 1.5, gives rise to a function called a *polynomial* of degree n:

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = \sum_{i=0}^{n} a_i x^i . \tag{2.8}$$

It is of special utility to know the roots of the polynomial, i.e. the values of x at which $f(x) = 0$. The reader must be familiar with the methods of finding the roots of the polynomials of the degree 1 (linear) and 2 (quadratic), see Sect. 1.5. Although cubic and quartic equations have been solved in the fifteenth/sixteenth centuries in radicals (i.e. using rational functions and roots),[1] the solutions look complicated and are of little practical use. In fact, it was later proven[2] that it is not possible to express the roots of general polynomial equations with degree higher than 4 in radicals, so that in practice numerical solutions (or analytical ones in simple cases) are usually made. What is important for us here is to state that equation $f(x) = 0$ always has at least one solution, and the maximum possible number of different roots is n. Some of the roots may be complex. However, in that case they will always come in complex conjugate pairs, $a + ib$ and $a - ib$. More precisely, and saying this a bit differently as well, one may say that this equation always has n roots some of which (or all) could be the same (repeated). This fact can be simply put by writing the function $f(x)$ in a factorised form:

$$f(x) = a_n (x - x_1)^{k_1} (x - x_2)^{k_2} \cdots (x - x_m)^{k_m} , \tag{2.9}$$

which states explicitly that $f(x)$ has m distinct roots x_1 (repeated k_1 times), x_2 (repeated k_2 times), and so on ($n = k_1 + k_2 + \cdots + k_m$). Multiplying the brackets in the above equation, a polynomial of degree n is obtained; comparing the same powers of x in it with the definition in Eq. (2.8), algebraic relationship between the roots $\{x_i\}$ and the coefficients of the polynomial $\{a_i\}$ can be established.

For instance, for the quadratic polynomial,

$$a_2 (x-x_1)(x-x_2) = a_2 \left[x^2 - (x_1+x_2)x + x_1 x_2\right] \implies \frac{a_1}{a_2} = -x_1 - x_2 , \quad \frac{a_0}{a_2} = x_1 x_2 .$$

[1]This was done by Italian mathematicians Lodovico Ferrari, Niccolo Fontana Tartaglia and Gerolamo Cardano.

[2]In 1824 by Italian mathematician Paolo Ruffini and Norwegian mathematician Niels Henrik Abel.

Problem 2.10. *Show that such a relationship for the cubic polynomial, assuming all roots are different and also that $a_3 = 1$, is:*

$$a_2 = -x_1 - x_2 - x_3; \quad a_1 = x_1x_2 + x_1x_3 + x_2x_3 \quad \textit{and} \quad a_0 = -x_1x_2x_3.$$

Problem 2.11. *The same for the quartic polynomial:*

$$a_3 = -x_1 - x_2 - x_3 - x_4; \quad a_2 = x_1x_2 + x_1x_3 + x_1x_4 + x_2x_3 + x_2x_4 + x_3x_4;$$

$$a_1 = -x_1x_2x_3 - x_1x_2x_4 - x_1x_3x_4 - x_2x_3x_4; \quad a_0 = x_1x_2x_3x_4.$$

Note that all coefficients are fully symmetric with respect to an arbitrary permutation of the roots. Moreover, one can see that each coefficient has all possible combinations of roots indices. For instance, consider the quartic case. a_1 is a sum (with the minus sign) of four terms corresponding to $\binom{4}{3} = 4!/3! = 4$ possible combinations of choosing 3 different indices from a pool of four, a_2 is the sum of 6 terms due to $\binom{4}{2} = 4!/(2!2!) = 6$ combinations, while a_3 contains $\binom{4}{1} = 4!/3! = 4$ terms and a_0 contains only $\binom{4}{4} = 1$.

Since the complex roots always come in complex conjugate pairs, the factorisation of a polynomial (2.9) can also be written entirely via real factors. Indeed, let $x = a + ib$ and $x^* = a - ib$ be two such roots, then the product of their factors,

$$(x - a - ib)(x - a + ib) = x^2 - 2ax + \left(a^2 + b^2\right) = x^2 + px + q,$$

becomes entirely real. Therefore, in general, a real polynomial is represented in a factorised form as

$$f(x) = a_n (x - x_1)^{k_1} \cdots (x - x_m)^{k_m} \left(x^2 + p_1x + q_1\right)^{l_1} \cdots \left(x^2 + p_rx + q_r\right)^{l_r}. \tag{2.10}$$

Here we first have m simple factors $x - x_i$ corresponding to real roots of $f(x)$, which are followed by square polynomials $x^2 + p_ix + q_i$ for each i-th pair of the complex conjugate roots.

Graphs of several integer power functions $y = x^n$ are shown in Fig. 2.4(a). Functions with even n are even and hence symmetrical with respect to the y axis, while functions with odd values of n are odd, i.e. for them $(-x)^n = -x^n$.

2.3.2 Rational Functions

Rational functions consist of the ratio of polynomials. In many applications it is necessary to decompose such a function into simpler functions. It is done by writing

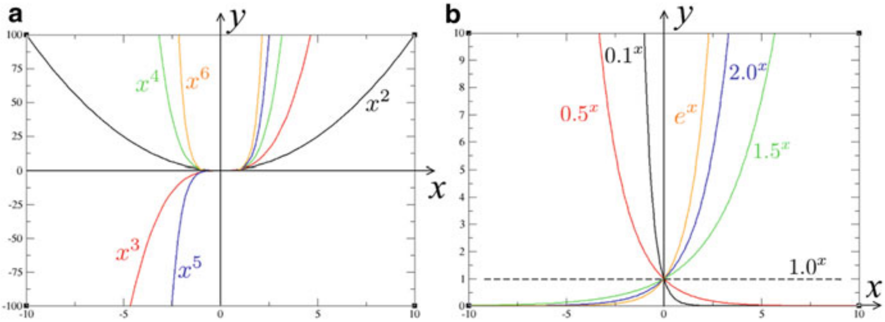


Fig. 2.4 (**a**) Power functions $y = x^n$ for selected values of n; (**b**) exponential functions $y = a^x$ for several values of the base a

down a desired decomposition with *undetermined coefficients* and then finding the coefficients by comparing both sides (as the equality should hold for any value of x). As an example, consider

$$f(x) = \frac{1}{x^2 - 1}.$$

Since $x^2 - 1$ can be factored as $(x-1)(x+1)$, we attempt to decompose $f(x)$ as follows:

$$\frac{1}{x^2-1} = \frac{A}{x-1} + \frac{B}{x+1},$$

with A and B being yet to be determined constants. Summing up the two terms above in the right-hand side, we obtain a fraction with the denominator identical to the original one, $x^2 - 1$, standing on the left. Let us then compare the numerators in the two sides:

$$1 = A(x+1) + B(x-1). \tag{2.11}$$

Since this equation is to hold for any x, we may, for instance, put two values of x to obtain two independent equations for the two constants. For instance, take $x = 1$ and $x = -1$, yielding

$$\begin{cases} 1 = 2A \\ 1 = -2B \end{cases} \implies \begin{cases} A = 1/2 \\ B = -1/2 \end{cases},$$

so that

$$\frac{1}{x^2-1} = \frac{1}{2}\left(\frac{1}{x-1} - \frac{1}{x+1}\right). \tag{2.12}$$

Another way of approaching the problem of solving Eq. (2.11) is to rearrange the right-hand side first in a form of a polynomial:

$$1 = (A - B) + (A + B)\,x.$$

In the left-hand side the polynomial is equal to just a constant, i.e. when we compare the coefficients in both sides to the same powers of x, we obtain two equations

$$\begin{cases} 1 = A - B \\ 0 = A + B \end{cases} \Longrightarrow \begin{cases} A = 1/2 \\ B = -1/2 \end{cases},$$

resulting in the same solution. This method might be preferable to the first one as it clearly uses the fact that the identity (2.11) must hold for any x.

Problem 2.12. *Decompose the following fraction:*

$$\frac{1}{x\,(x-1)\,(x-2)} \quad \left[\textit{Answer: } \frac{1}{2x} - \frac{1}{x-1} + \frac{1}{2\,(x-2)}\right].$$

Generally, if the function $f(x) = P_n(x)/Q_m(x)$ is given as a ratio of two polynomials of degrees n and m, two cases must be distinguished: when $n < m$ and $n \geq m$. The latter case can always be manipulated into the first one by singling out the denominator in the numerator at its highest power term(s), as, e.g., done in the following example:

$$\frac{3x^3+2x+1}{x^2+1} = \frac{3x\left[\left(x^2+1\right)-1\right]+2x+1}{x^2+1} = \frac{3x\left(x^2+1\right)-x+1}{x^2+1} = 3x - \frac{x-1}{x^2+1}. \tag{2.13}$$

This way one can always represent P_n/Q_m with $n \geq m$ as

$$\frac{P_n}{Q_m} = R_r + \frac{G_l}{Q_m}, \tag{2.14}$$

where R_r and G_l are polynomials and $l < m$.

Problem 2.13. *Represent the following polynomials in the simpler form using manipulations similar to the ones applied in Eq. (2.14):*

$$\frac{x^5+3x^2+3x+1}{x^2+5x+1} \quad \left[\textit{Answer: } \left(x^3-5x^2+24x\right) - \frac{112x^2+21x-1}{x^2+5x+1}\right]; \tag{2.15}$$

$$\frac{x^3+12x^2-x+1}{x+1} \quad \left[\textit{Asnwer: } \left(x^2+11x\right) - \frac{12x-1}{x+1}\right]. \tag{2.16}$$

The rational function P_n/Q_m with $n < m$ can then be decomposed further, as we have seen above, into a sum of simpler fractions. A general statement (which we shall not prove here) is the following: if the denominator $Q_m(x)$ is decomposed in a product of simple factors as in Eq. (2.10), then the fraction P_n/Q_m $(n < m)$ can be written as a sum of simpler fractions using the following rules:

- each single root a which is repeated r times gives rise to the following sum of terms:
$$\frac{A_1}{x-a} + \frac{A_2}{(x-a)^2} + \cdots + \frac{A_r}{(x-a)^r}$$
with r arbitrary constants $A_1, \ldots, A_r$;
- each pair of complex roots $z = a + ib$ and $z^* = a - ib$ corresponding to the square polynomial $x^2 + px + q$ results in the sum of terms
$$\frac{B_1 x + C_1}{x^2 + px + q} + \frac{B_2 x + C_2}{(x^2 + px + q)^2} + \cdots + \frac{B_l x + C_l}{(x^2 + px + q)^l}$$
with $2l$ arbitrary constants $B_1, C_1, \ldots, B_l, C_l$, where l is the repetition number for the given pair of roots.

Example 2.5. ► Decompose

$$R = \frac{3x^2 + x - 1}{(x-1)^2 (x^2+1)}.$$

Solution. Using the rules, we write:

$$R = \frac{A_1}{x-1} + \frac{A_2}{(x-1)^2} + \frac{Bx+C}{x^2+1} = \frac{A_1 (x-1)\left(x^2+1\right) + A_2\left(x^2+1\right) + (Bx+C)(x-1)^2}{(x-1)^2 (x^2+1)}.$$

Opening up the brackets in the numerator above, grouping terms with the same powers of x and comparing them with the corresponding coefficients of the original fraction, we obtain the following four equations for the four unknown coefficients to x^3, x^2, x^1 and x^0, respectively:

$$\begin{cases} x^3: & A_1 + B = 0 \\ x^2: & -A_1 + A_2 + C - 2B = 3 \\ x^1: & A_1 - 2C + B = 1 \\ x^0: & -A_1 + A_2 + C = -1 \end{cases} \Longrightarrow \begin{cases} A_1 = 2 \\ A_2 = 3/2 \\ B = -2 \\ C = -1/2 \end{cases},$$

i.e. we finally obtain the decomposition sought for:

$$\frac{3x^2 + x - 1}{(x-1)^2 (x^2+1)} = \frac{2}{x-1} + \frac{3}{2(x-1)^2} - \frac{4x+1}{2(x^2+1)}. \blacktriangleleft \tag{2.17}$$

Problem 2.14. *Prove the following identities corresponding to the decomposition into fractions:*

$$\frac{x^2-x+5}{(2x-1)\left(x^2+4x+4\right)} = \frac{1}{25}\left[\frac{19}{2x-1}+\frac{3}{x+2}-\frac{55}{(x+2)^2}\right];$$

$$\frac{x^3+x^2+1}{(3x-1)^2\left(2x^2+3\right)} = \frac{1}{2523}\left[\frac{137}{3x-1}+\frac{899}{(3x-1)^2}+\frac{189x+237}{2x^2+3}\right];$$

$$\frac{2x-3}{\left(x^2+1\right)^2\left(x^2+3\right)} = \frac{1}{4}\left[\frac{2x-3}{x^2+3}+\frac{-2x+3}{x^2+1}+\frac{4x-6}{\left(x^2+1\right)^2}\right].$$

[*Hint*: note that in the first fraction x^2+4x+4 should be factored first.]

Problem 2.15. *Demonstrate the following decompositions into the fractions:*

$$\frac{1}{x^4+1} = \frac{1}{2\sqrt{2}}\left(\frac{x+\sqrt{2}}{x^2+\sqrt{2}x+1}+\frac{-x+\sqrt{2}}{x^2-\sqrt{2}x+1}\right);$$

$$\frac{x^2}{x^4+1} = \frac{1}{2\sqrt{2}}\left(\frac{-x}{x^2+\sqrt{2}x+1}+\frac{x}{x^2-\sqrt{2}x+1}\right).$$

[*Hint*: note that

$$x^4+1 = \left(x^4+2x^2+1\right)-2x^2 = \left(x^2+1\right)^2-\left(\sqrt{2}x\right)^2 = \left(x^2+\sqrt{2}x+1\right)\left(x^2-\sqrt{2}x+1\right),$$

so that the denominator in each case can be factorised. Of course, the four roots of $x^4+1=0$ can also be formally found and then this expression be fully factored; but the roots are all complex and this method requires a better knowledge of complex numbers.]

2.3.3 General Power Function

The general power functions is $y = x^\alpha$ with parameter α being from $\mathbb{R}$ (a real number). Definition of the power function to any real power is done in stages. First of all, we introduce an integer power $\alpha = n$ as $y = x^n = x \cdot x \cdot \ldots \cdot x$, i.e. via a repeated multiplication. In addition, we define that $x^0 = 1$ for any x. It is easy to see that the power function is monotonically increasing. Indeed, consider first $n = 2$. Then, if $x_2 > x_1$, then

$$y_2 = x_2^2 = x_2x_2 > x_1x_2 > x_1x_1 = x_1^2 = y_1\ ,$$

i.e. $y_2 > y_1$. Similarly, that property can be demonstrated for any integer n.

Then, we define the power being an inverse integer, $\alpha = 1/n$, as the solution for y of the equation $y^n = x$, which we write $y = x^{1/n}$ or $y = \sqrt[n]{x}$. In other words, $y = x^{1/n}$ is defined as an inverse function to $y = x^n$. Note that in the cases of even values of n we only accept positive solutions for y and $x \geq 0$; this method would prevent us from arriving at multi-valued functions, as was discussed at the end of Sect. 2.1 using parabola as an example. Since the direct function, $y = x^n$, is monotonically increasing, so is the inverse integer power function, $y = x^{1/n}$, as was proven in the same section.

At the next step we define the power being a general rational number $\alpha = n/m$ via:

$$y = x^{n/m} = \left(x^{1/m}\right)^n$$

This is an example of the composition of two functions mentioned at the beginning of Sect. 2.1: indeed, we combined two functions, $y = w^n$ and $w = x^{1/m}$, into one: $y = w^n = \left(x^{1/m}\right)^n$. As both functions involved here are monotonically increasing, so is their composition. Hence, the function $y = x^\alpha$ for any rational α increases monotonically.

Now we can extend the definition of the power function to all *positive* real numbers α from $\mathbb{R}$. Since we already covered general rational numbers (which are represented by either finite or periodic decimal fractions, Sect. 1.2), we only need to consider general irrational values of α which are represented by infinite non-periodic decimal fractions. The main idea here is to set a lower and upper bounds, r_{lower} and r_{upper}, to the irrational number α by two close rational ones: $r_{\text{lower}} < \alpha < r_{\text{upper}}$. For instance, consider α being $\sqrt{2} = 1.1414213562\ldots$. This irrational number can be bound, e.g., by $r_{\text{lower}} = 1.1414$ and $r_{\text{upper}} = 1.1415$. Then, since the function $y = x^r$ with rational r is increasing monotonically, then $x^{r_{\text{lower}}} < x^{r_{\text{upper}}}$. One then *defines* that x^α should lie between these two values to obey the monotonic behaviour: $x^{r_{\text{lower}}} < x^\alpha < x^{r_{\text{upper}}}$, or in the case of the square root of two, we write: $x^{1.1414} < x^{\sqrt{2}} < x^{1.1415}$. Of course, these two bounds are only approximations to the actual value of $\sqrt{2}$, so we could do much better by enclosing it in tighter bounds, e.g. $x^{1.14142} < x^{\sqrt{2}} < x^{1.14143}$, then $x^{1.141421} < x^{\sqrt{2}} < x^{1.141422}$, and so on, i.e. this process can be continued by moving the boundaries on the left and on the right of $\sqrt{2}$ closer to it simultaneously from both directions. This way we create two *numerical sequences* of numbers approaching $\sqrt{2}$ from the left (1.1414, 1.14142, 1.141421, etc.) and from the right (1.1415, 1.14143, 1.141422, etc.). The sequence on the left is constructed by adding every time the next extra digit of $\sqrt{2}$ and hence is increasing gradually approaching $\sqrt{2}$ from below. The sequence on the right is composed also by adding an extra digit at the same position as in the first sequence; however, every time that digit is larger by 1 than in the previous case. Obviously, this sequence is decreasing gradually approaching $\sqrt{2}$ from above. These two sequences which bound $x^{\sqrt{2}}$ on both sides converge to the same limit which, by *definition*, is taken as the exact value of $x^{\sqrt{2}}$. From the practical point of view, bounding $x^{\sqrt{2}}$ on both sides using rational approximations to $\sqrt{2}$ gives a method for calculating $x^{\sqrt{2}}$ with any required precision.

Finally, a real *negative* power is defined as follows: $y = x^{-\alpha} = 1/x^{\alpha}$ with $\alpha > 0$. This way we have defined the general power function $y = x^{\alpha}$ for *any* real α.

Let us now consider properties of the power function. For integer n and m, we obviously have $x^n x^m = x^{n+m}$ and $(x^n)^m = x^{nm}$, due to the definition of the integer power. For the inverse integer powers, we also have $\left(x^{1/n}\right)^{1/m} = x^{1/(nm)}$ since this is the inverse function of the integer power. Indeed, on the one hand, from $(y^n)^m = y^{nm} = x$ follows that $y = x^{1/(nm)}$. On the other hand, from $(y^n)^m = x$ we can first write $y^n = x^{1/m}$ and then $y = \left(x^{1/n}\right)^{1/m}$. This means that $\left(x^{1/n}\right)^{1/m} = x^{1/(nm)}$, as required. Now we can consider the case of both powers being arbitrary rational numbers (n, m, p and q are positive integers):

$$\left(x^{n/m}\right)^{p/q} = \left\{\left[\left(x^{1/m}\right)^n\right]^p\right\}^{1/q} = \left[\left(x^{1/m}\right)^{np}\right]^{1/q} = \left[\left(x^{np}\right)^{1/m}\right]^{1/q} = \left(x^{np}\right)^{1/(mq)} = x^{np/(mq)} ,$$

as required. The other property (when the powers sum up) is proven by the following manipulation:

$$x^{n/m}x^{p/q} = x^{nq/(mq)}x^{pm/(mq)} = \left(x^{1/(mq)}\right)^{nq}\left(x^{1/(mq)}\right)^{pm} = \left(x^{1/(mq)}\right)^{nq+pm}$$
$$= x^{(nq+pm)/(mq)} = x^{(n/m)+(p/q)} .$$

Extension of these properties to arbitrary positive real numbers is achieved using the method of bounding rational sequences as already has been demonstrated above.

Problem 2.16. *Extend the above properties of the powers (multiplication and summation of powers) to all real powers including negative ones.*

Hence, we conclude that for any real α and β:

$$x^{\alpha}x^{\beta} = x^{\alpha\beta} \quad \text{and} \quad \left(x^{\alpha}\right)^{\beta} = x^{\alpha\beta} , \tag{2.18}$$

where generally we must consider $x \geq 0$.

2.3.4 *Number* e

Let us now consider a very famous sequence $x_n = \left(1 + \frac{1}{n}\right)^n$. We shall show that it tends to a limit $e = 2.71828182845\ldots$ (the base of the natural logarithm).[3] To show that this sequence tends to a certain limit, let us first apply to it the binomial formula (1.71):

[3]This is actually an irrational number although we won't be able to demonstrate this here.

$$
\begin{aligned}
x_n &= \left(1+\frac{1}{n}\right)^n = \sum_{k=0}^{n}\binom{n}{k}1^{n-k}\left(\frac{1}{n}\right)^k = \sum_{k=0}^{n}\binom{n}{k}\frac{1}{n^k} \\
&= 1+n\frac{1}{n}+\frac{n(n-1)}{2}\frac{1}{n^2}+\frac{n(n-1)(n-2)}{3!}\frac{1}{n^3}+\cdots \\
&\quad +\frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}\frac{1}{n^k}+\cdots+\frac{n!}{n!}\frac{1}{n^n},
\end{aligned} \tag{2.19}
$$

where we used the fact that the binomial coefficients (1.72) can be written as

$$
\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}
$$

for any $k \geq 1$. Consider now the k-th binomial coefficient above: it has exactly k multipliers in the numerator, so that the k-th term in the expansion (2.19) may be written as follows for any $k \geq 1$:

$$
\binom{n}{k}\frac{1}{n^k} = \frac{1}{k!}\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\left(\frac{n-2}{n}\right)\cdots\left(\frac{n-k+1}{n}\right) = \frac{1}{k!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right),
$$

so that

$$
\begin{aligned}
x_n &= 1+1+\frac{1}{2!}\left(1-\frac{1}{n}\right)+\frac{1}{3!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)+\frac{1}{4!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\left(1-\frac{3}{n}\right)+\cdots \\
&\quad +\frac{1}{k!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)+\cdots+\frac{1}{n!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{n-1}{n}\right),
\end{aligned} \tag{2.20}
$$

where above we have presented the last term of (2.19) also in the same form as the others. Note that there are $n+1$ terms in the expression (2.20) of x_n. Now let us write in the same way x_{n+1} using (2.20):

$$
\begin{aligned}
x_{n+1} &= 1+1+\frac{1}{2!}\left(1-\frac{1}{n+1}\right)+\frac{1}{3!}\left(1-\frac{1}{n+1}\right)\left(1-\frac{2}{n+1}\right) \\
&\quad +\frac{1}{4!}\left(1-\frac{1}{n+1}\right)\left(1-\frac{2}{n+1}\right)\left(1-\frac{3}{n+1}\right)+\cdots \\
&\quad +\frac{1}{k!}\left(1-\frac{1}{n+1}\right)\left(1-\frac{2}{n+1}\right)\cdots\left(1-\frac{k-1}{n+1}\right)+\cdots \\
&\quad +\frac{1}{(n+1)!}\left(1-\frac{1}{n+1}\right)\left(1-\frac{2}{n+1}\right)\cdots\left(1-\frac{n}{n+1}\right).
\end{aligned} \tag{2.21}
$$

This expansion of x_{n+1} has $n+2$ terms. Let us compare the two expansions. It is clear that for any $k \geq 1$

$$1-\frac{k}{n+1} > 1-\frac{k}{n}.$$

Therefore, the third term of x_{n+1} is greater than the third term of x_n, the fourth term of x_{n+1} is greater than the fourth term of x_n, and so on. In addition, the last term in x_{n+1} is absent in the expansion of x_n, but this extra term is obviously positive. Therefore, we strictly have $x_{n+1} > x_n$ for any value of n. This means that our sequence x_n is increasing with n; moreover, we see that $x_n > 2$ (notice $1+1$ as the first two terms in the expansions of x_n or x_{n+1}).

At the same time, the sequence is limited from above. Indeed, since $1-k/n < 1$ for any $k \geq 1$, then

$$x_n < 1+1+\frac{1}{2!}+\frac{1}{3!}+\cdots+\frac{1}{n!}.$$

To estimate the sum of the inverse factorials, we notice that for any $k \geq 1$

$$\frac{1}{k!} = \frac{1}{1\cdot 2\cdot 3\cdot \ldots \cdot k} < \frac{1}{1\cdot 2\cdot 2\cdot \ldots \cdot 2} = \frac{1}{2^{k-1}},$$

so that

$$x_n < 1+1+\frac{1}{2}+\frac{1}{2^2}+\cdots+\frac{1}{2^{n-1}} < 1+\sum_{k=0}^{\infty} 2^{-k},$$

where in the last passage we replaced the final sum of the inverse powers of 2 with their infinite sum, which made the inequality even stronger (as all terms are positive). However, the infinite sum represents an infinite geometric progression which can be calculated as we discussed above by virtue of Eq. (2.7):

$$\sum_{k=0}^{\infty} 2^{-k} = 1+\frac{1}{2}+\frac{1}{2^2}+\cdots+\frac{1}{2^{n-1}}+\cdots = \frac{1}{1-\frac{1}{2}} = 2\,,$$

so that

$$x_n < 1+\sum_{k=1}^{\infty} 2^{-k} = 1+2 = 3.$$

Hence, we showed that $2 < x_n < 3$ and x_n increases with n. Therefore, it approaches a certain limit[4] which is the number e. Hence, we can state that

$$\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = e. \tag{2.22}$$

[4] A numerical sequence which is bounded from below and above by two numbers always has a limit lying between those numbers, see the note at the end of Sect. 2.2.2.

From this it is easy to see that, for any positive integer m,

$$\lim_{n\to\infty}\left(1+\frac{m}{n}\right)^n=\lim_{n\to\infty}\left(1+\frac{1}{n/m}\right)^n=\lim_{n\to\infty}\left[\left(1+\frac{1}{n/m}\right)^{n/m}\right]^m.$$

Here n/m may be either integer or generally rational, and we have made use of the fact that $(x^\alpha)^\beta=x^{\alpha\beta}$. However, when we take the limit $n\to\infty$, we can run n only over a subset of integer numbers which are divisible by m. Then, $n/m=k$ will be an integer, and the limit $n\to\infty$ will be replaced with $k\to\infty$ on the subset (however, the limit should be the same, Theorem 2.4). This way we have another useful result:

$$\lim_{n\to\infty}\left(1+\frac{m}{n}\right)^n=\lim_{k\to\infty}\left[\left(1+\frac{1}{k}\right)^k\right]^m=\left[\lim_{k\to\infty}\left(1+\frac{1}{k}\right)^k\right]^m=e^m. \tag{2.23}$$

Next, let us consider negative $n=-m<0$ (and thus $m>0$):

$$\left(1+\frac{1}{n}\right)^n=\left(1+\frac{1}{-m}\right)^{-m}=\left(\frac{-m+1}{-m}\right)^{-m}=\left(\frac{-m}{-m+1}\right)^m$$
$$=\left(\frac{-m+1-1}{-m+1}\right)^m=\left(1+\frac{1}{m-1}\right)^m=\left(1+\frac{1}{m-1}\right)^{m-1}\left(1+\frac{1}{m-1}\right).$$

Consider now the limit of $m\to\infty$, which corresponds to the limit of $n\to-\infty$. We immediately see that

$$\lim_{m\to\infty}\left(1+\frac{1}{m-1}\right)^{m-1}\left(1+\frac{1}{m-1}\right)=\lim_{m\to\infty}\left(1+\frac{1}{m}\right)^m\left(1+\frac{1}{m-1}\right)=e,$$

since $\left(1+\frac{1}{m-1}\right)\to 1$. Therefore, we obtain that

$$\lim_{n\to-\infty}\left(1+\frac{1}{n}\right)^n=e. \tag{2.24}$$

Problem 2.17. *Prove that*

$$\lim_{n\to\infty}\left(1-\frac{m}{n}\right)^n=e^{-m}. \tag{2.25}$$

2.3.5 Exponential Function

Consider a general definition of $y = a^x$ for any real positive a (and avoiding the trivial case of $a = 1$) and arbitrary x from $\mathbb{R}$. The power of a for arbitrary x is defined as above by first considering rational x and then using the method of bounding numerical sequences to extend these definitions to all real numbers $-\infty < x < \infty$. It is important in this respect to establish first that the exponential function is monotonic.

Indeed, consider the case of $a > 1$. First we prove that $a^\lambda > 1$ for any positive λ. This follows immediately for $\lambda = n$ being an integer. We also have that $a^{1/m} > 1$ for any positive integer m. Indeed, assume the opposite, i.e. $a^{1/m} < 1$, i.e. $a^{1/m} = 1 - \alpha$ with some $0 < \alpha < 1$; this means that $a = (1-\alpha)^m$. But the power function is increasing monotonically, i.e. $a = (1-\alpha)^m < 1^m = 1$ as $1 - \alpha < 1$; hence, we arrive at $a < 1$ which contradicts our assumption of $a > 1$, and that proves the statement made above. Now it is easy to see that $a^\lambda > 1$ for any rational $\lambda = n/m$. Indeed, $a^{n/m} = (a^n)^{1/m} > 1$ since $a^n > 1$. The proof for any positive real λ is extended by applying the bounding numerical sequences.

Now returning back to the exponential function, we write for $x_2 > x_1$:

$$a^{x_2} - a^{x_1} = a^{x_1}\left(a^{x_2 - x_1} - 1\right) > 0,$$

since $a^\lambda > 1$ for any positive λ as was shown above. This means that indeed the exponential function $y = a^x$ is monotonically increasing for $a > 1$.

Problem 2.18. *Prove that the exponential function for $0 < a < 1$ is monotonically decreasing.*

If, as the base of the power, the number e introduced in Sect. 2.2 is used, then we have the function $y = e^x$.

Due to the properties of powers, Eq. (2.18), we also have:

$$a^{x_1} a^{x_2} = a^{x_1 + x_2} \quad \text{and} \quad \left(a^{x_1}\right)^{x_2} = a^{x_1 x_2}. \tag{2.26}$$

Exponential functions $y = a^x$ for selected values of the base a are shown in Fig. 2.4(b). We see that for $0 < a < 1$ the exponential function monotonically decreases and becomes more and more horizontal as a approaches the value of 1.0 when it becomes indeed horizontal (and hence a constant); then for $a > 1$, the function monotonically increases; the larger the value of a, the steeper the increase.

2.3.6 Hyperbolic Functions

These are combinations of exponential functions:

$$\sinh(x) = \frac{1}{2}\left(e^x - e^{-x}\right); \tag{2.27}$$

$$\cosh(x) = \frac{1}{2}\left(e^x + e^{-x}\right); \tag{2.28}$$

$$\tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}}; \tag{2.29}$$

$$\coth(x) = \frac{\cosh(x)}{\sinh(x)} = \frac{e^x + e^{-x}}{e^x - e^{-x}}. \tag{2.30}$$

It can be shown that these functions are closely related to trigonometric functions $\sin(x)$, $\cos(x)$, $\tan(x)$ and $\cot(x)$, and are called, respectively, hyperbolic sine, cosine, tangent and cotangent functions.

Problem 2.19. *Using the definitions of the hyperbolic functions given above, prove the following identities:*

$$\cosh^2(x) - \sinh^2(x) = 1, \tag{2.31}$$

$$\sinh(2x) = 2\sinh(x)\cosh(x), \tag{2.32}$$

$$\cosh(2x) = \cosh^2(x) + \sinh^2(x), \tag{2.33}$$

$$\sinh(x+y) = \sinh(x)\cosh(y) + \cosh(x)\sinh(y). \tag{2.34}$$

These identities have close analogues in trigonometric functions; for instance, the first one, Eq. (2.31), is analogous to (1.9), while the others are similar to the corresponding identities of the sine and cosine (see below).

One can also define inverse hyperbolic functions. To define those, we need to consider the logarithm first.

2.3.7 Logarithmic Function

The logarithmic function is defined as an inverse to the exponential function: from $y = a^x$ we write $x = \log_a y$, or using conventional notations, $y = \log_a x$. This function is called the logarithm of x to base a. From the properties of the exponential function, we can easily deduce the properties of the logarithmic function. First, if $y_1 = a^{x_1}$ and $y_2 = a^{x_2}$, then

$$y_1 y_2 = a^{x_1} a^{x_2} = a^{x_1+x_2} = y \quad \Rightarrow \quad \log_a y = \log_a (y_1 y_2) = x_1+x_2 = \log_a y_1 + \log_a y_2 . \tag{2.35}$$

Then, from the definition of the logarithmic function, it follows:

$$a^{\log_a x} = x. \tag{2.36}$$

Raise both sides of this equality into power z:

$$\left(a^{\log_a x}\right)^z = a^{z \log_a x} = x^z \quad \Rightarrow \quad z \log_a x = \log_a (x^z) . \tag{2.37}$$

Finally, we write

$$x = a^{\log_a x} = \left(b^{\log_b a}\right)^{\log_a x} = b^{\log_a x \cdot \log_b a},$$

from which it follows that

$$\log_b x = \log_a x \cdot \log_b a, \tag{2.38}$$

which is the formula which connects different bases of the logarithmic function.

If $a = e$ is used as the base, then the logarithmic function is denoted simply $\ln x = \log_e x$ and is called *natural logarithm*. This one is most frequently used. The natural logarithm $y = \ln x$ and the exponential function $y = e^x$, which is the inverse to it, are compared in Fig. 2.5. As expected, the graphs are symmetric with respect to the dashed line which bisects the first quadrant.

Problem 2.20. *Show that the inverse hyperbolic cosine function accepts two branches given by:*

$$\cosh^{-1}(x) = \ln\left(x \pm \sqrt{x^2-1}\right), \quad x \geq 1. \tag{2.39}$$

Normally the plus sign is used leading to positive values of the function for any $x \geq 1$. *Explain the reason behind these two possibilities. Sketch both branches of the function. Show that other inverse hyperbolic functions are single-valued and given by the following relations:*

$$\sinh^{-1}(x) = \ln\left(x + \sqrt{x^2+1}\right), \quad -\infty < x < \infty, \tag{2.40}$$

(continued)

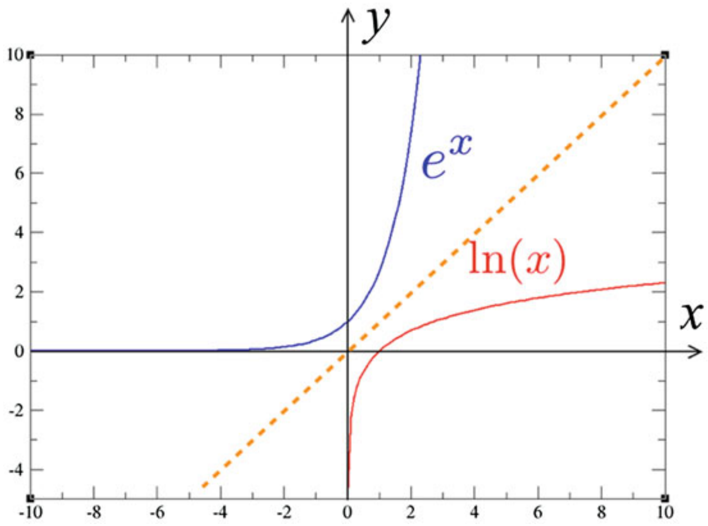


Fig. 2.5 Mutually inverse $y = e^x$ and $y = \ln x$ functions

Problem 2.20 (continued)

$$\tanh^{-1}(x) = \frac{1}{2}\ln\frac{1+x}{1-x}, \quad |x| < 1. \tag{2.41}$$

$$\coth^{-1}(x) = \frac{1}{2}\ln\frac{x+1}{x-1}, \quad |x| > 1. \tag{2.42}$$

Using the fact that the logarithmic function is only defined for its positive argument, explain the chosen x intervals in each case.

2.3.8 Trigonometric Functions

We have already defined sine and cosine functions in Sect. 1.5 and considered some of their properties. In particular, we found that these two functions are periodic with the period of 2π. The two functions are plotted in Fig. 2.6(a). Basically, the sine function is the cosine function shifted by $\pi/2$ to the right. Both functions are periodic with the period of 2π. The tangent, $y = \tan x$, and cotangent, $y = \cot x$, functions are shown in Fig. 2.6(b). They both have the period of π and are singular at some periodically repeated points: $\tan(x)$ is singular when $\cos x = 0$ ($x = \frac{\pi}{2} + \pi n$ with $n = 0, \pm 1, \pm 2, \ldots$), while $\cot x$ is singular when $\sin x = 0$ (which happens at $x = \pi n$ with $n = 0, \pm 1, \pm 2, \ldots$).

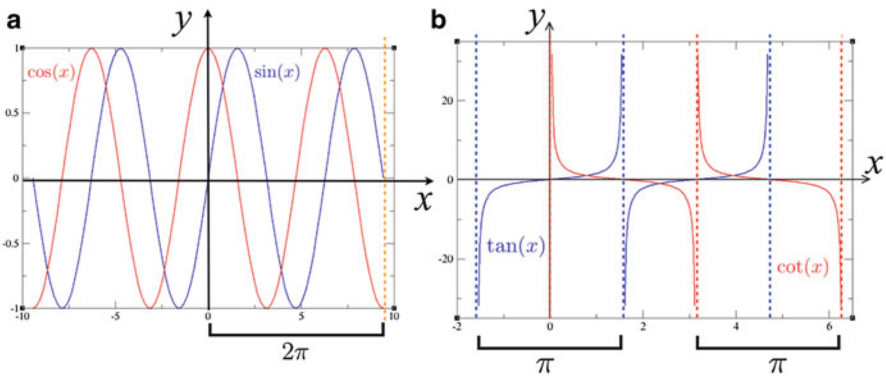
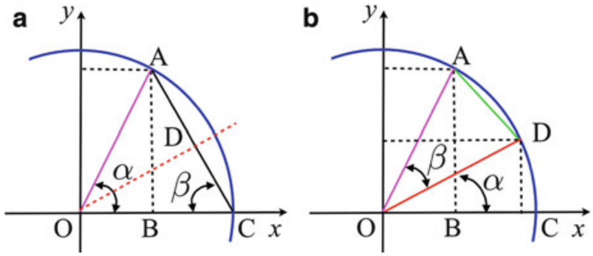


Fig. 2.6 (**a**) Functions $y = \sin(x)$ and $y = \cos(x)$ with the period of 2π; (**b**) functions $y = \tan x$ and $y = \cot x$ with the period of π. The period in each case is indicated at the *bottom* of the graphs

Fig. 2.7 Illustration for the proof of double angle (**a**) and sum-of-angles (**b**) trigonometric identities



We shall now consider other important properties of these functions.

First, we shall prove double angle identities. Consider a circle of unit radius as shown in Fig. 2.7(a). The triangle $\triangle AOC$ is an isosceles triangle since it has two its sides, OC and OA, equal (in fact, equal to 1), hence two angles $\angle OAC = \beta$ and $\angle OCA$ are also equal, and since all three angles of the triangle sum up to π, we have $2\beta + \alpha = \pi$. Also, the height of the triangle $AB = \sin\alpha$. Then, draw the bisector of the $\angle AOC$ which crosses AC at the point D; then, $AC = 2DC = 2\sin\frac{\alpha}{2}$. On the other hand,

$$AC = \frac{AB}{\sin\beta} = \frac{\sin\alpha}{\sin\beta} = \frac{\sin\alpha}{\sin\left(\frac{\pi}{2} - \frac{\alpha}{2}\right)} = \frac{\sin\alpha}{\cos\frac{\alpha}{2}},$$

which is equal to $2\sin\frac{\alpha}{2}$ if only $\sin\alpha = 2\sin\frac{\alpha}{2}\cos\frac{\alpha}{2}$, or, written in a more conventional form:

$$\sin(2\alpha) = 2\sin\alpha\cos\alpha. \tag{2.43}$$

This one can be compared with an analogous identity (2.32) we discussed above for the hyperbolic functions.

Similarly, looking again at Fig. 2.7(a), we can write:

$$OB = \cos\alpha = 1 - BC = 1 - AC\cos\beta = 1 - \frac{\sin\alpha}{\cos\frac{\alpha}{2}}\cos\left(\frac{\pi}{2} - \frac{\alpha}{2}\right)$$
$$= 1 - \frac{\sin\alpha}{\cos\frac{\alpha}{2}}\sin\frac{\alpha}{2} = 1 - 2\sin^2\frac{\alpha}{2},$$

where we decomposed $\sin\alpha$ using (2.43). Written in a more conventional form:

$$\cos(2\alpha) = 1 - 2\sin^2\alpha = \cos^2\alpha - \sin^2\alpha = -1 + 2\cos^2\alpha, \tag{2.44}$$

where we have also used Eq. (1.9). The previous formulae enable one to express the sine and cosine function of the angle α via sine (cosine) functions of either double or half angle, 2α or $\alpha/2$. Formula (2.33) for the hyperbolic functions is an analog of this identity.

Next, we consider a more general formula for a sum of two angles. Referring this time to Fig. 2.7(b), we write a square of AD in two different ways. The coordinates of points A and D obviously are: $A\left(\cos(\alpha+\beta), \sin(\alpha+\beta)\right)$ and $D(\cos\alpha, \sin\alpha)$, and hence

$$AD^2 = (\cos(\alpha+\beta) - \cos\alpha)^2 + (\sin(\alpha+\beta) - \sin\alpha)^2$$
$$= 2 - 2\left[\cos(\alpha+\beta)\cos\alpha + \sin(\alpha+\beta)\sin\alpha\right]$$

after opening the squares and using Eq. (1.9). At the same time, similarly to the previous case, $AD = 2\sin\frac{\beta}{2}$, and hence we can write:

$$2-2\left[\cos(\alpha+\beta)\cos\alpha+\sin(\alpha+\beta)\sin\alpha\right] = \left(2\sin\frac{\beta}{2}\right)^2 = 4\sin^2\frac{\beta}{2} = 2(1-\cos\beta),$$

after transforming the right-hand side using the double-angle formula (2.44). Opening the brackets and replacing $\gamma = \alpha + \beta$, we obtain:

$$\cos(\gamma - \alpha) = \cos\gamma\cos\alpha + \sin\gamma\sin\alpha. \tag{2.45}$$

Consequently, replacing α with $-\alpha$ in the above formula and recalling that sine and cosine functions are odd and even functions, respectively, we obtain:

$$\cos(\gamma + \alpha) = \cos\gamma\cos\alpha - \sin\gamma\sin\alpha. \tag{2.46}$$

Problem 2.21. *Using one of the equations above, prove similar identities for the sine function as well:*

$$\sin(\gamma \pm \alpha) = \sin\gamma\cos\alpha \pm \cos\gamma\sin\alpha. \tag{2.47}$$

Problem 2.22. *By adding (subtracting) Eqs. (2.45)–(2.47) and changing notations for the angles, derive the inverse relations:*

$$\sin\alpha + \sin\beta = 2\sin\left(\frac{\alpha+\beta}{2}\right)\cos\left(\frac{\alpha-\beta}{2}\right); \tag{2.48}$$

$$\sin\alpha - \sin\beta = 2\cos\left(\frac{\alpha+\beta}{2}\right)\sin\left(\frac{\alpha-\beta}{2}\right); \tag{2.49}$$

$$\cos\alpha + \cos\beta = 2\cos\left(\frac{\alpha+\beta}{2}\right)\cos\left(\frac{\alpha-\beta}{2}\right); \tag{2.50}$$

$$\cos\alpha - \cos\beta = -2\sin\left(\frac{\alpha+\beta}{2}\right)\sin\left(\frac{\alpha-\beta}{2}\right). \tag{2.51}$$

Problem 2.23. *Rearrange the above identities to get the following expansions of a product of the sine and cosine functions:*

$$\sin\alpha\sin\beta = \frac{1}{2}\left[\cos(\alpha-\beta) - \cos(\alpha+\beta)\right]; \tag{2.52}$$

$$\sin\alpha\cos\beta = \frac{1}{2}\left[\sin(\alpha+\beta) + \sin(\alpha-\beta)\right]; \tag{2.53}$$

$$\cos\alpha\sin\beta = \frac{1}{2}\left[\sin(\alpha+\beta) - \sin(\alpha-\beta)\right]; \tag{2.54}$$

$$\cos\alpha\cos\beta = \frac{1}{2}\left[\cos(\alpha-\beta) + \cos(\alpha+\beta)\right]. \tag{2.55}$$

Using these identities, one can express trigonometric functions of multiple angles via those of only a single one. This may be useful in calculating integrals as demonstrated in later chapters. For instance, using Eqs. (2.47) and (2.46), as well as the particular cases of these for equal angles $\gamma = \alpha$, we obtain:

$$\begin{aligned}\sin(3\alpha) &= \sin(2\alpha+\alpha) = \sin(2\alpha)\cos\alpha + \cos(2\alpha)\sin\alpha \\ &= 2\sin\alpha\cos^2\alpha + \cos^2\alpha\sin\alpha - \sin^3\alpha = 3\sin\alpha\cos^2\alpha - \sin^3\alpha,\end{aligned} \tag{2.56}$$

$$\begin{aligned}\cos(3\alpha) &= \cos(2\alpha+\alpha) = \cos(2\alpha)\cos\alpha - \sin(2\alpha)\sin\alpha \\ &= \cos^3\alpha - \sin^2\alpha\cos\alpha - 2\sin^2\alpha\cos\alpha = \cos^3\alpha - 3\sin^2\alpha\cos\alpha.\end{aligned} \tag{2.57}$$

Problem 2.24. *Prove the following identities using the above method:*

$$\sin(4\alpha) = 4\sin\alpha\cos^3\alpha - 4\sin^3\alpha\cos\alpha; \tag{2.58}$$

$$\cos(4\alpha) = \cos^4\alpha - 6\sin^2\alpha\cos^2\alpha + \sin^4\alpha; \tag{2.59}$$

[There is also a simpler way for both above.]

$$\sin(5\alpha) = \sin^5\alpha - 10\sin^3\alpha\cos^2\alpha + 5\sin\alpha\cos^4\alpha; \tag{2.60}$$

$$\cos(5\alpha) = \cos^5\alpha - 10\sin^2\alpha\cos^3\alpha + 5\sin^4\alpha\cos\alpha. \tag{2.61}$$

Problem 2.25. *Using the following identity,*

$$\sin(k\Delta) = \frac{1}{2\sin\frac{\Delta}{2}}\left\{\cos\left[\left(k-\frac{1}{2}\right)\Delta\right] - \cos\left[\left(k+\frac{1}{2}\right)\Delta\right]\right\}, \tag{2.62}$$

which follows from Eq. (2.51), derive the formula below for the finite sum of sine functions:

$$\sum_{k=1}^{n}\sin(\Delta k) = \frac{1}{\sin\frac{\Delta}{2}}\sin\frac{(n+1)\Delta}{2}\sin\frac{n\Delta}{2}. \tag{2.63}$$

[*Hint*: use (2.62) for the sine functions in the sum and then observe that only the first and the last terms remain, all other terms cancel out exactly.]

Problem 2.26. *Prove the same type of identity for the cosine functions:*

$$\sum_{k=1}^{n}\cos(\Delta k) = \frac{1}{2\sin\frac{\Delta}{2}}\left[\sin\left(\Delta\left(n+\frac{1}{2}\right)\right) - \sin\frac{\Delta}{2}\right] = \frac{1}{\sin\frac{\Delta}{2}}\cos\frac{(n+1)\Delta}{2}\sin\frac{n\Delta}{2}. \tag{2.64}$$

Problem 2.27. *When studying integrals containing a rational function of sine and cosine, a special substitution replacing both of them with a tangent function is frequently found useful. Prove, using relationships between the sine and cosine functions of angles* α *and* $\alpha/2$*, the following identities:*

$$\sin\alpha = \frac{2\tan\frac{\alpha}{2}}{1+\tan^2\frac{\alpha}{2}} \quad \textit{and} \quad \cos\alpha = \frac{1-\tan^2\frac{\alpha}{2}}{1+\tan^2\frac{\alpha}{2}}. \tag{2.65}$$

Problem 2.28. *Prove the following identity:*

$$\cot(x+y) - \cot(x-y) = \frac{2\sin 2y}{\cos 2x - \cos 2y}. \tag{2.66}$$

Problem 2.29. *Show using formula (1.50) that when we multiply two complex numbers, their absolute values multiply and the phases sum up.*

Problem 2.30. *Similarly, show that when we divide two complex numbers, their absolute values are divided and the phases are subtracted.*

Problem 2.31. *Correspondingly, derive the following de Moivre's formula:*

$$(\cos\phi + i\sin\phi)^n = \cos(n\phi) + i\sin(n\phi).$$

2.3.9 Inverse Trigonometric Functions

For each of the four trigonometric functions, $\sin x$, $\cos x$, $\tan x$ and $\cot x$, one can define their corresponding inverse functions, denoted, respectively, as $\arcsin x$, $\arccos x$, $\arctan x$ and $\text{arccot}\, x$. These functions have the meaning inverse to that of the direct functions. For instance, $y = \arcsin x$ corresponds to $\sin y = x$, i.e. $y = \arcsin x$ is the angle sine of which is equal to x.

Because of the periodicity of the direct functions, there are different values of y (the angles) possible that give the same value of the functions x. Indeed, consider, e.g., the sine function and its inverse. Then, for all angles given by $y = y_0 + 2\pi n$ with n from all integer numbers $\mathbb{Z}$ the same value of the sine function $x = \sin y = \sin y_0$ is obtained. This means that the inverse function is not well defined as there exist different values of the function $y = \arcsin x$ for the same value of its argument, x. This difficulty is similar to the one we came across before in Sect. 2.1 when considering the inverse of the $y = x^2$ function, and we realised there that by restricting the values of the argument of the direct function, y, the one-to-one correspondence can be established. The same method will be used here.

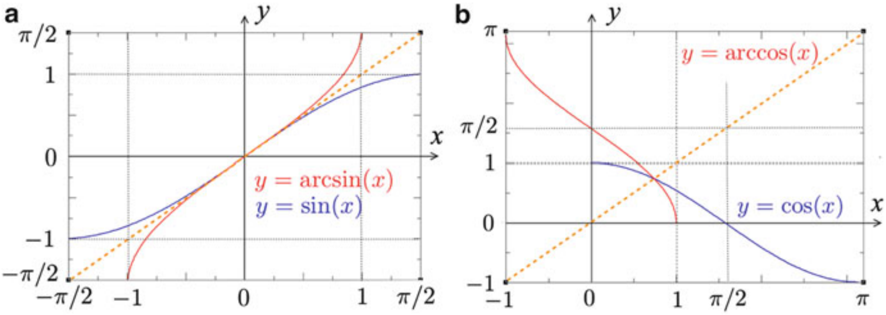


Fig. 2.8 Functions (**a**) $y = \sin x$, $y = \arcsin x$ and (**b**) $y = \cos x$, $y = \arccos x$

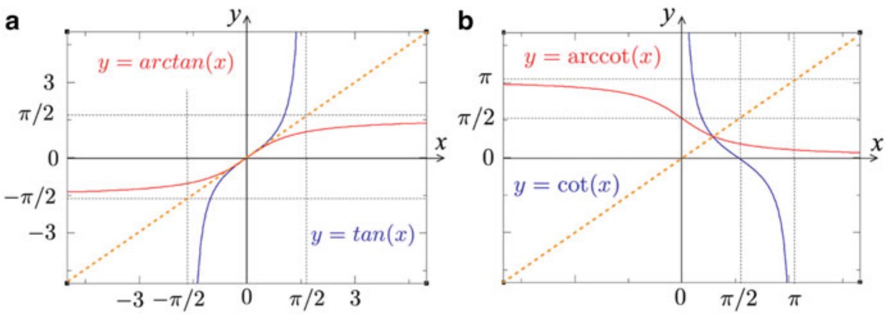


Fig. 2.9 Functions (**a**) $y = \tan x$, $y = \arctan x$ and (**b**) $y = \cot x$, $y = \mathrm{arccot}\, x$

Since the sine function varies between -1 and 1 while the angles run between $-\pi/2$ and $\pi/2$, the arcsine is defined as follows: it gives an angle $-\pi/2 \le y \le \pi/2$ such that sine of it is equal to x lying in the interval $-1 \le x \le 1$. The graphs of both functions are shown in Fig. 2.8(a). One can see that they are indeed symmetrical with respect to the diagonal of the I and III quadrants as we would expect (see the general discussion in Sect. 2.1).

Similarly, since the cosine function varies between -1 and 1 when the angle changes between 0 and π, the arccosine function and its inverse are well defined within these intervals. Namely, $y = \arccos x$ is defined for $-1 \le x \le 1$ and $0 \le y \le \pi$. Both functions are drawn in Fig. 2.8(b).

Finally, functions $y = \arctan x$ and $y = \mathrm{arccot}\, x$ are both defined for $-\infty < x < \infty$, and the values of the arctangent lie within the interval $-\pi/2 < y < \pi/2$, while the values of the arccotangent—in the interval $0 < y < \pi$. The graphs of these two functions together with the corresponding direct functions are shown in Fig. 2.9.

The inverse trigonometric functions are useful in writing general solutions of trigonometric equations. For instance, equation $\sin x = a$ (with $-1 \le a \le 1$) has the general solution $x = \arcsin a + 2\pi n$, where $n = 0, \pm 1, \pm 2, \ldots$ is from $\mathbb{Z}$.

Let us now establish main properties of the inverse trigonometric functions. Obviously, it follows from their definitions that

$$\sin(\arcsin x) = x, \ \cos(\arccos x) = x \text{ and } \tan(\arctan x) = x. \tag{2.67}$$

Other properties are all derived from their definitions, i.e. from the properties of the corresponding direct functions. Consider, for instance, the sine function: $y = \sin(-x) = -\sin x$, from where it follows that, on the one hand, $-x = \arcsin y$, but on the other, $\arcsin(-y) = x$. This means that the arcsine function is antisymmetric (odd):

$$\arcsin(-x) = -\arcsin x. \tag{2.68}$$

Problem 2.32. *Using the properties of the trigonometric functions, prove the following properties of the inverse trigonometric functions:*

$$\arctan(-x) = -\arctan x, \tag{2.69}$$

$$\operatorname{arccot}(-x) = \pi - \operatorname{arccot} x, \tag{2.70}$$

$$\arccos(-x) = \pi - \arccos x, \tag{2.71}$$

$$\arcsin x + \arccos x = \frac{\pi}{2}, \tag{2.72}$$

$$\arctan x + \operatorname{arccot} x = \frac{\pi}{2}, \tag{2.73}$$

$$\sin(\arccos x) = \cos(\arcsin x) = \sqrt{1-x^2}, \tag{2.74}$$

$$\arctan\frac{1}{x} = \operatorname{arccot} x = \arcsin\frac{1}{\sqrt{1+x^2}}, \tag{2.75}$$

$$\cos(\arctan x) = \frac{1}{\sqrt{1+x^2}}. \tag{2.76}$$

2.4 Limit of a Function

2.4.1 Definitions

Limit of a function is one of the fundamental notions in mathematics. Several different cases need to be considered.

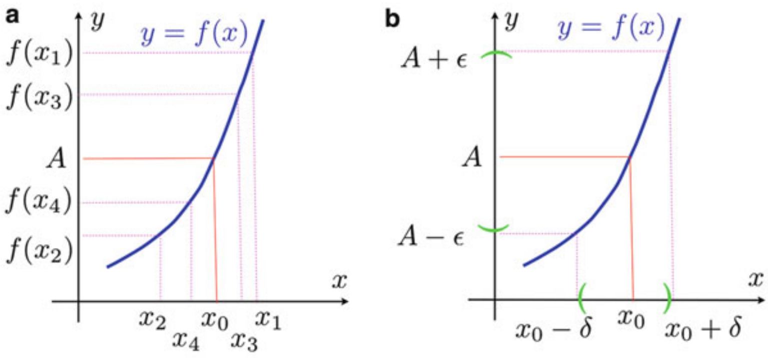


Fig. 2.10 Illustration of two equivalent definitions of the limit of a function $y = f(x)$ at the point x_0: (**a**) via an infinite numerical sequence $x_1, x_2, x_3, \ldots$ converging at x_0 and (**b**) using ϵ and δ

2.4.1.1 Limit of a Function at a Finite Point

We shall start by considering two definitions of a finite limit of a function at a point $-\infty < x < \infty$. Both definitions are completely equivalent although they sound somewhat different.

Let $f(x)$ be a function of a real variable x defined on some part of the real axis.

Definition 2.1. of the limit of a function $f(x)$ at x tending to x_0 (written: $x \to x_0$).[5] If we take a numerical sequence of numbers x_n which is different from x_0 and converges to x_0, then it generates another sequence, a sequence of the values of the function,

$$f(x_1), f(x_2), f(x_3), \ldots, f(x_n), \ldots, \tag{2.77}$$

as illustrated in Fig. 2.10(a). Then A is the limit of the function $f(x)$, i.e.

$$\lim_{x \to x_0} f(x) = A,$$

if *arbitrary* sequence of numbers converging at x_0 generates the functional sequence (2.77) converging *to the same* limit A.

Note that each sequence, if it has a limit, will converge to a unique limit (Sect. 2.2); however, what we are saying here is that if *any* sequence, converging at x_0, results in the same limit A of the generated functional sequence, only then the function $f(x)$ does have a limit at x_0. It follows from here that the function $f(x)$, if it has a limit at x_0, may only have *one* limit.

---

[5]Due to Heinrich Eduard Heine.

As an example, consider a function $f(x) = x$. If we take an arbitrary sequence x_n of its argument x which converges at x_0, then the corresponding functional sequence (2.77) would coincide with the sequence x_n since $f(x) = x$, and hence the functional sequence for any sequence x_n will converge to the same limit x_0.

Take now the parabolic function $f(x) = x^2$. The corresponding functional sequence becomes x_1^2, x_2^2, x_3^2, etc. This sequence can be considered as a product of two identical sequences x_n, and it tends to the limit x_0^2 (Theorem 2.2 in Sect. 2.2), where x_0 is the limit of the sequence x_n. So, if we consider an arbitrary sequence x_n converging to x_0, the functional sequence would have a limit which is x_0^2. Hence, the function $f(x) = x^2$ has a limit at x_0, equal to x_0^2.

For our next example we consider the function $f(x) = \sin(1/x)$ which is defined everywhere except $x = 0$. Consider a specific numerical sequence $x_n = 1/(n\pi)$, where $n = 1, 2, 3, \ldots$. Obviously, the sequence tends to 0 as $n \to \infty$, and hence the functional sequence $f(x_n) = \sin(\pi n) = 0$ results in the constant sequence of zeros, converging to zero. One may think then that our function $f(x)$ tends to zero when $x \to 0$. However, this function in fact does not have a limit, and to prove that it is sufficient to build at least one example of a sequence x_n, also converging to zero, which generates the corresponding functional sequence converging to a *different* limit. And indeed, in this case this is easy to do. Consider the second sequence, $x_n = 1/(\pi/2 + 2\pi n)$, which is obviously converges to zero as well. However, generated by it functional sequence $f(x_n) = \sin(\pi/2 + 2\pi n)$ converges to 1 instead. Hence, the function $f(x) = \sin(1/x)$ does not have a limit at $x = 0$ (however, it does have a well-defined limit at any other point).

Now we shall give another, equivalent, definition of the limit of a function.

Definition 2.2. of the limit of a function $f(x)$ at x_0.[6] The limit of the function $f(x)$ is A when $x \to x_0$, if for any $\epsilon > 0$ one can indicate a vicinity of x_0 such that for any x from that vicinity $f(x)$ is different from A by no more than ϵ. Rephrased in a more formal way: for any $\epsilon > 0$ one can find such $\delta > 0$, so that for any x satisfying $|x - x_0| < \delta$ it follows $|f(x) - A| < \epsilon$. This is still exactly the same formulation: $|x - x_0| < \delta$ means that x belongs to some δ-vicinity of x_0, i.e. $x_0 - \delta < x < x_0 + \delta$, and $|f(x) - A| < \epsilon$ similarly means that $f(x)$ is in the vicinity of A by no more than ϵ as illustrated in Fig. 2.10(b). In other words, no matter how close we would like to have $f(x)$ to A, we shall always find the corresponding values of x close to x_0 for which this condition will be satisfied. This second definition is sometimes called "$\epsilon - \delta$" definition of the limit of a function or the definition in terms of intervals. Note that $\delta = \delta(\epsilon)$ is a function of ϵ, i.e. it depends on the choice of ϵ.

As we said before, the two definitions are equivalent.[7] Let us now illustrate the second definition on some examples.

[6]Due to Augustin-Louis Cauchy.

[7]We shall not prove their equivalence here.

Consider first the limit of the function $f(x) = x$ at x_0. Choose some positive ϵ; then, by taking $\delta = \epsilon$, we immediately see that for any x satisfying $|x - x_0| < \delta$ we also have $|f(x) - x_0| = |x - x_0| < \delta = \epsilon$, i.e. $f(x)$ has the limit of x_0 at this point.

Consider now the limit of $f(x) = x^2$ at x_0. Let us prove that the limit exists and it is equal to x_0^2. Indeed, consider some vicinity $|x - x_0| < \delta$ of x_0. Then, for all x in that vicinity, we can estimate:

$$\left|f(x)-x_0^2\right| = \left|x^2-x_0^2\right| = |x-x_0|\,|(x-x_0)+2x_0| \leq \delta\,(|x-x_0|+2\,|x_0|) \leq \delta\,(\delta+2\,|x_0|)\,.$$

If we now take δ as a solution of the equation $\epsilon = \delta\,(\delta + 2\,|x_0|)$ and choose its positive root, $\delta = -|x_0| + \sqrt{|x_0|^2 + \epsilon}$, then we can say that whatever value of ϵ we take, we can always find δ, e.g. the one given above, that the required condition $\left|x^2 - x_0^2\right| < \epsilon$ is indeed satisfied for any x from the δ-vicinity of x_0. So, $f(x) = x^2$ has the limit x_0^2 when $x \to x_0$.

Next, let us consider $f(x) = \sin x$. We would like to prove that it tends to $\sin x_0$ when $x \to x_0$. First of all, let us make an estimate, assuming that $|x - x_0| < \delta$. We can write:

$$\sin x - \sin x_0 = 2\sin\frac{x - x_0}{2}\cos\frac{x + x_0}{2},$$

so that

$$|\sin x - \sin x_0| = 2\left|\sin\frac{x-x_0}{2}\right|\cdot\left|\cos\frac{x+x_0}{2}\right| < 2\left|\frac{x-x_0}{2}\right|\cdot\left|\cos\frac{x+x_0}{2}\right| < |x-x_0| < \delta,$$

where we used the first part of the inequality (1.93) and the fact that the absolute value of the cosine is not larger than 1. The result we have just obtained clearly shows that for any ϵ there exists such $\delta = \epsilon$, so that from $|x - x_0| < \delta$ immediately follows $|\sin x - \sin x_0| < \epsilon$, which is exactly what was required.

Finally, let us prove that

$$\lim_{x\to 0} a^x = 1 \text{ for } a > 1. \tag{2.78}$$

Here $x_0 = 0$, i.e. we need to prove that from $|x| < \delta$ follows $|a^x - 1| < \epsilon$, or that

$$-\epsilon + 1 < a^x < \epsilon + 1. \tag{2.79}$$

Since $a^x > 0$ for all values of x, it is sufficient to consider $0 < \epsilon < 1$. Then, as the logarithm is a monotonically increasing function, the inequality (2.79) can also be written as $\log_a(1-\epsilon) < x < \log_a(1+\epsilon)$ (note that $\log_a a^x = x$), where the left boundary, $\log_a(1-\epsilon)$, is negative (since $0 < 1-\epsilon < 1$), while the right one, $\log_a(1+\epsilon)$, is positive ($1+\epsilon > 1$). We see that if we take δ as the smallest absolute value of the two boundaries,

$$\delta = \min\left\{\left|\log_a(1-\epsilon)\right|, \left|\log_a(1+\epsilon)\right|\right\},$$

then we obtain $-\delta < x < \delta$, i.e. $\delta = \delta(\epsilon)$ can be found from the equation above. This proves the limit (2.78).

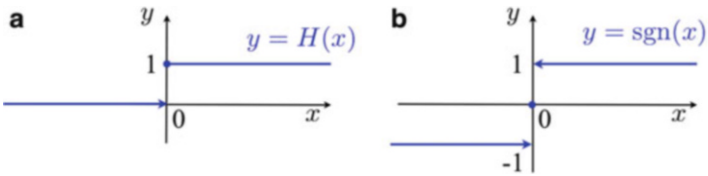


Fig. 2.11 Heaviside (**a**) and sign (**b**) functions experience a jump at $x = 0$. They do not have a certain limit at this point, although both left and right limits at $x = 0$ exist in both cases. The *arrows* on some lines indicate that the lines just approach the point $x = 0$, but never reach it. The *fat dot* indicates the value of the function at $x = 0$ according to Eqs. (2.1) and (2.2)

> **Problem 2.33.** *Prove, using Eq. (2.78), that* $\lim_{x\to x_0} a^x = a^{x_0}$.
>
> **Problem 2.34.** *Prove using the definition by intervals that* $\lim_{x\to x_0} \sqrt{x} = \sqrt{x_0}$ *for any* $x_0 > 0$.

2.4.1.2 One-Side Limits

Some functions may experience a jump at point x_0. For instance, the Heaviside function $H(x)$, defined in Eq. (2.1), makes a jump at $x = 0$ from zero to one, see Fig. 2.11(a). Similarly, the sign function, Eq. (2.2), experiences a jump as well, see Fig. 2.11(b). For these and other cases it is useful to introduce *one-side limits*, when x tends to x_0 either from the left, being always smaller than x_0 (written: $x \to x_0 - 0$ or $x \to x_0^-$ or $x \to -x_0$), or from the right, when $x > x_0$ (written: $x \to x_0 + 0$ or $x \to x_0^+$ or $x \to +x_0$). Both definitions of the limit are easily extended to this case. For instance, using the language of intervals of the "$\epsilon - \delta$" definition, the fact that the left limit of $f(x)$ at x_0 is equal to A (written: $\lim_{x\to x_0-0} f(x) = A$) can be formulated as follows: for each $\epsilon > 0$ there exists $\delta > 0$ such that $x_0 - \delta < x < x_0$ implies $|f(x) - A| < \epsilon$.

> **Problem 2.35.** *Give the "$\epsilon - \delta$" definition of the right limit,* $\lim_{x\to x_0^+} f(x) = A$.

Both the left and right limits are illustrated in Fig. 2.12. It is clear that if the limit of $y = f(x)$ exists at x_0 and is equal to A, then both the left and right limits exist at this point as well and both are equal to the same value A. Inversely, if both the left and right limits exist and both are equal to A, then there exists the limit of $f(x)$ at x_0 also equal to A. If, however, at point x the left and right limits exist, but are different,

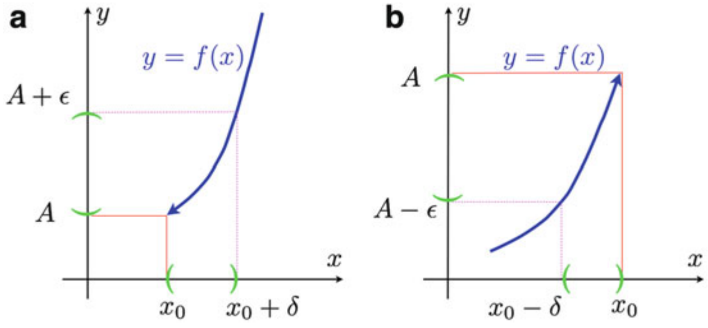


Fig. 2.12 Right (**a**) and left (**b**) limits of a function $y = f(x)$ at $x = x_0$ using the language of intervals

then at this point the limit of the function $f(x)$ does not exist. The functions $H(x)$ and sgn (x) shown in Fig. 2.11 do not have a limit at $x = 0$, although both their left and right limits exist.

As an example, let us prove that $\lim_{x\to 0^+} \sqrt{x} = 0$. This function is only defined for $x \geq 0$ and hence we can only consider the right limit. We need to prove that for any $\epsilon > 0$ one can always find such $\delta > 0$ that $0 < x < \delta$ implies $\left|\sqrt{x} - 0\right| = \sqrt{x} < \epsilon$. It is clear that this inequality is indeed fulfilled if $\delta = \epsilon^2$.

Problem 2.36. *Prove that the left and right limits of the Heaviside function around* $x = 0$ *are 0 and 1, respectively.*

Problem 2.37. *Prove that the left and right limits of the sign function around* $x = 0$ *are −1 and 1, respectively.*

Problem 2.38. *Prove that the left and right limits of the function* $y = |x|$ *are the same and equal to 0.*

2.4.2 Main Theorems

Similarly to the numerical sequences, one can operate almost algebraically with the limits of functions. This follows immediately from the first definition of the limit of a function based on numerical sequences and the fact that one can operate with sequences algebraically, as we have seen in Sect. 2.2. Therefore, similar to sequences, analogous theorems exist for the limits of the functions.

Let us formulate these theorems here for convenience; we shall do it for the limit at a point, one-side limit theorems are formulated almost identically.

Theorem 2.9. *If the function $y = f(x)$ has a limit at x_0, it is unique.*

It is proven by contradiction.

Theorem 2.10. *If two functions, $f(x)$ and $g(x)$, have limits F and G, respectively, at point x_0, then the functions $f(x) + g(x)$, $f(x)g(x)$ and $f(x)/g(x)$ have definite limits at this point, equal, respectively, to $F + G$, $F \cdot G$ and F/G (it is assumed in the last case that $G \neq 0$).*

These are easily proven using the language of intervals.

Theorem 2.11. *If $\lim_{x \to x_0} f(x) = A$ and $f(x) \geq F$ (or $f(x) \leq F$) in some vicinity of x_0, then $A \geq F$ (or $A \leq F$).*

Proven by contradiction.

Theorem 2.12. *If in some vicinity of x_0 the two functions $f(x) \leq g(x)$, and both have definite limits at x_0, i.e. $\lim_{x \to x_0} f(x) = F$ and $\lim_{x \to x_0} g(x) = G$, then $F \leq G$.*

This is a consequence of the previous theorem since $h(x) = f(x) - g(x) \leq 0$ and $\lim_{x \to x_0} h(x) = F - G \leq 0$.

It then follows from the last two theorems that if $f(x)$ is bounded by two functions, $\phi(x) \leq f(x) \leq h(x)$, and both functions $\phi(x)$ and $h(x)$ tend to the same limit A at $x \to x_0$, so is the function $f(x)$.

All these theorems can be proven similarly to those for numerical sequences. As an example, let us prove the multiplication theorem. Since the two functions have definite limits F and G, then for any positive ϵ_1 and ϵ_2 one can always find such positive δ_1 and δ_2, so that from $|x - x_0| < \delta_1$ follows $|f(x) - F| < \epsilon_1$ and from $|x - x_0| < \delta_2$ follows $|g(x) - G| < \epsilon_2$. Consider now $\delta = \min\{\delta_1, \delta_2\}$ and any value of the x within the interval $|x - x_0| < \delta$. We can then write:

$$
\begin{aligned}
|f(x)g(x) - FG| &= |(f(x) - F)(g(x) - G) + F(g(x) - G) + G(f(x) - F)| \\
&< \epsilon_1\epsilon_2 + |F|\,\epsilon_2 + |G|\,\epsilon_1 ,
\end{aligned}
$$

which gives an expression for the ϵ in the right-hand side above, as required.

Problem 2.39. *Prove all the above theorems using the "$\epsilon - \delta$" language.*

We shall need some other theorems as well.

Theorem 2.13. *Consider a function $f(x)$ which is* limited *within some vicinity of x_0, i.e. $|f(x)| \leq M$. Then, if $\lim_{x \to x_0} g(x) = 0$, then $\lim_{x \to x_0} [f(x)g(x)] = 0$ as well. Note that $f(x)$ may not even have a definite limit at x_0.*

Proof. Since $g(x)$ has the zero limit, then for any x within $|x - x_0| < \delta$ we have $|g(x) - 0| = |g(x)| < \epsilon'$. Therefore, for the same values of x,

$$|f(x)g(x) - 0| = |f(x)g(x)| \leq M\,|g(x)| < M\epsilon',$$

i.e. $\epsilon = M\epsilon'$. **Q.E.D.**

As an example, consider $\lim_{x \to 0} [x \sin(1/x)]$. The sine function is not well defined at $x = 0$ (in fact, it does not have a definite limit there, recall a discussion on sequences in Sect. 2.2); however, it is bounded from below and above, i.e. it is limited: $|\sin(1/x)| \leq 1$. Therefore, since the limit of x at zero is zero, so is the limit of its product with the sine function: $\lim_{x \to 0} [x \sin(1/x)] = 0$.

At variance with the sequences, two more operations exist with functions: the composition and inverse; hence, two more limit theorems.

Theorem 2.14 (Limit of Composition). *Consider two functions $y = f(t)$ and $t = g(x)$; their composition is the function $h(x) = f(g(x))$. If $\lim_{x \to x_0} g(x) = t_0$ and $\lim_{t \to t_0} f(t) = F$, then $\lim_{x \to x_0} h(x) = F$.*

Proof. since $g(x)$ has a definite limit, then there exists δ such that $|x - x_0| < \delta$ implies $|g(x) - t_0| < \epsilon_1$. On the other hand, $f(t)$ also has a definite limit at t_0, i.e. $|t - t_0| < \delta_1$ implies $|f(t) - F| < \epsilon$. However, since $t = g(x)$, the inequality $|t - t_0| < \delta_1$ is equivalent to $|g(x) - t_0| < \delta_1$. Therefore, by taking $\delta_1 = \epsilon_1$, we can write:

$$|x - x_0| < \delta \;\Rightarrow\; |g(x) - t_0| = |t - t_0| < \delta_1 \;\Rightarrow\; |f(t) - F| = |f(g(x)) - F| < \epsilon,$$

proving the theorem, **Q.E.D.**

This result can be used to find limits of complex functions by considering them as compositions of simpler ones, for instance,

$$\lim_{x\to x_0} \sin^2 x = \left(\lim_{x\to x_0} \sin x \right)^2 = \sin^2 x_0.$$

Theorem 2.15 (Limit of the Inverse Function). *Let $f^{-1}(x)$ be an inverse function to $f(x)$; the latter has a limit at x_0 equal to F, i.e.* $\lim_{x\to x_0} f(x) = F$. *Then,* $\lim_{x\to F} f^{-1}(x) = x_0$.

Proof. if $y = f(x)$ is the direct function, then $x = f^{-1}(y)$. Since the direct function has the limit, then for each ϵ we can find δ such that $|x - x_0| < \delta$ implies $|f(x) - F| = |y - F| < \epsilon$. We need to prove that for any ϵ_1 there exists δ_1 such that for any y where $|y - F| < \delta_1$ we may write $|x - x_0| < \epsilon_1$. By taking $\epsilon_1 = \delta$, we can see from the existence of the limit for the direct function that we can always find $\delta_1 = \epsilon$, as required, **Q.E.D.**

2.4.3 Continuous Functions

Intuitively, it seems that if one can draw a function $y = f(x)$ with a continuous line between $x = a$ and $x = b$, then the function is thought to be *continuous*. A more rigorous consideration, however, requires some care.

One of the definitions of continuity of the function $f(x)$ within an interval $a < x < b$ is that its limit at any point x_0 within this interval coincides with the value of the function there:

$$\lim_{x\to x_0} f(x) = f(x_0). \tag{2.80}$$

It appears, however, that this definition follows from a less obvious statement that it is sufficient for the function $f(x)$ to have a definite limit for any x within some interval to be continuous there. Indeed, let us prove that Eq. (2.80) indeed follows from this statement. We shall prove this by contradiction: assume the opposite that the limit of $f(x)$ at x_0 is $A \neq f(x_0)$. We start by assuming that $\lim_{x\to x_0} f(x) = A > f(x_0)$. This means that for any $\epsilon > 0$ there exists $\delta > 0$ such that $|x - x_0| < \delta$ implies $|f(x) - A| < \epsilon$. Let us choose the positive ϵ as follows: $\epsilon = \frac{1}{2}(A - f(x_0))$. Then, we have:

$$|f(x) - A| < \epsilon = \frac{1}{2}(A - f(x_0)),$$

which, if calculated at the point $x = x_0$, gives:

$$|f(x_0) - A| = A - f(x_0) < \epsilon = \frac{1}{2}(A - f(x_0)),$$

which is obviously wrong, i.e. our assumption of A being larger than $f(x_0)$ is incorrect. Similarly, one can consider the case of $A < f(x_0)$ which also is easily shown to be wrong. Hence, $A = f(x_0)$, **Q.E.D.**

This means that the function $f(x)$ is continuous within some interval of x if it has well-defined limits at any point x belonging to this interval. Moreover, the limit of the continuous function coincides with the value of it at that point.

Yet another feeling for the continuity of a function one can get if we rewrite Eq. (2.80) in a slightly different form. Let us write $x = x_0 + \Delta x$, then $f(x) = f(x_0 + \Delta x)$ and hence instead of Eq. (2.80) we arrive at:

$$\lim_{\Delta x \to 0} [f(x_0 + \Delta x) - f(x_0)] = \lim_{\Delta x \to 0} \Delta f(x_0) = 0. \tag{2.81}$$

Here $\Delta x = x - x_0$ is the change of the argument x_0 of the function, while $\Delta f(x_0) = f(x_0 + \Delta x) - f(x_0)$ is the corresponding change of the function, see Fig. 2.13. As Δx becomes smaller, the point B on the graph of $f(x)$ which corresponds to the coordinates $x_0 + \Delta x$ and $f(x_0 + \Delta x)$, moves continuously towards the point A corresponding to x_0 and $f(x_0)$; as Δx becomes smaller, so is $\Delta y = f(x_0 + \Delta x) - f(x_0)$. Clearly, as the limit (2.81) exists for any point x (within some interval), this kind of continuity exists at each such point; as we change x continuously between two values x_1 and $x_2 > x_1$, the function $y = f(x)$ runs continuously over all values between $f(x_1)$ and $f(x_2)$.

Since a continuous function has well-defined limits for all values of x within some interval, the function cannot jump there, i.e. it cannot be discontinuous. This means that the limits from the left and right at each such point x should exist and be equal to the value of the function at that point. If this condition is not satisfied at some point x_0, the function $f(x)$ is not continuous at this point. There are two types

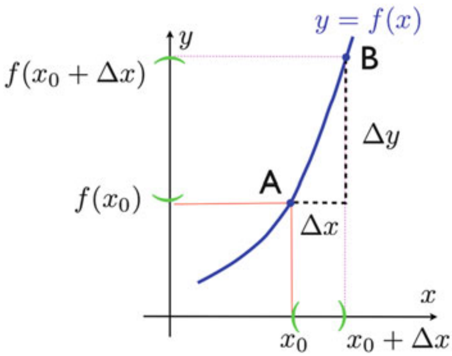
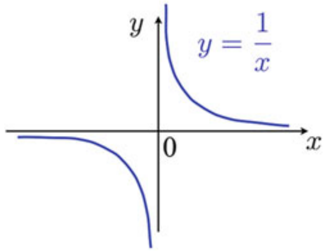


Fig. 2.13 Change of a function Δy tends to zero as change Δx of x tends to zero

Fig. 2.14 The hyperbolic function $y = 1/x$



of *discontinuities*. If the function $f(x)$ has well-defined left and right limits at point x_0, which are not equal to each other,

$$\lim_{x \to x_0^+} f(x) = f(x_0 + 0) \neq \lim_{x \to x_0^-} f(x) = f(x_0 - 0), \tag{2.82}$$

then it is said that it has the *discontinuity of the first kind*. All other discontinuities are of the *second kind*.

Functions $H(x)$ and $\mathrm{sgn}\,(x)$ have at $x = 0$ discontinuity of the first kind as both functions have well-defined limits from the left and right of this point. On the other hand, the function $y = 1/x$ has infinite limits on the left (equal to $-\infty$) and on the right (equal to $+\infty$) of the point $x = 0$, and hence has the discontinuity of the second kind there, see Fig. 2.14.

Now we are going to establish continuity of all elementary functions. Since this property is based on the functions having a well defined limit at each point, the limit theorems of Sect. 2.4.2 would allow us to formulate a set of very simple statements about the continuity of functions which are invaluable in establishing that property in complex cases: if two functions $f(x)$ and $g(x)$ are continuous in some interval of the values of their argument x, then their sum, difference, product and division ($f(x) \pm g(x), f(x)g(x)$ and $f(x)/g(x)$) are also continuous (except at the points of the roots of $g(x) = 0$ in the latter case). The same can be said about the composition of the two functions, $y = f\,(g(x))$. Moreover, if the function $f(x)$ is monotonically increasing or decreasing and is continuous, then its inverse (which will also be monotonic) is also continuous.

Let us now look at the elementary functions and establish their continuity.

Polynomials. Consider first a constant function $y = C$. Obviously, this is continuous as its limit at any $-\infty < x < \infty$ is equal to C which is the value of the function at any of these points. Then, let us consider an integer power function $y = x^n$. Using the binomial formula, we obtain for the change of the function at x:

$$\Delta y = (x + \Delta x)^n - x^n = \sum_{i=0}^{n} \binom{n}{i} (\Delta x)^i x^{n-i} - x^n = \sum_{i=1}^{n} \binom{n}{i} (\Delta x)^i x^{n-i}, \tag{2.83}$$

where in the second passage above we noticed that the very first term in the sum corresponding to $i = 0$ is exactly equal to x^n since the binomial coefficient $\binom{n}{0} =$

1 for any n (see Sect. 1.10); this means that x^n cancels out exactly. Now it is seen that each term in the sum above contains powers of Δx, and hence $\Delta y \to 0$ as $\Delta x \to 0$. This means that the integer power function is continuous.

A polynomial is a linear combination of such functions, i.e. a sum of products of integer powers with constant functions. Since arithmetic operations involving continuous functions result in a continuous function, any polynomial is a continuous function.

General Power Functions. Consider now $y = x^\alpha$, a general power function, where α is an arbitrary real number. The proof of this function being continuous follows the same logic we used to define this function in the first place in Sect. 2.3. We first consider $\alpha = 1/n$ with n being positive integer. The function $y = x^{1/n}$ is an inverse to $y = x^n$ and hence is continuous. The function $y = x^{n/m}$ (with $\alpha = n/m$ being a positive rational number) is a composition of two functions just considered, $y = \left(x^{1/m}\right)^n$, and hence is also a continuous function. Then, consider α being an irrational number. Following the same method as in Sect. 2.3, we consider two numerical sequences α_i' and α_i'' ($i = 1, 2, 3, \ldots$) of rational numbers bracketing the irrational α for each i (i.e. $\alpha_i' < \alpha < \alpha_i''$) and converging to α from both sides. Correspondingly, the two numerical sequences $x^{\alpha_i'}$ and $x^{\alpha_i''}$ bracket the power x^α for each i and any value of $x \geq 0$. Consider now the $x \to x_0$ limit. Since the rational power is a continuous function, each term in the two sequences tends to $x_0^{\alpha_i'}$ and $x_0^{\alpha_i''}$, i.e. for any i we can write $x_0^{\alpha_i'} < \lim_{x \to x_0} x^\alpha < x_0^{\alpha_i''}$. Taking now $i \to \infty$ and using the fact that both sequences on the left and right converge to the same limit x_0^α, we arrive at the conclusion that $\lim_{x \to x_0} x^\alpha = x_0^\alpha$, i.e. any positive real power function is indeed continuous. Next, the function $y = 1/x$ is obviously continuous for any $x \neq 0$:

$$\Delta y = \frac{1}{x + \Delta x} - \frac{1}{x} = -\frac{\Delta x}{x\,(x + \Delta x)}$$
$$\Longrightarrow \quad \lim_{\Delta x \to 0} \Delta y = -\lim_{\Delta x \to 0} \frac{\Delta x}{x\,(x + \Delta x)} = -\lim_{\Delta x \to 0} \frac{\Delta x}{x^2} = 0.$$

Since for negative power functions ($\alpha < 0$) one can always write $y = x^\alpha = 1/x^{-\alpha}$ which is a convolution of two continuous functions (note that $-\alpha > 0$), negative power functions are also continuous. We finally conclude that the power function with any real power α is continuous.

Trigonometric Functions. We have already considered in Sect. 2.4.1.1 the limit of the sine function, $y = \sin x$, and proved that its limit at $x \to x_0$ is equal to $\sin x_0$, i.e. to the value of the sine function at this point. This proves that the sine function is continuous. The cosine function $y = \cos x$ can be composed out of the sine function, e.g. $y = \sin(\pi/2 - x)$, and hence is also continuous. The tangent and cotangent are obtained via division of two continuous functions, sine and cosine, and hence are also continuous.

Inverse Trigonometric Functions. These are also continuous since they are inverse of the corresponding trigonometric functions which are continuous.

Exponential Functions. For any positive $a > 0$ we can write

$$\lim_{\Delta x \to 0}\left(a^{x+\Delta x} - a^x\right) = \lim_{\Delta x \to 0} a^x\left(a^{\Delta x} - 1\right) = a^x \lim_{\Delta x \to 0}\left(a^{\Delta x} - 1\right) = a^x\left(\lim_{\Delta x \to 0} a^{\Delta x} - 1\right) = 0,$$

since, as we have seen above in Eq. (2.78), $\lim_{\Delta x \to 0} a^{\Delta x} = 1$. Therefore, the exponential function $y = a^x$ is continuous.

Logarithmic Function. This is continuous as an inverse of the continuous exponential function.

Hyperbolic Functions. These functions are derived from the exponential functions via a finite number of arithmetic operations and hence are all continuous.

Problem 2.40. *Prove that the sine and cosine functions are continuous also by showing that their change* $\Delta y \to 0$ *as the change of the variable* $\Delta x \to 0$.

Problem 2.41. *Prove that the logarithmic function is continuous by demonstrating that its change* $\Delta y \to 0$ *as the change of the variable* $\Delta x \to 0$.

Problem 2.42. *Prove by direct calculation of* Δy *that the hyperbolic sine, cosine and tangent functions are continuous.*

Concluding this section, we see that any function constructed either as a composition and/or via any number of arithmetic operations from the elementary functions considered above is continuous for all values of x where the function is defined.

2.4.4 Several Famous Theorems Related to Continuous Functions

There are a number of (actually quite famous) theorems which clarify the meaning of continuity of functions. We shall formulate them and give an idea of the proof without going into subtle mathematical details in some cases.

Theorem 2.16 (Conservation of Sign). *If a function* $f(x)$ *is continuous around point* x_0, *then there always exists such vicinity of* x_0 *where* $f(x)$ *has the same sign as* $f(x_0)$.

Proof. If the function is continuous, then for any x from a vicinity of x_0 we have $\lim_{x \to x_0} f(x) = f(x_0)$. This means that for any x satisfying $|x - x_0| < \delta$ with $\delta > 0$ we have $|f(x) - f(x_0)| < \epsilon$, which in turn means that $-\epsilon + f(x_0) < f(x) < \epsilon + f(x_0)$. Consider first the case of $f(x_0) > 0$. Then by choosing $\epsilon = f(x_0)$ one obtains $0 < f(x) < 2f(x_0)$, i.e. within (possibly) small vicinity $|x - x_0| < \delta$ of x_0 we have $f(x)$ positive. If, however, $f(x_0) < 0$, then we choose $\epsilon = -f(x_0) > 0$ yielding $-2f(x_0) < f(x) < 0$, i.e. within some vicinity of x_0 the function $f(x)$ is negative. **Q.E.D.**

Theorem 2.17 (The First Theorem of Bolzano-Cauchy—Existence of a Root). *If a continuous function $f(x)$ has different signs at the boundaries c and d of the interval $c \leq x \leq d$, then there exists a point x_0 belonging to the same interval where $f(x_0) = 0$.*

Proof. Suppose for definiteness that $f(c) < 0$ and $f(d) > 0$, see Fig. 2.15. The idea of this "constructive" proof is to build a sequence of smaller and smaller intervals around x_0 such that it converges to the point x_0 itself after an infinite number of steps (if not earlier).[8] Choose the point x_1 in the middle of the interval, $x_1 = (c + d)/2$.

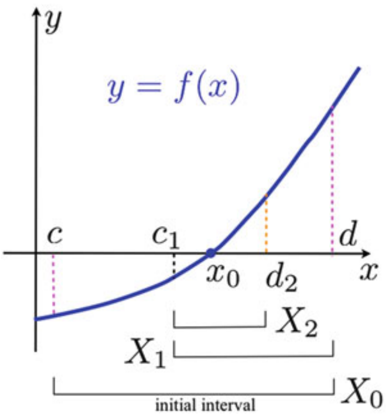


Fig. 2.15 To the "constructive" proof of the first theorem of Bolzano-Cauchy. We start with the original interval X_0 between points c and d. The first middle point between c and d is c_1 which replaces the left boundary, i.e. after the first step we have the new interval $c_1 \leq x \leq d_1$ (denoted X_1) with $c_1 = (c + d)/2$ and $d_1 = d$. At the second step we divide the new interval into two equal parts again and choose the middle point d_2 to replace the right boundary thereby arriving at the next interval $c_2 \leq x \leq d_2$ (denoted X_2) with $c_2 = c_1$ and $d_2 = (c_1 + d_1)/2$. Continuing this process leads to narrower and narrower intervals sequence $X_0, X_1, X_2, \ldots$, which all contain the point x_0

[8] In fact, this so-called "division by two" method is actually used in practical numerical calculations for finding roots of functions $f(x)$.

If $f(x_1) = 0$, then $x_0 = x_1$ and we stop here. Otherwise, we have to analyse the value of the function at the new point: if $f(x_1) < 0$, then we accept x_1 as the new left boundary c_1 of the interval; if $f(x_1) > 0$, we replace the right boundary d_1 with x_1. This way we arrive at a new interval $c_1 \leq x \leq d_1$ which is two times shorter, i.e. its length is $l_1 = (d - c)/2$. Then, we repeat the previous step: choose the middle point $x_2 = (c_1 + d_1)/2$ and by analysing the value of the function at point x_2 we either stop (and the theorem is proven) or construct a new interval $c_2 \leq x \leq d_2$ by replacing one of the boundary points, so that the function would still have different signs at them. The length $l_2 = l_1/2 = (d - c)/2^2$ of the new interval is again two times shorter. Repeating this process, we construct a (possibly) infinite sequence of intervals $c_n \leq x \leq d_n$ of reducing lengths $l_n = (d - c)/2^n$. It is essential that, by construction, all of them have their boundaries of different signs and their length becomes smaller and smaller tending to zero as $1/2^n$. Hence, in the limit, the intervals would shrink to a point, let us call it x_0, at which the function is equal to zero. The fact that such a point exists follows from the previous theorem of conservation of sign of a function: if we assume that $f(x_0) \neq 0$, it is either positive or negative and hence there will exist an interval around x_0 where the functions keep its sign; this, however, contradicts our construction upon which the function changes its sign at the boundaries; we have to accept that $f(x_0) = 0$. **Q.E.D.**

Theorem 2.18 (The Second Theorem of Bolzano-Cauchy—Existence of Any Intermediate Value). *Suppose a continuous function $f(x)$ at the boundaries of the interval $c \leq x \leq d$ has values $f(c) = C$ and $f(d) = D$. If A is any number between C and D, then there exists a point a within the interval, $c \leq a \leq d$, where $f(a) = A$.*

Problem 2.43. *Prove this theorem by defining a new function $g(x) = f(x) - A$ and then use the "existence of root" theorem.*

Basically, this theorem states that the function $f(x)$ takes on all intermediate values between C and D within the interval. In other words, one can draw the graph of the function between c and d without lifting the pen off the paper.

Theorem 2.19 (Existence of a Limit at a Region Boundary for Monotonically Increasing Functions). *Consider a function $f(x)$ which is monotonically increasing within some interval $a \leq x < b$, i.e. for any $x_1 < x_2$ inside the interval $f(x_1) < f(x_2)$. If the function is limited from above within the interval, $f(x) \leq M$, then when $x \to b$ the function has a limit which is equal to M.*

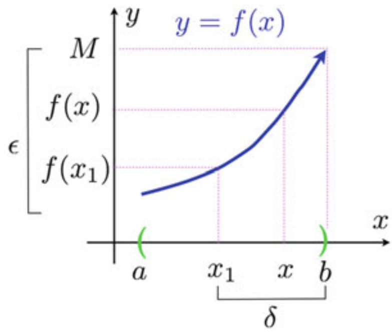


Fig. 2.16 To the proof of Theorem 2.19: the monotonically increasing function $f(x)$ is limited within the interval $a \leq x < b$ by the number M

Proof. (see Fig. 2.16) Let us choose some positive ϵ. Since the function $f(x)$ increases monotonically, one can always find such value x_1 that $f(x_1) > M - \epsilon$. This value of x_1 will be by some δ on the left from the boundary value b, i.e. $x_1 = b - \delta$. Then, for any $x > x_1$, i.e. for any x satisfying $|x - b| < \delta$, we would have $f(x) > f(x_1) > M - \epsilon$. However, on the other hand, $f(x) \leq M < M + \epsilon$, so that

$$M - \epsilon < f(x) < M + \epsilon \quad \Longrightarrow \quad |f(x) - M| < \epsilon,$$

as required, i.e. the limit of the function exists and is equal to the value of M. **Q.E.D.**

Problem 2.44. *Prove the analogous theorem for a monotonically increasing function at its left boundary (i.e. for $f(x) > M > 0$), as well as for a monotonically decreasing function at either of its boundaries.*

2.4.5 Infinite Limits and Limits at Infinities

The limit of a function may not necessarily be finite. If for any $\epsilon > 0$ there exists $\delta > 0$ such that for any x within $|x - x_0| < \delta$ one has[9] $f(x) > \epsilon$, then it is said that the function $f(x)$ tends to $+\infty$ as $x \to x_0$. Indeed, if for any vicinity of x_0 (small δ) one can always find values of the $f(x)$ larger than any given ϵ (however large), then the function $f(x)$ grows indefinitely to infinity.

[9]Note that here ϵ is not longer assumed "small", it is "large".

Problem 2.45. *Formulate the definition for the function to tend to $-\infty$ when $x \to x_0$.*

We can also consider behaviour of a function at both infinities, i.e. when x_0 is either $+\infty$ or $-\infty$. These are basically one-side limits which are formulated as follows. The function $f(x)$ tends to F when $x \to +\infty$, if for any ϵ there exists $\delta > 0$ such that for any $x > \delta$ we have[10] $|f(x) - F| < \epsilon$.

Problem 2.46. *Formulate definitions for the function $f(x)$ to have: (i) a finite limit F when $x \to -\infty$, (ii) an infinite limit $F = +\infty$ (or $-\infty$) when $x \to +\infty$ (or $-\infty$).*

Problem 2.47. *Prove the analogue of Theorem 2.19 of the previous subsection for the limits at $+\infty$ and $-\infty$ for both monotonically increasing and decreasing functions.*

Consider, for instance, the function $f(x) = 1/x^2$ at $x \to 0$. In this case $x_0 = 0$; we see that for any x within $|x - 0| = |x| < \delta$ we shall have $f(x) = 1/x^2 > 1/\delta^2 = \epsilon$, i.e. one should take $\delta = 1/\sqrt{\epsilon}$ in this case. And indeed, the function grows indefinitely as we approach $x = 0$ either from the left or from the right.

Now, if $f(x) = x^2$, then at $x \to +\infty$ it tends to $+\infty$ as well. Indeed, for any $x > \delta$ we have $f(x) = x^2 > \delta^2 = \epsilon$, i.e. for any ϵ we can indeed find $\delta = \sqrt{\epsilon}$.

Similarly one can formulate infinite limits for a function at $x \to \pm\infty$.

Problem 2.48. *Prove that the left and right limits of the function $y = 3^{1/x}$ at $x = 0$ are 0 and ∞, respectively, and hence this function does not have a definite limit at $x = 0$.*

Problem 2.49. *Prove using the "$\epsilon - \delta$" language that $\lim_{x\to+\infty} e^{-x^2}/x = 0$.*

Let us prove that

$$\lim_{x\to+\infty}\left(1+\frac{1}{x}\right)^x = e \tag{2.84}$$

This is a generalisation to any real x of the special cases we considered above in Sect. 2.3.4 for the base of the natural logarithm, the number e. To prove this result, we note that any positive x can be placed between two *consecutive* integers: $n \leq$

[10] δ is assumed to be indefinitely "large" here.

$x \leq n + 1$. Then,

$$\frac{1}{n+1} \leq \frac{1}{x} \leq \frac{1}{n} \quad \Longrightarrow \quad 1 + \frac{1}{n+1} \leq 1 + \frac{1}{x} \leq 1 + \frac{1}{n}.$$

Since the power function is a monotonically increasing function, we can also write:

$$\left(1 + \frac{1}{n+1}\right)^n \leq \left(1 + \frac{1}{x}\right)^x \leq \left(1 + \frac{1}{n}\right)^{n+1}$$

or

$$\frac{\left(1 + \frac{1}{n+1}\right)^{n+1}}{1 + \frac{1}{n+1}} \leq \left(1 + \frac{1}{x}\right)^x \leq \left(1 + \frac{1}{n}\right)^n \left(1 + \frac{1}{n}\right).$$

As we now tend x to $+\infty$, n will also tend to the same limit. However, from Eq. (2.22) it follows that both limits on the left and right of the last inequality tend to the same limit e:

$$\frac{\left(1+\frac{1}{n+1}\right)^{n+1}}{1+\frac{1}{n+1}} \to \left(1+\frac{1}{n+1}\right)^{n+1} \to e \quad \text{and} \quad \left(1+\frac{1}{n}\right)^n \left(1+\frac{1}{n}\right) \to \left(1+\frac{1}{n}\right)^n \to e,$$

so that, the function $\left(1 + \frac{1}{x}\right)^x$ will tend to the same limit. **Q.E.D.**

Problem 2.50. *Prove that*

$$\lim_{x \to +\infty} \left(1 + \frac{y}{x}\right)^x = e^y, \tag{2.85}$$

where y is any real number (including $y < 0$).

2.4.6 Dealing with Uncertainties

The case for the number e considered above is an example of an uncertainty. Indeed, $1 + 1/x$ tends to 1 as $x \to +\infty$, however, when raised in an infinite power $x \to +\infty$ the result is not 1, but the number $e = 2.718\ldots$. Several other types of uncertainties exist and often happen in applications, so we shall consider them here.

Firstly, let us consider yet another famous limit:

$$\lim_{x\to 0}\frac{\sin x}{x}=1. \tag{2.86}$$

There is an uncertainty: indeed, both $\sin x$ and x tend to zero as x tends to zero, so that we encounter the case of $0/0$ here. The result of the limit, as we shall see imminently is finite and equal to one. The starting point is the inequality (1.93), $\sin x < x < \tan x$. Dividing both sides of this inequality by $\sin x$ (and assuming $0 < x < \pi/2$), we get:

$$1<\frac{x}{\sin x}<\frac{\tan x}{\sin x}=\frac{1}{\cos x}\quad\text{or}\quad\cos x<\frac{\sin x}{x}<1.$$

As $x \to 0$, the cosine tends to 1, i.e. our function of interest becomes bracketed between two equal limits of one, and hence tends to the same limit itself. **Q.E.D.**

The following examples deal with other cases of uncertainties.

Uncertainty ∞/∞:

$$\lim_{x\to+\infty}\frac{2x^2-1}{x^2+2x+1}=\lim_{x\to+\infty}\frac{2-1/x^2}{1+2/x+1/x^2}=\frac{2-0}{1+0+0}=2.$$

Uncertainty $\infty-\infty$:

$$\begin{aligned}\lim_{x\to+\infty}\left(\sqrt{x^2-1}-\sqrt{x+1}\right)&=\lim_{x\to+\infty}\left[\left(\sqrt{x^2-1}-\sqrt{x+1}\right)\frac{\sqrt{x^2-1}+\sqrt{x+1}}{\sqrt{x^2-1}+\sqrt{x+1}}\right]\\&=\lim_{x\to+\infty}\frac{\left(x^2-1\right)-(x+1)}{\sqrt{x^2-1}+\sqrt{x+1}}\\&=\lim_{x\to+\infty}\frac{x^2-x-2}{\sqrt{x^2-1}+\sqrt{x+1}}\\&=\lim_{x\to+\infty}\frac{x-1-2/x}{\sqrt{1-1/x^2}+\sqrt{1/x+1/x^2}}=+\infty.\end{aligned}$$

Uncertainty $0/0$:

$$\begin{aligned}\lim_{x\to 0}\frac{\ln(1+x)}{x}&=\lim_{x\to 0}\left[\frac{1}{x}\ln(1+x)\right]=\lim_{x\to 0}\ln(1+x)^{1/x}=\left|\begin{array}{c}\text{change}\\ y=1/x\end{array}\right|\\&=\lim_{y\to\infty}\ln\left(1+\frac{1}{y}\right)^y=\ln\left[\lim_{y\to\infty}\left(1+\frac{1}{y}\right)^y\right]=\ln e=1.\end{aligned}$$

Note that here we were able to take the limit inside the logarithm since the latter is a continuous function.

Uncertainty $0 \cdot \infty$:

$$\lim_{x\to 0} x\cot x = \lim_{x\to 0} x\frac{\cos x}{\sin x} = \left(\lim_{x\to 0}\frac{x}{\sin x}\right)\left(\lim_{x\to 0}\cos x\right) = \left(\lim_{x\to 0}\frac{\sin x}{x}\right)^{-1}\cdot 1 = 1.$$

Problem 2.51. *Prove the following limits:*

$$\lim_{x\to+\infty}\frac{5^x-5^{-x}}{5^x+5^{-x}} = 1; \quad \lim_{x\to 1+0}\left(\frac{3}{x-1}-\frac{x+3}{x^2-1}\right) = +\infty;$$

$$\lim_{x\to 1-0}\left(\frac{3}{x-1}-\frac{x+3}{x^2-1}\right) = -\infty; \lim_{x\to+\infty}\left(\sqrt{x}-\sqrt{x+2}\right) = 0;$$

$$\lim_{x\to 0}\frac{1-\cos x}{x^2} = \frac{1}{2}; \quad \lim_{x\to+\infty}\sqrt[x]{x} = 1.$$

[*Hint*: in the latter case transform first the expression using an exponential function.]

Part II
Basics

Chapter 3
Derivatives

3.1 Definition of the Derivative

Very often we would like to know an *instantaneous* rate of change of something. For instance, consider a one-dimensional motion of a particle along the x axis: at each time t we have the position of the particle given by a function $x(t)$. If at the initial moment $t = 0$ the particle was at the position $x_0 = x(0)$, then at time t it arrives at the point with the coordinate $x(t)$. If the particle moved with a constant velocity v_c, then its coordinate would change according to $x(t) = x_0 + v_c t$. This means that in this case the rate of change of $x(t)$, i.e. the velocity, can be calculated via $v_c = (x - x_0)/t$. If, however, the particle travelled with different velocities along its trajectory, i.e. at some instances travelled faster and at some slower, then this expression would only give an average velocity $v_{\text{av}} = (x - x_0)/t$. Although we do very often use this notion of slower or faster velocity describing a single continuous motion and seem to understand intuitively its meaning, we probably subconsciously mean average velocities over some rather short periods of time during the trajectory. For instance, we can split the whole journey time t into N equal slices of length Δt, i.e. $t_0 = 0$, $t_1 = \Delta t$, $t_2 = 2\Delta t$, etc., $t_N = N\Delta t = t$ (i.e. $t_i = i\Delta t$ with $\Delta t = t/N$ and $i = 0, 1, 2, \dots, N$), when positions of the particle are correspondingly given by its coordinates $x_0, x_1, x_2, \dots, x_i, \dots, x_N = x_t$, and then define average velocities

$$v_{\text{av},i} = \frac{x_{i+1} - x_i}{\Delta t} = \frac{x(t_{i+1}) - x(t_i)}{\Delta t} = \frac{x(t_i + \Delta t) - x(t_i)}{\Delta t} \tag{3.1}$$

for each time interval $\Delta t = t_{i+1} - t_i$. Of course, this calculation would be able to tell us if the velocity changed along the journey; however, the actual quantitative answer would obviously depend on the number of time slices N we used: the more slices we use, the more correct this calculation would be. This simple minded exercise brings us to a method which would be able to determine the instantaneous velocity of the particle at any time along the journey: one has to tend the number of slices

L. Kantorovich, *Mathematics for Natural Scientists*, Undergraduate Lecture Notes in Physics, DOI 10.1007/978-1-4939-2785-2_3

N to infinity ($N \to \infty$, which is equivalent to taking the time length of each slice $\Delta t \to 0$). Therefore, we arrive at the following conclusion: if we would like to calculate the velocity $v(t)$ at some time t along the journey, we need to consider the positions of the particle $x(t)$ and $x(t + \Delta t)$ at the times t and $t + \Delta t$, and then calculate

$$v(t) \simeq \frac{x(t + \Delta t) - x(t)}{\Delta t}$$

(compare Eq. (3.1)). This expression becomes more and more precise as Δt gets smaller; it becomes exact in the limit of $\Delta t \to 0$.

This kind of reasoning can be applied to many phenomena around us. Consider electrons flowing in a thin wire with cross section S. Again, let us split the time into small and equal slices Δt and count the number of electrons ΔN_i passing through S during each slice $\Delta t = t_{i+1} - t_i$. This way we can approximately calculate the electric current passing through the wire at time t_i as $J_i \simeq e\Delta N_i/\Delta t$ (e is the electron charge). This will be an average current; of course, the smaller the time slices, the better our calculation would be and, if the current changes with time, we should be able to reproduce this dependence in detail. If $N(t)$ is the total number of electrons passed through S up to time t, then their total number passed over time Δt is $\Delta N = N(t+\Delta t) - N(t)$, and we arrive again at a formula for the current "around time t" as

$$J(t) \simeq \frac{N(t + \Delta t) - N(t)}{\Delta t},$$

which becomes more and more exact as the time slices Δt get smaller and smaller; in the limit of $\Delta t \to 0$, this formula becomes exact.

Similarly, one can define a linear density of a metal rod by dividing the rod of length L into N slices each of length Δx (so that $L = N\Delta x$), and calculating the mass Δm_i of each i-th slice; then, the average density of the slice located near x_i would be

$$\rho_i \simeq \frac{\Delta m_i}{\Delta x} = \frac{m(x_i + \Delta x) - m(x_i)}{\Delta x},$$

where $m(x)$ is the total mass of the rod between points $x = 0$ and $x > 0$. Again, the formula becomes exact if we tend $\Delta x \to 0$ which corresponds to the infinite number of length slices. Exactly in the same way one can define the linear charge density of a charged rod via the limit of its total charge within a thin slice divided by its length.

Although many more examples can be given (and the reader will find them in this book, as well as in any physics textbook), all of them can be summarised as follows: if we have a function $y = y(x)$ and we are interested in its *rate of change* at the point x, we calculate the limit

$$y'(x) = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{y(x + \Delta x) - y(x)}{\Delta x}, \tag{3.2}$$

called the *derivative* of the function $y(x)$ at point x. Here $\Delta y = y(x + \Delta x) - y(x)$ is the change of the function $y(x)$ over the change Δx of its argument from x to $x + \Delta x$. Sometimes, the derivative is also denoted as y' or y'_x; in the latter case, the index x emphasises specifically that the derivative is taken with respect to the variable x.

Example 3.1. ► Calculate the derivative of the following functions: $y = c$ (a constant), $y = x$ and $y = x^2$.

Solution. In the first case $\Delta y = c - c = 0$ and hence the derivative is zero, i.e. $(c)' = 0$. In the second case $\Delta y = y(x + \Delta x) - y(x) = (x + \Delta x) - x = \Delta x$, so that $\Delta y / \Delta x = \Delta x / \Delta x = 1$ and hence $(x)' = 1$ for any value of the x. In the last case, we need to calculate first the change of the function:

$$\Delta y = y(x + \Delta x) - y(x) = (x + \Delta x)^2 - x^2 = 2x\Delta x + (\Delta x)^2 .$$

Therefore, the derivative of $y = x^2$ is

$$\left(x^2\right)' = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{2x\Delta x + (\Delta x)^2}{\Delta x} = \lim_{\Delta x \to 0} (2x + \Delta x) = 2x. \blacktriangleleft \tag{3.3}$$

Note that the change of the function due to the change Δx of its argument in this case,

$$\Delta y = 2x\Delta x + (\Delta x)^2 = y'(x)\Delta x + \alpha(\Delta x)\Delta x, \tag{3.4}$$

is composed of a linear part containing the derivative and an additional term which tends to zero faster than linearly (in fact, as the square of Δx in this case) since $\alpha(\Delta x) = \Delta x$ is some function which tends to zero as $\Delta x \to 0$.

Problem 3.1. *Show using the binomial expansion (see Sect. 1.10) that*

$$(x^n)' = nx^{n-1} \tag{3.5}$$

for any integer $n \geq 1$.

The derivative of a function has also a very simple geometric interpretation. Consider a function $y(x)$ plotted in Fig. 3.1(a). Let us try to find an equation $y = a + kx$ for the tangent line (shown in dashed green) at the point A with coordinates x and $y(x)$. Our interest here is in obtaining the *slope* of the curve $y = y(x)$ at point A which is given by the constant k (it will in general depend on the x). The constant a can then be easily determined using the known coordinates of the point A. We shall be using the following procedure: let us first pick an arbitrary point B_1 on the curve of the function and draw a line connecting this point with the point A. This line may have a very different direction to the one we are seeking; however, its slope given by the tangent of AB_1 to the x axis may already serve as the first approximation to the contstant k. Let us now take a point B_2 on the graph

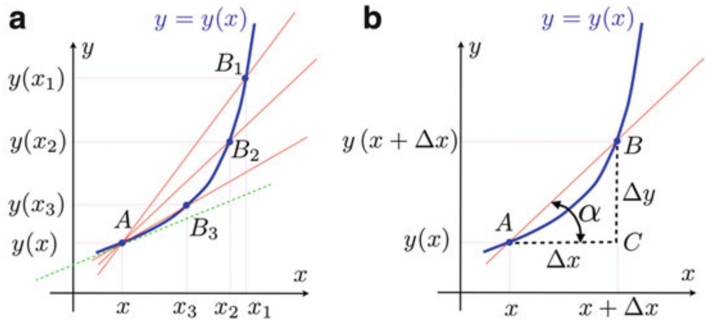


Fig. 3.1 For the determination of the tangent line to a curve of the function $y = y(x)$ at point A

of the function which is somewhere between the points B_1 and A; it is now closer to the point of interest. This time the line AB_2 may serve as a better approximation to the tangent line: although it still crosses the curve of the function in two close points A and B_2 (in the vicinity of A the tangent line will only *touch* the curve at the point A; of course, it may cross the curve at some other more distant point(s) if the curve sufficiently changes its direction, e.g. oscillates); still this straight line AB_2 is a much better approximation to the tangent line than the line AB_1. This process can be continued: we take a point B_3 on the curve lying between A and B_2, then a point B_4 between A and B_3, etc., each time taking the point B_N closer and closer to the point A and thereby obtaining better and better approximation to the tangent line. In the limit when the point B_N tends to the point A, our procedure would give the correct tangent line at the point A. It is easy to see that this process is equivalent to taking a point B on the curve of the function $y(x)$, see Fig. 3.1(b), and moving it towards A. Therefore, to obtain the slope k at point A (or at x), we can calculate the tangent of the line AB,

$$\tan\alpha = \frac{y\,(x + \Delta x) - y\,(x)}{\Delta x} = \frac{\Delta y}{\Delta x},$$

and then take the limit as $\Delta x \to 0$ (which would exactly correspond to the point B moving towards the point A). So, in the limit, we shall get the exact value of the slope k of the tangent line at point A. Again, as one can see, we relate the slope to the derivative of the function $y(x)$ at point x, i.e. $k = y'(x)$.

Similarly to the right and left limits defined in the previous Chapter, we can also define left and right derivatives, denoted $y'_-(x) = y'(x - 0)$ and $y'_+(x) = y'(x + 0)$, respectively. In fact, the derivative defined above is exactly the right derivative as it contains only the values of the function on the right of x, the point of interest. The left derivative is defined only via points on the left of x as follows:

$$y'_-(x) = \lim_{\Delta x \to 0} \frac{y\,(x) - y\,(x - \Delta x)}{\Delta x}, \tag{3.6}$$

where $\Delta x > 0$. This expression can be written in a form very similar to that of the right derivative in Eq. (3.2) by assuming that $\Delta x < 0$ and tends to zero always being negative (this is denoted as $\Delta x \to 0 - 0$ or simply $\Delta x \to -0$):

$$y'_{-}(x) = \lim_{|\Delta x| \to 0} \frac{y(x) - y(x - |\Delta x|)}{|\Delta x|} = \lim_{\Delta x \to -0} \frac{y(x + \Delta x) - y(x)}{\Delta x}. \tag{3.7}$$

If the derivative of $y(x)$ exists at point x, then both the right and left derivatives coincide there; otherwise, the derivative does not exist.

> **Problem 3.2.** *Consider a line tangent to a curve of the function $y = y(x)$ at point x_0 which is given by the equation $y = k_\tau (x - x_0) + y_0$, with $k_\tau = y'(x_0)$ and $y_0 = y(x_0)$. Prove that the line perpendicular to the tangent line and passing through the same point (x_0, y_0) is given by the equation $y = k_\perp (x - x_0) + y_0$, with $k_\perp = -1/k_\tau = -1/y'(x_0)$.*

3.2 Main Theorems

The notion of the derivative of a function is closely related to its continuity. To establish this relationship, let us first consider the change Δy of the function $y(x)$ at point x due to change Δx of its argument in a bit more detail. If the derivative $y'(x)$ exists (in other words, if the limit (3.2) exists), then one can write:

$$\Delta y = y'(x)\Delta x + \alpha \Delta x, \tag{3.8}$$

where α is an explicit function of Δx (we have seen an example of this relation above, Eq. (3.4), when considering the square function). It is essential that α tends to zero as $\Delta x \to 0$. Indeed, only in this case does the change Δy of y above yields the derivative:

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = y'(x) + \lim_{\Delta x \to 0} \alpha = y'(x).$$

The formula (3.8) states, quite importantly, that for any Δx the change of the function is in the first approximation given by the product of its derivative and Δx; the correction term, equal to $\alpha \Delta x$, is of a higher order with respect to Δx, and for small Δx may be neglected. Therefore, the main part of the change of the function, Δy, due to the change Δx of its argument x can be written as $\Delta y \simeq y'(x)\Delta x$. This change is called the *differential* of the function $y = y(x)$ and is denoted dy. Using dx

for the differential of the argument (this definition is consistent with the differential of the linear function $y = x$, as for it $dy = (x)' \, \Delta x = \Delta x$ and, since $y = x$, we have $dx = \Delta x$ as well), we can write:

$$dy = y'(x)dx. \tag{3.9}$$

For instance, if $y = x^2$, then $dy = \left(x^2\right)' dx = 2xdx$ or simply $d\left(x^2\right) = 2xdx$. Using the notion of differentials of the function and its argument, we can also write the derivative of the function as their ratio:

$$y'(x) = \frac{dy}{dx}. \tag{3.10}$$

This formula has a clear meaning of the slope of the tangent line to the curve of the function $y(x)$ as using differentials already implies performing the limit $\Delta x \to 0$. In addition, the ratio of differentials also serves as a very convenient notation for the derivative which we will frequently use.

Example 3.2. ► Derive an explicit expression for the function $\alpha\,(\Delta x)$ for $y = x^n$ with positive integer $n \geq 1$.

Solution. using the binomial expansion, we write (see Eq. (2.83))

$$\begin{aligned}
\Delta y &= (x + \Delta x)^n - x^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} (\Delta x)^k - x^n = \sum_{k=1}^{n} \binom{n}{k} x^{n-k} (\Delta x)^k \\
&= \binom{n}{1} x^{n-1} \Delta x + \sum_{k=2}^{n} \binom{n}{k} x^{n-k} (\Delta x)^k \\
&= nx^{n-1} \Delta x + \left[\sum_{k=2}^{n} \binom{n}{k} x^{n-k} (\Delta x)^{k-1} \right] \Delta x,
\end{aligned}$$

which is of the required form (3.8) since $(x^n)' = nx^{n-1}$, see Eq. (3.6). Therefore, the expression in the square brackets must be the function $\alpha\,(\Delta x)$ we seek for; it is clearly seen that it tends to zero as $\Delta x \to 0$ (the sum starts from $k = 2$).◄

Now we are ready to formulate the relationship between the continuity of a function and the existence of its derivative:

Theorem 3.1. *If a function $y = y(x)$ has a derivative at point x, it is continuous there.*

Proof. Indeed, if $y'(x)$ exists, then one can write expression (3.8) for Δy. Taking the limit as $\Delta x \to 0$ we immediately obtain that $\Delta y \to 0$ as well at this point, which is one of the definitions of the continuity of a function, as explained in Sect. 2.4.3. **Q.E.D.**

Note that the inverse statement is not true. Indeed, consider the function $y = |x|$. This function is continuous, since the limits from left and right at $x = 0$ give one and the same number 0, and this coincides with the value of the function at $x = 0$. At the same time, the derivative at $x = 0$ does not exist. Indeed, for $x > 0$ the function is $y = x$ and hence $y' = (x)' = 1$. On the other hand, for $x < 0$ we have $y = -x$ and hence

$$y' = (-x)' = \frac{[-(x + \Delta x)] - (-x)}{\Delta x} = -1.$$

We see that the derivatives from left and right around $x = 0$ are different, and hence the derivative at $x = 0$ does not exist.

Theorem 3.2 (Sum Rule). *if $y(x) = u(x) \pm v(x)$, then $y'(x) = u'(x) \pm v'(x)$, i.e.*

$$[u(x) \pm v(x)]' = u'(x) \pm v'(x). \tag{3.11}$$

Proof. Consider the function $y(x)$ which can be written as a sum of two functions $u(x)$ and $v(x)$. Then, according to the definition of the derivative (3.2), we can write:

$$\begin{aligned}
\frac{\Delta y}{\Delta x} &= \frac{1}{\Delta x}\left[u(x + \Delta x) + v(x + \Delta x) - u(x) - v(x)\right] \\
&= \frac{u(x + \Delta x) - u(x)}{\Delta x} + \frac{v(x + \Delta x) - v(x)}{\Delta x} = \frac{\Delta u}{\Delta x} + \frac{\Delta v}{\Delta x}.
\end{aligned}$$

After taking the limit as $\Delta x \to 0$, we obtain the required result. Similarly, one proves $(u - v)' = u' - v'$. **Q.E.D.**

Theorem 3.3 (Product Rule). *If $y(x) = u(x)v(x)$, then*

$$y'(x) = [u(x)v(x)]' = u'(x)v(x) + u(x)v'(x). \tag{3.12}$$

Proof. We need to calculate the change of the function $y(x)$ due to the change Δx of its argument:

$$\begin{aligned}
\Delta y &= u(x + \Delta x)\, v(x + \Delta x) - u(x)\, v(x) \\
&= [u(x + \Delta x) - u(x)]\, v(x + \Delta x) + u(x)\,[v(x + \Delta x) - v(x)] \\
&= (\Delta u)\, v(x + \Delta x) + u(x)\, \Delta v.
\end{aligned}$$

Dividing Δy by Δx and taking the limit, we obtain the required result:

$$\lim_{\Delta x\to 0}\frac{\Delta y}{\Delta x} = \lim_{\Delta x\to 0}\left[\frac{\Delta u}{\Delta x}v\left(x+\Delta x\right)\right] + \lim_{\Delta x\to 0}\left[u\left(x\right)\frac{\Delta v}{\Delta x}\right]$$
$$= \lim_{\Delta x\to 0}\frac{\Delta u}{\Delta x}\lim_{\Delta x\to 0}v\left(x+\Delta x\right) + u\left(x\right)\lim_{\Delta x\to 0}\frac{\Delta v}{\Delta x} = u'(x)v\left(x\right) + u\left(x\right)v'\left(x\right)$$

Q.E.D.

Problem 3.3. *Prove that derivative of a linear combination of two functions is equal to the linear combination of their derivatives, i.e.*

$$\left[\alpha u\left(x\right) + \beta v\left(x\right)\right]' = \alpha u'\left(x\right) + \beta v'\left(x\right). \tag{3.13}$$

Problem 3.4. *Generalise the product rule for more than two functions:*

$$\left[u_1(x)u_2\left(x\right)\ldots u_n\left(x\right)\right]' = u_1'u_2\ldots u_n + u_1u_2'\ldots u_n + \cdots + u_1u_2\ldots u_n'$$
$$= \sum_{i=1}^{n}\left[u_i'\prod_{j=1,j\neq i}^{n}u_j\right], \tag{3.14}$$

where the product indicated by the symbol $\prod$ *is taken over all values of* j *between* 1 *and* n *except for* $j=i$.

Theorem 3.4 (Quotient Rule). *If* $y(x) = u(x)/v(x)$ *and* $v(x)\neq 0$, *then*

$$y'(x) = \left[\frac{u\left(x\right)}{v\left(x\right)}\right]' = \frac{1}{v^2}\left[u'v - uv'\right]. \tag{3.15}$$

Proof. Here we write

$$\Delta y = \frac{u\left(x+\Delta x\right)}{v\left(x+\Delta x\right)} - \frac{u\left(x\right)}{v\left(x\right)} = \frac{u\left(x+\Delta x\right)v\left(x\right) - u\left(x\right)v\left(x+\Delta x\right)}{v\left(x+\Delta x\right)v\left(x\right)}$$
$$= \frac{v(x)\left[u\left(x+\Delta x\right) - u\left(x\right)\right] - u\left(x\right)\left[v\left(x+\Delta x\right) - v\left(x\right)\right]}{v\left(x+\Delta x\right)v\left(x\right)} = \frac{v\left(x\right)\Delta u - u\left(x\right)\Delta v}{v\left(x+\Delta x\right)v\left(x\right)},$$

which, after dividing by Δx and taking the limit, gives:

$$\lim_{\Delta x\to 0}\frac{\Delta y}{\Delta x} = \frac{v(x)\left(\lim_{\Delta x\to 0}\frac{\Delta u}{\Delta x}\right) - u\left(x\right)\left(\lim_{\Delta x\to 0}\frac{\Delta v}{\Delta x}\right)}{\lim_{\Delta x\to 0}\left[v\left(x+\Delta x\right)v\left(x\right)\right]} = \frac{vu' - uv'}{v^2}, \text{ **Q.E.D.**}$$

Problem 3.5. *Show that the formula (3.5) for the derivative of positive powers can be generalised to negative integer powers as well, i.e. that* $(x^n)' = nx^{n-1}$ *for any* n *from* $\mathbb{Z}$ *(i.e.* $n = 0, \pm 1, \pm 2, \pm 3, \ldots$*).*

Theorem 3.5 (Chain Rule). *Consider a composite function* $y = u(v)$ *and* $v = v(x)$*, i.e.* $y = u(v(x))$*. Then, its derivative is obtained as follows:*

$$[u(v(x))]' = u'_v v', \tag{3.16}$$

where u'_v *corresponds to the derivative of the function* $u(v)$ *with respect to* v *(not* x*).*

Proof. Again, we need to calculate the change Δy of the function $y(x)$:

$$\Delta y = u(v(x + \Delta x)) - u(v(x)) .$$

Here $v(x + \Delta x) = \Delta v + v(x)$ and we can similarly write $u(v + \Delta v) = \Delta_v u + u(v)$, where $\Delta_v u$ is the corresponding change of $u(v)$ when v changes from v to $v + \Delta v$. Therefore,

$$\begin{aligned} \frac{\Delta y}{\Delta x} &= \frac{u(v(x+\Delta x)) - u(v(x))}{\Delta x} = \frac{u(v+\Delta v) - u(v)}{\Delta x} \\ &= \frac{u(v+\Delta v) - u(v)}{\Delta v}\frac{\Delta v}{\Delta x} = \frac{\Delta_v u}{\Delta v}\frac{\Delta v}{\Delta x}, \end{aligned}$$

so that, after taking the limit, we arrive at Eq. (3.16). **Q.E.D.**

Problem 3.6. *Using the chain rule, show that* $\left[u(x)^{-1}\right]' = -u^{-2}u'$*. Does this formula agree with the quotient rule (3.15)? Then, using this result and the product rule, prove the quotient rule in a general case.*

Problem 3.7. *A one-dimensional movement of a particle of mass* m *along the* x *axis is governed by the Newtonian equation of motion,* $\dot{p} = -dU(x)/dx$*, where* $x(t)$ *is the particle coordinate,* $U(x)$ *is its potential energy,* $p = m\dot{x}$ *its momentum, and dots above* x *and* p *denote the time derivative. Show that the total energy* $H = p^2/2m + U(x)$ *is conserved in time. (The function* H *is also called Hamiltonian of the particle.)* [*Hint*: assume that both p and x are some functions of time which would make H a composite function, so that you may use the chain rule.]

Theorem 3.6 (Inverse function rule). *If $y = y(x)$ is the inverse of $x = x(y)$, then the derivatives of these two functions are related simply by*

$$[y(x)]' = \frac{1}{x'(y)}. \tag{3.17}$$

Proof. Here

$$y' = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{1}{\Delta x / \Delta y} = \frac{1}{\lim_{\Delta x \to 0} \frac{\Delta x}{\Delta y}}.$$

Since the direct function $x = x(y)$ is continuous, $\Delta x \to 0$ corresponds to $\Delta y \to 0$ as well, so that the limit in the denominator of the above formula can be replaced with $\Delta y \to 0$, which would give the derivative $x'(y) = x'_y$ of $x(y)$ with respect to y, which proves the theorem. **Q.E.D.**

As an example, let us derive a formula for the derivative of the square root function $y = \sqrt{x}$. Indeed, in this case $y(x)$ is the inverse of the function $x(y) = y^2$. Therefore, $x'_y = 2y$. Here we have used (3.5) for the derivative of the integer power function ($n = 2$ in our case). Hence, $y' = 1/x'_y = 1/(2y) = 1/\left(2\sqrt{x}\right)$. Note that this result can also be written as $\left(x^{1/2}\right)' = \frac{1}{2}x^{-1/2}$, which formally agrees with Eq. (3.5) used for $n = 1/2$.

Problem 3.8. *Show, using a combination of various rules for differentiation discussed above, that the formula (3.5) proved for positive powers can in fact be generalised to any rational powers as well, i.e. that $(x^\alpha)' = \alpha x^{\alpha-1}$, where $\alpha = p/q$ is a rational number (p and q are positive and/or negative integers, $q \neq 0$).* [*Hint*: the rational power $x^{p/q}$ can always be considered as a composite function $\left(x^{1/q}\right)^p$ with positive q. Prove first, using the inverse function rule, that $\left(x^{1/q}\right)' = \frac{1}{q}x^{1/q-1}$. Then, combine this result with the chain rule to extend Eq. (3.5) to all rational powers.]

3.3 Derivatives of Elementary Functions

The rules for differentiation considered above provide us with enough muscle to tackle any function written in analytical form. But first we need to derive formulae for the derivatives of all elementary functions.

General Power Function. $y = x^\alpha$ (α is a real number).

We have already considered positive and negative values of α and proved that in these cases $(x^\alpha) = \alpha x^{\alpha-1}$. Using the definition of the inverse function and the inverse function rule, we also discussed that this formula is valid for root powers $1/q$ $(q \neq 0)$ as well, and then, employing the chain rule, we extended this result for any rational power $\alpha = p/q$. It is clear then that it must be valid for any real powers α since, as we already discussed in Sects. 2.3 and 2.4.3, any irrational number can be bracketed by two infinite sequences of rational numbers converging to it, and so this can be done for the derivative as well. So, for any real α,

$$(x^\alpha)' = \alpha x^{\alpha-1}. \tag{3.18}$$

Problem 3.9. *Prove that*

$$n2^{n-1} = \sum_{k=1}^{n} k \binom{n}{k}.$$

[*Hint*: differentiate the binomial expansion of $(1+x)^n$ with respect to x.]

Logarithmic Functions. $y = \log_a x$ and $y = \ln x$.

We write:

$$(\log_a x)' = \lim_{\Delta x \to 0} \frac{1}{\Delta x} \left[\log_a (x + \Delta x) - \log_a x\right] = \lim_{\Delta x \to 0} \frac{1}{\Delta x} \log_a \frac{x + \Delta x}{x}$$
$$= \lim_{\Delta x \to 0} \frac{1}{\Delta x} \log_a \left(1 + \frac{\Delta x}{x}\right) = \lim_{\Delta x \to 0} \log_a \left(1 + \frac{\Delta x}{x}\right)^{1/\Delta x}.$$

Denoting $\lambda = 1/\Delta x$, which tends to infinity as $\Delta x \to 0$, we can rewrite the limit as

$$(\log_a x)' = \lim_{\lambda \to \infty} \log_a \left(1 + \frac{1/x}{\lambda}\right)^\lambda = \log_a \left[\lim_{\lambda \to \infty} \left(1 + \frac{1/x}{\lambda}\right)^\lambda\right].$$

We have already calculated the limit inside the square brackets before in Eq. (2.85), and it is equal to $e^{1/x}$. Therefore,

$$(\log_a x)' = \log_a e^{1/x} = \frac{1}{x} \log_a e = \frac{1}{x \ln a}. \tag{3.19}$$

In the last step we made use of Eq. (2.38) to express $\log_a e$ via the logarithm with respect to the base e, i.e. $\ln e = \log_a e \ln a$, which immediately gives the required relationship due to the fact that $\ln e = 1$. In a very important case of the base $a = e$, we then obtain:

$$(\ln x)' = \frac{1}{x}. \tag{3.20}$$

Exponential Functions. $y = a^x$ and $y = e^x$.

To work out the derivative of $y = a^x$, it is convenient to recall that it can be considered as an inverse of the logarithmic function $x = \log_a y$. Therefore, using Eq. (3.17), we immediately have:

$$(a^x)' = \frac{1}{(\log_a y)'_y} = \frac{1}{1/(y \ln a)} = y \ln a = a^x \ln a. \tag{3.21}$$

In particular, if $a = e$, we obtain a very important result:

$$(e^x)' = e^x, \tag{3.22}$$

i.e. the derivative of the e^x function is equal to itself. It is the only function which has this peculiar property.

> **Problem 3.10.** *Prove formula (3.18) by representing $y = x^\alpha$ as*
>
> $$y = e^{\ln(x^\alpha)} = e^{\alpha \ln x}.$$
>
> This is another way of proving the formula for the derivative of the power function with arbitrary real power α.

Trigonometric Functions. Let us first consider the sine function:

$$\Delta y = \sin(x + \Delta x) - \sin x = 2\cos\left(x + \frac{\Delta x}{2}\right)\sin\frac{\Delta x}{2},$$

where we have used Eq. (2.49) for the difference of two sine functions. Therefore,

$$\begin{aligned}(\sin x)' &= \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0}\left[\cos\left(x + \frac{\Delta x}{2}\right)\frac{\sin(\Delta x/2)}{\Delta x/2}\right] \\ &= \left[\lim_{\Delta x \to 0}\cos\left(x + \frac{\Delta x}{2}\right)\right]\left[\lim_{\Delta x \to 0}\frac{\sin(\Delta x/2)}{\Delta x/2}\right] = \cos x\left[\lim_{\lambda \to 0}\frac{\sin\lambda}{\lambda}\right],\end{aligned}$$

where in the last passage we replaced the $\Delta x \to 0$ limit with the equivalent $\lambda = \Delta x/2 \to 0$ limit. The last limit we have already seen before in Eq. (2.86), and it is equal to 1. Therefore, we finally obtain:

$$(\sin x)' = \cos x. \tag{3.23}$$

The derivative of the cosine function can be obtained using the chain rule. Indeed, $\cos x = \sin(\pi/2 - x)$, i.e. $y = \cos x$ can be written as $y = \sin v$ and $v = \pi/2 - x$, so that

$$(\cos x)' = (\sin v)'_v \left(\frac{\pi}{2} - x\right)' = (\cos v)(-1) = -\cos v = -\cos\left(\frac{\pi}{2} - x\right) = -\sin x. \tag{3.24}$$

Problem 3.11. *Prove the same formula using directly the definition of the derivative and formula (2.51).*

Problem 3.12. *Prove that*

$$(\tan x)' = \frac{1}{\cos^2 x} \tag{3.25}$$

and

$$(\cot x)' = -\frac{1}{\sin^2 x} \tag{3.26}$$

using the definitions of the tangent and cotangent functions and the basic rules for differentiation.

Inverse Trigonometric Functions. We shall first consider $y = \arcsin x$. It is the inverse function to $x = \sin y$. Therefore, using Eq. (3.17), we obtain:

$$(\arcsin x)' = \frac{1}{x'_y} = \frac{1}{(\sin y)'_y} = \frac{1}{\cos y} = \frac{1}{\cos(\arcsin x)} = \frac{1}{\sqrt{1-x^2}}, \tag{3.27}$$

where we have used Eq. (2.74).

Problem 3.13. *Prove that*

$$(\arccos x)' = -\frac{1}{\sqrt{1-x^2}}, \tag{3.28}$$

$$(\arctan x)' = \frac{1}{1+x^2}. \tag{3.29}$$

You may need identities (2.74) and (2.76) here.

Hyperbolic Functions. These functions are composed of exponential functions, and one can use the rules of differentiation to calculate the required derivatives. As an example, let us consider the hyperbolic sine, $y = \sinh(x)$:

$$[\sinh(x)]' = \left[\frac{1}{2}\left(e^x - e^{-x}\right)\right]' = \frac{1}{2}\left[(e^x)' - (e^{-x})'\right] = \frac{1}{2}\left(e^x + e^{-x}\right) = \cosh(x). \tag{3.30}$$

Problem 3.14. *Prove the following formulae:*

$$[\cosh(x)]' = \sinh(x), \tag{3.31}$$

$$[\tanh(x)]' = \frac{1}{\cosh^2(x)}, \tag{3.32}$$

$$[\coth(x)]' = -\frac{1}{\sinh^2(x)}. \tag{3.33}$$

Problem 3.15. *Prove the following formulae for the inverse hyperbolic functions:*

$$\left[\sinh^{-1}(x)\right]' = \frac{1}{\sqrt{x^2+1}},$$

$$\left[\cosh^{-1}(x)\right]' = \frac{1}{\sqrt{x^2-1}},$$

$$\left[\tanh^{-1}(x)\right]' = \left[\coth^{-1}(x)\right]' = \frac{1}{1-x^2}.$$

By comparing these equations with the corresponding formulae for the trigonometric functions, one can appreciate their similarity. The reason for this is quite profound and lies in the fact, proven in theory of complex functions, that the exponential and trigonometric functions are closely related.

3.4 Approximate Representations of Functions

Using formula (3.9) we can derive a number of very useful approximate representations of functions. These can be used for their numerical calculation, but also for all sorts of derivations.

As an example, let us show that for small x

$$\sin x \simeq x. \tag{3.34}$$

Indeed, consider the change of the function $y = \sin x$ when x changes by Δx:

$$\Delta y = \sin(x + \Delta x) - \sin x \simeq \frac{dy}{dx}\Delta x = (\cos x)\,\Delta x$$
$$\Rightarrow \quad \sin(x + \Delta x) \simeq \sin x + (\cos x)\,\Delta x.$$

Taking $x = 0$ in the last formula, we obtain $\sin \Delta x \simeq \Delta x$ as required.

Problem 3.16. *Prove the following approximate formulae valid for $x \ll 1$:*

$$\sqrt{1+x} \simeq 1 + \frac{x}{2}, \tag{3.35}$$

$$\frac{1}{1 \pm x} \simeq 1 \mp x, \tag{3.36}$$

$$\tan x \simeq x, \tag{3.37}$$

$$e^x \simeq 1 + x, \tag{3.38}$$

$$\ln(1+x) \simeq x. \tag{3.39}$$

We shall learn later on how one can build up systematic corrections to the approximate expressions (3.34)–(3.39) given above; these contain additional terms with higher powers of x and hence are valid not only for very small values of x, but also for x deviating significantly from zero.

3.5 Differentiation in More Difficult Cases

Direct specification of a function via $y = y(x)$ is not the only one possible. A function $y(x)$ can also be specified parametrically, e.g.

$$\begin{cases} x = x(t) \\ y = y(t) \end{cases} . \tag{3.40}$$

These types of equations may correspond to the x and y coordinates of a particle moving on a plane, in which case t (the parameter) is time. If we are interested in the derivative $y'(x)$, then it can be calculated as follows:

$$\begin{aligned} y'(x) &= \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{y(t+\Delta t) - y(t)}{x(t+\Delta t) - x(t)} = \lim_{\Delta t \to 0} \frac{\frac{y(t+\Delta t)-y(t)}{\Delta t}}{\frac{x(t+\Delta t)-x(t)}{\Delta t}} \\ &= \frac{\lim_{\Delta t \to 0} \frac{y(t+\Delta t)-y(t)}{\Delta t}}{\lim_{\Delta t \to 0} \frac{x(t+\Delta t)-x(t)}{\Delta t}} = \frac{y'(t)}{x'(t)} . \end{aligned} \tag{3.41}$$

Here at the first step we explicitly indicated that the change of both y and x is due to the change of t by Δt; that is why Δy was replaced by $y(t+\Delta t) - y(t)$ and Δx by $x(t+\Delta t) - x(t)$. At the second step we changed the limit from $\Delta x \to 0$ to $\Delta t \to 0$ which is possible to do because we assume that the functions $y(t)$ and $x(t)$ are continuous (recall one of the definitions of the continuity of a function!).

At the last step we applied the limit separately to the nominator and denominator of the fraction; this is allowed because of the properties of the limit discussed in the previous chapter.

Problem 3.17. *Prove the same result (3.41) by assuming that there exists an inverse $t = t(x)$ of the function $x = x(t)$, i.e. that the function $y = y(t) = y(t(x))$ is a composite function.* [*Hint*: use the chain rule (3.16) and Eq. (3.17) for the derivative of an inverse function.]

Example 3.3. ► An inter-galactic rocket moves along a spiral trajectory away from the Sun according to the equations $x(t) = e^t \cos t$ and $y(t) = e^t \sin t$. Determine the slope of the tangent line to the trajectory as a function of time.

Solution. The slope to the trajectory will be given by the derivative $y'(x)$. Therefore, we obtain:

$$y'(x) = \frac{y'(t)}{x'(t)} = \frac{e^t \sin t + e^t \cos t}{e^t \cos t - e^t \sin t} = -\frac{1 + \cot t}{1 - \cot t}. \blacktriangleleft$$

Problem 3.18. *Find the derivative $y'(x)$ of the function specified parametrically as $y = t^2 e^{-2t}$ and $x = \left(1 + t^2\right) e^{-2t}$.* [*Answer:*$dy/dx = t(1-t)/(-1+t-t^2)$.]

A function $y = y(x)$ can also be specified implicitly via an equation $F(x, y) = 0$ that cannot be solved with respect to y (and hence the direct calculation of the derivative is impossible). Still, calculation of the derivative $y'(x)$ can be performed even in this case (although it is not guaranteed that the derivative can be obtained in a closed form).

Example 3.4. ► Let the function $y = y(x)$ be specified via the equation $x^2+y^2 = R^2$ (a circle of radius R and centred at the origin). Calculate the derivative $y'(x)$.

Solution. We differentiate both sides of the equation with respect to x, treating the $y^2 = (y(x))^2$ as a composite function and hence using the chain rule. Then in the right-hand side of $x^2 + y^2 = R^2$ we have zero, while in the left-hand side $\left(x^2 + y^2\right)' = 2x + 2yy'(x)$, i.e. we obtain an equation

$$2x + 2yy'(x) = 0,$$

solving which with respect to the derivative finally yields: $y'(x) = -x/y(x)$. It is easy to see that this is the correct result. Indeed, in this particular case we can directly solve the equation of the circle and obtain $y(x)$ explicitly as $y(x) = \pm\sqrt{R^2 - x^2}$, where the plus sign corresponds to the upper part of the circle ($y \geq 0$), while the minus sign corresponds to the lower part of it ($y < 0$). This expression can be differentiated easily using the chain rule:

$$y' = \pm\frac{1}{2}\left(R^2 - x^2\right)^{-1/2}(-2x) = \mp\frac{x}{\sqrt{R^2 - x^2}} = -\frac{x}{\pm\sqrt{R^2 - x^2}} = -\frac{x}{y},$$

i.e. the same result.◀

Problem 3.19. *Find the derivative $y'(x)$ of the function specified via equation $y^4 + 2xy^2 + e^{-x}y = 0$.* [*Answer*: $y' = y\left(e^{-x} - 2y\right)/\left(4y^3 + 4xy + e^{-x}\right)$.]

Finally, there is a class of functions for which a special method is used to perform differentiation: these are the exponential functions containing the base which is a function of x itself, i.e. $y(x) = u(x)^{v(x)}$. In these cases one first takes a logarithm of both sides turning the equation into an implicit with respect to y, and then differentiates both sides.

For instance, let us calculate the derivative of $y = x^x$. After taking the natural logarithm of both sides, we obtain: $\ln y = x\ln x$. Differentiating both sides (remember to use the chain rule when differentiating the left-hand side as $y = y(x)$), we obtain:

$$\frac{1}{y}y' = \ln x + x\frac{1}{x} = \ln x + 1 \quad \Rightarrow \quad y' = (x^x)' = y\,(1 + \ln x) = x^x\,(1 + \ln x)\,.$$

Problem 3.20. *Derive the general formula*

$$\left(u(x)^{v(x)}\right)' = v'u^v \ln u + u'u^{v-1}v$$

for the derivative of the function $y = u(x)^{v(x)}$.

Problem 3.21. *Obtain the same result by representing the function $y(x) = u(x)^{v(x)}$ as an exponential.*

Problem 3.22. *Find the derivative $y'(x)$ of the functions specified in the following way: (a) $y = \ln t$, $x = e^{-2t}\sinh t$; (b) $y^{-3} + 2e^{-2xy} = 1$; (c) $y = \left(1 + 2x^2\right)^{3-3x^2}$; (d) $y = \sqrt{x + \sqrt{x + \sqrt{x + y}}}$ (in the latter case consider two methods: first, differentiate directly using the chain rule applied several times; second, rearrange the equation in such a way as to get rid of the radicals completely and then differentiate).* [*Answers*: (a) $e^{2t}/\left[t\cosh(t) - 2t\sinh(t)\right]$; (b) $-y/\left[x + (3/4)\,e^{2xy}y^{-4}\right]$; (c) $6xy\left[-\ln\left(1 + 2x^2\right) + 2\left(1 - x^2\right)/\left(1 + 2x^2\right)\right]$; (d) A/B, where $B = 1 - 8yg\sqrt{x + y}$, $A = -\left[1 + 2\sqrt{x + y}\,(2g + 1)\right]$, and $g = y^2 - x$.]

3.6 Higher Order Derivatives

The derivative $y'(x)$ is a function of x as well and hence it can also be differentiated, which is denoted $y'' = (y')'$ ("double prime"). The following terminology is used: y' is called the first order derivative, while y'' is the second order. Obviously, this process can be continued and one can talk about derivatives of the n-th order, denoted $y^{(n)}(x)$ (n is enclosed in brackets to distinguish it from the n-th power). Obviously, $y^{(n+1)} = \left(y^{(n)}\right)'$. Similar notations exist for the higher order derivatives based on the symbols of differentials:

$$y'' = \frac{d^2y}{dx^2}\ ,\ y^{(3)} = y''' = \frac{d^3y}{dx^3}\ ,\ \dots\ ,\ y^{(n)} = \frac{d^ny}{dx^n}.$$

Higher order derivatives are frequently used in applications. Many processes are described by equations (the so-called differential equations) involving derivatives of different orders. In mechanics acceleration $a(t) = v'(t)$ is the rate of change of the velocity; at the same time, since the velocity $v(t) = x'(t)$ relates to the rate of change of the position, the acceleration can also be considered as the second order derivative of the position: $a(t) = v'(t) = (x'(t))' = x''(t)$.

Another important application of first and second order derivatives is in analysing functions. We shall consider this particular point in detail later on in Sect. 3.10.

Example 3.5. ► Calculate the 5-th order derivative of $y = \cos x$. Then, derive the formula for the n-th order derivative of the cosine function.

Solution. we have: $(\cos x)' = -\sin x$; $(\cos x)'' = (-\sin x)' = -\cos x$; $(\cos x)''' = (-\cos x)' = \sin x$, then $(\cos x)^{(4)} = (\sin x)' = \cos x$, and, finally, $(\cos x)^{(5)} = (\cos x)' = -\sin x$. We see that the 5-th derivative is the same as the first. A general formula can be derived if we notice that $(\cos x)' = -\sin x = \cos(x + \pi/2)$, $(\cos x)'' = -\cos x = \cos(x + 2\pi/2)$, $(\cos x)^{(3)} = \sin x = \cos(x + 3\pi/2)$, etc. We observe that each time a phase of $\pi/2$ is added to the cosine function. Using the method of induction, we can now prove the general formula:

$$(\cos x)^{(n)} = \cos\left(x + \frac{\pi}{2}n\right). \tag{3.42}$$

Indeed, the formula works for the cases of $n = 1, 2, 3$. We assume that it also works for some arbitrary n. Let us check if it would work for the next value $n + 1$. We have:

$$\begin{aligned}(\cos x)^{(n+1)} &= \left((\cos x)^{(n)}\right)' = \left(\cos\left(x + \frac{\pi}{2}n\right)\right)' \\ &= -\sin\left(x + \frac{\pi}{2}n\right) = \cos\left(x + \frac{\pi}{2}n + \frac{\pi}{2}\right) = \cos\left(x + \frac{\pi}{2}(n+1)\right),\end{aligned}$$

which is exactly the desired result, i.e. (3.42) for $n \to n + 1$. The formula is proven. ◄

Problem 3.23. *Prove the following formulae:*

$$(\sin x)^{(n)} = \sin\left(x + \frac{\pi}{2}n\right), \tag{3.43}$$

$$\left(e^{\alpha x}\right)^{(n)} = \alpha^n e^{\alpha x}, \tag{3.44}$$

and, assuming a general real α*,*

$$\left(x^{\alpha}\right)^{(n)} = \alpha\,(\alpha - 1)\,(\alpha - n)\cdots(\alpha - n + 1)\,x^{\alpha - n}. \tag{3.45}$$

If $\alpha = m$ *is a positive integer, then, in particular,*

$$\left(x^m\right)^{(n)} = \frac{m!}{(m-n)!}x^{m-n} \quad \textit{if} \quad m \geq n \quad \textit{and} \quad \left(x^m\right)^{(n)} = 0 \quad \textit{if} \quad m < n.$$

Problem 3.24. *Prove the formula:*

$$\lim_{z\to 0} \frac{d^{2n}}{dz^{2n}}\left(z^2 - 1\right)^{2n} = (-1)^n \left(\frac{(2n)!}{n!}\right)^2. \tag{3.46}$$

Very often it is necessary to calculate the n-th derivative of a product of two functions. We shall now consider a formula due to Leibniz, which allows one to perform such a calculation. This result is especially useful if one of the functions in the product is a finite order polynomial as will become apparent later on.
Let $f(x) = u(x)v(x)$. Then we shall show that

$$\frac{d^n(uv)}{dx^n} = (uv)^{(n)} = uv^{(n)} + \binom{n}{1}u'v^{(n-1)} + \binom{n}{2}u''v^{(n-2)} + \cdots + \binom{n}{k}u^{(k)}v^{(n-k)}$$
$$+ \cdots + u^{(n)}v = \sum_{k=0}^{n}\binom{n}{k}u^{(k)}v^{(n-k)}, \tag{3.47}$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ are *the binomial coefficients* which we met in Sect. 1.10, and we implied that $u^{(0)} = u$ and $v^{(0)} = v$ are the functions themselves. In fact, we see that the Leibniz formula looks strikingly similar to the *binomial theorem* (1.71) and therefore is sometimes written in the following symbolic form:

$$\frac{d^n(uv)}{dx^n} = (uv)^{(n)} = (u + v)^{(n)}. \tag{3.48}$$

Let us first prove this result using a method very similar to the one we used in Sect. 1.10 when proving the binomial theorem. It is based on induction. Firstly, we check that the formula works for some initial values of $n = 1, 2, 3$ (in fact, only one value, e.g. $n = 1$, would suffice):

$$\begin{aligned}
(uv)' &= u^{(1)}v + uv^{(1)} = u^{(1)}v^{(0)} + u^{(0)}v^{(1)} \quad \text{(i.e. familiar } u'v + uv'\text{)},\\
(uv)'' &= u^{(2)}v^{(0)} + 2u^{(1)}v^{(1)} + u^{(0)}v^{(2)},\\
(uv)''' &= \left(u^{(3)}v^{(0)} + u^{(2)}v^{(1)}\right) + 2\left(u^{(2)}v^{(1)} + u^{(1)}v^{(2)}\right) + \left(u^{(1)}v^{(2)} + u^{(0)}v^{(3)}\right)\\
&= u^{(3)}v^{(0)} + 3u^{(2)}v^{(1)} + 3u^{(1)}v^{(2)} + u^{(0)}v^{(3)},
\end{aligned}$$

which are all consistent with Eqs. (3.47) or (3.48), since

$$\binom{3}{0} = \frac{3!}{0!3!} = 1,\ \binom{3}{1} = \frac{3!}{1!2!} = 3,\ \binom{3}{2} = \frac{3!}{2!1!} = 3 \text{ and } \binom{3}{3} = \frac{3!}{0!3!} = 1.$$

Recall that $0! = 1$ by definition. Next we assume that the formula (3.47) is valid for some value n. We should now prove from it that Eq. (3.47) holds also for the next value of n, i.e. for $n \to n + 1$. Differentiating both sides of Eq. (3.47), we get:

$$\begin{aligned}
(uv)^{(n+1)} &= \left[\sum_{k=0}^{n}\binom{n}{k}u^{(k)}v^{(n-k)}\right]' = \sum_{k=0}^{n}\binom{n}{k}\left[u^{(k+1)}v^{(n-k)} + u^{(k)}v^{(n-k+1)}\right]\\
&= \sum_{k=0}^{n}\binom{n}{k}u^{(k+1)}v^{(n-k)} + \sum_{k=0}^{n}\binom{n}{k}u^{(k)}v^{(n-k+1)}.
\end{aligned}$$

In the first term we make a substitution of the summation index, $l = k + 1$, which gives (after replacing l with k again):

$$(uv)^{(n+1)} = \sum_{k=1}^{n+1}\binom{n}{k-1}u^{(k)}v^{(n-k+1)} + \sum_{k=0}^{n}\binom{n}{k}u^{(k)}v^{(n-k+1)}.$$

Then, separate out the last (with $k = n + 1$) term in the first sum and the first (the one with $k = 0$) term in the second; combine the others together:

$$\begin{aligned}
(uv)^{(n+1)} = \binom{n}{n}u^{(n+1)}v^{(0)} + \sum_{k=1}^{n}\left[\binom{n}{k-1} + \binom{n}{k}\right]u^{(k)}v^{(n-k+1)}\\
+ \binom{n}{0}u^{(0)}v^{(n+1)}.
\end{aligned} \tag{3.49}$$

Because of Eqs. (1.74) and (1.75), we can finally rewrite Eq. (3.49) as follows:

$$
\begin{aligned}
(uv)^{(n+1)} &= \binom{n+1}{0} u^{(0)} v^{(n+1)} + \sum_{k=1}^{n} \binom{n+1}{k} u^{(k)} v^{(n+1-k)} \\
&\quad + \binom{n+1}{n+1} u^{(n+1)} v^{(0)} = \sum_{k=0}^{n+1} \binom{n+1}{k} u^{(k)} v^{(n+1-k)},
\end{aligned}
$$

which is the desired result as it looks exactly as Eq. (3.47) written for $n \to n+1$.

There is another proof of the Leibniz formula which is worth giving here as it is based on rather different ideas. Differentiation of a function, $f'(x)$, can be formally considered as an action of an *operator*, $\frac{d}{dx}$, on the function $f(x)$, i.e.

$$
f'(x) = \frac{df}{dx} = \frac{d}{dx} f(x).
$$

The n-th derivative, $f^{(n)}(x)$, can then be regarded as an action of the same operator n times:

$$
f^{(n)}(x) = \left(\frac{d}{dx}\right)^n f(x).
$$

Let us now use these types of ideas to prove the Leibniz formula. Let us introduce two operators, $D_u = \left(\frac{d}{dx}\right)_u$ and $D_v = \left(\frac{d}{dx}\right)_v$, which differentiate only the functions $u(x)$ and $v(x)$, respectively. Then, the first derivative of the product uv can be written as

$$
(uv)' = (D_u + D_v)\, uv = (D_u u)\, v + u\, (D_v v) = u^{(1)} v^{(0)} + u^{(0)} v^{(1)};
$$

the second derivative is obtained by acting with the same operator twice:

$$
\begin{aligned}
(uv)'' &= (D_u + D_v)(D_u + D_v)\, uv = (D_u + D_v)^2\, uv \\
&= \left(D_u^2 + D_u D_v + D_v D_u + D_v^2\right) uv = \left(D_u^2 + 2 D_u D_v + D_v^2\right) uv,
\end{aligned}
$$

where in the last step we recognised that the order in which operators D_u and D_v act on the product of functions u and v does not matter (since they act on different functions anyway), and hence we combined them together: $D_u D_v + D_v D_u = 2 D_u D_v$. Thus, we obtain:

$$
\begin{aligned}
(uv)'' &= \left(D_u^2 + 2 D_u D_v + D_v^2\right) uv \\
&= \left(D_u^2 u\right) v + 2\,(D_u u)\,(D_v v) + u\left(D_v^2 v\right) = u^{(2)} v^{(0)} + 2 u^{(1)} v^{(1)} + u^{(0)} v^{(2)}.
\end{aligned}
$$

This process can be continued; when we want to calculate the n-th derivative, we will have to act with the operator $(D_u + D_v)^n$ on the product uv. It must be clear now what we ought to do in the general case: since the order of operators is not important, we can treat them like numbers and expand the operator $(D_u + D_v)^n$ formally using the binomial formula. Then, associating $D_u^k u$ with $u^{(k)}$ and similarly $D_v^k v$ with $v^{(k)}$, we can write:

$$(uv)^{(n)} = (D_u + D_v)^n \, uv = \sum_{k=0}^{n} \binom{n}{k} D_u^k D_v^{n-k} uv$$
$$= \sum_{k=0}^{n} \binom{n}{k} \left(D_u^k u\right) \left(D_v^{n-k} v\right) = \sum_{k=0}^{n} \binom{n}{k} u^{(k)} v^{(n-k)},$$

which is exactly the correct expression (3.47). This particular proof shows quite clearly why the Leibniz formula is so similar to the binomial formula. At the same time, we should not forget that this resemblance might be misleading: if in the binomial formula for $(a+b)^n$ we have powers and hence a^0 and b^0 are both equal to one and therefore do not need to be written explicitly, in the Leibniz formula for $(uv)^{(n)}$ the "powers" in $u^{(k)}$ and $v^{(n-k)}$ correspond to the order of the derivatives and thus $u^{(0)}$ and $v^{(0)}$ coincide with the functions $u(x)$ and $v(x)$ themselves and are *not* equal to one; these should be written explicitly whenever they appear.

Example 3.6. ► Using Leibniz formula, obtain a formula for the n-th derivative of the function $f(x) = \left(x^2 + 1\right) v(x)$.

Solution. Here $u(x) = x^2 + 1$ and hence can only be differentiated twice: $u^{(1)} = 2x$ and $u^{(2)} = 2$, while all higher derivatives give zero. Therefore, the Leibniz formula contains only the first three terms corresponding to the summation index $k = 0, 1, 2$:

$$f^{(n)}(x) = \binom{n}{0} u^{(0)} v^{(n)} + \binom{n}{1} u^{(1)} v^{(n-1)} + \binom{n}{2} u^{(2)} v^{(n-2)}$$
$$= uv^{(n)} + 2nxv^{(n-1)} + n(n-1)v^{(n-2)},$$

where we have used explicit expressions for the first three binomial coefficients: $\binom{n}{0} = 1$, $\binom{n}{1} = n$ and $\binom{n}{2} = n\,(n-1)\,/2$. ◄

Problem 3.25. *Consider a function $\varphi(x) = (1 - x^3)f(x)$. The n-th derivative of $\varphi(x)$ can be written in the following form:*

$$\varphi^{(n)}(x) = a_0(1 - x^3)f^{(n)} + a_1 x^2 f^{(n-1)} + a_2 x f^{(n-2)} + a_3 f^{(n-3)}.$$

Show, using the Leibniz formula (3.47), that the numerical values of the coefficients are: $a_0 = 1$, $a_1 = -3n$, $a_2 = -3n(n-1)$ and $a_3 = -n(n\text{-}1)(n-2)$.

Problem 3.26. *Show that the n-th derivative of the function $f(x) = x^3 \sin(2x)$, assuming n is even, is:*

$$f^{(n)}(x) = (-1)^{n/2}\, 2^{n-3} \left\{2x\left[4x^2 - 3n(n-1)\right]\sin 2x \right. \\ \left. +n\left[-12x^2 + (n-1)(n-2)\right]\cos 2x\right\}.$$

In some cases the calculation of the n-th derivative could involve some tricks. We shall illustrate this by considering the n-th derivative of the $y = \arcsin(x)$ function. The first derivative $y' = \left(1 - x^2\right)^{-1/2}$. The calculation of the second derivative would result in two terms as one has to use the chain rule; more terms arise as we attempt to calculate more derivatives and there seems to be no general rule as how to do the n-th derivative (try this!). Direct application of the Leibniz formula here does not seem to help either. The solution to this problem appears to be possible if we notice that

$$y'(x) = (1-x)^{-1/2}\,(1+x)^{-1/2} = u(x)v(x),$$

so that, applying the Leibniz formula now, we can write:

$$y^{(n+1)}(x) = \left(y'\right)^{(n)} = (uv)^{(n)} = \sum_{k=0}^{n} \binom{n}{k} u^{(k)} v^{(n-k)},$$

where (e.g. by induction)

$$u^{(k)} = \left[(1-x)^{-1/2}\right]^{(k)} = \frac{1}{2}\cdot\frac{3}{2}\cdot\ldots\cdot\frac{2k-1}{2}\,(1-x)^{-(2k+1)/2} \\ = \frac{(2k-1)!!}{2^k}\,(1-x)^{-(2k+1)/2}.$$

Here the double factorial means a product of all odd integers, i.e. $(2k-1)!! = 1\cdot 3\cdot 5\cdot\ldots\cdot(2k-1)$. We also adopt that $(-1)!! = 1$, which makes the formula for $u^{(k)}$ also formally valid for $k = 0$, i.e. $u^{(0)} = (1-x)^{-1/2} = u(x)$. The double factorial can always be expressed via single factorials and powers of 2 by inserting between odd integers the corresponding even ones:

$$(2k-1)!! = 1\cdot 3\cdot 5\cdot\ldots\cdot(2k-1) \\ = \frac{1\cdot 2\cdot 3\cdot 4\cdot 5\cdot 6\cdot\ldots\cdot(2k-1)\cdot(2k)}{2\cdot 4\cdot 6\cdot\ldots\cdot(2k)} = \frac{(2k)!}{2^k k!}, \tag{3.50}$$

which yields

$$u^{(k)} = \frac{(2k)!}{4^k k!}\,(1-x)^{-(2k+1)/2} = \frac{1}{\sqrt{1-x}}\,\frac{(2k)!}{4^k k!}\,(1-x)^{-k}\,.$$

Correspondingly,

$$\begin{aligned} v^{(k)} &= \left[(1+x)^{-1/2}\right]^{(k)} = \left(-\frac{1}{2}\right)\cdot\left(-\frac{3}{2}\right)\cdot\ldots\cdot\left(-\frac{2k-1}{2}\right)(1+x)^{-(2k+1)/2} \\ &= (-1)^k\,\frac{(2k-1)!!}{2^k}\,(1+x)^{-(2k+1)/2} = \frac{(-1)^k}{\sqrt{1+x}}\,\frac{(2k)!}{4^k k!}\,(1+x)^{-k}\,, \end{aligned}$$

and therefore

$$v^{(n-k)} = \frac{(-1)^{n-k}}{\sqrt{1+x}}\,\frac{(2n-2k)!}{4^{n-k}\,(n-k)!}\,(1+x)^{-n+k}\,.$$

Combining the obtained derivatives of $u(x)$ and $v(x)$ in the Leibniz formula, we finally obtain:

$$\begin{aligned} [\arcsin(x)]^{(n+1)} &= \frac{1}{\sqrt{1-x^2}}\sum_{k=0}^{n}\binom{n}{k}\frac{(2k)!}{4^k k!}\,\frac{(2n-2k)!}{4^{n-k}\,(n-k)!}\times \\ &\quad\times(-1)^{n-k}\,(1-x)^{-k}\,(1+x)^{-n+k} \\ &= \frac{(1+x)^{-n}}{\sqrt{1-x^2}}\,\frac{(2n)!}{4^n n!}\sum_{k=0}^{n}(-1)^{n-k}\binom{n}{k}^2\binom{2n}{2k}^{-1}\left(\frac{1+x}{1-x}\right)^k, \end{aligned} \tag{3.51}$$

where we expressed all factorials via the binomial coefficients.

3.7 Taylor's Formula

Taylor's formula[1] allows representing a function as a finite degree polynomial and a remainder term to be considered as "small". It manifests itself as one of the main results of mathematical analysis. It is used in numerous analytical proofs and estimates. It is also useful for approximate calculations of functions.

Let us first do some preliminary work which will help later on. Consider a function $f(x)$ which is a polynomial of degree n:

$$f(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_n x^n. \tag{3.52}$$

[1] Named after Brook Taylor.

This polynomial contains powers of x, but can alternatively be written via powers of $(x-a)$ for any a as follows:

$$f(x) = A_0 + A_1(x-a) + A_2(x-a)^2 + \cdots + A_n(x-a)^n \,. \tag{3.53}$$

Problem 3.27. *Using the binomial formula for each of the powers $(x-a)^k$, $k = 1, \ldots, n$, it must be possible to have the $f(x)$ written via powers of x again. By comparing terms of the same powers of x with Eq. (3.52), derive explicit expressions relating the coefficients α_i and A_k. Use this method to show that*

$$\alpha_k = \sum_{i=k}^{n} A_i \binom{i}{k} (-a)^{i-k} \,.$$

Problem 3.28. *Using Eq. (3.53), it is seen that $f(a) = A_0$, then*

$$\begin{aligned}
f'(x) &= A_1 + 2A_2(x-a) + 3A_3(x-a)^2 + \cdots \implies f'(a) = A_1, \\
f''(x) &= 2A_2 + 3\cdot 2A_3(x-a) + 4\cdot 3A_4(x-a)^2 + \cdots \implies f''(a) = 2A_2, \\
f'''(x) &= 3\cdot 2A_3 + 4\cdot 3\cdot 2A_4(x-a) + 5\cdot 4\cdot 3A_4(x-a)^2 + \cdots \\
&\implies f'''(a) = 3\cdot 2A_3 = 6A_3,
\end{aligned}$$

etc. Show using induction that for any $k = 1, 2, \ldots, n$ one can write:

$$\begin{aligned}
f^{(k)}(x) &= k!A_k + \frac{(k+1)!}{1!}A_{k+1}(x-a) + \frac{(k+2)!}{2!}A_{k+2}(x-a)^2 + \cdots \\
&\implies f^{(k)}(a) = k!A_k,
\end{aligned}$$

so that

$$A_1 = f'(a) = \frac{f'(a)}{1!},\; A_2 = \frac{f''(a)}{2!}, \ldots,\; A_k = \frac{f^{(k)}(a)}{k!}, \ldots,\; A_n = \frac{f^{(n)}(a)}{n!}. \tag{3.54}$$

Problem 3.29. *Differentiate $f(x)$ in Eq. (3.52) to express derivatives of $f(x)$ at $x = a$ via the a_k coefficients. Hence, show that the coefficients A_k can be expressed via α_i as follows:*

$$A_k = \sum_{i=k}^{n} \alpha_i \binom{i}{k} a^{i-k}.$$

It follows from Eq. (3.54) that any polynomial of degree n can be actually written as a sum of powers of $(x-a)$ for any a:

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n$$
$$= \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k .$$

Of course, if $f(x)$ is not a polynomial, it can never be written via a finite sum of powers of $(x-a)$. Let us define the following n-th degree polynomial (Taylor's polynomial)

$$F_n(x,a) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n$$
$$= \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k , \qquad (3.55)$$

which is a function of two variables, x and a, and satisfies an obvious property $F_n(a,a) = f(a)$. Then it is clear that formally *any* function $f(x)$ can be written as

$$f(x) = F_n(x,a) + R_{n+1}, \qquad (3.56)$$

where R_{n+1}, called the *remainder term*, accounts for the full error we are making by replacing $f(x)$ with the Taylor's polynomial of degree n given by Eq. (3.55); of course, R_{n+1} depends on x as well. Taylor's theorem to be discussed below enables one to write an explicit formula for the remainder term for any $f(x)$.

But before formulating and proving Taylor's theorem, we need one more statement to be made, which, we hope, is intuitively obvious.[2] Consider some *continuous* function $f(x)$ within the interval $a \le x \le b$, and let the function have *identical* values at both ends: $f(a) = f(b)$. Assume first that the function $f(x)$ is not a constant (equal to its edge value). Then it is obvious that it will behave within the interval between a and b in such a way that it would have at least one minimum or maximum as illustrated in Fig. 3.2. Indeed, for instance, if the function goes down from the point $x = a$, it must eventually go up towards the point $x = b$ as its values at the edges are the same. If the function has a minimum or maximum at a point $x = \xi$, then its derivative is zero at that point, $f'(\xi) = 0$, hence the slope of the tangent line at that point is zero (it is horizontal). This is also illustrated in the figure (more discussion on the extrema will be given in the next Sect. 3.10). To prove this statement, let us suppose that the function has a maximum at point $x = \xi$. This means that $f(\xi) \ge f(\xi - \epsilon)$ and $f(\xi) \ge f(\xi + \epsilon)$ for a sufficiently small $\epsilon > 0$,

[2] In mathematics literature it is called Rolle's theorem after Michel Rolle.

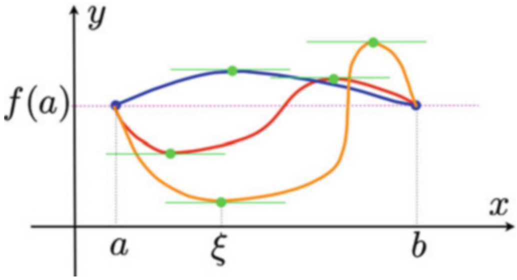


Fig. 3.2 A schematic which illustrates the point that a non-constant function $f(x)$ which has equal values at both ends of the interval $a \leq x \leq b$ will experience at least one (may be more) extremum within that interval. Here three functions are sketched, shown by *different colours*, all having the same values at the edge points a and b. The extrema (minima and/or maxima) of the functions are indicated by *green dots*; the corresponding tangent lines at these points are also shown by *horizontal green lines*

i.e. at the neighbouring points on the right and on the left of the point $x = \xi$ the function $f(x)$ has smaller or (possibly) equal values (definitely it cannot be larger at the neighbouring points). Therefore, we have:

$$\frac{f(\xi+\epsilon)-f(\xi)}{\epsilon} \leq 0 \quad \text{and} \quad \frac{f(\xi)-f(\xi-\epsilon)}{\epsilon} \geq 0.$$

Note that the first expression is non-positive, while the second is non-negative. In the limit of $\epsilon \to 0$ both of these expressions should give the derivatives $f'(\xi)$ from the right and left. Since the function $f(x)$ is continuous, these two derivatives should be equal to each other, and that means that they should be equal to zero as otherwise it is impossible to satisfy the two inequalities written above. A similar proof is given in the case of a minimum. Note that if the function $f(x)$ is constant on an interval, all points between its edges have zero derivative and this does not contradict the above made statement.

What is essential for us right now is that if a continuous function $f(x)$ does have equal values at two points, then there must be at least one point ξ between them where the first derivative of the function is zero: $f'(\xi) = 0$.

Taylor's theorem is based on properties of the following function:

$$\rho(t) = f(x) - F_n(x,t) - \frac{(x-t)^{n+1}}{(x-a)^{n+1}}\left[f(x) - F_n(x,a)\right]. \tag{3.57}$$

Let us first calculate this function at the points $t = a$ and $t = x$:

$$\begin{aligned}\rho(a) &= \left[f(x) - F_n(x,a)\right] - \frac{(x-a)^{n+1}}{(x-a)^{n+1}}\left[f(x) - F_n(x,a)\right] \\ &= \left[f(x) - F_n(x,a)\right] - \left[f(x) - F_n(x,a)\right] = 0\end{aligned}$$

and

$$\rho(x) = [f(x) - F_n(x,x)] - \frac{(x-x)^{n+1}}{(x-a)^{n+1}}[f(x) - F_n(x,a)] = f(x) - F_n(x,x) = 0,$$

as $F_n(x,x) = f(x)$ from its definition (3.55), as was already mentioned. We see that $\rho(a) = \rho(x) = 0$. Next, we calculate the derivative of $\rho(t)$:

$$\rho'(t) = -\frac{dF_n(x,t)}{dt} + (n+1)\frac{(x-t)^n}{(x-a)^{n+1}}[f(x) - F_n(x,a)], \tag{3.58}$$

where

$$\begin{aligned}
\frac{dF_n(x,t)}{dt} &= \frac{d}{dt}\left[f(t) + \sum_{k=1}^{n}\frac{f^{(k)}(t)}{k!}(x-t)^k\right] = f'(t) + \sum_{k=1}^{n}\frac{d}{dt}\left[\frac{f^{(k)}(t)}{k!}(x-t)^k\right] \\
&= f'(t) + \sum_{k=1}^{n}\left[\frac{f^{(k+1)}(t)}{k!}(x-t)^k + \frac{f^{(k)}(t)}{k!}(-k)(x-t)^{k-1}\right] \\
&= f'(t) + \sum_{k=1}^{n}\left[\frac{f^{(k+1)}(t)}{k!}(x-t)^k - \frac{f^{(k)}(t)}{(k-1)!}(x-t)^{k-1}\right] \\
&= f'(t) + \left[f''(t)(x-t) - f'(t)\right] + \left[\frac{1}{2}f'''(t)(x-t)^2 - f''(t)(x-t)\right] \\
&\quad + \left[\frac{1}{3!}f^{(4)}(t)(x-t)^3 - \frac{1}{2}f'''(t)(x-t)^2\right] + \cdots \\
&\quad + \left[\frac{f^{(n+1)}(t)}{n!}(x-t)^n - \frac{f^{(n)}(t)}{(n-1)!}(x-t)^{n-1}\right].
\end{aligned}$$

It is seen that the first term here is cancelled out with the third (the second term in the first square bracket), the second with the fifth, the fourth with the seventh, and so on; all terms cancel out this way apart from the first term in the last bracket, i.e.

$$\frac{dF_n(x,t)}{dt} = \frac{f^{(n+1)}(t)}{n!}(x-t)^n,$$

and hence from (3.58) it follows that

$$\begin{aligned}
\rho'(t) &= -\frac{f^{(n+1)}(t)}{n!}(x-t)^n + (n+1)\frac{(x-t)^n}{(x-a)^{n+1}}[f(x) - F_n(x,a)] \\
&= -\frac{f^{(n+1)}(t)}{n!}(x-t)^n + (n+1)\frac{(x-t)^n}{(x-a)^{n+1}}R_{n+1},
\end{aligned} \tag{3.59}$$

since R_{n+1} is the remainder term between the function $f(x)$ and the polynomial $F_n(x,a)$ by definition of Eq. (3.56).

Now we are ready to formulate and then prove the famous Taylor's theorem (formula).

Theorem 3.7 (Taylor). *Let $y = f(x)$ have at least $m + 1$ derivatives (i.e. $f^{(k)}(x)$ for $k = 1, \ldots, m + 1$) in some vicinity of point a. Then this function at some point x can be written as the following $(n + 1)$-degree polynomial (assuming $n \leq m$):*

$$f(x) = \left[f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n \right] + R_{n+1}(x) = F_n(x,a) + R_{n+1}(x), \quad (3.60)$$

where

$$R_{n+1}(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-a)^{n+1} \quad (3.61)$$

is the remainder term *with ξ being some point between x and a; of course, the point ξ generally depends on the choice of n.*

Proof. Let us choose a point $x > a$ in the vicinity of a (the case of $x < a$ is considered similarly) and let $F_n(x,a)$ be the first part of Taylor's formula before the remainder term. Then, consider a function $\rho(t)$ defined in Eq. (3.57). Since the function $\rho(t)$ has equal values (equal to zero) at the boundaries of the interval $a \leq t \leq x$, it has to have at least one extremum there, either a minimum or a maximum, i.e. there should exist a point ξ (where $a < \xi < x$) such that $\rho'(\xi) = 0$.

Setting the derivative from expression (3.59) at $t = \xi$ to zero, we obtain:

$$\rho'(\xi) = -\frac{f^{(n+1)}(\xi)}{n!}(x-\xi)^n + (n+1)\frac{(x-\xi)^n}{(x-a)^{n+1}}R_{n+1} = 0$$

$$\Longrightarrow \quad R_{n+1} = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-a)^{n+1},$$

which proves the theorem in view of Eq. (3.56). **Q.E.D.**

Note that it is often useful to write the intermediate point ξ explicitly as follows:

$$\xi = a + \vartheta(x-a), \quad \text{where} \quad 0 < \vartheta < 1.$$

This form guarantees that ξ lies between a and x. Note that the number ϑ depends on the points x and a.

There are several important particular cases of Taylor's formula worth mentioning. In the special case of $a = 0$ Taylor's formula takes a simplified form:

$$f(x) = f(0) + \frac{f'(0)}{1!}x + \frac{f''(0)}{2!}x^2 + \frac{f''(0)}{3!}x^2 + \cdots + \frac{f^{(n)}(0)}{n!}x^n + R_{n+1}(x), \tag{3.62}$$

which is called Maclaurin's formula. Here the remainder term is given by

$$R_{n+1}(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}x^{n+1} = \frac{f^{(n+1)}(\vartheta x)}{(n+1)!}x^{n+1}, \tag{3.63}$$

where $0 < \vartheta < 1$. The other important formula (due to Lagrange) is obtained from the Taylor's formula in the case of $n = 0$. Indeed, in this case the remainder term $R_1(x) = f'(\xi)(x-a)$, so that we can write:

$$\begin{aligned} &f(x) = f(a) + f'(\xi)(x-a) = f(a) + f'(a + \vartheta(x-a))(x-a), \\ &a < \xi < x, \ 0 < \vartheta < 1. \end{aligned} \tag{3.64}$$

This formula is found to be very useful when proving various results and we shall be using it often starting from the next section.

Let us now apply the Maclaurin formula to some elementary functions.

Example 3.7. ► Apply the Maclaurin formula to the exponential function $y = e^x$.

Solution. Since $(e^x)' = e^x$, it is clear that $(e^x)^{(i)} = e^x$ for any $i = 1, 2, \ldots, n$. Therefore, applying formula (3.62) and noting that $e^0 = 1$, we obtain:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots + \frac{x^n}{n!} + \frac{e^{\vartheta x}x^{n+1}}{(n+1)!} = \sum_{k=0}^{n}\frac{x^k}{k!} + \frac{e^{\vartheta x}x^{n+1}}{(n+1)!}. \blacktriangleleft \tag{3.65}$$

Problem 3.30. *Prove the following Maclaurin formulae:*

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots + (-1)^{n+1}\frac{x^n}{n} + R_{n+1} = \sum_{k=1}^{n}(-1)^{k+1}\frac{x^k}{k} + R_{n+1} \tag{3.66}$$

with

$$R_{n+1} = \frac{(-1)^n}{n+1}\frac{x^{n+1}}{(1+\vartheta x)^{n+1}};$$

$$\begin{aligned} \sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots + \frac{(-1)^{n+1}x^{2n-1}}{(2n-1)!} + R_{2n+1} \\ &= \sum_{k=1}^{n}\frac{(-1)^{k-1}x^{2k-1}}{(2k-1)!} + R_{2n+1} \end{aligned} \tag{3.67}$$

(continued)

Problem 3.30 (continued)
with

$$R_{2n+1} = \frac{(-1)^n \cos(\vartheta x)}{(2n+1)!} x^{2n+1};$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots + \frac{(-1)^n x^{2n}}{(2n)!} + R_{2n+2}$$
$$= \sum_{k=0}^{n} \frac{(-1)^k x^{2k}}{(2k)!} + R_{2n+2} \tag{3.68}$$

with

$$R_{2n+2} = \frac{(-1)^{n+1} \cos(\vartheta x)}{(2n+2)!} x^{2n+2};$$

$$(1+x)^\alpha = 1+\alpha x+\frac{\alpha(\alpha-1)}{2}x^2+\cdots+\frac{\alpha(\alpha-1)(\alpha-2)\cdots(\alpha-n+1)}{n!}x^n + R_{n+1}$$
$$= \sum_{k=0}^{n} \binom{\alpha}{k} x^k + R_{n+1}, \tag{3.69}$$

where

$$\binom{\alpha}{k} = \frac{\alpha(\alpha-1)(\alpha-2)\cdots(\alpha-k+1)}{k!} \tag{3.70}$$

are called generalised binomial coefficients *and*

$$R_{n+1} = \binom{\alpha}{n+1} (1+\vartheta x)^{\alpha-n-1} x^{n+1}.$$

The notation $\binom{\alpha}{k}$ *for the coefficients is not accidental. Indeed, show that for positive integer* $\alpha = m$ *the formula (3.69) becomes the binomial formula.* [*Hint*: first, check that the coefficients $\binom{m}{k}$ coincide with the binomial coefficients (1.72), and second, verify that the remainder term is zero in this case, i.e. the expansion contains a finite number of terms.]

In the following example we shall consider a much trickier case of $y = \arcsin(x)$. It is easy to see that a direct calculation of the derivatives of the arcsine function at $x = 0$ from Eq. (3.51) contains complicated finite sums with binomial coefficients and hence this method seems to lead us nowhere. The trick is in noticing that

$$y' = \frac{1}{\sqrt{1-x^2}} \quad \text{and} \quad y'' = \frac{x}{(1-x^2)^{3/2}} = y' \frac{x}{1-x^2} \quad \Longrightarrow \quad \left(1-x^2\right) y'' = xy'.$$

Problem 3.31. *Apply the Leibniz formula to both sides of the last identity to derive the following relationship:*

$$\left(1-x^2\right) y^{(n+2)} = (2n+1)\, xy^{(n+1)} + n^2 y^{(n)}.$$

This is a recurrence relation *as it allows calculating derivatives of higher orders recursively (i.e. one after another) from the previous ones.*

Now apply $x = 0$ in the above recurrence relation to get $y^{(n+2)}(0) = n^2 y^{(n)}(0)$, which can be used to derive recursively all the necessary derivatives at $x = 0$ needed to derive the Maclaurin formula for the arcsine function. We see that $y^{(2)}(0) = 0$ (corresponds to $n = 0$), $y^{(4)}(0) = 2^2 y^{(2)}(0) = 0$, etc., i.e. all the even order derivatives are equal to zero. Then, we only need to consider odd order derivatives. We know that $y^{(1)}(0) = 1$ from $y' = \left(1-x^2\right)^{-1/2}$; next, $y^{(3)}(0) = 1^2 y^{(1)}(0) = 1^2$, $y^{(5)}(0) = 3^2 y^{(3)}(0) = 3^2 1^2$, $y^{(7)}(0) = 5^2 y^{(5)}(0) = 5^2 3^2 1^2$. This process can be continued, but it should be clear by now (you can prove this rigorously by induction), that

$$y^{(2k+1)}(0) = [(2k-1)!!]^2 = \left[\frac{(2k)!}{2^k k!}\right]^2,$$

where we have used our previous result (3.50) for the double factorial. Now we are ready to write the Maclaurin formula for the arcsine function (recall, that we only need to keep the odd order derivatives and hence only odd powers of x are to be retained):

$$\arcsin(x) = \sum_{k=0}^{n} y^{(2k+1)}(0) \frac{x^{2k+1}}{(2k+1)!} + R_{2k+2} = \sum_{k=0}^{n} \frac{(2k)!}{4^k (2k+1) (k!)^2} x^{2k+1} + R_{2k+2}. \tag{3.71}$$

The constant term is missing here since $\arcsin(0) = 0$. The expression for the remainder term is rather cumbersome and is not given here. It can be worked out using Eq. (3.51) if desired.

Problem 3.32. *Apply the same method for the arccosine function* $y = \arccos(x)$ *and prove the Maclaurin formula for it:*

$$\arccos(x) = \frac{\pi}{2} - \sum_{k=0}^{n} \frac{(2k)!}{4^k (2k+1) (k!)^2} x^{2k+1} + R_{2k+2}. \tag{3.72}$$

Problem 3.33. *Apply a similar method for obtaining the Maclaurin series for the arctangent function:*

$$\arctan(x) = \sum_{k=0}^{n} \frac{(-1)^k}{2k+1} x^{2k+1} + R_{2k+2}. \tag{3.73}$$

[*Hint*: since $y' = [\arctan(x)]' = \left(1+x^2\right)^{-1}$ does not contain radicals, it is sufficient to consider the identity $\left(1+x^2\right) y' = 1$ and then differentiate both sides of it n times by means of the Leibniz formula (3.47).]

Compare the two Maclaurin formulae for the arcsine and arccosine with the identity (2.72): they would be identical if not for the different remainder terms. The apparent contradiction will be resolved in Sect. 7.3.3 when we shall consider infinite Taylor series and it will become clear that identity (2.72) is satisfied exactly.

3.8 Approximate Calculations of Functions

Taylor's formula is frequently used for approximate calculations of various functions. The idea is based on taking an n-term Taylor's formula (the Taylor polynomial (3.55)) and omitting the remainder term. If a better approximation for a function is needed, more terms in the Taylor's formula are considered, i.e. the function is replaced by a higher order Taylor polynomial.

As an illustration of this method, we shall consider a numerical calculation of the numbers π and e. To calculate e, we shall use Taylor's formula for the exponential function (3.65) at $x = 1$:

$$\begin{aligned} e &\simeq 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots + \frac{1}{n!} \\ &= \sum_{k=0}^{n} \frac{1}{k!} = 2 + \sum_{k=2}^{n} e_k \quad \text{with} \quad e_{k+1} = \frac{1}{(k+1)!} = \frac{e_k}{k+1}. \end{aligned} \tag{3.74}$$

The recurrent formula for the terms in the sum given above allows calculating every next term, e_{k+1}, from the previous one, e_k, and is much more convenient for the practical calculation on a computer than the direct calculation from their definition via the factorial. The dependence of the sum on the number of terms n is shown in Fig. 3.3(a). It is really surprising that only seven terms in the sum are already sufficient to get the number $e \simeq 2.718281828459$ with the precision corresponding to 12 digits.

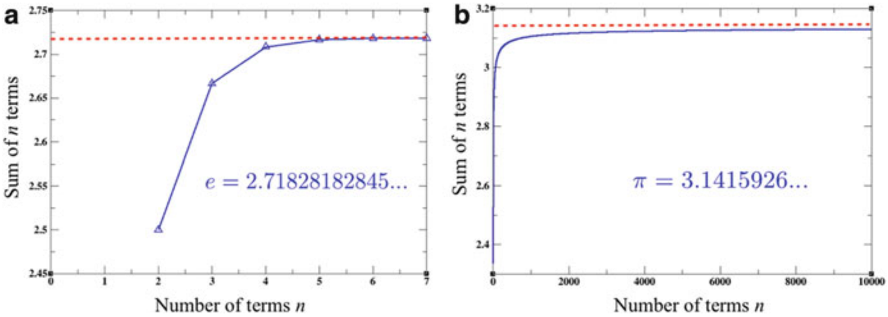


Fig. 3.3 Calculation of the numbers e (**a**) and π (**b**) using the corresponding Maclaurin finite sums (3.74) and (3.75), respectively. If the convergence of the sum for the number e is very fast, this is not so in the case of the arcsine Maclaurin series for the number π. In both cases the exact values of the numbers e and π are shown by the *horizontal red dashed lines*

To calculate the number π, we consider the Maclaurin formula (3.71) for the arcsine function at $x = 1$ since $\arcsin(1) = \pi/2$:

$$\pi = 2\arcsin(1) \simeq 2\sum_{k=0}^{n} \frac{(2k)!}{4^k\,(2k+1)\,(k!)^2} x^{2k+1}$$

$$= \sum_{k=0}^{n} a_k \quad \text{with} \quad a_{k+1} = a_k \frac{(2k+1)^2}{2\,(k+1)\,(2k+3)}. \tag{3.75}$$

Summing this series up for bigger and bigger values of n allows obtaining better and better approximations to the number π. However, as it follows from the numerical calculation shown in Fig. 3.3(b), this time the series converges much slower; even with the 10^5 terms kept in the sum the obtained approximation 3.1303099 is still with the considerable error of the order of 10^{-2}. Much better methods exist for the numerical calculation of the number π based on different principles which we cannot go into here.

Taylor's formula is frequently used for obtaining approximate representations for the functions in the vicinity of a point of interest. In other words, if one needs a simple analytical representation of a function $f(x)$ around point a, then the Taylor polynomial could be a good starting point. In Fig. 3.4 we show how this works for the exponential function $y = e^x$. Its Taylor polynomial around $x = 0$ is given by (3.65) if we drop the remainder term. It is clearly seen from the plotted graphs in (a) that a reasonable representation of this function can indeed be achieved using a small number of terms (a low order polynomial). However, the quality of the representation depends very much on the required precision (acceptable error). This is illustrated with the zoom-in image in (b) where the errors are better seen. The error gets bigger when the values of x which are further away from $x = 0$ are considered.

Generally, the higher the order of the polynomial, the better representation of the function is obtained. Also, it is a good idea to expand the function around the point of interest, not just around $x = 0$ as we have done above for the exponential

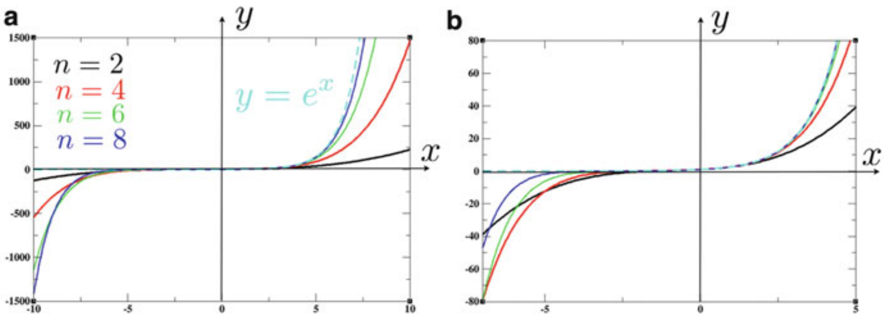


Fig. 3.4 Approximate representations of the exponential function $y = e^x$ using n-order Taylor's polynomials (3.65) around $x = 0$ (i.e. the Maclaurin formula with the remainder term dropped) for $n = 2, 4, 6, 8$ are shown together with the exact function (*cyan*). The central area of the graph in (**a**) is shown in (**b**) in more detail

function, as this would require a smaller number of terms to achieve the same precision. However, the wider the region of x for which the analytical representation is sought, the more terms will be required.

Problem 3.34. *Consider a mass m in the gravitational field of a very large object (e.g. the Earth) of mass M and radius R. The potential energy of the mass m at a distance d from the centre of the large object is $U(d) = -GmM/d$, where G is the gravitational constant. Show that, up to a constant, the potential energy of the mass m at the distance z above the big object surface (up to an unimportant constant) is given by $U(z) \simeq mgz$ if $z \ll R$, and derive the expression for the constant g.* [*Answer:* $g = GM/R^2$.]

3.9 Calculating Limits of Functions in Difficult Cases

In Sect. 2.4.6 we looked at some simple cases of uncertainties when taking a limit. Here we shall develop a general approach to all such cases. Taylor's formula is especially useful in calculating limits of functions in difficult cases of uncertainties $0/0$, ∞/∞ or $0 \cdot \infty$.

As an example, consider the limit (2.86) by expanding the sine function in the Maclaurin series:

$$\lim_{x\to 0} \frac{\sin x}{x} = \lim_{x\to 0} \frac{1}{x}\left[x - \frac{x^3}{3!} + \alpha(x)\right] = \lim_{x\to 0}\left[1 - \frac{x^2}{3!} + \frac{\alpha(x)}{x}\right] = 1,$$

where $\alpha(x)$ corresponds to all other terms in the expansion of the sine function which do not need here to be written explicitly; what is essential for us here is that function $\alpha(x)$ tends to zero at least as x^5 when $x \to 0$. We have obtained the same result as in Sect. 2.4.6.

Problem 3.35. *Use this method to prove the following limits:*

$$\lim_{x\to 0}\frac{\cos x-1}{x^2}=-\frac{1}{2};\quad \lim_{x\to 0}\frac{\sqrt{1+x}-1}{x}=\frac{1}{2}.$$

In fact, Taylor's theorem offers an extremely powerful general way of calculating limits with uncertainties (some special methods have already been considered in Sect. 2.4.6). Indeed, consider the limit of $f(x)/g(x)$ as $x\to x_0$ in the special case of both functions being equal to zero at the point x_0 (this is the case of $0/0$). Expanding both functions using Taylor's formula up to the $n=1$ term and using the fact that $f(x_0)=g(x_0)=0$, we have

$$\begin{aligned}\lim_{x\to x_0}\frac{f(x)}{g(x)} &= \lim_{x\to x_0}\frac{f(x_0)+f'(x_0)(x-x_0)+\lambda(x)(x-x_0)^2}{g(x_0)+g'(x_0)(x-x_0)+\mu(x)(x-x_0)^2}\\ &= \lim_{x\to x_0}\frac{f'(x_0)(x-x_0)+\lambda(x)(x-x_0)^2}{g'(x_0)(x-x_0)+\mu(x)(x-x_0)^2}\\ &= \lim_{x\to x_0}\frac{f'(x_0)+\lambda(x)(x-x_0)}{g'(x_0)+\mu(x)(x-x_0)}=\frac{f'(x_0)}{g'(x_0)},\end{aligned}\tag{3.76}$$

where the functions $\lambda(x)$ and $\mu(x)$ which arise in the remainder terms R_2 of Taylor's formula applied to the two functions $f(x)$ and $g(x)$, respectively, are assumed to be non-zero at x_0. The obtained result is known as L'Hôpital's rule.

This rule may need to be applied several times. This happens if the functions $\lambda(x)$ and $\mu(x)$ used above for $n=1$ are zeroes at x_0. In other words, Taylor's formula needs to be applied to $f(x)$ and $g(x)$ keeping up to the first non-zero term in the expansion.

Problem 3.36. *Let the functions $f(x)$ and $g(x)$ together with their first k derivatives be equal to zero at $x=x_0$. Prove that in this case*

$$\lim_{x\to x_0}\frac{f(x)}{g(x)}=\frac{f^{(k+1)}(x_0)}{g^{(k+1)}(x_0)},$$

which can be considered a generalisation of L'Hôpital's rule. In practice, however, it is more convenient to apply the rule (3.76) several times until the uncertainty is fully removed.

Example 3.8. ► Calculate the limit

$$L = \lim_{x \to 0} \frac{\sin^3 x}{x^2}$$

Solution. Apply L'Hôpital's rule

$$L = \lim_{x \to 0} \frac{\left(\sin^3 x\right)'}{\left(x^2\right)'} = \lim_{x \to 0} \frac{3 \sin^2 x \cos x}{2x} = \frac{3}{2} \lim_{x \to 0} \frac{\sin^2 x}{x}.$$

Now apply the rule again as we still have here $0/0$:

$$L = \frac{3}{2} \lim_{x \to 0} \frac{\left(\sin^2 x\right)'}{(x)'} = \frac{3}{2} \lim_{x \to 0} \frac{2 \sin x \cos x}{1} = \frac{3}{2} \cdot 2 \cdot \lim_{x \to 0} \sin x = 0. \blacktriangleleft$$

Above we considered the $0/0$ case in applying L'Hôpital's rule. However, the other two cases, ∞/∞ and $0 \cdot \infty$, can in fact be related to the considered $0/0$ uncertainty and hence similar rules be developed. For instance, in the ∞/∞ case, when $f(x_0) = g(x_0) = \infty$, we write:

$$\lim_{x \to x_0} \frac{f(x)}{g(x)} = \lim_{x \to x_0} \frac{1/g(x)}{1/f(x)}.$$

The functions $1/f(x)$ and $1/g(x)$ tend to zero values when $x \to x_0$, and hence L'Hôpital's rule can be applied directly to them:

$$\lim_{x \to x_0} \frac{f(x)}{g(x)} = \lim_{x \to x_0} \frac{[1/g(x)]'}{[1/f(x)]'} = \left. \frac{[1/g(x)]'}{[1/f(x)]'} \right|_{x=x_0}.$$

In the same way the $0 \cdot \infty$ uncertainty is considered: if $f(x_0) = 0$ and $g(x_0) = \infty$, then in this case

$$\lim_{x \to x_0} f(x)g(x) = \lim_{x \to x_0} \frac{f(x)}{1/g(x)} = \left. \frac{f'(x)}{[1/g(x)]'} \right|_{x=x_0}.$$

L'Hôpital's rule is also useful in finding the limits at $\pm\infty$ in any of the cases of uncertainties. Indeed, as an example let $f(x)$ and $g(x)$ be two functions which both tend to zero as $x \to +\infty$; we would like to calculate the limit of $f(x)/g(x)$. This case can be transformed into the familiar case of the limit at zero if we introduce a new variable $t = 1/x$. In that case the limit of $x \to +\infty$ becomes equivalent to $t \to 0$, and we can write:

$$\begin{aligned} \lim_{x \to +\infty} \frac{f(x)}{g(x)} &= \lim_{t \to 0} \frac{f(x(t))}{g(x(t))} = \lim_{t \to 0} \frac{F(t)}{G(t)} = \lim_{t \to 0} \frac{[F(t)]'}{[G(t)]'} \\ &= \lim_{t \to 0} \frac{f'(x)\, x'(t)}{g'(x)\, x'(t)} = \lim_{t \to 0} \frac{f'(x)}{g'(x)} = \lim_{x \to +\infty} \frac{f'(x)}{g'(x)}, \end{aligned} \tag{3.77}$$

where $F(t) = f(x(t))$ and $G(t) = g(x(t))$ are functions of t, and we used the chain rule when differentiating them with respect to t. As this simple derivation shows, the L'Hôpital's rule can be used even when working out the limits at infinities as well.

Problem 3.37. *Prove that formula (1.61) for a finite geometrical progression is formally valid for $q = 1$. That would mean that it is actually valid for any q.*

Problem 3.38. *Using the L'Hôpital's rule, prove the following limits:*

$$(a)\quad \lim_{x\to 0}\frac{\sqrt{1+x^2}-\sqrt{1-x}}{x}=\frac{1}{2}\,,\qquad (b)\quad \lim_{x\to 0}\frac{\ln(1+x)}{x}=1\,,$$

$$(c)\quad \lim_{x\to +\infty}\frac{\ln(1+x)}{x}=0,\qquad (d)\quad \lim_{x\to 0}\frac{e^{\alpha x}-1}{\sin x}=\alpha\,,$$

$$(e)\quad \lim_{x\to +\infty}\sqrt[x]{x}=1\,,\qquad (f)\quad \lim_{x\to 0}\frac{\cos x-1}{x^2}=-\frac{1}{2}\,,$$

$$(g)\quad \lim_{x\to\infty}x\left[e^{-\alpha/x}-1\right]=-\alpha,\qquad (h)\quad \lim_{x\to 0}\frac{xe^x-\sin x}{2x^2\cos x}=\frac{1}{2}\,,$$

$$(i)\quad \lim_{x\to\infty}x\left[1-x\ln\left(1+\frac{1}{x}\right)\right]=\frac{1}{2}\,,\qquad \lim_{x\to\infty}\left(x\arctan\frac{y}{x+a}\right)=y.$$

[*Hint*: in the case of the $x\to\infty$ limit it is convenient to introduce $t = 1/x$ so that the limit $t \to 0$ can be considered instead.]

3.10 Analysing Behaviour of Functions

The first and second derivatives are very useful in analysing graphs of functions $y = y(x)$: one can quickly see in which intervals the function increases or decreases, or where it has extrema (minima or maxima) and saddle points.

Increase/Decrease of Functions. It seems obvious that if a continuous function has a positive derivative in some interval $a \le x \le b$, then it increases there. The rigorous proof is provided in the following problem.

Problem 3.39. *Show using the Lagrange equation (3.64) that if $f'(x) > 0$ for all x within the interval $a \le x \le b$, then $f(x)$ monotonically increases there.* [*Hint*: prove that $f(x+\Delta x) > f(x)$ for any $\Delta x > 0$.] *Similarly show that if the first derivative is negative within some interval, the function $f(x)$ monotonically decreases there.*

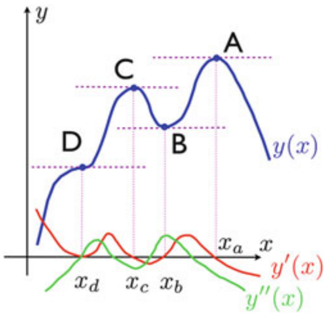


Fig. 3.5 Cartoon of a function (*blue*) $y = y(x)$ which has two maxima at points A and C, one minimum at point B and a saddle point at D. Its first and second derivatives, $y'(x)$ (*red*) and $y''(x)$ (*green*), are schematically drawn as well. Note that the tangent lines (*dashed purple*) at minima, maxima and saddle points are horizontal, i.e. have zero slope

These statements have a very simple geometrical interpretation: if the function increases, then the tangent line at any point x will make a positive angle with the x axis; if it decreases, then this angle will be negative.

Critical Points: Minima, Maxima and Saddle Points. At the beginning, let us consider an example of a function sketched in Fig. 3.5. It has two maxima (points A and C), one minimum (at B) and one saddle point (at D). At all these points the slope of the tangent lines is zero, i.e. the tangent lines are horizontal. The saddle point, which we shall discuss below in more detail, is specifically unusual: the tangent line there is horizontal, however, it is neither a minimum nor a maximum. How can we define and distinguish different types of such points?

Let us first analyse the behaviour of a function around point A (with $x = x_a$) where it has a maximum. Although we already proved in the previous Sect. 3.7 that the function has zero derivative at this point, it is worthwhile to revisit this proof again at a slightly different angle. When moving from some $x > x_a$ to the left towards x_a, our function increases and its slope is non-positive, i.e. $y'(x) = [y(x + dx) - y(x)]/dx \leq 0$ (we imply the limit here as we use dx instead of Δx). If we now move towards the point x_a from the left (i.e. from some $x < x_a$), then $y(x)$ increases again; however, in this case, the slope is non-negative: $y'(x) = [y(x + dx) - y(x)]/dx \geq 0$. In other words, the slope (the first derivative of $y(x)$) *changes sign* when crossing the point x_a of the maximum being exactly equal to zero at that point; at the point, x_a where the first derivative is equal to zero the tangent line (with the zero slope) is parallel to the x axis. It will be clear in a moment that this is a necessary condition for a maximum (but not yet sufficient).

Importantly, if you look at the graph of $y'(x)$ around the maximum point x_a (the red curve in Fig. 3.5), the first derivative decreases all the way starting somewhere between points x_a and x_b, passing through zero at x_a and then going down still beyond that point. In other words, the derivative of $y'(x)$, i.e. the second derivative $y''(x)$ of the original function $y(x)$, is negative in the vicinity of the maximum of a function, i.e. $y''(x_a) < 0$. We shall immediately see this as a very convenient condition of distinguishing between various critical points.

Note that the same result is valid for the other maximum at point x_c: the first derivative $y'(x_c) = 0$ there, while the second derivative $y''(x_c) < 0$ is strictly negative.

The same type of analysis can be performed around the minimum point x_b as well.

Problem 3.40. *Perform similarly a general analysis of the behaviour of a function $y(x)$ around a minimum point x_b and show that: (i) $y'(x_b) = 0$ and (ii) $y''(x_b) > 0$.*

We see that minima and maxima of a function (i.e. its *extrema*) can be easily found by simply solving the equation $y'(x) = 0$ with respect to x; then in many cases one can determine whether the obtained solutions correspond to the minimum or maximum by looking at the sign of the second derivative at that point: it is positive at the minimum point(s) and negative at the maximum point(s). This would all be good, but unfortunately for some functions this method does not work as the second derivative at the extremum point of a function is zero. The method must be slightly modified as is illustrated by the following example.

Consider the function $y = x^2$. It has at the point $x = 0$ (where $y'(x) = 2x = 0$) the second derivative $y''(x) = 2 > 0$, which then definitely corresponds to a minimum. Then, consider the function $y = x^4$ shown in Fig. 3.6(b) together with its first two derivatives. The equation $y'(x) = 4x^3 = 0$ gives $x = 0$ as the point of the extremum, however, the second derivative $y''(x) = 12x^2$ at this point is zero and hence it is not possible to say by just looking at $y''(0)$ if $x = 0$ corresponds to the minimum or maximum of the function. In this case one can look at the sign of the derivative on both sides of the point $x = 0$ or at the sign of the second derivative in the vicinity of this point. Indeed, consider two points very close to the point $x = 0$ of the extremum lying on both sides of it: $x = \pm\epsilon$ (with $\epsilon \to +0$). The first derivative

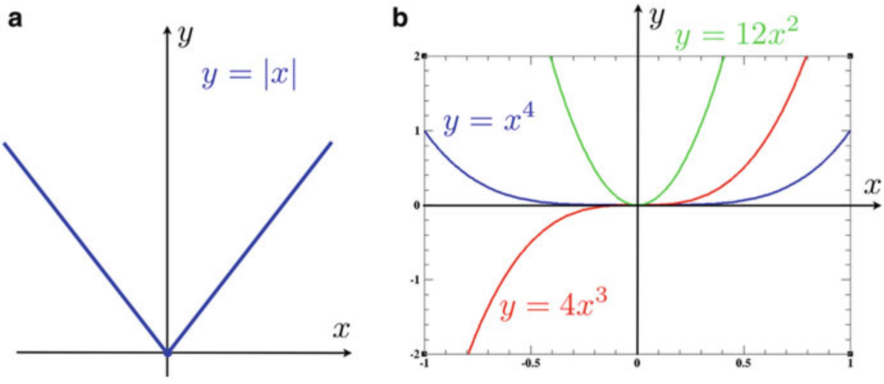


Fig. 3.6 Functions (**a**) $y = |x|$ and (**b**) $y = x^4$. In the latter case the first ($y' = 4x^3$) and second ($y'' = 12x^2$) derivatives are also shown

at these points, $y'(-\epsilon) = -4\epsilon^3 < 0$ and $y'(\epsilon) = 4\epsilon^3 > 0$, changes its sign from negative to positive when moving from the left ($x = -\epsilon$) to the right ($x = \epsilon$) points, which is characteristic of the minimum. At the same time, the second derivatives at these points $y''(\pm\epsilon) = 12\epsilon^2 > 0$ are both positive, and this again is characteristic of the minimum.

Care is needed in applying the condition $y'(x) = 0$ for finding extrema of a function. This is because, firstly, the derivative may not exist. This is, e.g., the case for $y = |x|$, which has a minimum at $x = 0$ where the first derivative is not defined, Fig. 3.6(a). Secondly, there could also be another type of critical points which also have a horizontal tangent line (of zero slope). These are called *saddle points* and one such point (point D) is contained on the graph of our schematic function in Fig. 3.5. As the slope is horizontal, the first derivative of the function at the saddle point $y'(x_d)$ is also equal to zero. However, the first derivative does *not* change its sign when passing this point; in addition, the second derivative at this point is zero as well: $y(x)$ increases from the left towards x_d, and then continues increasing on the right of it, i.e. the first derivative is always non-negative (and equal to zero at x_d). Hence, the second derivative is negative on the left of x_d and positive on the right of it, and passes through zero at the saddle point. An important conclusion from this consideration is that the roots of the first derivative equation $y'(x) = 0$ may give any type of the critical points (minima, maxima and the saddle), and additional consideration is necessary to characterise them explicitly.

Convex/Concave Behaviour of a Function. We know intuitively that around a minimum any function is convex, and around a maximum—concave. There exists a very simple criterion to determine if the function is concave or convex and within which interval (which may not include neither minimum nor maximum, the example being, e.g., the exponential function e^x which everywhere convex, but does not possess extrema, see below). Investigating the convexity of functions appears to be quite useful when analysing their general behaviour prior to plotting/sketching their graph (see below).

Note that if the function $f(x)$ is convex, its negative, $g(x) = -f(x)$, is concave. It is sufficient therefore only to consider the first (convex) case to work out the corresponding criterion.

We start by giving definitions. Let us look at Fig. 3.7(a). If we plot a line AB crossing the convex function at two points, $(a, f(a))$ and $(b, f(b))$, then all points of the function between a and b would lie not higher than the corresponding values on the crossing line: $f(x) \leq f_l(x)$ for any $a \leq x \leq b$. Let us write down the corresponding condition the function should satisfy to be convex (concave) according to this definition. The crossing line passes through two points A and B and hence has the following analytical form:

$$\begin{aligned} f_l(x) &= f(a) + \frac{f(b) - f(a)}{b - a}(x - a) \\ &= \left[1 - \frac{x - a}{b - a}\right] f(a) + \frac{x - a}{b - a} f(b) = (1 - \lambda) f(a) + \lambda f(b), \end{aligned} \tag{3.78}$$

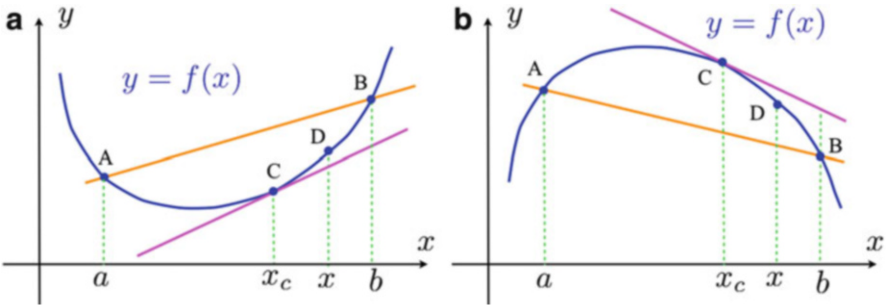


Fig. 3.7 To the two definitions of a convex (**a**) and concave (**b**) function. The tangent lines at point C are shown with *magenta*, while lines AB crossing the function $f(x)$ with *orange* in both cases

where we introduced a quantity $\lambda = (x-a)/(b-a)$ lying between zero and one: $0 < \lambda < 1$. Since $x = (1-\lambda)a + \lambda b$, the condition $f(x) \leq f_l(x)$ can equivalently be written as follows:

$$f((1-\lambda)a + \lambda b) \leq (1-\lambda)f(a) + \lambda f(b) , \quad \text{for any} \quad 0 < \lambda < 1. \tag{3.79}$$

In particular, for $\lambda = 1/2$:

$$f\left(\frac{a+b}{2}\right) \leq \frac{1}{2}[f(a) + f(b)] . \tag{3.80}$$

This condition shows that the convex function at the mean point between a and b is always not larger than the average of the function at both points.

The sign in these inequalities is to be reversed for the concave functions in the given interval $a \leq x \leq b$.

In fact, a simpler condition exists which can tell immediately if the function is either convex or concave. Let us look at the convex function sketched in Fig. 3.7(a) one more time: if we plot a tangent line to any point within the interval $a \leq x \leq b$, then it is clear that the tangent line cannot be above the graph of the function itself. In the figure we plotted a tangent line at point C with coordinates $(x_c, f(x_c))$, and it is readily seen to be everywhere below any point on the graph of the function $f(x)$ itself. Hence, for any point x within the convex interval, we can simply write $f(x) \geq f_t(x)$, where

$$f_t(x) = f(x_c) + f'(x_c)(x - x_c) \tag{3.81}$$

is the tangent line at the point x_c (it passes through the point C and has the slope equal to $f'(x_c)$). Similarly, the function is concave within some interval if for any point x we have $f(x) \leq f_t(x)$, and this situation is shown schematically in Fig. 3.7(b).

To work out the required criterion for the function to be convex, we shall use the $n = 1$ Taylor's formula (3.60):

$$
\begin{aligned}
f(x) &= f(x_c) + f'(x_c)(x - x_c) + R_2 \\
&= f_t(x) + R_2(x) \quad \text{with} \quad R_2(x) = \frac{1}{2} f''(\xi)(x - x_c)^2 ,
\end{aligned}
$$

where we have used (3.61) for the remainder term in this case and ξ is somewhere between x and x_c. Therefore, from the second definition of the convex curve given above, $f(x) - f_t(x) \geq 0$, we obtain:

$$
f(x) - f_t(x) = R_2(x) = \frac{1}{2} f''(\xi)(x - x_c)^2 \geq 0 \quad \Longrightarrow \quad f''(\xi) \geq 0, \tag{3.82}
$$

i.e. the second derivative should not be negative. Since we used arbitrary points x_c and x, it is clear that at any point within the convex interval the second derivative of the function is not negative. Correspondingly, the second derivative is not positive within the concave interval. These are the required criteria we have been looking for.

The exponential function $y = e^x$ is convex everywhere since $y'' = e^x > 0$. The logarithmic function $y = \ln x$ is everywhere[3] concave since $y'' = (\ln x)'' = (1/x)' = -1/x^2 < 0$.

The point at which the second derivative is zero could be an *inflection point* where concave (convex) behaviour goes over into being convex (concave). Recall that $y = x^4$ has the second derivative equal to zero at $x = 0$, but has the minimum there, not the inflection point. The sufficient condition for the point to be the inflection point is that the second derivative changes sign when passing it (which is not the case for $y = x^4$).

As an example, consider the function

$$
y = 4x^3 + 3x^2 - 6x + 1. \tag{3.83}
$$

Its first derivative

$$
y' = 12x^2 + 6x - 6 = 6(2x - 1)(x + 1) \tag{3.84}
$$

suggests that there are critical points at $x_1 = 1/2$ and $x_2 = -1$. To investigate the nature of these points, we calculate the second derivative: $y'' = 24x + 6$. At the point x_1 we have a minimum since $y''(1/2) = 18$ is positive, while at the point x_2 there is a maximum as $y''(-1) = -18$ is negative. The convex interval is obtained by solving the inequality $y'' = 24x + 6 > 0$ yielding $x > -1/4$, while the concave interval is given by $y'' = 24x + 6 < 0$ yielding $x < -1/4$. The point $x_3 = -1/4$ is therefore the inflection point. At $x \to -\infty$ the function tends to $-\infty$, while at

[3] To be more precise, where it is defined, i.e. for $x > 0$.

$x \to +\infty$ it tends to the $+\infty$. The function increases where $y' > 0$ which gives two possibilities, see Eq. (3.84): either

$$\begin{cases} 2x - 1 > 0 \\ x + 1 > 0 \end{cases} \implies \begin{cases} x > 1/2 \\ x > -1 \end{cases} \implies x > \frac{1}{2}$$

or

$$\begin{cases} 2x - 1 < 0 \\ x + 1 < 0 \end{cases} \implies \begin{cases} x < 1/2 \\ x < -1 \end{cases} \implies x < -1,$$

and the function decreases when $y' < 0$, which also gives two possibilities: either

$$\begin{cases} 2x - 1 > 0 \\ x + 1 < 0 \end{cases} \implies \begin{cases} x > 1/2 \\ x < -1 \end{cases} \implies \text{impossible},$$

or

$$\begin{cases} 2x - 1 < 0 \\ x + 1 > 0 \end{cases} \implies \begin{cases} x < 1/2 \\ x > -1 \end{cases} \implies -1 < x < \frac{1}{2}.$$

Consequently, now we should be able to sketch the function. Moving from $-\infty$ to the right towards $+\infty$, it first increases up to the point $x = -1$, where it experiences a maximum; after that point, it decreases up to the point $x = 1/2$ where it has a minimum; moving further, it increases again to $+\infty$. It is convex at $x > -1/4$ and concave for $x < -1/4$. The function is plotted in Fig. 3.8.

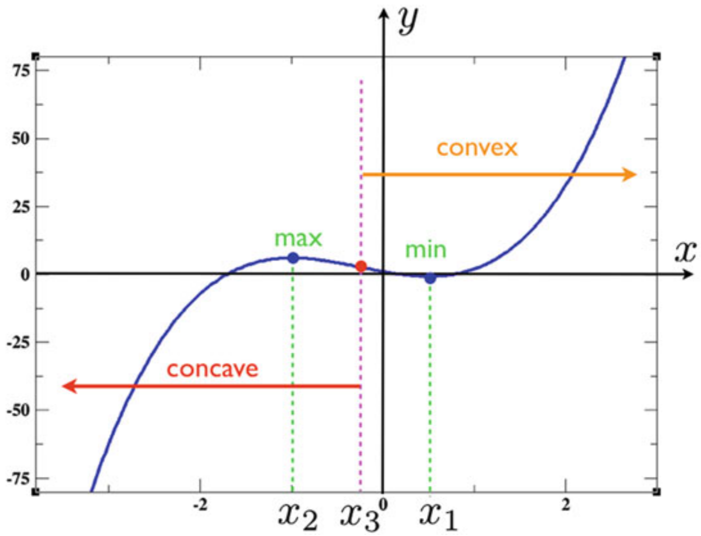


Fig. 3.8 Graph of the function (3.83). It has a maximum at x_2, a minimum at x_1 and an inflection point at x_3

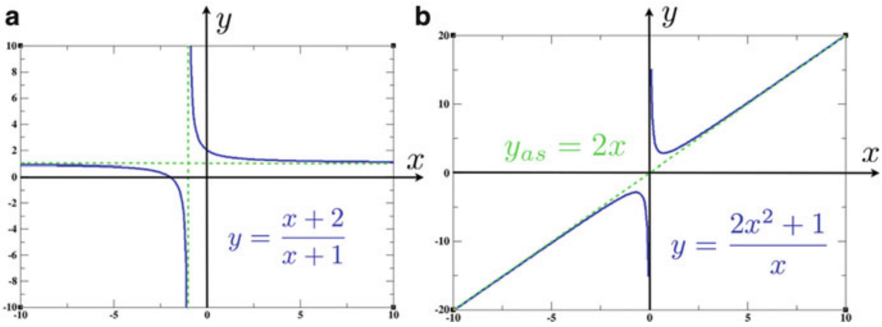


Fig. 3.9 Functions (**a**) $y = (x+2)/(x+1)$ and (**b**) $y = \left(2x^2+1\right)/x$ (*blue*) and their asymptotes (*green*)

Asymptotes of a Function. It order to analyse a function $y = y(x)$ at $x = \pm\infty$ and also at points x where $y(x)$ is not defined (or equal to $\pm\infty$), it is sometimes possible to study its *asymptotes* (or *asymptotic curves*); they do not always exist, but if they do it helps to sketch the function and better understand its behaviour. Before giving the corresponding definition, let us first consider two functions:

$$y_1(x) = \frac{x+2}{x+1} = 1 + \frac{1}{x+1} \quad \text{and} \quad y_2(x) = \frac{2x^2+1}{x} = 2x + \frac{1}{x}$$

to illustrate the point. Consider first $y_1(x)$; its graphs is shown in Fig. 3.9(a). This function is not defined at $x = -1$, and its graph approaches the (green) vertical line $x = -1$ as x tends to -1 from both sides, i.e. when $x \to -1 - 0$ (from the left) and $x \to -1 + 0$ (from the right). As x gets closer to the value of -1, the graph of the function gets closer as well to the green vertical line. This vertical line is called the *vertical asymptote* of the function $y_1(x)$. Let us now analyse the behaviour of this function at $x \to \pm\infty$ using its alternative representation as a sum of one and the function $1/(x+1)$ which will guide us in this discussion. We can see that on the right (for $x \to +\infty$) the function approaches the value of 1 being all the time higher than it (we say: it approaches it from above); similarly, on the left (when $x \to -\infty$) the same limiting value is being approached, this time from below, i.e. the function y_1 is always lower than this value. The green horizontal line $y = 1$ serves as a useful guide for this asymptotic behaviour at $\pm\infty$ and is called the *horizontal asymptote*.

Asymptotes do not need to be necessarily either vertical or horizontal, they may be of arbitrary direction as well. This is illustrated by our second function $y_2(x)$ shown in Fig. 3.9(b). This function is not defined at $x = 0$ and hence the vertical line $x = 0$ serves as its vertical asymptote. However, from the way the function $y_2(x)$

is written, it is clear that as $x \to \pm\infty$, the term $1/x \to 0$ and the behaviour of the function is entirely determined by the first part $2x$, which is a straight line shown in green in the figure. Hence, the line $y_{as} = 2x$ serves as a convenient guide to describe the behaviour of this function at $\pm\infty$ as $y_2(x)$ approaches this straight line either from above (when $x \to +\infty$) or below (when $x \to -\infty$). The line y_{as} is called an *oblique asymptote*. The horizontal asymptote can be considered as a particular case of the oblique one with the zero slope, while the vertical one—with the 90° slope.

Now we are ready to give the definitions of the asymptotes. The vertical asymptote is a vertical line $x = x_0$ with x_0 being a point at which the function $y(x)$ is not defined (tends to $\pm\infty$). The function $y(x)$ approaches the vertical asymptote as $x \to x_0$ from the left and right, i.e. the distance between the curve of the function and the asymptote tends to zero as $x \to x_0$ from either side. A horizontal or oblique asymptote $y_{as} = kx + b$ is a straight line which approaches the curve of the function $y(x)$ at either $x \to +\infty$ or $-\infty$ (the asymptotes at both limits could be different). We shall now derive a formula for the coefficients k and b of the asymptote (obviously, $k = 0$ for the horizontal asymptote) in the case of $x \to +\infty$. To this end, let us choose an arbitrary point x as shown in Fig. 3.10. We need to calculate the distance AB between the function $y(x)$ and the asymptote $y_{as}(x) = kx + b$ at this point. We have already solved this problem in Sect. 1.13.5 in Eq. (1.104), and the distance is given by

$$d = \frac{|kx - y(x) + b|}{\sqrt{1+k^2}}.$$

Since d should tend to zero as $x \to +\infty$, then $kx - y(x) + b$ must be equal to some function $\alpha(x)$ which tends to zero at this limit:

$$\lim_{x\to+\infty} \left[kx - y(x) + b - \alpha(x)\right] = 0$$
$$\Longrightarrow \quad k = \lim_{x\to+\infty} \frac{y(x) - b + \alpha(x)}{x} = \lim_{x\to+\infty} \frac{y(x)}{x}, \tag{3.85}$$

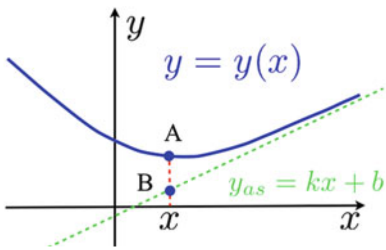


Fig. 3.10 To the derivation of the oblique asymptote for the $x \to +\infty$ case

which is the required formula for k. The constant b is obtained similarly:

$$\lim_{x\to+\infty}\left[kx - y(x) + b - \alpha(x)\right] = 0$$
$$\Longrightarrow \quad b = \lim_{x\to+\infty}\left[-kx + y(x) + \alpha(x)\right] = \lim_{x\to+\infty}\left[y(x) - kx\right]. \quad (3.86)$$

In exactly the same way the limit $x \to -\infty$ is considered and exactly the same formulae are obtained in which $x \to -\infty$ is assumed.

Example 3.9. ► Work out the asymptotes of the function

$$y = \frac{2x^2 - 3x + 6}{x + 1}. \quad (3.87)$$

Solution. First we note that $x = -1$ is the vertical asymptote of the function (note that the numerator $2x^2 - 3x + 6$ is not equal to zero at $x = -1$). To find the oblique or horizontal asymptotes, we calculate the limits given by Eqs. (3.85) and (3.86) at $+\infty$ and $-\infty$:

$$k_{+\infty} = \lim_{x\to+\infty}\frac{2x^2 - 3x + 6}{x\,(x+1)} = \lim_{x\to+\infty}\frac{2 - 3/x + 6/x^2}{1 + 1/x} = 2,$$

$$b_{+\infty} = \lim_{x\to+\infty}\left[\frac{2x^2 - 3x + 6}{x+1} - 2x\right] = \lim_{x\to+\infty}\frac{2x^2 - 3x + 6 - 2x\,(x+1)}{x+1}$$
$$= \lim_{x\to+\infty}\frac{-5x + 6}{x+1} = \lim_{x\to+\infty}\frac{-5 + 6/x}{1 + 1/x} = -5,$$

and the same limits are obviously obtained at $x \to -\infty$, i.e. one and the same asymptote $y_{as} = 2x - 5$ serves at both infinities. There are no horizontal asymptotes here as we did not find any asymptote with $k = 0$. The function together with its asymptotes is plotted in Fig. 3.11. ◄

Using critical points of the function (where extrema are) and its asymptotes, together with some additional information (e.g. crossing points with x and y axes) it is often possible to analyse the general behaviour of the function in some detail and sketch the function.

A usual procedure consists of the following steps:

- find the domain of values of x where the function $y = y(x)$ is defined;
- check if there is any symmetry, e.g. the function is odd or even, or is periodic;
- find limits of the function at the boundaries of the domain (e.g. at $x \to \pm\infty$);
- calculate $y'(x)$ and solve the equation $y'(x) = 0$ to determine all critical points;
- determine whether the function at each critical point has a minimum, maximum or it is a saddle point;
- determine the intervals where the function increases and decreases by solving inequalities $y' > 0$ and $y' < 0$, respectively;

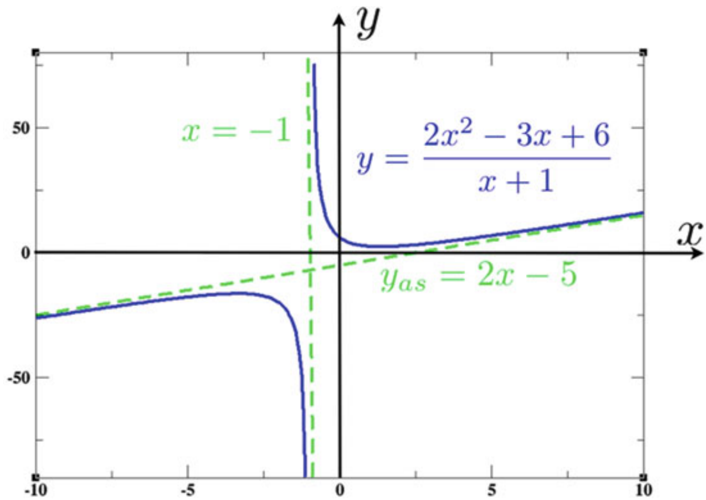


Fig. 3.11 Function $y(x)$ of Eq. (3.87) (*blue*) and its asymptotes (*green*)

- calculate $y''(x)$ and determine the inflection point; then, determine the intervals where the function is convex ($y'' \geq 0$) and concave ($y'' \leq 0$);
- calculate the value of the function at some selected points, especially at all critical points, the inflection points; determine the points x where the function crosses the x axis (roots of the function, $y(x) = 0$);
- make a sketch of the function.

Problem 3.41. *Analyse the function* $y(x) = 3(x-3)(x+1)(2x-1)$ *by finding all critical and inflection points, and then making a sketch. Check your findings by plotting the function using any available software. Does this function have asymptotes at* $x \to \pm\infty$*?* [*Answer*: maximum at $x = -1/3$, minimum at $x = 2$, inflection point $x = 5/6$, no asymptotes.]

Problem 3.42. *The same for* $y(x) = \left(x^2 + x + 1\right)/(x+1)$. *Does this function have an inflection point?* [*Answer*: maximum at $x = -2$, minimum at $x = 0$, no inflection point, asymptote $y = x$.]

Problem 3.43. *The same for* $y(x) = (x-2)^3/(x+2)^2$. [*Answer*: maximum at $x = -10$, inflection point at $x = 2$, asymptotes $x = -2$ and $y = x - 10$.]

Problem 3.44. *The same for* $y = e^{-x}\sin x$. [*Answer*: critical points at $x = \frac{\pi}{4} \pm \pi n$, where minima are for $n = 1, 3, \ldots$, while maxima for $n = 0, 2, 4, \ldots$, inflection points are at $x = \frac{\pi}{2} \pm \pi n$, where $n = 0, 1, 2, \ldots$; no asymptotes.]

Problem 3.45. *Consider the van der Waals equation of state of a gas*

$$\left(P + \frac{a}{v}\right)(v - b) = RT,$$

where P is pressure, T temperature, R gas constant, $v = V/N$ the volume per particle (N is the number of particles, V total volume), and a and b are constants: a indicates the strength of attraction between particles, while b corresponds to the volume a particle excludes from V (note that $v > b$). We shall consider the case of constant T to investigate the dependence $P(v)$, i.e. the isotherms of the gas.

(a) Show that the derivative

$$\frac{dP}{dv} = \frac{2a}{(v-b)^2} f(v) \ , \quad where \quad f(v) = \frac{(v-b)^2}{v^3} - \frac{RT}{2a}.$$

(b) Show that the derivative $f'(v)$ has a single maximum at $v = 3b$, where at the maximum $f(3b) < 0$ if $8a/27b < RT$ and $f(3b) > 0$ if $8a/27b > RT$.

(c) Then show that in either case $f(\pm\infty) = -RT/2a$ and is negative.

(d) Therefore, argue that in the first case, when $8a/27b < RT$, pressure P always decreases with the increase of the volume v.

(e) Show that in the second case when $8a/27b > RT$, two volumes v_1 and v_2 must exist at which $f(v) = 0$ (and hence $dP/dv = 0$), where $b < v_1 < 3b$ and $3b < v_2 < \infty$ (a cubic equation needs to be solved to find them, but the good thing is that our general analysis does not require knowing these two critical volumes explicitly).

(f) By analysing the sign of dP/dv in each of the intervals, $b < v < v_1$, $v_1 < v < v_2$ and $v_2 < v < \infty$, show that P decreases with v in the first interval, then when passing a minimum at v_1 starts to increase, and continues doing so until it reaches a maximum at v_2; after this point, it decreases again.

(g) The point at which two roots v_1 and v_2 coincide, i.e. when $8a/27b = RT$, is called the critical point of the van der Waals equation. Show that in this case the curve $P(v)$ has an inflection point at $v = 3b$ with the pressure $P = a/27b^2$.

(h) Sketch the graphs of the van der Waals equation using, e.g., $a = 1$ and trying various values of the other two parameters (T and b).

Problem 3.46. *A second order phase transition in a ferromagnetic material between paramagnetic and ferromagnetic phases can be described, within the so-called Landau–Ginzburg theory, using the following phenomenological expression for the free energy density of the material:*

$$f(m,T) = \frac{1}{2}\alpha\left(T - T_c\right)m^2 + f_4 m^4, \tag{3.88}$$

where m is magnetisation, T temperature, T_c phase transition temperature, f_4 and α positive constants.

(a) Show that for $T > T_c$ the free energy has a single stable state (i.e. a single minimum) with zero magnetisation only.

(b) Show that at $T < T_c$ the zero magnetisation state $m = 0$ becomes unstable (a maximum), while two states with non-zero magnetisation $m_\pm = \pm\sqrt{\alpha\left(T_c - T\right)/4f_4}$ become stable (f is minimum there). Note that the magnetisation changes continuosuly between the two solutions when the temperature passes the transition temperature T_c. This is characteristic for the second order phase transition.

(c) Show that in the latter case of $T < T_c$ there are two inflection points at $m = \pm\sqrt{\alpha\left(T_c - T\right)/12f_4} = \pm\left|m_\pm\right|/\sqrt{3}$.

Problem 3.47. *The same theory can be applied to the second order phase transition between paraelectric (zero polarisation, $P = 0$) and ferroelectric ($P \neq 0$) phases of a ferroelectric material. The free energy is identical to Eq. (3.88) of a ferromagnetic if one replaces m with P. Hence, as follows from the solution of the previous problem, the paraelectric phase exists at $T > T_c$, while ferroelectric solution becomes stable at $T < T_c$. Here we shall consider a more general case of the material in an external electric field E:*

$$f(P,T) = \frac{1}{2}\alpha\left(T - T_c\right)P^2 + f_4 P^4 - EP. \tag{3.89}$$

Show, that the susceptibility $\chi = \partial P/\partial E$ of the paraelectric phase ($T > T_c$) at $E = 0$ is $\chi = \left[\alpha\left(T - T_c\right)\right]^{-1}$, while in the ferroelectric phase ($T < T_c$) we have $\chi = \left[2\alpha\left(T_c - T\right)\right]^{-1}$, i.e. the susceptibility jumps at the transition temperature T_c.

Problem 3.48. *Consider a problem of crystal growth from the viewpoint of classical nucleation theory. In a supersaturated solution of particles (e.g., atoms) a nucleus of a crystalline phase may appear which will then may grow into the bulk phase. Whether or not a spherical nucleus of radius R will continue growing or will decompose into individual particles depends on the balance of its free energy* ΔG^* *which is composed of two terms:*

$$\Delta G^* = \frac{4}{3}\pi R^3 \rho_B \Delta\mu + 4\pi R^2 \gamma.$$

The first term is negative; it accounts for the decrease of the system free energy due to growth of a more energetically favourable solid phase with $\Delta\mu < 0$ *being the corresponding gain in the free energy per particle, and* ρ_B *the concentration of particles. This term favours formation of the crystalline (solid) phase. The second term is positive; it corresponds to the energy penalty due to formation of the surface of the nucleus with the surface energy* $\gamma > 0$ *per unit area. It works against the formation of the nucleus.*

(a) Show that $\Delta G^*(R)$ *has two extrema: at* $R = 0$ *and* $R^* = 2\gamma/\rho_B |\Delta\mu|$.
(b) Show that the first one corresponds to a minimum while the second to a maximum.
(c) Show that at $R > 3R^*/2$ *the free energy becomes negative.*
(d) Show that the energy barrier for nucleation

$$\Delta G^*\left(R^*\right) = \frac{16\pi}{3}\frac{\gamma^3}{\rho^2 |\Delta\mu|^2}.$$

(d) Sketch the $\Delta G^*(R)$.

For nuclei sizes smaller than R^* *the growth is unfavourable; however, as R becomes larger than this critical nucleus size the free energy starts decreasing and becomes negative after* $3R^*/2$ *which corresponds to a very high growth rate.*

Chapter 4
Integral

A notion of the integral is one of the main ideas of mathematical analysis (or simply "analysis"). In some literature integrals are introduced in two steps: initially, the so-called *indefinite integral* is defined as an inverse operation to differentiation; then *a definite integral* is introduced via some natural applications such as an area under a curve. In this chapter we shall use a much clearer and more rigorous route based on defining the definite integral first as a limit of an integral sum. Then the indefinite integral is introduced in a natural way simply as a convenience tool.

4.1 Definite Integral: Introduction

Consider initially a non-negative function $f(x) \geq 0$ specified in the interval $a \leq x \leq b$ as sketched in Fig. 4.1. What we would like to do is to find a way of calculating the area under the curve of the function between the boundary points a and b. If the function was a constant, $f(x) = C$, then the area is easily be calculated as $A = C(b-a)$. The area is also easily calculated in the case of a linear function,

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

passing through the points $(a, f(a))$ and $(b, f(b))$, in which case we have a trapezoid, and hence its area $A = [f(a) + f(b)](b-a)/2$.

How can one calculate the area under a general curve $f(x)$ which is not a straight line? We can try to calculate the area approximately using the following steps. First, we divide the interval $a \leq x \leq b$ into n subintervals by points $x_1, x_2, \ldots, x_{n-1}$ such that

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-2} < x_{n-1} < x_n = b,$$

L. Kantorovich, *Mathematics for Natural Scientists*, Undergraduate Lecture Notes in Physics, DOI 10.1007/978-1-4939-2785-2_4

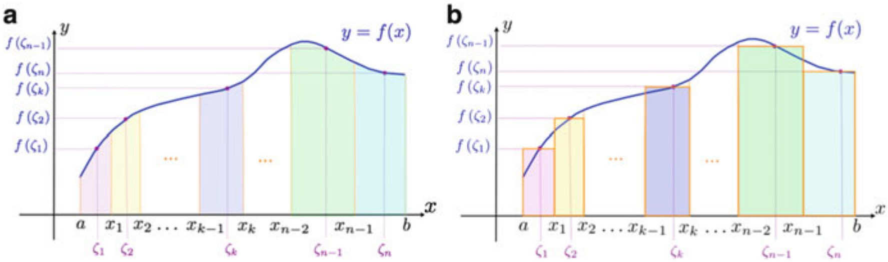


Fig. 4.1 Definition of a definite integral between a and b. The interval is divided up by n subintervals $\Delta x_k = x_k - x_{k-1}$ ($k = 1, 2, \ldots, n$), and the area of each stripe in (**a**) is replaced by the area of a rectangular of height $f(\zeta_k)$ and the width Δx_k, as shown in (**b**)

where for convenience we included the boundary points a and b into the set $\{x_k\}$ as x_0 and x_n, respectively. Next, we choose an *arbitrary* point ζ_1 within the first subinterval, $x_0 \leq \zeta_1 < x_1$, then another arbitrary point ζ_2 between x_1 and x_2, and so on. Altogether n points ζ_k are chosen such that $x_{k-1} \leq \zeta_k < x_k$ for all $k = 1, 2, \ldots, n$ as shown in Fig. 4.1(a). This way we divided up the area under the curve $y = f(x)$ into n strips (shown in different colours in the figure), and hence we may calculate the whole area by summing up areas of every such strip. The calculation of the area A_k of every strip $x_{k-1} \leq x < x_k$ can be done approximately by simply choosing the value of the function $f(\zeta_k)$ at the chosen point ζ_k within the strip as the strip height, see Fig. 4.1(b), i.e. by writing $A_k \simeq f(\zeta_k)(x_k - x_{k-1}) = f(\zeta_k)\Delta x_k$, where Δx_k is the width of the strip k. This way we replace the actual strip k having a (generally) curved line at the top with a rectangular of height $f(\zeta_k)$. The error we are making with this may be not so large if the strip is narrow enough, i.e. its width Δx_k is relatively small (compared to the width of the whole interval $\Delta = b - a$). Summing up all contributions, we obtain an approximate value of the whole area under the curve as a sum

$$A \simeq \sum_{k=1}^{n} f(\zeta_k)\Delta x_k. \tag{4.1}$$

This is called *the integral* or *Riemann's sum* (due to Riemann). It is clear that if we make more strips, they would become narrower, the function $f(x)$ within each such interval Δx_k would change less and the approximation we are making by replacing the function within each strip with a constant $f(\zeta_k)$, less severe. So, by making more strips, one can calculate the area with better and better precision. An important point is that the *exact* result can be obtained by making an infinite number of strips (by taking *all* widths of the strips Δx_k to zero at the same time which would require the number of strips $n \to \infty$), or ensuring that the widest of the intervals, i.e. $\max\{|\Delta x_k|\}$, goes to zero:

$$A = \lim_{\max\{|\Delta x_k|\} \to 0} \left(\sum_{k=1}^{n} f(\zeta_k)\Delta x_k \right). \tag{4.2}$$

This limit (if exists) is called a *definite (Riemann) integral* and is written as follows:

$$A = \int_a^b f(x)\mathrm{d}x, \tag{4.3}$$

where the sign $\int$ means literally "a sum" (it indeed looks like the letter "S"), and the numbers a and b are called the bottom and top limits of the integral, while the function under the integral sign the *integrand*. Note that $b > a$, i.e. the top limit is larger than the bottom one. If the limit of the integral sum exists, the function $f(x)$ is said to be *integrable*. Note that there are also more general Riemann–Stieltjes and Lebesgue integrals; we will not consider them here.

The definition for the definite integral (4.2) given above is immediately generalised for functions that are not necessarily positive. In this case the same formula (4.2) is implied, although, of course, the relation to the "area under the curve" is lost.

Moreover, the integral we considered above had a "direction" from a to b (where $b > a$), i.e. from left to right. It is also possible to consider the integral in the opposite direction, from right to left. In this case the points x_k are numbered from right to left instead and hence the differences $\Delta x_k = x_k - x_{k-1}$ become all negative. Still, the same integral sum as above can be written, which is denoted

$$A' = \int_b^a f(x)dx,$$

i.e. it looks the same apart from the fact that the top and the bottom limits changed places: now the bottom limit b is actually larger than the top limit a since $b > a$. Assume now that $f(x)$ is positive. Then, it is clear that the above integral consists of the same "areas under the curve" as in the previous case considered above (when we integrated from a to b), but calculated with negative $\triangle x_k = -|\Delta x_k|$. Therefore, by summing up all the contributions in the integral sum we should arrive at the same numerical value of the integral as when calculating it from a to b, only with the negative sign. It is clear now that A' must be equal to $-A$, i.e. we can write the first property of the definite integral (*direction*):

$$\int_a^b f(x)dx = -\int_b^a f(x)dx. \tag{4.4}$$

Consider now a point c somewhere between a and b, and let us choose one of the division points x_k exactly at c. Then, when we add more and more points to make the subintervals smaller and smaller, we can always keep the point c as a division point. It is clear that the integral sum will split into two, one from a and c, and another from c and b; at the same time, this must be still the integral from a to b. This is because the limit of the integral sum must exist no matter how the interval

$a < x < b$ is divided. Hence, we obtain the second property of the definite integral (*additivity*):

$$\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx. \tag{4.5}$$

Problem 4.1. *Prove that this identity still holds even if $c > b$ or $c < a$.*

Next, let the function $f(x)$ be a linear combination of two other functions, $u(x)$ and $v(x)$, i.e. $f(x) = \alpha u(x) + \beta v(x)$ with some real constants α and β. Then,

$$\lim_{\max\{|\Delta x_k|\}\to 0}\left[\sum_{k=1}^{n} f(\zeta_k)\Delta x_k\right] = \lim_{\max\{|\Delta x_k|\}\to 0}\left\{\sum_{k=1}^{n}\left[\alpha u(\zeta_k) + \beta v(\zeta_k)\right]\Delta x_k\right\}$$
$$= \lim_{\max\{|\Delta x_k|\}\to 0}\left[\sum_{k=1}^{n}\alpha u(\zeta_k)\Delta x_k\right] + \lim_{\max\{|\Delta x_k|\}\to 0}\left[\sum_{k=1}^{n}\beta v(\zeta_k)\Delta x_k\right]$$
$$= \alpha \lim_{\max\{|\Delta x_k|\}\to 0}\left[\sum_{k=1}^{n} u(\zeta_k)\Delta x_k\right] + \beta \lim_{\max\{|\Delta x_k|\}\to 0}\left[\sum_{k=1}^{n} v(\zeta_k)\Delta x_k\right],$$

i.e., using notations for integrals, we have:

$$\int_a^b \left[\alpha u(x) + \beta v(x)\right] dx = \alpha \int_a^b u(x)dx + \beta \int_a^b v(x)dx. \tag{4.6}$$

Of course, the proven *linearity* property holds for a linear combination of any number of functions. In particular, it follows from this property that a constant pre-factor in the integrand can always be taken outside the integral sign.

Finally, the integral between two identical points is obviously equal to zero:

$$\int_a^a f(x)dx = 0.$$

Problem 4.2. *Consider two functions: one is even, $f(-x) = f(x)$, and one odd, $g(-x) = -g(x)$. Then demonstrate, using the definition of the integral, that the following identities are valid when these functions are integrated over a symmetric interval $-a \le x \le a$:*

$$\int_{-a}^a f(x)dx = 2\int_0^a f(x)dx\ , \ \textit{if} \quad f(-x) = f(x), \tag{4.7}$$

$$\int_{-a}^a g(x)dx = 0\ , \ \textit{if} \quad g(-x) = -g(x). \tag{4.8}$$

It is sometimes useful to remember that any function $f(x)$ can always be written as a sum of even and odd functions:

$$f(x) = f_{\text{even}}(x) + f_{\text{odd}}(x) \ , \ f_{\text{even}}(x) = \frac{1}{2}\left[f(x) + f(-x)\right] \ , \ f_{\text{odd}}(x) = \frac{1}{2}\left[f(x) - f(-x)\right]. \tag{4.9}$$

The integral defined above depends on the function $f(x)$ and the interval boundaries a and b. We did mention that the integral (if exists) should not depend on the way division of the interval is made. It might also be important how the points ζ_k (at which the function is calculated) are chosen within each subinterval. Therefore, it is clear that the limit written above is not a usual limit we considered in the previous chapters, and care is needed in taking it. However, before we start looking into these subtle complications, it is instructive to consider some examples first.

Example 4.1. ► Evaluate the definite integrals

$$J = \int_0^1 dx = 1 \quad \text{and} \quad I = \int_0^1 x dx,$$

using subintervals of the same length and points ζ_k chosen at the end of each subinterval.

Solution. In the first integral the function $f(x) = 1$ is constant, so for any division of the interval $0 \le x \le 1$ into subintervals we have

$$J = \lim_{\max\{|\Delta x_k|\}\to 0}\left[\sum_{k=1}^{n} \Delta x_k\right] = \lim_{\max\{|\Delta x_k|\}\to 0} \Delta x = \Delta x = 1,$$

where $\Delta x = 1$ is the whole integration interval. (Note that the calculation does not depend on the choice of the points ζ_k within the subintervals because the integrand is a constant.)

In the second case $f(x) = x$. We divide the interval $0 \le x \le 1$ into n subintervals of the same length $\Delta = 1/n$ with points $x_k = k\Delta$ $(k = 1, 2, \ldots, n)$, and within each one of them we choose the point $\zeta_k = x_k$ right at the end of it. Then, consider the integral sum:

$$I_n = \sum_{k=1}^{n} f(\zeta_k)\Delta = \sum_{k=1}^{n} x_k \Delta = \sum_{k=1}^{n} k\Delta^2 = \Delta^2 \sum_{k=1}^{n} k = \Delta^2 \frac{1+n}{2} n,$$

where in the last passage we have used the formula for a sum of an arithmetic progression (Sect. 1.9, Eq. (1.65)). Since $\Delta = 1/n$ by construction, we get

$$I_n = \frac{1+n}{2n} = \frac{1}{2n} + \frac{1}{2}.$$

The limit of the widths of all subintervals going to zero corresponds to the limit of $n \to \infty$. Taking this limit in the calculated integral sum, I_n, we finally obtain $I = 1/2$. ◄

Problem 4.3. *Perform the calculation of the same integral I of the previous example by still using subintervals of the same length* $\Delta = 1/n$*; however, this time choose the points* ζ_k *either at the beginning,* $\zeta_k = (k-1)\,\Delta$*, or in the middle,* $\zeta_k = (k+1/2)\,\Delta$*, of each subinterval. Make sure that you get the same result after taking the* $n \to \infty$ *limit, in spite of the fact that the integral sums for the given value of n are different in each case.*

Problem 4.4. *Generalise the above results for the intervals* $a \le x \le b$*:*

$$\int_a^b dx = b - a \quad \textit{and} \quad \int_a^b x dx = \frac{1}{2}\left(b^2 - a^2\right). \tag{4.10}$$

Problem 4.5. *Using the same method (subintervals of the same length and a particular, but simple choice of the points* ζ_k *within them), show that the integral*

$$\int_0^1 x^2 dx = \frac{1}{3}.$$

[*Hint*: you will need expression (1.66) for the sum of the squares of integer numbers.] *Try several different choices of the points* ζ_k *within the subintervals, and make sure that the result is always the same in the* $\Delta \to 0$ *limit!*

Problem 4.6. *Using the subintervals of the same length, prove the following identity:*

$$\int_{-1}^1 e^{\alpha x} dx = \frac{1}{\alpha}\left(e^{\alpha} - e^{-\alpha}\right).$$

[*Hint*: you will have to sum a geometrical progression here, see (1.61), and calculate a limit at $n \to \infty$, which contains $0/0$ uncertainty.]

Problem 4.7. *Using the subintervals of the same length, prove the following identity:*

$$\int_0^{\pi} \sin x dx = 2.$$

[*Hint*: use Eq. (2.63).]

4.2 Main Theorems

Of course, the above direct method of calculating integrals seems to be extremely difficult. In many cases, analytical calculation of the required finite integral sums may be not so straightforward or even impossible to do. Fortunately, there is a much easier way whose essence is in relating the integral to the derivative which manifests itself in the main theorem of calculus to be discussed later on. However, in order to discuss it, we first have to understand in more detail the meaning of the definite integral itself.

We start from a more rigorous definition. A subtle point about existence of the limit of the integral sum (and hence of the integral) is that the same result must be obtained no matter how the division of each interval is made (e.g. equal subintervals for each particular n or of different widths) and how the points ζ_k are chosen within the subintervals. By adopting a particular way of dividing the original interval $a \leq x \leq b$ into n subintervals and choosing the points ζ_k within each of them we shall obtain a particular value of the integral sum

$$A(n) = \lim_{\Delta(n) \to 0} \sum_{k=1}^{n} f(\zeta_k) \Delta_k,$$

where $\Delta(n) = \max\{\Delta_k\}$ is the maximum width of all subintervals for the given n (assuming for definiteness that any $\Delta_k > 0$). The obtained numerical sequence, $A(1), A(2), \ldots, A(n), \ldots$, of the values of the integral sums may tend to some limit A as $n \to \infty$.

The first definition of the integral then would sound similar to the limit of the numerical sequences in Sect. 2.2: the definite integral of a function $f(x)$ on the interval $a \leq x \leq b$ exists (or the function $f(x)$ is *integrable* there) if any numerical sequence $\{A(n)\}$ of the integral sums obtained by choosing various ways of dividing the original interval into n subintervals, such that $\lim_{n\to\infty} \Delta(n) = 0$, and choosing points ζ_k inside the subintervals in different ways, tends to the same limit A. This definition allows us to adopt immediately all main theorems related to the limit of numerical sequences of Sect. 2.2. For instance, the limit of a sum is equal to the sum of the limits, and hence the integral of a sum of two functions is equal to the sum of individual integrals calculated for each of them.

It is possible to give an appropriate definition of the integral using the "$\epsilon - \delta$" language as well: the integral on the interval $a \leq x \leq b$ exists and equal A (or the function $f(x)$ is integrable there with the value of the integral A) if for any $\epsilon > 0$ there exists such $\delta > 0$ so that for any division into subintervals satisfying $\Delta(n) < \delta$ (i.e. any of them is shorter than δ) we have

$$\left| \sum_{k=1}^{n} f(\zeta_k) \Delta_k - A \right| < \epsilon \tag{4.11}$$

for arbitrary selection of points ζ_k within the subintervals. This definition is basically saying that if we choose any precision ϵ with which we would like to estimate the value of the integral via the appropriate integral sum, then it is always possible to select an appropriate division into subintervals which are small enough to obtain the integral sum with the given precision, i.e. satisfying the condition

$$A - \epsilon < \sum_{k=1}^{n} f(\zeta_k)\,\Delta_k < A + \epsilon. \tag{4.12}$$

Both definitions of the limit of the integral sum are equivalent and clearly state that the integral depends only on the function $f(x)$ and the numbers a and b.

Once the rigorous definitions are given, we have to understand what conditions the functions must satisfy to be integrable.

Theorem 4.1. *To be integrable, the function $f(x)$ must be limited on the interval $a \leq x \leq b$, i.e. two real numbers m and $M > m$ must exist such that $m \leq f(x) \leq M$. This condition is the* necessary *one.*

Proof (by contradiction):. assume that the function $f(x)$ is unlimited from above (i.e. it can be as *large* as desired) within the given interval. Let us assume that it is infinite at least at one point c within the interval. Then, when we divide the interval into subinterval. One such subinterval (lets its number is k) would contain that point c and hence we can always choose the point ζ_k within that subinterval very close to c, with the value of the function $f(\zeta_k)$ there as large as we like. It is clear then that this way we can make the contribution $f(\zeta_k)\,\Delta_k$ of that subinterval into the integral sum as large as necessary to disobey the inequality (4.11). **Q.E.D.**

Next, we introduce new definitions due to French mathematician Darboux, called Darboux sums. Consider a specific division of the interval $a \leq x \leq b$ into n subintervals, and let m_k and M_k be the smallest and largest values of the function $f(x)$ within the k-th subinterval $x_{k-1} \leq x < x_k$, i.e. $m_k \leq f(x) \leq M_k$. The corresponding integrals sums,

$$A^{\text{low}}(n) = \sum_{k=1}^{n} m_k \Delta_k, \tag{4.13}$$

$$A^{\text{up}}(n) = \sum_{k=1}^{n} M_k \Delta_k, \tag{4.14}$$

are called lower and upper Darboux sums, respectively. These sums correspond to a specific choice of the points ζ_k within each subinterval as illustrated in Fig. 4.2: for the lower sum each point ζ_k is chosen where the function $f(x)$ takes on the lowest

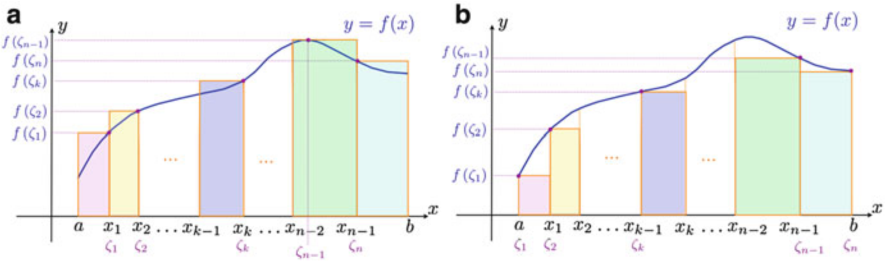


Fig. 4.2 To the definition of the upper (**a**) and lower (**b**) Darboux sums

value within each subinterval, and similarly for the upper Darboux sum. Since any other choice would give the values of the function not lower than m_k and not larger than M_k, it is clear that for an arbitrary choice of the points ζ_k and for the given division of the original interval the corresponding integral sum $A(n)$ will be not lower than $A^{\text{low}}(n)$ and not larger than $A^{\text{up}}(n)$:

$$A^{\text{low}}(n) \leq A(n) \leq A^{\text{up}}(n). \tag{4.15}$$

Note that the Darboux sums only depend on the function $f(x)$ and on the division points x_k. Now let us add one more division point x'_k between the points x_{k-1} and x_k. This way we should have $n+1$ divisions. The values of m_i and M_i for all i except for $i = k$ will be the same, however, the k-th subinterval now consists of two, $x_{k-1} \leq x < x'_k$ and $x'_k \leq x < x_k$, with the lowest values of the function m'_k and m''_k and its largest values M'_k and M''_k, respectively. It is clear, however, that at least one of the lowest values will be larger than the lowest value m_k for the combined subinterval $x_{k-1} \leq x < x_k$ and one of the largest values will be smaller than M_k. For instance, if say m_k is taken by the function $f(x)$ at some point $x = \zeta'$ between x_{k-1} and x'_k, then in the other subinterval, $x'_k \leq x < x_k$, the function will be greater than $m_k = f(\zeta')$. Therefore, the corresponding lower Darboux sum becomes larger, while the upper sum becomes smaller. This should remain true if we add an arbitrary number of new divisions. We have just proved this

Theorem 4.2. *If we add more division points to the given ones, the lower Darboux sum can only increase, while the upper sum only decrease.*

We also need another theorem.

Theorem 4.3. *Any lower Darboux sum (for arbitrary division of the interval) can never be larger than any upper sum corresponding to the same or a different division.*

Proof. Consider a particular division with the lower and upper Darboux sums being A_1^{low} and A_1^{up} (we omitted the number of divisions n here for convenience). Consider next a completely different division in which some or all division points are different from those in the first division. For the second division the Darboux sums are, respectively, A_2^{low} and A_2^{up}. Let us construct the third division by adding all new points from the second division to the first. The Darboux sums become A_3^{low} and A_3^{up}. From Theorem 4.2 it follows that $A_3^{\text{low}} \geq A_1^{\text{low}}$ and $A_3^{\text{up}} \leq A_1^{\text{up}}$. However, exactly the same combined division can be constructed by adding all new division points of the first division to the second one. Again using Theorem 4.2, we can also write: $A_3^{\text{low}} \geq A_2^{\text{low}}$ and $A_3^{\text{up}} \leq A_2^{\text{up}}$. On the other hand, the low Darboux sum is always not larger than the upper one for the same division. Therefore, we have:

$$A_1^{\text{low}} \leq A_3^{\text{low}} \quad \text{and} \quad A_3^{\text{low}} \leq A_3^{\text{up}} \quad \text{and} \quad A_3^{\text{up}} \leq A_2^{\text{up}} \quad \Longrightarrow \quad A_1^{\text{low}} \leq A_2^{\text{up}}. \tag{4.16}$$

This shows that the lower sum is always not larger than the upper one, even if the latter corresponds to a completely different set of divisions (of course, for the same divisions, by definition, $A^{\text{low}} \leq A^{\text{up}}$). Similarly,

$$A_1^{\text{up}} \geq A_3^{\text{up}} \quad \text{and} \quad A_3^{\text{up}} \geq A_3^{\text{low}} \quad \text{and} \quad A_3^{\text{low}} \geq A_2^{\text{low}} \quad \Longrightarrow \quad A_1^{\text{up}} \geq A_2^{\text{low}}, \tag{4.17}$$

i.e. the upper sum is always not smaller than any lower sum from the same or any other division. **Q.E.D.**

The proven theorem shows that whatever divisions we make, the lower Darboux sums are always smaller than (or equal to) any of the upper ones, and, vise versa, the upper ones are always larger than (or equal to) any of the lower ones. This means that all possible (i.e. obtained using various divisions) lower sums are limited from above, and all possible upper sums are limited from below. Hence we denote by A_* the upper value for all the lower sums and by A^* the lowest value for all the upper sums. These are called lower and upper Darboux integrals, respectively. Obviously, $A_* \leq A^*$, and for any lower and upper sums:

$$A^{\text{low}} \leq A_* \leq A^* \leq A^{\text{up}}. \tag{4.18}$$

Note that A^{low} and A^{up} may correspond to two completely different divisions of the interval.

Theorem 4.4 (Existence of the Definite Integral). *For the definite integral to exist, it is necessary and sufficient that*

$$\lim_{\Delta \to 0} \left(A^{low} - A^{up}\right) = 0, \tag{4.19}$$

i.e. for any positive ϵ there exists such positive δ that for any divisions satisfying $\Delta = \max\{\Delta_k\} < \delta$ we have $0 < A^{up} - A^{low} < \epsilon$.

Proof. We have to prove both sides of the theorem, i.e. that this condition is both necessary and sufficient. To prove the necessary condition, we should assume that the integral exists and hence the condition (4.19) follows as a consequence. If the integral exists and equals A than for *any* division satisfying $\Delta < \delta$ follows

$$A - \frac{\epsilon}{4} < \alpha < A + \frac{\epsilon}{4} \quad \Longrightarrow \quad \alpha - A < \frac{\epsilon}{4} \quad \text{and} \quad A - \alpha < \frac{\epsilon}{4}, \tag{4.20}$$

where α is the corresponding integral sum for the chosen division and we introduced $\epsilon/4$ instead of ϵ for convenience. Note that this inequality holds for any choice of internal points ζ_k inside the divisions (subintervals). On the other hand, $\alpha > A^{\text{low}}$ and any other choice of the internal points ζ_k would result in the integral sum exceeding A^{low}. Still, for the given ϵ, we can select the internal points in such a way that the corresponding integral sum α' would satisfy

$$\alpha' < A^{\text{low}} + \frac{\epsilon}{4} \quad \Longrightarrow \quad \alpha' - A^{\text{low}} < \frac{\epsilon}{4}. \tag{4.21}$$

Similarly, with any choice of the internal points the integral sum would not exceed A^{up}, however, one can always make a selection of the internal points such that the corresponding integral sum α'' would satisfy

$$\alpha'' > A^{\text{up}} - \frac{\epsilon}{4} \quad \Longrightarrow \quad A^{\text{up}} - \alpha'' < \frac{\epsilon}{4}. \tag{4.22}$$

We mentioned that inequalities (4.20) are valid for any choice of the internal points. In particular, they are valid for the two internal point selections we have made above, hence one can also write for them:

$$\alpha'' - A < \frac{\epsilon}{4} \quad \text{and} \quad A - \alpha' < \frac{\epsilon}{4}. \tag{4.23}$$

Therefore, we have:

$$A^{\text{up}} - A^{\text{low}} = \left(A^{\text{up}} - \alpha''\right) + \left(\alpha'' - A\right) + \left(A - \alpha'\right) + \left(\alpha' - A^{\text{low}}\right) < \frac{\epsilon}{4} + \frac{\epsilon}{4} + \frac{\epsilon}{4} + \frac{\epsilon}{4} < \epsilon,$$

which means that (4.19) is indeed satisfied.

Now we prove the sufficiency condition, which means that by assuming that (4.19) is satisfied we have to show that the integral exists. For any division the lower and upper sums satisfy $A^{\text{low}} \leq A_*$ and $A^* \leq A^{\text{up}}$, so that

$$A^{\text{low}} + A^* \leq A_* + A^{\text{up}} \quad \Longrightarrow \quad A^* - A_* \leq A^{\text{up}} - A^{\text{low}}.$$

Also, $A^* \geq A_*$, from which it follows that

$$A^* - A_* \geq 0 \quad \text{and} \quad A^* - A_* \leq A^{\text{up}} - A^{\text{low}} \quad \Longrightarrow \quad 0 \leq A^* - A_* \leq A^{\text{up}} - A^{\text{low}}.$$

On the other hand, since the condition (4.19) is satisfied, we have $A^{\text{up}} - A^{\text{low}} < \epsilon$, so that $0 \le A^* - A_* < \epsilon$ for any ϵ, meaning that $A^* = A_*$ which we shall denote simply as A. Then, we can also write

$$A^{\text{low}} \le A_* = A = A^* \le A^{\text{up}}. \tag{4.24}$$

On the other hand, for any choice of the internal points ζ_k we have the integral sum α satisfying $A^{\text{low}} \le \alpha \le A^{\text{up}}$, which, when combined with inequality (4.24), gives:

$$\alpha \le A^{\text{up}} \quad \text{and} \quad A \ge A^{\text{low}} \quad \Longrightarrow \quad \alpha - A \le A^{\text{up}} - A^{\text{low}}$$

and

$$\alpha \ge A^{\text{low}} \quad \text{and} \quad A \le A^{\text{up}} \quad \Longrightarrow \quad \alpha - A \ge A^{\text{low}} - A^{\text{up}} = -\left(A^{\text{up}} - A^{\text{low}}\right),$$

which simply means that $|\alpha - A| \le A^{\text{up}} - A^{\text{low}}$. However, $A^{\text{up}} - A^{\text{low}} < \epsilon$ for any division obeying $\Delta < \delta$ due to condition (4.19), which yields $|\alpha - A| < \epsilon$ for any such division. But, by virtue of condition (4.11), this is exactly what we mean by saying that the integral exists and its value is A. **Q.E.D.**

At this point we are ready to formulate a sufficient condition for a function $f(x)$ to be integratable.

Theorem 4.5. *Functions which are continuous within an interval are integrable there, i.e. the definite integral (4.3) exists if $f(x)$ is continuous between a and b.*

Proof. Consider the difference of the upper and lower sums for a given division:

$$A^{\text{up}} - A^{\text{low}} = \sum_{k=1}^{n} M_k \Delta x_k - \sum_{k=1}^{n} m_k \Delta x_k = \sum_{k=1}^{n} \Delta f_k \Delta x_k, \tag{4.25}$$

where m_k and M_k are the minimum and maximum values of the function $f(x)$ within the subinterval $\Delta x_k = x_k - x_{k-1}$, and $\Delta f_k = M_k - m_k$ characterises the change of the function within the same subinterval. Because the function $f(x)$ is continuous, for any $\epsilon' > 0$ one can find $\delta > 0$ such that $|x' - x''| < \delta$ implies $|f(x') - f(x'')| < \epsilon'$. Let x' be a point where $f(x)$ takes on its minimum value m_k within the interval Δx_k and x'' its maximum value M_k. Then, we can say that

$$\left|f\left(x'\right) - f\left(x''\right)\right| = |m_k - M_k| = M_k - m_k = \Delta f_k < \epsilon' \quad \text{as long as} \quad \Delta x_k < \delta.$$

Choosing $\epsilon' = \epsilon / (b - a)$, we can rewrite (4.25) as follows:

$$A^{\text{up}} - A^{\text{low}} = \sum_{k=1}^{n} \Delta f_k \Delta x_k < \sum_{k=1}^{n} \frac{\epsilon}{b-a} \Delta x_k = \frac{\epsilon}{b-a} \sum_{k=1}^{n} \Delta x_k = \frac{\epsilon}{b-a} (b-a) = \epsilon,$$

which corresponds to the sufficient condition for the integral to exist (from the previous Theorem 4.4) to be obeyed. Therefore, continuous functions do fully satisfy that theorem. **Q.E.D.**

In fact it appears that piecewise continuous functions are also integrable. This is because the points of discontinuity can always be chosen as some of the division points. Moreover, the value of the integral does not change if the value of the function $f(x)$ is changed in any number of *isolated points* within the integration region (see Sect. 4.5.3).

Theorem 4.6 (Average Value Theorem). *If the function $f(x)$ is continuous between a and b, then*

$$\int_a^b f(x)dx = f(\zeta)\,(b-a)\,, \tag{4.26}$$

where ζ is some point within the interval.

Proof. Indeed, since the function is continuous, it should reach its minimum, m, and maximum, M, values somewhere within that interval; these could also be the boundaries of the interval. Let us now estimate the integral sum from below,

$$\sum_{k=1}^{n} f(\zeta_k)\,\Delta_k \geq \sum_{k=1}^{n} m\Delta_k = m \sum_{k=1}^{n} \Delta_k = m\,(b-a) = m\mu,$$

and above,

$$\sum_{k=1}^{n} f(\zeta_k)\,\Delta_k \leq \sum_{k=1}^{n} M\Delta_k = M \sum_{k=1}^{n} \Delta_k = M\,(b-a) = M\mu,$$

where $\mu = b - a$ is the length of the interval. Taking the limit $\Delta \to 0$, we then obtain

$$m\mu \leq \int_a^b f(x)dx \leq M\mu \quad \Longrightarrow \quad m \leq \frac{\int_a^b f(x)dx}{\mu} \leq M, \tag{4.27}$$

i.e. the value of the integral divided by the lengths of the interval μ, let us call this ratio λ, is bracketed by the minimum and maximum values of the function on the interval: $m \leq \lambda \leq M$. According to the second Bolzano–Cauchy theorem (Theorem 2.18), there should exist a point ζ within the interval at which the function $f(x)$ would take on exactly that value λ, i.e. $\lambda = f(\zeta)$, since continuous functions take on all values continuously within the interval. Recalling what the number λ is, we recover the desired result (4.26). **Q.E.D.**

This theorem is called the average value theorem because it shows that $f(\zeta)$ is equal to the integral between a and b divided by the length of the interval:

$$f(\zeta) = \frac{\int_a^b f(x)dx}{b-a}. \tag{4.28}$$

Theorem 4.7 (Absolute Value Estimate). *It is usually necessary to be able to put an upper and lower boundaries to an integral. For this purpose the following inequality is useful:*

$$\left|\int_a^b f(x)dx\right| \leq \int_a^b |f(x)|\, dx \;\; or \;\; -\int_a^b |f(x)|\, dx \leq \int_a^b f(x)dx \leq \int_a^b |f(x)|\, dx. \tag{4.29}$$

Proof. This follows immediately from the integral being an integral sum and the inequality (1.92):

$$\left|\sum_{k=1}^{n} f(\zeta_k)\, \Delta x_k\right| \leq \sum_{k=1}^{n} |f(\zeta_k)|\, \Delta x_k.$$

Performing the limit $\Delta \to 0$, we obtain the result sought for. **Q.E.D.**

4.3 Main Theorem of Integration: Indefinite Integrals

We mentioned above that it is possible to relate the definite integral to a derivative, and this relationship forms a very powerful tool for calculating integrals which bypasses calculation of the integral sums. To this end, we first have to define an integral as a function. Indeed, consider a continuous function $f(x)$ and let us integrate it between the limits a and x treating the upper limit $x > a$ as a *variable*:

$$F(x) = \int_a^x f(x')\mathrm{d}x'. \tag{4.30}$$

We used x' inside the integral instead of x in order to distinguish it from the upper limit x; note that the integration variable (x' in our case) is a dummy variable similar to the summation index, and hence any letter can be used for it without affecting the value of the integral. It is always a good idea to follow this simple rule to avoid mistakes.

Now, the integral defined above has its upper limit as a variable and hence it is a function of that variable; we denoted this function as $F(x)$. What we would like to do is to calculate its derivative. According to the definition of the derivative,

$$F'(x) = \lim_{\Delta x \to 0} \frac{\Delta F}{\Delta x},$$

where

$$\Delta F = F(x + \Delta x) - F(x) = \int_a^{x+\Delta x} f(x')dx' - \int_a^x f(x')dx'. \tag{4.31}$$

Because of the additivity property (4.5) of the integrals, this difference is simply an integral between x and $x + \Delta x$:

$$\Delta F = \int_x^{x+\Delta x} f(x')dx'. \tag{4.32}$$

This result is easy to understand in terms of the integral being an "area under the curve of the function" concept and is helped by Fig. 4.3: the integral from a to $x + \Delta x$ gives an area between these two points; when we subtract from it the integral between a and x, we obtain only the area between points x and $x + \Delta x$, i.e. exactly the integral (4.32). Next, we shall use the fact that the function $f(x)$ is continuous, and hence we can benefit from Theorem 4.6 of the previous subsection to write:

$$\Delta F = \int_x^{x+\Delta x} f(x')dx' = f(\zeta)\left[(x + \Delta x) - x\right] = f(\zeta)\,\Delta x, \tag{4.33}$$

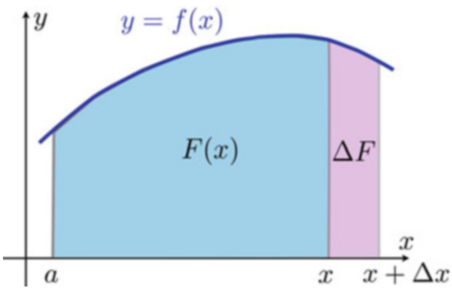


Fig. 4.3 An illustration to formula (4.32)

where ζ is somewhere between x and $x+\Delta x$, i.e. $\zeta = x + \vartheta\Delta x$ with ϑ being some number between 0 and 1. Hence,

$$F'(x) = \lim_{\Delta x\to 0}\frac{\Delta F}{\Delta x} = \lim_{\Delta x\to 0}\frac{f(\zeta)\,\Delta x}{\Delta x} = \lim_{\Delta x\to 0} f(\zeta) = \lim_{\Delta x\to 0} f(x+\vartheta\Delta x) = f(x). \tag{4.34}$$

Thus, the derivative of the integral $F(x)$ considered as a function of its upper limit gives the integrand $f(x)$. In this sense, the integration is *inverse* to differentiation.

This very important result is the key for calculating integrals. Indeed, what it tells us is that in order to calculate an integral of $f(x)$ with the upper limit being x we simply need to find a function $F(x)$ which derivative is exactly equal to $f(x)$. Of course, if such a function is found, then any function $F(x) + C$, where C is a constant, would also satisfy the required condition $F'(x) = f(x)$. This general relationship between the integrand $f(x)$ and the function $F(x)$ whose derivative is equal to $f(x)$ is usually expressed via the so-called *indefinite integral*:

$$\int f(x)dx = F(x) + C \quad \Longleftrightarrow \quad F'(x) = f(x). \tag{4.35}$$

Here $F(x)$ is any function whose derivative is equal to $f(x)$ and C is an arbitrary number (a constant). As we can calculate derivatives of arbitrary continuous functions, we can, by means of the inverse calculation, find the corresponding indefinite integrals of many functions.

Problem 4.8. *Let $F(x)$ be a function whose derivative is equal to $f(x)$. Prove that other functions whose derivative is equal to $f(x)$ can only differ by a constant.* [*Hint*: prove by contradiction. Also note that the only function whose derivative is equal to zero is a constant function. The latter follows from the Lagrange formula (3.64).]

Consider several examples of solving indefinite integrals. Since

$$\left(\frac{x^{\alpha+1}}{\alpha+1}\right)' = x^{\alpha} \quad \text{for any} \quad \alpha \neq -1,$$

then we can immediately write this also as

$$\int x^{\alpha}dx = \frac{x^{\alpha+1}}{\alpha+1} + C, \quad \alpha \neq -1. \tag{4.36}$$

The case of $\alpha = -1$ is a special one, since $(\ln x)' = x^{-1}$. To make our integral valid also for negative values of x, this formula is generalised: $(\ln|x|)' = x^{-1}$. Indeed, if $x > 0$, then we get the familiar result. If, however, $x < 0$, then

$$(\ln|x|)' = (\ln(-x))' = \frac{1}{-x}(-1) = \frac{1}{x},$$

i.e. the same result. Therefore, we arrive at the following indefinite integral:

$$\int \frac{dx}{x} = \ln|x| + C. \tag{4.37}$$

Problem 4.9. *Prove the following frequently encountered indefinite integrals:*

$$\int \cos x dx = \sin x + C \ , \quad \int \sin x dx = -\cos x + C, \tag{4.38}$$

$$\int e^x dx = e^x + C, \tag{4.39}$$

$$\int \frac{dx}{1+x^2} = \arctan x + C \ , \quad \int \frac{dx}{\sqrt{1-x^2}} = \arcsin x + C, \tag{4.40}$$

$$\int \frac{dx}{\cos^2 x} = \tan x + C \ , \quad \int \frac{dx}{\sin^2 x} = -\cot x + C. \tag{4.41}$$

$$\int \frac{dx}{\sqrt{1+x^2}} = \ln\left|x + \sqrt{1+x^2}\right| + C. \tag{4.42}$$

Now we are finally at the position to derive the formula for calculating definite integrals. For this, we consider again the definite integral with the variable upper limit. We found above that its derivative, $F'(x)$ is equal to the integrand, $f(x)$. It is therefore clear that the actual value of the integral should be

$$\int_a^x f(x')dx' = F(x) + C.$$

Here the constant C must be somehow related to the bottom limit a. To find C, we set $x = a$ in the formula above and use the fact that the integral between identical limits is equal to zero. We obtain then that $C = -F(a)$. This brings us to

$$\int_a^x f(x')dx' = F(x) - F(a).$$

It is now convenient to replace x as the upper limit with b, and we arrive at the *main formula of the integral calculus* (the so-called Newton–Leibniz formula):

$$\int_a^b f(x)dx = F(b) - F(a) = F(x)|_a^b \, . \tag{4.43}$$

Above a convenient notation has been introduced in the last passage for the difference of the function $F(x)$ between the end points of the integration interval, a and b; we shall be using it frequently.

This is our central result: calculation of a definite integral requires finding the function $F(x)$ whose derivative is equal to the integrand $f(x)$; after that the difference of the two values of $F(x)$ at b and a provides us with the final result. This means that we need to solve first for the indefinite integral, take its result $F(x)$ (ignoring any constant C) and calculate the difference of it at b and a.

As an example, let us calculate the integral between 0 and 1 of $f_1(x) = x^2$ and $f_2(x) = 2x^2 - 1$. In the case of $f_1(x)$, we have from (4.36): $F_1(x) = x^3/3 + C$ (this is obtained with $\alpha = 2$). We can ignore the constant to obtain:

$$\int_0^1 x^2 dx = \left.\frac{x^3}{3}\right|_0^1 = \left(\frac{x^3}{3}\right)_{x=1} - \left(\frac{x^3}{3}\right)_{x=0} = \frac{1^3}{3} - \frac{0^3}{3} = \frac{1}{3},$$

which is the same result as we got in Sect. 4.1 using the method of the integral sum. The function $f_2(x)$ can be considered as a linear combination of $f_1(x)$ and $f_3(x) = 1$ (a constant function). Since the integral of a linear combination of functions is a linear combination of integrals, we can write:

$$\int_0^1 \left(2x^2 - 1\right) dx = 2\int_0^1 x^2 dx - \int_0^1 dx = 2\cdot\frac{1}{3} - x|_0^1 = \frac{2}{3} - (1-0) = \frac{2}{3} - 1 = -\frac{1}{3}.$$

Problem 4.10. *Verify the following results:*

$$\int_{-1}^1 \left(-2e^x + x^3\right) dx = -2e + 2e^{-1}; \quad \int_0^a \left(x^2 - b\right)^3 dx = \frac{a^7}{7} - \frac{3a^5 b}{5} + a^3 b^2 - ab^3.$$

Formula (4.43) was proven above only for continuous functions. It may fail for functions which are piece-wise continuous. As the first example to illustrate this point, consider an integral of the sign function (2.2) which is discontinuous at $x = 0$, as shown in Fig. 2.1(b). For negative x, the sign function is equal to -1 and hence the indefinite integral is equal to $-x + C$; at positive x the sign function is $+1$ and the indefinite integral is $x + C$. Using formally Eq. (4.43), we then would have an integral between -1 and $+1$ equal to zero:

$$\int_{-1}^1 \mathrm{sgn}(x) dx = F(x)|_{-1}^1 = x|_{x=1} - (-x)|_{x=-1} = 1 - 1 = 0,$$

where we chose to have identical arbitrary constants for both indefinite integral functions, at positive and negative x. The obtained result makes perfect sense as the sign function is an odd function, and integration of an odd function between symmetric limits should give zero, see (4.8). Moreover, we can also split the integral into two and get the same result:

$$\int_{-1}^{1} \operatorname{sgn}(x)dx = \int_{-1}^{0} (-1)\, dx + \int_{0}^{1} (+1)\, dx = (-x)|_{-1}^{0} + x|_{0}^{1} = -1 + 1 = 0.$$

One may think that this example indicates the applicability of formula (4.43) even for discontinuous functions. However, remember that the correct result in our example was only obtained if the arbitrary constants were assumed to be the same when considering the indefinite integrals for negative and positive x, and there is no reason to believe that this must be the case generally, and hence one would be posed with the problem of choosing the constants in each case. As our second example, let us consider a function

$$f(x) = \begin{cases} f_1(x)\,, & \text{when} \quad x \geq 0 \\ f_2(x)\,, & \text{when} \quad x < 0 \end{cases} = \begin{cases} e^x\,, & \text{when} \quad x \geq 0 \\ 2e^x\,, & \text{when} \quad x < 0 \end{cases},$$

which makes a jump at $x = 0$. The indefinite integral of $f_1(x)$ is $F_1(x) = e^x + C$, while it is $F_2(x) = 2e^x + C$ for $f_2(x)$. Taking again identical constants for the two functions, we find for the definite integral between -1 and $+1$ the following value using directly formula (4.43):

$$\int_{-1}^{1} f(x)dx = e^x|_{x=1} - 2e^x|_{x=-1} = e - 2e^{-1}.$$

On the other hand, if we split the integral into two, we would get a different answer:

$$\begin{aligned} \int_{-1}^{1} f(x)dx &= \int_{-1}^{0} 2e^x dx + \int_{0}^{1} e^x dx = 2e^x|_{-1}^{0} + e^x|_{0}^{1} = 2\left(1 - e^{-1}\right) + (e - 1) \\ &= 1 - 2e^{-1} + e, \end{aligned}$$

which differs by a constant from the result found initially with the Newton–Leibniz formula. So, a general conclusion from these examples must be that if a function makes a finite jump (has a discontinuity) at some point c which is within the integration region, $a < c < b$, then one has to split the interval into regions where the function is continuous ($a \leq x < c$ and $c \leq x \leq b$), and then integrate separately within each such region as we did above; applying the Newton–Leibniz formula in those cases may give incorrect results.

There could only be one arbitrary constant in a single indefinite integral, even though it may consist of several functions as illustrated by this example:

$$I = \int \left(2x^2 + \frac{3}{x} - 2\sin x\right) dx.$$

We split the integral into three and take each of them individually, so that three different constants emerge:

$$I = \int 2x^2 dx + \int \frac{3}{x} dx - \int 2\sin x dx = 2\int x^2 dx + 3\int \frac{dx}{x} - 2\int \sin x dx$$
$$= \left(2\frac{x^3}{3} + C_1\right) + (3\ln|x| + C_2) - (-2\cos x + C_3)$$
$$= \frac{2}{3}x^3 + 3\ln|x| + 2\cos x + C,$$

where, as we can see, all the constants combine into one: $C = C_1 + C_2 + C_3$.

In many cases already simple algebraic manipulations of the integrand may result in a combination of elementary functions which can be integrated:

$$\int \frac{x+2}{x} dx = \int \left(1 + \frac{2}{x}\right) dx = \int dx + 2\int \frac{dx}{x} = x + 2\ln|x| + C,$$
$$\int \frac{e^{-x}+1}{e^{-x}} dx = \int \left(1 + \frac{1}{e^{-x}}\right) dx = \int dx + \int e^x dx = x + e^x + C.$$

Problem 4.11. *Prove the following identities:*

$$\int \cot^2 x dx = -x - \cot x + C\ , \quad \int \tan^2 x dx = -x + \tan x + C.$$

[*Hint*: express cotangent and tangent functions via sine and cosine functions.]

$$\int (2x+1)^3\, dx = 2x^4 + 4x^3 + 3x^2 + x + C\ ,$$
$$\int \sin(x+a)\, dx = -\cos(x+a) + C.$$

[*Hint*: use the binomial formula in the first and Eqs. (2.46) and (2.47) in the second case.]

$$\int \sin^2 \frac{x}{2} dx = \frac{1}{2}(x - \sin x) + C\ , \quad \int \cos^2 \frac{x}{2} dx = \frac{1}{2}(x + \sin x) + C.$$

[*Hint*: use Eq. (2.44).]

However, in many cases more involved methods must be used. Some of the most frequent ones will be considered in the next section.

4.4 Indefinite Integrals: Main Techniques

As we learned in the previous section, calculation of a definite integral requires, first of all, solving the corresponding indefinite integral. This requires finding a function which, when differentiated, gives the integrand itself. Of course, one may select all sorts of functions and then differentiate them to get a "big" table of derivatives of various functions. That may be useful since by looking at this table "in reverse" it is possible to find solutions of many indefinite integrals (and it is a good idea for the reader to make his/her own table of various indefinite integrals while reading the rest of this chapter). The problem is that there might be no solution for a particular integral of the actual interest in the table as it is impossible to choose "all possible functions" for differentiation due to an infinite number of possibilities. We need to develop various techniques for solving integrals. This is the main theme of this section. It is of course sufficient to discuss indefinite integrals, and this is what we shall be doing in the first instance.

4.4.1 Change of Variables

Sometimes by changing the integration variable it is possible to transform the integrand into something which can be recognised as a linear combination of functions for which indefinite integrals are known (and hopefully listed in our big table of integrals).

Consider an integral:

$$\int f(x)dx = F(x) + C,$$

where $F'(x) = f(x)$. Let $x = x(t)$ be a function of a variable t, so that $F(x) = F(x(t))$. Using the chain rule, we can write:

$$\frac{d}{dt}F(x(t)) = \frac{dF}{dx}\frac{dx}{dt} = F'(x)x'(t) = f(x)x'(t) = f(x(t))\,x'(t).$$

This expression simply means that if we want to integrate the function $\phi(t) = f(x(t))\,x'(t)$ *with respect to* t, the result must be $F(x(t)) + C$, and this is true for any choice of the function $x = x(t)$. We conclude:

$$\int f(x(t))\,x'(t)dt = F(x(t)) + C = F(x) + C = \int f(x)dx,$$

i.e. it is exactly the same integral we are seeking to solve. The formal procedure we have performed can simply be considered as a replacement of x with $x(t)$ and of

dx with $x'(t)dt$. We note that the differential for $x(t)$ is exactly the latter expression $x'(t)dt$ and it is indeed denoted dx. This simple argument proves the consistency of the notations for the integral we have been using.[1]

Before we consider various cases in a more systematic way, let us first look at some simple examples.

Let us calculate the integral

$$I = \int xe^{x^2} dx.$$

In this integral the function x^2 in the exponent is a cumbersome one, so a wise choice here is to make a substitution. The simplest choice is to introduce a new variable $t = x^2$ which turns the rather complex exponential function e^{x^2} into a simpler one, e^t. We have $dt = d\left(x^2\right) = \left(x^2\right)' dx = 2xdx$, from where it follows that $xdx = dt/2$. Alternatively, we can solve $t = x^2$ for $x(t) = \pm t^{1/2}$, yielding $dx = \left(\pm t^{1/2}\right)' dt = \pm\frac{1}{2}t^{-1/2}dt$ and hence $xdx = \left(\pm t^{1/2}\right)\left(\pm\frac{1}{2}t^{-1/2}\right)dt = \frac{1}{2}dt$, the same result. Here we explicitly made use of the fact that dx is the differential of x. With this substitution, we can proceed to the integral itself:

$$I = \int e^{x^2}(xdx) = \int e^t \frac{1}{2}dt = \frac{1}{2}\int e^t dt = \frac{1}{2}e^t + C = \frac{1}{2}e^{x^2} + C.$$

It is easy to check that, indeed, $\frac{1}{2}\left(e^{x^2}\right)' = \frac{1}{2}e^{x^2}\left(x^2\right)' = \frac{1}{2}e^{x^2}2x = e^{x^2}x$, exactly the integrand of I.

Next, we calculate the integral

$$I = \int \frac{x^2 - 1}{2x - 5} dx.$$

In this case we use $t = 2x - 5$, so that $x = (t + 5)/2$ and hence $x^2 - 1 = t^2/4 + 5t/2 + 21/4$. This substitution allows us to manipulate the numerator in the integrand as follows:

$$\begin{aligned} I &= \left| \begin{array}{c} t = 2x - 5 \\ dt = 2dx \end{array} \right| = \int \frac{\frac{1}{4}t^2 + \frac{5}{2}t + \frac{21}{4}}{t}\frac{dt}{2} = \int \left(\frac{1}{8}t + \frac{5}{4} + \frac{21}{8}\frac{1}{t}\right) dt \\ &= \frac{t^2}{16} + \frac{5t}{4} + \frac{21}{8}\ln|t| + C \\ &= \frac{(2x-5)^2}{16} + \frac{5(2x-5)}{4} + \frac{21}{8}\ln|2x-5| + C. \end{aligned}$$

[1] Which is due to Gottfried Leibniz and Jean Baptiste Joseph Fourier.

In this example a very convenient form of working out the integral by a change of the variable has been proposed, and we urge the reader to make full use of it: all the working out needed to change the integration variable is placed between two vertical lines; we only wrote there the actual change of the variable ($t = t(x)$) and the connection of the differentials dt and dx. However, the inverse transformation $x = x(t)$ (of course, only if needed and possible) together with other relevant algebraic manipulations (if not too cumbersome) may also be written there. This way of writing is convenient as it allows constructing the whole derivation without interrupting the lines of algebra and in a short a clear way. We shall be using this method frequently.

Problem 4.12. *Prove the following identities using the appropriate change of variables:*

$$\int a^x dx = \frac{a^x}{\ln a} + C \quad (\textit{with } a > 0), \quad \int \sin(ax+b)\,dx = -\frac{1}{a}\cos(ax+b) + C,$$

$$\int (ax+b)^n\,dx = \frac{(ax+b)^{n+1}}{a(n+1)} + C, \quad \int \cos(ax+b)\,dx = \frac{1}{a}\sin(ax+b) + C,$$

$$\int \frac{dx}{\sqrt{a^2-x^2}} = \arcsin\frac{x}{a} + C,$$

$$\int \frac{dx}{ax+b} = \frac{1}{a}\ln|ax+b| + C, \quad \int \tan x dx = -\ln|\cos x| + C,$$

$$\int \cot x dx = \ln|\sin x| + C,$$

$$\int x^3\sqrt{1+x^2}dx = \frac{1}{15}\left(1+x^2\right)^{3/2}\left(3x^2-2\right), \quad \int \frac{dx}{x^2+a^2} = \frac{1}{a}\arctan\frac{x}{a} + C,$$

$$\int \frac{1}{x^2}\sin\frac{1}{x}dx = \cos\frac{1}{x} + C, \quad \int \frac{1}{x^2}\cos\frac{1}{x}dx = -\sin\frac{1}{x} + C,$$

$$\int \frac{dx}{(x+3)^4} = -\frac{(x+3)^{-3}}{3} + C.$$

Problem 4.13. *Let the function $f(x)$ be periodic with the period of X, i.e. for any x we have $f(x+X) = f(x)$. Prove that for any a*

$$\int_a^{a+X} f(x)dx = \int_0^X f(x)dx. \tag{4.44}$$

Consider specifically $f(x) = \sin(x)$ for which $X = 2\pi$, sketch the function and reconcile with this rule.

4.4.2 Integration by Parts

This method is based on the following observation: if $u(x)$ and $v(x)$ are two functions, then, as we know well, $(uv)' = uv' + u'v$. Therefore, the function $F(x) = u(x)v(x) + C$ is an indefinite integral of the function $f(x) = u(x)v'(x) + u'(x)v(x)$:

$$\int_a^b \left[uv' + u'v\right] dx = uv|_a^b ,$$

or, since the integral of a sum is a sum of integrals:

$$\int_a^b uv'dx = uv|_a^b - \int_a^b vu'dx. \tag{4.45}$$

This result is usually written in a somewhat shorter form:

$$\int_a^b udv = uv|_a^b - \int_a^b vdu, \tag{4.46}$$

where we introduced the differentials $du = u'dx$ and $dv = v'dx$ of the functions $u(x)$ and $v(x)$, respectively. These were formulae for the definite integral. For indefinite integrals the analogue expressions are the following:

$$\int udv = uv - \int vdu \quad \text{or} \quad \int uv'dx = uv - \int vu'dx. \tag{4.47}$$

There is no need here to introduce the constant as it can be done at the last stage when taking the integral $\int vdu$ in the right-hand side.

The usefulness of this trick "by parts" is in that in the right-hand side the function $u(x)$ is differentiated and, at the same time, instead of the function v', we have $v(x)$. This interchange may help in calculating the original integral in the left-hand side. Of course, in the original integral we have just a single function $f(x)$, and various "divisions" or "factorisations" of it into $u(x)$ and $v(x)$ may be possible; it is not known which one would succeed, so sometimes different divisions are needed to try and then see which one would work. In many cases neither of "reasonable" choices works and one has to think of some other "trick". In some cases more than one integration by parts is required, one after another. In some other cases algebraic equations for the integral in question may be obtained by using this method which, when solved, gives the required answer.

As an example, let us calculate the indefinite integral $I = \int xe^x dx$. The idea is to have in the right-hand side just the exponential function on its own since we know perfectly well how to integrate it. If first we try taking $u = e^x$ and $dv = xdx$, then $u' = e^x$ and $v = x^2/2$ (remember, the constant C is introduced at the last step and hence is now ignored), so that we get

$$\int xe^x dx = \int \underbrace{e^x}_{u} \underbrace{xdx}_{dv} = \underbrace{e^x}_{u} \underbrace{\frac{x^2}{2}}_{v} - \int \underbrace{\frac{x^2}{2}}_{v} \underbrace{e^x dx}_{du} = \frac{1}{2}x^2 e^x - \frac{1}{2}\int x^2 e^x dx.$$

We see that, instead of getting rid of the x in the integrand, we raised its power to two and got x^2 instead. This result shows that our choice of u and dv was not successful, and we should try another one to see if it would work. However, our unsuccessful result may give us a hint as to how to proceed: it shows (if we look from right to left) that the integral $\int x^2 e^x dx$ can in fact be made related to the integral $\int xe^x dx$ with the lower power of x:

$$\int x^2 e^x dx = x^2 e^x - 2\int xe^x dx, \tag{4.48}$$

if x^2 is made the $u(x)$, so that upon differentiation in the right-hand side we would get $u' = 2x$, i.e. one power less. So, coming back to our original integral, this time we make the following decision:

$$\int xe^x dx = \left| \begin{array}{cc} u = x \ , & du = 1 \cdot dx = dx \\ dv = e^x dx \ , & v = e^x \end{array} \right| = xe^x - \int e^x dx = xe^x - e^x + C,$$

which is the final result. It is easy to see that $(xe^x - e^x)' = (xe^x)' - (e^x)' = e^x + xe^x - e^x = xe^x$, i.e. the integrand of the original integral; our second choice succeeded. If we now need to calculate the integral with x^2e^x, then two integrations by parts would be needed: one to reduce the power of the x from two to one as in (4.48), and then the calculation of $\int xe^x dx$ as we did above:

$$\begin{aligned}
\int x^2 e^x dx &= \begin{vmatrix} u = x^2 \ , & du = 2xdx \\ dv = e^x dx \ , & v = e^x \end{vmatrix} = x^2 e^x - 2\int xe^x dx \\
&= \begin{vmatrix} u = x \ , & du = dx \\ dv = e^x dx \ , & v = e^x \end{vmatrix} = x^2 e^x - 2\left[xe^x - \int e^x dx \right] \\
&= x^2 e^x - 2xe^x + 2e^x + C = e^x \left(x^2 - 2x + 2 \right) + C .
\end{aligned}$$

Let us now calculate this integral between 0 and 10:

$$\int_0^{10} x^2 e^x dx = e^x \left(x^2 - 2x + 2 \right) \Big|_0^{10} = e^{10} \left(10^2 - 20 + 2 \right) - e^0 2 = 82e^{10} - 2 .$$

Problem 4.14. *Calculate the following integrals by parts:*

$$\int (2x+5)^3 e^{-2x} dx \quad \left[\textit{Answer:} \quad -\left(4x^3 + 36x^2 + 111x + 118 \right) e^{-2x} + C \right];$$

$$\int x^3 e^x dx \quad \left[\textit{Answer:} \quad \left(x^3 - 3x^2 + 6x - 6 \right) e^x + C \right];$$

$$\int x^4 e^x dx \quad \left[\textit{Answer:} \quad \left(x^4 - 4x^3 + 12x^2 - 24x + 24 \right) e^x + C \right].$$

Problem 4.15. *Using integration by parts, derive a recurrence relation $I_n = x^n e^x - nI_{n-1}$ for the integral $I_n = \int x^n e^x dx$ (where $n = 0, 1, 2, \ldots$). Then by repeated application of the recurrence formula, establish the following general result:*

$$\int x^n e^x dx = e^x \sum_{k=0}^{n} (-1)^k \frac{n!}{(n-k)!} x^{n-k} + C .$$

Using induction, prove that this formula is indeed correct.

Problem 4.16. *Calculate*

$$\int x\sin x dx \quad [\textit{Answer:} \quad \sin x - x\cos x + C]\,;$$

$$\int x\cos x dx \quad [\textit{Answer:} \quad \cos x + x\sin x + C]\,;$$

$$\int \frac{1}{x^3}\sin\frac{1}{x}dx \quad \left[\textit{Answer:}\ -\sin\frac{1}{x} + \frac{1}{x}\cos\frac{1}{x} + C\right]\,;$$

$$\int \frac{1}{x^4}\sin\frac{1}{x}dx \quad \left[\textit{Answer:}\ -\frac{1}{x^2}\left[2x\sin\frac{1}{x} + \left(2x^2-1\right)\cos\frac{1}{x}\right] + C\right]\,;$$

$$\int x^2\cos x dx \quad \left[\textit{Answer:} \quad 2x\cos x + \left(x^2-2\right)\sin x + C\right]\,;$$

$$\int x^3\cos x dx \quad \left[\textit{Answer:} \quad 3\left(x^2-2\right)\cos x + x\left(x^2-6\right)\sin x + C\right]\,.$$

These examples should convince us that the integration by parts can be used for calculating integrals $\int P_n(x)e^{\alpha x}dx$, $\int P_n(x)\sin(ax+b)\,dx$ and $\int P_n(x)\cos(ax+b)\,dx$, where $P_n(x)$ is the n-th degree polynomial. Note that more general arguments to the exponential, sine and cosine functions can indeed be used as these arguments can always be replaced by t using the appropriate change of variables; the polynomial will remain a polynomial of the same degree (but with different coefficients).

Fortunately, integration by parts has a much wider scope. It can also be used for integrating functions which are products of a polynomial and $\ln x$ or any of the inverse trigonometric functions as illustrated below:

$$\int \arctan x dx = \left| \begin{array}{c} u = \arctan x\,, \quad du = \frac{1}{1+x^2}dx \\ dv = dx\,, \quad v = x \end{array} \right| = x\arctan x - \int \frac{xdx}{1+x^2}$$

$$= x\arctan x - \frac{1}{2}\ln\left(x^2+1\right) + C,$$

where in the last integral we have made a change of variable $t = 1 + x^2$, $dt = 2xdx$.

Problem 4.17. *Calculate*

$$\int x \arcsin x dx \quad \left[\textit{Answer:} \quad \frac{x}{4}\sqrt{1-x^2} + \frac{2x^2-1}{4}\arcsin x + C \right];$$

$$\int x^2 \arcsin x dx \quad \left[\textit{Answer:} \quad \frac{x^2+2}{9}\sqrt{1-x^2} + \frac{x^3}{3}\arcsin x + C \right];$$

$$\int \arccos x dx \quad \left[\textit{Answer:} \quad -\sqrt{1-x^2} + x\arccos x + C \right];$$

$$\int x \arccos x dx \quad \left[\textit{Answer:} \quad -\frac{x}{4}\sqrt{1-x^2} + \frac{x^2}{2}\arccos x + \frac{1}{4}\arcsin x + C \right];$$

$$\int x^3 \arccos x dx \quad \left[\textit{Answer:}. \quad -\frac{\left(2x^2+3\right)x}{32}\sqrt{1-x^2} + \frac{x^4}{4}\arccos x + \frac{3}{32}\arcsin x + C \right];$$

$$\int \ln x dx \quad [\textit{Answer:} \quad x\ln x - x + C] \ ;$$

$$\int x \ln x dx \quad \left[\textit{Answer:} \quad -\frac{x^2}{4}(1-2\ln x) + C \right].$$

The last example which we shall consider is of the following two integrals:

$$I_S = \int e^x \sin x dx = \left| \begin{matrix} u = e^x \ , & du = e^x dx \\ dv = \sin x dx \ , & v = -\cos x \end{matrix} \right| = -e^x \cos x + \int e^x \cos x dx$$
$$= -e^x \cos x + I_C$$

and

$$I_C = \int e^x \cos x dx = \left| \begin{matrix} u = e^x \ , & du = e^x dx \\ dv = \cos x dx \ , & v = \sin x \end{matrix} \right| = e^x \sin x - \int e^x \sin x dx$$
$$= e^x \sin x - I_S.$$

It is seen that a single integration by parts relates one integral to the other. Therefore, double integration by parts results in an algebraic equation for one of the integrals:

$$I_S = -e^x \cos x + I_C = -e^x \cos x + [e^x \sin x - I_S] = e^x(-\cos x + \sin x) - I_S,$$

which, when solved with respect to I_S, finally yields:

$$I_S = \frac{1}{2}e^x\left(-\cos x + \sin x\right).$$

As a by-product, we also have

$$I_C = e^x \sin x - I_S = \frac{1}{2}e^x\left(\cos x + \sin x\right).$$

A product of a polynomial $P_n(x)$ with several of exponential or trigonometric functions is dealt with in more or less the same way:

$$\int xe^x \sin x dx = \left| \begin{array}{cc} u = x\,, & du = dx \\ dv = e^x \sin x dx\,, & v = I_S \end{array} \right| = xI_S - \frac{1}{2}\int e^x\left(\sin x - \cos x\right)dx,$$

where we again denoted the intermediate integral $\int e^x \sin x dx$ as I_S for convenience. The integral we have arrived at is a difference of the familiar I_S and I_C integrals, which allows us to finally write:

$$\int xe^x \sin x dx = xI_S - \frac{1}{2}\left(I_S - I_C\right) = \left(x - \frac{1}{2}\right)I_S + \frac{1}{2}I_C$$
$$= \frac{1}{2}e^x\left[(1-x)\cos x + x\sin x\right] + C.$$

Problem 4.18. *Calculate*

$$\int xe^x \cos x dx \quad \left[\textit{Answer:}\ \frac{1}{2}e^x\left\{(x-1)\sin x + x\cos x\right\} + C\right];$$

$$\int x^2 e^x \sin x dx \quad \left[\textit{Answer:}\ \frac{1}{2}e^x\left(x-1\right)\left\{-\left(x-1\right)\cos x + \left(x+1\right)\sin x\right\} + C\right];$$

$$\int x^2 e^x \cos x dx \quad \left[\textit{Answer:}\ \frac{1}{2}e^x\left(x-1\right)\left\{\left(x-1\right)\sin x + \left(x+1\right)\cos x\right\} + C\right].$$

Problem 4.19. *Show that for any positive integers n and m,*

$$I_{n,m} = \int_0^1 x^n \ln^m x dx = \frac{(-1)^m\, m!}{(n+1)^{m+1}}.$$

[*Hint*: first of all, establish the recurrence relation $I_{n,m} = -\left[m/\left(n+1\right)\right] I_{n,m-1}$.]

Problem 4.20. *Similarly, show that for any positive integers n and m,*

$$J_{n,m} = \int_0^1 (1-x)^n x^m dx = \frac{n!m!}{(n+m+1)!}.$$

[*Hint*: first of all, establish the recurrence relation $J_{n,m} = [n/(n+m+1)]$ $J_{n-1,m}$ by writing $(1-x) = (1/x - 1)\,x$.]

Problem 4.21. *Show, using integration by parts, that $\int f(x)f'(x)dx = f(x)^2 + C$. This simple result can also be considered as an application of the differentials, since $df = (df/dx)\,dx = f'(x)dx$. Indeed,*

$$\int f(x)f'(x)dx = \int fdf = \frac{f^2}{2} + C = \frac{f(x)^2}{2} + C.$$

Convince yourself that this is true also by differentiating the right-hand side.

4.4.3 Integration of Rational Functions

In the case of rational functions $f(x) = P_n(x)/Q_m(x)$ with polynomials $P_n(x)$ and $Q_m(x)$ of degrees n and m, respectively, the corresponding indefinite integrals can always be calculated. As we saw in Sect. 2.3.2, a rational function can always be decomposed into a sum of simpler functions: a polynomial (which is only present if $n \geq m$) and the following functions:

$$f_k(x) = \frac{1}{(x-a)^k} \quad \text{and/or} \quad g_k(x) = \frac{Ax+B}{(x^2+px+q)^k}$$

with $k = 1, 2, \ldots$. Before we consider an example of more complex rational functions, let us first make sure that we know how to integrate the above two functions.

Integration of $f_1(x)$ (for $k = 1$) is trivial and leads to a logarithm. Integration of f_k for $k > 1$ is done using the substitution $t = x - a$ and is also simple; both of these integrals we have seen before. The integration of $g_1(x)$ can be done by a substitution which follows from making the complete square of the denominator:

$$x^2 + px + q = \left[x^2 + 2\left(\frac{p}{2}\right)x + \left(\frac{p}{2}\right)^2\right] - \left(\frac{p}{2}\right)^2 + q = \left(x + \frac{p}{2}\right)^2 - D,$$

where $D = p^2/4 - q$ is definitely negative because only in this case the square polynomial $x^2 + px + q$ (with D being its discriminant) corresponds to two complex conjugate roots (see Sects. 1.5 and 2.3.2).[2] Now, the substitution $t = x + p/2$ is made and the function $g_1(x)$ is easily integrated.

Consider, for instance, this integral:

$$\int \frac{2x+1}{x^2-2x+2}dx = \int \frac{2x+1}{(x-1)^2+1}dx = \left| \begin{array}{l} t = x-1 \\ dt = dx \end{array} \right| = \int \frac{2t+3}{t^2+1}dt$$
$$= 2\int \frac{tdt}{t^2+1} + 3\int \frac{dt}{t^2+1}.$$

The first integral is calculated using the substitution $y = t^2 + 1$, while the second integral leads to $\arctan t$:

$$\int \frac{2x+1}{x^2-2x+2}dx = \ln\left(1+t^2\right) + 3\arctan t + C$$
$$= \ln\left(x^2-2x+2\right) + 3\arctan(x-1) + C.$$

Problem 4.22. *Calculate*

$$\int \frac{x-1}{x^2+4x+6}dx \quad \left[\textit{Answer:} \quad -\frac{3}{\sqrt{2}}\arctan\frac{x+2}{\sqrt{2}} + \frac{1}{2}\ln\left(x^2+4x+6\right) + C\right];$$

$$\int \frac{-3x+2}{x^2-6x+10}dx \quad \left[\textit{Answer:} \ 7\arctan(3-x) - \frac{3}{2}\ln\left(x^2-6x+10\right) + C\right].$$

Calculation of the integral $g_k(x)$ with $k \geq 2$ is a bit trickier. Firstly, we build the complete square as in the previous case of $k = 1$ and make the corresponding change of variables. Two types of integrals are then encountered which require our attention:

$$h_k = \int \frac{tdt}{\left(t^2+a^2\right)^k} \quad \text{and} \quad r_k = \int \frac{dt}{\left(t^2+a^2\right)^k}.$$

The first of these, h_k, is trivially calculated with the substitution $y = t^2 + a^2$:

$$h_k = \int \frac{tdt}{\left(t^2+a^2\right)^k} = \frac{1}{2(1-k)}\left(t^2+a^2\right)^{1-k}.$$

[2] Of course, when $D < 0$ the parabola appears completely above the horizontal x axis and hence does not have real roots.

The second integral, r_k, is calculated with the following trick leading to a recurrence relation whereby r_k is expressed via the same type of integrals with lower values of k:

$$r_k = \int \frac{dt}{(t^2+a^2)^k} = \frac{1}{a^2}\int \frac{\left[\left(t^2+a^2\right)-t^2\right]dt}{(t^2+a^2)^k}$$
$$= \frac{1}{a^2}\left[\int \frac{dt}{(t^2+a^2)^{k-1}} - \int \frac{t^2dt}{(t^2+a^2)^k}\right] = \frac{r_{k-1}}{a^2} - \frac{1}{a^2}J_k,$$

where the integral J_k we take by parts:

$$J_k = \int \frac{t^2dt}{(t^2+a^2)^k} = \left| \begin{array}{cc} u = t\,, & du = dt \\ dv = t\left(t^2+a^2\right)^{-k}dt\,, & v = h_k \end{array} \right| = th_k - \int h_k dt$$
$$= \frac{t}{2(1-k)}\left(t^2+a^2\right)^{1-k} - \frac{1}{2(1-k)}\int \left(t^2+a^2\right)^{1-k}dt$$
$$= \frac{t}{2(1-k)}\left(t^2+a^2\right)^{1-k} - \frac{r_{k-1}}{2(1-k)},$$

which result finally in the following recurrence relation for r_k:

$$r_k = \frac{r_{k-1}}{a^2} - \frac{1}{a^2}J_k = \frac{r_{k-1}}{a^2}\left[1 + \frac{1}{2\,(1-k)}\right] - \frac{t}{2(1-k)a^2}\left(t^2+a^2\right)^{1-k}.$$

It is seen that r_k is directly related to r_{k-1}. This method allows expressing recursively the given r_k to r_{k-1}, then the latter to r_{k-2}, etc., until one arrives at r_1 which we have already encountered above: $r_1 = (1/a)\arctan(x/a)$. This way it is possible to express any of the g_k integrals for any integer k via elementary functions.

Let us consider an example:

$$\int \frac{2x+1}{(x^2-2x+2)^2}dx = \int \frac{2x+1}{\left[(x-1)^2+1\right]^2}dx = \left| \begin{array}{c} t = x-1 \\ dt = dx \end{array} \right|$$
$$= \int \frac{2t+3}{(t^2+1)^2}dt = 2h_2 + 3r_2.$$

Here

$$h_2 = \int \frac{tdt}{(t^2+1)^2} = \left| \begin{array}{c} y = t^2+1 \\ dy = 2tdt \end{array} \right| = \frac{1}{2}\int \frac{dy}{y^2} = -\frac{1}{2y} = -\frac{1}{2\,(t^2+1)}$$

and

$$r_2 = \int \frac{dt}{(t^2+1)^2} = \int \frac{\left(t^2+1\right)-t^2}{(t^2+1)^2}dt = \int \frac{dt}{t^2+1} - \int \frac{t^2dt}{(t^2+1)^2} = \arctan t - J_2,$$

where the integral J_2 we calculate by parts:

$$J_2 = \int \frac{t^2dt}{(t^2+1)^2} = \left| \begin{array}{cc} u = t\,, & du = dt \\ dv = t\left(t^2+1\right)^{-2}dt\,, & v = h_2(t) \end{array} \right|$$
$$= -\frac{t}{2\left(t^2+1\right)} + \frac{1}{2}\int \frac{dt}{t^2+1} = -\frac{t}{2\left(t^2+1\right)} + \frac{1}{2}J_1 = -\frac{t}{2\left(t^2+1\right)} + \frac{1}{2}\arctan t.$$

Combining all contributions together, we can finally state the result:

$$\int \frac{2x+1}{(x^2-2x+2)^2}dx = \frac{3x-5}{2\left(x^2-2x+2\right)} + \frac{3}{2}\arctan(x-1) + C.$$

Problem 4.23. *Calculate*

$$\int \frac{x-1}{(x^2+4x+6)^3}dx \quad \left[\textit{Answer:} \quad -\frac{9\sqrt{2}}{64}\arctan\frac{x+2}{\sqrt{2}} - \frac{9x^3+54x^2+138x+140}{32\left(x^2+4x+6\right)^2} + C\right];$$

$$\int \frac{-3x+2}{(x^2-6x+10)^2}dx \quad \left[\textit{Answer:} \quad \frac{7}{2}\arctan(3-x) - \frac{24-7x}{2\left(x^2-6x+10\right)} + C\right].$$

Once we have discussed all the necessary elementary integrals, we are now ready to consider integrals of arbitrary rational functions. The calculation consists of first decomposing the rational function into elementary fractions as discussed at the beginning of this section and especially in Sect. 2.3.2, and then, second, integrating each elementary fraction as we have just demonstrated.

As an example, we calculate the integral

$$J = \int \frac{3x^2+x-1}{(x-1)^2\left(x^2+1\right)}dx.$$

We start by decomposing the fraction. This was done in Eq. (2.17); therefore, we can immediately write:

$$J = 2\int \frac{1}{x-1}dx + \frac{3}{2}\int \frac{dx}{(x-1)^2} - \frac{1}{2}\int \frac{4x+1}{x^2+1}dx$$
$$= 2\ln|x-1| - \frac{3}{2(x-1)} - 2\int \frac{xdx}{x^2+1} - \frac{1}{2}\int \frac{dx}{x^2+1}$$
$$= 2\ln|x-1| - \frac{3}{2(x-1)} - \ln\left(x^2+1\right) - \frac{1}{2}\arctan x + C.$$

Problem 4.24. *Calculate the following integrals of rational functions:*

$$\int \frac{x^2-x+5}{(2x-1)(x^2+4x+4)}dx \qquad \left[\textit{Answer:}\quad \frac{11}{5(x+2)} + \frac{3}{25}\ln|2+x| + \frac{19}{50}\ln|2x-1| + C\right];$$

$$\int \frac{x^3+x^2+1}{(x-1)^2(x^2+4)}dx \qquad \left[\textit{Answer:}\quad -\frac{3}{5(x-1)} + \frac{41}{50}\arctan\frac{x}{2} + \frac{3}{25}\ln\left(x^2+4\right) + \frac{19}{25}\ln|x-1| + C\right];$$

$$\int \frac{dx}{1-x^3} \qquad \left[\textit{Answer:}\quad \frac{1}{\sqrt{3}}\arctan\frac{2x+1}{\sqrt{3}} + \frac{1}{6}\ln\frac{1+x+x^2}{(x-1)^2} + C\right];$$

$$\int \frac{dx}{1+x^4} \qquad \left[\textit{Answer:}\quad -\frac{1}{2\sqrt{2}}\left[\arctan\left(1-\sqrt{2}x\right) - \arctan\left(1+\sqrt{2}x\right)\right] - \frac{1}{4\sqrt{2}}\ln\frac{1-\sqrt{2}x+x^2}{1+\sqrt{2}x+x^2} + C\right];$$

$$\int \frac{x^2dx}{1+x^4} \qquad \left[\textit{Answer:}\quad -\frac{1}{2\sqrt{2}}\left[\arctan\left(1-\sqrt{2}x\right) - \arctan\left(1+\sqrt{2}x\right)\right] + \frac{1}{4\sqrt{2}}\ln\frac{1-\sqrt{2}x+x^2}{1+\sqrt{2}x+x^2} + C\right];$$

$$\int \frac{x^5+2x^3-1}{x^2+4x+4}dx \qquad \left[\textit{Answer:}\quad \frac{x^4}{4} - \frac{4x^3}{3} + 7x^2 - 40x + \frac{49}{x+2} + 104\ln|x+2| + C\right].$$

4.4.4 Integration of Trigonometric Functions

Integrals containing a rational function $R(\sin x, \cos x)$ of the sine and cosine of the same angle x can always be turned into an integral of a rational function, which, as we know from the previous subsection, can always be calculated. The trick is based on replacing the sine and cosine functions by rational expressions using the substitution

$$t = \tan\frac{x}{2}\,, \quad \text{so that} \quad dt = \frac{dx}{2\cos^2\frac{x}{2}} = \frac{\sin^2\frac{x}{2} + \cos^2\frac{x}{2}}{2\cos^2\frac{x}{2}}dx$$
$$= \frac{1}{2}\left(t^2+1\right)dx \quad \Longrightarrow \quad dx = \frac{2dt}{1+t^2}, \tag{4.49}$$

so that both sine and cosine (see Eq. (2.65)) are expressed via t:

$$\sin x = \frac{2\tan\frac{x}{2}}{1+\tan^2\frac{x}{2}} = \frac{2t}{1+t^2} \quad \text{and} \quad \cos x = \frac{1-\tan^2\frac{x}{2}}{1+\tan^2\frac{x}{2}} = \frac{1-t^2}{1+t^2}. \tag{4.50}$$

It is clear that any rational function of the sine and cosine functions will be transformed by means of this substitution into some rational function of t. This substitution is sometimes unnecessarily cumbersome since other methods (if known) may give the answer in a somewhat simpler way; still, it is good to know that it always leads to a solution.

As an example we shall calculate the following integral:

$$\int \frac{1+\sin x}{1+\cos x}dx = \left| \begin{array}{c} t = \tan\frac{x}{2} \\ dx = 2dt/\left(1+t^2\right) \end{array} \right| = \int \frac{1+2t/\left(1+t^2\right)}{1+\left(1-t^2\right)/\left(1+t^2\right)}\frac{2dt}{1+t^2}$$
$$= \int \frac{1+t^2+2t}{1+t^2}dt = \int \left(1+\frac{2t}{1+t^2}\right)dt$$
$$= t + \int \frac{2tdt}{1+t^2} = \left| \begin{array}{c} y = 1+t^2 \\ dy = 2tdt \end{array} \right| = t + \int \frac{dy}{y} = t + \ln|y| = t + \ln\left|1+t^2\right|$$
$$= \tan\frac{x}{2} + \ln\left|1+\tan^2\frac{x}{2}\right| = \tan\frac{x}{2} + \ln\left|\frac{1}{\cos^2\frac{x}{2}}\right| = \tan\frac{x}{2} - 2\ln\left|\cos\frac{x}{2}\right| + C.$$

We have introduced the constant C only at the very last step.

Problem 4.25. *Calculate the following integrals:*

$$\int \frac{1+\sin^2 x}{2+\sin x+\cos x}dx \quad \left[\textit{Answer:} \quad \frac{3}{\sqrt{2}}\arctan\left(\frac{1+\tan\frac{x}{2}}{\sqrt{2}}\right)\right.$$

$$\left. -\frac{1}{2}(\sin x+\cos x)+\ln(2+\sin x+\cos x)+C\right];$$

$$\int \frac{dx}{1+\sin x} \quad \left[\textit{Answer:} \quad \frac{2}{1+\cot\frac{x}{2}}+C\right]$$

$$\int \frac{dx}{(1+\sin x)^2} \quad \left[\textit{Answer:} \quad \frac{3\sin\frac{x}{2}-\cos\frac{3x}{2}}{3\left(\cos\frac{x}{2}+\sin\frac{x}{2}\right)^3}+C\right];$$

$$\int \frac{dx}{1+\cos x} \quad \left[\textit{Answer:} \quad \tan\frac{x}{2}+C\right]$$

$$\int \frac{dx}{(1+\cos x)^2} \quad \left[\textit{Answer:} \quad \frac{(2+\cos x)\sin x}{3(1+\cos x)^2}+C\right];$$

$$\int \frac{1+\cos^2 x}{1+\sin^2 x}dx \quad \left[\textit{Answer:} \quad -x+\frac{3}{\sqrt{2}}\arctan\left(\sqrt{2}\tan x\right)+C\right];$$

$$\int \frac{1+\cos^2 x}{\left(1+\sin^2 x\right)^2}dx \quad \left[\textit{Answer:} \quad \frac{5}{4\sqrt{2}}\arctan\left(\sqrt{2}\tan x\right)\right.$$

$$\left. -\frac{3}{4}\frac{\sin(2x)}{\cos(2x)-3}+C\right].$$

In many cases, as was mentioned, other substitutions may be more convenient. For instance, if the rational function $R(\sin x, \cos x)$ is odd with respect to the cosine function, i.e. $R(\sin x, -\cos x) = -R(\sin x, \cos x)$, then the substitution $t = \sin x$ can be used, for instance:

$$\int \frac{\cos x}{2+\sin x}dx = \left|\begin{matrix} t=\sin x \\ dt=\cos x dx\end{matrix}\right| = \int \frac{dt}{2+t} = \ln|2+t|+C = \ln(2+\sin x)+C.$$

Here $R(\sin x, \cos x) = \cos x/(2+\sin x)$ and is obviously odd with respect to the cosine function. If, instead, the rational function is odd with respect to the sine function, $R(-\sin x, \cos x) = -R(\sin x, \cos x)$, then the substitution $t = \cos x$ is found to be useful:

$$\int \frac{\sin^3 x}{\cos^4 x}dx = \left| \begin{array}{c} t = \cos x \\ dt = -\sin x dx \end{array} \right| = \int \frac{1-t^2}{t^4}(-dt)$$
$$= -\int t^{-4}dt + \int t^{-2}dt = \frac{1}{3}t^{-3} - t^{-1} + C = \frac{1}{3\cos^3 x} - \frac{1}{\cos x} + C.$$

Problem 4.26. *Calculate the following integrals:*

$$\int \sin^2 x \cos^3 x dx \quad \left[\textit{Answer:} \quad \frac{1}{3}\sin^3 x - \frac{1}{5}\sin^5 x + C \right];$$

$$\int \sin^3 x \cos^2 x dx \quad \left[\textit{Answer:} \quad -\frac{1}{3}\cos^3 x + \frac{1}{5}\cos^5 x + C \right].$$

In many other cases usage of trigonometric identities such as those in Eqs. (2.43)–(2.61) may be found useful, especially if the trigonometric functions of different arguments are present. For instance, using the identity (2.52), we can easily take the following integral:

$$\int \sin(11x)\sin(9x)\,dx = \int \frac{1}{2}[\cos(2x) - \cos(20x)]\,dx$$
$$= \frac{1}{4}\sin(2x) - \frac{1}{40}\sin(20x) + C.$$

Sometimes, by reducing the power of the sine or cosine functions via sine and/or cosine functions of a different argument, integrals containing powers of the trigonometric functions can be calculated:

$$\int \cos^2 x dx = \int \frac{1}{2}[1 + \cos(2x)]\,dx = \frac{x}{2} + \frac{1}{4}\sin(2x) + C,$$
$$\int \sin^2 x dx = \int \frac{1}{2}[1 - \cos(2x)]\,dx = \frac{x}{2} - \frac{1}{4}\sin(2x) + C,$$

where we made use of Eq. (2.44). If we use Eq. (2.56), the following integral may be easily calculated:

$$\int \frac{\sin(3x)}{\sin x} dx = \int \frac{3\sin x \cos^2 x - \sin^3 x}{\sin x} dx$$
$$= \int \left(3\cos^2 x - \sin^2 x\right) dx = x + \sin(2x) + C.$$

As a rule, if there are trigonometric functions of different arguments in the integrand, one has to first convert them into trigonometric functions of the same argument. This can always be done if the arguments are of the form nx with n being an integer.

Problem 4.27. *Calculate the following integrals:*

$$\int \sin(6x)\cos(5x)\,dx \quad \left[\textit{Answer:} \quad -\frac{1}{2}\cos x - \frac{1}{22}\cos(11x) + C\right];$$

$$\int \frac{\sin(5x)}{\sin x} dx \quad \left[\textit{Answer:} \quad x + \sin(2x) + \frac{1}{2}\sin(4x) + C\right].$$

4.4.5 Integration of a Rational Function of the Exponential Function

A rational function $R(e^x)$ of the exponential function can also be transformed, using the new variable $t = e^x$, into an integral of the rational function $R_1(t) = R(t)/t$, and hence can always be integrated. For instance, consider:

$$I = \int \frac{1+2e^x}{1-2e^x} dx = \left| \begin{array}{c} t = e^x \\ dt = e^x dx = t dx \end{array} \right| = \int \frac{1+2t}{t(1-2t)} dt.$$

The last integral is calculated, e.g., by decomposing the rational function into elementary functions A/t and $B/(1-2t)$ (with coefficients $A = 1$ and $B = 4$), and then integrating:

$$I = \int \left(\frac{1}{t} + \frac{4}{1-2t}\right) dt = \ln|t| - 2\ln|1-2t| = \ln e^x - 2\ln|1-2e^x|$$
$$= x - \ln(1-2e^x)^2 + C.$$

Problem 4.28. *Calculate the following integrals:*

$$\int \sinh(x)dx \quad [Answer: \ \cosh(x) + C]\,; \quad \int \cosh(x)dx \quad [Answer: \ \sinh(x) + C]\,;$$

$$\int \tanh(x)dx \quad \left[Answer: \ \ln(\cosh x) + C = -x + \ln\left(1 + e^{2x}\right) + C\right];$$

$$\int \coth(x)dx \quad \left[Answer: \ \ln|\sinh x| + C = -x + \ln\left|e^{2x} - 1\right| + C\right];$$

$$\int \tanh^2(x)dx \quad [Answer: \ x - \tanh x + C]\,;$$

$$\int \tanh^3(x)dx \quad \left[Answer: \ \ln(\cosh x) + \frac{1}{2\cosh^2 x} + C\right].$$

4.4.6 Integration of Irrational Functions

Irrational functions, i.e. those which contain radicals, are generally more difficult to integrate. However, there exists a class of irrational functions which, upon an appropriate substitution, can be transformed into an integral with respect to a rational function, and hence are integrable. Consider integrals of a function $R\left(x, y(x)^{1/m}\right) = R\left(x, \sqrt[m]{y(x)}\right)$, which is rational with respect to its arguments x and $\sqrt[m]{y(x)}$ (only the latter argument of course makes the whole function irrational). Here m is an integer and $y(x) = (ax+b)/(gx+f)$ is a simple rational function of x with some constant coefficients a, b, g and f. Note that in the function R there could be an integer power of $y^{1/m}$, i.e. $y(x)$ could be in any rational power n/m (with integer n).

These types of integrals are calculated using the following substitution:

$$t^m = \frac{ax+b}{gx+f}\,, \quad \text{leading to} \quad x = \omega(t) = \frac{ft^m - b}{-gt^m + a}, \tag{4.51}$$

which turns the original integral over x into a t-integral containing a rational function of t. Indeed, in this case,

$$\int R\left(x, \sqrt[m]{\frac{ax+b}{gx+f}}\right)dx = \int R\left(\omega(t), t\right)\omega'(t)dt, \tag{4.52}$$

proving the above made statement as, obviously, the derivative $\omega'(t)$ is a rational function of t and R is rational with respect to both of its arguments.

As an example, consider the following integral where $R(x, z) = (z-1)/(z+1)$, $z = y^{1/2}$ (i.e. $m = 2$) and $y = x + 1$:

$$\int \frac{\sqrt{x+1}-1}{\sqrt{x+1}+1}dx = \left|\begin{array}{c} t^2 = x+1 \\ 2tdt = dx \end{array}\right| = \int \frac{t-1}{t+1}2tdt = \left|\begin{array}{c} y = t+1 \\ dy = dt \end{array}\right|$$
$$= \int \frac{2}{y}(y-1)(y-2)\,dy = 2\int \left(y - 3 + \frac{2}{y}\right)dy = y^2 - 6y + 4\ln|y|$$
$$= x - 4\left(1+\sqrt{x+1}\right) + 4\ln\left(1+\sqrt{x+1}\right) + C,$$

since $y = 1 + t = 1 + \sqrt{x+1}$.

Sometimes, the irrational function in the integrand looks different to the one given above; however, it can still be manipulated into this form as demonstrated by the following example:

$$\int \frac{dx}{\sqrt{(x-1)(x+2)}} = \int \sqrt{\frac{x+2}{x-1}}\frac{dx}{x+2}$$
$$= \left| t^2 = \frac{x+2}{x-1}\,,\ x = \frac{t^2+2}{t^2-1}\,,\ dx = \frac{-6t}{(t^2-1)^2}dt\,,\ x+2 = \frac{3t^2}{t^2-1} \right|$$
$$= \int t\left(\frac{t^2-1}{3t^2}\right)\left(\frac{-6t}{(t^2-1)^2}dt\right) = -2\int \frac{dt}{t^2-1} = \ln\left|\frac{1+t}{1-t}\right|$$
$$= \ln\left|\frac{\sqrt{x-1}+\sqrt{x+2}}{\sqrt{x-1}-\sqrt{x+2}}\right|$$
$$= \ln\left|\frac{\sqrt{x-1}+\sqrt{x+2}}{\sqrt{x-1}-\sqrt{x+2}}\cdot\frac{\sqrt{x-1}+\sqrt{x+2}}{\sqrt{x-1}+\sqrt{x+2}}\right|$$
$$= \ln\left|\frac{\left(\sqrt{x-1}+\sqrt{x+2}\right)^2}{(x-1)-(x+2)}\right|$$
$$= \ln\left|\frac{1+2x+2\sqrt{(x-1)(x+2)}}{-3}\right|$$
$$= \ln\left|1+2x+2\sqrt{(x-1)(x+2)}\right| - \ln 3$$
$$= \ln\left|1+2x+2\sqrt{(x-1)(x+2)}\right| + C,$$

where the constant $\ln 3$ has been dropped as we have to add an arbitrary constant C (introduced at the last step) anyway.

Problem 4.29. *Calculate the following integrals:*

$$\int \frac{dx}{\sqrt[3]{(x+1)(x-1)^2}} \quad \left[\textit{Answer: } \sqrt{3}\arctan\frac{2t+1}{\sqrt{3}} + \frac{1}{2}\ln\frac{1+t+t^2}{(t-1)^2} + C\,, \right.$$

$$\left. \textit{where } t^3 = \frac{x-1}{x+1} \right];$$

$$\int \left(\frac{x-1}{x+1}\right)^{2/3} dx \quad \left[\textit{Answer: } -\frac{2t^2}{t^3-1} + \frac{4}{\sqrt{3}}\arctan\frac{2t+1}{\sqrt{3}} \right.$$

$$\left. -\frac{2}{3}\ln\frac{1+t+t^2}{(t-1)^2} + C\,,\ \textit{where } t^3 = \frac{x-1}{x+1} \right].$$

Another class of integrals which can also be transformed into integrals of rational functions are the ones containing a rational function of x and a square root of a quadratic polynomial, i.e.

$$I = \int R\left(x, \sqrt{ax^2+bx+c}\right) dx. \tag{4.53}$$

Euler studied these types of integrals and proposed general methods for their calculation based on special substitutions which bear Euler's name. The particular substitution to be used depends on the form of the quadratic polynomial under the square root. Three possibilities can be considered, although in practice only two would be needed.

If $a > 0$, then the substitution

$$\sqrt{ax^2+bx+c} = t \pm \sqrt{a}x \tag{4.54}$$

can be used (with either sign).

Problem 4.30. *When applying this substitution, one needs expressions of x via t, dx via dt, and of the square root of the polynomial also via t; moreover, all these expressions should be* rational *functions in t. Demonstrate by the direct calculation that the following identities are valid:*

(continued)

Problem 4.30 (continued)

$$x = \frac{t^2 - c}{b \mp 2\sqrt{a}t}, \quad dx = 2\frac{bt \mp \sqrt{a}t^2 \mp \sqrt{a}c}{\left(b \mp 2\sqrt{a}t\right)^2}dt \quad and$$

$$\sqrt{ax^2 + bx + c} = \frac{bt \mp \sqrt{a}t^2 \mp \sqrt{a}c}{b \mp 2\sqrt{a}t}. \tag{4.55}$$

After the t integral is calculated, one has to replace it with an appropriate expression via x, which follows immediately from Eq. (4.54) as $t = \sqrt{ax^2 + bx + c} \mp \sqrt{a}x$.

Consider the following integral, for which we apply the substitution (4.54) using the minus sign for definiteness:

$$\int \sqrt{x^2 + 2x + 2}dx = \left| \sqrt{x^2 + 2x + 2} = \frac{2t + t^2 + 2}{2(1+t)}, \; dx = 2\frac{2t + t^2 + 2}{4(1+t)^2}dt \right|$$

$$= \int \left(\frac{2t + t^2 + 2}{2(1+t)}\right)\left(\frac{2t + t^2 + 2}{2(1+t)^2}dt\right) = \frac{1}{4}\int \frac{\left(2t + t^2 + 2\right)^2}{(1+t)^3}dt = \left| \begin{array}{c} y = 1 + t \\ dy = dt \end{array} \right|$$

$$= \frac{1}{4}\int \left(y^2 + 1\right)^2 \frac{dy}{y^3}$$

$$= \frac{1}{4}\int \left(y + \frac{2}{y} + \frac{1}{y^3}\right)dy = \frac{y^2}{8} - \frac{1}{8y^2} + \frac{1}{2}\ln|y| + C,$$

where $y = 1 + t = \sqrt{x^2 + 2x + 2} + 1 + x$.

Problem 4.31. *Using this method, calculate the following integrals:*

$$\int \frac{dx}{\sqrt{x^2 + 2x + 2}} \quad \left[\textit{Answer:} \quad \ln\left|1 + x + \sqrt{x^2 + 2x + 2}\right| + C\right];$$

$$\int \frac{xdx}{\sqrt{x^2 + 2x + 2}} \quad \left[\textit{Answer:} \quad \sqrt{x^2 + 2x + 2} \right.$$

$$\left. - \ln\left|1 + x + \sqrt{x^2 + 2x + 2}\right| + C\right].$$

If the constant c in the quadratic polynomial is positive, Euler proposed another substitution which also leads to a rational function integral:

$$\sqrt{ax^2+bx+c} = xt \pm \sqrt{c} \tag{4.56}$$

Problem 4.32. *Prove that in this case the required transformations are given by:*

$$x = \frac{b \mp 2\sqrt{c}t}{t^2-a}\ , \quad dx = 2\frac{-bt \pm \sqrt{c}t^2 \pm \sqrt{c}a}{\left(t^2-a\right)^2}dt \quad \textit{and}$$
$$\sqrt{ax^2+bx+c} = \frac{bt \mp \sqrt{c}t^2 \mp \sqrt{c}a}{t^2-a}. \tag{4.57}$$

and $t = \left(\sqrt{ax^2+bx+c} \mp \sqrt{c}\right)/x$.

Finally, the third Euler's substitution can be applied using the roots, x_1 and x_2, of the quadratic polynomial (we only consider here the case when these roots are real):

$$ax^2+bx+c = a\left(x-x_1\right)\left(x-x_2\right).$$

In this case the following substitution is applied:

$$\sqrt{ax^2+bx+c} = t\left(x-x_1\right) \tag{4.58}$$

(the other root may be used as well).

Problem 4.33. *Prove that in this case the required transformations are as follows:*

$$x = \frac{t^2x_1-ax_2}{t^2-a},\ \ dx = \frac{2a\left(x_2-x_1\right)t}{\left(t^2-a\right)^2}dt \ \textit{ and } \ \sqrt{ax^2+bx+c} = \frac{a\left(x_1-x_2\right)t}{t^2-a}, \tag{4.59}$$

and $t = \sqrt{ax^2+bx+c}/\left(x-x_1\right)$.

Note that if $x_1 = x_2$ (repeated roots), then $\sqrt{ax^2+bx+c} = \sqrt{a}\left|x-x_1\right|$ (a positive value of the square root is to be assumed) and the square root disappears, i.e. the integrand is a rational function without any additional transformations. Also note that the third Euler's substitution is equivalent to the one we used at the beginning of this subsection since one can always write (as we did in one of the examples above):

$$\sqrt{ax^2+bx+c} = \sqrt{a(x-x_1)(x-x_2)} = \sqrt{a(x-x_1)(x-x_2)}\sqrt{\frac{x-x_2}{x-x_2}}$$
$$= \sqrt{a}\,(x-x_2)\sqrt{\frac{x-x_1}{x-x_2}},$$

i.e. the substitution of the type (4.51) with $m = 2$ can be used which will bring the integrand into a rational form with respect to the variable $t = \sqrt{(x-x_1)/(x-x_2)}$. Finally, it is easy to see that Euler's second substitution is in fact redundant. Indeed, if $a > 0$, then one can always use the first substitution, even when $c > 0$. If, however, $a < 0$, then the parabola corresponding to the square polynomial is either located entirely below the x axis (or just touches it at a single point) or its top is above it. In the former case the polynomial is everywhere negative (except may be at a single point), and this case can be disregarded within the manifold of real numbers we are interested in so far. In the latter case there must be two real roots x_1 and x_2, in which case the third substitution becomes legitimate. In the case of the two roots x_1 and x_2 being complex, the third Euler's substitution can always be used.

The consideration above leads us to a conclusion that all integrals with the function $R\left(x, \sqrt{ax^2+bx+c}\right)$ can be transformed into a rational form and hence integrated in elementary functions.

In some rather simple cases the Euler's method is not required, as much simpler substitutions exist. For instance, transforming to the complete square and then performing a straightforward change of the variable may suffice as illustrated in the example we have already considered above using a different method:

$$\int \frac{dx}{\sqrt{x^2+2x+2}} = \int \frac{dx}{\sqrt{(x+1)^2+1}} = \left| \begin{array}{c} t = x+1 \\ dt = dx \end{array} \right| = \int \frac{dt}{\sqrt{1+t^2}}$$
$$= \ln\left|t+\sqrt{1+t^2}\right| = \ln\left|1+x+\sqrt{x^2+2x+2}\right| + C,$$

which is the same result as in Problem 4.31.

Problem 4.34. *Using this method, calculate the following integral:*

$$\int \frac{(2x-1)\,dx}{\sqrt{x^2+2x+2}} \qquad \left[\textit{Answer:}\quad 2\sqrt{x^2+2x+2} - 3\ln\left|1+x+\sqrt{x^2+2x+2}\right| + C\right].$$

Finally, integrals of the functions $R\left(x, \sqrt{a^2 - x^2}\right)$, $R\left(x, \sqrt{x^2 - a^2}\right)$ and $R\left(x, \sqrt{x^2 + a^2}\right)$ can be transformed into integrals containing rational functions of the trigonometric functions using the so-called trigonometric substitutions. Of course, these particular integrals can always be taken using the Euler's substitutions, but simpler trigonometric substitutions may result in less cumbersome algebra. Indeed, in the case of $\sqrt{a^2 - x^2}$ one uses $x = a \sin t$ (or $x = a \cos t$), e.g.

$$\int \sqrt{a^2 - x^2} dx = \left| \begin{matrix} x = a \sin t \ , \ \cos t = \sqrt{1 - (x/a)^2} \\ dx = a \cos t dt \end{matrix} \right| = a^2 \int \sqrt{1 - \sin^2 t} \cos t dt$$

$$= a^2 \int \cos^2 t dt = a^2 \left[\frac{t}{2} + \frac{1}{4} \sin(2t) \right] = \frac{a^2}{2} \left[t + \sin t \cos t \right]$$

$$= \frac{a^2}{2} \left[\arcsin \frac{x}{a} + \frac{x}{a} \sqrt{1 - \left(\frac{x}{a}\right)^2} \right] = \frac{a^2}{2} \arcsin \frac{x}{a} + \frac{x}{2} \sqrt{a^2 - x^2} + C. \quad (4.60)$$

Note that it is legitimate here to choose the positive value of the cosine function during intermediate manipulations since $-a \leq x \leq a$ and hence one can choose $-\pi/2 \leq t \leq \pi/2$ when $\cos t \geq 0$.

In the case of $\sqrt{x^2 - a^2}$, one similarly uses $x = a/ \sin t$, while in the case of $\sqrt{x^2 + a^2}$ the tangent substitution, $x = a \tan t$, can be used instead since $1 + \tan^2 t = 1/ \cos^2 t$ and hence the square root disappears.

Problem 4.35. *Calculate the following integral:*

$$\int \sqrt{x^2 - a^2} dx \quad \left[\textit{Answer:} \quad \frac{x}{2} \sqrt{x^2 - a^2} - \frac{a^2}{2} \ln \left| x + \sqrt{x^2 - a^2} \right| + C \right].$$

Problem 4.36. *Calculate the integral:*

$$\int \sqrt{x^2 + a^2} dx \quad \left[\textit{Answer:} \quad \frac{x}{2} \sqrt{x^2 + a^2} + \frac{a^2}{2} \ln \left| x + \sqrt{x^2 + a^2} \right| + C \right]. \quad (4.61)$$

Problem 4.37. *Show that*

$$\int \frac{dx}{\sqrt{x^2 - 1}} = \ln \left| x + \sqrt{x^2 - 1} \right| + C = \cosh^{-1}(x) + C. \quad (4.62)$$

4.5 More on Calculation of Definite Integrals

As we have already discussed in Sect. 4.3, it is easy to calculate a definite integral if the corresponding indefinite integral is known, and many methods to help with this task have been reviewed in the previous Sect. 4.4. Still, more needs to be said about some subtle points and extensions, which we are going to do in this section.

4.5.1 *Change of Variables and Integration by Parts in Definite Integrals*

We learned two important techniques of calculating indefinite integrals in Sect. 4.4, a change in the integration variable and integration by parts, and of course these can be used for calculating definite integrals as well. However, there are some points we must discuss, otherwise wrong results may be obtained.

Change of Variables. Consider a definite integral $\int_a^b f(x)dx$. As we know from studying indefinite integrals, if a substitution $x = x(t)$ is used, then an integral with respect to the variable t is obtained, and, if we are lucky, the t-integral is calculated, the t is replaced back (by solving the equation $x = x(t)$ with respect to t), and that is it. Two comments are in order here as far as definite integrals are concerned.

Firstly, when a definite integral is calculated using a substitution, it is easier to work out the limits of the new variable t alongside usual steps of expressing dx via dt and x via t everywhere in the integrand. The benefit of this is that one can avoid calculating t via x and hence replacing t back with the x; instead, the boundary values of t can be immediately used after the t-integral is calculated to get the required result (which is a number). For instance,

$$\int_0^a xe^{-x^2}dx = \left| \begin{array}{c} t = x^2 \ , \ dt = 2xdx \ , \ xdx = dt/2 \\ t = 0 \text{ when } x = 0 \ ; \ t = a^2 \text{ when } x = a \end{array} \right| = \frac{1}{2}\int_0^{a^2} e^{-t}dt$$
$$= \frac{1}{2}\left(-e^{-t}\right)\Big|_0^{a^2} = \frac{1}{2}\left(-e^{-a^2}+1\right).$$

Note that we replaced the limits in the t-integral when changing the variables and then used those limits directly without going back to the original x variable.

Secondly, one has to make sure that the substitution $x = x(t)$ is single-valued, continuous and its derivative $x'(t)$ is also continuous. If these conditions are not satisfied within the whole integration interval $a \leq x \leq b$, the result of the substitution could be wrong. The following examples illustrate this subtle point.

Consider the integral ($a > 0$):

$$I_1 = \int_{-a}^{a} xdx = \frac{x^2}{2}\Big|_{-a}^{a} = 0.$$

Of course, this integral is zero because the integrand is an odd function and the limits of integration form a symmetric interval around $x = 0$, see also Eq. (4.8), i.e. the above obtained result makes a lot of sense. On the other hand, let us make the substitution defined by the equation $x^2 = t$:

$$I_1 = \left| \begin{array}{c} x^2 = t \ , \ x = \sqrt{t} \\ dx = dt/2\sqrt{t} \end{array} \right| = \int_{a^2}^{a^2} \sqrt{t} \frac{dt}{2\sqrt{t}} = \frac{1}{2} \int_{a^2}^{a^2} dt = 0,$$

since the two t-limits are now the same and equal to a^2. The same result as before has been obtained; good! However, let us now consider the integral of x^2 instead:

$$I_2 = \int_{-a}^{a} x^2 dx = \left. \frac{x^3}{3} \right|_{-a}^{a} = \frac{2}{3} a^3,$$

while if we used the substitution $t = x^2$, the result would have been zero as the two limits will be the same again! This paradoxical result originates from the fact that the substitution we used defines a multivalued function $t = \pm\sqrt{x}$, and this contradicts one of the necessary conditions needed for the substitution to work. In fact, when integrating between $-a$ and 0, when x is negative, we must have used $x = -\sqrt{t}$, while $x = \sqrt{t}$ must be used for positive x. The same is true for I_1; the correct result we obtained for I_1 was accidental. That means that the integration interval must be split into two and the correct single valued substitution is to be used in each interval individually.

Another caution needs to be given concerning the requirement that both the function $x(t)$ and its derivative $x'(t)$ are to be continuous within the integration interval.

The above points are illustrated by calculating the following integral:

$$I_3 = \int_{-\pi}^{\pi} dx = x|_{-\pi}^{\pi} = 2\pi.$$

Alternatively, let us make a substitution $t = \tan x$. When $x = -\pi$ we have $t = 0$; however, when $x = \pi$, the same result $t = 0$ is obtained. Therefore, inevitably, the integral becomes zero after the substitution is made:

$$I_3 = \left| \begin{array}{c} t = \tan x \\ dt = dx/\cos^2 x = \left(1 + t^2\right) dx \end{array} \right| = \int_0^0 \frac{dt}{1 + t^2} = 0.$$

This result is obviously wrong, and the reason is related to the fact that the function $t = \tan x$ is not continuous over the whole range of the x values between $-\pi$ and π, and experiences two discontinuities at $x = \pm\pi/2$. Therefore, formula (4.43) is not longer valid: one has to split the x-integral into three with intervals: $-\pi \leq x < -\pi/2$, $-\pi/2 < x < \pi/2$ and $\pi/2 < x \leq \pi$ (the points $\pm\pi/2$ have to be excluded from integration, see Sect. 4.5.3 for more details). Then the correct result is obtained:

$$\int_{-\pi}^{-\pi/2} dx = \int_0^{\infty} \frac{dt}{1+t^2} = \arctan x|_0^{\infty} = \frac{\pi}{2},$$

$$\int_{-\pi/2}^{\pi/2} dx = \int_{-\infty}^{\infty} \frac{dt}{1+t^2} = \arctan x|_{-\infty}^{\infty} = \pi,$$

$$\int_{\pi/2}^{\pi} dx = \int_{-\infty}^{0} \frac{dt}{1+t^2} = \arctan x|_{-\infty}^{0} = \frac{\pi}{2},$$

yielding the correct result of 2π. The subtle points here are that in the first integral the upper limit corresponds to $x = -\frac{\pi}{2} - 0$ (as $x < -\frac{\pi}{2}$) which results in $t = \tan(-\pi/2 - 0) = +\infty$ for the upper limit of the t-integral; similarly, in the second integral the bottom limit is $-\frac{\pi}{2} + 0$ and the upper is $\frac{\pi}{2} - 0$, while in the third integral the bottom limit is $\frac{\pi}{2} + 0$.

The above examples indicate that care is needed when applying a substitution; the latter should satisfy a number of conditions, and must satisfy them along the whole integration interval as otherwise the result may be wrong!

Integration by Parts. This is another important method we found for the indefinite integrals in Sect. 4.4.2. Of course, it can also be used directly when calculating definite integrals, e.g.

$$\int_{-1}^{1} x\cos x dx = \left| \begin{array}{c} u = x\ ,\ du = dx \\ dv = \cos x dx\ ,\ v = \sin x \end{array} \right| = x\sin x|_{-1}^{1} - \int_{-1}^{1} \sin x dx$$
$$= [\sin 1 - (-1)\sin(-1)] - (-\cos x)|_{-1}^{1} = 0,$$

which is to be expected as the function $x\cos x$ is odd and the interval is symmetric.

The method of integration by parts may be found especially useful in working out general expressions for various definite integrals containing integer n as a parameter. Consider, for instance, the following integral:

$$I_n = \int_0^{\pi/2} \cos^n x dx, \tag{4.63}$$

where $n = 0, 1, 2, \ldots$ Integrating it by parts, we can obtain a recurrence relation connecting I_n with the integral associated with a smaller value of n:

$$\begin{aligned} I_n &= \int_0^{\pi/2} \cos^{n-1} x \cos x dx = \left| \begin{array}{c} u = \cos^{n-1} x\ ,\ du = (n-1)\cos^{n-2} x\,(-\sin x)\, dx \\ dv = \cos x dx\ ,\ v = \sin x \end{array} \right| \\ &= \underbrace{\cos^{n-1} x \sin x\Big|_0^{\pi/2}}_{=0} + (n-1)\int_0^{\pi/2} \cos^{n-2} x \sin^2 x dx \\ &= (n-1)\int_0^{\pi/2} \cos^{n-2} x \left(1 - \cos^2 x\right) dx \\ &= (n-1)\, I_{n-2} - (n-1)\, I_n, \end{aligned}$$

which gives an equation for I_n, namely:

$$I_n = \frac{n-1}{n} I_{n-2}.$$

Since $I_0 = \int_0^{\pi/2} dx = \pi/2$, we obtain, for instance: $I_2 = (2-1)\, I_0/2 = \pi/4$, $I_4 = (4-1)\, I_2/4 = 3\pi/16$, etc.

For odd values of n, our recurrence relation is valid as well; however, the first (the smallest) value of n must be $n = 1$ instead for which

$$I_1 = \int_0^{\pi/2} \cos x dx = \sin x|_0^{\pi/2} = 1,$$

and hence $I_3 = (3-1)\, I_1/3 = 2/3$, $I_5 = (5-1)\, I_3/5 = 8/15$, and so on.

Problem 4.38. *Prove that for even $n = 2p$,*

$$I_{2p} = \frac{\pi}{4^p} \frac{(2p-1)!}{p!\,(p-1)!} = \frac{\pi}{4^p} \binom{2p-1}{p},$$

while for odd values of $n = 2p + 1$,

$$I_{2p+1} = \frac{4^p\,(p!)^2}{(2p+1)!}.$$

Problem 4.39. *Show that the same recurrence relation is obtained for the integral*

$$J_n = \int_0^{\pi/2} \sin^n x dx,$$

and that for both even and odd values of n the same results are obtained as for I_n.

4.5.2 Integrals Depending on a Parameter

In many applications either the limits of a definite integral $a(\lambda)$ and $b(\lambda)$ and/or the integrand $f(x, \lambda)$ itself may depend on some parameter, let us call it λ,

$$I(\lambda) = \int_{a(\lambda)}^{b(\lambda)} f(x; \lambda) dx, \tag{4.64}$$

and it is required to calculate the derivative of the value of the integral, which is thus a function of λ, with respect to it. We know from Sect. 4.3 that if only the upper limit depends on λ as $b = \lambda$, one obtains $I'(\lambda) = f(\lambda)$, i.e. the value of the integrand at $x = \lambda$. Here we shall derive the most general formula for the λ-derivative of the integral given by Eq. (4.64) where both limits and the integrand itself may depend on λ in a general way.

We start by writing the corresponding definition of the derivative:

$$I'(\lambda) = \lim_{\Delta\lambda \to 0} \frac{I(\lambda + \Delta\lambda) - I(\lambda)}{\Delta\lambda} = \lim_{\Delta\lambda \to 0} \frac{\Delta I}{\Delta\lambda},$$

where

$$\begin{aligned}
\Delta I &= \int_{a+\Delta a}^{b+\Delta b} f(x;\lambda + \Delta\lambda)dx - \int_a^b f(x;\lambda)dx \\
&= \int_{a+\Delta a}^{a} f(x;\lambda + \Delta\lambda)dx + \int_a^b f(x;\lambda + \Delta\lambda)dx \\
&\quad + \int_b^{b+\Delta b} f(x;\lambda + \Delta\lambda)dx - \int_a^b f(x;\lambda)dx \\
&= \int_a^b \left[f(x;\lambda + \Delta\lambda) - f(x;\lambda)\right] dx - \int_a^{a+\Delta a} f(x;\lambda + \Delta\lambda)dx \\
&\quad + \int_b^{b+\Delta b} f(x;\lambda + \Delta\lambda)dx,
\end{aligned}$$

with $\Delta a = a'(x)\Delta\lambda$ and $\Delta b = b'(x)\Delta\lambda$. According to our discussion in Sect. 4.3 (see specifically Eq. (4.33)), we can write for the second and the third integrals:

$$\begin{aligned}
\int_a^{a+\Delta a} f(x;\lambda + \Delta\lambda)dx &= f(a_1;\lambda + \Delta\lambda)\Delta a = \left[f(a_1;\lambda) + f'_\lambda(a_1;\xi)\Delta\lambda\right] a'(\lambda)\Delta\lambda, \\
\int_b^{b+\Delta b} f(x;\lambda + \Delta\lambda)dx &= f(b_1;\lambda + \Delta\lambda)\Delta b = \left[f(b_1;\lambda) + f'_\lambda(b_1;\zeta)\Delta\lambda\right] b'(\lambda)\Delta\lambda,
\end{aligned}$$

where $a < a_1 < a + \Delta a$, $b < b_1 < b + \Delta b$ and in the second passage in both lines we have made use of the Lagrange formula (3.64) with respect to the variable λ:

$$\begin{aligned}
f(a_1;\lambda + \Delta\lambda) &= f(a_1;\lambda) + f'_\lambda(a_1;\xi)\Delta\lambda \quad \text{and} \\
f(b_1;\lambda + \Delta\lambda) &= f(b_1;\lambda) + f'_\lambda(b_1;\zeta)\Delta\lambda
\end{aligned}$$

with $\lambda < \xi < \lambda + \Delta\lambda$ and $\lambda < \zeta < \lambda + \Delta\lambda$. Here f'_λ denotes the derivative of the function $f(x;\lambda)$ with respect to its second variable, λ.[3] Combining all expressions for the integrals into ΔI and dividing it by $\Delta\lambda$, we obtain:

$$\frac{\Delta I}{\Delta\lambda} = \int_a^b \frac{f(x;\lambda+\Delta\lambda) - f(x;\lambda)}{\Delta\lambda} dx - \left[f(a_1;\lambda) + f'_\lambda(a_1;\xi)\Delta\lambda\right] a'(\lambda)$$
$$+ \left[f(b_1;\lambda) + f'_\lambda(b_1;\zeta)\Delta\lambda\right] b'(\lambda).$$

Now, when taking the limit $\Delta\lambda \to 0$, the points $a_1 \to a$ and $b_1 \to b$, this finally yields[4]:

$$\frac{d}{d\lambda}\int_{a(\lambda)}^{b(\lambda)} f(x;\lambda)dx = \int_a^b \frac{df(x;\lambda)}{d\lambda}dx + f\,(b;\lambda)\,b'(\lambda) - f\,(a;\lambda)\,a'(\lambda). \tag{4.65}$$

It is seen that this result indeed generalises the one we obtained earlier in Sect. 4.3: if a and f do not depend on λ and $b = \lambda$, formula (4.34) is recovered.

Sometimes, by introducing a parameter artificially, the integral of interest may be related to another; if the expression for the latter integral is known, this trick allows us to calculate the former. Suppose we need to calculate the following integral

$$I = \int_0^{\pi/2} x^3 \sin(\omega x)\, dx.$$

Of course, this can be calculated by doing three integrations by parts; instead, we shall consider a function

$$J(\omega) = \int_0^{\pi/2} \cos(\omega x)\, dx = \frac{1}{\omega}\sin\left(\omega\frac{\pi}{2}\right),$$

and notice that upon differentiation of the cosine function $\cos(\omega x)$ in $J(\omega)$ three times with respect to ω, we get $x^3\sin(\omega x)$, which is exactly the required integrand of I. Therefore, we can write:

$$I = \frac{d^3}{d\omega^3}J(\omega) = \frac{d^3}{d\omega^3}\left[\frac{1}{\omega}\sin\left(\omega\frac{\pi}{2}\right)\right]$$
$$= \left(\frac{1}{\omega}\right)^{(3)}\sin\left(\omega\frac{\pi}{2}\right) + 3\left(\frac{1}{\omega}\right)^{(2)}\left[\sin\left(\omega\frac{\pi}{2}\right)\right]^{(1)}$$
$$+3\left(\frac{1}{\omega}\right)^{(1)}\left[\sin\left(\omega\frac{\pi}{2}\right)\right]^{(2)} + \frac{1}{\omega}\left[\sin\left(\omega\frac{\pi}{2}\right)\right]^{(3)}$$

[3] We shall learn rather soon in Sect. 5.3 when considering functions of many variables that this is called a partial derivative with respect to λ of the function $f(x,\lambda)$ of two variables.

[4] We shall see later on that the derivative of f with respect to λ in Eq. (4.65) should really be denoted as $\partial f/\partial\lambda$ using the symbol ∂ rather than d, since it is actually a partial derivative of f with respect to its other variable λ as was already mentioned.

$$= \frac{-6}{\omega^4}\sin\left(\omega\frac{\pi}{2}\right) + \frac{6}{\omega^3}\frac{\pi}{2}\cos\left(\omega\frac{\pi}{2}\right) + \frac{3}{\omega^2}\left(\frac{\pi}{2}\right)^2\sin\left(\omega\frac{\pi}{2}\right)$$
$$-\frac{1}{\omega}\left(\frac{\pi}{2}\right)^3\cos\left(\omega\frac{\pi}{2}\right),$$

which is the required expression (which can still be simplified further if desired). When performing differentiation, we have conveniently used here the Leibnitz formula (3.47).

Problem 4.40. *The Newton's equation of motion of a Brownian particle of mass m moving in a liquid can be written as the following Langevin equation:*

$$m\frac{dv}{dt} = f - \int_0^t K(t-\tau)\, v(\tau)d\tau,$$

where f is the force due to liquid molecules hitting the particle at random, $v(t)$ particle velocity, and $K(t) = A\exp(-\alpha t)$, the so-called friction kernel which we assume decays exponentially with time. Note that the integral term stands up for its name of the friction force since it is proportional to the particle velocity and comes with the minus sign. The integral here means that the particle motion depends on all its previous velocities, i.e. the memory of its motion is essential to correctly describe its dynamics. This may be the case, e.g., if the environment relaxation time is of the same order or longer than the characteristic time of the particle. Now, denote the integral term as a function $y(t)$ and show that the single Langevin equation above is equivalent to the following two equations:

$$m\frac{dv}{dt} = f - y \quad \text{and} \quad \frac{dy}{dt} = Av - \alpha y.$$

4.5.3 Improper Integrals

So far our definite integrals were taken between finite limits a and b; moreover, we assumed that the integrated function $f(x)$ does not have singularities anywhere within the integration interval $a \le x \le b$, including the boundary points themselves. We can now define more general integrals in which these conditions are relaxed.

We shall start by considering the upper limit being $+\infty$. To this end, let us consider an integral

$$I(b) = \int_a^b f(x)dx,$$

which we assume exists for any finite b. This integral can be considered as a function, $I(b)$, with respect to its upper limit b. If we now consider larger and larger values of the upper limit b, then in the limit of $b \to +\infty$ we shall arrive at the so-called *improper integral* with the upper limit being $+\infty$:

$$\int_a^{\infty} f(x)dx = \lim_{b\to\infty} \int_a^b f(x)dx. \tag{4.66}$$

Let us state the definition of its convergence. The above improper integral $\int_a^{\infty} f(x)dx$ is said to converge to a number F, if for any $\epsilon > 0$ there always exists such $A > a$, that

$$\left| F - \int_a^A f(x)dx \right| < \epsilon.$$

This means that the smaller the value of ϵ we take, the larger the value of the upper limit A need to be taken to have the integral $\int_a^A f(x)dx$ be close (within ϵ) to F. This condition can also be formally rewritten as

$$\left| \int_A^{\infty} f(x)dx \right| < \epsilon,$$

stating that the integral $\int_A^{\infty} f(x)dx$ tends to zero as $A \to \infty$.

Similarly, one defines the improper integrals with the bottom limit being $-\infty$,

$$\int_{-\infty}^b f(x)dx = \lim_{a\to-\infty} \int_a^b f(x)dx, \tag{4.67}$$

or when both limits are at $\pm\infty$:

$$\int_{-\infty}^{\infty} f(x)dx = \lim_{b\to\infty} \left(\lim_{a\to-\infty} \int_a^b f(x)dx \right). \tag{4.68}$$

Problem 4.41. *Give the definition of the improper integral with the bottom limit being* $-\infty$.

In the case of Eq. (4.68) both limits should exist (taken in any order and independently of each other). In each case the integrals may either converge (the limits exist) or diverge (are infinite). The geometrical interpretation of the improper integrals is exactly the same as for proper ones: if $f(x)$ is a non-negative function, then any of the above improper integrals corresponds to the area under the curve of the function within the corresponding intervals $a \le x < +\infty$, $-\infty < x \le b$ or $-\infty < x < +\infty$, respectively. If any of the integrals converge, then the area under the curve is finite.

When calculating the improper integrals one can still use the Newton–Leibnitz formula (4.43). Indeed, if $F'(x) = f(x)$, then, for instance,

$$\int_a^\infty f(x)dx = \lim_{b\to\infty}\int_a^b f(x)dx = \lim_{b\to\infty}[F(b)-F(a)]$$
$$= \lim_{b\to\infty} F(b) - F(a) = F(+\infty) - F(a), \tag{4.69}$$

and

$$\int_{-\infty}^\infty f(x)dx = \lim_{b\to\infty}\left(\lim_{a\to-\infty}\int_a^b f(x)dx\right) = \lim_{b\to\infty}\left[\lim_{a\to-\infty}(F(b)-F(a))\right]$$
$$= \lim_{b\to\infty} F(b) - \lim_{a\to-\infty} F(a) = F(+\infty) - F(-\infty). \tag{4.70}$$

Let us consider some examples of these types of improper integrals. As our first example let us calculate the integral which has a finite upper limit:

$$I(b) = \int_0^b e^{-x}dx = -e^{-x}|_0^b = 1 - e^{-b}.$$

Integration of the same function between 0 and $+\infty$, according to the definition given above, requires us to take the $b \to \infty$ limit. Therefore,

$$\int_0^\infty e^{-x}dx = \lim_{b\to\infty}\left(1 - e^{-b}\right) = 1.$$

As the limit exists (i.e. it is finite), it is to be taken as the value of the integral.

As another example, consider the integral

$$I_\lambda = \int_1^\infty \frac{dx}{x^\lambda}, \tag{4.71}$$

where λ is positive. Let us first investigate the case of $\lambda > 1$ (strictly larger than one!). In this case

$$I_{\lambda>1} = \lim_{b\to\infty}\int_1^b \frac{dx}{x^\lambda} = \lim_{b\to\infty}\left.\frac{x^{-\lambda+1}}{-\lambda+1}\right|_1^b = \frac{1}{1-\lambda}\lim_{b\to\infty}\left(\frac{1}{b^{\lambda-1}} - 1\right)$$
$$= -\frac{1}{1-\lambda} = \frac{1}{\lambda-1},$$

since the power $\lambda - 1$ of b in $1/b^{\lambda-1}$ is positive and hence this term tends to zero in the limit. In the case of $\lambda = 1$, we similarly write:

$$I_{\lambda=1} = \lim_{b\to\infty}\int_1^b \frac{dx}{x} = \lim_{b\to\infty}\ln x|_1^b = \lim_{b\to\infty}(\ln b - \ln 1) = \lim_{b\to\infty}\ln b = \infty,$$

i.e. in this case the integral diverges. Finally, if $0 < \lambda < 1$, then

$$I_{\lambda<1} = \lim_{b\to\infty}\int_1^b \frac{dx}{x^\lambda} = \lim_{b\to\infty} \left.\frac{x^{-\lambda+1}}{-\lambda+1}\right|_1^b = \frac{1}{1-\lambda}\lim_{b\to\infty}\left(b^{1-\lambda}-1\right) = \infty,$$

and the integral diverges again. We conclude that the integral (4.71) only converges for strictly $\lambda > 1$.

Let us now investigate whether the following integral exists:

$$\int_0^\infty \sin x dx = \lim_{b\to\infty} \left(-\cos x\right)\Big|_0^b = \lim_{b\to\infty}\left(1-\cos b\right) = 1 - \lim_{b\to\infty}\cos b.$$

Since the cosine function oscillates and hence does not have a definite limit as $b \to \infty$ (although its value is bounded from above and below), the integral in question does not converge.

The above examples tell us that an improper integral can be calculated in exactly the same way as the corresponding proper ones, and the value of the integral is still given as the difference of the function $F(x)$ between the upper and lower limits. The only difference is in working out the value(s) of $F(\pm\infty)$, which requires taking the appropriate limit(s).

Problem 4.42. *Check the values of the following improper integrals using their appropriate definitions:*

$$\int_0^\infty e^{-x}\sin x dx = \frac{1}{2}\ ; \quad \int_0^\infty \frac{dx}{1+x^2} = \frac{\pi}{2}\ ;$$

$$\int_{-\infty}^0 e^x dx = 1\ ; \int_{1/\pi}^\infty \frac{1}{x^2}\sin\frac{1}{x}dx = 2.$$

Problem 4.43. *Using integration by parts, derive a recurrence relation for the following integral*[5]*:*

$$\Gamma(n) = \int_0^\infty x^{n-1}e^{-x}dx.$$

Show by direct calculation for several values of $n = 1, 2, 3$ *and then using induction that* $\Gamma(n+1) = n!$.

[5]This is a particular case of one of the special functions, the so-called Gamma function.

It is also possible to define an integral of a function which has a singularity[6] either at the bottom and/or top limits, and/or between the limits. If a function $f(x)$ is singular at the top limit b, i.e. $f(b) = \pm\infty$, but is finite just before it, we define a positive ϵ and define the integral via the limit as:

$$\int_a^b f(x)dx = \lim_{\epsilon \to 0} \int_a^{b-\epsilon} f(x)dx. \tag{4.72}$$

Here the function is finite for $a \leq x \leq b - \epsilon$ for any $\epsilon > 0$, and hence the integral before the limit is well defined. If the limit exists, then the integral converges. Similarly, if the singularity of $f(x)$ coincides with the bottom limit, $f(a) = \pm\infty$, then we consider

$$\int_a^b f(x)dx = \lim_{\epsilon \to 0} \int_{a+\epsilon}^{b} f(x)dx. \tag{4.73}$$

Finally, if the singularity occurs at a point c somewhere between a and b, one breaks the whole integration interval into two subintervals $a \leq x < c$ and $c < x \leq b$, so that the initial integral is split into two, and then the above definitions are applied to each of the integrals separately:

$$\int_a^b f(x)dx = \lim_{\epsilon_1 \to 0} \int_a^{c-\epsilon_1} f(x)dx + \lim_{\epsilon_2 \to 0} \int_{c+\epsilon_2}^{b} f(x)dx. \tag{4.74}$$

If both limits exist, the integral in the left-hand side is said to converge (or to exist). It is essential here that both limits are taken independently; this is indicated by using $\epsilon_1 \neq \epsilon_2$ in the two integrals.[7] This type of reasoning also explains why changing the value of the function $f(x)$ at a single isolated point $x = c$ will not change the value of the integral. Indeed, in this case the integral can be defined by virtue of Eq. (4.74); clearly, the value of the function $f(x)$ at $x = c$ does not appear in this case at all; it could in fact be different from either of the two limits $\lim_{x \to c \pm 0} f(x)$.

As an example of a singularity inside the integration interval, consider the function $f(x) = 1/x^\lambda$ with positive λ. This function is singular at $x = 0$, so that whether it is integratable around this point needs to be specifically investigated. Firstly, let $\lambda > 1$ and consider the integral

$$I_{\lambda>1} = \int_0^1 \frac{dx}{x^\lambda} = \lim_{\epsilon \to 0} \int_{0+\epsilon}^1 \frac{dx}{x^\lambda} = \lim_{\epsilon \to 0} \frac{1}{1-\lambda}\left[1 - \frac{1}{\epsilon^{\lambda-1}}\right].$$

[6]That is the function is equal to $\pm\infty$ at an isolated point between $-\infty$ and $+\infty$.

[7]Compare with Sect. 4.5.4.

Since $\lambda - 1 > 0$, the limit of $1/\epsilon^{\lambda-1}$ is equal to infinity, i.e. the integral does not converge for any $\lambda > 1$. In the case of $\lambda = 1$, we have

$$I_{\lambda=1} = \int_0^1 \frac{dx}{x} = \lim_{\epsilon\to 0}\int_\epsilon^1 \frac{dx}{x} = \lim_{\epsilon\to 0}(\ln 1 - \ln\epsilon) = -\lim_{\epsilon\to 0}\ln\epsilon = \infty,$$

i.e. the limit is not finite again, and hence the integral diverges[8] for $\lambda = 1$. Finally, consider the case of $0 < \lambda < 1$. In this case

$$I_{\lambda<1} = \int_0^1 \frac{dx}{x^\lambda} = \lim_{\epsilon\to 0}\int_\epsilon^1 \frac{dx}{x^\lambda} = \lim_{\epsilon\to 0}\frac{1}{1-\lambda}\left[1-\epsilon^{1-\lambda}\right] = \frac{1}{1-\lambda},$$

since $1 - \lambda > 0$ and hence the limit of $\epsilon^{1-\lambda}$ is well defined and equal to zero. So, the above integral only converges for $0 < \lambda < 1$.

Problem 4.44. *Check whether the following integrals converge and then calculate them if they do:*

$$\int_0^1 \frac{dx}{\sqrt{1-x^2}}\ ; \quad \int_{-1}^1 \frac{dx}{\sqrt{1-x^2}}\ ; \quad \int_0^2 \ln x dx\ ; \quad \int_1^2 \frac{1}{x\ln x}dx.$$

[*Hint*: use the substitution $t = \ln x$ in the last two cases. *Answers*: $\pi/2$, π, $2(\ln 2 - 1)$, diverges at $x = 1$.]

Problem 4.45. *Consider a function $f(x)$ on the interval $a \leq x \leq b$ which has singularities in a finite number of points $c_1, c_2, \dots, c_n$ between a and b. If the integral of $f(x)$ between these boundaries exists and the indefinite integral $F(x)$ is continuous in any of the singular points ($F'(x) = f(x)$), show that formally the Newton–Leibnitz formula (4.43) is still valid, i.e. that*

$$\int_a^b f(x)dx = F(b) - F(a). \tag{4.75}$$

[*Hint*: first consider a single singularity ($n = 1$), and then generalise your result for their arbitrary finite number n.]

[8] It is said that it diverges logarithmically.

Problem 4.46. *An essential condition for the above formula to be valid is that the indefinite integral $F(x)$ is continuous at all singularity points as otherwise the limits from the left and right at the singularity points would be different and not cancel out. This point is illustrated by the following problem. Show using the definition of the improper integral of a function with singularity at $x = 0$ that*

$$\int_{-1}^{1} x^{-1/3} dx = 0.$$

On the other hand, using the Newton–Leibnitz formula (4.75), we obtain the same result:

$$\int_{-1}^{1} x^{-1/3} dx = \frac{3}{2} x^{2/3} \bigg|_{-1}^{1} = \frac{3}{2} \left[1^{2/3} - (-1)^{2/3} \right] = \frac{3}{2} (1 - 1) = 0.$$

This is because the indefinite integral $F(x) = x^{2/3}$ is continuous at $x = 0$.

The integral of $f(x)$ may not always be calculated using the Newton–Leibnitz formula as this requires knowing the indefinite integral, i.e. the function $F(x)$ such that $F'(x) = f(x)$. Still, it is frequently useful to know if the given improper integral converges even though its analytical calculation is difficult or even impossible. Therefore, it is useful to be able to assess the convergence of an integral without actually calculating it. Below we shall consider several useful criteria which can be used to test the convergence of improper integrals.

Theorem 4.8 (Comparison Test). *Consider two continuous functions $f(x)$ and $g(x)$ such that $0 \leq f(x) \leq g(x)$ for all x within the interval $a < x < b$, where a could be either finite or $-\infty$ and b is also either finite or equal to $+\infty$. Both functions may also have a finite number of singularities within the interval. Then, if the integral $\int_a^b g(x)dx$ converges, then the integral $\int_a^b f(x)dx$ also converges; also if the latter integral diverges, the former one diverges as well.*

Proof. For definiteness consider both integrals with a finite bottom limit a and the infinite upper limit $b = +\infty$. We define two functions of the upper limit:

$$F(X) = \int_a^X f(x)dx \quad \text{and} \quad G(X) = \int_a^X g(x)dx.$$

Since $f(x)$ and $g(x)$ are positive, both functions $F(X)$ and $G(X)$ are increasing monotonically. Indeed, for instance by taking any positive $\Delta X > 0$ we can write:

$$F(X+\Delta X) = \int_a^{X+\Delta X} f(x)dx = \int_a^X f(x)dx + \int_X^{X+\Delta X} f(x)dx$$
$$= F(X) + \int_X^{X+\Delta X} f(x)dx = F(X) + \Delta F \geq F(X),$$

where the integral ΔF is obviously positive since $f(x) > 0$ within the interval $X \leq x \leq X + \Delta X$.

Now since $f(x) \leq g(x)$, then $F(X) \leq G(X)$ as integrals are defined as the integral sums. Next, assume that the integral of $g(x)$ converges. This means that the limit of $G(X)$ when $X \to +\infty$ exists; moreover, since $G(X)$ increases with the increase of X, it must be limited from above, i.e. $G(+\infty) < M$ with some finite M. The function $F(X)$ is never larger than $G(X)$, and hence is also limited from above, e.g. by the same value M. Hence, the function $F(X)$ must have a definite limit when $x \to +\infty$ (see Theorem 2.19 in Sect. 2.4.4), i.e. $F(+\infty)$ can only be finite, i.e. the integral of $f(x)$ also converges. This proves the first part of the theorem that from convergence of the integral of $g(x)$ follows convergence of the integral with $f(x)$.

The second part of the theorem is proven by contradiction: let the integral of $f(x)$ diverges. Then, if we assume that the integral of $g(x)$ converges, the integral of $f(x)$ will converge as well by virtue of the first part of the theorem. However, this contradicts with the fact that the integral of $f(x)$ actually diverges. Therefore, our assumption concerning the convergence of the integral of $g(x)$ is incorrect; this integral diverges. **Q.E.D.**

Other cases (the bottom limit is $-\infty$ and/or the functions have singularities between $-\infty$ and $+\infty$) are considered similarly. Note that the theorem cannot establish convergence of the integral of $f(x)$ if the integral of $g(x)$ diverges; the former integral may either diverge or converge and this requires further consideration.

As an example consider the convergence of the integral of the function $f(x) = e^{-x}\sin^2 x$ containing the infinite upper limit. Since $\sin^2 x \leq 1$, we can introduce a positive function $g(x) = e^{-x}$ such that for any x we have $f(x) \leq g(x)$ since $\sin^2 x \leq 1$. Then, because the integral

$$\int_a^\infty e^{-x}dx = -e^{-x}\Big|_a^\infty = e^a$$

converges (here a is any finite number), so must be true for the integral $\int_a^\infty e^{-x}\sin^2 x dx$.

Problem 4.47. *Investigate the convergence of the integral*

$$\int_1^\infty \frac{x^{-\alpha}}{1+x^2}dx$$

for various values of a positive parameter α. [*Hint*: find the upper limit of the function $1/\left(1+x^2\right)$ for $x \geq 1$. *Answer*: $\alpha > -1$.]

Problem 4.48. *Investigate the convergence of the integral*

$$\int_{-1}^{1} \frac{\sin^2 x}{\sqrt{x}}dx.$$

[*Answer*: converges.]

When investigating convergence, it is useful to approximate the integrand around the singularity point to see if the integral with the limits just around that singularity converges. For instance, in the above two problems, we have the integrands given by $f_1(x) = x^{-\alpha}/\left(1+x^2\right)$ and $f_2(x) = x^{-1/2}\sin^2 x$. In the first case we have to investigate the convergence at $x \to \infty$. For large x one may approximate $1 + x^2$ simply with x^2, leading to the integral of $x^{-\alpha-2}$ which, when integrated, results in a function proportional to $x^{-\alpha-1}$. This function is well defined at $x \to \infty$ (and tends to zero there) only if $\alpha > -1$. In the second case the singularity occurs at $x = 0$. For small x we can replace $\sin x$ with x, which results in the integrand proportional to $x^{-1/2+2} = x^{3/2}$, which is integrable around $x = 0$.

The criterion stated by Theorem 4.8 requires the two functions to be necessarily non-negative. If the given function $f(x)$ changes its sign, then this test cannot be applied.

Theorem 4.9. *Consider an improper integral* $\int_a^b f(x)dx$ *(where either the function* $f(x)$ *has singularities within the interval and/or one or both limits* a *and* b *are infinite). This integral converges if the integral* $\int_a^b |f(x)|\, dx$ *of the absolute value of* $f(x)$ *converges.*

Proof. Indeed, since the integral of $|f(x)|$ converges, then $\int_a^b |f(x)|\, dx < M$ with some positive M. On the other hand, the integral is a limit of the integral sum, and hence one can write (see Eq. (4.29) in Sect. 4.2)

$$\left|\int_a^b f(x)dx\right| \leq \int_a^b |f(x)|\, dx < M,$$

i.e. the improper integral under study, $\int_a^b f(x)dx$, is limited and hence converges. **Q.E.D.**

The proven theorem tells us that from the convergence of the integral $\int_a^b |f(x)|\, dx$ follows convergence of $\int_a^b f(x)dx$; in this case, it is said the latter integral converges *absolutely*. This condition of convergence is stronger than an ordinary convergence of the integral $\int_a^b f(x)dx$, since the integral $\int_a^b |f(x)|\, dx$ may not converge even if the integral $\int_a^b f(x)dx$ does.

Problem 4.49. *Consider an improper integral $\int_a^b f(x)dx$ that converges absolutely and a function $g(x)$ bounded from below and above, $|g(x)| < M$. Prove that the improper integral $\int_a^b f(x)g(x)dx$ also converges absolutely.*

Problem 4.50. *In particular, investigate the convergence of the integral $\int_1^\infty g(x)x^{-\lambda}dx$, where $\lambda > 0$ and the function $g(x)$ is limited from below and above.* [*Answer*: converges for $\lambda > 1$.]

Problem 4.51. *Similarly, investigate the convergence of the integral $\int_0^\infty g(x)e^{-x}dx$.* [*Answer*: converges.]

Problem 4.52. *Investigate the convergence of the following integrals:*

$$\int_0^\infty \frac{\sin x}{1+x^2}dx \quad \textit{and} \quad \int_1^\infty \frac{\cos x}{x^\lambda}dx \quad \textit{with} \quad \lambda > 0.$$

[*Answers*: converges; converges for $\lambda > 1$.]

4.5.4 Cauchy Principal Value

Consider a function $f(x)$ which lies within the interval $a < x < b$ and has there a singularity at an internal point c (here we assume that a and b are either finite or infinite). The corresponding improper integral is defined by Eq. (4.74) with the two limits taken independently of each other, i.e. with $\epsilon_1 \neq \epsilon_2$. If this improper integrals diverges, it might in some cases be useful to constrain the limiting procedure to having $\epsilon_1 = \epsilon_2 = \epsilon$ and hence perform a *single* limit $\epsilon \to +0$. If this limit exists, the integral is said to converge in the sense of the *Cauchy principal value*:

$$\mathcal{P}\int_a^b f(x)dx \quad \text{or} \quad ⨍_a^b f(x)dx = \lim_{\epsilon\to 0}\left(\int_a^{c-\epsilon} f(x)dx + \int_{c+\epsilon}^b f(x)dx\right). \tag{4.76}$$

This value may be assigned to some of the improper integrals which otherwise diverge. It is a special form of *regularisation* of the otherwise ill-defined integrals.

Consider, for instance, the following improper integral ($a < c < b$):

$$I = \int_a^b \frac{dx}{x-c}.$$

In the sense of the formal definition of Sect. 4.5.3 this integral is at least ill-defined:

$$\begin{aligned} I &= \lim_{\epsilon_1\to 0}\int_a^{c-\epsilon_1}\frac{dx}{x-c}+\lim_{\epsilon_2\to 0}\int_{c+\epsilon_2}^{b}\frac{dx}{x-c}=\lim_{\epsilon_1\to 0}\ln\frac{\epsilon_1}{c-a}+\lim_{\epsilon_2\to 0}\ln\frac{b-c}{\epsilon_2} \\ &= \ln\frac{b-c}{c-a}+\left[\lim_{\epsilon_1\to 0}\ln\epsilon_1-\lim_{\epsilon_2\to 0}\ln\epsilon_2\right]=\ln\frac{b-c}{c-a}+[(-\infty)-(-\infty)] \\ &= \ln\frac{b-c}{c-a}+(-\infty+\infty)\,. \end{aligned}$$

At the same time, its principal value is well defined:

$$I=\lim_{\epsilon\to 0}\left[\int_a^{c-\epsilon}\frac{dx}{x-c}+\int_{c+\epsilon}^{b}\frac{dx}{x-c}\right]=\lim_{\epsilon\to 0}\left[\ln\frac{\epsilon}{c-a}+\ln\frac{b-c}{\epsilon}\right]=\ln\frac{b-c}{c-a},$$

as the two terms with $\ln\epsilon$ cancel out exactly.

Problem 4.53. *Calculate the principal value of $\int_a^b (x-c)^{-n}\,dx$ with $n>1$ being a positive integer and c lying between a and b. Consider the cases of n odd and even. When does this integral exist?* [*Answer*: n must be odd; $\left[(b-c)^{1-n}-(a-c)^{1-n}\right]/\,(1-n)$.]

It can also be shown that a more general integral

$$I=\int_a^b\frac{f(x)}{x-c}dx,$$

in which the function $f(x)$ is continuous everywhere between a and b, is well defined in the principal value sense. Indeed, let us rewrite it as follows:

$$I=\int_a^b\frac{f(x)-f(c)}{x-c}dx+f(c)\int_a^b\frac{dx}{x-c}.$$

Here the first integral is not improper anymore since the integrand is a continuous function at $x=c$ with the limiting value of $f'(c)$ (use L'Hôpital's rule, Sect. 3.9, to check this!); the second integral is well defined in the principal sense.

The regularisation we performed above, when discussing the improper integrals with the integrand having a singularity, is not the only possible. Indeed, consider the integral (4.68) with both infinite boundaries. Its usual definition also contains two limits and in some cases these may yield divergence if applied independently.

However, the following regularisation, also called the principal value, may result in a well-defined limit:

$$\mathcal{P}\int_{-\infty}^{\infty} f(x)dx = \lim_{a\to\infty}\int_{-a}^{a} f(x)dx. \tag{4.77}$$

As an example, consider the following integral:

$$\begin{aligned} I = \int_{-\infty}^{\infty}\frac{xdx}{x^2+1} &= \lim_{a\to-\infty}\left\{\lim_{b\to+\infty}\int_a^b \frac{xdx}{x^2+1}\right\} \\ &= \lim_{a\to-\infty}\left\{\lim_{b\to+\infty}\frac{1}{2}\left[\ln\left(b^2+1\right)-\ln\left(a^2+1\right)\right]\right\} \\ &= \lim_{b\to\infty}\ln\left(b^2+1\right)-\lim_{a\to-\infty}\ln\left(a^2+1\right). \end{aligned}$$

This expression is ill-defined as it contains a difference of two infinities. At the same time, the principal value of this integral is well defined:

$$I = \lim_{a\to+\infty}\int_{-a}^{a}\frac{xdx}{x^2+1} = \lim_{a\to+\infty}\frac{1}{2}\ln\frac{1+a^2}{1+a^2} = 0,$$

which in fact is what one would assign to this integral as the integrand is an odd function.

Problem 4.54. *Investigate whether improper integrals* $\int_{-\infty}^{\infty}\sin x dx$ *and* $\int_{-\infty}^{\infty}\cos x dx$ *exist in the principal value sense.* [*Answer*: 0; diverges.]

Problem 4.55. *Show that*

$$\mathcal{P}\int_{-E}^{E}\frac{dx}{(x-a)(x-b)} = \frac{1}{a-b}\ln\left|\frac{(E-a)(E+b)}{(E+a)(E-b)}\right|$$

irrespective of whether the points a and/or b are inside or outside the integration interval.

4.6 Applications of Definite Integrals

Here we shall consider some applications of definite integrals in geometry and physics. The concept of integration is so general and widespread in engineering and physics (as well as, e.g., in chemistry and biology where physical principles are applied), that it is absolutely impossible to give here a complete account of all possible uses of it; only some most obvious examples will be given.

The underlying idea in all these applications is the same; therefore, it is essential to get the general principle, and then it would be straightforward to apply it in any other situation: we have to sum up small quantities of something to get the whole of it, e.g. small lengths along a curved line to calculate the length of the whole line, small volumes within some big volume bounded by some curved surface or the work done over a small distance summed over the whole distance. There is always a parameter in the problem at hand, let us call it the "driving coordinate" and denote by q, which, if its change Δq_{all} is multiplied with some other quantity, say F, would give the required result: $A = F\Delta q_{\text{all}}$. The problem here is that the formula $A = F\Delta q_{\text{all}}$ would only make sense if F was constant; if this is not the case and $F(q)$ is a function of q, the simple product $A = F\Delta q_{\text{all}}$ would not make any sense. So, the idea is to discretise the values of q and construct the Riemann sum similar to the one in Eq. (4.1) which we discussed when doing the calculation of an area under the curve. So, we choose n points $q_0 = q_{\text{ini}}, q_1, q_2, \ldots, q_n = q_{\text{fin}}$ between the initial, q_{ini}, and final, q_{fin}, values of the parameter q. Most easily, these points can be chosen equidistant, i.e. $q_i = q_{\text{ini}} + i\Delta q$, where $i = 0, 1, 2, \ldots, n$ and $\Delta q = (q_{\text{fin}} - q_{\text{ini}})/n$. Then, the quantity of interest corresponding to Δq is approximately calculated as $\Delta A_i \simeq F(q_i)\Delta q$, this is to be associated with a small change $\Delta q = q_{i+1} - q_i$ of q, since for sufficiently small values of Δq, i.e. between sufficiently close values q_i and q_{i+1}, the function $F(q)$ can be approximated as a constant $F(q_i)$. The final quantity is obtained by summing up all the contributions,

$$A \simeq \sum_{i=1}^{n} F(q_i)\Delta q,$$

and then taking the limit of $n \to \infty$ (or $\Delta q \to 0$) which would give the exact result in the case of the variable function $F(q)$.

4.6.1 Length of a Curved Line

Consider a curved line in the $x - y$ plane specified by the function $y = f(x)$ within the interval $a \le x \le b$ as shown in Fig. 4.4(a). We would like to calculate the length of the line. To do this, we first divide the line into n small segments by points A_i, where $i = 0, 1, 2, \ldots, n$, and then connect the points by straight lines (shown as dashed). The length of each segment between two neighbouring points A_i and A_{i+1} is then approximated by $\Delta l_i = \sqrt{(\Delta x_i)^2 + (\Delta y_i)^2}$, where $\Delta x_i = x_{i+1} - x_i$ and $\Delta y_i = y_{i+1} - y_i$ are the corresponding increments of the x and y coordinates of the two points, as shown in Fig. 4.4(b). Since the change Δy_i can be related to the derivative of the function $y = f(x)$ at point x_i, i.e. $\Delta y_i \simeq f'(x_i)\Delta x_i$ (recall our discussion in Sect. 3.2, Eq. (3.9), and Sect. 3.8), we can write

$$\Delta l_i \simeq \sqrt{(\Delta x_i)^2 + (f'(x_i)\Delta x_i)^2} = \sqrt{1 + (f'(x_i))^2}\,\Delta x_i.$$

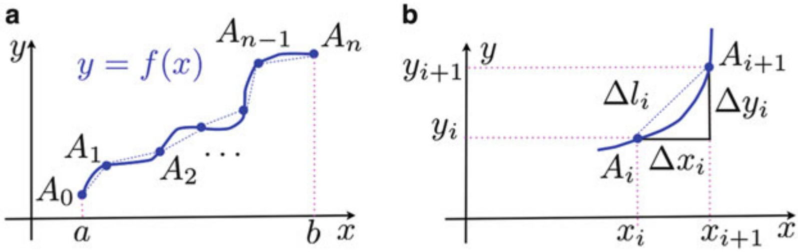


Fig. 4.4 (**a**) A *curved line* between points $(a,f(a))$ and $(b,f(b))$ is divided by points A_i $(i = 0,\ldots,n)$ into n curved segments that are then replaced by straight lines (shown by the *dashed lines*). (**b**) The i-th such segment is shown. Its length is approximated by the length Δl_i of the straight line connecting the points A_i and A_{i+1}

Summing up these lengths along the whole curve and taking the limit of $n \to \infty$ (or $\max_i \{\Delta x_i\} \to 0$), we arrive at the useful exact formula for calculating the length of a curved line:

$$l = \int_a^b \sqrt{1 + \left(\frac{dy}{dx}\right)^2} dx = \int_a^b \sqrt{1 + (f'(x))^2} dx. \tag{4.78}$$

The role of the *driving* coordinate is here served by x. To make sure that this calculation always yields a positive result, one has to integrate from the smallest to the largest values of the x.

As an example, let us calculate the length of a circle of radius R. The equation of a circle centred at the origin is $x^2 + y^2 = R^2$. Its length is most easily calculated by considering a quarter of the circle in the first quadrant, where both $x > 0$ and $y > 0$ and $y = \sqrt{R^2 - x^2}$. This will have to be multiplied by 4 to give the whole circle circumference. We have $y' = -x/\sqrt{R^2 - x^2}$ and hence

$$\frac{l}{4} = \int_0^R \sqrt{1 + (y')^2} dx = R \int_0^R \frac{dx}{\sqrt{R^2 - x^2}} = R \int_0^1 \frac{dt}{\sqrt{1 - t^2}} = R \arcsin t|_0^1 = R\frac{\pi}{2},$$

which yields for the whole length a well-known expression of $l = 2\pi R$.

Note that the same integral over x can also be calculated using the substitution $x = R\cos\phi$:

$$\frac{l}{4} = R \int_0^R \frac{dx}{\sqrt{R^2 - x^2}} = \left| \begin{array}{c} x = R\cos\phi \\ dx = -R\sin\phi d\phi \end{array} \right| = R \int_{\pi/2}^0 \frac{-R\sin\phi d\phi}{R\sin\phi}$$

$$= R \int_0^{\pi/2} d\phi = R\frac{\pi}{2},$$

which of course is the same result. Although what we have just done can formally be considered as yet another substitution, we have actually used the polar coordinates (see Sect. 1.13.2) with the ϕ being the angle, since in these coordinates $x = R\cos\phi$ and $y = R\sin\phi$.

This consideration brings us to the following point: a line can also be considered by means of the polar coordinates in which case the angle ϕ is used as the driving coordinate along the line. In polar coordinates the line is specified via the equation (Sect. 1.13.3) $r = r(\phi)$. Since $x = r(\phi)\cos\phi$ and $y = r(\phi)\sin\phi$, we can write:

$$\begin{aligned} dx &= x'(\phi)d\phi = \left[r'(\phi)\cos\phi - r\sin\phi\right]d\phi \ , \\ dy &= y'(\phi)d\phi = \left[r'(\phi)\sin\phi + r\cos\phi\right]d\phi, \end{aligned} \tag{4.79}$$

and therefore

$$l = \int_{\phi_{\min}}^{\phi_{\max}} dl = \int_{\phi_{\min}}^{\phi_{\max}} \sqrt{dx^2 + dy^2} = \int_{\phi_{\min}}^{\phi_{\max}} \sqrt{r'(\phi)^2 + r^2}d\phi. \tag{4.80}$$

Check this simple result using Eq. (4.79).

Let us apply this rather powerful formula to the same problem of calculating the circumference of the circle of radius R. The equation of the circle is simply $r = R$ (i.e. r does not depend on ϕ), so that $r' = 0$ and

$$l = \int_0^{2\pi} \sqrt{0 + R^2}d\phi = 2\pi R.$$

Problem 4.56. *Show that the length of the spiral in Fig. 1.20(d) given in the polar coordinates as* $r(\phi) = e^{-\phi}$ *(with* ϕ *changing from* 0 *to* ∞*) is equal to* $\sqrt{2}$.

Problem 4.57. *Show that the length of the single revolution of the Archimedean spiral, Fig. 1.20(c), given in polar coordinates as* $r = a\phi$ *(with* ϕ *changing from* 0 *to* 2π*), is equal to*

$$l = \frac{a}{2}\left[2\pi\sqrt{1+4\pi^2} + \ln\left(2\pi + \sqrt{1+4\pi^2}\right)\right].$$

Problem 4.58. *Show that the length of the snail curve specified via* $r = a(1+\cos\phi)$*, see Fig. 1.20(b), is equal to* $8a$.

More generally, if the curve on the $x - y$ plane is specified parametrically as $x = x(t)$ and $y = y(t)$, then

$$l = \int_{t_{\min}}^{t_{\max}} \sqrt{x'(t)^2 + y'(t)^2}dt, \tag{4.81}$$

since $dx = x'(t)dt$ and $y = y'(t)dt$. This result can easily be generalised to a curve in three-dimensions which is specified parametrically as $x = x(t)$, $y = y(t)$ and $z = z(t)$. In this case the length dl of a small section of the curve obtained by incrementing the three coordinates by dx, dy and dz is obtained as

$$dl = \sqrt{dl_1^2 + dz^2} = \sqrt{(dx^2 + dy^2) + dz^2} = \sqrt{x'(t)^2 + y'(t)^2 + z'(t)^2}dt \tag{4.82}$$

(see Fig. 4.5) and therefore the length of the line

$$l = \int_{t_{\min}}^{t_{\max}} \sqrt{x'(t)^2 + y'(t)^2 + z'(t)^2}dt. \tag{4.83}$$

Problem 4.59. *Show that the length of the three revolutions of the spiral, specified via $x(\phi) = R\cos\phi$, $y(\phi) = R\sin\phi$ and $z(\phi) = b\phi$ and shown in Fig. 4.6(a), is given by $l = 6\pi\sqrt{R^2 + b^2}$.*

Problem 4.60. *Derive a formula for the maximum length of a wire of radius r that can be coiled around a cylinder of radius R and length L in a single layer. Then, consider the second, third, and so on layers wound one after another, see Fig. 4.6(b).*

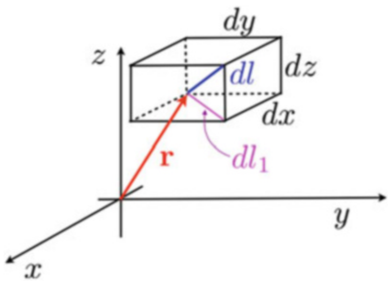


Fig. 4.5 For the derivation of the diagonal dl of a three-dimensional cuboid formed by three sides (increments along the three Cartesian axes) dx, dy and dz

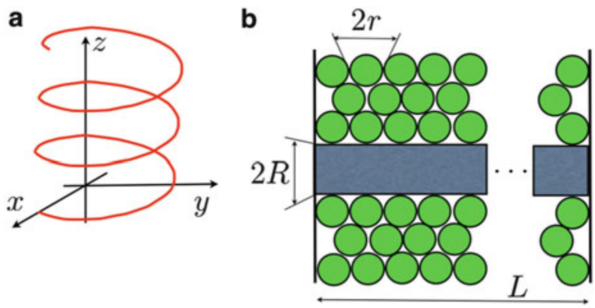


Fig. 4.6 (**a**) Three revolutions of the spiral. (**b**) Three coils of wire of radius r wound around a cylinder of radius R

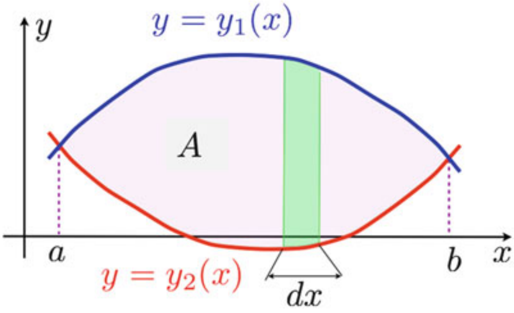


Fig. 4.7 Area between two curves $y = y_1(x)$ and $y = y_2(x)$ intercepting at $x = a$ and $x = b$

4.6.2 Area of a Plane Figure

We know that if $y = y(x)$ is a positive function, then the area under it (i.e. the area bounded by the curve above and the x axis below) and between the vertical lines $x = a$ and $x = b$ is given by the definite integral $\int_a^b y(x)dx$. This result is easily generalised for an area bounded by two curves $y = y_1(x)$ and $y = y_2(x)$ as (see Fig. 4.7):

$$A = \int_a^b [y_2(x) - y_1(x)]\, dx. \tag{4.84}$$

> **Problem 4.61.** *Write a similar formula for a figure enclosed between two curves $x = x_1(y)$ and $x = x_2(y)$.*

These formulae, however, are not very convenient if the figure has a more complex shape. For instance, this may happen if the figure (i.e. the closed curve which bounds it) is specified parametrically as $x = x(t)$ and $y = y(t)$, see schematics in Fig. 4.8(a). To derive the corresponding formula in this case, we have to carefully consider regions where for the same value of x there are several possible values of y, i.e. the shaded regions in Fig. 4.8(a). To this end, let us analyse several possible cases noticing that with the chosen direction of the traverse of the bounding curve (which corresponds to the increase of the parameter t) the body of the figure is always on the right; regions which are on the left of the traverse direction must not be included in the calculation. The green shaded region in case A in Fig. 4.8(b) is to be fully included, and its area is $dA = y(t)\,[x(t + dt) - x(t)] = y(t)x'(t)dt$ and is positive. The to-be-excluded area in case B can formally be written using the same expression, but dA in this case will be negative (since $x(t + dt) < x(t)$ and will then be subtracted). Similar analysis shows that the area is negative in case C and positive

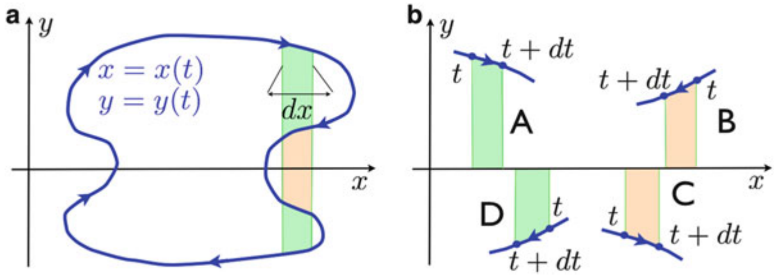


Fig. 4.8 For the calculation of the area of a figure whose bounding curve is specified by parametric equations $x = x(t)$ and $y = y(t)$ with the parameter t. (**a**) A sketch of a figure with a complex shape; *green shaded areas* should be included into the calculated area, while the *orange shaded areas* are to be excluded. (**b**) Several particular cases (see text). The direction in which the bounding curve is traversed when the parameter t is increased is also indicated

in case D. It is clear now that if we sum up all these areas, the correct expression for the whole area enclosed by the bounding curve is obtained:

$$A = \left| \int_{t_a}^{t_b} y(t)x'(t)dt \right| , \tag{4.85}$$

where t_a and t_b are the initial (usually, zero) and final values of the t required to draw the complete figure. The regions that need to be excluded are excluded in this case automatically (explain, why). We use the absolute value in the above equation since, depending on the direction of the traverse, the result may be negative.

As an example, let us calculate the area enclosed by a circle of radius R centred at the origin. In this case the parametric equations for the circle are $x(\phi) = R\cos\phi$ and $y(\phi) = R\sin\phi$, so that we obtain:

$$\begin{aligned} A_{\text{circle}} &= \left| \int_0^{2\pi} (R\sin\phi)(-R\sin\phi)\, d\phi \right| = R^2 \int_0^{2\pi} \sin^2\phi d\phi \\ &= R^2 \int_0^{2\pi} \frac{1}{2}\left[1 - \cos(2\phi)\right] d\phi = R^2 \frac{1}{2}(2\pi - 0) = \pi R^2. \end{aligned}$$

> **Problem 4.62.** *Show that the area of an ellipse, given parametrically via $x(\phi) = a\cos\phi$ and $y(\phi) = b\sin\phi$, is equal to $A_{ellipse} = \pi ab$. Note that when $a = b$, the ellipse turns into a circle of radius a, and the correct expression for its area, πa^2, is indeed recovered.*

Finally, we shall consider yet another useful result for the area of a figure specified in polar coordinates via an equation $r = r(\phi)$ (see Sect. 1.13.3).

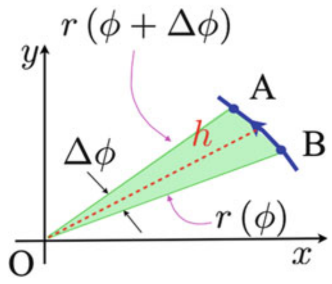


Fig. 4.9 For the derivation of formula (4.86) for the area of a plane figure specified via the equation $r = r(\phi)$ in polar coordinates

In this case one can use Eq. (4.85) with the polar angle ϕ as the parameter (see Problem 4.64). However, a simpler method exists. Consider the green shaded sector in Fig. 4.9 corresponding to a small change $\Delta\phi$ of the polar angle ϕ from ϕ to $\phi + d\phi$. The area of the sector is then calculated assuming that the figure OAB can be treated approximately as a triangle with the height $h \simeq r(\phi)$ which is perpendicular to the side $AB \simeq 2r(\phi)\tan(\Delta\phi/2) \simeq r(\phi)\Delta\phi$ of the triangle (recall that for small angles, $\tan\alpha \simeq \alpha$, see Eq. (3.37)). Therefore, the area of the triangle $\Delta A \simeq h \cdot AB/2 = r^2(\phi)\Delta\phi/2$. Summing up these areas and tending $\Delta\phi \to 0$, i.e. integrating over the angle ϕ, an exact expression is obtained:

$$A = \frac{1}{2}\int_{\phi_1}^{\phi_b} r^2(\phi)d\phi. \tag{4.86}$$

It is easy to see that for a circle $r(\phi) = R$ is a constant, the angle ϕ changes between 0 and 2π, so that the same expression for the circle area is obtained as above, $A = \frac{1}{2}R^2 2\pi = \pi R^2$.

Problem 4.63. *Here we shall perform a more careful derivation of Eq. (4.86). Let (x, y) are the two Cartesian coordinates of the point B in Fig. 4.9, where $x = r(\phi)\cos\phi$ and $y = r(\phi)\sin\phi$, while the point A similarly has coordinates corresponding to the angle $\phi + \Delta\phi$. Calculate the area of the triangle OAB by considering the absolute value of the vector product of the vectors $\overrightarrow{OA}$ and $\overrightarrow{OB}$, and then expand the obtained expression in terms of $\Delta\phi$ recovering $\Delta A = r^2(\phi)\Delta\phi/2$ to the first order in $\Delta\phi$.*

Problem 4.64. *Derive Eq. (4.86) from Eq. (4.85).* [*Hint*: use $x(\phi) = r(\phi)\cos\phi$ and $y(\phi) = r(\phi)\sin\phi$ and integration by parts (only in the term containing $r'(\phi)$); also make use of the fact that the angle ϕ changes by an integer amount of π when traversing around the whole figure.]

Problem 4.65. *Show that the area enclosed by the three-leaf rose shown in Fig. 1.20(a) is $A = \pi a^2/4$.*

4.6.3 Volume of Three-Dimensional Bodies

Later on in Chap. 6 after developing concepts of a function of several variables and of double and triple integrals, we shall learn how to calculate volumes of general three-dimensional bodies. However, in simple cases it is possible to calculate the volume of some figures using a single definite integral. Here we shall consider some of the most common cases.

We start with the simplest case when one can select a direction through the figure (let us draw the z axis along it) such that the area $A(z)$ of the cross section at each value of z is known, see Fig. 4.10(a). The volume inside a small cylinder of height dz enclosed between cross sections at z and $z + dz$ is obviously equal to $A(z)dx$, so that the whole volume is given by:

$$V = \int_{z_1}^{z_2} A(z)dz, \tag{4.87}$$

where z_1 and z_2 and the smallest and the largest values of the z which are necessary to specify the object under discussion.

As a simple application of this result, let us first consider a cylinder of radius R and length (height) h. If we choose the z axis along the symmetry axis of the cylinder, then the cross section at any point z along the cylinder height will be the a circle of the same radius R with the area $A(z) = \pi R^2$. It does not depend on the z, leading to the volume

$$V_{\text{cylinder}} = \int_0^h \pi R^2 dz = \pi R^2 h. \tag{4.88}$$

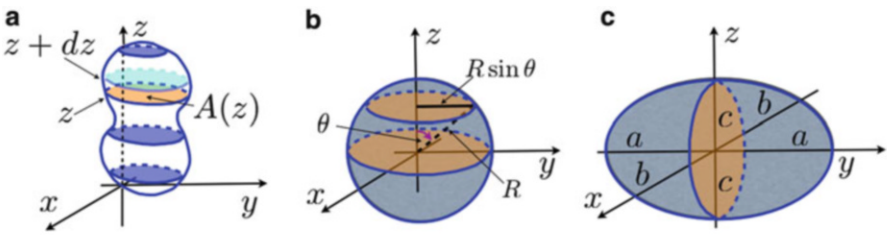


Fig. 4.10 Three-dimensional figures for which the area of its cross section at each value of z is known to be $A(z)$. (**a**) A general figure. (**b**) A sphere of radius R. (**c**) An ellipsoid with parameters a, b and c

As a less trivial example, consider now a sphere of radius R centred at the origin, Fig. 4.10(b). A cross section of the sphere perpendicular to the z axis is a circle of radius $r = R\sin\theta$, where the angle θ is related to z via $z = R\cos\theta$, i.e. $r = R\sqrt{1-(z/R)^2}$, so that the area of the cross section $A(z) = \pi r^2 = \pi\left(R^2 - z^2\right)$, and we obtain for the volume:

$$V_{\text{sphere}} = \int_{-R}^{R} \pi\left(R^2 - z^2\right) dz = \pi\left(2R^3 - 2\frac{R^3}{3}\right) = \frac{4}{3}\pi R^3. \tag{4.89}$$

Problem 4.66. *Show that the volume of an ellipsoid shown in Fig. 4.10(c) is*

$$V_{ellipsoid} = \frac{4}{3}\pi abc. \tag{4.90}$$

Note that the equation of the ellipsoid is given by

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 + \left(\frac{z}{c}\right)^2 = 1. \tag{4.91}$$

The simple result (4.90) makes perfect sense: when the ellipsoid is fully symmetric, i.e. when $a = b = c$, it becomes a sphere, and the volume in this case goes directly into the expression (4.89) for the sphere.

Problem 4.67. *Show that the volume of a cone shown in Fig. 4.11(a) is*

$$V_{cone} = \frac{1}{3}\pi R^2 h = \frac{1}{3}\pi h^3 \tan^2\alpha. \tag{4.92}$$

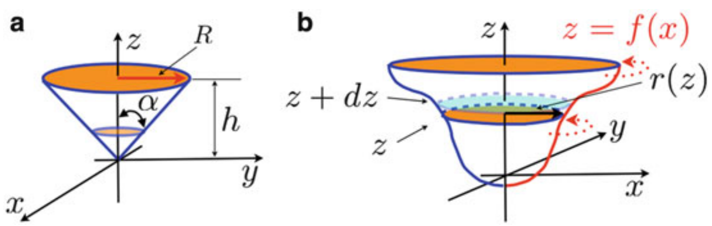


Fig. 4.11 (**a**) A cone of radius R and height h. (**b**) A solid obtained by a full revolution around the z axis of the curve $z = f(x)$

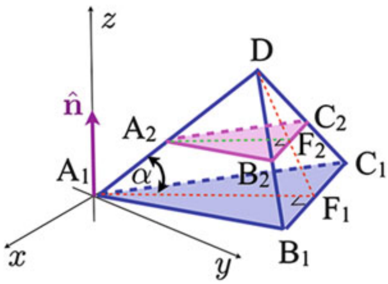


Fig. 4.12 For the calculation of the volume of a triangular pyramid (*tetrahedron*) $A_1B_1C_1D$

Problem 4.68. *Show that the volume of a tetrahedron $A_1B_1C_1D$ (also called a triangular pyramid), shown in Fig. 4.12, is given by the following formula:*

$$V_{tetrahedron} = \frac{1}{3} S_{\triangle} c \sin\alpha, \tag{4.93}$$

where $S_{\triangle}$ is the area of its base (i.e. of the triangle $\triangle A_1B_1C_1$), $c = A_1D$ is one of its sides starting from the base and going towards the pyramid summit, and α is the angle between the side and the base. [*Hint*: Let the base $A_1B_1C_1$ of the pyramid be in the $x-y$ plane. Consider a point A_2 along the side $A_1D = c$, and let $A_1A_2 = t$, where t changes from 0 to c. Demonstrate, using the similarity of the triangles $\triangle A_1B_1C_1$ and $\triangle A_2B_2C_2$, that the area of the intermediate triangle $\triangle A_2B_2C_2$, expressed as a function of the parameter t, is given by $S_{\triangle}(t) = \left(ah/2c^2\right)(c-t)^2$, where $h = A_1F_1$ is the height of the base drawn towards the side $a = B_1C_1$. Then relate dz and dt, and obtain the volume by integrating over dt.]

Problem 4.69. *Next consider the triangular pyramid of the previous problem as created by three vectors: $\overrightarrow{A_1B_1}$, $\overrightarrow{A_1C_1}$ and $\overrightarrow{A_1D}$. Show that the volume of the pyramid (4.93) can also be written as*

$$V_{tetrahedron} = \frac{1}{6}\left|\left(\overrightarrow{A_1D}\cdot\left[\overrightarrow{A_1B_1}\times\overrightarrow{A_1C_1}\right]\right)\right| = \frac{1}{6}\left|\left[\overrightarrow{A_1B_1},\overrightarrow{A_1C_1},\overrightarrow{A_1D}\right]\right|, \tag{4.94}$$

i.e. it can be related to the mixed product of the three vectors. Since the triple product corresponds to the volume of the parallelepiped formed by the three vectors, see Eq. (1.41), we see that the volume of the tetrahedron is one sixth of the volume of the corresponding parallelepiped.

The volume can also be easily expressed via a definite integral for another class of symmetric (in fact, cylinder-like) three-dimensional objects which are obtained by rotating a two-dimensional curve $z = f(x)$ around the z axis, the so-called *solids of revolution*, see Fig. 4.11(b). In this case one can directly use Eq. (4.87) in which the radius of the circle in the cross section, $r(z)$, is obtained by solving the equation $z = f(x)$ for the x, i.e. $r = x = f^{-1}(z)$. Therefore,

$$V = \int_{z_1}^{z_2} \pi x^2 dz , \quad x = f^{-1}(z). \tag{4.95}$$

The cone in Fig. 4.11(a) is the simplest example of such a figure as it can be obtained by rotating the straight line $z = kx$ with $k = \cot\alpha$ around the z axis. Therefore, in this case $x(z) = z/k = (\tan\alpha)\, z$ and the integration above yields exactly the same result as in Eq. (4.92).

Problem 4.70. *Show that the volume of a solid of revolution obtained by rotating around the z axis the parabola $z = ax^2$ with $0 \le x \le R$ is $V = \frac{1}{2}\pi a R^4$.*

Problem 4.71. *Show that the volume of a solid of revolution obtained by rotating around the z axis the curve $z = ax^4$ with $0 \le x \le R$ is $V = (2\pi/3)\, R^6 a$.*

4.6.4 A Surface of Revolution

A general consideration of the surface area of three-dimensional bodies requires developing a number of special mathematical instruments, and we shall do it in Chap. 6. These will allow us to calculate the area of the outer surface of arbitrary bodies. Here we only show that in some simple cases this problem can be solved with the tools which are at our disposal already.

Consider an object created by rotating around the z axis a curve $z = f(x)$, see Fig. 4.11(b). Above we have considered its volume. Now we turn to the calculation of its outer surface. Looking at the surface layer created between the cross sections at z and $z + dz$, we can calculate its area as a product of $2\pi r(z)$ (which is the length of the circumference of the cross section at z with $r(z) = x(z) = f^{-1}(z)$, as above) and the layer width, dl. The latter is not just dz as the curve $z = f(x)$ may have a curvature. The calculation of dl has already been done in Sect. 4.6.1 (and, specifically, see Fig. 4.4(b)) when considering the length of a curve. Hence, we can write down the result immediately:

$$dl = \sqrt{(dx)^2 + (dz)^2} = \sqrt{1 + x'(z)^2} dz.$$

Therefore, the surface area of the solid of revolution enclosed between z_1 and z_2 is obtained by summing up all these contributions, i.e. by integrating:

$$A = \int_{z_1}^{z_2} 2\pi x(z) \sqrt{1 + (x'(z))^2} dz. \tag{4.96}$$

As an example, consider the surface of a sphere. The equation of a sphere of radius R and centred at the origin, see Fig. 4.10(b), is expressed by $x^2 + y^2 + z^2 = R^2$. It is convenient to consider the upper part of the sphere ($z > 0$) and then multiply the result by 2. The upper hemisphere of the sphere can be considered as a revolution of the curve $z = \sqrt{R^2 - x^2}$ around the z axis. The latter equation is easily solved with respect to the radius of the cross section: $x(z) = \sqrt{R^2 - z^2}$. Therefore, $x'(z) = -z/\sqrt{R^2 - z^2}$ and we can write:

$$\begin{aligned} S_{\text{sphere}} &= 2\int_0^R 2\pi\sqrt{R^2 - z^2}\sqrt{1 + \frac{z^2}{R^2 - z^2}}dz \\ &= 4\pi\int_0^R \sqrt{R^2 - z^2}\sqrt{\frac{R^2}{R^2 - z^2}}dz = 4\pi R\int_0^R dz = 4\pi R^2. \end{aligned} \tag{4.97}$$

Problem 4.72. *Show that the area of the side surface of a cone in Fig. 4.11(a), obtained by rotating the line $z = kx$ with $k = \cot\alpha$, is*

$$A_{cone} = \frac{\pi h^2}{k^2}\sqrt{1 + k^2} = \pi h^2\frac{\sin\alpha}{\cos^2\alpha} = \pi R\sqrt{h^2 + R^2}. \tag{4.98}$$

Problem 4.73. *Show that the area of the side surface of a solid of revolution obtained by rotating around the z axis the parabola $z = ax^2$ with $0 \le x \le R$ is*

$$A = \frac{\pi}{6a^2}\left[(1 + 4ah)^{3/2} - 1\right],$$

where $h = aR^2$ is the height of the object.

Problem 4.74. *A torus of radius c with the tube of radius R, see Fig. 4.13, can be formed by rotating the circle of radius R (shown in the figure as the cross section of the torus tube in the $x - z$ plane) around the z axis. The centre of the circle is displaced from the origin by c along the x axis. The equation of the circle $(x - c)^2 + z^2 = R^2$, which gives $r = x = c \pm \sqrt{R^2 - z^2}$ for the radius of the cross section perpendicular to the z axis. The plus sign corresponds to the outer part of the surface obtained by rotating the right half of the circle with $c \le x \le c + R$, while the minus sign gives the inner surface of the torus due to revolution of the left side of the circle, $c - R \le x \le c$, around the z axis. Because the function $x(z)$ is multivalued, the two surfaces have to be considered separately. Show that their areas are given by $A_\pm = 2\pi R(\pi c \pm 2R)$, where the plus sign corresponds to the outer and the minus to the inner surfaces. Hence, the total surface of the torus appears to be*

$$A_{torus} = 4\pi^2 Rc. \tag{4.99}$$

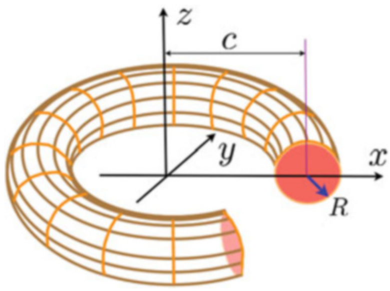


Fig. 4.13 A torus of radius c with the tube of radius R can be obtained by rotating a coloured circle of radius R displaced from the origin by the distance c along the x axis

Problem 4.75. *Consider a solid of revolution created by the line $z = f(x)$. Let the line be given parametrically as $z = z(t)$ and $x = x(t)$. Show that the area of the surface of revolution can be calculated via the integral:*

$$A = \int_{t_1}^{t_2} 2\pi x(t) \sqrt{(x'(t))^2 + (z'(t))^2} dt, \qquad (4.100)$$

where t_1 and t_2 are the corresponding initial and final values of the parameter t. Note that this formula is in a sense more general than (4.96) since it can also be used when a rotation of the curve is performed around an arbitrary axis. It could be any line in space, not necessarily the Cartesian axes, if the appropriate parametric equations are provided.

Problem 4.76. *Consider the torus of Fig. 4.13 again, but this time let us specify the equation of the circle parametrically: $x = c + R\cos\phi$ and $z = R\sin\phi$, where $0 \le \phi \le 2\pi$. The convenience of this representation is shown when using Eq. (4.100) instead of (4.96) the calculation of the surface area can be performed in a single step, there is no need to consider the inner and outer torus surfaces separately as we did in the problem above. Show that the same result (4.99) for the surface area is obtained.*

4.6.5 Simple Applications in Physics

Of course, the language of physics is mathematics and hence integrals (comprising an essential part of the language) play an extremely important role in it. Here we shall give some obvious examples to illustrate this point.

Mechanics. Consider a particle performing a one-dimensional movement along the x axis. Suppose we know the velocity of the particle v, then the distance the particle would travel between times t_1 and t_2 is obviously $s = v(t_1 - t_2)$. The problem with this simple formula is that it can only be used if the velocity of the particle was constant over the whole distance. To generalise this result for the case of variable velocity, we can use the concept of integration: (i) divide up the time into small intervals Δt_i, (ii) assuming that the velocity v_i within each interval is constant, calculate the distance $\Delta x_i \simeq v_i \Delta t_i$ the particle passed over each such small time interval ($\triangle x_i$ calculated that way is approximate for finite Δt_i), and (iii) then sum these distances up to get the whole distance passed over the whole time. In the limit of all $\Delta t_i \to 0$ (or infinite number of divisions of the whole time) the calculation becomes exact with the distance expressed via the integral:

$$s \simeq \sum_{i=1}^{n} v_i \Delta t_i \quad \Longrightarrow \quad s = \int_{t_1}^{t_2} v(t)dt. \tag{4.101}$$

For instance, consider a particle moving with the speed $v(t) = v_0 + at$, which linearly increases ($a > 0$) or decreases ($a < 0$) with time from the value of v_0 (at $t = 0$). In this case the distance s is

$$s = \int_0^t (v_0 + a\tau)\, d\tau = v_0 t + \frac{1}{2}at^2.$$

Here the acceleration $a = F/m$ is assumed to be constant, which happens when the force F acting on the particle of mass m is constant; e.g. during a vertical motion under the Earth's gravity or because of the friction during the sliding motion on a surface.

Another example is related to calculating work on a particle done by a variable force that moves it along the x axis: if the force F changes during a process, then, again, the work ΔW after a small distance Δx can be considered approximately constant yielding $\Delta W \simeq F(x)\Delta x$. To calculate the total work done by the force on the particle in displacing it from the initial position x_1 to its final position x_2, we integrate:

$$W = \int_{x_1}^{x_2} F dx. \tag{4.102}$$

Problem 4.77. *A body of mass m is taken from the surface of the Earth to infinity. The Earth's mass is M and radius R. The gravitational force acting between two bodies which are a distance x apart is $F = -GmM/x^2$. Show that the work needed to be done against the Earth's gravity is $W = -GmM/R$. What is the minimum possible velocity the rocket needs to develop in order to overcome the Earth's gravity?*

If two functions are equal for all values of x within some interval, $f_1(x) = f_2(x)$, then obviously their integrals are equal as well, $\int_a^b f_1(x)dx = \int_a^b f_2(x)dx$, where the points a and b lie within the interval. The passage from the former equality to the latter is usually described as "integrating the equality" or "integrating both sides of" $f_1(x) = f_2(x)$ with respect to x from a to b.

Problem 4.78. *Consider a particle of mass m moving along the x axis under the influence of the force $F(x)$. The Newtonian equation of motion is $\dot{p} = F(x)$, where $\dot{p}$ is the time derivative of the particle's momentum, p. Integrate both sides of this equation with respect to x between the initial, x_0, and final, x_1, points along a particle trajectory to show that*

$$\frac{p_1^2}{2m} - \frac{p_0^2}{2m} = \int_{x_0}^{x_1} F(x)dx.$$

Interpret your result. [*Hint*: express $\dot{p}$ via the derivative of p over x first.]

Another class of problems is related to the calculation of the mass of an object from its mass density. For linear objects such as a rod of length L oriented along the x axis and positioned between coordinates $x = 0$ and $x = L$, the linear density is defined as the derivative $\rho(x) = dm/dx$. Here $m(x)$ is the mass of a part of the rod between 0 and x. Correspondingly, $\Delta m \simeq \rho(x)\Delta x$ will be the mass of a small piece of the rod of length Δx. The total mass is calculated by summing up these contributions along the whole length of the rod and taking the limit $\Delta x \to 0$, i.e. by the integral

$$m = \int_{x_1}^{x_2} \rho(x)dx. \tag{4.103}$$

If we are interested in calculating the centre of mass of the rod, this can also be related to the appropriate definite integral. For a system of n particles of mass m_i arranged along the x axis at the positions x_i, the centre of mass is determined as

$$x_{\mathrm{cm}} = \frac{1}{m}\sum_{i=1}^{n} x_i m_i$$

with $m = \sum_{i=1}^{n} m_i$ being the total mass of all particles. For a continuous rod, we split it into small sections of length Δx and mass $\Delta m \simeq \rho(x)\Delta x$ that are positioned

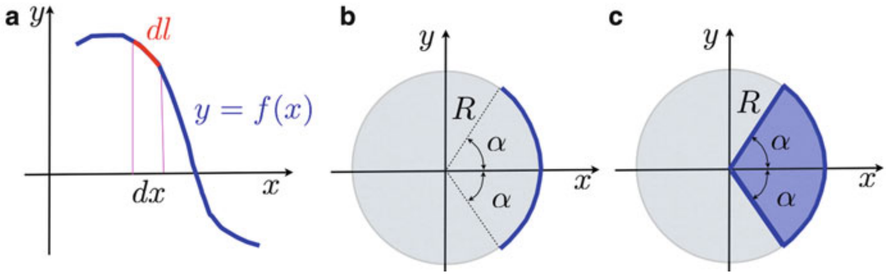


Fig. 4.14 To the calculation of the centre of mass of: (**a**) a *thin curved line* specified by the equation $y = f(x)$; (**b**) an arch of the angle 2α and radius R, as well as (**c**) a corresponding sector of the circle

between x and $x + \Delta x$. Each of these sections can be treated as a point mass for sufficiently short sections. Hence, using the above formula with substitutions $m_i \to \Delta m$ and $x_i \to x$ and taking the limit $\Delta x \to 0$, we arrive at the corresponding expression for the continuous rod:

$$x_{cm} = \frac{1}{m}\int_{x_1}^{x_2} x\rho(x)dx. \tag{4.104}$$

Let us consider calculating the centre of mass of a thin curved line specified on the plane by the equation $y = f(x)$, see Fig. 4.14(a). As the line is in two-dimensions, there will be two coordinates of the centre of mass, (x_{xm}, y_{cm}), which are determined by considering small lengths

$$dl = \sqrt{(dx)^2 + (dy)^2} = \sqrt{1 + (f'(x))^2}dx$$

along the line and then summing up all the contributions:

$$x_{cm} = \frac{1}{m}\int_{x_1}^{x_2} x\rho(x)\sqrt{1 + (f'(x))^2}dx \quad \text{and}$$
$$y_{cm} = \frac{1}{m}\int_{x_1}^{x_2} f(x)\rho(x)\sqrt{1 + (f'(x))^2}dx, \tag{4.105}$$

where m is the mass of the line and $\rho(x)$ is the line density of the rod at the point with coordinates $(x, f(x))$.

Problem 4.79. *Show that if the line is specified parametrically in the three-dimensional space as $x = x(t)$, $y = y(t)$ and $z = z(t)$ (or, simply, in the vector form, as $\mathbf{r} = \mathbf{r}(t)$), then the position $\mathbf{r}^{(cm)} = \left(r_\alpha^{(cm)}\right) = (x_{cm}, y_{cm}, z_{cm})$ of the centre of mass is given by*

$$r_\alpha^{(cm)} = \frac{1}{m}\int_{t_1}^{t_2} r_\alpha(t)\rho(t)\sqrt{(x'(t))^2 + (y'(t))^2 + (z'(t))^2}dt,$$

where $\alpha = x, y, z$ designates the three Cartesian components of the vectors and $\rho(t)$ is the corresponding line density.

Problem 4.80. *Show that the coordinates of the centre of mass of an arch of a unit linear density, radius R and angle 2α, shown in Fig. 4.14(b), is given by $(x_{cm}, 0)$, where $x_{cm} = R\sin\alpha/\alpha$.*

Problem 4.81. *Show that the coordinates of the centre of mass of a uniform sector of radius R and angle 2α, shown in Fig. 4.14(c), is given by $(x_{cm}, 0)$, where*

$$x_{cm} = \frac{2R\sin\alpha}{3\alpha}.$$

Problem 4.82. *Consider a cylinder of the uniform density ρ, radius R and height h, rotating around its symmetry axis with the angular velocity ω. Show that the kinetic energy of the cylinder is $E_{KE} = \pi\rho h R^4\omega^2/4$.* [*Hint*: particles of the cylinder positioned distance r from its symmetry axis has the velocity $v(r) = \omega r$.]

Problem 4.83. *Consider a circular disk of radius R placed in water vertically such that its centre is at the depth of $h > R$. If ρ is the water density, show that the total force exerted on any side surface of the disk is $F = \pi h\rho g R^2$, where g is the Earth's gravitational constant.*

Problem 4.84. *Generalise the result of the previous problem for the disk which is only partially vertically immersed into water ($h < R$). Show that in this case*

$$F = \rho g\left\{hR^2\left(\frac{\pi}{2} + \arcsin\frac{h}{R}\right) + h^2\sqrt{R^2 - h^2} + \frac{2}{3}\left(R^2 - h^2\right)^{3/2}\right\}.$$

However, the mass and the centre of mass of more complex two- and even three-dimensional objects are defined via their corresponding surface and volume densities. As such, these are much easier to consider later on after we learn a bit more necessary mathematics.

Gas Theory. An equation of the ideal gas reads $PV = nRT$, where P is pressure, V volume, T temperature, n the number of moles and R is the gas constant. In a process in which the volume of the gas changes from V_1 to V_2, the work done by the gas is $W = P(V_2 - V_1)$. Provided, of course, that the pressure is the same during the process in question. This happens in the isobaric process when the V changes due to a change in temperature, but the pressure remains the same. However, if the process is not isobaric, this formula is not longer applicable and one has to take into consideration explicitly the change of the pressure during the gas expansion (or contraction). Again, the solution to this problem can easily be found using the integration: divide the whole process into small volume changes ΔV during which the pressure can approximately be considered constant, calculate the work $\Delta W \simeq P\Delta V$ done during this small volume change, and then sum up all these changes and take the limit $\Delta V \to 0$, which results in the integral:

$$W_{\text{gas}} = \int_{V_1}^{V_2} P dV, \tag{4.106}$$

where the pressure P is a function of the volume, $P = P(V)$.

Problem 4.85. *Show that the work done by the gas in an adiabatic process, $PV^{\gamma} = Const$, where the adiabatic index $\gamma = C_P/C_V$ is given as the ratio of the respective specific heats, is*

$$W_{adiab} = \frac{P_1 V_1}{\gamma - 1}\left[1 - \left(\frac{V_1}{V_2}\right)^{\gamma - 1}\right].$$

Problem 4.86. *Consider a gas of molecules of mass m under influence of the Earth's gravity. The number density of the gas would be greatest near the Earth's surface and then it will get smaller and smaller as the distance z from the surface is increased:*

$$\rho(z) = \rho_0 \exp\left(-\frac{mgz}{k_B T}\right) = \rho_0 \exp\left(-\frac{Mgz}{RT}\right),$$

where k_B is the Bolztmann's constant, $R = k_B N_A$ the universal gas constant, N_A number of atoms in a mole of gas (Avogadro's number), $M = mN_A$ the mole mass, T absolute temperature and g the standard acceleration near the Earth's surface. Show that the weight $\mathcal{P} = m_{gas} g$ of the air in a vertical tube above the Earth's surface of cross section area A and height H, where m_{gas} is the total mass of air in the tube, is $\mathcal{P} = A(P_0 - P_H)$, where P_0 and P_H are the gas pressures at the bottom and top of the tube. Rewrite this equation as $\mathcal{P}/A + P_H = P_0$ and then interpret this result.

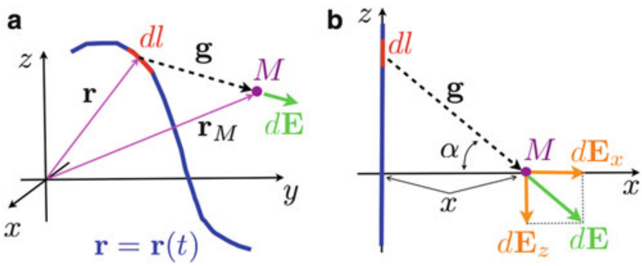


Fig. 4.15 For the calculation of the electrostatic field at point M of a charged line: (**a**) a general case of a line $\mathbf{r} = \mathbf{r}(t)$ specified parametrically, and (**b**) a vertical uniformly charged straight line with the field calculated distance x away from it

Electrostatics and Magnetism. Consider a continuous distribution of charge along a line, specified parametrically as $\mathbf{r} = \mathbf{r}(t)$ in a three-dimensional space. The linear charge density $\rho(t) = dq/dl$ may vary along the line. Here dq is amount of charge contained along a small length

$$dl = \sqrt{(x'(t))^2 + (y'(t))^2 + (z'(t))^2}dt$$

along the line, see Fig. 4.15(a). The electrostatic field $d\mathbf{E}$ at point M, whose position is given by the vector $\mathbf{r}_M$, is equal to $d\mathbf{E} = dq\mathbf{g}/g^3$, where $\mathbf{g} = \mathbf{r}_M - \mathbf{r}$ is the vector connecting dl and the point M. The total field is obtained by summing all contributions coming from all little fragments dl along the whole charged line, i.e. by integrating along the length of the line:

$$\begin{aligned} \mathbf{E}(\mathbf{r}_M) &= \int_{\text{line}} \frac{\mathbf{g}}{g^3} dq = \int_{\text{line}} \frac{\mathbf{g}}{g^3} \rho dl \\ &= \int_{t_0}^{t_1} \frac{\mathbf{r}_M - \mathbf{r}(t)}{|\mathbf{r}_M - \mathbf{r}(t)|^3} \rho(t) \sqrt{(x'(t))^2 + (y'(t))^2 + (z'(t))^2} dt. \end{aligned} \quad (4.107)$$

We have ignored here the irrelevant pre-factor $1/4\pi\epsilon_0$.

As an example of the application of this method, consider the field due to an infinite uniformly charged straight line distance x from it as shown in Fig. 4.15(b). Let the element $dl = dz$ of the line with the charge $dq = \rho dl = \rho dz$ be distance z from the point on the line which is closest to the point M. One can see that $z = g \sin\alpha = x \tan\alpha$. It is clear that the field components along the z axis from the two line elements dl symmetrically placed at both sides of the connecting line at z and $-z$ would cancel each other. Hence, after integrating, there will only be the component of the field along the x axis. Then, the vector $\mathbf{g} = (x, 0, -z)$ and the required component of the field E_x is:

$$E_x = \int_{-\infty}^{\infty} \frac{\rho x}{(x^2 + z^2)^{3/2}} dz = \left| \begin{array}{c} z = x \tan \alpha \\ dz = x d\alpha / \cos^2 \alpha \end{array} \right|$$

$$= \rho x \int_{-\pi/2}^{\pi/2} \frac{1}{(x^2 + x^2 \tan^2 \alpha)^{3/2}} \frac{x d\alpha}{\cos^2 \alpha}$$

$$= \frac{\rho}{x} \int_{-\pi/2}^{\pi/2} \frac{\cos^3 \alpha}{\cos^2 \alpha} d\alpha = \frac{\rho}{x} \int_{-\pi/2}^{\pi/2} \cos \alpha d\alpha = \frac{2\rho}{x},$$

i.e. the field decays with the distance x from the charged line as $1/x$.

Similarly one can consider the magnetic field $\mathbf{B}$ due to a curved wire given parametrically as $\mathbf{r} = \mathbf{r}(t)$ in a three-dimensional space, see Fig. 4.16(a). According to the Biot–Savart law, the magnetic field $d\mathbf{B}$ at point M due to directed element $d\mathbf{l} = \mathbf{n} dl$ ($\mathbf{n}$ is the unit vector directed along the current J), which is separated from the point M by the vector $\mathbf{g}$, is

$$d\mathbf{B} = \frac{J \left[d\mathbf{l} \times \mathbf{g} \right]}{g^3} = \frac{J \left[\mathbf{n} \times \mathbf{g} \right]}{g^3} dl$$

(up to an irrelevant pre-factor), so that, integrating along the whole length of the wire, we obtain for the field due to the entire wire:

$$\mathbf{B} \left(\mathbf{r}_M \right) = \int_{\text{line}} \frac{J \left[\mathbf{n} \times \mathbf{g} \right]}{g^3} dl = J \int_{t_0}^{t_1} \frac{\left[\mathbf{n}(t) \times \mathbf{g}(t) \right]}{g(t)^3} \sqrt{(x'(t))^2 + (y'(t))^2 + (z'(t))^2} dt. \tag{4.108}$$

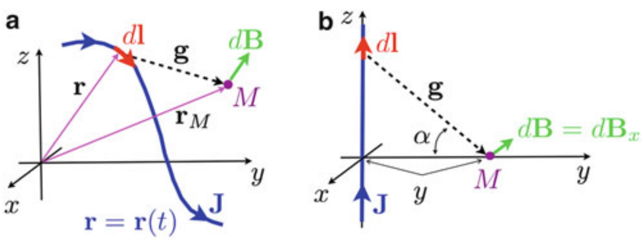


Fig. 4.16 For the calculation of the magnetic field $\mathbf{B}$ at point M due to the wire in which the current J is running: (**a**) a general case of a line $\mathbf{r} = \mathbf{r}(t)$ given parametrically, and (**b**) a vertical wire, and the observation point M which is the distance y from the wire

Problem 4.87. *Show that the magnetic field at distance y from the infinitely long wire with the current J, as shown in Fig. 4.16(b), is directed along the x axis and is equal to*

$$\mathbf{B} = -\frac{2J}{y}\mathbf{i},$$

where **i** *is the unit vector along the x axis.*

Problem 4.88. *Consider a single coil of wire of radius R placed within the $x-y$ plane as shown in Fig. 4.17(a). The current in the wire is equal to J. Show that the magnetic field* **B** *at a point M along the symmetry axis of the circle (i.e. along the z axis) at a distance z from it is*

$$\mathbf{B} = (0, 0, B_z) \quad \textit{with} \quad B_z = \frac{2\pi J R^2}{(R^2 + z^2)^{3/2}}.$$

[*Hint*: the vector $\mathbf{n} = \mathbf{e}_\phi$ is tangential to the circle as shown in Fig. 4.17(b); also, $\mathbf{g} = \mathbf{z} - \mathbf{r}$, where $\mathbf{r} = R\mathbf{e}_r$ ($\mathbf{e}_r$ being a unit vector along **r**), and $\mathbf{z} = z\mathbf{k}$ is the vector along the z axis. The two unit vectors $\mathbf{e}_r$ and $\mathbf{e}_\phi$ are perpendicular to each other; due to symmetry, only the z component of the field remains; hence, only the z component of the vector $[\mathbf{n} \times \mathbf{g}] = \left[\mathbf{e}_\phi \times (z\mathbf{k} - R\mathbf{e}_r)\right]$ is needed.]

Examples given above serve to illustrate the general concept of integration which is in summing up small contributions of some cumulative quantity such as length of a curve, surface area, electric field, etc. Physics presents numerous examples of such quantities, but in order to consider them we need to build up our mathematical muscle by considering, in the first instance, functions of many variables and multiple integration. Moreover, integration also appears when solving

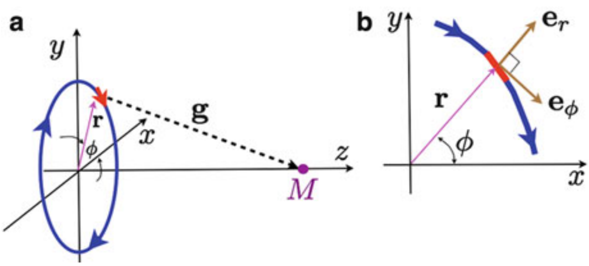


Fig. 4.17 For the calculation of the magnetic field at a point M along the z axis due to a single coil of wire placed within the $x-y$ plane and with its centre at the origin (**a**). The close-up of a piece of the wire within the $x-y$ plane is shown in (**b**). Orthogonal unit vectors $\mathbf{e}_r$ and $\mathbf{e}_\phi$ are along radial and tangential directions, respectively, to the circle of the wire

the so-called differential equations which connect derivatives of functions of interest and the functions themselves into a single (or a set of) equation(s). Since most physical laws are given via these kind of equations, one may say that solutions of equations in physics involve integration in a natural way. You may appreciate now that differentiation and integration, together with infinite series and differential equations (still yet to be considered), represent the basic language of the calculus used in physics.

4.7 Summary

As it should have become clear from the previous chapter, any "well-behaved" function written in an analytical form can be differentiated, however complex. This is not the case for integration: there is no magical general recipe to integrate an arbitrary function.

There are many methods to integrate specific classes of functions, some of which we considered in this chapter:

- the change of variable, and there are a number of different ones suited for particular types (classes) of functions;
- integration by parts, several of them may be needed;
- sometimes integration by parts establishes an equation for the integral being calculated;
- some integrals can be calculated by introducing parameters in the integrand; if the differentiation with respect to the parameter yields a known integral, then the required integral is obtained by differentiating it; not one but many differentiations are sometimes required.

More methods exist for solving integrals. For instance, indefinite integrals can be expanded into infinite series; some *definite integrals* can be calculated using more sophisticated methods such as integral transforms or complex calculus. Examples of indefinite integrals which cannot be expressed as a finite number of elementary functions (meaning: *solved*): $\int e^{\alpha x^2} dx$, $\int \cos x^2 dx$, $\int \frac{\sin x}{x} dx$, $\int \frac{1}{x} e^{\alpha x} dx$.

Many integrals have been studied and solved over last centuries, and several comprehensive tables of integrals exist which list these results. For example, the following book is an absolutely excellent reference: I.S. Gradshteyn and I.M. Ryzhik, *Table of Integrals, Series and Products,* Academic Press, 1980. It is a good idea to consult this and other books when needed. There are also computer programs designed to solve integrals, both indefinite and definite, such as Mathematica. Finally, it is a good idea to remember that definite integrals of almost any complexity can be calculated numerically.

Chapter 5
Functions of Many Variables: Differentiation

Mathematical ideas of the previous chapters, such as limits, derivatives and one-dimensional integrals, are the tools we developed for functions $y = f(x)$ of a single variable. These functions, however, present only particular cases. In practical applications functions of more than one variable are frequently encountered. For instance, a temperature of an extended object is a function of all its coordinates (x, y, z), and may also be a function of time, t, i.e. it depends on four variables. Electric and magnetic fields, **E** and **B**, in general will also depend on all three coordinates and the time.[1] Therefore, one has to generalise our definitions of the limits and derivatives for function of many variables. It is the purpose of this chapter to lie down the necessary principles of the theory of differentiation extending the foundations we developed so far to functions of many variables. In the next chapter the notion of integration will be extended to this case as well.

5.1 Specification of Functions of Many Variables

If a *unique* number z corresponds to a pair x and y of variables, defined within their respective intervals, then it is said that a function of two variables is defined. This correspondence is written either as $z = z(x, y)$ or $z = f(x, y)$. This is a direct way in which a function can be defined, e.g. by a formula like $z = \left(x^2 + y\right)^2$. Alternatively, a function of two variables can be defined by an equation $f(x, y, z) = 0$, a solution z of which for every pair of x and y provides the required correspondence (if proven unique, i.e. that there is one and only one value of z for every pair of the variables x and y). A function of two variables $z = z(x, y)$ can also be specified parametrically as

[1] In fact, in the latter case we have three functions, not one, for either of the fields, as each component of the field is an independent function of all these variables.

L. Kantorovich, *Mathematics for Natural Scientists*, Undergraduate Lecture Notes in Physics, DOI 10.1007/978-1-4939-2785-2_5

$x = x(u, v)$, $y = y(u, v)$ using *two* parameters u and v defined within their respective intervals. Functions of more than two variables are defined in a similar way.

When a function of a single variable is defined, $y = f(x)$, it can be represented as a graph of y vs. x, i.e. the function defines a line in a two-dimensional space (on a plane). Correspondingly, functions $z = f(x, y)$ of two variables x and y can be represented in the *three-dimensional space* by points with the coordinates (x, y, z). It is clear that a surface will be obtained this way: as the coordinates x and y change, only a single z value is found each time.

Before making a general statement on functions of arbitrary number of variables, let us consider here some examples of functions of two variables, which is the simplest case. More specifically, we shall dwell on *surfaces of the second order* which are generally given by the equation

$$a\left(x - x_0\right)^2 + b\left(y - y_0\right)^2 + c\left(z - z_0\right)^2 = 1.$$

The equation defines not one but two functions $z = f(x, y)$ given explicitly as

$$z = z_0 \pm \sqrt{\frac{1}{c}\left[1 - a\left(x - x_0\right)^2 - b\left(y - y_0\right)^2\right]}.$$

Recall that only a single value of the z can be put into correspondence to any pair of x and y, and hence there are two functions: one gives a single surface in the upper $z \geq z_0$ half-space, while the other in the lower $z \leq z_0$ half-space. Depending on the coefficients a, b and c, different types of surfaces are obtained. The point (x_0, y_0, z_0) defines the central point of the surfaces and can be chosen to coincide with the origin of the coordinate system, i.e. when $x_0 = y_0 = z_0 = 0$; their non-zero values simply correspond to a shift of the surfaces by the vector $\mathbf{r}_0 = (x_0, y_0, z_0)$. Hence, we shall assume in the following that $x_0 = y_0 = z_0 = 0$, and consider a simpler equation

$$ax^2 + by^2 + cz^2 = 1. \tag{5.1}$$

5.1.1 Sphere

If $a = b = c = 1/R^2$, we obtain the equation of a sphere of radius R and centred at the origin: $x^2 + y^2 + z^2 = R^2$. There are two surfaces here describing the upper $(z \geq 0)$ and the lower $(z \leq 0)$ hemispheres via $z = \pm\sqrt{R^2 - x^2 - y^2}$.

5.1.2 Ellipsoid

A more elaborate example is presented by a choice of still positive but non-equal coefficients via $a = 1/\alpha^2$, $b = 1/\beta^2$ and $c = 1/\gamma^2$, which yields the equation

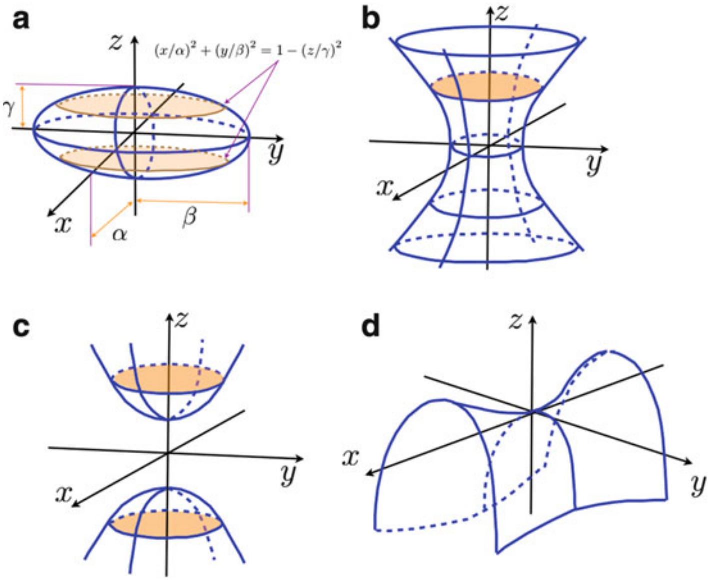


Fig. 5.1 Some of the second order surfaces: (**a**) ellipsoid; (**b**) one-pole (one sheet) hyperboloid; (**c**) two-pole (two sheets) hyperboloid; (**d**) saddle-like elliptic paraboloid

$$\left(\frac{x}{\alpha}\right)^2 + \left(\frac{y}{\beta}\right)^2 + \left(\frac{z}{\gamma}\right)^2 = 1 \tag{5.2}$$

of an ellipsoid shown in Fig. 5.1(a). To convince ourselves of this, let us consider cross sections of this shape with various planes perpendicular to the Cartesian axes. First of all we note that the absolute values of x/α, y/β and z/γ must be all less or equal to one, as otherwise the left-hand side in Eq. (5.2) would be larger than the right-hand side. Therefore, we conclude that $-\alpha \le x \le \alpha$, $-\beta \le y \le \beta$ and $-\gamma \le z \le \gamma$. Then, consider cross sections of the ellipsoid with planes $z = g$ (where $0 < g < \gamma$) which are parallel to the $x - y$ plane. In the cross section by either of these planes we obtain a line with the equation $(x/\alpha)^2 + (y/\beta)^2 = d^2$, where $d^2 = 1 - (g/\gamma)^2 > 0$ is some positive number. This line is an *ellipse*

$$\left(\frac{x}{A}\right)^2 + \left(\frac{y}{B}\right)^2 = 1$$

with the semi-axes $A = \alpha d$ and $B = \beta d$. As g increases from 0 to its maximum value of γ, the parameter d decreases from 1 to 0, i.e. the semi-axes decrease from their corresponding maximum values (equal to α and β, respectively) to zeros. Therefore, at $z = \gamma$ the ellipse is simply a point, but then, as the plane $z = g$ moves down closer to the $x - y$ plane, we observe the ellipse in the cross section whose size gradually increases as we approach the $x - y$ plane. As the crossing plane moves further towards negative values of g, the size of the ellipse in the cross section gradually decreases and turns into a point at $g = -\gamma$.

5.1.3 One-Pole (One Sheet) Hyperboloid

If now $a = 1/\alpha^2 > 0$, $b = 1/\beta^2 > 0$, but $c = -1/\gamma^2 < 0$, then we obtain the equation

$$\left(\frac{x}{\alpha}\right)^2 + \left(\frac{y}{\beta}\right)^2 - \left(\frac{z}{\gamma}\right)^2 = 1, \tag{5.3}$$

which corresponds to the figure shown in Fig. 5.1(b) called one-pole (or one sheet) hyperboloid.

5.1.4 Two-Pole (Two Sheet) Hyperboloid

When choosing -1 in the right-hand side of Eq. (5.3), a very different figure is obtained called two-pole (or two sheet) hyperboloid which is shown in Fig. 5.1(c). Its equation is

$$\left(\frac{x}{\alpha}\right)^2 + \left(\frac{y}{\beta}\right)^2 - \left(\frac{z}{\gamma}\right)^2 = -1. \tag{5.4}$$

Problem 5.1. *Using the method of sections with planes perpendicular to the coordinate axes, investigate the shape of the figure given by Eq. (5.3).*

Problem 5.2. *Investigate the shape of the figure given by Eq. (5.4). Are there any limitations for the values of z?*

5.1.5 Hyperbolic Paraboloid

This figure is described by the equation

$$\left(\frac{x}{\alpha}\right)^2 - \left(\frac{y}{\beta}\right)^2 = z. \tag{5.5}$$

Let us investigate its shape. Considering first crossing planes $x = g$ (which are perpendicular to the x axis), we obtain the lines $z = -y^2/\beta^2 + (g/\alpha)^2$, which are parabolas facing downwards; as g increases from zero in either the positive or negative direction, the top of the parabolas lifts up. Similarly, cutting with planes $y = g$ (which are perpendicular to the y axis) results in parabolas $z = x^2/\alpha^2 - (g/\beta)^2$ which face upwards; increasing the value of g from zero results in the parabolas shifting downwards. Finally, planes $z = g > 0$ (parallel to the $x - y$ plane) result in hyperbolas $(x/A)^2 - (y/B)^2 = 1$ with $A = \alpha\sqrt{g}$ and $B = \beta\sqrt{g}$. When $z = -g < 0$, then hyperbolas $-(x/A)^2 + (y/B)^2 = 1$ are obtained which are rotated by 90° with respect to the first set. The obtained figure has a peculiar shape of a *saddle* shown in Fig. 5.1(d): the centre of the coordinate system $x = y = z = 0$ is a minimum along the x axis direction, but it is a maximum when one moves along the perpendicular direction. This point is usually called the saddle point or, in physics, the *transition state* (TS).

This situation is very common in many physical processes: the total energy E of a system is a function of atomic coordinates, and may contain many minima.[2] Each minimum corresponds to a stable configuration of the system. If we are interested in transitions between different states, i.e. between two neighbouring minima, then we have to consider the minimum energy path connecting the two minima. The energy barrier Δ along the path would determine the rate with which such a transition can happen as the rate is approximately proportional to $\exp(-\Delta/k_BT)$, where k_B is the Boltzmann's constant and T the absolute temperature. Sometimes there is only a single so-called "reaction" coordinate connecting the two minima along the minimum energy path, and going along this direction requires overcoming the energy barrier, i.e. along this direction one has to move over a maximum. At the same time, since this is the minimum energy path, each point on it is a minimum with respect to the rest of the coordinates of the system. In other words, the crossing point is nothing but a saddle on the complicated potential energy surface E considered as a function of its all coordinates.

The simplest example of this situation can be found in atomic diffusion on a metal surface: the energy $E(x, y, z)$ of an adatom on the surface can be considered merely as a function of its two Cartesian coordinates, x and y, determining its lateral position, and of its height z above the surface as schematically shown in Fig. 5.2. Assuming the simplest square lattice of the metal surface with the distance between nearest atoms to be a, and the adatom occupying a hollow site with distance z_0 above the surface as shown in the figure, the transition between two such sites $(0, 0, z_0)$ and $(a, 0, z_0)$ (which are equivalent) would require the adatom to move along the x axis overcoming an energy barrier Δ at the transition point $(a/2, 0, z_{TS})$ which is exactly half way through; it corresponds to the bridge position of the adatom. Here z_{TS} is the height of the adatom at the transition state, and the x coordinate appears to serve as the reaction coordinate. In practice, however, a single Cartesian coordinate rarely serves as the reaction coordinate for the transition and is instead specified in a more complicated way.

[2] We shall define more accurately minimum (and maximum) points of functions of many variables later on as points where the function has the lowest (the highest) values within a finite vicinity of their arguments.

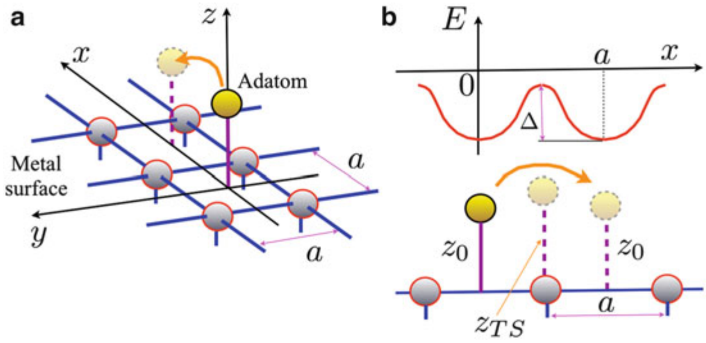


Fig. 5.2 Transition of an adatom between two hollow sites on a simple cubic lattice. The distance between the nearest atoms on the surface is a. (**a**) Three-dimensional view in which the initial and final (*dimmed*) positions of the adatom are shown. (**b**) Minimum energy path along the x direction with the energy barrier Δ between the two positions indicated. The x coordinate serves as the reaction coordinate here

5.2 Limit and Continuity of a Function of Several Variables

Let us recall how we defined the limit of a function $y = f(x)$ of a single variable at a point x_0 (Sect. 2.4.1.1). Actually, we had two definitions which are equivalent. One definition was based on building a sequence of points $\{x_n, n = 1, 2, 3, \ldots\} = \{x_1, x_2, x_3, \ldots, x_n, \ldots\}$ which converges to the point x_0; this sequence generates the corresponding functional sequence of values $\{f(x_i), i = 1, 2, 3, \ldots\}$ of $f(x)$, and the limit of the function as $x \to x_0$ is the limit of this functional sequence. Note that no matter which particular sequence is chosen, as long as it converges to the same value of x_0, the *same limit* of the functional sequence is required for the limit to exist. Another definition was based on the "$\epsilon - \delta$" language: we called A the limit of $f(x)$ when $x \to x_0$ if for any $\epsilon > 0$ one can always find a positive $\delta = \delta(\epsilon)$, such that for any x within the distance δ from x_0 (i.e. for any x satisfying $|x - x_0| < \delta$) the number A differs from $f(x)$ by no more than ϵ, i.e. $|f(x) - A| < \epsilon$ is implied. Note that one might take the values of x either below or above x_0, still the same limit should be achieved.

Both these definitions are immediately generalised to the case of a functions of more than one variable if we note that a function $f(x_1, x_2, \ldots, x_N)$ of N variables can formally be considered as a function of an N-dimensional vector $\mathbf{x} = (x_1, \ldots, x_N)$, i.e. as $f(\mathbf{x})$, and hence we are interested in the definition of the limit of $f(\mathbf{x})$ when $\mathbf{x} \to \mathbf{x}_0$, with $\mathbf{x}_0$ being some other vector. The generalisation is simply done by replacing a single variable x in the definitions corresponding to the cases of a single variable function, with the N-dimensional vector $\mathbf{x} = (x_1, \ldots, x_N)$; correspondingly, the distance $|x - x_0|$ between the points x and x_0 in the one-dimensional space is replaced with the distance (1.48) between vectors $\mathbf{x}$ and $\mathbf{x}_0$, i.e. with $d(\mathbf{x}, \mathbf{x}_0) = |\mathbf{x} - \mathbf{x}_0|$. With these substitutions, either definition of the limit of the function $f(\mathbf{x})$ when $\mathbf{x} \to \mathbf{x}_0$ sounds exactly the same.

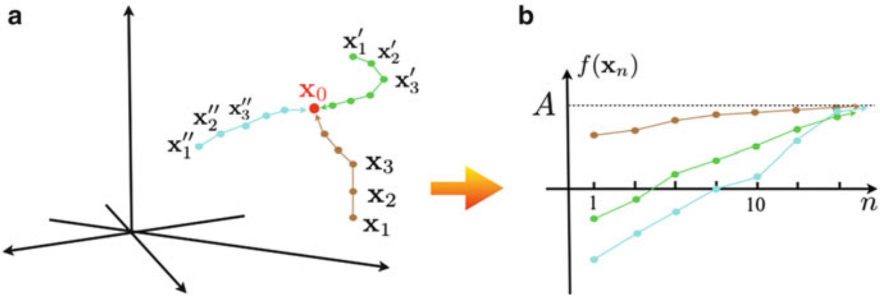


Fig. 5.3 This illustration shows that for a function $f(\mathbf{x})$, which has a well-defined limit when $\mathbf{x} \to \mathbf{x}_0$, any path to the limit point $\mathbf{x}_0$ made out of a sequence of points in the N-dimensional space, as shown in (**a**), results in the same limit A for the functional sequence presented schematically in (**b**)

The first definition—via a sequence—has a simple geometrical interpretation: points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \ldots$ in the sequence form a line in the N-dimensional space which converges to a single point $\mathbf{x}_0$. The limit would exist if and only if any *path* towards $\mathbf{x}_0$, as depicted in Fig. 5.3, generates the corresponding functional sequence converging to the same limit A.

This condition is not always satisfied. Consider as our first example the limit of the function of two variables $f(x, y) = xy/(x + y)$ at the point $(x, y) = (0, 0)$. We shall do this by constraining the y using $y = kx$ with some constant k when $x \to 0$. By changing the constant k different paths towards the centre of the coordinate system are taken; however, a well-defined limit exists for any value of k. Indeed,

$$\lim_{x\to 0, y\to 0} \frac{xy}{x+y} = \lim_{x\to 0} \frac{kx^2}{x(1+k)} = \frac{k}{1+k} \lim_{x\to 0} x = 0.$$

Consider now, using the same set of paths, the function $f(x, y) = xy/\left(x^2 + y^2\right)$. For this function the limit does depend on the value of the constant k:

$$\lim_{x\to 0, y\to 0} \frac{xy}{x^2+y^2} = \lim_{x\to 0} \frac{kx^2}{x^2(1+k^2)} = \frac{k}{1+k^2}.$$

Therefore, different limiting values can be obtained using various values of the k and hence, we must conclude, the limit of this function at the point $(0, 0)$ does *not* exist.

Problem 5.3. *Investigate whether the function* $f(x, y) = \arctan(x/y)$ *has well-defined limits at the following* (x, y) *points: (a)* $(1, 1)$*; (b)* $(0, 1)$*; (c)* $(1, 0)$*; (d)* $(0, 0)$. *Calculate the limits when they do exist.* [*Answers*: (a) $\pi/4$; (b) 0; (c,d) do not exist.]

Most of the limit theorems discussed in Sect. 2.4.2 are immediately generalised to the case of the functions of many variables, e.g. if the limit exists it is unique and that one can manipulate limits algebraically.

The notion of *continuity* of a function of many variables is introduced similarly to the case of a function of a single variable (see Sect. 2.4.3): namely, the function $f(\mathbf{x})$ is continuous at the point $\mathbf{x}_0$ if its limit at $\mathbf{x} \to \mathbf{x}_0$ coincides with the value of the function at the point $\mathbf{x}_0$, i.e.

$$\lim_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x}) = f(\mathbf{x}_0). \tag{5.6}$$

Points where the function $f(\mathbf{x})$ is not continuous are called points of discontinuity. For continuous functions the change Δy of the function $y = f(\mathbf{x})$ by going from the point $\mathbf{x}$ to the point $\mathbf{x}' = \mathbf{x} + \Delta\mathbf{x}$ should approach zero when the point $\mathbf{x}'$ approaches $\mathbf{x}$, i.e. when $\Delta\mathbf{x} \to \mathbf{0}$:

$$\lim_{\Delta\mathbf{x}\to\mathbf{0}} \Delta y = \lim_{\Delta\mathbf{x}\to\mathbf{0}} \left[f\left(\mathbf{x} + \Delta\mathbf{x}\right) - f(\mathbf{x})\right] = 0. \tag{5.7}$$

Many properties of continuous functions of several variables are also analogous to those of functions of a single variable. For instance, if a function $f(\mathbf{x})$ is continuous within some region in the N-dimensional space (where $\mathbf{x}$ is defined) and takes there two values A and B, then it takes in this region all values between A and B, i.e. for any C lying between A and B there is a point $\mathbf{x}_C$ such that $f(\mathbf{x}_C) = C$. In particular, if $A < 0$ and $B > 0$, then there exists at least one point $\mathbf{x}_0$ where the function $f(\mathbf{x}_0) = 0$.

Problem 5.4. *Consider the function $z = f(x, y) = xy^2$. Show that the change of the function Δz corresponding to the change Δx and Δy of its variables can be written in the form:*

$$\Delta z = y^2\Delta x + 2xy\Delta y + \alpha\Delta x + \beta\Delta y,$$

where both α and β tend to zero as $(\Delta x, \Delta y) \to \mathbf{0}$. Note that the choice of α and β here is not unique.

5.3 Partial Derivatives: Differentiability

When in Sect. 3.1 we considered a function of a single variable, $y = f(x)$, we defined its derivative as the ratio $\Delta y/\Delta x$ of the change of the function Δy due to the change Δx of its variable, taken at the limit $\Delta x \to 0$. This definition cannot be directly generalised to the case of functions of many variables as $\Delta\mathbf{x}$ in this case is an N-dimensional vector.

At the same time, we can easily turn a function of many variables into a function of just one variable by "freezing" all other variables. Indeed, let us fix all variables but x_1; one can then consider the change of the function $y = f(x_1, x_2, \ldots, x_N) = f(\mathbf{x})$ due to its single variable x_1, the so-called partial change,

$$\Delta_1 y = f(x_1 + \Delta x_1, x_2, \ldots, x_N) - f(x_1, x_2, \ldots, x_N),$$

and then define the "usual" derivative

$$\lim_{\Delta x_1 \to 0} \frac{\Delta_1 y}{\Delta x_1} = \left.\frac{dy}{dx}\right|_{x_2,\ldots,x_N}.$$

The latter notation, consisting of the variables written as a subscript to the vertical line, explicitly specifies that the variables $x_2, x_3, \ldots, x_N$ are fixed. This derivative is called *partial derivative* of the function y with respect to its first variable x_1, and is denoted $\partial y/\partial x_1$. It shows the rate of change of the function $y = f(\mathbf{x})$ due to its single variable x_1.

Similarly, one can consider the partial change $\Delta_i y$ of the function with respect to any variable x_i, where $i = 1, 2, \ldots, N$. Correspondingly, dividing the partial change of the function $\Delta_i y$ by the change of the corresponding variable, Δx_i, and taking the limit, we arrive at the partial derivative

$$\frac{\partial y}{\partial x_i} = \lim_{\Delta x_i \to 0} \frac{\Delta_i y}{\Delta x_i} = \left.\frac{dy}{dx}\right|_{x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_N}$$

with respect to the variable x_i. The partial derivative $\partial y/\partial x_i$ is sometimes also denoted as y'_{x_i} or $\partial_x y$.

Problem 5.5. *Calculate all partial derivatives of the functions:*

$$f(x,y) = \sin x \sin y \quad \left[\textit{Answer: } \frac{\partial f}{\partial x} = \cos x \sin y,\ \frac{\partial f}{\partial y} = \sin x \cos y\right];$$

$$f(x,y) = e^{x^2+y^2} \quad \left[\textit{Answer: } \frac{\partial f}{\partial x} = 2xe^{x^2+y^2},\ \frac{\partial f}{\partial y} = 2ye^{x^2+y^2}\right];$$

$$f(x,y,z) = \frac{1}{\sqrt{x^2+y^2+z^2}}$$

$$\left[\textit{Answer: } \frac{\partial f}{\partial x} = -\frac{x}{r^3},\ \frac{\partial f}{\partial y} = -\frac{y}{r^3},\ \frac{\partial f}{\partial z} = -\frac{z}{r^3},\ r = \sqrt{x^2+y^2+z^2}\right].$$

Problem 5.6. *The three components F_x, F_y and F_z of the force acting on a particle in an external field (or potential) $U(\mathbf{r}) = U(x, y, z)$ are obtained via partial derivatives of the minus potential, i.e. $F_x = -\partial U/\partial x$ and similarly for the other two components. A satellite of mass M which is orbiting the Earth of mass M_0 experiences the potential field $U(\mathbf{r}) = -\alpha/|\mathbf{r} - \mathbf{r}_0|$, where $\mathbf{r}_0$ is the position vector of the Earth. Determine the force acting on the satellite.*

Partial derivatives can be considered as derivatives along the directions of the Cartesian axes; it is also possible to define a derivative along an arbitrary direction; this will be done later on in Sect. 5.8.

In Sect. 3.2 we obtained a general expression for a change Δy of the function $y = f(x)$ of a single variable and related this to the differentiability of the function. There exists a generalisation of this result for a function of many variables. For simplicity of notations, we shall consider the case of a function of two variables, $z = f(x, y)$; a more general case is straightforward.

We say that the function $z = f(x, y)$ is differentiable at (x, y) if its change Δz due to the changes Δx and Δy of its variables can be written, similarly to Eq. (3.8), as:

$$\Delta z = A\Delta x + B\Delta y + \alpha\Delta x + \beta\Delta y, \tag{5.8}$$

where α and β tend to zero as both Δx and Δy tend to zero, while A and B depend only on the point (x, y), but not on Δx and Δy. This formula states that the first two terms in the right-hand side are of the first order with respect to the changes of the variables, Δx and Δy, while the last two terms are at least of the second order (cf. Problem 5.4).

Problem 5.7. *Consider the function $z = x^2y + xy^2$ and derive expressions for A, B, α and β in the above formula (5.8). Check explicitly that both α and β tend to zero as both Δx and Δy tend to zero.*

It is obvious from this definition that if the function is differentiable, i.e. if its change can be recast in the form (5.8), then $\Delta z \to 0$ when Δx and Δy tend to zero. I.e. the function $z = f(x, y)$ is continuous. It is also clear that both partial derivatives exist at the point (x, y). Indeed,

$$\begin{aligned}\frac{\partial z}{\partial x} &= \lim_{\Delta x\to 0} \left.\frac{\Delta z}{\Delta x}\right|_{\Delta y=0} = \lim_{\Delta x\to 0}\left[A + B\frac{\Delta y}{\Delta x} + \alpha + \beta\frac{\Delta y}{\Delta x}\right]_{\Delta y=0} \\ &= \lim_{\Delta x\to 0}(A+\alpha)_{\Delta y=0} = A + \lim_{\Delta x\to 0}\alpha|_{\Delta y=0} = A,\end{aligned}$$

and similarly

$$\frac{\partial z}{\partial y}=\lim_{\Delta y\to 0}\left.\frac{\Delta z}{\Delta y}\right|_{\Delta x=0}=\lim_{\Delta y\to 0}\left[A\frac{\Delta x}{\Delta y}+B+\alpha\frac{\Delta x}{\Delta y}+\beta\right]_{\Delta x=0}=B+\lim_{\Delta y\to 0}\beta|_{\Delta x=0}=B,$$

where, while calculating the partial derivative, we kept the other variable constant, i.e. considered its change to be equal to zero *before* taking the limit. Therefore, both partial derivatives exist and hence the change of the function (5.8) can be written in a more detailed form as

$$\Delta z=\frac{\partial z}{\partial x}\Delta x+\frac{\partial z}{\partial y}\Delta y+\alpha\Delta x+\beta\Delta y. \tag{5.9}$$

The first two terms in the above formula can be used to estimate numerically the change of the function due to small changes Δx, Δy of its arguments.

In a general case of a function $y=f(\mathbf{x})$ of N variables $\mathbf{x}=(x_1,\ldots,x_N)$ the previous formula is straightforwardly generalised as follows:

$$\Delta y=y\left(\mathbf{x}+\Delta\mathbf{x}\right)-y(\mathbf{x})=\sum_{i=1}^{N}\left(\frac{\partial y}{\partial x_i}\Delta x_i+\alpha_i\Delta x_i\right), \tag{5.10}$$

where all α_i tend to zero when the vector $\Delta\mathbf{x}\to\mathbf{0}$.

Importantly, existence of the partial derivatives does not yet guarantee the differentiability of the function, i.e. the reverse statement may not be true. For instance, the function $z=\sqrt{x^2+y^2}$ is continuous at the point $(0,0)$, however its partial derivatives do not exist there: indeed, when, for instance, we are interested in the $\partial z/\partial x$, then at this point $y=0$ and hence $z=\sqrt{x^2}=|x|$, which does not have the derivative at $x=0$ (since the derivatives from the left and right are different). Since the partial derivatives do not exist at this point, the function is not differentiable there. Also, similarly to the case of functions of a single variable, differentiability is a stronger condition than continuity: continuity of a function at a point is not yet sufficient for it to be differentiable there.

Theorem 5.1 (Sufficient Condition of Differentiability). *For the function $z=f(x,y)$ to be differentiable at the point (x,y) it is sufficient that z has continuous partial derivatives there.*

Proof. Consider a total change $\Delta z=f(x+\Delta x,y+\Delta y)-f(x,y)$ of the function z due to changes Δx and Δy of its both variables, which we shall rewrite as follows:

$$\begin{aligned}\Delta z&=[f\left(x+\Delta x,y+\Delta y\right)-f\left(x,y+\Delta y\right)]+[f\left(x,y+\Delta y\right)-f\left(x,y\right)]\\&=\left.\Delta_x z\right|_{y+\Delta y}+\left.\Delta_y z\right|_{x}.\end{aligned} \tag{5.11}$$

Here the first term is the partial change $\Delta_x z$ of the function z with its second variable kept constant at the value of $y + \Delta y$, while the second term gives the partial change $\Delta_y z$ with the first variable kept constant at x. Since in both cases we effectively deal with functions of a single variable (as the other variable is frozen), and since we assumed that partial derivatives exist, we can apply the Lagrange formula (3.64) for the corresponding partial changes:

$$\Delta_x z\big|_{y+\Delta y} = f_x'\left(x + \vartheta\Delta x, y + \Delta y\right)\Delta x, \quad \Delta_y z\big|_x = f_y'\left(x, y + \theta\Delta y\right)\Delta y, \tag{5.12}$$

where ϑ and θ are both somewhere between zero and one. Note that derivatives above are, in fact, the partial derivatives, and we have used simplified notations for them: f_x' or f_y'. (The subscript indicates which variable is being differentiated upon.) As noted above, these kinds of simplified notations are frequently met in the literature and we shall be using them frequently as well.

Since the partial derivatives are continuous, we can write:

$$f_x'\left(x+\vartheta\Delta x, y+\Delta y\right) = f_x'(x,y)+\alpha=\frac{\partial z}{\partial x}+\alpha \quad \text{and}$$

$$f_y'\left(x, y+\theta\Delta y\right) = f_y'(x,y)+\beta=\frac{\partial z}{\partial y}+\beta,$$

where both functions, α and β, tend to zero when Δx or Δy tend to zero. Substituting the above results into Eqs. (5.12) and (5.11) leads exactly to the required form (5.9), proving the differentiability of the function z. **Q.E.D.**

Hence, if the partial derivatives exist and are continuous, the function of two variables is differentiable.

Problem 5.8. *Prove the above theorem for a function of three variables, and then for any number of variables.*

Above we have introduced first order partial derivatives. However, a partial derivative of a function is a function itself and hence may also be differentiated. Taking the partial derivative with respect to the variable x of the x-partial derivative yields the second order partial derivative:

$$\frac{\partial}{\partial x}\left(\frac{\partial z}{\partial x}\right) = \frac{\partial^2 z}{\partial x^2} = f_{xx}''.$$

However, if we take a partial derivative with respect to y of the partial derivative $\partial z/\partial x$, or vise versa, two mixed second order partial derivatives are obtained:

$$\frac{\partial}{\partial y}\left(\frac{\partial z}{\partial x}\right) = \frac{\partial^2 z}{\partial y\partial x} = f_{yx}'' \quad \text{and} \quad \frac{\partial}{\partial x}\left(\frac{\partial z}{\partial y}\right) = \frac{\partial^2 z}{\partial x\partial y} = f_{xy}''.$$

If this process is continued, various partial derivatives can be obtained.

As an example, consider $z(x, y) = \exp[\alpha(x - y)]$. We have:

$$z'_x = \alpha e^{\alpha(x-y)};\; z'_y = -\alpha e^{\alpha(x-y)};\; z''_{xy} = \frac{\partial}{\partial x}\left(z'_y\right) = -\alpha^2 e^{\alpha(x-y)};$$

$$z''_{yx} = \frac{\partial}{\partial y}\left(z'_x\right) = -\alpha^2 e^{\alpha(x-y)};$$

$$z''_{xx}=\frac{\partial}{\partial x}\left(z'_x\right)=\alpha^2 e^{\alpha(x-y)};\; z''_{yy}=\frac{\partial}{\partial y}\left(z'_y\right) = \alpha^2 e^{\alpha(x-y)};\; z'''_{xyy} = \frac{\partial}{\partial x}\left(z''_{yy}\right) = \alpha^3 e^{\alpha(x-y)};$$

$$z'''_{yxy} = \frac{\partial}{\partial y}\left(z''_{xy}\right) = \alpha^3 e^{\alpha(x-y)};\; z'''_{yyx} = \frac{\partial}{\partial y}\left(z''_{yx}\right) = \alpha^3 e^{\alpha(x-y)}, \quad \text{etc.}$$

Problem 5.9. *Calculate all non-zero partial derivatives of the function* $z(x, y) = x^3y^2 + x^2y^3$.

Problem 5.10. *In a one-dimensional (along the x axis) Brownian motion the probability to find a particle between x and x + dx at time t is given by*

$$w(x,t) = \frac{1}{\sqrt{4\pi Dt}} e^{-x^2/4Dt}, \tag{5.13}$$

where D is the diffusion constant. Check by direct differentiation that this function satisfies the following (the so-called partial differential*) equation:*

$$\frac{1}{D}\frac{\partial w}{\partial t} = \frac{\partial^2 w}{\partial x^2}. \tag{5.14}$$

This is called the diffusion equation.

It follows from the example and the problem, that the partial derivatives can be taken in any order, the result does not depend on it. The validity of this important statement follows from the following theorem which we shall formulate and prove for the case of a function of two variables and for the derivatives of the second order.

Theorem 5.2 (On Mixed Derivatives). *Consider a function* $z = f(x, y)$. *If its mixed second order partial derivatives* f''_{xy} *and* f''_{yx} *exist and are continuous in some vicinity of a point* (x, y), *then they are equal at that point:*

$$f''_{xy} = f''_{yx}. \tag{5.15}$$

Proof. The first partial derivative f'_x requires calculation of the partial change of the function

$$\Delta_x f = f(x+\Delta x, y) - f(x,y) = \phi(y),$$

which we can formally consider as a function $\phi(y)$ of the other variable y. The mixed derivative f''_{yx} will then require calculating the partial difference of the function $\phi(y)$. I.e. the total change needed for the mixed derivative will be

$$\begin{aligned}\Delta_{yx} f = \Delta_y \phi = \phi(y+\Delta y) - \phi(y) = &[f(x+\Delta x, y+\Delta y) - f(x, y+\Delta y)] \\ &- [f(x+\Delta x, y) - f(x,y)] \, .\end{aligned} \tag{5.16}$$

This expression, after dividing by $\Delta x \Delta y$ and taking the limits $\Delta x \to 0$ and $\Delta y \to 0$, results in the mixed derivative f''_{yx}.

Similarly, let us construct the full change $\Delta_{xy} f$ of the function f required for the calculation of the mixed derivative f''_{xy}. We first calculate the change

$$\Delta_y f = f(x, y+\Delta y) - f(x,y) = \varphi(x),$$

and then the additional change with respect to the other variable:

$$\begin{aligned}\Delta_{xy} f = \Delta_x \varphi = \varphi(x+\Delta x) - \varphi(x) = &[f(x+\Delta x, y+\Delta y) - f(x+\Delta x, y)] \\ &- [f(x, y+\Delta y) - f(x,y)] \, .\end{aligned} \tag{5.17}$$

Dividing this expression by $\Delta x \Delta y$ and taking the limits would give the other mixed derivative f''_{xy}. One can see now that the two expressions (5.16) and (5.17) are exactly the same, $\Delta_{xy} f = \Delta_{yx} f$ (although with terms arranged in a different order), which already suggests that the mixed derivatives must be the same.

To proceed more rigorously, we use the Lagrange formula (3.64) for each of the changes of the function:

$$\phi(y) = \Delta_x f = f'_x(x+\theta_1 \Delta x, y) \quad \text{and}$$
$$\Delta_{yx} f = \Delta_y \phi = \phi'_y(y+\theta_2 \Delta y) = f''_{yx}(x+\theta_1 \Delta x, y+\theta_2 \Delta y)$$

and, similarly,

$$\varphi(x) = \Delta_y f = f'_y(x, y+\vartheta_1 \Delta y) \quad \text{and}$$
$$\Delta_{xy} f = \Delta_x \varphi = \varphi'_x(x+\vartheta_2 \Delta x) = f''_{xy}(x+\vartheta_2 \Delta x, y+\vartheta_1 \Delta y),$$

where numbers ϑ_1, ϑ_2, θ_1 and θ_2 are each between 0 and 1. Note that the application of the Lagrange formula for each individual step in either of the expressions above is perfectly legitimate since each step corresponds to the change of the function with respect to a different variable.

Since both changes of the function $f(x, y)$ are the same, we can write:

$$f''_{xy}(x + \vartheta_2 \Delta x, y + \vartheta_1 \Delta y) = f''_{yx}(x + \theta_1 \Delta x, y + \theta_2 \Delta y).$$

Both mixed derivatives are continuous functions, hence taking the limits $\Delta x \to 0$ and $\Delta y \to 0$ yields

$$\lim_{\Delta x \to 0, \Delta y \to 0} f''_{yx}(x + \theta_1 \Delta x, y + \theta_2 \Delta y) = f''_{yx}(x, y),$$

$$\lim_{\Delta x \to 0, \Delta y \to 0} f''_{xy}(x + \vartheta_2 \Delta x, y + \vartheta_1 \Delta y) = f''_{xy}(x, y),$$

which establishes the required result, $f''_{xy}(x, y) = f''_{yx}(x, y)$, **Q.E.D.**

Problem 5.11. *Generalise this theorem for a function of $N > 2$ variables.* [*Hint*: try to make use of the already proven theorem.]

Problem 5.12. *Generalise this theorem for mixed derivatives of any order.*

5.4 A Surface Normal. Tangent Plane

As an application of the developed formalism, let us derive an expression for the normal $\mathbf{n}$ to a surface $z = z(x, y)$ at point $M_0(x_0, y_0, z_0)$, see Fig. 5.4(a). To this end, we shall consider lines lying within the surface. These lines are obtained by keeping either x or y at their values x_0 or y_0, respectively. The line DC is obtained by changing x and keeping $y = y_0$, its equation is $z = z(x, y_0)$. Similarly, the line AB with equation $z = z(x_0, y)$ is obtained by keeping $x = x_0$ and changing y. Both lines cross exactly at the point M_0.

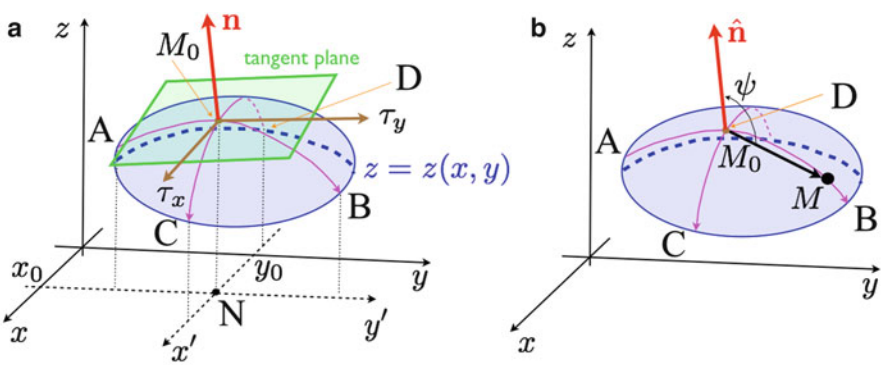


Fig. 5.4 To the derivation of the normal (*red*) to the surface (*blue*), $z = z(x, y)$, at point M_0. The plane tangential to the surface at point M_0 is shown in *light green*

Any point on our surface can be represented by the vector $\mathbf{r} = x\mathbf{i} + y\mathbf{j} + z(x, y)\mathbf{k}$. Consider a point M (with coordinates x_0, y, $z\,(x_0, y)$) along the AB curve ($x = x_0$, y is variable), see Fig. 5.4(b). The vector $\overrightarrow{M_0M}$ makes an angle ψ with the normal $\mathbf{n}$. As the point M approaches the point M_0, the angle ψ must approach the value of $\pi/2$. The cosine of ψ can be obtained by considering the dot product of vectors $\mathbf{n} = \left(n_x, n_y, n_z\right)$ and $\overrightarrow{M_0M} = (0, \Delta y, \Delta z)$, where $\Delta y = y - y_0$ and $\Delta z = z\,(x_0, y) - z\,(x_0, y_0)$:

$$\cos\psi = \frac{\overrightarrow{M_0M}\cdot\mathbf{n}}{\left|\overrightarrow{M_0M}\right|\,|\mathbf{n}|} = \frac{n_y\Delta y + n_z\Delta z}{|\mathbf{n}|\sqrt{\Delta y^2 + \Delta z^2}} = \frac{n_y + n_z\frac{\Delta z}{\Delta y}}{|\mathbf{n}|\sqrt{1 + \left(\frac{\Delta z}{\Delta y}\right)^2}}.$$

In the limit when the point M approaches M_0 (or, which is equivalent, $\Delta y \to 0$), the angle $\psi \to \pi/2$, i.e. $\cos\psi \to 0$. Hence,

$$\lim_{\Delta y\to 0}\frac{n_y + n_z\frac{\Delta z}{\Delta y}}{|\mathbf{n}|\sqrt{1 + \left(\frac{\Delta z}{\Delta y}\right)^2}} = \frac{n_y + n_z\lim_{\Delta y\to 0}\frac{\Delta z}{\Delta y}}{|\mathbf{n}|\sqrt{1 + \left(\lim_{\Delta y\to 0}\frac{\Delta z}{\Delta y}\right)^2}} = \frac{n_y + n_z\frac{\partial z}{\partial y}}{|\mathbf{n}|\sqrt{1 + \left(\frac{\partial z}{\partial y}\right)^2}} = 0,$$

which means that $n_y + n_z\frac{\partial z}{\partial y} = 0$. Similarly, considering a point M somewhere on the DC line, when $y = y_0$ and x is allowed to change, we obtain an equation $n_x + n_z\frac{\partial z}{\partial x} = 0$. Assuming $n_z \neq 0$ (which corresponds to the surface $z = z(x, y)$ which is not parallel to the z axis anywhere in the vicinity of the given point), we obtain for the normal:

$$\mathbf{n} = \left(n_x, n_y, n_z\right) = \left(-n_z\frac{\partial z}{\partial x}, -n_z\frac{\partial z}{\partial y}, n_z\right) = -n_z\left(\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}, -1\right)$$
$$\Longrightarrow \quad \mathbf{n} = \left(\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}, -1\right), \tag{5.18}$$

where in the last step we removed n_z as this would only affect the (non-important) length of the normal vector, not its direction.

Once we have an expression for the normal, we can also write an equation for the tangent plane touching our surface at point M_0 as shown in Fig. 5.4(a). According to Eq. (1.101), the equation for the plane has the form:

$$n_x\,(x - x_0) + n_y\,(y - y_0) + n_z\,(z - z_0) = 0,$$

where the components of the normal are given by Eq. (5.18). Therefore, we can rewrite the equation for the normal plane in more detail as follows:

$$z - z_0 = (x - x_0)\left(\frac{\partial z}{\partial x}\right)_{M_0} + (y - y_0)\left(\frac{\partial z}{\partial y}\right)_{M_0}. \tag{5.19}$$

Note that derivatives above are calculated at point M_0 and hence are constants.

As an example of the application of this result, consider the cone specified by the equation $z = \sqrt{x^2 + y^2}$. Using Eq. (5.18), we have: $\mathbf{n} = (x/z, y/z, -1)$.

Problem 5.13. *Show that the normal vector to the sphere $x^2 + y^2 + z^2 = R^2$ is always directed radially from the sphere centre, i.e. it goes along the vector $\mathbf{r} = (x, y, z)$.*

Problem 5.14. *Obtain the normal vector to the one-pole hyperboloid (5.3) at the point (x_0, y_0, z_0) which is on its surface.*

Problem 5.15. *The same for the two-pole hyperboloid (5.4).*

Problem 5.16. *The same for the saddle (5.5).*

Problem 5.17. *Derive the result (5.18) using a slightly different method. First, consider the vectors[3] $\tau_x = \partial \mathbf{r}/\partial x$ and $\tau_y = \partial \mathbf{r}/\partial y$, where $\mathbf{r} = x\mathbf{i} + y\mathbf{j} + z(x, y)\mathbf{k}$. These vectors, see Fig. 5.4(a), are tangential to the curves DC and AB, respectively, at the point M_0. Therefore, the normal is obtained by calculating their vector product: $\mathbf{n} = \tau_x \times \tau_y$. Up to a scale factor and (possibly) a sign this is exactly the same result as in Eq. (5.18).*

Here we have considered a specific surface $z = z(x, y)$. Using the same method one can also consider surfaces $x = x(y, z)$ and $y = y(z, x)$. An arbitrary surface requires a somewhat more general approach; this will be done later on in Sect. 6.4.1.

5.5 Exact Differentials

The linear part of the change of the function $z = f(x, y)$ due to change of its variables dx and dy, see (5.9), is called its *exact differential* (cf. Sect. 3.2). It is written as follows:

$$dz = \frac{\partial z}{\partial x}dx + \frac{\partial z}{\partial y}dy. \tag{5.20}$$

In the general case of a function $y = f(\mathbf{x})$ of N variables $\mathbf{x} = (x_1, \ldots, x_N)$ the differential is

$$dy = \sum_{i=1}^{N} \frac{\partial f}{\partial x_i}dx_i. \tag{5.21}$$

The independence of the mixed derivatives on the order in which differentiation is performed allows establishing a simple necessary condition for a form such as

[3] Notation $\partial \mathbf{r}/\partial \lambda$ corresponds to the vector $(\partial x/\partial \lambda, \partial y/\partial \lambda, \partial z/\partial \lambda)$.

$A(x,y)dx + B(x,y)dy$ to be the exact differential of *some* function $z = z(x,y)$. This means that there exists such a function $z = z(x,y)$ whose differential (5.20) is equal exactly to $A(x,y)dx + B(x,y)dy$. Of course, it is not at all obvious that for any functions A and B this would be necessarily true; if this is not true, the form is called *inexact differential*. We shall now establish the necessary and sufficient condition for the form to be the exact differential.

Theorem 5.3. *The form $A(x,y)dx + B(x,y)dy$ is the exact differential dz of some function $z = z(x,y)$ if and only if*

$$\frac{\partial A}{\partial y} = \frac{\partial B}{\partial x}. \tag{5.22}$$

Proof. We first prove that this is a necessary condition. Indeed, if there was a function z such that its differential dz is given precisely by formula (5.20), then we should have that $A(x,y) = \partial z/\partial x$ and $B(x,y) = \partial z/\partial y$. Let us now differentiate A with respect to y and B with respect to x:

$$\frac{\partial A}{\partial y} = \frac{\partial}{\partial y}\left(\frac{\partial z}{\partial x}\right) = \frac{\partial^2 z}{\partial y \partial x} \quad \text{and} \quad \frac{\partial B}{\partial x} = \frac{\partial}{\partial x}\left(\frac{\partial z}{\partial y}\right) = \frac{\partial^2 z}{\partial x \partial y}.$$

We have arrived at two mixed derivatives where differentiation is performed in the opposite order. However, if the function $z = z(x,y)$ satisfies the required conditions of Theorem 5.2, these mixed derivatives must be the same, i.e. the functions A and B must then satisfy the condition (5.22).

To prove the sufficiency, we assume that the condition (5.22) is satisfied. We have to show that there exists a function $z = z(x,y)$ such that its differential dz has exactly the given form $Adx + Bdy$. Let us define a function $g(x,y)$ such that

$$A(x,y) = \frac{\partial g}{\partial x}. \tag{5.23}$$

This can always be done. Then we can write:

$$\frac{\partial A}{\partial y} = \frac{\partial^2 g}{\partial y \partial x} = \frac{\partial^2 g}{\partial x \partial y} = \frac{\partial}{\partial x}\left(\frac{\partial g}{\partial y}\right).$$

On the other hand, because of our assumption (5.22), this is also equal to $\partial B/\partial x$, which means that

$$\frac{\partial B}{\partial x} = \frac{\partial}{\partial x}\left(\frac{\partial g}{\partial y}\right) \quad \Longrightarrow \quad \frac{\partial}{\partial x}\left(B - \frac{\partial g}{\partial y}\right) = 0 \quad \Longrightarrow \quad B = \frac{\partial g}{\partial y} + h(y), \tag{5.24}$$

where $h(y)$ must be some function of y only. The latter result is a direct consequence of our assumption (5.22). Next, we define a function $p(y)$ such that $p'(y) = h(y)$. The function $p(y)$ can be always found, at least in principle, as it is an indefinite integral of $h(y)$.

Now we are ready to provide a working guess for the function $z(x, y)$ as follows: $z(x, y) = g(x, y) + p(y)$. This function satisfies the required properties. Indeed,

$$\frac{\partial z}{\partial x} = \frac{\partial g}{\partial x} = A \quad \text{and} \quad \frac{\partial z}{\partial y} = \frac{\partial g}{\partial y} + p'(y) = \frac{\partial g}{\partial y} + h(y) = B,$$

as required. **Q.E.D.**

The second part of the theorem suggests also a constructive method of finding the function $z = z(x, y)$, provided, of course, that the necessary condition (5.22) is satisfied.

We shall illustrate the method by considering the differential form $\left(1 + 3x^2y^2\right) dx + 2x^3ydy$. Here $A(x, y) = 1 + 3x^2y^2$ and $B(x, y) = 2x^3y$. We would like to know if this form is an exact differential, and if it is, of which function $z(x, y)$. We have:

$$\frac{\partial A}{\partial y} = 6x^2y \quad \text{and} \quad \frac{\partial B}{\partial x} = 6x^2y,$$

which are the same and hence there must be a function z such that its differential $dz = \left(1 + 3x^2y^2\right) dx + 2x^3ydy$. To find this function, we first consider the condition $\partial z/\partial x = A = 1 + 3x^2y^2$. It follows from here that z is an indefinite integral of A with respect to the variable x, with y considered as a constant since we have a partial derivative of z with respect to x. Thus,

$$z(x, y) = \int A(x, y)dx = x + x^3y^2 + C(y),$$

where we have added an arbitrary function $C(y)$ since its derivative with respect to x is obviously zero. Now, we shall use the expression for B and the condition that it must be equal to $\partial z/\partial y$, i.e.

$$\frac{\partial z}{\partial y} = \frac{\partial}{\partial y}\left[x + x^3y^2 + C(y)\right] = 2x^3y + C'(y) = B = 2x^3y,$$

from which it follows that $C'(y) = 0$, so $C(y)$ does not depend on y at all and must be a constant C. This results in the final expression for the function sought for: $z = x + x^3y^2 + C$. It is easily checked that, indeed, $\partial z/\partial x = 1 + 3x^2y^2$ and $\partial z/\partial y = 2x^3y$, as it should be.

Although in the above example $C'(y) = 0$, in other cases it may not be equal to zero, so that $C(y)$ may still be some function of y.

Problem 5.18. *Determine if the expression* $2xydx + \left(x^2 + y^2\right) dy$ *corresponds to the exact differential* dz *of some function* $z = z(x, y)$*, and then find this function.* [*Answer*: $z = x^2y + y^3/3 + C$.]

Problem 5.19. *The same for the form* $\left(6xy + 5y^2\right) dx + \left(3x^2 + 10xy\right) dy$. [*Answer*: $z = 3x^2y + 5xy^2 + C$.]

Problem 5.20. *Show that the criterion (5.22) for the case of a function of* N *variables is generalised as follows: for the expression*

$$\sum_{i=1}^{N} A_i(\mathbf{x})dx_i$$

to be the exact differential of some function $y = f(\mathbf{x})$*, the following conditions need to be satisfied:*

$$\frac{\partial A_i}{\partial x_j} = \frac{\partial A_j}{\partial x_i} \quad \textit{for any} \quad i \neq j.$$

Argue that there will be exactly $N(N-1)/2$ *such equations to be satisfied.*

5.6 Derivatives of Composite Functions

Consider a differentiable function $y = f(\mathbf{x})$ of N variables $\mathbf{x} = (x_1, \ldots, x_N)$. Let each of the variables be a continuous function of a single variable t, i.e. $x_i = x_i(t)$, and each of these functions can be differentiated with respect to t. This makes $y = y(t)$ a function of a single variable t and it is reasonable to ask a question of how to find its derivative dy/dt. In order to proceed, we need to calculate the change of the function $\Delta y = y(t + \Delta t) - y(t)$ due to the change Δt of its variable t. Once t is changed by Δt, each of the functions $x_i(t)$ is changed as well by $\Delta x_i = x_i(t + \Delta t) - x_i(t)$. Because the function $y = f(\mathbf{x})$ is differentiable, we can use Eq. (5.10) to relate the change in y to the changes in x_i:

$$\Delta y = \sum_{i=1}^{N} \left(\frac{\partial y}{\partial x_i} + \alpha_i \right) \Delta x_i,$$

and hence

$$y'(t) = \lim_{\Delta t \to 0} \frac{\Delta y}{\Delta t} = \lim_{\Delta t \to 0} \sum_{i=1}^{N} \left(\frac{\partial y}{\partial x_i} + \alpha_i \right) \frac{\Delta x_i}{\Delta t} = \sum_{i=1}^{N} \left(\frac{\partial y}{\partial x_i} + \lim_{\Delta t \to 0} \alpha_i \right) \lim_{\Delta t \to 0} \frac{\Delta x_i}{\Delta t}.$$

When $\Delta t \to 0$, all $\Delta x_i \to 0$ since the functions $x_i(t)$ are continuous. Also, all $\alpha_i \to 0$ when $\Delta \mathbf{x} \to 0$, and hence we obtain:

$$y'(t) = \frac{dy}{dt} = \sum_{i=1}^{N} \frac{\partial y}{\partial x_i} \lim_{\Delta t \to 0} \frac{\Delta x_i}{\Delta t} = \sum_{i=1}^{N} \frac{\partial y}{\partial x_i} \frac{dx_i}{dt}. \tag{5.25}$$

This formula has a very simple meaning: if arguments of a function of N variables in turn depend on a single variable t, one should differentiate y with respect to each of its direct variables x_i (partial derivatives with ∂), multiply them by the derivatives $x_i'(t)$ and then sum up all the contributions. Each term in the sum would be exactly the same if the function y was considered as a function of a *single* variable x_i. Indeed, in this case y is a composition $y = y(x_i) = y\,(x_i(t))$, and its derivative is $y'(t) = y'(x_i)x'(t)$ according to the chain rule (3.16), the result for a function of a single variable. Therefore, formula (5.25) tells us that one has to use the chain rule for each argument x_i of the function $y = f(\mathbf{x})$ individually (i.e. assuming that other arguments are constants, and hence using the partial derivative symbol for the $y'(x_i) \Rightarrow \partial y/\partial x_i$), and then summing up all such contributions. It is clear from this discussion that our result (5.25) is the straightforward generalisation of Eq. (3.16) for functions of more than one variable, i.e. it is a chain rule for this case.

As an example, consider $z = x^2y^2 + \sin x \sin y$ with $x = t^2$ and $y = t^3$. Using the formula just derived, we write:

$$\begin{aligned}
\frac{dz}{dt} &= \frac{\partial}{\partial x}\left(x^2y^2 + \sin x \sin y\right)\frac{d\left(t^2\right)}{dt} + \frac{\partial}{\partial y}\left(x^2y^2 + \sin x \sin y\right)\frac{d\left(t^3\right)}{dt} \\
&= \left(2xy^2 + \cos x \sin y\right)2t + \left(2x^2y + \sin x \cos y\right)3t^2 \\
&= 2t\left(2t^8 + \cos t^2 \sin t^3\right) + 3t^2\left(2t^7 + \sin t^2 \cos t^3\right) = 10t^9 + 2t\cos t^2 \sin t^3 \\
&\quad + 3t^2 \sin t^2 \cos t^3 .
\end{aligned}$$

It is instructive to see that exactly the same result is obtained if we replace the functions x and y in $z = z(x, y)$ with their respective expressions via t at the very beginning and then differentiate with respect to t:

$$z(t) = \left(t^2\right)^2\left(t^3\right)^2 + \sin t^2 \sin t^3 = t^{10} + \sin t^2 \sin t^3,$$

which, as can easily be checked by direct differentiation (do it!), results exactly in the same expression as when using the chain rule.

As a more physical example, let us consider a gas of atoms. The atoms in the gas all have different positions and momenta (velocity times atom mass, m) which change with time t. Statistical properties of such a gas can be described by the distribution function $f(\mathbf{r}, \mathbf{p}, t)$ defined in the following way. We define a six-dimensional phase space formed by the particles coordinates $\mathbf{r} = (x, y, z)$ (the coordinate subspace) and momentum $\mathbf{p} = \left(p_x, p_y, y_z\right)$ (the momentum subspace).

Then $f(\mathbf{r}, \mathbf{p}, t)d\mathbf{r}d\mathbf{p}$ gives the number of atoms of the gas at time t which coordinates fall within a small box between $\mathbf{r}$ and $\mathbf{r} + d\mathbf{r}$ of the coordinate subspace (i.e. within the Cartesian coordinates ranging between x and $x + dx$, y and $y + dy$, z and $z + dz$), and their momenta are within a small box in the momentum subspace between $\mathbf{p}$ and $\mathbf{p} + d\mathbf{p}$.

The total change of the number of atoms in that six-dimensional box in time can be due to a possible intrinsic time dependence of the distribution function itself, and because the atoms move in and out of the box. The total change, per unit time, is given by the total derivative df/dt of the function $f(\mathbf{r}, \mathbf{p}, t)$ with respect to time, considering all seven its arguments $f(\mathbf{r}, \mathbf{p}, t) = f\left(x, y, z, p_x, p_y, p_z, t\right)$. Therefore, f can actually be written as $f(\mathbf{x})$, where $N = 7$ and $\mathbf{x} = (\mathbf{r}(t), \mathbf{p}(t), t)$ is a 7−dimensional vector. The time dependence of the $\mathbf{r}$ and $\mathbf{p}$ follow from the definition of the velocity and Newton's equations of motion: $\dot{\mathbf{r}}(t) = d\mathbf{r}/dt = \mathbf{p}/m$ and $\dot{\mathbf{p}}(t) = d\mathbf{p}/dt = \mathbf{F}$, where $\mathbf{F}$ is the force (due to an external field and interaction between atoms). Therefore, the total derivative of this composite function, according to Eq. (5.25), is:

$$\left(\frac{df}{dt}\right)_{\text{tot}} = \frac{\partial f}{\partial t}\frac{dt}{dt} + \sum_{\alpha=1}^{3}\frac{\partial f}{\partial r_\alpha}\frac{dr_\alpha}{dt} + \sum_{\alpha=1}^{3}\frac{\partial f}{\partial p_\alpha}\frac{dp_\alpha}{dt} = \frac{\partial f}{\partial t} + \sum_{\alpha=1}^{3}\frac{\partial f}{\partial r_\alpha}\frac{p_\alpha}{m} + \sum_{\alpha=1}^{3}\frac{\partial f}{\partial p_\alpha}F_\alpha,$$

where $\alpha = 1, 2, 3$ designates the three Cartesian components of the vectors. The expression above can be written in a more compact vector form as follows:

$$\left(\frac{df}{dt}\right)_{\text{tot}} = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial \mathbf{r}}\cdot\frac{\mathbf{p}}{m} + \frac{\partial f}{\partial \mathbf{p}}\cdot\mathbf{F}. \tag{5.26}$$

Here the second and the third terms in the right-hand side are scalar products of vectors: $\partial f/\partial\mathbf{r}$ is the vector $(\partial f/\partial x, \partial f/\partial y, \partial f/\partial z)$ and $\partial f/\partial\mathbf{p}$ is similarly the vector $\left(\partial f/\partial p_x, \partial f/\partial p_y, \partial f/\partial p_z\right)$.[4] The change of f in time is due to collisions between atoms. Equating the obtained expression for the derivative of f to the so-called collision integral results in the (classical) Boltzmann's kinetic equation for the gas of atoms.

Problem 5.21. *Calculate $z'(t)$ for the following function: $z = \exp\left(-x^2 - y^2\right)$, where $x = e^{-t}$ and $y = 2e^{-2t}$.* [*Answer*: $2ze^{-2t}\left(1 + 8e^{-2t}\right)$.]

Problem 5.22. *The Hamiltonian*[5] *of a particle in an external field $U(\mathbf{r}, t)$ is given as a function of both particle momentum $\mathbf{p}$ and its position $\mathbf{r}$, i.e. $H(\mathbf{r}, \mathbf{p}) = \mathbf{p}^2/2m + U(\mathbf{r}, t)$. Show that the Hamiltonian function is conserved in time only if the external potential does not depend explicitly on time.* [*Hint*: the Newton's equations of motion read $\dot{\mathbf{p}} = \mathbf{F} = -\partial U/\partial\mathbf{r}$, where $\dot{\mathbf{p}} = d\mathbf{p}/dt$.]

[4] We shall learn in Sect. 5.8 that these vectors are called gradients.

[5] Due to Sir William Rowan Hamilton.

The derivative (5.25) is often called *total derivative* of the function y with respect to t since it includes contributions from all its variables into its overall change with the t. Sometimes this formula is written in a symbolic form via *operators*:

$$\frac{dy}{dt} = \sum_{i=1}^{N} \frac{dx_i}{dt}\frac{\partial y}{\partial x_i} = \left(\sum_{i=1}^{N} \frac{dx_i}{dt}\frac{\partial}{\partial x_i}\right) y = \hat{D}y. \tag{5.27}$$

Here the operator

$$\hat{D} = \sum_{i=1}^{N} \frac{dx_i}{dt}\frac{\partial}{\partial x_i} = x_1'(t)\frac{\partial}{\partial x_1} + x_2'(t)\frac{\partial}{\partial x_2} + \cdots + x_N'(t)\frac{\partial}{\partial x_N} \tag{5.28}$$

contains a sum of partial derivatives with respect to all variables x_1, x_2, etc., multiplied by t-derivatives of the variables themselves. Frequently a hat above a symbol is used to indicate that it is an operator.

Now we can consider a more general case of the composite functions. Let us first look at the simplest case of a function of only two variables, $z = f(x, y)$, when the variables are both functions not of one (as we have done before) but two variables: $x = x(u, v)$ and $y = y(u, v)$. Then, effectively, the function z becomes a function of two variables, u and v, and it is legitimate to ask ourselves what would be its partial (this time!) derivatives with respect to these two variables. Consider first the partial derivative of $z = z(x(u, v), y(u, v))$ with respect to u, i.e. the variable v is kept constant. In that case z may be thought of as effectively depending only on a single variable u, and hence our general result (5.25) can be directly applied:

$$\frac{\partial z}{\partial u} = \frac{\partial z}{\partial x}\frac{\partial x}{\partial u} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial u} \quad \text{or} \quad z_u' = z_x' x_u' + z_y' y_u', \tag{5.29}$$

where we have again used simplified notations for the partial derivatives, e.g. $y_u' = \partial y / \partial u$, etc.

Similarly,

$$\frac{\partial z}{\partial v} = \frac{\partial z}{\partial x}\frac{\partial x}{\partial v} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial v} \quad \text{or} \quad z_v' = z_x' x_v' + z_y' y_v'. \tag{5.30}$$

This result is easily generalised to a more general case of functions of more than two variables each of which is also a function of arbitrary number of variables.

Problem 5.23. *Derive analogous formulae for the partial derivatives z_u' and z_v' of $z = z(x)$, where $x = x(u, v)$.*

Problem 5.24. *Similarly for $z = z(u, x)$, where $x = x(u, v)$.*

Problem 5.25. *Similarly for $z = z(u, v, x)$, where $x = x(u, v)$.*

Problem 5.26. *Consider $z = z(x, y)$ with $x = x(u, v)$ and $y = y(u, v)$. Prove the following expressions for the second order derivatives of z with respect to u and v:*

$$z''_{uu} = z'_x x''_{uu} + z'_y y''_{uu} + \left[z''_{xx} \left(x'_u\right)^2 + 2z''_{xy} x'_u y'_u + z''_{yy} \left(y'_u\right)^2 \right], \tag{5.31}$$

$$z''_{vv} = z'_x x''_{vv} + z'_y y''_{vv} + \left[z''_{xx} \left(x'_v\right)^2 + 2z''_{xy} x'_v y'_v + z''_{yy} \left(y'_v\right)^2 \right], \tag{5.32}$$

$$z''_{uv} = z'_x x''_{uv} + z'_y y''_{uv} + \left[z''_{xx} x'_u x'_v + z''_{xy} \left(x'_u y'_v + x'_v y'_u\right) + z''_{yy} y'_u y'_v \right], \tag{5.33}$$

where the simplified notations for the second order partial derivatives were again used, e.g. $z''_{xy} = \partial^2 z / \partial x \partial y$.

Problem 5.27. *Polar coordinates on the plane (r, φ) are defined via $x = r \cos \varphi$ and $y = r \sin \varphi$. Show that differential operators with respect to the Cartesian coordinates and polar coordinates are related as follows:*

$$\frac{\partial}{\partial r} = \cos \varphi \frac{\partial}{\partial x} + \sin \varphi \frac{\partial}{\partial y} \quad \textit{and} \quad \frac{\partial}{\partial \varphi} = -r \sin \varphi \frac{\partial}{\partial x} + r \cos \varphi \frac{\partial}{\partial y}. \tag{5.34}$$

[*Hint*: assume a function $z = f(x, y)$ with $x = x(r, \varphi)$ and $y = y(r, \varphi)$ as given by the equations for the polar coordinates, and then use Eqs. (5.29) and (5.30).] *Then apply these relationships to $z = x^2 y^2$ and calculate its partial derivatives with respect to r and φ. Verify your result by differentiating directly the function $z = x^2 y^2 = r^4 \cos^2 \varphi \sin^2 \varphi$.*

Finally, let us consider *implicit functions* and their derivatives. Yet again, we start from the simplest case of a function of two variables, $z = z(x, y)$, which is specified by means of a function of three variables $f(x, y, z)$ via the following algebraic equation:

$$f\left(x, y, z\left(x, y\right)\right) = 0. \tag{5.35}$$

If we could solve this equation for the function $z = z(x, y)$, then there would have been no problem as the required partial derivatives $\partial z / \partial x$ and $\partial z / \partial y$ would be available for the calculation directly. However, this may not be the case (i.e. the equation above cannot be solved analytically) and hence a general method is required. The idea of this method is based on the fact that the function f is a constant (equal to zero), so that its derivative with respect to either x or y (not forgetting the fact that its third argument z is also a function of these two) is zero. Therefore, one can write:

$$\left.\frac{\partial f}{\partial x}\right|_{\mathrm{tot}} = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial z}\frac{\partial z}{\partial x} = 0 \quad \Longrightarrow \quad \frac{\partial z}{\partial x} = -\frac{\partial f/\partial x}{\partial f/\partial z}. \tag{5.36}$$

Note that the partial derivative $\partial f/\partial x|_{\mathrm{tot}}$ in the left-hand side corresponds to the total derivative of f with respect to the variable x, including its dependence on x via its third variable z. In the right-hand side of the final formula for $\partial z/\partial x$, however, the partial derivative $\partial f/\partial x$ is taken only with respect to the first (explicit) argument of f, i.e. keeping not only y but also z constant. Similarly one obtains

$$\frac{\partial z}{\partial y} = -\frac{\partial f/\partial y}{\partial f/\partial z}. \tag{5.37}$$

Problem 5.28. *Show that the results above are formally correct even if a function of more than two variables is similarly specified via an algebraic equation, e.g.* $f(x, y, z, w) = 0$, *where* $w = w(x, y, z)$.

Problem 5.29. *Consider a function* $w = w(x, y) = h(x, y, z(x, y))$. *As we already know, the first order derivative is given by:*

$$\frac{\partial w}{\partial x} = \frac{\partial h}{\partial x} + \frac{\partial h}{\partial z}\frac{\partial z}{\partial x}. \tag{5.38}$$

Show that for the second order derivatives we get:

$$\frac{\partial^2 w}{\partial x^2} = \frac{\partial h}{\partial z}\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 h}{\partial x^2} + 2\frac{\partial^2 h}{\partial x \partial z}\frac{\partial z}{\partial x} + \frac{\partial^2 h}{\partial z^2}\left(\frac{\partial z}{\partial x}\right)^2, \tag{5.39}$$

and similarly for the derivatives with respect to y. Also, for the mixed derivative,

$$\frac{\partial^2 w}{\partial x \partial y} = \frac{\partial h}{\partial z}\frac{\partial^2 z}{\partial x \partial y} + \frac{\partial^2 h}{\partial x \partial y} + \frac{\partial^2 h}{\partial x \partial z}\frac{\partial z}{\partial y} + \frac{\partial^2 h}{\partial y \partial z}\frac{\partial z}{\partial x} + \frac{\partial^2 h}{\partial z^2}\frac{\partial z}{\partial x}\frac{\partial z}{\partial y}. \tag{5.40}$$

As an example, consider the function $z = z(x, y)$ specified via the equation $x^2 + y^2 + z^2 = 1$ and the condition $z \geq 0$. In this case $f = x^2 + y^2 + z^2 - 1$, and we obtain:

$$\frac{\partial z}{\partial x} = -\frac{\partial f/\partial x}{\partial f/\partial z} = -\frac{2x}{2z} = -\frac{x}{z} \quad \text{and} \quad \frac{\partial z}{\partial y} = -\frac{y}{z}.$$

This result can be compared with that obtained via the direct calculation: solving the algebraic equation for z, we obtain $z = +\sqrt{1-x^2-y^2}$, so that the required partial derivatives can be immediately calculated. The result, as the reader no doubt can easily check, is exactly the same as above (do it!).

Problem 5.30. *Consider equations $x = r\cos\varphi$ and $y = r\sin\varphi$, defining the polar coordinates. Differentiate both sides of the equations with respect to x and y, remembering that both r and φ are indirect functions of x and y. Solve the obtained algebraic equations with respect to the derivatives of r and φ, and hence show that:*

$$\frac{\partial r}{\partial x} = \cos\varphi, \quad \frac{\partial r}{\partial y} = \sin\varphi, \quad \frac{\partial \varphi}{\partial x} = -\frac{\sin\varphi}{r} \quad \text{and} \quad \frac{\partial \varphi}{\partial y} = \frac{\cos\varphi}{r}. \tag{5.41}$$

Using the above Eq. (5.34), we can also derive relations relating second order *operator derivatives* with respect to r and φ with those with respect to x and y. Indeed,

$$\frac{\partial^2 f}{\partial r^2} = \frac{\partial}{\partial r}\left(\frac{\partial f}{\partial r}\right) = \frac{\partial}{\partial r}\left(\cos\varphi\frac{\partial f}{\partial x} + \sin\varphi\frac{\partial f}{\partial y}\right).$$

Since we differentiate with respect to r, the other independent variable φ is to be kept constant, and hence the cosine and sine functions need to be left alone:

$$\frac{\partial^2 f}{\partial r^2} = \cos\varphi\frac{\partial}{\partial r}\left(\frac{\partial f}{\partial x}\right) + \sin\varphi\frac{\partial}{\partial r}\left(\frac{\partial f}{\partial y}\right).$$

The functions $\partial f/\partial x$ and $\partial f/\partial y$ are some functions of x and y, and hence we can again use Eq. (5.34), yielding

$$\begin{aligned}\frac{\partial^2 f}{\partial r^2} &= \cos\varphi\left(\cos\varphi\frac{\partial}{\partial x} + \sin\varphi\frac{\partial}{\partial y}\right)\frac{\partial f}{\partial x} + \sin\varphi\left(\cos\varphi\frac{\partial}{\partial x} + \sin\varphi\frac{\partial}{\partial y}\right)\frac{\partial f}{\partial y} \\ &= \cos^2\varphi\frac{\partial^2 f}{\partial x^2} + \cos\varphi\sin\varphi\frac{\partial^2 f}{\partial y\partial x} + \sin\varphi\cos\varphi\frac{\partial^2 f}{\partial x\partial y} + \sin^2\varphi\frac{\partial^2 f}{\partial y^2},\end{aligned}$$

and, since the mixed derivatives do not depend on the order in which the partial derivatives are taken, we obtain:

$$\frac{\partial^2 f}{\partial r^2} = \cos^2\varphi\frac{\partial^2 f}{\partial x^2} + 2\cos\varphi\sin\varphi\frac{\partial^2 f}{\partial y\partial x} + \sin^2\varphi\frac{\partial^2 f}{\partial y^2},$$

which can be written symbolically with operators as

$$\frac{\partial^2}{\partial r^2} = \cos^2\varphi \frac{\partial^2}{\partial x^2} + 2\cos\varphi \sin\varphi \frac{\partial^2}{\partial y \partial x} + \sin^2\varphi \frac{\partial^2}{\partial y^2} = \left(\cos\varphi \frac{\partial}{\partial x} + \sin\varphi \frac{\partial}{\partial y}\right)^2 . \tag{5.42}$$

This is of course the same as repeating twice the operator $\partial/\partial r$, Eq. (5.34), i.e. $(\partial/\partial r)^2$, as it should be.

Problem 5.31. *Show that*

$$\frac{\partial^2}{\partial \varphi^2} = -r\left(\cos\varphi \frac{\partial}{\partial x} + \sin\varphi \frac{\partial}{\partial y}\right) + \left(-r\sin\varphi \frac{\partial}{\partial x} + r\cos\varphi \frac{\partial}{\partial y}\right)^2 , \tag{5.43}$$

$$\begin{aligned} \frac{\partial^2}{\partial r \partial \varphi} &= -\sin\varphi \frac{\partial}{\partial x} + \cos\varphi \frac{\partial}{\partial y} \\ &\quad + \left(\cos\varphi \frac{\partial}{\partial x} + \sin\varphi \frac{\partial}{\partial y}\right)\left(-r\sin\varphi \frac{\partial}{\partial x} + r\cos\varphi \frac{\partial}{\partial y}\right) = \frac{\partial^2}{\partial \varphi \partial r}. \end{aligned} \tag{5.44}$$

Problem 5.32. *Calculate all three second derivatives of the function $z(x,y) = x\sin y + y\sin x$ with respect to the polar coordinates using Eqs. (5.42)–(5.44). Verify your expressions by replacing x and y directly in z and performing differentiation.*

Problem 5.33. *Using the above formulae, demonstrate by a direct calculation that*

$$\frac{\partial^2 f}{\partial r^2} + \frac{1}{r}\frac{\partial f}{\partial r} + \frac{1}{r^2}\frac{\partial^2 f}{\partial \varphi^2} = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}. \tag{5.45}$$

Here the expression in the right-hand side represents a two-dimensional Laplacian[6] of the function $f(x,y)$. This identity shows how the Laplacian can be equivalently written using polar coordinates instead. There are much more powerful methods to change variables in this kind of differential expressions.

[6] Named after Pierre-Simon, marquis de Laplace.

Problem 5.34. *Consider the spherical coordinates*

$$x = r\sin\theta\cos\phi, \quad y = r\sin\theta\sin\phi \quad and \quad z = r\cos\theta.$$

Show by differentiating these equations with respect to x, y and z that

$$\frac{\partial r}{\partial x} = \cos\phi, \quad \frac{\partial \theta}{\partial x} = \frac{1}{r}\cos\theta\cos\phi, \quad \frac{\partial \phi}{\partial x} = -\frac{1}{r}\frac{\sin\phi}{\sin\theta},$$
$$\frac{\partial r}{\partial y} = \sin\theta\sin\phi, \quad \frac{\partial \theta}{\partial y} = \frac{1}{r}\cos\theta\sin\phi, \quad \frac{\partial \phi}{\partial y} = \frac{1}{r}\frac{\cos\phi}{\sin\theta},$$
$$\frac{\partial r}{\partial z} = \cos\theta, \quad \frac{\partial \theta}{\partial z} = -\frac{1}{r}\sin\theta, \quad \frac{\partial \phi}{\partial z} = 0.$$

Problem 5.35. *In quantum mechanics Cartesian components of the operator of the angular momentum of an electron are defined as follows:*

$$\hat{L}_x = -i\hbar\left(y\frac{\partial}{\partial z} - z\frac{\partial}{\partial y}\right), \quad \hat{L}_y = -i\hbar\left(z\frac{\partial}{\partial x} - x\frac{\partial}{\partial z}\right), \quad \hat{L}_z = -i\hbar\left(x\frac{\partial}{\partial y} - y\frac{\partial}{\partial x}\right).$$

Show that in the spherical coordinates the operators are (these operators act on functions depending only on the angles so that the r dependence can be ignored):

$$\hat{L}_x = i\hbar\left(\sin\phi\frac{\partial}{\partial\theta} + \frac{\cos\phi}{\tan\theta}\frac{\partial}{\partial\phi}\right), \quad \hat{L}_y = i\hbar\left(-\cos\phi\frac{\partial}{\partial\theta} + \frac{\sin\phi}{\tan\theta}\frac{\partial}{\partial\phi}\right),$$
$$\hat{L}_z s = -i\hbar\frac{\partial}{\partial\phi}.$$

Finally, demonstrate that the square of the operator of the total angular momentum is:

$$\hat{\mathbf{L}}^2 = \hat{L}_x^2 + \hat{L}_y^2 + \hat{L}_z^2 = -\hbar^2\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\phi^2}\right].$$

Consider now, quite generally, three quantities x, y and z which are related by a single equation $f(x, y, z) = 0$. This means that any one of them may be considered as a function of the other two. Then, writing differentials for $z = z(x, y)$ and $x = x(y, z)$, we obtain:

$$dz = \left(\frac{\partial z}{\partial x}\right)_y dx + \left(\frac{\partial z}{\partial y}\right)_x dy \tag{5.46}$$

and

$$dx = \left(\frac{\partial x}{\partial y}\right)_z dy + \left(\frac{\partial x}{\partial z}\right)_y dz. \tag{5.47}$$

Here the variable written as a subscript to the partial derivatives is considered constant while differentiating. Use now dz from (5.46) in the second term of dx above:

$$\begin{aligned} dx &= \left(\frac{\partial x}{\partial y}\right)_z dy + \left(\frac{\partial x}{\partial z}\right)_y \left[\left(\frac{\partial z}{\partial x}\right)_y dx + \left(\frac{\partial z}{\partial y}\right)_x dy\right] \\ &= \left(\frac{\partial x}{\partial z}\right)_y \left(\frac{\partial z}{\partial x}\right)_y dx + \left[\left(\frac{\partial x}{\partial y}\right)_z + \left(\frac{\partial x}{\partial z}\right)_y \left(\frac{\partial z}{\partial y}\right)_x\right] dy. \end{aligned}$$

Examine the coefficient to dx. The variable y is kept constant, and hence x can be considered as a function of only z, while z can be considered as a function of x via the inverse function; however, as we know from Eq. (3.17), $z'_x = 1/x'_z$, which proves that the coefficient to dx is in fact equal to one. To have the left-hand side equal to the right one, we then need the expression in the square brackets to be zero which gives us an interesting identity between various partial derivatives in this case:

$$\left(\frac{\partial x}{\partial y}\right)_z + \left(\frac{\partial x}{\partial z}\right)_y \left(\frac{\partial z}{\partial y}\right)_x = 0. \tag{5.48}$$

Problem 5.36. *Prove that Eq. (5.48) can also be written in a more symmetric form:*

$$\left(\frac{\partial x}{\partial z}\right)_y \left(\frac{\partial z}{\partial y}\right)_x \left(\frac{\partial y}{\partial x}\right)_z = -1. \tag{5.49}$$

Next, let us examine relationships between derivatives of four functions x, y, z and t, from which only two (any two) are independent, i.e. any other two can be considered as functions of them. Then, we can write, considering t as a function of x and y and using dx from (5.47):

$$\begin{aligned} dt &= \left(\frac{\partial t}{\partial x}\right)_y dx + \left(\frac{\partial t}{\partial y}\right)_x dy = \left(\frac{\partial t}{\partial x}\right)_y \left[\left(\frac{\partial x}{\partial y}\right)_z dy + \left(\frac{\partial x}{\partial z}\right)_y dz\right] + \left(\frac{\partial t}{\partial y}\right)_x dy \\ &= \left[\left(\frac{\partial t}{\partial x}\right)_y \left(\frac{\partial x}{\partial y}\right)_z + \left(\frac{\partial t}{\partial y}\right)_x\right] dy + \left(\frac{\partial t}{\partial x}\right)_y \left(\frac{\partial x}{\partial z}\right)_y dz. \end{aligned}$$

In the last term y is kept constant and hence t can be considered a function of $x = x(z)$; hence, the coefficient to dz can be recognised to be the chain rule for $(\partial t/\partial z)_y$:

$$dt = \left[\left(\frac{\partial t}{\partial x}\right)_y \left(\frac{\partial x}{\partial y}\right)_z + \left(\frac{\partial t}{\partial y}\right)_x\right] dy + \left(\frac{\partial t}{\partial z}\right)_y dz.$$

Now, we compare this result with the one in which $t = t(y, z)$:

$$dt = \left(\frac{\partial t}{\partial y}\right)_z dy + \left(\frac{\partial t}{\partial z}\right)_y dz.$$

Comparison immediately yields a useful identity:

$$\left(\frac{\partial t}{\partial y}\right)_z = \left(\frac{\partial t}{\partial y}\right)_x + \left(\frac{\partial t}{\partial x}\right)_y \left(\frac{\partial x}{\partial y}\right)_z. \tag{5.50}$$

Numerous applications of the derived identities (5.48) and (5.50) can be found in thermodynamics, and several typical examples of these the reader can find in the next section.

5.7 Applications in Thermodynamics

The formalism we have developed is quite useful in thermodynamics. To illustrate these kinds of applications, let us assume that the state of a system is uniquely determined by a pair of variables selected from the following four of them: pressure P, temperature T, volume V and entropy S. Once a particular pair of variables is chosen, any other variable appears a function of these two. For each pair the so-called *thermodynamic potential* is defined which conveniently describes the state of the system. The thermodynamic potential $A = A(X, Y)$ is a unique function of the chosen two variables X and Y and hence its total differential has the form

$$dA = \left(\frac{\partial A}{\partial X}\right)_Y dX + \left(\frac{\partial A}{\partial Y}\right)_X dY. \tag{5.51}$$

Above, for convenience, as was done in the previous section and is customarily done in thermodynamics, we have written partial derivatives with a subscript indicating which particular variable is kept constant. We shall constantly use these notations throughout this section. Note that above dA is the exact differential with respect to the corresponding variables X and Y.

We shall now consider four most useful thermodynamic potentials indicating in each case the appropriate pair of variables corresponding to them: (i) the Helmholtz free energy

$$F = U - TS \quad \text{[variables } T, V\text{]}, \tag{5.52}$$

where U is the internal energy; (ii) the Gibbs free energy

$$G = U + PV - TS \quad \text{[variables } T, P\text{]}; \tag{5.53}$$

(iii) the enthalpy

$$H = U + PV \quad \text{[variables } S, P\text{]}; \tag{5.54}$$

and, finally, (iv) internal energy U for which the appropriate variables are S and V. Note that entropy can also be considered as a thermodynamic potential.

From the second law of thermodynamics it follows that $TdS = dU + PdV$, meaning physically that the applied heat $\Delta Q = TdS$ during an infinitesimal process in general goes towards increasing the system free energy dU and for doing some work PdV. Note that the heat Q is not a thermodynamic potential and its change $\Delta Q = TdS$ is *not* an exact differential (that is why instead of dQ we write ΔQ), but $dS = \Delta Q/T$ is.[7]

Let us now write explicitly total differentials for each thermodynamic potential and establish simple exact relationships between various thermodynamic quantities. Subtracting[8] $d(TS) = TdS + SdT$ from dU, we obtain $dU - d(TS) = d(U - TS) = dF$, which, from the second law above, yields

$$dF = dU - d(TS) = (TdS - PdV) - (TdS + SdT) = -SdT - PdV. \tag{5.55}$$

Since dF is an exact differential, the following identities must then be rigorously obeyed:

$$S = -\left(\frac{\partial F}{\partial T}\right)_V \quad \text{and} \quad P = -\left(\frac{\partial F}{\partial V}\right)_T . \tag{5.56}$$

Note that the last equation is in fact an *equation of state* of the system which relates P, T and V together, making any one of them a function of the other two. Since the mixed derivatives do not depend on the order in which each individual partial derivative is taken, one more identity can be established here. Indeed,

[7]The factor $1/T$, which turns an inexact differential into an exact one, is called the integrating factor. We shall come across these when considering differential equations.

[8]We are using here the product rule written directly for the differential of the function TS.

$$\left(\frac{\partial S}{\partial V}\right)_T = -\frac{\partial^2 F}{\partial V \partial T}, \quad \left(\frac{\partial P}{\partial T}\right)_V = -\frac{\partial^2 F}{\partial T \partial V} \quad \Longrightarrow \quad \left(\frac{\partial S}{\partial V}\right)_T = \left(\frac{\partial P}{\partial T}\right)_V . \tag{5.57}$$

This last formula is called the Maxwell relation.

Problem 5.37. *Establish similarly exact differentials and the corresponding identities for the internal energy:*

$$dU = TdS - PdV, \quad T = \left(\frac{\partial U}{\partial S}\right)_V, \quad P = -\left(\frac{\partial U}{\partial V}\right)_S, \quad \left(\frac{\partial T}{\partial V}\right)_S = -\left(\frac{\partial P}{\partial S}\right)_V . \tag{5.58}$$

Problem 5.38. *Similarly for the Gibbs free energy:*

$$dG = VdP - SdT, \quad V = \left(\frac{\partial G}{\partial P}\right)_T, \quad S = -\left(\frac{\partial G}{\partial T}\right)_P, \quad \left(\frac{\partial V}{\partial T}\right)_P = -\left(\frac{\partial S}{\partial P}\right)_T . \tag{5.59}$$

Problem 5.39. *Similarly for the enthalpy:*

$$dH = TdS + VdP, \quad T = \left(\frac{\partial H}{\partial S}\right)_P, \quad V = \left(\frac{\partial H}{\partial P}\right)_S, \quad \left(\frac{\partial T}{\partial P}\right)_S = \left(\frac{\partial V}{\partial S}\right)_P . \tag{5.60}$$

Problem 5.40. *Prove the following relationships:*

$$U = F - T\left(\frac{\partial F}{\partial T}\right)_V; \quad \left(\frac{\partial U}{\partial V}\right)_T = -P + T\left(\frac{\partial P}{\partial T}\right)_V; \quad \left[\frac{\partial}{\partial T}\left(\frac{F}{T}\right)\right]_V = -\frac{U}{T^2}; \tag{5.61}$$

$$H = G - T\left(\frac{\partial G}{\partial T}\right)_P; \quad \left(\frac{\partial H}{\partial P}\right)_T = V - T\left(\frac{\partial V}{\partial T}\right)_P; \quad \left[\frac{\partial}{\partial T}\left(\frac{G}{T}\right)\right]_P = -\frac{H}{T^2}. \tag{5.62}$$

At the next step we introduce various thermodynamic observable quantities which can be measured experimentally. These are: (i) heat capacities at constant volume and pressure, c_V and c_P, defined via

$$c_V = \left(\frac{\partial Q}{\partial T}\right)_V = \left(\frac{\partial U}{\partial T}\right)_V = T\left(\frac{\partial S}{\partial T}\right)_V \quad \text{and}$$
$$c_P = \left(\frac{\partial Q}{\partial T}\right)_P = \left(\frac{\partial H}{\partial T}\right)_P = T\left(\frac{\partial S}{\partial T}\right)_P \tag{5.63}$$

(note that when V is constant, $\Delta Q = dU$, while when P is constant, $\Delta Q = dH$, as follows from Eqs. (5.58) and (5.60)); (ii) thermal expansion and thermal pressure coefficients

$$\alpha_V = \frac{1}{V}\left(\frac{\partial V}{\partial T}\right)_P \quad \text{and} \quad \alpha_P = \frac{1}{P}\left(\frac{\partial P}{\partial T}\right)_V ; \tag{5.64}$$

(iii) isothermal and adiabatic compressibilites

$$\gamma_T = -\frac{1}{V}\left(\frac{\partial V}{\partial P}\right)_T \quad \text{and} \quad \gamma_S = -\frac{1}{V}\left(\frac{\partial V}{\partial P}\right)_S . \tag{5.65}$$

Simple relationships exist between all these quantities which can be established using the mathematical devices we developed at the end of the previous section, see Eqs. (5.48) and (5.50). We start by considering V as a function of P and T, i.e. $V = V(T, P)$, and applying Eq. (5.48) with the correspondence: $(V, T, P) \Longrightarrow (x, z, y)$. We have:

$$\left(\frac{\partial V}{\partial T}\right)_P \left(\frac{\partial T}{\partial P}\right)_V + \left(\frac{\partial V}{\partial P}\right)_T = 0 \quad \Longrightarrow \quad \frac{(\partial V/\partial T)_P}{(\partial P/\partial T)_V} + \left(\frac{\partial V}{\partial P}\right)_T = 0,$$

where we inverted the derivative $(\partial T/\partial P)_V$ using the inverse function relationship. Upon employing the definitions for α_V, γ_T and α_P given above, this immediately yields:

$$\frac{V\alpha_V}{P\alpha_P} - V\gamma_T = 0 \quad \Longrightarrow \quad \alpha_V = P\alpha_P\gamma_T .$$

Another useful relationship is obtained by first applying (5.48) with the correspondence $(S, P, V) \Longrightarrow (x, y, z)$, giving

$$\left(\frac{\partial S}{\partial V}\right)_P \left(\frac{\partial V}{\partial P}\right)_S + \left(\frac{\partial S}{\partial P}\right)_V = 0 \quad \Longrightarrow \quad \left(\frac{\partial V}{\partial P}\right)_S = -\frac{(\partial S/\partial P)_V}{(\partial S/\partial V)_P}; \tag{5.66}$$

similarly, using the correspondence $(P, V, T) \Longrightarrow (y, z, x)$, we get

$$\left(\frac{\partial T}{\partial V}\right)_P \left(\frac{\partial V}{\partial P}\right)_T + \left(\frac{\partial T}{\partial P}\right)_V = 0 \Longrightarrow \left(\frac{\partial V}{\partial P}\right)_T = -\frac{(\partial T/\partial P)_V}{(\partial T/\partial V)_P} = -\frac{(\partial V/\partial T)_P}{(\partial P/\partial T)_V}. \tag{5.67}$$

In the last passage we reversed both derivatives using the inverse function relation. Therefore,

$$\frac{\gamma_S}{\gamma_T} = \frac{(\partial V/\partial P)_S}{(\partial V/\partial P)_T} = \frac{(\partial S/\partial P)_V\,(\partial P/\partial T)_V}{(\partial S/\partial V)_P\,(\partial V/\partial T)_P} = \frac{(\partial S/\partial T)_V}{(\partial S/\partial T)_P} = \frac{c_V}{c_P}.$$

So far we have used formula (5.48) which connects three quantities, so that any two are independent. Let us now illustrate an application of Eq. (5.50) which connects four functions. We choose $(S, T, P, V) \Longrightarrow (t, y, z, x)$, which gives:

$$\left(\frac{\partial S}{\partial T}\right)_P = \left(\frac{\partial S}{\partial T}\right)_V + \left(\frac{\partial S}{\partial V}\right)_T \left(\frac{\partial V}{\partial T}\right)_P = \left(\frac{\partial S}{\partial T}\right)_V + \left(\frac{\partial P}{\partial T}\right)_V \left(\frac{\partial V}{\partial T}\right)_P ,$$

where we have used (5.57) in the last passage. Using now definitions of the heat capacities c_V and c_P, as well as of α_P and α_V, we finally obtain:

$$c_P = c_V + TVP\alpha_V\alpha_P = c_V + \frac{TV\alpha_V^2}{\gamma_T}.$$

Problem 5.41. *Prove the following relationship:*

$$\gamma_T = \gamma_S + \frac{TV\alpha_V^2}{c_P}.$$

[*Hint*: first, use formula (5.50) with the correspondence $(V, P, S, T) \Longrightarrow (t, y, z, x)$; then employ Eq. (5.48) for $(T, P, S) \Longrightarrow (x, y, z)$ and the last formula in Eq. (5.59).]

5.8 Directional Derivative and the Gradient of a Scalar Field

Here we shall generalise further our definition of the partial derivative. Indeed, the partial derivatives introduced above for a function of many variables correspond to the rate of change of the function along the Cartesian axes. In fact, it is also possible to define the partial derivative along an arbitrary direction in space. This is what we shall accomplish here.

Suppose there is a scalar function $U(\mathbf{r}) = U(x, y, z)$ (it is frequently said a "*scalar field*") that is continuous in some region V in a three-dimensional space (generalisation to spaces of arbitrary dimensions is straightforward). Consider a change (per unit length) of U along some curved line L between points M_0 and M:

$$\frac{\Delta U}{\Delta l} = \frac{U(M) - U(M_0)}{\Delta l},$$

where Δl is the length M_0M along the line L. Now, suppose the line L is defined parametrically in a natural way as $x = x(l)$, $y = y(l)$ and $z = z(l)$ via the length l along the curve (measured relative to some arbitrary fixed point on the curve); the value l corresponds to the point M_0, while $l + \Delta l$ to the point M as shown in Fig. 5.5. This means that the change of U along the curve can be written, in the limit of the

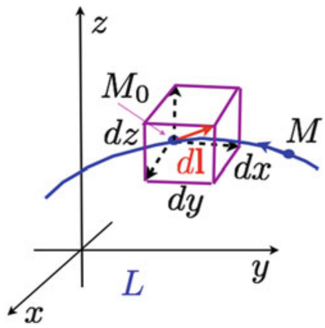


Fig. 5.5 To the calculation of the directional derivative of the scalar field $U(\mathbf{r})$ along the curve L

point M approaching M_0, as

$$\frac{\partial U}{\partial l}=\lim_{\Delta l\to 0}\frac{\Delta U}{\Delta l}=\lim_{\Delta l\to 0}\frac{U\left(\mathbf{r}(l+\Delta l)\right)-U\left(\mathbf{r}(l)\right)}{\Delta l}=\frac{\partial U}{\partial x}\frac{dx}{dl}+\frac{\partial U}{\partial y}\frac{dy}{dl}+\frac{\partial U}{\partial z}\frac{dz}{dl}$$

by the rule of differentiation of a composite function. The derivatives dx/dl, dy/dl and dz/dl correspond to the cosines of the tangent vector $d\mathbf{l}$ to the curve L at the point M_0 with the x, y and z axes, see Fig. 5.5. Therefore,

$$\frac{\partial U}{\partial l}=\frac{\partial U}{\partial x}l_x+\frac{\partial U}{\partial y}l_y+\frac{\partial U}{\partial z}l_z=\frac{\partial U}{\partial x}\cos\alpha+\frac{\partial U}{\partial y}\cos\beta+\frac{\partial U}{\partial z}\cos\gamma, \tag{5.68}$$

where $\mathbf{l}=(l_x,l_y,l_z)$ is the unit vector tangent to the curve L at point M_0, and α, β and γ are the angles it forms with the three Cartesian axes.

Problem 5.42. *Prove that*

$$\cos^2\alpha+\cos^2\beta+\cos^2\gamma=1,$$

i.e. *the vector* $\mathbf{l}$ *is indeed a unit vector.*

The derivative considered above is called *directional derivative* of the scalar field U along the curve L. It depends only on the point M_0 and the chosen direction $\mathbf{l}$. For some directions the directional derivative at the same point M_0 can be smaller, for some larger. Let us find the direction of the maximum *increase* of U at the point M_0. If we introduce the *gradient* of the scalar field U as the vector

$$\mathbf{g}=\operatorname{grad}U(\mathbf{r})=\frac{\partial U}{\partial x}\mathbf{i}+\frac{\partial U}{\partial y}\mathbf{j}+\frac{\partial U}{\partial z}\mathbf{k}, \tag{5.69}$$

then it follows from Eq. (5.68) that the directional derivative can be written as a scalar product

$$\frac{\partial U}{\partial l} = (\text{grad}\, U, \mathbf{l}) = \mathbf{g} \cdot \mathbf{l} = |\mathbf{g}| \cos \vartheta, \tag{5.70}$$

where ϑ is the angle between $\mathbf{g}$ and $\mathbf{l}$. It is seen that the maximum value the directional derivative is obtained when the direction $\mathbf{l}$ is chosen exactly along the gradient vector $\mathbf{g}$. Thus, the gradient of a scalar field points in the direction of the most rapid increase of the field, and its magnitude gives the rate of the increase. Correspondingly, the most rapid decrease of U is provided by $-\text{grad}\, U$, i.e. the direction which is opposite to the gradient.

The gradient of U can also be written, quite formally, as a product of the so-called *del operator*

$$\nabla = \frac{\partial}{\partial x}\mathbf{i} + \frac{\partial}{\partial y}\mathbf{j} + \frac{\partial}{\partial z}\mathbf{k} \tag{5.71}$$

and the field U. Indeed,

$$\nabla U = \left(\frac{\partial}{\partial x}\mathbf{i} + \frac{\partial}{\partial y}\mathbf{j} + \frac{\partial}{\partial z}\mathbf{k} \right) U = \frac{\partial U}{\partial x}\mathbf{i} + \frac{\partial U}{\partial y}\mathbf{j} + \frac{\partial U}{\partial z}\mathbf{k},$$

which is exactly grad U. This notation is quite useful and is frequently used.

As an example, let us calculate the gradient of $U(\mathbf{r}) = f(r)$, where $f(r)$ is some function and $r = |\mathbf{r}|$. The calculation of the gradient requires calculating derivatives of the function $f(r)$ with respect to Cartesian components x, y and z; these are easily done since $r = \sqrt{x^2 + y^2 + z^2}$:

$$\frac{\partial f}{\partial x} = \frac{df}{dr}\frac{\partial r}{\partial x} = \frac{df}{dr}\frac{x}{r}, \quad \frac{\partial f}{\partial y} = \frac{df}{dr}\frac{y}{r} \quad \text{and} \quad \frac{\partial f}{\partial z} = \frac{df}{dr}\frac{z}{r},$$

so that

$$\nabla U = \frac{\partial f}{\partial x}\mathbf{i} + \frac{\partial f}{\partial y}\mathbf{j} + \frac{\partial f}{\partial z}\mathbf{k} = \frac{df}{dr}\frac{x\mathbf{i} + y\mathbf{j} + z\mathbf{k}}{r} = \frac{df}{dr}\frac{\mathbf{r}}{r},$$

i.e. the gradient of a scalar field which depend only on the distance to the centre of the coordinate system has a radial character, i.e. it is directed along the vector $\mathbf{r}$.

Problem 5.43. *The gravitational potential of a galaxy decays with the distance r from it as $\varphi(r) = -\alpha/r$, where α is a positive constant. Calculate the force $F = -m\nabla\varphi(r)$ with which the galaxy acts on a spaceship of mass m at distance r. What is the direction of the force? State whether the interaction between the two objects corresponds to repulsion or attraction.* [*Answer*: $\mathbf{F} = -m\alpha\mathbf{r}/r^3$.]

Problem 5.44. *Calculate the directional derivative of* $\phi(x, y, z) = xyz$ *at the point* $(1, -1, 0)$ *in the direction* $\mathbf{e} = (1, 1, 1)$. [Answer*:* $-1/\sqrt{3}$.]

Problem 5.45. *Consider the scalar field* $\phi(x, y, z) = 2xy^3 + z^2$*. (a) Evaluate the derivative of* ϕ *along the direction* $\mathbf{e} = (1, 2, 0)$ *at arbitrary point* (x, y)*; (b) find the direction of the most rapid* increase *of* ϕ *at the point* $(2, 2, 0)$*; (c) what is the direction of the most rapid* decrease *of* ϕ *at the same point?* [Answer*: (a)* $2y^2 (y + 6x) / \sqrt{5}$*; (b)* $(16, 48, 0)$*; (c)* $(-16, -48, 0)$.]

Problem 5.46. *Given* $\phi = x^2 - y^2 z$*, calculate* $\nabla\phi$ *at point* $(1, 1, 1)$*. Hence, find the directional derivative of* ϕ *at that point in the direction* $\mathbf{e} = (1, -2, 1)$. [*Answer*: $(2, -2, -1)$ and $5/\sqrt{6}$.]

The gradient (in fact, its generalisation to many variables) is useful when one wants to find a minimum (or maximum) of a function $f(\mathbf{r})$ in p dimensions. Indeed, if we start from some initial point $\mathbf{r}_0$, then $\mathbf{g}_0 = -\mathrm{grad} f(\mathbf{r})$ would indicate the direction of the most rapid decrease of f from that point. Take a step $\delta\mathbf{r} = \lambda_1 \mathbf{g}_0$ along $\mathbf{g}_0$ with some λ_1 (the step length) and arrive at the point $\mathbf{r}_1 = \mathbf{r}_0 + \delta\mathbf{r}$. Calculate the gradient there, $\mathbf{g}_1 = -\mathrm{grad} f(\mathbf{r}_1)$, and then move to the next point $\mathbf{r}_2 = \mathbf{r}_1 + \lambda_2 \mathbf{g}_1$, and so on. Near the minimum, the gradient will be very small (exactly equal to zero at the minimum, see Sect. 5.10), the steps $\delta\mathbf{r}$ would become smaller and smaller, so that after some finite number of steps a point very close to the exact minimum will be reached. This method is called *steepest descent*. Note that the steps λ_1, λ_2, etc. should be taken small enough as compared to the characteristic dimension of the well containing the minimum point.

For instance, this method may be used to calculate the mechanical equilibrium of a collection of atoms (e.g. in a molecule or a solid). This is because in physics the gradient of an external potential $U(\mathbf{r})$ is related to the force $\mathbf{F}$ acting on a particle moving in it via $\mathbf{F} = -\mathrm{grad}\,\mathbf{U} = -\partial U/\partial\mathbf{r} = -\nabla U$ (all these notations are often used). If we can calculate the total potential energy of a set of atoms, then the forces acting on atoms at the given atomic configuration are determined by the components of the minus gradient of the energy. Moving atoms in the direction of forces leads to a lower energy. To find mechanical equilibrium of the system of atoms at zero temperature, one has to move atoms in the direction of forces unless the forces become less than some predefined tolerance. This idea is at the heart of all modern methods of material modelling.

To finish this section, we shall mention one more application of the gradient. Consider a surface in 3D that is specified by the equation $f(x, y, z) = C$, where C is a constant (e.g. for a sphere $x^2 + y^2 + z^2 = R^2$).

Theorem 5.4. *The normal vector to the surface* $f(x, y, z) = 0$ *can be obtained from the gradient of the function* $f(x, y, z)$.

Proof. Because of the equation $f(x, y, z) = 0$, only two coordinates are independent on the surface. Let us assume that the coordinates x and y are independent, i.e. $z = z(x, y)$. Then, consider the gradient of f:

$$\nabla f = \frac{\partial f}{\partial x}\mathbf{i} + \frac{\partial f}{\partial y}\mathbf{j} + \frac{\partial f}{\partial z}\mathbf{k}.$$

To calculate the necessary derivatives, we first differentiate both sides of the equation $f(x, y, z) = f(x, y, z(x, y)) = 0$ with respect to x and y, taking into account the $z = z(x, y)$ dependence explicitly:

$$\left(\frac{\partial f}{\partial x}\right)_{\text{tot}} = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial z}\frac{\partial z}{\partial x} = 0 \quad \text{and} \quad \left(\frac{\partial f}{\partial y}\right)_{\text{tot}} = \frac{\partial f}{\partial y} + \frac{\partial f}{\partial z}\frac{\partial z}{\partial y} = 0$$

(zeros in the right-hand sides are due to the zero right-hand side of $f(x, y, z) = 0$). Hence, the partial derivatives entering the gradient above are:

$$\frac{\partial f}{\partial x} = -\frac{\partial f}{\partial z}\frac{\partial z}{\partial x} \quad \text{and} \quad \frac{\partial f}{\partial y} = -\frac{\partial f}{\partial z}\frac{\partial z}{\partial y}\ ,$$

and hence the gradient becomes:

$$\nabla f = \frac{\partial f}{\partial z}\left(-\frac{\partial z}{\partial x}\mathbf{i} - \frac{\partial z}{\partial y}\mathbf{j} + \mathbf{k}\right).$$

Comparing this result with Eq. (5.18), we immediately see that ∇f is indeed directed along the normal $\mathbf{n}$ to the surface $z = z(x, y)$ (or $f(x, y, z) = 0$). **Q.E. D.**

As an example of this application of the gradient, we calculate the normal to the surface $x^2 + y^2 = z$ at the point $A(1, 1, 2)$. Here $f(x, y, z) = x^2 + y^2 - z$ and $\nabla f = 2x\mathbf{i} + 2y\mathbf{j} - \mathbf{k}$, so that the normal (of unit length)

$$\mathbf{n} = \frac{\nabla f}{|\nabla f|} = \frac{2x\mathbf{i} + 2y\mathbf{j} - \mathbf{k}}{\sqrt{4x^2 + 4y^2 + 1}},$$

which at the point A takes on the value of $\mathbf{n}(A) = 2\left(\mathbf{i} + \mathbf{j}\right)/3 - \mathbf{k}/3$.

Problem 5.47. *Using the definition of the gradient, prove its following properties:*

$$\nabla\left(U + V\right) = \nabla U + \nabla V, \quad \nabla\left(UV\right) = U\nabla V + V\nabla U, \quad \nabla U(V) = \frac{dU}{dV}\nabla V, \tag{5.72}$$

where U and V are scalar fields.

5.9 Taylor's Theorem for Functions of Many Variables

Similarly to the case of functions of a single variable considered in Sect. 3.7, there is an equivalent formula for functions of many variables. For simplicity, we shall only consider the case of a function $z = f(x, y)$ of two variables; the formulae for functions of more than two variables can be obtained similarly although this may be quite cumbersome.

What we would like to do is to write an expansion for $f(x_0 + \Delta x, y_0 + \Delta y)$ around the point (x_0, y_0) in terms of $\Delta x = x - x_0$ and $\Delta y = y - y_0$ which are variations of the function arguments around that point. Instead of generalising the method we applied in Sect. 3.7, we shall use a very simple trick which makes use of the Taylor formula (3.60) for the functions of a single variable which we already know. To this end, we construct the following auxiliary function

$$F(t) = f(x_0 + t\Delta x, y_0 + t\Delta y) \tag{5.73}$$

of a single variable $0 \le t \le 1$. The required change Δz of our function $z = f(x, y)$ is obtained by calculating $F(1)$, i.e. by taking $t = 1$ in $F(t)$, and then taking away $f(x_0, y_0)$. Sine $F(t)$ is a function of a single variable, it can be expanded according to the Maclaurin formula (3.62), i.e. the Taylor's formula around the point $t = 0$, in terms of $\Delta t = t$:

$$F(t) = F(0) + F'(0)t + \frac{F''(0)t^2}{2} + \cdots + \frac{F^{(n)}(0)t^n}{n!} + \frac{F^{(n+1)}(\xi)t^{n+1}}{(n+1)!}, \tag{5.74}$$

where the last term is the remainder $R_{n+1}(\xi)$ with $\xi = \vartheta t$ and $0 < \vartheta < 1$. By taking $t = 1$ the function $F(1)$ becomes equal to $f(x_0 + \Delta x, y_0 + \Delta y) = f(x, y)$, and the expansion above turns into the required Taylor expansion for the function $f(x, y)$ around the point (x_0, y_0).

Obviously, $F(0) = f(x_0, y_0)$. To calculate $F'(t)$, we notice that $F(t)$ can be considered as a composite function $F = F(x, y)$ with the arguments being linear functions of t, i.e. $x(t) = x_0 + t\Delta x$ and $y(t) = y_0 + t\Delta y$ with $\partial x/\partial t = \Delta x$ and $\partial y/\partial t = \Delta y$. Therefore, applying formula (5.25) for the derivative of a composite function of two variables ($N = 2$), we have:

$$F'(t) = \frac{\partial F}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial F}{\partial y}\frac{\partial y}{\partial t} = \frac{\partial F}{\partial x}\Delta x + \frac{\partial F}{\partial y}\Delta y = \left(\Delta x\frac{\partial}{\partial x} + \Delta y\frac{\partial}{\partial y}\right)F(t). \tag{5.75}$$

Here in the last passage we employed a simplified notation using an operator (within the brackets) acting on the function $F(t)$. This notation, as will be seen presently, proves to be very convenient in writing higher order derivatives. Indeed, the second derivative can be calculated by differentiating the just obtained first derivative and noticing that both functions $\partial F/\partial x$ and $\partial F/\partial y$ are again composite functions of t:

$$
\begin{aligned}
F''(t) &= \frac{d}{dt}\left(\frac{\partial F}{\partial x}\Delta x + \frac{\partial F}{\partial y}\Delta y\right) = \Delta x \frac{d}{dt}\left(\frac{\partial F}{\partial x}\right) + \Delta y \frac{d}{dt}\left(\frac{\partial F}{\partial y}\right) \\
&= \Delta x \left(\frac{\partial^2 F}{\partial x^2}\Delta x + \frac{\partial^2 F}{\partial y \partial x}\Delta y\right) + \Delta y \left(\frac{\partial^2 F}{\partial x \partial y}\Delta x + \frac{\partial^2 F}{\partial y^2}\Delta y\right) \\
&= (\Delta x)^2 \frac{\partial^2 F}{\partial x^2} + 2\Delta x \Delta y \frac{\partial^2 F}{\partial x \partial y} + (\Delta y)^2 \frac{\partial^2 F}{\partial y^2} = \left(\Delta x \frac{\partial}{\partial x} + \Delta y \frac{\partial}{\partial y}\right)^2 F(t),
\end{aligned}
$$

where again we have written the result of the differentiation using an operator; in fact, this is the same operator as in our result (5.75) for the first derivative, but squared.[9]

Problem 5.48. *Use induction to prove that generally*

$$
F^{(n)}(t) = \left(\Delta x \frac{\partial}{\partial x} + \Delta y \frac{\partial}{\partial y}\right)^n F(t) \tag{5.76}
$$

for any $n = 1, 2, 3, \ldots$.

The obtained formulae allow us to write:

$$
F^{(n)}(0) = \left(\Delta x \frac{\partial}{\partial x} + \Delta y \frac{\partial}{\partial y}\right)^n F(t)\bigg|_{t=0} = \left(\Delta x \frac{\partial}{\partial x} + \Delta y \frac{\partial}{\partial y}\right)^n f(x, y)\bigg|_{x=x_0, y=y_0}, \tag{5.77}
$$

while the remainder term at $t = 1$ yields

$$
\begin{aligned}
R_{n+1}(\vartheta) &= \frac{1}{(n+1)!}\left(\Delta x \frac{\partial}{\partial x} + \Delta y \frac{\partial}{\partial y}\right)^{n+1} F(\vartheta t)\bigg|_{t=1} \\
&= \frac{1}{(n+1)!}\left(\Delta x \frac{\partial}{\partial x} + \Delta y \frac{\partial}{\partial y}\right)^{n+1} f\left(x_0 + \vartheta \Delta x, y_0 + \vartheta \Delta y\right).
\end{aligned} \tag{5.78}
$$

Correspondingly, the Taylor expansion, after taking $t = 1$ in Eq. (5.74), reads:

$$
f(x, y) = f\left(x_0, y_0\right) + F'(0) + \frac{F''(0)}{2} + \cdots + \frac{F^{(n)}(0)}{n!} + R_{n+1}, \tag{5.79}
$$

which is our final result.

[9] Recall, that a similar trick with operators we have already used in Sect. 3.6 when illustrating the derivation of the Leibnitz formula (3.47).

As an important example, let us derive an analogue of the Lagrange formula (3.64) for the case of a function of two variables. Applying Eqs. (5.76)–(5.79) for $n = 0$, we get:

$$f(x, y) = f(x_0, y_0) + \left(\Delta x \frac{\partial}{\partial x} + \Delta y \frac{\partial}{\partial y}\right) f(x_0 + \vartheta \Delta x, y_0 + \vartheta \Delta y)$$
$$= f(x_0, y_0) + f'_x(x', y') \Delta x + f'_y(x', y') \Delta y, \tag{5.80}$$

which is the required result. Here we differentiate $f(x, y)$ with respect to x and y and then replace the variables according to $x \to x' = x_0 + \vartheta \Delta x$ and $y \to y' = y_0 + \vartheta \Delta x$, respectively. Another important formula which we shall need below corresponds to the $n = 1$ case:

$$f(x, y) = f(x_0, y_0) + \left[f'_x(x_0, y_0) \Delta x + f'_y(x_0, y_0) \Delta y\right]$$
$$+ \frac{1}{2}\left[f''_{xx}(x', y') (\Delta x)^2 + 2 f''_{xy}(x', y') \Delta x \Delta y + f''_{yy}(x', y') (\Delta y)^2\right], \tag{5.81}$$

where the second derivatives are calculated at $x' = x_0 + \vartheta \Delta x$ and $y' = y_0 + \vartheta \Delta y$.

Problem 5.49. *Derive the $n = 2$ Taylor expansion of $f(x, y) = e^{x+y}$.*

The formulae we have obtained here can in fact be generalised for functions of any number of variables; this, however, requires developing more algebraic results; in particular, a generalisation of the Binomial expansion to forms $(a_1 + \cdots + a_N)^n$ containing $N > 2$ terms (this has been done in Sect. 1.11) is required and investigation of the properties of the corresponding generalised multinomial coefficients. Then, the method developed above can be basically repeated almost without change. The final formula appears to be extremely cumbersome.

5.10 Introduction to Finding an Extremum of a Function

Now we are in the position to discuss finding maxima and minima (extrema) of functions of more than one variables. We shall only consider here in detail the case of functions of two variables, $z = f(x, y)$.

Similarly to functions of a single variable, we define the function $z = f(x, y)$ to have a maximum at point $M_0(x_0, y_0)$, if there exists a positive radius R such that for *any* point $M(x, y)$ for which the distance to the point M_0 is smaller than R (i.e. $d(M, M_0) = \sqrt{(x - x_0)^2 + (y - y_0)^2} < R$), the value of the function is $f(x, y) < f(x_0, y_0)$. The definition of the minimum is defined similarly.

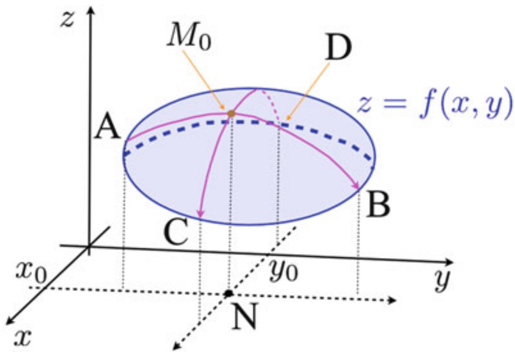


Fig. 5.6 For the prove of the necessary condition of a maximum of a function $z = f(x, y)$ of two variables. Here the function has a maximum at the point M_0 and the lines AB and DC along the surface $z = f(x, y)$, which is associated with the function, correspond to one variable to be fixed and another allowed to change

5.10.1 Necessary Condition: Stationary Points

The necessary condition for the minimum or maximum can easily be established. Consider Fig. 5.6 where the function $z = f(x, y)$ has a maximum at point M_0. If we fix the first variable x of the function at x_0 and allow only the other variable to change, we would draw the line DC on the surface associated with our function; this line will correspond to the function $g(y) = f(x_0, y)$ which is a function of only a single variable y. Since for any point y within some vicinity of y_0 the value of that function is smaller than its value $g(y_0) = f(x_0, y_0)$ (since according to our assumption the point M_0 is a maximum), the function $g(y)$ has a maximum at the point y_0. Therefore, from the familiar necessary condition for the maximum of a function of one variable, the derivative $g'(y)$ is equal to zero at $y = y_0$. But this derivative is in fact the partial derivative $f'_y(x_0, y)$ calculated at the point $y = y_0$. Similarly, fixing $y = y_0$ and considering the function $h(x) = f(x, y_0)$, we arrive at the line AB in Fig. 5.6 and the necessary condition $h'(x_0) = f'_x(x_0, y_0) = 0$. Therefore, we conclude that the necessary condition for a function to have a maximum is that its partial derivatives are equal to zero at that point:

$$f'_x(x_0, y_0) = f'_y(x_0, y_0) = 0. \tag{5.82}$$

In the case of a minimum, a similar analysis yields exactly the same condition, but in this case each line $x = x_0$ or $y = y_0$ would correspond to a minimum of a function of a single variable. We see that the conditions (5.82) generalise directly the corresponding condition for functions of single variables of Sect. 3.10. Solving two equations (5.82), gives the *stationary point* (x_0, y_0) which might be a minimum or maximum of the function. More than one solution is possible (or none[10]). Note that the stationary point(s) obtained may be either minimum, maximum or neither of these (see below) since these conditions are only the necessary ones.

[10]If a function is defined within a finite region, than a maximum or minimum may still exist at the region boundary.

It is also easy to realise that this simple result is immediately generalised to the case of a function of any number of variables $y = f(\mathbf{x})$, where $\mathbf{x} = (x_1, \dots, x_N)$. Indeed, if at the point $\mathbf{x}_0 = \left(x_1^0, \dots, x_N^0\right)$ the function has a minimum (maximum), then by fixing all variables but one, say x_i, we arrive at a function of a single variable $g(x_i) = f\left(x_1^0, \dots, x_{i-1}^0, x_i, x_{i+1}^0, \dots, x_N^0\right)$, which has a minimum (maximum) at the point x_i^0, and hence the corresponding partial derivative must be zero:

$$g'\left(x_i^0\right) = f'_{x_i}\left(\mathbf{x}^0\right) = 0. \tag{5.83}$$

By repeating this process for every $i = 1, \dots, N$ we arrive at N such conditions which give N equations; solving these equations yields the stationary point(s) $\mathbf{x}^0$ which might correspond to extrema of the function. Note that the total differential of the function at the stationary point is zero as all its partial derivatives vanish:

$$df = \sum_{i=1}^{N} \frac{\partial f}{\partial x_i} dx_i = 0. \tag{5.84}$$

As an example, let us find the minimum distance d between two lines specified by the vector equations $\mathbf{r} = \mathbf{r}_1 + t_1\mathbf{a}_1$ and $\mathbf{r} = \mathbf{r}_2 + t_2\mathbf{a}_2$, where t_1 and t_2 are the corresponding parameters. We need to minimise a function of the square of the distance,

$$\begin{aligned} d^2(t_1, t_2) &= (\mathbf{r}_1 + t_1\mathbf{a}_1 - \mathbf{r}_2 - t_2\mathbf{a}_2)^2 \\ &= (x_1 + t_1 a_{1x} - x_2 - t_2 a_{2x})^2 + \left(y_1 + t_1 a_{1y} - y_2 - t_2 a_{2y}\right)^2 \\ &\quad + (z_1 + t_1 a_{1z} - z_2 - t_2 a_{2z})^2, \end{aligned} \tag{5.85}$$

with respect to the two parameters t_1 and t_2. Calculating the partial derivatives and setting them to zero, two equations are obtained for the parameters:

$$\frac{\partial d^2}{\partial t_1} = 2\mathbf{a}_1 \cdot (\mathbf{r}_1 + t_1\mathbf{a}_1 - \mathbf{r}_2 - t_2\mathbf{a}_2) = 0 \quad \Longrightarrow \quad t_1 a_1^2 - t_2\mathbf{a}_1 \cdot \mathbf{a}_2 = -\mathbf{a}_1 \cdot \mathbf{r}_{12},$$

$$\frac{\partial d^2}{\partial t_2} = -2\mathbf{a}_2 \cdot (\mathbf{r}_1 + t_1\mathbf{a}_1 - \mathbf{r}_2 - t_2\mathbf{a}_2) = 0 \quad \Longrightarrow \quad t_1\mathbf{a}_1 \cdot \mathbf{a}_2 - t_2 a_2^2 = -\mathbf{a}_2 \cdot \mathbf{r}_{12},$$

where $\mathbf{r}_{12} = \mathbf{r}_1 - \mathbf{r}_2$, $a_1^2 = \mathbf{a}_1 \cdot \mathbf{a}_1$ and $a_2^2 = \mathbf{a}_2 \cdot \mathbf{a}_2$. Solving these equations with respect to the two parameters (e.g. solve for t_1 in the first equation and substitute into the second), we obtain

$$t_1 = \frac{-a_2^2(\mathbf{a}_1 \cdot \mathbf{r}_{12}) + (\mathbf{a}_1 \cdot \mathbf{a}_2)(\mathbf{a}_2 \cdot \mathbf{r}_{12})}{a_1^2 a_2^2 - (\mathbf{a}_1 \cdot \mathbf{a}_2)^2}, \quad t_2 = \frac{a_1^2(\mathbf{a}_2 \cdot \mathbf{r}_{12}) - (\mathbf{a}_1 \cdot \mathbf{a}_2)(\mathbf{a}_1 \cdot \mathbf{r}_{12})}{a_1^2 a_2^2 - (\mathbf{a}_1 \cdot \mathbf{a}_2)^2}.$$

Problem 5.50. *Substituting the found parameters into the distance square in (5.85), show that the minimum distance between the two lines is indeed given correctly by Eq. (1.109).*

5.10.2 Characterising Stationary Points: Sufficient Conditions

Of course, similar to the case of a single variable function, we still have to establish the sufficient conditions for the minimum and maximum. Recall that the function $y = x^3$ does have its first derivative $y' = 3x^2$ equal to zero at $x = 0$, however this point is neither minimum nor maximum as its second derivative $y'' = 6x$ is neither positive nor negative at $x = 0$. Analogue to this situation is the saddle in Fig. 5.1(d). Indeed, the function $z = x^2 - y^2$ of the particular case of the hyperbolic paraboloid of Eq. (5.5) has both its partial derivatives $z'_x = 2x$ and $z'_y = -2y$ equal to zero at $x = y = 0$, however, this point is neither minimum nor maximum: it is a minimum along some directions and a maximum along the others. If we, however, consider a similar function $z = x^2 + y^2$, then at the same point it obviously has a minimum: indeed, $z'_x = 2x$ and $z'_y = 2y$ give a single solution $x = y = 0$ at which $z(0, 0) = 0$, and at any other point the function $z(x, y) > 0$, i.e. the function at any point in the vicinity of the point $(0, 0)$ is definitely larger than zero and hence this point must be a minimum. Hence, the necessary conditions (5.83) must be supplemented with the corresponding *sufficient* conditions which should tell us if each point found by solving Eqs. (5.83) is a minimum, maximum or neither of these.

Consider a function $z = f(x, y)$ which has both partial derivatives $f'_x(x_0, y_0) = f'_y(x_0, y_0) = 0$ at some point (x_0, y_0). The change $\Delta z = f(x, y) - f(x_0, y_0)$ of the function at the point $M(x, y)$ can then be written using the $n = 1$ case Taylor expansion (5.81) in which the first order terms should be dropped as the corresponding partial derivatives are zero:

$$\Delta z = \frac{1}{2}\left[F_{xx}(\Delta x)^2 + 2F_{xy}\Delta x \Delta y + F_{yy}(\Delta y)^2\right], \tag{5.86}$$

where for convenience we introduced the following notations:

$$F_{xx} = f''_{xx}(x_0 + \vartheta\Delta x, y_0 + \vartheta\Delta y)\,,\ F_{xy} = f''_{xy}(x_0 + \vartheta\Delta x, y_0 + \vartheta\Delta y)\,,$$
$$F_{yy} = f''_{yy}(x_0 + \vartheta\Delta x, y_0 + \vartheta\Delta y)\,. \tag{5.87}$$

We shall rearrange the terms in (5.86) in such a way as to build the full square (assuming $F_{xx} \neq 0$):

$$\Delta z = \frac{F_{xx}}{2}\left\{\left[(\Delta x)^2 + 2\frac{F_{xy}}{F_{xx}}\Delta x\Delta y + \left(\frac{F_{xy}}{F_{xx}}\Delta y\right)^2\right] + \left[-\left(\frac{F_{xy}}{F_{xx}}\Delta y\right)^2 + \frac{F_{yy}}{F_{xx}}(\Delta y)^2\right]\right\}$$
$$= \frac{F_{xx}}{2}\left[\left(\Delta x + \frac{F_{xy}}{F_{xx}}\Delta y\right)^2 + \mathcal{D}\left(\frac{\Delta y}{F_{xx}}\right)^2\right], \qquad (5.88)$$

where $\mathcal{D} = -\left(F_{xy}\right)^2 + F_{xx}F_{yy}$. Note that here F_{xx}, F_{xy} and F_{yy} (and hence $\mathcal{D}$) still depend on the particular point (x, y) via ϑ, Δx and Δy. To proceed we have to assume that all three second derivatives of the function $f(x, y)$ are continuous around the point (x_0, y_0). Specifically, at this point

$$F^0_{xx} = f''_{xx}(x_0, y_0), \; F^0_{xy} = f''_{xy}(x_0, y_0), \; F^0_{yy} = f''_{yy}(x_0, y_0) \quad \text{and} \quad \mathcal{D}^0 = F^0_{xx}F^0_{yy} - \left(F^0_{xy}\right)^2.$$

Assume first that $F^0_{xx} > 0$ and $\mathcal{D}^0 > 0$. Since the derivatives are continuous, we can write:

$$\lim_{\Delta x, \Delta y \to 0} F_{xx} = F^0_{xx} > 0 \quad \text{and} \quad \lim_{\Delta x, \Delta y \to 0} \mathcal{D} = \mathcal{D}^0 > 0,$$

and hence because of the continuity of the second derivatives, there will exist a small vicinity of the point (x_0, y_0) (i.e. small enough Δx and Δy) such that $F_{xx} > 0$ and $\mathcal{D} > 0$ for any (x, y) within that vicinity. It is seen now from Eq. (5.88) that $\Delta z > 0$ and hence the point (x_0, y_0) must be a minimum. Conversely if $F_{xx} < 0$ but still $\mathcal{D} > 0$, then we have a maximum.

If, however, $\mathcal{D} < 0$, then this must be neither minimum nor maximum as we shall specifically investigate presently. Indeed, let us go along a particular direction in the $x - y$ plane, given by $\Delta y = k\Delta x$ with some fixed k. Along this direction, the change of the function (5.86) reads

$$\Delta z = \frac{1}{2}(\Delta x)^2\left[F_{xx} + 2F_{xy}k + F_{yy}k^2\right].$$

The square polynomial in the brackets has two real roots $k_{1,2} = \left(-F_{xy} \pm \sqrt{D}\right)/F_{yy}$ since its determinant $D = -\mathcal{D} = F^2_{xy} - F_{xx}F_{yy}$ is positive. This means that the parabola $g(k) = F_{xx} + 2F_{xy}k + F_{yy}k^2$ definitely crosses the k axis at two points k_1 and k_2. Then, depending on the sign of F_{yy}, the parabola either has a minimum or maximum as a function of k, but either way, it is positive for some values of k and negative for some others. Therefore, going along any direction of the k in which the parabola is positive, Δz increases (as $\Delta z \sim (\Delta x)^2$), while going along other directions in which the parabola is negative we have $\Delta z \sim -(\Delta x)^2$ and it is decreasing. Hence, in some directions the function $z(x, y)$ goes up as we depart from the point (x_0, y_0), as in the other it is going down, i.e. the function has neither minimum nor maximum at this point.

Summarising, the question of whether the function $z = f(x, y)$ has a minimum or a maximum at a point (x_0, y_0) where it has zero partial derivatives is solved in the following way: it is a minimum if

$$F^0_{xx} = f''_{xx} > 0 \quad \text{and} \quad \mathcal{D} = F^0_{xx}F^0_{yy} - \left(F^0_{xy}\right)^2 = \begin{vmatrix} f''_{xx} & f''_{xy} \\ f''_{xy} & f''_{yy} \end{vmatrix} > 0 \quad \text{(Min)} \tag{5.89}$$

at the point (x_0, y_0), while we have a maximum there if

$$F^0_{xx} = f''_{xx} < 0 \quad \text{and} \quad \mathcal{D} = F^0_{xx}F^0_{yy} - \left(F^0_{xy}\right)^2 = \begin{vmatrix} f''_{xx} & f''_{xy} \\ f''_{xy} & f''_{yy} \end{vmatrix} > 0 \quad \text{(Max).} \tag{5.90}$$

The function does not have an extremum if $\mathcal{D} < 0$.

Generalisation of this result to functions of more than two variables will be (rather briefly) formulated in Sect. 1.2.15 of Volume II.

As an example of the application of the developed method, let us investigate the extrema of the function $z = 5x^2 - 3xy + 6y^2$. Setting its first derivatives $z'_x = 10x - 3y$ and $z'_y = -3x + 12y$ to zero, we obtain a possible point for a minimum or a maximum as $x = y = 0$. The function $z = 0$ at this point. To characterise this point, we calculate the second derivatives, $z''_{xx} = 10$, $z''_{xy} = -3$ and $z''_{yy} = 12$, and the determinant $\mathcal{D} = 10 \cdot 12 - (-3)^2 = 111$. Since $\mathcal{D} > 0$ and $z''_{xx} > 0$, we have a minimum at the point $(0, 0)$. It is easy to see that this result makes perfect sense. Indeed, a simple manipulation of our function yields:

$$z = 5\left(x - \frac{3}{10}y\right)^2 + \frac{111}{20}y^2,$$

which clearly shows that the function is positive everywhere around the $(0, 0)$ point and hence has a minimum there.

The case of $\mathcal{D} = 0$ must be specifically investigated as in this case an extremum may or may not exist. To illustrate this point, let us consider two simple functions: $z = x^4 + y^4$ and $h = x^3 + y^3$. In both cases the partial derivatives method suggests the stationary point for a minimum or maximum as $(0, 0)$, where the function is equal to zero and (please, check!) all second order derivatives are equal to zero as well. Hence, $f''_{xx} = \mathcal{D} = 0$ in both cases. However, as is easily seen, the function $z > 0$ away from the point $(0, 0)$ and hence has a well-defined minimum, while the function $h(x, y)$ changes sign when crossing zero (e.g. take $y = 0$ and move x from negative to positive values) and hence has neither minimum nor maximum.

Problem 5.51. *Show that the function* $z = 2x^2 + 2xy + 3y^2 - 2x - y$ *has a minimum at the point* $(1/2, 0)$.

Problem 5.52. *Show that the function* $z = -9x^2 - 5y^2 + 6xy + 12x - 5$ *has a maximum at the point* $(5/6, 1/2)$.

Problem 5.53. *Show that the function* $z = 27x^2 - 25y^2 + 30xy - 60x$ *does not have an extremum at the point* $(5/6, 1/2)$.

Problem 5.54. *Show that the function* $w = x^2 + y^2 + (1 - x - y)^2$ *has a minimum at the point* $(1/3, 1/3)$.

Problem 5.55. (The Least Square Method) *Suppose an experiment is performed N times by measuring a property y for different values of some parameter x. This way N pairs of data points (x_i, y_i) for $i = 1, \ldots, N$ become available. Assuming a linear dependence of y over x, i.e. $y = ax + b$, the "best" fit to the measured data is obtained by minimising the error function*

$$\epsilon = \sum_{i=1}^{N} (y_i - ax_i - b)^2 . \tag{5.91}$$

Demonstrate by direct calculation that the "best" constants a and b are determined by solving the equations

$$S_{xx}a + S_x b = S_{xy} \quad \textit{and} \quad S_x a + Nb = S_y ,$$

where

$$S_x = \sum_{i=1}^{N} x_i , \quad S_y = \sum_{i=1}^{N} y_i , \quad S_{xx} = \sum_{i=1}^{N} x_i^2 \quad \textit{and} \quad S_{xy} = \sum_{i=1}^{N} x_i y_i .$$

Then show that the error $\epsilon(a, b)$ has the minimum at the corresponding values of a and b. [*Hint*: note that

$$NS_{xx} - S_x^2 = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(x_i - x_j\right)^2$$

and is hence always positive.]

5.10.3 Finding Extrema Subject to Additional Conditions

Very often it is necessary to find a minimum (or maximum) of a function "with restrains", i.e. when the variables are constrained by some conditions, for instance, when an extremum of $h = h(x, y, z)$ is sought on a surface specified by the equation $g(x, y, z) = 0$.

Sometimes, this problem can simply be solved by relating one of the variables to the others via the equation of the condition, e.g. $z = z(x, y)$, and hence formulating the problem without a constraint for a function of a smaller number of variables, e.g. $w(x, y) = h(x, y, z(x, y))$. As an example, consider an extremum of $h = x^2 + y^2 + z^2$ on the plane specified by $x + y + z = 1$. From this condition, $z = 1 - x - y$, and hence we have to find an extremum of the function of two variables, $w = x^2 + y^2 + (1 - x - y)^2$, without any constraints. We know how to do this, see Problem 5.54: at $x = y = 1/3$ we have a minimum; at this point $z = 1/3$ as well. Hence, at the point $(1/3, 1/3, 1/3)$ the constrained function $h(x, y, z)$ experiences a minimum.

Problem 5.56. *Find a conditional extremum of the function* $h(x, y, z) = -x^2 - y^2 - z^2 + xy + yz + xz + 3(x + y + z)$ *on the plane* $x + y + z = 1$. [*Answer*: a maximum at $(1/3, 1/3)$.]

Problem 5.57. *Find a conditional extremum of the function* $h(x, y, z) = x^2 + y^2 + z^2 - 2(x + y + z)$ *on the cone* $z = x^2 + y^2$. [*Answer*: a minimum at $\left(x_0, x_0, 2x_0^2\right)$, where $x_0 \simeq 0.7607$ and is determined by solving the algebraic equation $4x^3 - x - 1 = 0$.]

The method above is only useful if the constraint equation(s) can be solved analytically with respect to some of the variables so that these could be eliminated from the function explicitly. In many cases this may be either impossible or inconvenient to do, and hence another method must be developed. To this end, we consider a problem of finding extrema of a function $h = h(x, y, z)$ subject to a general condition $g(x, y, z) = 0$. The condition formally gives $z = z(x, y)$ and hence defines a function of two variables $w(x, y) = h(x, y, z(x, y))$. Therefore, one can use formula (5.38) for calculating the required first order derivatives needed for finding the stationary point(s). Next, formulae (5.39) and (5.40) can be employed for calculating the corresponding second order derivatives and hence characterising the stationary point(s). What is still missing is that we need derivatives of z with respect to x and y, but we do not know an explicit dependence $z = z(x, y)$. This, however, should not be a problem as we can differentiate the equation of the condition with respect to x and y and get the required derivatives. Differentiating both sides of the condition equation $g(x, y, z) = 0$ with respect to x and y, we obtain:

$$\frac{\partial g}{\partial x} + \frac{\partial g}{\partial z}\frac{\partial z}{\partial x} = 0 \quad \Longrightarrow \quad z'_x = -g'_x/g'_z, \tag{5.92}$$

$$\frac{\partial g}{\partial y} + \frac{\partial g}{\partial z}\frac{\partial z}{\partial y} = 0 \quad \Longrightarrow \quad z'_y = -g'_y/g'_z, \tag{5.93}$$

which, of course, are exactly the same as Eqs. (5.36) and (5.37) in Sect. 5.6. The second order derivatives can then be calculated by differentiating the condition equation twice and using Eqs. (5.38)–(5.40):

$$\frac{\partial g}{\partial z}\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 g}{\partial x^2} + 2\frac{\partial^2 g}{\partial x \partial z}\frac{\partial z}{\partial x} + \frac{\partial^2 g}{\partial z^2}\left(\frac{\partial z}{\partial x}\right)^2 = 0$$

$$\Longrightarrow \quad z''_{xx} = -\frac{1}{g'_z}\left[g''_{xx} + 2g''_{xz}z'_x + g''_{zz}\left(z'_x\right)^2\right], \tag{5.94}$$

$$\frac{\partial g}{\partial z}\frac{\partial^2 z}{\partial y^2} + \frac{\partial^2 g}{\partial y^2} + 2\frac{\partial^2 g}{\partial y \partial z}\frac{\partial z}{\partial y} + \frac{\partial^2 g}{\partial z^2}\left(\frac{\partial z}{\partial y}\right)^2 = 0$$

$$\Longrightarrow \quad z''_{yy} = -\frac{1}{g'_z}\left[g''_{yy} + 2g''_{yz}z'_y + g''_{zz}\left(z'_y\right)^2\right], \tag{5.95}$$

$$\frac{\partial g}{\partial z}\frac{\partial^2 z}{\partial x \partial y} + \frac{\partial^2 g}{\partial x \partial y} + \frac{\partial^2 g}{\partial x \partial z}\frac{\partial z}{\partial y} + \frac{\partial^2 g}{\partial y \partial z}\frac{\partial z}{\partial x} + \frac{\partial^2 g}{\partial z^2}\frac{\partial z}{\partial x}\frac{\partial z}{\partial y} = 0$$

$$\Longrightarrow \quad z''_{xy} = -\frac{1}{g'_z}\left[g''_{xy} + g''_{xz}z'_y + g''_{yz}z'_x + g''_{zz}z'_x z'_y\right]. \tag{5.96}$$

The obtained equations fully solve the problem. Indeed, the stationary point(s) is(are) obtained from Eq. (5.38) yielding $w'_x = h'_x + h'_z z'_x = 0$, so that we can write:

$$h'_x - h'_z\frac{g'_x}{g'_z} = 0 \quad \Longrightarrow \quad \frac{h'_x}{g'_x} = \frac{h'_z}{g'_z},$$

and similarly for the y derivatives. Hence, we can write the required equations for the stationary point(s) as follows:

$$\frac{h'_x}{g'_x} = \frac{h'_y}{g'_y} = \frac{h'_z}{g'_z}. \tag{5.97}$$

These two equations are to be solved together with the condition $g(x, y, z) = 0$ giving one or more stationary points (or none).

As an example, let us consider again the same problem as at the beginning of this section: the function $h = x^2 + y^2 + z^2$ with the constraint on the plane specified by $x + y + z = 1$. Here $g = x + y + z - 1$, and hence $g'_x = g'_y = g'_z = 1$ with all second order derivatives of g equal to zero. Therefore, $z'_x = z'_y = -1$ with all second order derivatives equal to zero as well. The stationary point is obtained by calculating $h'_x = 2x$, $h'_y = 2y$ and $h'_z = 2z$, yielding $2x/1 = 2y/1 = 2z/1$, i.e. $x = y = z$, which, together with the constraint equation, gives the same point as we obtained above using the direct calculation: $(1/3, 1/3, 1/3)$. Now, we need to calculate the second order derivatives to characterise the stationary point. Since the second order derivatives of $z(x, y)$ are equal to zero, we obtain from Eqs. (5.39) and (5.40):

$$w''_{xx} = h''_{xx} + 2h''_{xz}z'_x + h''_{zz}\left(z'_x\right)^2 = 2 + 2 \cdot 0 \cdot (-1) + 2\,(-1)^2 = 4,$$

similarly $w''_{yy} = 4$, and

$$w''_{xy} = h''_{xy} + h''_{xz}z'_y + h''_{yz}z'_x + h''_{zz}z'_x z'_y = 0 + 0 \cdot z'_y + 0 \cdot z'_x + 2\,(-1)\,(-1) = 2,$$

which are again exactly the same as given by the direct calculation performed above, and hence we made the same conclusion that the stationary point found corresponds to a minimum.

Problem 5.58. *Find a conditional extremum of the function $h(x, y, z) = 2x^3y^2z$ on the plane $x + y + z = 1$ using the above method.* [*Answer*: a maximum at $(1/2, 1/3, 1/6)$.]

Problem 5.59. *Find a conditional extremum of the function $h(x, y, z) = x^2 + y^2 + z^2 - 2\,(x + y + z)$ on the cone $z = x^2 + y^2$ using the above method. Compare the result with that obtained in Problem 5.57.*

5.10.4 Method of Lagrange Multipliers

Let us now have a closer look at Eq. (5.97) (essentially, there are two of them) which are to be solved (together with the constraint equations) to give the stationary point(s). Note that sometimes it is more convenient to consider three equations instead of the two by introducing the fourth unknown λ equal to either of the fractions in (5.97):

$$h'_x = \lambda g'_x, \quad h'_y = \lambda g'_y, \quad h'_z = \lambda g'_z. \tag{5.98}$$

The three equations above together with the condition equation are sufficient to find the required four unknowns: x, y, z and λ. Alternatively, we can consider an auxiliary function

$$\Phi(x, y, z) = h(x, y, z) - \lambda g(x, y, z). \tag{5.99}$$

Treating now all three variables x, y and z as being independent, we obtain the same three equations (5.98) for the stationary point(s), e.g. $\Phi'_x = h'_x - \lambda g'_x = 0$, etc. The factor λ is called a Lagrange multiplier, and this method of *Lagrange multipliers* is frequently used in practical calculations since it is extremely convenient: it treats all the variables on an equal footing and also allows the introduction of more than one condition easily. It is also applicable to functions of arbitrary number of variables. Therefore, it is essential that we consider this method for a general case.

Consider a function $w = w\left(x_1, \ldots, x_{n+p}\right)$ of $n+p = N$ variables. We would like to determine all its stationary points subject to p (where $1 \le p \le N-1$) conditions $g_k\left(x_1, \ldots, x_{n+p}\right) = 0$, where $k = 1, \ldots, p$. If the point $\mathbf{x} = \left(x_1, \ldots, x_{n+p}\right)$ corresponds to an extremum of the function w, then the total differential of w at this point must vanish (see Eq. (5.84)):

$$dw = \sum_{i=1}^{n+p} \frac{\partial w}{\partial x_i} dx_i = 0. \tag{5.100}$$

The total differential of every condition must also be zero because $g_k = 0$:

$$dg_k = \sum_{i=1}^{n+p} \frac{\partial g_k}{\partial x_i} dx_i = 0, \quad k = 1, \ldots, p. \tag{5.101}$$

Therefore, an arbitrary linear combination of the differentials will also be zero:

$$dw + \sum_{k=1}^{p} \lambda_k dg_k = \sum_{i=1}^{n+p} \left[\frac{\partial w}{\partial x_i} + \sum_{k=1}^{p} \lambda_k \frac{\partial g_k}{\partial x_i} \right] dx_i = \sum_{i=1}^{n+p} \frac{\partial \Phi}{\partial x_i} dx_i = 0, \tag{5.102}$$

where λ_k (with $k = 1, \ldots, p$) are arbitrary numbers and we introduced an auxiliary function

$$\Phi = w\left(x_1, \ldots, x_N\right) + \sum_{k=1}^{p} \lambda_k g_k\left(x_1, \ldots, x_N\right). \tag{5.103}$$

We determine the multipliers λ_k by assuming that the last p variables $x_{n+1}, \ldots, x_{n+p}$ depend on the first n of them (i.e. on $x_1, \ldots, x_n$), and hence the latter can be considered as independent. Then, we request that the contribution to the differential $d\Phi$ in (5.102) due to each "dependent" variable dx_i (for $i = n+1, n+2, \ldots, n+p$) to be individually equal to zero, i.e. that

$$\frac{\partial w}{\partial x_i} + \sum_{k=1}^{p} \lambda_k \frac{\partial g_k}{\partial x_i} = \frac{\partial \Phi}{\partial x_i} = 0, \quad i = n+1, \ldots, n+p. \tag{5.104}$$

Then, we are left with the condition

$$\sum_{i=1}^{n} \left[\frac{\partial w}{\partial x_i} + \sum_{k=1}^{p} \lambda_k \frac{\partial g_k}{\partial x_i} \right] dx_i = \sum_{i=1}^{n} \frac{\partial \Phi}{\partial x_i} dx_i = 0.$$

Since the first n variables $x_1, \ldots, x_n$ are independent and the above expression should be zero for any dx_i, we arrive at the n equations:

$$\frac{\partial \Phi}{\partial x_i} = 0, \quad i = 1, \ldots, n. \tag{5.105}$$

We see by inspecting Eqs. (5.104) and (5.105) that both conditions can be written simply as

$$\frac{\partial \Phi}{\partial x_i} = 0, \quad i = 1, \ldots, n + p, \tag{5.106}$$

i.e. in order to find the stationary point(s), we can consider the auxiliary function (5.103) instead of the original function $w(\mathbf{x})$ ignoring the fact that some variables depend on the others due to constraints (conditions). In other words, all variables can be considered as independent and all constraints (conditions) combined into the auxiliary function (5.103); then, the problem of finding the stationary point(s) is solved by considering Φ. Note that there are exactly $n + 2p = N + p$ equations for N variables $\{x_i\}$ and p Lagrange multipliers: N equations (5.106) and p conditions $g_k\left(x_1, \ldots, x_{n+p}\right) = 0$ (where $k = 1, \ldots, p$).

As an example of using Lagrange multipliers, let us find the minimum distance between a point $M(x_M, y_M, z_M)$ and a plane $Ax + By + Cz = D$. If $P(x, y, z)$ is a point on the plane, then we have to find the minimum of the distance squared,

$$d\,(M, P)^2 = (x - x_M)^2 + (y - y_M)^2 + (z - z_M)^2\,,$$

subject to the additional condition that the point P is on the plane, i.e. it satisfies the equation $Ax + By + Cz = D$. As we only have a single condition, there will be one Lagrange multiplier λ, and we consider the auxiliary function

$$\Phi = (x - x_M)^2 + (y - y_M)^2 + (z - z_M)^2 - \lambda\,(Ax + By + Cz - D)\,,$$

for which the stationary point is determined by three equations:

$$\begin{aligned} \frac{\partial \Phi}{\partial x} &= 2\,(x - x_M) - \lambda A = 0, \quad \frac{\partial \Phi}{\partial y} = 2\,(y - y_M) - \lambda B = 0, \\ \frac{\partial \Phi}{\partial z} &= 2\,(z - z_M) - \lambda C = 0, \end{aligned} \tag{5.107}$$

and the equation of the plane. Multiplying the first equation above by A, the second by B and the third by C, summing them up and using the equation of the plane, we obtain an explicit expression for the Lagrange multiplier:

$$\lambda = -\frac{2\,(Ax_M + By_M + Cz_M - D)}{A^2 + B^2 + C^2}. \tag{5.108}$$

Using now Eqs. (5.107) we can obtain the coordinates of the point P on the plane which is nearest to the point M:

$$x = x_M + \frac{1}{2}\lambda A, \quad y = y_M + \frac{1}{2}\lambda B, \quad z = z_M + \frac{1}{2}\lambda C. \tag{5.109}$$

It is easy to see that the distance d to the plane coincides with that given by Eq. (1.106).

Chapter 6
Functions of Many Variables: Integration

6.1 Double Integrals

We recall that one-dimensional definite integral introduced in Sect. 4.1 is related to functions of a single variable. Here we shall generalise the idea of integration for functions of more than one variable. We start from the simplest case—integration of functions of two variables.

6.1.1 Definition and Intuitive Approach

Consider a thin 2D plate lying in the $x - y$ plane, Fig. 6.1(a), the area density of which is $\rho(x, y)$. We would like to calculate the total mass of the plate. Here the density $\rho(x, y) = \rho(M)$ is the function of a point M inside the plate. If the plate was uniform, then its total mass could be obtained simply by multiplying its surface density ρ by its area. However, since the density may be highly non-uniform across the plate, we divide the plate area by horizontal and vertical lines into small area segments (regions) as shown in Fig. 6.1(b): points x_k divide the x axis with divisions Δx_k, while the points y_j divide the y axis into intervals Δy_j, so each little rectangular segment has a double index with the area $\Delta A_{kj} = \Delta x_k \Delta y_j$. Within each such segment the density can be approximately treated as a constant yielding its mass equal to $\rho\left(M_{kj}\right)\Delta A_{kj}$, where the point $M_{kj}\left(\xi_k, \zeta_j\right)$ is chosen somewhere inside the (kj)-segment with the coordinates ξ_k and ζ_j, where $x_{k-1} \leq \xi_k < x_k$ and $y_{j-1} \leq \zeta_j < y_j$. The total mass of the whole plate is then equal to the sum over all segments:

$$\sum_{k,j} \rho\left(M_{kj}\right)\Delta A_{kj} = \sum_{k,j} \rho\left(\xi_k, \zeta_j\right)\Delta x_k \Delta y_j .$$

L. Kantorovich, *Mathematics for Natural Scientists*, Undergraduate Lecture Notes in Physics, DOI 10.1007/978-1-4939-2785-2_6

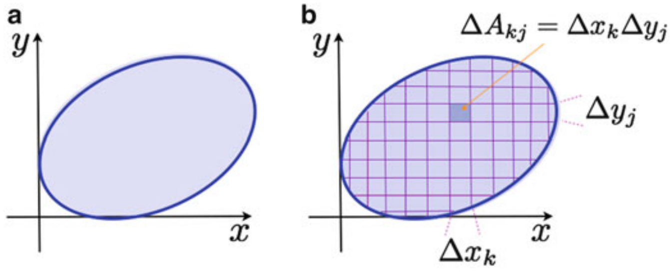
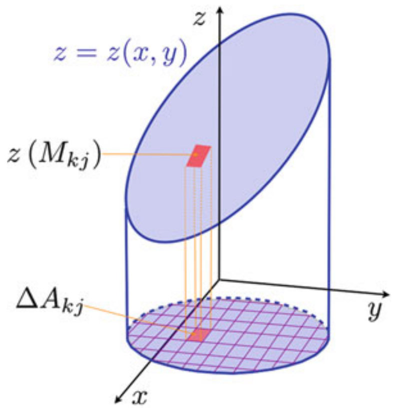


Fig. 6.1 2D plate lying in the $x - y$ plane

Fig. 6.2 For the calculation of the volume of a cylinder oriented along the z axis



Then, we consider all segments becoming smaller and smaller (so that their number increases), and then *in the limit* of infinitesimally small segments we arrive at the *double integral*

$$\int\int_A \rho(x, y)dxdy = \lim_{\max \Delta A_{kj} \to 0} \sum_{k,j} \rho\left(M_{kj}\right) \Delta A_{kj}, \qquad (6.1)$$

which gives the *exact* mass of the plate. Above, the symbol A standing by the symbol of the double integral denotes the two-dimensional region in the $x - y$ plane corresponding to the plate.

Another example where a similar concept is employed is encountered when we would like to calculate the volume of a cylindrical body bounded by the $x - y$ plane at the bottom and some surface $z = z(x, y) > 0$ at the top, see Fig. 6.2. Again, we divide the bottom surface into small regions of the area ΔA_{kj}, choose a point $M_{kj}(\xi_k, \zeta_j)$ inside each such region, calculate the height of the cylinder at that point, $z\left(M_{kj}\right)$, and then estimate the volume as a sum

$$V \simeq \sum_{k,j} z\left(M_{kj}\right) \Delta A_{kj} = \sum_{k,j} z\left(\xi_k, \zeta_j\right) \Delta x_k \Delta y_j,$$

that in the limit of infinitesimally small regions tends to the double integral

$$V = \int\int_A z(x, y)dxdy,$$

that gives the exact volume of the cylindrical body.

Since the double integral is defined as a limit of an integral sum, it has the same properties as a usual (one-dimensional) definite integral. In particular, in order for the integral sum to converge it is *necessary* that the function which is integrated be bounded within the integration region A. Theory of the Darboux sum developed in Sect. 4.2 can be directly applied here as well almost without change.[1] In particular, a (rather mild) sufficient *condition* for a function $f(x, y)$ to be integrable in region A is its continuity, although a wider class of functions can also be integrated. Also, the average value Theorem 4.6 in Sect. 4.2 is valid for double integrals, i.e.

$$\int\int_A f(x, y)dxdy = f(\xi, \zeta)A, \tag{6.2}$$

where A is the total area of region A and the function in the right-hand side is calculated at some point $M(\xi, \zeta)$ belonging to the region A, i.e. the point M lies somewhere inside A.

6.1.2 Calculation via Iterated Integral

Double integrals $\int\int_A f(x, y)dxdy$ can be calculated simply by considering one definite integral after another, and this, as we shall argue presently, may be done in any order in the cases of *proper* integrals (for improper integrals see Sect. 6.1.3).

Indeed, consider a plane region A bounded from the bottom and the top by functions $y_1(x)$ and $y_2(x)$, respectively, as depicted in Fig. 6.3(a). Horizontally, the

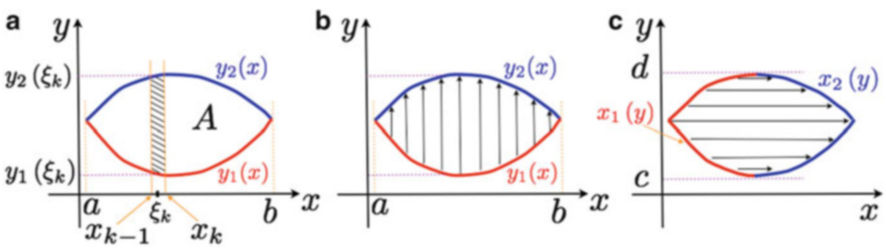


Fig. 6.3 For the discussion concerning the order of integration in double integrals

[1]After all, in Sect. 4.2 we did not actually made use of the fact that what we sum corresponds to a one-dimensional integral; we simply considered the limit of an integral sum.

whole region stretches from $x = a$ to $x = b$. The integral (6.1) is considered as a limit of an integral sum, and obviously, the sum would only make sense if it does not depend on the order in which we add contributions from each region. This is ensured if the integral sum converges, which we assume is the case. However, as we shall show below, a certain *order* in summing up the regions will allow us to devise a simple method of calculating the double integral via usual definite integrals calculated one after another.

Let us divide up the whole integration region A into small regions as we did above with the index k numerating divisions along the x axis, while j along the y. Let m_{kj} and M_{kj} be the minimum and the maximum values of the function $f(x, y)$ within the region (kj), i.e. $m_{kj} \leq f(x, y) \leq M_{kj}$ for any x, y within the region (kj). Then, integrating this inequality with respect to dy over the vertical span of the small (kj) region (i.e. between y_{j-1} and y_j) we arrive at the inequality:

$$m_{kj}\Delta y_j \leq \int_{y_{j-1}}^{y_j} f(\xi_k, y)\, dy \leq M_{kj}\Delta y_j \ , \quad \text{where} \quad x_{k-1} \leq \xi_k < x_k,$$

i.e. the integral in the middle of the inequality is bounded from the bottom and the top.[2] Since this is the case for any j, we can sum these inequalities for all j values, giving:

$$\sum_j m_{kj}\Delta y_j \leq \int_{y_1(\xi_k)}^{y_2(\xi_k)} f(\xi_k, y)\, dy \leq \sum_j M_{kj}\Delta y_j$$
$$\Longrightarrow \quad \sum_j m_{kj}\Delta y_j \leq g(\xi_k) \leq \sum_j M_{kj}\Delta y_j, \tag{6.3}$$

where the integral in the middle is now taken between the two y-values at the lower and upper boundaries of the region, $y_1(\xi_k)$ and $y_2(\xi_k)$, corresponding to the given value of $x = \xi_k$.

The integral in the middle of the inequality is some function $g(x)$ at $x = \xi_k$. Since the inequality is valid for any choice of the value of ξ_k (as long as it lies within the interval between x_{k-1} and x_k), it can be replaced simply by x lying within the interval:

$$\sum_j m_{kj}\Delta y_j \leq g(x) \leq \sum_j M_{kj}\Delta y_j \ , \quad \text{where} \quad x_{k-1} \leq x < x_k.$$

[2]Note that on the left and right of the inequality prior to integration we had constant functions m_{kj} and M_{kj} which do not depend on x, so that integration simply corresponds to multiplying by Δy_j in either part.

Next, we integrate both sides of the above inequality with respect to dx between x_{k-1} and x_k, giving:

$$\left(\sum_j m_{kj}\Delta y_j\right)\Delta x_k \le \int_{x_{k-1}}^{x_k} g\left(x\right)dx \le \left(\sum_j M_{kj}\Delta y_j\right)\Delta x_k.$$

Since this is valid for any k, we can sum up these inequalities over all values of k, which yields:

$$\sum_k\sum_j m_{kj}\Delta x_k\Delta y_j \le \int_a^b g(x)dx \le \sum_k\sum_j M_{kj}\Delta x_k\Delta y_j.$$

On the left and on the right we have the lower and upper Darboux sums. Now we take the limit of all regions $\Delta A_{kj} = \Delta x_k\Delta y_j$ tending to zero; both Darboux sums converge to the value of the double integral, since according to our assumption the double integral exists. Therefore, the integral in the middle of the inequality is bounded by the same value and hence exists and is equal to the double integral:

$$\int\int_A f(x,y)dxdy = \int_a^b g(x)dx$$
$$= \int_a^b\left(\int_{y_1(x)}^{y_2(x)} f(x,y)dy\right)dx = \int_a^b dx\int_{y_1(x)}^{y_2(x)} f(x,y)dy. \tag{6.4}$$

Here we recalled that the $g(x)$ is an integral of $f(x,y)$ over dy taken between $y_1(x)$ and $y_2(x)$.

Hence, in order to calculate the double integral, we have to perform integration over y for a fixed value of x first between the smallest, $y_1(x)$, and the largest, $y_2(x)$, values of the y for the value of x in question, and then integrate the result over all possible values of x. Two ways of writing the double integral shown on the right in Eq. (6.4) demonstrate this explicitly in two different ways (both are frequently used). The order of integration in the last passage in the above equation is from right to left: the x-integration is performed *after* the y-integration.

The method we just described basically suggests that in order to calculate the double integral, we may first integrate over a (hatched) vertical stripe in region A shown in Fig. 6.3(a), which corresponds to the limit of the corresponding integral sum over j for the fixed ξ_k:

$$\Delta S(\xi_k) = \Delta x_k\sum_j f\left(\xi_k,\zeta_j\right)\Delta y_j \Longrightarrow \Delta x_k\int_{y_1(\xi_k)}^{y_2(\xi_k)} f(\xi_k,y)dy = \Delta x_k g\left(\xi_k\right),$$

and then sum up contributions from all such vertical stripes:

$$\sum_k \Delta S(\xi_k) = \sum_k g(\xi_k)\Delta x_k.$$

Finally, taking the limit $\max \Delta x_k \to 0$, we arrive at a definite integral with respect to x taken between a and b which gives the value of the double integral we seek. The arrows in Fig. 6.3(b) serve to indicate symbolically that the first integration is performed in the direction of the arrows (i.e. along y), and then the contributions from all such integrations are summed up by means of taking the x integral.

This way of calculating the double integral via a sequence of definite integrals is called an *iterated integral*. The obtained formula provides us with a practical recipe for calculating double integrals: first, we fix the value of x and calculate the y integral between $y_1(x)$ and $y_2(x)$ (these give the span of y for this particular value of x); then, we perform the x integration between a and b.

Above we assumed that the integration region is such that it can be bounded from below and above by two functions $y_1(x)$ and $y_2(x)$ for which there is only one pair $y_1(\xi_k)$ and $y_2(\xi_k)$ of boundary values for the given division of the x axis ($y_1(\xi_k)$ and $y_2(\xi_k)$ may coincide, e.g. at the boundaries). The consideration is easily generalised to more complex integration regions when this is not possible: in those cases the region should be divided up into smaller regions each satisfying the required condition. Since the integral is a sum, the final result is obtained by simply summing up the results of the integration over each of the regions.

Example 6.1. ► Calculate the volume of the hemisphere of radius R centred at zero with $z \geq 0$, see Fig. 6.4(a).

Solution. For the given x, y, the height of the hemisphere is $z(x,y) = \sqrt{R^2 - x^2 - y^2}$ as the equation of the sphere is $x^2 + y^2 + z^2 = R^2$. Thus, the volume of the hemisphere is obtained from the double integral

$$V = \int\int z(x,y)dxdy = \int_{-R}^{R} dx \int_{-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} \sqrt{R^2 - x^2 - y^2}dy.$$

Here x changes between $-R$ to R in total. For the given x, the variable y spans between $-\sqrt{R^2 - x^2}$ and $+\sqrt{R^2 - x^2}$, see Fig. 6.4(b), since at the base of the

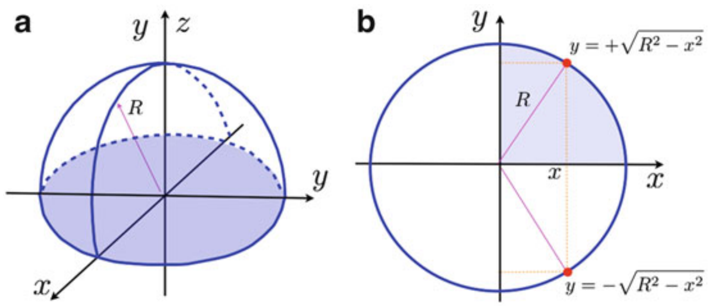


Fig. 6.4 (**a**) A hemisphere bounded by the surfaces $z = 0$ and $z = \sqrt{R^2 - x^2 - y^2}$ used in Example 6.1. (**b**) The base of the hemisphere showing the span of y (the *vertical thin dashed line*) for the given value of x. The *shaded quarter of the circle* is used for the actual integration

hemisphere we have a circle $x^2 + y^2 = R^2$. Also, due to symmetry, we can only consider a quarter of the whole volume in the actual integration (and hence a quarter of the area of the base, i.e. the integration over x will run between 0 and R, while the integration over y between 0 and $+\sqrt{R^2 - x^2}$, see the shaded region in Fig. 6.4(b)) and then introduce a factor of four[3]:

$$\begin{aligned}
V &= 4\int_0^R dx \int_0^{\sqrt{R^2-x^2}} \sqrt{R^2 - x^2 - y^2}dy \\
&= 4\int_0^R dx \left\{ \frac{1}{2}\left(y\sqrt{R^2 - x^2 - y^2} + \left(R^2 - x^2\right)\arcsin\frac{y}{\sqrt{R^2 - x^2}} \right) \right\}_{y=0}^{y=\sqrt{R^2-x^2}} \\
&= 4\int_0^R \frac{1}{2}\left\{ \left(R^2 - x^2\right)\frac{\pi}{2} \right\} dx = \pi\left(R^3 - \frac{R^3}{3} \right) = \frac{2\pi}{3}R^3,
\end{aligned}$$

as it supposed to be: the volume of the whole sphere is two times larger, i.e. $4\pi R^3/3$. ◄

Above, we performed the y integration first, that followed by the x integration. Alternatively, we can do the integrations in the reverse order. Indeed, assuming that the same region in Fig. 6.3(a) can also be described by the boundary lines $x = x_1(y)$ and $x = x_2(y)$ bounding the whole region from the left and right, respectively, see Fig. 6.3(c), we can start by summing up contributions along the x axis first, i.e. from all little regions that have the same value of ζ_j lying somewhere between the division points y_{j-1} and y_j, and then sum up all these contributions as shown schematically in Fig. 6.3(c). Obviously, this way we arrive at an alternative formula:

$$\begin{aligned}
J &= \int\int_A f(x,y)dxdy = \int_c^d r(y)dy \\
&= \int_c^d \left(\int_{x_1(y)}^{x_2(y)} f(x,y)dx \right) dy = \int_c^d dy \int_{x_1(y)}^{x_2(y)} f(x,y)dx, \qquad (6.5)
\end{aligned}$$

where the whole span of y is between c and d. Note that the horizontal integration (over x) gives a function $r(y)$.

Thus, either order of integration can be used (either over x first and then over y, or vice versa) and the same result will be obtained. In some cases useful general relationships can be obtained by using this "symmetry" property. We can also use it to our advantage to simplify the calculation of a double integral, e.g. when it is easier to recognise the bounding functions being $x_1(y)$ and $x_2(y)$ on the left and right rather than $y_1(x)$ and $y_2(x)$ as the lower and upper boundaries, respectively.

Example 6.2. ►Calculate the area bounded by curves shown in Fig. 6.5.

---

[3]We will use the following result: $\int \sqrt{a^2 - t^2}dt = \frac{1}{2}\left[t\sqrt{a^2 - t^2} + a^2\arcsin\frac{t}{a} \right] + C$, see Eq. (4.60).

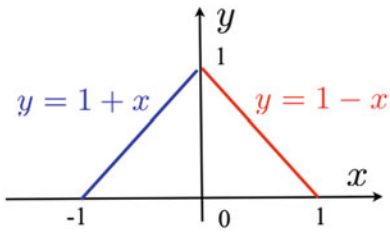


Fig. 6.5 The integration region in Example 6.2

Solution. First of all, we note that it is a triangle, and hence its area should be equal to one (the horizontal base is equal to 2 and the vertical height to 1). Let us now apply the double integrals to do this calculation. We first try the method of Fig. 6.3(b). The integral is to be split into two regions, $-1 \leq x \leq 0$ and $0 \leq x \leq 1$, since the upper boundary line is different in each region as indicated in the figure:

$$\int\int dxdy = \int_{-1}^{0} dx \int_{0}^{1+x} dy + \int_{0}^{1} dx \int_{0}^{1-x} dy$$
$$= \int_{-1}^{0} (1+x)\, dx + \int_{0}^{1} (1-x)\, dx = 1,$$

the correct result. Now, if using the other method of Fig. 6.3(c), we only need to calculate a single double integral as the whole integration region can be dealt with at once. Indeed, if we draw a horizontal line at any value of y between 0 and 1, then it will cross the lines $y = 1 + x$ and $y = 1 - x$ at values of $x = y - 1$ and $x = 1 - y$, respectively; no need to split the integration into two. Hence, we obtain:

$$\int_0^1 dy \int_{y-1}^{1-y} dx = \int_0^1 [1 - y - (y-1)]\, dy = 2\int_0^1 (1-y)\, dy = 1.$$

The result is the same, although the calculation may seem easier. ◄

Problem 6.1. *Calculate the area enclosed between the curves* $y = x^2$ *and* $y = 2 - x^2$ *using the double integral. Use both methods, when either the y or the x-integrations are performed first.* [*Answer*: 8/3.]

Problem 6.2. *Calculate the double integral:*

$$\int_0^1 dy \int_y^1 \frac{ye^x}{x} dx.$$

[*Hint*: Change the order of the integration so that the y integral goes first, paying special attention to the limits, and sketch the way the integration region is "scanned" in both cases as in Fig. 6.3(b, c).][*Answer*: 1/2.]

6.1.3 Improper Integrals

So far we have considered double integrals with finite limits. We have argued that one can do the integration in any order for these integrals: either over x first (the fast variable) and over y second (the slow), or other way round. In practice it is often necessary to deal with double integrals in which one (both) integration(s) is (are) performed using one or both limits to be infinities, i.e. one (both) of the integrals is (are) improper, Sect. 4.5.3. Here we shall briefly consider the questions of whether one can interchange the order of integration if one of the integrals is improper and under which conditions it is possible to do so.

For that let us consider the following improper integral with respect to t containing the function $f(x,t)$ of two variables:

$$F(x) = \int_a^\infty f(x,t)\,dt. \tag{6.6}$$

This integral can be treated as an improper integral depending on the parameter x (cf. Sect. 4.5.2). We shall assume that x may take any value between finite numbers x_1 and x_2, i.e. $x_1 \leq x \leq x_2$. We are going to investigate such improper integrals here.

First of all, let us state an important definition of the so-called *uniform convergence* of improper integrals which is central to our discussion which is to follow. We gave already a definition of the convergence of improper integrals in Sect. 4.5.3. Let us apply this definition to our case: the integral (6.6) is said to converge to $F(x)$ for the given value of x, if for any $\epsilon > 0$ there always exists $A > a$, such that

$$\left| F(x) - \int_a^A f(x,t)dt \right| < \epsilon \quad \text{or} \quad \left| \int_A^\infty f(x,t)dt \right| < \epsilon. \tag{6.7}$$

This condition states that the integral converges to $F(x)$ for the given x, and there is no reason to believe in general that A will not depend on x, in fact, in many cases it will. However, in some cases it won't. If the value of A (for the given ϵ) can be chosen the same for all values of x between x_1 and x_2, then the integral (6.6) is said to converge to $F(x)$ *uniformly* with respect to x.

There is a very simple test one can apply in order to check whether the convergence is uniform or not, due to Weierstrass.

Theorem 6.1 (Weierstrass Test). *If the function $f(x,t)$ which is continuous with respect to both of its variables can be estimated from above as $|f(x,t)| \leq M(t)$ for all values of x and t, and the function $M(t)$ is also continuous, then the integral (6.6) converges uniformly with respect to x.*

Proof. Indeed, what we need to prove is that the residue of the integral appearing in Eq. (6.7) can be estimated by some chosen ϵ using the bottom limit A which is independent of x. Indeed,

$$\left|\int_A^\infty f(x,t)dt\right| \leq \int_A^\infty |f(x,t|\, dt \leq \int_A^\infty M(t)dt = \epsilon.$$

It is seen that for the given ϵ one can always choose such value of A that the integral of $M(t)$ between A and ∞ is equal to ϵ. The value of A chosen in this way does not depend on x, and hence, the integral converges uniformly. **Q.E.D.**

This simple test can be illustrated on the integral $\int_0^\infty e^{-xt}dt$. In this case $f(x,t) = e^{-xt}$ and can be estimated from above for any $x > 1$ as $f(x,t) \leq e^{-t}$. Therefore, $M(t) = e^{-t}$ and

$$\int_A^\infty e^{-t}dt = e^{-A} \quad \Longrightarrow \quad e^{-A} = \epsilon \quad \Longrightarrow \quad A = -\ln \epsilon.$$

It is seen that A can be chosen for the given ϵ in the same way for any value of x for $x > 1$. The integral therefore converges uniformly within this interval of x.

Problem 6.3. *Check uniform convergence of the following improper integrals:*

$$(a)\ \int_\alpha^\infty \frac{dt}{x^2+t^2}\ (\alpha > 0)\,; \qquad (b)\ \int_0^\infty \frac{\sin^2(xt)}{a^2+t^2}dt;$$

$$(c)\ \int_\alpha^\infty \frac{\sin(xt)}{x^2+t^2}dt\ \ (\alpha > 0)\,; \qquad (d)\ \int_0^\infty e^{-x^2t^2}\cos^3(xt)\,dt\ (x > 1);$$

$$(e)\ \int_0^\infty \frac{\sin(xt)}{t^\nu}dt\ \ (\nu > 1)\,; \qquad (f)\ \int_\alpha^\infty \frac{\cos\left(x^2t^3\right)}{(x+t)^\nu}dt\ \ (\nu > 1)\,.$$

[*Answers*: yes in all cases.]

Problem 6.4. *Show that the integral*

$$E_1(x) = \int_x^\infty e^{-t}\frac{dt}{t}$$

converges uniformly with respect to $x > 1$. $E_1(x)$ *is a special function called the exponential integral.* [*Hint*: make a substitution $u = t/x$ and then estimate the function e^{-xu}/u for $u > 1$ by e^{-u}.]

Next we shall prove a theorem which states under which conditions the function $F(x)$ defined by the improper integral (6.6) is continuous.

Theorem 6.2. *If $f(x,t)$ is continuous with respect both of its variables, and the integral (6.6) converges uniformly, then it defines a continuous function $F(x)$.*

Proof. Indeed, the function $F(x)$ is continuous at a point x_0 if $\lim_{x\to x_0} F(x) = F(x_0)$. This means that for any $\epsilon > 0$ one can find a positive δ such that for any x in the proximity to x_0, i.e. $|x - x_0| < \delta$, one has $|F(x) - F(x_0)| < \epsilon$. Let us split the integration between a and $+\infty$ by some A which is larger than a and does not depend on x. We then have:

$$\begin{aligned}
|F(x) - F(x_0)| &= \left| \int_a^\infty f(x,t)dt - \int_a^\infty f(x_0,t)\,dt \right| \\
&= \left| \int_a^A [f(x,t) - f(x_0,t)]\,dt + \int_A^\infty f(x,t)dt - \int_A^\infty f(x_0,t)\,dt \right| \\
&\le \int_a^A |f(x,t) - f(x_0,t)|\,dt + \left| \int_A^\infty f(x,t)dt \right| + \left| \int_A^\infty f(x_0,t)\,dt \right| .
\end{aligned}$$

Since $f(x,t)$ is continuous, for any $\epsilon_1 > 0$ one can find such $\delta > 0$ that $|f(x,t) - f(x_0,t)| < \epsilon_1$. This fact allows estimating the first integral in the right-hand side above as less than or equal to $\epsilon_1 (A - a)$. The other two integrals can be both estimated using some $\epsilon_2 > 0$ due to uniform convergence of the improper integral in question. Hence,

$$|F(x) - F(x_0)| \le \epsilon_1 (A - a) + 2\epsilon_2 = \epsilon,$$

as required. Indeed, for the chosen ϵ one can define separately $\epsilon_2 < \epsilon$ and determine the value of A from the condition of the uniform convergence; then, ϵ_1 is determined from the above equation relating all these quantities. Once ϵ_1 is known, then the value of δ is determined from the continuity of the function $f(x,t)$. **Q.E.D.**

The other theorem we would like to prove is related to the fact that one can change the order of integration in a double integral if one of the integrals is improper with respect to its upper limit.

Theorem 6.3. *If $f(x,t)$ is continuous with respect to both of its variables and the integral (6.6) $F(x)$ converges uniformly, then*

$$\int_{x_1}^{x_2} F(x)dx = \int_a^{\infty} dt \left[\int_{x_1}^{x_2} dx f(x,t) \right] .$$

Proof. Here

$$\int_{x_1}^{x_2} F(x)dx = \int_{x_1}^{x_2} dx \int_a^{\infty} dt f(x,t) .$$

We need to prove that the difference

$$L_A = \int_{x_1}^{x_2} F(x)dx - \int_a^{A} dt \int_{x_1}^{x_2} dx f(x,t)$$

tends to zero as $A \to \infty$. Consider this difference carefully:

$$\begin{aligned} L_A &= \int_{x_1}^{x_2} dx \int_a^{\infty} dt f(x,t) - \int_a^{A} dt \int_{x_1}^{x_2} dx f(x,t) \\ &= \int_{x_1}^{x_2} dx \int_a^{A} dt f(x,t) + \int_{x_1}^{x_2} dx \int_A^{\infty} dt f(x,t) - \int_a^{A} dt \int_{x_1}^{x_2} dx f(x,t) . \end{aligned}$$

The first and the last double integrals are the proper ones, and since a proper double integral does not depend on the order of integration, they cancel out. We obtain

$$L_A = \int_{x_1}^{x_2} dx \int_A^{\infty} dt f(x,t) . \tag{6.8}$$

Let us estimate this double integral:

$$|L_A| = \left| \int_{x_1}^{x_2} dx \int_A^{\infty} dt f(x,t) \right| \le \int_{x_1}^{x_2} dx \underbrace{\left| \int_A^{\infty} dt f(x,t) \right|}_{\le \epsilon_1} \le \epsilon_1 (x_2 - x_1) = \epsilon ,$$

since the integral over t of $f(x,t)$ converges uniformly. Therefore, by choosing arbitrary $\epsilon > 0$, we calculate $\epsilon_1 = \epsilon / (x_2 - x_1)$ and then determine the appropriate value of A to satisfy the inequality above. This means that for any $\epsilon > 0$ one can always find such $A > a$ that $|L_A| < \epsilon$. This simply means that L_A in the expression (6.8) converges to zero in the $A \to \infty$ limit, as required. **Q.E.D.**

Problem 6.5. *Prove that (see also Problem 6.4)*

$$\int_0^\infty E_1(x)dx = 1 \ , \quad where \quad E_1(x) = \int_x^\infty t^{-1}e^{-t}dt.$$

[*Hint*: change the order of the integration.]

The theorems considered above are of course valid for the bottom infinite limit or for both integration limits in one or both integrals being $\pm\infty$. Moreover, they are also valid for the improper integrals due to singularity of the integrand $f(x,t)$ somewhere inside the considered x and t intervals.

6.1.4 Change of Variables: Jacobian

We know from Chap. 4 that the change of variables in single integrals is an extremely powerful tool which enables their calculation in many cases. Remarkably, it is also possible to perform a change of variables in a double integral, but this time we have to change two variables at once. To explain how this can be done is the purpose of this section.

Let us assume that the calculation of a double integral can be simplified by introducing new variables u and v via some so-called *connection equations*:

$$\begin{cases} x = x(u,v) \\ y = y(u,v) \end{cases} . \tag{6.9}$$

If the new variables are chosen properly, the integration function (the integrand) may be significantly simplified; moreover, and more importantly, the integration region on the $x-y$ plane changes to a very different shape in the (u,v) coordinate system. If the latter becomes more like a rectangular, this transformation may allow the calculation of the integral, as illustrated in Fig. 6.6.

In order to change the integration variables $(x,y) \rightarrow (u,v)$, we have to find the relationship between the areas $dxdy$ and $dudv$ in the two systems. To this end,

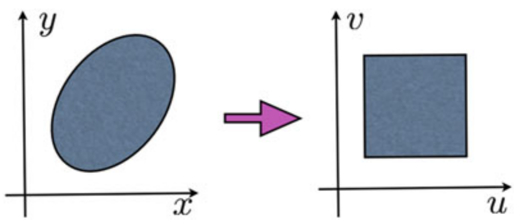


Fig. 6.6 The shape of the integration region in the (u,v) system may be much simpler than in the (x,y) system, and thus may make the calculation of the integral possible

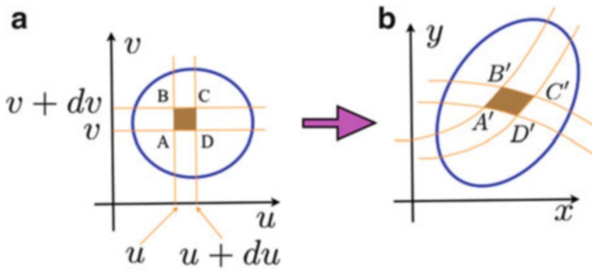


Fig. 6.7 For the calculation of the relationship between the small areas in the two coordinate systems: (**b**) a planar region is mapped onto another planar region in the $u-v$ plane shown in (**a**)

consider a parallelepiped $ABCD$ in the (u, v) system, Fig. 6.7(a). The coordinates of the points of its four vertices are: $A(u, v)$, $B(u, v+dv)$, $C(u+du, v+dv)$ and $D(u+du, v)$, and they form a parallelogram with the area $dudv$. These points will transform into points A', B', C' and D', respectively, in the (x, y) system which coordinates are obtained explicitly via the connection relations (6.9). If the point A' is given by the vector $\mathbf{r} = \mathbf{r}(A) = (x, y, 0)$, where $x = x(u, v)$ and $y = y(u, v)$, then the other three points are:

$$\mathbf{r}\left(B'\right) = \mathbf{r} + \frac{\partial \mathbf{r}}{\partial v}dv\ , \quad \mathbf{r}\left(C'\right) = \mathbf{r} + \frac{\partial \mathbf{r}}{\partial v}dv + \frac{\partial \mathbf{r}}{\partial u}du \quad \text{and} \quad \mathbf{r}\left(D'\right) = \mathbf{r} + \frac{\partial \mathbf{r}}{\partial u}du. \tag{6.10}$$

Indeed, e.g. the point B' is the image of the point B in the (u, v) system and is obtained by advancing point A by dv there; therefore,

$$x\left(B'\right) = x\left(u, v+dv\right) = x(u, v) + \frac{\partial x}{\partial v}dv,$$

and similarly for the y coordinate. Therefore, in the vector form we obtain the first relation in (6.10). All other points are obtained accordingly, and in the vector notations their coordinates are given by Eq. (6.10). Once we know the points of the chosen small area in the (x, y) system, which can approximately be considered as a parallelogram, we can calculate the vectors corresponding to its sides:

$$\overrightarrow{A'B'} = \mathbf{r}\left(B'\right) - \mathbf{r}\left(A'\right) = \frac{\partial \mathbf{r}}{\partial v}dv = \left(\frac{\partial x}{\partial v}dv, \frac{\partial y}{\partial v}dv, 0\right)$$

and

$$\overrightarrow{A'D'} = \mathbf{r}\left(D'\right) - \mathbf{r}\left(A'\right) = \frac{\partial \mathbf{r}}{\partial u}du = \left(\frac{\partial x}{\partial u}du, \frac{\partial y}{\partial u}du, 0\right),$$

$$\overrightarrow{B'C'} = \frac{\partial \mathbf{r}}{\partial u}du = \left(\frac{\partial x}{\partial u}du, \frac{\partial y}{\partial u}du, 0\right) = \overrightarrow{A'D'}$$

and

$$\overrightarrow{D'C'} = \frac{\partial \mathbf{r}}{\partial v} dv = \left(\frac{\partial x}{\partial v} dv, \frac{\partial y}{\partial v} dv \right) = \overrightarrow{A'B'},$$

Since the shaded region in the (x, y) system, which is the image of the corresponding shaded rectangular region in the (u, v) system in Fig. 6.7, is a parallelogram, its area can be calculated using the absolute value of the vector product of its two adjacent sides:

$$\overrightarrow{A'B'} \times \overrightarrow{A'D'} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ (\partial x/\partial v)\, dv & (\partial y/\partial v)\, dv & 0 \\ (\partial x/\partial u)\, du & (\partial y/\partial u)\, du & 0 \end{vmatrix} = \begin{vmatrix} (\partial x/\partial v)\, dv & (\partial y/\partial v)\, dv \\ (\partial x/\partial u)\, du & (\partial y/\partial u)\, du \end{vmatrix} \mathbf{k},$$

and hence the area $dA = \left| \overrightarrow{A'B'} \times \overrightarrow{A'D'} \right|$ becomes:

$$dA = \begin{vmatrix} (\partial x/\partial v)\, dv & (\partial y/\partial v)\, dv \\ (\partial x/\partial u)\, du & (\partial y/\partial u)\, du \end{vmatrix} = \left| \frac{\partial x}{\partial v}\frac{\partial y}{\partial u} - \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} \right| dudv = |J(u, v)|\, dudv,$$

where

$$J(u, v) = \frac{\partial(x, y)}{\partial(u, v)} = \frac{\partial x}{\partial v}\frac{\partial y}{\partial u} - \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} = \begin{vmatrix} \partial x/\partial v & \partial y/\partial v \\ \partial x/\partial u & \partial y/\partial u \end{vmatrix} \tag{6.11}$$

is called *Jacobian*[4] of the transformation. The notation $\frac{\partial(x,y)}{\partial(u,v)}$ for the Jacobian is very convenient as it shows symbolically how it should be calculated. Thus, the areas in the two systems, which correspond to each other by means of the transformation (6.9), are related via the Jacobian:

$$dA = \left| \frac{\partial(x, y)}{\partial(u, v)} \right| dudv, \tag{6.12}$$

and the change of the variables in the double integral is thus performed using the following rule:

$$\int\int_{\Sigma} f(x, y) dxdy = \int\int_{\Delta} f\left(x(u, v), y(u, v)\right) \left| \frac{\partial(x, y)}{\partial(u, v)} \right| dudv, \tag{6.13}$$

where the region Σ in the (x, y) system transforms into the region Δ in the (u, v) system. Note that the *absolute value* of the Jacobian must be used in the formula since the area must be positive.

As an example, let us calculate the integral $I = \int\int_D xydxdy$ over the region D shown in Fig. 6.8(b). We shall transform the integral to the polar coordinates (r, ϕ) in

[4] Named after Carl Gustav Jacob Jacobi.

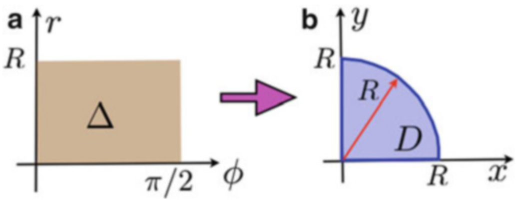


Fig. 6.8 A quarter of a circle in the (x, y) system (**b**) becomes a rectangular in the (r, ϕ) system of the polar coordinates (**a**)

which the integration region D (a quarter of a circle) turns into a rectangle with sides R and $\pi/2$, shown in Fig. 6.8(a). Indeed, if we take any value of r between 0 and R within region D in the $x - y$ plane, we shall draw a quarter of a circle by changing the angle ϕ from 0 to $\pi/2$. Hence, the span of ϕ is the same for any r, and hence in the (r, ϕ) coordinate system we shall have a rectangular. The old coordinates are related to the new ones via $x = r\cos\phi$, $y = r\sin\phi$, so that the Jacobian can be calculated as

$$\frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \partial x/\partial r & \partial y/\partial r \\ \partial x/\partial\phi & \partial y/\partial\phi \end{vmatrix} = \begin{vmatrix} \cos\phi & \sin\phi \\ -r\sin\phi & r\cos\phi \end{vmatrix} = r\cos^2\phi + r\sin^2\phi = r,$$

and the double integral becomes:

$$I = \int\int (r\cos\phi)(r\sin\phi)(rdrd\phi)$$

with r changing from 0 to R and ϕ from 0 to $\pi/2$. The integrand is a product of two functions, the limits are constants, so that it does not matter which integral is calculated first:

$$I = \left(\int_0^R r^3 dr\right)\left(\int_0^{\pi/2} \cos\phi\sin\phi d\phi\right) = \left(\frac{R^4}{4}\right)\left.\frac{\sin^2\phi}{2}\right|_0^{\pi/2} = \frac{R^4}{8}.$$

Let us remember: for the polar coordinate system the Jacobian is equal to r and the area element in this system

$$dA = rdrd\phi. \tag{6.14}$$

It is instructive to obtain this result also using a very simple geometrical consideration illustrated in Fig. 6.9. The area dA of the hatched region $ABCD$ (which can approximately be considered rectangular) is a product of the sides AB and AD, where $AB = dr$ and $AD = r\sin(d\phi) = rd\phi$, so that $dA = rdrd\phi$, exactly as in (6.14).

Example 6.3. ► Find the volume of the hemisphere in Fig. 6.4 using the change of variables method.

Solution. It is wise to introduce polar coordinates in this problem as well, since in these coordinates $x^2 + y^2 = r^2$. The Jacobian is $J = r$, and the volume becomes:

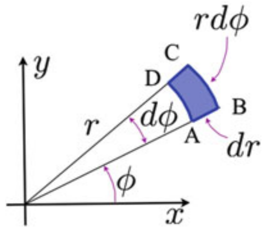


Fig. 6.9 For the derivation of the surface area in polar coordinates

$$V = \int\int \sqrt{R^2 - x^2 - y^2}dxdy = \int\int \sqrt{R^2 - r^2}rdrd\phi$$
$$= \int_0^R \sqrt{R^2 - r^2}rdr \int_0^{2\pi} d\phi = 2\pi \int_0^R \sqrt{R^2 - r^2}rdr.$$

By making the substitution $\zeta = R^2 - r^2$, the r-integral can then be calculated:

$$V = -\pi \int_{R^2}^0 \zeta^{1/2}d\zeta = \pi \left(\frac{\zeta^{3/2}}{3/2}\right)\Big|_0^{R^2} = \frac{2\pi R^3}{3},$$

i.e. the same result as we obtained in Example 6.1 in the previous section using direct integration over x and y. However, the calculation appeared to be extremely simple this time! ◄

Problem 6.6. *Calculate the integral*

$$\int\int \frac{e^{-\alpha\sqrt{x^2+y^2}}}{\sqrt{x^2+y^2}}dxdy$$

over the entire $x-y$ plane using an appropriate change of coordinates. [*Answer*: $2\pi/\alpha$.]

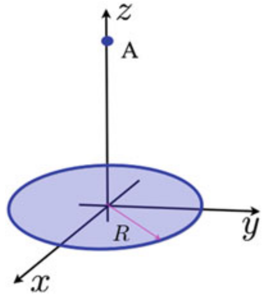
Problem 6.7. *Using polar coordinates, show that*

$$\int\int_S \frac{dxdy}{x^2+y^2} = 2\pi \ln \frac{R_2}{R_1},$$

where region S is bound by two concentric circles of radii R_1 and $R_2 > R_1$, both centred at the origin.

Problem 6.8. *Show using Cartesian coordinates, that the double integral $\int\int_S ydxdy = 0$, where S is the $x \geq 0$ half circle of radius R centred at the origin. Then repeat the calculation in polar coordinates.*

Fig. 6.10 For the calculation of the electrostatic potential from a disk in Problem 6.10



Problem 6.9. *Show that the integral of the previous problem is equal to* $2R^3/3$ *if* S *is the* $y \geq 0$ *half circle. Do the calculation using both methods as well.*

Problem 6.10. *Calculate the electrostatic potential* ϕ *of a circular disk of radius* R *at the point* A *along its symmetry axis, separated by the distance* z *from the disk, as shown in Fig. 6.10. The disk is uniformly charged with the surface charge density* σ. [*Hint:* use polar coordinates when integrating over the disk surface.][*Answer*: $\phi(z) = 2\pi\sigma\left(\sqrt{z^2 + R^2} - z\right)$.]

Problem 6.11. *A thin disc of radius* R *placed in the* $x-y$ *plane with its centre at the origin has the surface mass density* $\rho(x, y) = \rho_0 e^{-\alpha r}$, *where* $r = \sqrt{x^2 + y^2}$. *Assuming that the density decays sufficiently quickly with increasing* r, *i.e. that* $\alpha \gg 1/R$, *calculate the total disc mass using the polar coordinates.* [*Answer*: $2\pi\rho_0/\alpha^2$.]

In physics applications, the following, the so-called *Gaussian integral,* is frequently used:

$$\int_{-\infty}^{\infty} e^{-\alpha x^2} dx = \sqrt{\frac{\pi}{\alpha}}. \tag{6.15}$$

There is a very nice method based on a double integral and polar coordinates which allows this integral to be calculated.[5] Consider a product of two Gaussian integrals, in which we shall use x as the integration variable in the first of them and y in the second:

$$G^2 = \int_{-\infty}^{\infty} e^{-\alpha x^2} dx \int_{-\infty}^{\infty} e^{-\alpha y^2} dy.$$

[5]The corresponding indefinite integral cannot be represented by a combination of elementary analytical functions.

Of course, it does not matter which letters are used as integration variables, but using these makes the argument more transparent: the product of these two integrals can be also considered as a double integral across the whole $x-y$ plane, in which case we can use the polar coordinates in an attempt of calculating it:

$$G^2 = \int\int_{-\infty}^{+\infty} e^{-\alpha(x^2+y^2)} dxdy = \int_0^{2\pi}\int_0^{\infty} e^{-\alpha r^2} r dr d\phi$$
$$= 2\pi \int_0^{\infty} e^{-\alpha r^2} r dr = \begin{vmatrix} t = r^2 \\ dt = 2rdr \end{vmatrix} = \pi \int_0^{\infty} e^{-\alpha t} dt = \frac{\pi}{\alpha},$$

so that $G = \sqrt{\pi/\alpha}$ as required.

Problem 6.12. *Prove that*

$$\int_{-\infty}^{\infty} \exp\left[-a(x-z)^2 - b(x-y)^2\right] dx = \sqrt{\frac{\pi}{a+b}} \exp\left[-\frac{ab}{a+b}(y-z)^2\right]. \tag{6.16}$$

Problem 6.13. *Prove that the diffusion probability for the Brownian motion of a particle in a solution, introduced by Eq. (5.13), satisfies the equation*

$$w(x-x_0, t-t_0) = \int_{-\infty}^{\infty} w\left(x-x', t-t'\right) w\left(x'-x_0, t'-t_0\right) dx'. \tag{6.17}$$

This equation shows that the probability to find the particle at the point x at time t, if it was at x_0 at time t_0, can be represented as a sum of all possible trajectories via an intermediate point x' at some intermediate time t', where $t_0 < t' < t$. The above equation lies at the heart of the so-called Markov stochastic processes and is called Einstein–Smoluchowski–Kolmogorov–Chapman equation.

6.2 Volume (Triple) Integrals

6.2.1 Definition and Calculation

Double integrals were justified by problems based on areas and two-dimensional planar objects. Similarly, triple integrals appear in problems related to volumes of three-dimensional objects. Similarly to all other cases, these are defined as a limit of an integral sum,

$$\sum_{k,j,l} f(\xi_k, \zeta_j, \eta_l)\Delta x_k \Delta y_j \Delta z_l,$$

where ξ_k, ζ_j and η_l are some points within the intervals Δx_k, Δy_j and Δz_l along the Cartesian axes. The integrals are calculated exactly in the same way as two-dimensional integrals by chaining one integration to another (the *iterated integral*). For *proper* integrals it does not matter which integration (over x, y or z) goes first, second and third, this is dictated entirely by convenience of the calculation (the integrand and the shape of the integration region). For instance, the volume integral of a function $f(\mathbf{r}) = f(x, y, z)$ over a volume region V,

$$I = \int\int\int_V f(x, y, z)dxdydz, \tag{6.18}$$

can be calculated in the following sequence of three ordinary definite integrals:

$$I = \int_a^b \left[\int_{y_1(x)}^{y_2(x)} \left(\int_{z_1(x,y)}^{z_2(x,y)} f(x, y, z)dz\right) dy\right] dx = \int_a^b dx \int_{y_1(x)}^{y_2(x)} dy \int_{z_1(x,y)}^{z_2(x,y)} f(x, y, z)dz, \tag{6.19}$$

where we first fix x and y and calculate the integral over z between the corresponding minimum and maximum z values z_1 and z_2 for the given x and y; after that, the y integration is performed for the fixed x value; the x integration is performed last.

Example 6.4. ► Calculate the volume of the hemisphere in Fig. 6.4 using the volume integral.

Solution. The volume is given by:

$$V = \int_0^R dz \int_{-\sqrt{R^2-z^2}}^{\sqrt{R^2-z^2}} dx \int_{-\sqrt{R^2-z^2-x^2}}^{\sqrt{R^2-z^2-x^2}} dy,$$

where the total span of z is from 0 to R. When fixing z (see Fig. 6.11), we arrive at a circular cross section of the hemisphere by the plane parallel to the $x-y$ plane at this z value. The radius of the circle in the cross section is $r = \sqrt{R^2 - z^2}$. So, the span of x (for the given z) will be between $\pm\sqrt{R^2 - z^2}$, while the span of y for the given (fixed) x and z is between $\pm\sqrt{R^2 - z^2 - x^2}$. By performing the calculation of each integral in order from right to left (and the dy integral is trivial), we get[6]:

[6]Once again, we need to use the integral of Eq. (4.60).

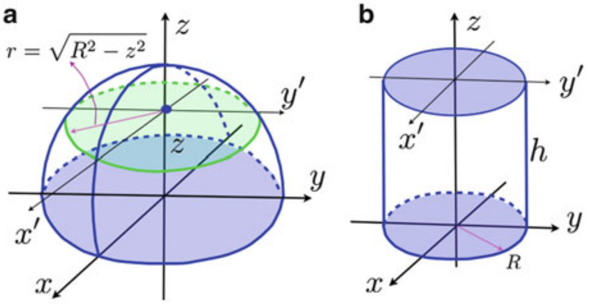


Fig. 6.11 For the calculation of the volume of a hemisphere (**a**) and the moment of inertia of a cylinder (**b**)

$$\begin{aligned}
V &= \int_0^R dz \int_{-\sqrt{R^2-z^2}}^{\sqrt{R^2-z^2}} 2\sqrt{R^2-x^2-z^2}dx \\
&= 2\int_0^R dz\frac{1}{2}\left\{x\sqrt{R^2-x^2-z^2}+\left(R^2-z^2\right)\arcsin\frac{x}{\sqrt{R^2-z^2}}\right\}\Bigg|_{x=-\sqrt{R^2-z^2}}^{x=\sqrt{R^2-z^2}} \\
&= \int_0^R dz\left\{\left(R^2-z^2\right)\arcsin(1)-\left(R^2-z^2\right)\arcsin(-1)\right\} \\
&= \int_0^R dz\left\{\left(R^2-z^2\right)\frac{\pi}{2}-\left(R^2-z^2\right)\left(-\frac{\pi}{2}\right)\right\} = \pi\int_0^R\left(R^2-z^2\right)dz = \pi\frac{2R^3}{3},
\end{aligned}$$

again, the same result as before. Note that this calculation performed in Cartesian coordinates was rather cumbersome. ◄

6.2.2 Change of Variables: Jacobian

Similar to the case of double integrals, we can perform a change of variables in volume integrals as well. Suppose, we would like to introduce new variables (u, v, g) instead of (x, y, z) using *connection equations*:

$$x = x(u, v, g), \quad y = y(u, v, g) \quad \text{and} \quad z = z(u, v, g). \tag{6.20}$$

The integration region V in the (x, y, z) system of coordinates will transform into a region Ω in the (u, v, g) system. To obtain the required relationship between the infinitesimal volumes dV and $d\Omega = dudvdg$ in the two systems, let us consider a cuboid in the (u, v, g) system that is obtained by advancing by du, dv and dg along the corresponding coordinate axes u, v and g from the point $A(u, v, g)$ (corresponding to the point $A'(x, y, z)$ in the (x, y, z)-system), see Fig. 6.12.

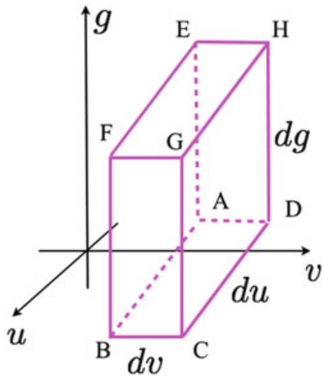


Fig. 6.12 A small cuboid in the (u, v, g) system

In the (x, y, z) system the cuboid $ABCDEFGH$ transforms into the parallelepiped $A'B'C'D'E'F'G'H'$. The vectors corresponding to its three sides and running from the point $A'(x, y, z)$ are:

$$\overrightarrow{A'B'} = \frac{\partial \mathbf{r}}{\partial u}du \ , \quad \overrightarrow{A'D'} = \frac{\partial \mathbf{r}}{\partial v}dv \quad \text{and} \quad \overrightarrow{A'E'} = \frac{\partial \mathbf{r}}{\partial g}dg.$$

Then the volume of the parallelepiped is given by the absolute value of the mixed product of these three vectors (see Eq. (1.41)):

$$dV = \left| \left(\frac{\partial \mathbf{r}}{\partial u}du \cdot \left[\frac{\partial \mathbf{r}}{\partial v}dv \times \frac{\partial \mathbf{r}}{\partial g}dg \right] \right) \right| = \begin{vmatrix} (\partial x/\partial u)\, du & (\partial y/\partial u)\, du & (\partial z/\partial u)\, du \\ (\partial x/\partial v)\, dv & (\partial y/\partial v)\, dv & (\partial z/\partial v)\, dv \\ (\partial x/\partial g)\, dg & (\partial y/\partial g)\, dg & (\partial z/\partial g)\, dg \end{vmatrix}$$
$$= |J(u, v, g)|\, dudvdg, \tag{6.21}$$

where

$$J(u, v, g) = \frac{\partial(x, y, z)}{\partial(u, v, g)} = \begin{vmatrix} \partial x/\partial u & \partial y/\partial u & \partial z/\partial u \\ \partial x/\partial v & \partial y/\partial v & \partial z/\partial v \\ \partial x/\partial g & \partial y/\partial g & \partial z/\partial g \end{vmatrix} \tag{6.22}$$

is the corresponding 3×3 Jacobian. As you can see, for three-dimensions it is defined in a very similar way to the two-dimensional case of Eq. (6.11). Then, the volume integral

$$V = \int\int\int_{\Omega} f\left(\mathbf{r}(u, v, g)\right) \left| \frac{\partial(x, y, z)}{\partial(u, v, g)} \right| dudvdg. \tag{6.23}$$

As an example, let us again calculate the volume of the hemisphere in Fig. 6.4. It is advantageous to use the spherical coordinates (r, θ, ϕ) here,

$$x = r \sin\theta \cos\phi \ , \quad y = r \sin\theta \sin\phi \quad \text{and} \quad z = r\cos\theta,$$

since in this system the hemisphere transforms into a parallelepiped, that can be easily integrated: $0 \le r \le R$, $0 \le \theta \le \pi/2$ and $0 \le \phi < 2\pi$. The Jacobian of the transformation

$$\frac{\partial(x,y,z)}{\partial(r,\theta,\phi)} = \begin{vmatrix} \partial x/\partial r & \partial y/\partial r & \partial z/\partial r \\ \partial x/\partial \theta & \partial y/\partial \theta & \partial z/\partial \theta \\ \partial x/\partial \phi & \partial y/\partial \phi & \partial z/\partial \phi \end{vmatrix}$$

$$= \begin{vmatrix} \sin\theta\cos\phi & \sin\theta\sin\phi & \cos\theta \\ r\cos\theta\cos\phi & r\cos\theta\sin\phi & -r\sin\theta \\ -r\sin\theta\sin\phi & r\sin\theta\cos\phi & 0 \end{vmatrix} = r^2\sin\theta.$$

It is a useful exercise to use the definition of the 3×3 determinant (Sect. 1.6) and verify the above result for the Jacobian. Hence, our integral in the spherical system becomes:

$$V = \int_0^R dr \int_0^{\pi/2} d\theta \int_0^{2\pi} \left| r^2 \sin\theta \right| d\phi = \int_0^R r^2 dr \int_0^{2\pi} d\phi \int_0^{\pi/2} \sin\theta d\theta$$
$$= \frac{R^3}{3} 2\pi \left(-\cos\theta\right)_0^{\pi/2} = \frac{2\pi}{3} R^3.$$

As you see, all three integrals are completely independent yielding the same result as in Examples 6.1, 6.3 and 6.4 again! This is by far the simplest calculation of the sphere volume.

It is useful to remember: the Jacobian for the spherical coordinate system is $r^2 \sin\theta$, and the volume element

$$dV = r^2 \sin\theta dr d\theta d\phi. \tag{6.24}$$

Problem 6.14. *The moment of inertia of a solid body of density $\rho(x,y,z)$ about the z axis is given by*

$$I_z = \int\int\int \left(x^2 + y^2\right) \rho(x,y,z) dx dy dz,$$

and

$$M = \int\int\int \rho(x,y,z) dx dy dz$$

is the body mass. Calculate the mass and the z-moment of inertia of a cylinder of uniform density $\rho = 1$ of radius R and height h whose symmetry axis is running along the z axis, Fig. 6.11(b). [*Hint*: use the cylinder coordinates (r,ϕ,z) with the connection equations $x = r\cos\phi$, $y = r\sin\phi$ and $z = z$.] [*Answer*: $M = \pi R^2 h$, $I_z = \frac{1}{2} M R^2$.]

Problem 6.15. *Calculate the moment of inertia about the z axis of a uniform sphere of radius R and density* $\rho = 1$ *centred at the origin using the cylindrical coordinates* (r, ϕ, z). [*Answer*: $M = \frac{4}{3}\pi R^3$, $I_z = \frac{2}{5}MR^2$.]

Problem 6.16. *A uniform sphere of radius R and density* ρ *is rotated around an axis passing through its centre with the angular velocity* ω. *Show that the kinetic energy of the sphere is* $E_{KE} = (4/15)\,\pi R^5 \rho\omega^2$.

6.3 Line Integrals

6.3.1 Line Integrals for Scalar Fields

The consideration in this section generalises the one we made in Sect. 4.6.1 when we considered the calculation of the length of a curved line.

Consider a scalar function (a *scalar field*) $f(\mathbf{r}) = f(x, y, z)$ that is specified in the three-dimensional space. This could be a linear mass density of a wire ρ, temperature T, potential energy U, and so on. Then, consider a curved line AB that is specified parametrically as

$$x = x(u)\ , \quad y = y(u)\ , \quad z = z(u), \tag{6.25}$$

where u is a parameter (e.g. time t in which case the line may be a particle trajectory in the 3D space). We assume that $a \leq u \leq b$. Then, let us divide the line into n sections by points M_1, M_2, ..., M_{n-1}; the beginning and the end of the line are marked by points $M_0 = A$ and $M_n = B$, respectively, see Fig. 6.13. Link the nearest points with straight lines (to form a broken line as shown in the figure), and choose a point N_i somewhere inside each section $M_{i-1}M_i$. Finally, consider the sum

$$\sum_{i=1}^{n} f(N_i)\Delta_i,$$

where $\Delta_i = |M_{i-1}M_i|$ is the length of the i-th section of the broken line, and take the limit of all $\Delta_i \to 0$ whereby the number of segments tends to infinity. If the limit exists, then it does not depend on the way the curved line is divided up and how the points N_i are chosen within each segment, and the sum converges to what is called a *line integral* between points A and B. It is denoted as follows:

$$I = \int_{AB} f(x, y, z)dl = \int_{AB} f(l)dl, \tag{6.26}$$

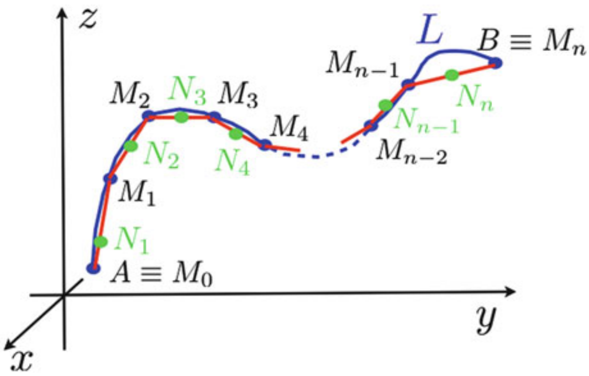


Fig. 6.13 A *curved line* between points A and B specified via the parameter u spanning between a and b is divided by points M_i $(i = 1, \dots, n-1)$ into n curved segments that are then replaced by straight segments. Points $N_1, N_2, \dots, N_n$ are chosen somewhere inside each such segment

where dl is the length of the differential section of the line related to the changes dx, dy and dz of the Cartesian coordinates due to the increment du of the parameter (e.g. $dx = (dx/du)\,du = x'(u)du$ and so on), and $f(l)$ is the value of the function $f(x, y, z)$ on the line at the point where the length l is reached (e.g. measured from the initial point A); obviously, the value of l is directly related to the value of u. This line integral is called the line integral for scalar fields or of the first kind.

If the line is *closed*, then the following notation is frequently used:

$$I = \oint_L f(l)dl.$$

To obtain a useful formula that can be used to calculate the integral in practice, consider the differential length (4.82) (see also Fig. 5.5):

$$dl = \sqrt{dx^2 + dy^2 + dz^2} = \sqrt{(x'(u))^2 + (y'(u))^2 + (z'(u))^2}du,$$

so that a working expression for the integral (6.26) can be obtained in the form of a usual definite integral as:

$$I = \int_a^b f(x(u), y(u), z(u))\sqrt{(x'(u))^2 + (y'(u))^2 + (z'(u))^2}du. \tag{6.27}$$

Note that the above defined line integral possesses all properties of definite integrals. There is only one exception: the integral does *not* change sign upon changing the direction of integration as we deal with the scalar field and the positive length dl:

$$\int_{AB} f(x, y, z)dl = \int_{BA} f(x, y, z)dl. \tag{6.28}$$

To make sure that this is the case in practice, you must follow the rule: if you integrate from smaller to larger values of the u, then the change $du > 0$ and hence dl will automatically be positive. However, if you integrate in the opposite direction, from larger to smaller values of the u, then $du < 0$ and hence the minus sign should then be assigned to it (and hence to the integral) to ensure that $dl > 0$.

Let us now consider some examples.

First, we shall consider the calculation of the length of a circle of radius R. The circle can be specified on the $x - y$ plane as $x = R\cos\phi$ and $y = R\sin\phi$ with ϕ changing from zero to 2π. So in this problem the role of the parameter u is played by the angle ϕ. Note that since this is a planar problem, $z'(\phi) = 0$. Then, the length is obtained as a line integral with $f(x, y, z) = 1$:

$$l = \int_0^{2\pi} \sqrt{(x'(\phi))^2 + (y'(\phi))^2}d\phi = \int_0^{2\pi} \sqrt{R^2\sin^2\phi + R^2\cos^2\phi}$$

$$d\phi = \int_0^{2\pi} Rd\phi = 2\pi R,$$

as expected.

In our second example we shall determine the mass of a non-uniform wire which has a shape of a spiral and is given in polar coordinates as $r = e^{-\phi}$, with the angle ϕ changing from 0 to ∞, see Fig. 1.20(d). We shall assume that the linear mass density of the spiral is proportional to the revolution angle: $\rho(\phi) = \alpha\phi$, where α is a parameter. The mass is given by the integral $M = \int \rho dl$. Introducing polar coordinates, we arrive at the following definite integral:

$$l = \alpha \int_0^{\infty} \phi\sqrt{(x'(\phi))^2 + (y'(\phi))^2}d\phi,$$

where $x(\phi) = r(\phi)\cos\phi = e^{-\phi}\cos\phi$ and $y(\phi) = e^{-\phi}\sin\phi$, so that

$$\begin{aligned} l &= \alpha \int_0^{\infty} \phi\sqrt{(-e^{-\phi}\cos\phi - e^{-\phi}\sin\phi)^2 + (-e^{-\phi}\sin\phi + e^{-\phi}\cos\phi)^2}d\phi \\ &= \alpha \int_0^{\infty} \phi e^{-\phi}\sqrt{2\left(\cos^2\phi + \sin^2\phi\right)}d\phi = \alpha\sqrt{2}\int_0^{\infty} \phi e^{-\phi}d\phi = \alpha\sqrt{2}. \end{aligned}$$

Interestingly, despite the fact that the wire becomes heavier and heavier as it revolves indefinitely around the zero of the coordinate system, its mass remains finite.

In our third example, we will calculate the line integral

$$I = \oint_L (x + 2y)\, dl$$

along the closed path L shown in Fig. 6.14. We split the integral into three contributions:

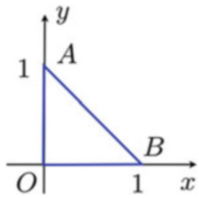


Fig. 6.14 Triangle $\triangle ABO$

$$\oint_L (x+2y)\,dl = \int_{OB} (x+2y)\,dl + \int_{BA} (x+2y)\,dl + \int_{AO} (x+2y)\,dl.$$

We shall use x as a parameter u for the line intervals OB and AB with the length increment

$$dl = \sqrt{dx^2+dy^2} = \sqrt{dx^2\left(1+\left(\frac{dy}{dx}\right)^2\right)} = \sqrt{1+y'(x)^2}dx,$$

while y should be used for AO as the parameter, with the line increment being

$$dl = \sqrt{dx^2+dy^2} = \sqrt{dy^2\left(1+\left(\frac{dx}{dy}\right)^2\right)} = \sqrt{1+x'(y)^2}dy.$$

Remember that the increment of length should always be positive, and the integral does not depend on the direction along the line! Along the line OB: $y = 0$ and $y'(x) = 0$:

$$\int_{OB} (x+2y)\,dl = \int_0^1 (x+2\cdot 0)\sqrt{1+0^2}dx = \int_0^1 xdx = \left.\frac{x^2}{2}\right|_0^1 = \frac{1}{2}.$$

Along the line BA we have: $y = -x+1$, $y'(x) = -1$, so that

$$\int_{BA} (x+2y)\,dl = \int_{AB} (x+2y)\,dl = \int_0^1 (x+2(1-x))\sqrt{1+(-1)^2}dx$$
$$= \sqrt{2}\int_0^1 (2-x)\,dx = 2\sqrt{2} - \sqrt{2}\frac{1}{2} = \frac{3}{\sqrt{2}}.$$

Note that we integrated from $x = 0$ to $x = 1$ to ensure that $dl > 0$. Finally, along the line AO we have: $x = 0$ and $x'(y) = 0$, so that (recall, that in this case y is the parameter):

$$\int_{AO} (x+2y)\,dl = \int_{OA} (x+2y)\,dl = \int_0^1 (0+2y)\,dy = \int_0^1 2ydy = 1.$$

Again, we integrated from the smaller ($y = 0$) to the larger ($y = 1$) value of the y here. Adding up all the three contributions, we obtain the final result: $I = \frac{1}{2} + \frac{3}{\sqrt{2}} + 1 = \frac{3}{2}\left(1 + \sqrt{2}\right)$.

Problem 6.17. *Find the length of a two-dimensional spiral specified parametrically in polar coordinates* (r, ϕ) *as* $r^2 = ae^{-\alpha\phi}$, *where* $a > 0$, $\alpha > 0$ *and* $0 \le \phi \le \infty$. [*Answer*: $2\sqrt{a}\left(1 + \alpha^2/4\right)/\alpha$.]

Problem 6.18. *Find the length of a line specified parametrically as* $x = e^{-\varphi}cos\varphi$, $y = e^{-\varphi}sin\varphi$ *and* $z = e^{-\varphi}$, *where* $0 < \varphi < \infty$. *[*Answer*:* $\sqrt{3}$.*]*

Problem 6.19. *Calculate the length of the parabola* $y = x^2$ *between points* $(x, y) = (0, 0)$ *and* $(1, 1)$. [*Hint*: the integral (4.61) may be of use.] [*Answer*: $\left[2\sqrt{5} + \ln\left(2 + \sqrt{5}\right)\right]/2$.]

Problem 6.20. *Calculate the mass of ten revolutions of the unfolding helix wire given parametrically by* $x = e^{\phi}\cos\phi$, $y = e^{\phi}\sin\phi$ *and* $z = e^{\phi}$ *and with the linear mass density* $\rho = \alpha\phi$. [*Answer*: $\alpha\sqrt{3}\left[(2\pi - 1)e^{20\pi} + 1\right]$.]

6.3.2 Line Integrals for Vector Fields

Above, a scalar field was considered. However, one may also consider a vector function $\mathbf{F}(\mathbf{r}) = \mathbf{F}(x, y, z)$ specified in the 3D space, the so-called *vector field*. For instance, it could be a force acting on a particle moving along some trajectory L. If we would like to calculate the work done by the force on the particle, we may split the trajectory into little *directional straight* segments $\Delta\mathbf{l}$ that would run along the actual trajectory (see again Fig. 6.13), calculate the work $\Delta A = \mathbf{F}(l) \cdot \Delta\mathbf{l}$ done along each of them, and then sum up all the contributions:

$$A \simeq \sum \mathbf{F}(l) \cdot \Delta\mathbf{l}.$$

Then, we can make the segments smaller and smaller, and as the limit of each of the segments length tends to zero, we shall arrive at the line integral

$$I = \int_L \mathbf{F}(l) \cdot d\mathbf{l} = \int_L F_x(l)dx + F_y(l)dy + F_z(l)dz, \tag{6.29}$$

where we wrote the dot product of $\mathbf{F}$ and $d\mathbf{l}$ explicitly in the second part (recall that dx, dy and dz form the three components of the vector $d\mathbf{l}$, see Fig. 5.5). Note that the increments dx, dy and dz are not independent, but correspond to the *change along* *the line*. This line integral is called a line integral for vector fields or of the second kind.

Two methods for the calculation of the line integral may be of use: using a parametric representation of the line L or in Cartesian coordinates. In the first method the line L is specified parametrically as $x = x(u)$, $y = y(u)$ and $z = z(u)$, and hence the integral is calculated simply by noticing that $dx = x'(u)du$, and similarly for dy and dz, so that

$$\int_L F_x(l)dx + F_y(l)dy + F_z(l)dz = \int_L \left[F_x(u)x'(u) + F_y(u)y'(u) + F_z(u)z'(u)\right] du, \tag{6.30}$$

where e.g. $F_x(u) = F_x\,(x(u), y(u), z(u))$ and similarly $F_y(u)$ and $F_z(u)$ are calculated *along* the line L. Of course, the integral can also be split into three independent integrals (since the whole integral is a limit of an integral sum, and the sum can be split into three), each of them to be calculated independently, e.g.

$$\int_L F_x(l)dx = \int_L F_x\,(x(u), y(u), z(u))\, x'(u)du = \int_L F_x(u)x'(u)du,$$

and similarly for the other two integrals. Either of these methods gives a simple recipe to calculate this type of the line integral, and these techniques are the most convenient in practice.

In the second (Cartesian) method, projections of the line L on the Cartesian planes $x - y$, $y - z$ and $x - z$ are used. Indeed, first the line integral is split into three:

$$\int_L F_x(l)dx + F_y(l)dy + F_z(l)dz = \int_L F_x(l)dx + \int_L F_y(l)dy + \int_L F_z(l)dz. \tag{6.31}$$

Here, each of the integrals in the right-hand side is completely independent and can be calculated separately. When calculating any of the three integrals, e.g. the first one $\int_L F_x dx$, we have to assume that the other two variables, y and z, are some functions of the x; this is possible as we integrate *alone* the line, i.e. x is simply used to traverse the line (instead of l or u). Basically, this method is similar to the first one; the main difference is that $u = x$ is used for the calculation of the dx integral, $u = y$ for the dy, and $u = z$ for the dz integrals, respectively, i.e. different parameters are employed for each part of the line integral.

It may seem that the line integral we just introduced is quite different from the one defined in the previous Sect. 6.3.1 by Eq. (6.26). Indeed, for instance, our new integral *does* change sign when the direction of the calculation along the same line is reversed:

$$\int_{AB} \mathbf{F}(l) \cdot d\mathbf{l} = -\int_{BA} \mathbf{F}(l) \cdot d\mathbf{l}, \tag{6.32}$$

since the directional differential length $d\mathbf{l}$ changes its sign in this case. However, the two integrals are still closely related. By noting that $d\mathbf{l} = \mathbf{m}dl$, where $\mathbf{m}$ is the unit vector directed tangentially to the line L at each point and along the direction of integration, and dl is the differential length, we can write:

$$\int_L \mathbf{F}(l) \cdot d\mathbf{l} = \int_L [\mathbf{F}(l) \cdot \mathbf{m}]\, dl, \tag{6.33}$$

where we have now arrived at the familiar line integral for the scalar field $f(l) = \mathbf{F}(l) \cdot \mathbf{m}$ in the right-hand side.

Consider several examples.

Example 6.5. ► Let us calculate the line integral

$$J = \int_{AB} xdx - ydy + zdz$$

along one revolution of a helical line in Fig. 4.6(a) that is specified parametrically as $x = R\cos\phi$, $y = R\sin\phi$ and $z = b\phi$.

Solution. One revolution corresponds to a change of ϕ from zero to 2π. We have:

$$dx = -R\sin\phi d\phi, \quad dy = R\cos\phi d\phi \quad \text{and} \quad dz = bd\phi,$$

so that

$$J = \int_0^{2\pi} [-xR\sin\phi - yR\cos\phi + zb]\, d\phi = \int_0^{2\pi} \left[-2R^2\sin\phi\cos\phi + b^2\phi\right] d\phi$$
$$= -2R^2\int_0^0 \sin\phi d(\sin\phi) + b^2\int_0^{2\pi}\phi d\phi = 0 + b^2\left.\frac{\phi^2}{2}\right|_0^{2\pi} = 2\pi^2 b^2. \quad ◄$$

Example 6.6. ► In our second example we shall calculate the integral

$$J = \oint_L 2xydx + x^2dy$$

along the closed loop shown in Fig. 6.14 which is traced anticlockwise.

Solution. Consider each of the three parts of the path separately. Along OB $dy = 0$ and

$$J_{OB} = \int_0^1 2xydx = 0 \ \text{ since } \ y = 0.$$

Along BA $y = 1 - x$, so, by considering x as the parameter u, we have $dy = y'(x)dx = -dx$, and hence

$$J_{BA} = \int_1^0 \left[2xydx + x^2(-dx)\right] = \int_1^0 \left[2x(1-x) - x^2\right]dx = \left(2\frac{x^2}{2} - 3\frac{x^3}{3}\right)\Bigg|_1^0 = 0.$$

Note that, opposite to the example in the previous subsection, the direction of integration here does matter,[7] and hence along the path BA the integral was calculated from $x = 1$ to $x = 0$ corresponding to the change of x when moving from point B to A.

Finally, along AO we have $dx = 0$ and, therefore,

$$J_{AO} = \int_1^0 x^2 dy = 0 \text{ since } x = 0.$$

Thus, the integral $J = 0$. ◄

Note that, as we shall see later on, this result is not accidental as the integrand is the total differential of the function $F(x, y) = x^2 y$. Indeed,

$$dF = \frac{\partial F}{\partial x}dx + \frac{\partial F}{\partial y}dy = 2xydx + x^2 dy \ .$$

Problem 6.21. *Calculate the line integral* $\int_L \mathbf{F} \cdot d\mathbf{l}$ *along the straight line connecting points* $(0, 0, 0)$ *and* $(1, 1, 1)$ *for the vector field* $\mathbf{F} = (z, y, x)$. [*Answer*: 3/2.]

Problem 6.22. *Calculate the work done by the force field* $F = (0, x, 0)$ *in moving a particle between two points* $A(1, 0, 0)$ *and* $B(0, 1, 0)$ *along a circle of unit radius and centred at the origin.* [*Answer*: $\pi/4$.]

Problem 6.23. *Evaluate the line integral* $\int -x^2 dx + 3xydy$ *along the sides of the triangle ABC in the anticlockwise direction. The triangle lies in the* $x - y$ *plane and its vertices are:* $A(0, 0)$, $B(2, 0)$ *and* $C(2, 2)$. [*Answer*: 4.]

Problem 6.24. *Evaluate* $\int \mathbf{F} \cdot d\mathbf{l}$ *along the parabola* $y = 2x^2$ *between points* $A(0, 0)$ *and* $B(1, 2)$ *if* $\mathbf{F} = xy\mathbf{i} - y^2\mathbf{j}$. [*Answer*: $-13/6$.]

Problem 6.25. *A particle moves in the anticlockwise direction along a square, see Fig. 6.15, by the force* $\mathbf{F} = xy\mathbf{i} + x^2y^2\mathbf{j}$. *Calculate the work done by the force along the closed loop.* [*Answer*: 0.]

[7] Although in this particular case, it does not affect the final zero result!

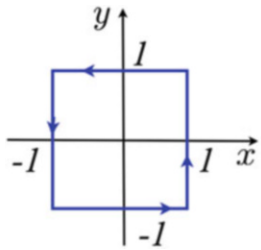


Fig. 6.15 A square region

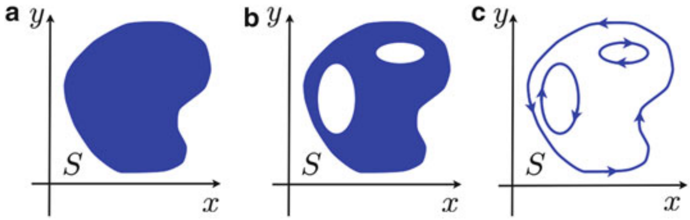


Fig. 6.16 Simply connected (**a**) and not simply connected (**b**) regions; (**c**) directions in the line integral in the Green's formula in the case of the region (**b**) which contains two holes

---

Problem 6.26. *Calculate the line integral* $\oint \mathbf{F} \cdot d\mathbf{l}$ *in the* clockwise *direction along the unit circle centred at the origin, if* $\mathbf{F} = (-y\mathbf{i} + x\mathbf{j}) / \left(x^2 + y^2\right)$. [*Answer*: -2π.]

Problem 6.27. *Evaluate the integral* $\oint_L \mathbf{F} \cdot d\mathbf{l}$ *taken in the anticlockwise direction along the circle of radius R centred at* $(a, 0)$ *if* $\mathbf{F} = xy\mathbf{i} + y^2\mathbf{j}$. [*Answer*: $-\pi a R^2$.]

Problem 6.28. *Calculate the work done by the force* $\mathbf{F} = (-y, x, z)$ *in moving a particle from* $(1,0,0)$ *to* $(-1, 0, \pi)$ *along the helical trajectory specified via* $x = \cos t$, $y = \sin t$, $z = t$. [*Answer*: $\pi\ (1 + \pi/2)$.]

6.3.3 Two-Dimensional Case: Green's Formula

A plane region is said to be *simply connected* if it does not have holes in it, see Fig. 6.16. This may happen if, for instance, a function has singularities at some points which must be removed. It is possible to relate a double integral over some simply connected plane region S to a line integral taken over the boundary of this region.

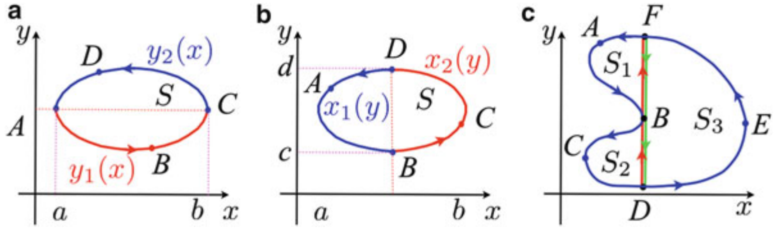


Fig. 6.17 For the derivation of the Green's formula

> **Theorem 6.4 (Green's Theorem).** *For any simply connected region S, the following* Green's formula *is valid:*
>
> $$\int\int_S \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dxdy = \oint_L P(x,y)dx + Q(x,y)dy, \tag{6.34}$$
>
> *where the double integral in the left-hand side is taken over the region S, while the line integral in the right-hand side is taken along its boundary L in the* anticlockwise *direction.*

Proof. Consider a simply connected region S bounded from the bottom and the top by two functions $y = y_1(x)$ and $y = y_2(x)$, Fig. 6.17(a), and two functions $P(x,y)$ and $Q(x,y)$ defined there. Note that the region S we are considering right now is very special: it is such that any line parallel to the coordinate axes would cross the boundary of S in *no more* than two points. Then, the double integral over the region of $\partial P/\partial y$ is

$$\begin{aligned}
\int\int_S \frac{\partial P}{\partial y} dxdy &= \int_a^b dx \int_{y_1(x)}^{y_2(x)} \frac{\partial P(x,y)}{\partial y} dy = \int_a^b \left[P\left(x, y_2(x)\right) - P\left(x, y_1(x)\right)\right] dx \\
&= \int_a^b P\left(x, y_2(x)\right) dx - \int_a^b P\left(x, y_1(x)\right) dx \\
&= -\int_b^a P\left(x, y_2(x)\right) dx - \int_a^b P\left(x, y_1(x)\right) dx \\
&= -\int_{CDA} P(x,y)dx - \int_{ABC} P(x,y)dx = -\oint_{ABCDA} P(x,y)dx,
\end{aligned}$$

where $ABCDA$ is the boundary of the region S. In the left-hand side we have a double integral, in the right—a line integral taken in the *anticlockwise* direction around the region's boundary.

On the other hand, this special region can be considered as bounded by functions $x = x_1(y)$ and $x = x_2(y)$ on the left and the right of the region S instead, as shown in Fig. 6.17(b). In this case, we can similarly write:

$$\begin{aligned}\int\int_S \frac{\partial Q}{\partial x} dxdy &= \int_c^d dy \int_{x_1(y)}^{x_2(y)} \frac{\partial Q(x,y)}{\partial x} dx = \int_c^d \left[Q\left(x_2(y), y\right) - Q\left(x_1(y), y\right) \right] dy \\ &= \int_c^d Q\left(x_2(y), y\right) dy - \int_c^d Q\left(x_1(y), y\right) dy \\ &= \int_c^d Q\left(x_2(y), y\right) dy + \int_d^c Q\left(x_1(y), y\right) dy \\ &= \int_{BCD} Q(x,y)dy + \int_{DAB} Q(x,y)dy = \oint_{ABCDA} Q(x,y)dy,\end{aligned}$$

where the line integral in the right-hand side is taken along the whole closed boundary $ABCDA$ of region S, traversed again in the *anticlockwise* direction. Now we subtract one expression from the other to get the required Eq. (6.34).

The formula we have just proved is rather limited as the derivation has been done for a very special region S shown in Fig. 6.17(a, b). This result is, however, valid for *any* region, since it is always possible to split the arbitrarily complicated region into subregions each of the same type as considered above, i.e. that vertical and horizontal lines cross the boundary of each part at two points at most. The double integrals over all regions sum up into one double integral taken over the whole region, while the line integrals over all internal lines used to break the whole region into subregions would cancel out, so that the line integral only over the outside boundary of S remains. To illustrate this idea of generalising the region shape, consider the region shown in Fig. 6.17(c): the region S is split into three regions S_1, S_2 and S_3 by a vertical line. Let us consider the line integrals along each of the boundaries: $ABFA$ for S_1, $BCDB$ for S_2 and $BDEFB$ for S_3. Thus, for each of the regions the above Green's formula has been proven to be valid:

$$\begin{aligned}\int\int_{S_1} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dxdy &= \oint_{ABFA} P(x,y)dx + Q(x,y)dy, \\ \int\int_{S_2} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dxdy &= \oint_{BCDB} P(x,y)dx + Q(x,y)dy, \\ \int\int_{S_3} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dxdy &= \oint_{BDEFB} P(x,y)dx + Q(x,y)dy.\end{aligned}$$

If we now sum up these equations, we will have the double integral over the whole region S in the left-hand side, and the line integral over all boundaries in the right-hand side. However, the line integral does change sign when the direction is changed. Therefore, the contributions from the added (vertical) line, which is run twice in the opposite directions, will exactly cancel out, and we arrive at the line integral taken only over the actual boundary of S. **Q.D.E.**

Green's formula is actually valid even for regions with holes, but in this case it is necessary to take the line integrals over the *inner boundary* of the region (the holes boundaries) as well going in the clockwise direction as shown in Fig. 6.16(c). The idea of the proof is actually simple: we first consider an extension of the functions $P(x, y)$ and $Q(x, y)$ inside the holes (any will do), then apply Green's theorem to the whole region including the holes and subtract from it Green's formula for each of the holes.

As a simple example which demonstrates the validity of Green's formula, let us verify it for functions $P(x, y) = 2xy$ and $Q(x, y) = x^2$ and the triangle region shown in Fig. 6.14. We know from Example 6.6 of the previous section, that the line integral is equal to zero. Consider now the left-hand side of Green's formula:

$$\int\int \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dxdy = \int\int (2x - 2x)\, dxdy = 0,$$

i.e. the same result as expected from Green's formula.

In our next example, we shall verify the Green's formula for $P(x, y) = x^2 + y^2$ and $Q(x, y) = 4xy$ and the region which is a semicircle of radius R lying in the $y \geq 0$ region as shown in Fig. 6.18. We will first calculate the double integral using polar coordinates $x = r\cos\phi$ and $y = r\sin\phi$ (recall that the corresponding Jacobian is r):

$$\begin{aligned}\int\int \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dxdy &= \int\int (4y - 2y)\, dxdy = \int\int 2ydxdy \\ &= \int\int (2r\sin\phi)\, rdrd\phi = 2\int_0^R r^2 dr \int_0^\pi \sin\phi d\phi \\ &= \frac{2R^3}{3} (-\cos\phi)_0^\pi = \frac{4R^3}{3}.\end{aligned}$$

On the other hand, consider the line integral

$$\oint \left(x^2 + y^2\right) dx + 4xydy$$

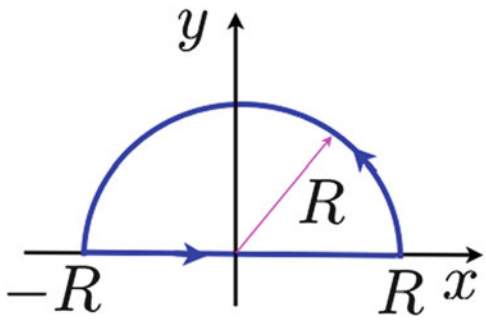


Fig. 6.18 Semicircle region

along the boundary of the region. The boundary consists of two parts: along the semicircle and along the x axis from $-R$ to R. The first integral is most easily calculated using the polar coordinates $x = R\cos\phi$ and $y = R\sin\phi$ (using ϕ as a parameter representing the semicircle):

$$\int_{\text{semicircle}} \left[\left(R^2\right)x'(\phi) + \left(4R^2\cos\phi\sin\phi\right)y'(\phi)\right]d\phi$$
$$= \int_0^{\pi} \left[-R^3\sin\phi + 4R^3\cos^2\phi\sin\phi\right]d\phi = R^3\int_0^{\pi}\left(4\cos^2\phi - 1\right)\sin\phi d\phi$$
$$= \left|\begin{array}{c} t = \cos\phi \\ dt = -\sin\phi d\phi \end{array}\right| = R^3\int_1^{-1}\left(4t^2 - 1\right)(-dt) = R^3\left(\frac{4t^3}{3} - t\right)\Big|_{-1}^{1} = \frac{2R^3}{3}.$$

The line integral along the bottom boundary of the semicircle is most easily calculated in the Cartesian coordinates since $y = 0$ (then $dy = 0$):

$$\int_{\text{bottom}} \left(x^2 + y^2\right)dx + 4xydy = \int_{-R}^{R} x^2 dx = \frac{2R^3}{3}.$$

Summing up the two contributions, we arrive at the same result as given by the double integral above.

Problem 6.29. *Consider a planar figure S with the boundary line L. Using the Green's formula, derive the following alternative expressions for the area of the figure via line integrals:*

$$A = \int\int_S dxdy = \oint_L -ydx = \oint_L xdy = \frac{1}{2}\oint_L -ydx + xdy. \qquad (6.35)$$

[*Hint*: in each case find appropriate functions $P(x, y)$ and $Q(x, y)$.]

Problem 6.30. *Suppose that a planar figure S is specified in polar coordinates as $r = r(\phi)$ and contains the origin inside it. Show that the area of S is*

$$A = \frac{1}{2}\int_{\phi_{\min}}^{\phi_{\max}} r^2(\phi)d\phi, \qquad (6.36)$$

where the integration is performed over the appropriate range of the angle ϕ. Derive this expression using both the first and the last formulae for the area given in (6.35).

Some problems related to the application of Eq. (6.36) can be found in Sect. 4.6.2.

6.3.4 Exact Differentials

We found in Example 6.6 that the line integral over a specific closed path was zero. However, it can be shown that this particular line integral will be zero for *any* closed path. One may wonder when line integrals over a closed loop are equal to zero for any choice of the contour L. More precisely, the necessary and sufficient conditions for this are established by the following:

Theorem 6.5. *Consider a simply connected region S in the $x-y$ plane. Then, the following four statements are equivalent to each other:*

1. *for any closed loop inside region S the integral*

$$\oint_L P(x,y)dx + Q(x,y)dy = 0; \tag{6.37}$$

2. *the line integral $\int_C P(x,y)dx + Q(x,y)dy$ does not depend on the integration curve C, it only depends on its starting and ending points;*
3. *$dU(x,y) = P(x,y)dx + Q(x,y)dy$ is the exact differential, and*
4. *the following condition is satisfied:*

$$\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}. \tag{6.38}$$

Proof. We shall prove this theorem by using the scheme: from 1 follows 2 (i.e. 1→2), then 2→3, 3→4 and then, finally, 4→1 again.

• 1→2. We assume that Eq. (6.37) is satisfied for *any* closed loop L. We have to prove from this that the statement 2 is true. Indeed, take a closed loop $ABCDA$ as shown in Fig. 6.19(a). We have:

$$\oint_{ABCDA} P(x,y)dx + Q(x,y)dy = 0,$$

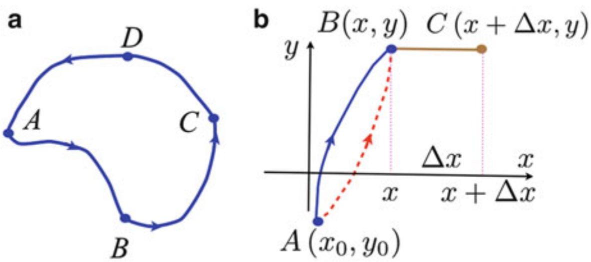


Fig. 6.19 (**a**) A closed loop *ABCDA*. (**b**) To the proof of the step 2→3 of Theorem 6.5

or

$$\int_{ABC} P(x,y)dx + Q(x,y)dy + \int_{CDA} P(x,y)dx + Q(x,y)dy = 0,$$
$$\int_{ABC} P(x,y)dx + Q(x,y)dy - \int_{ADC} P(x,y)dx + Q(x,y)dy = 0,$$
$$\int_{ABC} P(x,y)dx + Q(x,y)dy = \int_{ADC} P(x,y)dx + Q(x,y)dy,$$

i.e. the line integral is the same along the lines ABC and ADC. Since the initial closed loop was chosen arbitrarily, then the integral will be the same along any path from A to C.

• 2→3. Now we are given that the line integral does not depend on the particular path, and we have to prove that $dU = P(x,y)dx + Q(x,y)dy$ is the exact differential. Let us choose a point A (x_0, y_0). Then, for any point $B(x,y)$, we can calculate

$$U(x,y) = \int_{A(x_0,y_0)}^{B(x,y)} P(x_1,y_1)dx_1 + Q(x_1,y_1)dy_1. \tag{6.39}$$

Note that we have used integration variables x_1 and y_1 here to distinguish them from the coordinates of the point B. Now consider the partial derivatives of the function $U(x,y)$. For the x-derivative we have, see Fig. 6.19(b):

$$\frac{\partial U}{\partial x} = \lim_{\Delta x\to 0}\frac{\Delta U}{\Delta x} = \lim_{\Delta x\to 0}\frac{U(x+\Delta x,y) - U(x,y)}{\Delta x}$$
$$= \lim_{\Delta x\to 0}\frac{1}{\Delta x}\left(\int_{A(x_0,y_0)}^{C(x+\Delta x,y)} \dots - \int_{A(x_0,y_0)}^{B(x,y)} \dots\right),$$

where we omitted the integrand for simplicity of notations. The first integral above in the right-hand side is taken along the path from A to C, while the second one goes from A to B. Since, according to our assumption 2, the actual path does not matter, the path $A \to C$ in the first integral can be taken via point B as $A \to B \to C$, with the transition from B to C being passed horizontally along the x axis. Then, the part of the path $A \to B$ cancels out, and we have

$$\frac{\Delta U}{\Delta x} = \frac{1}{\Delta x}\int_{B(x,y)}^{C(x+\Delta x,y)} P(x_1,y_1)dx_1 + Q(x_1,y_1)dy_1 = \frac{1}{\Delta x}\int_{B(x,y)}^{C(x+\Delta x,y)} P(x_1,y)dx_1$$
$$= \frac{1}{\Delta x}\int_{x}^{x+\Delta x} P(x_1,y)dx_1 = P(\zeta,y),$$

where $\zeta = x + \theta\Delta x$ is some point between x and $x+\Delta x$ with $0 < \theta < 1$. Here we have used the (average value) Theorem 4.6 for definite integrals in Sect. 4.2. Note also that above the dy_1 part of the line integral disappeared since along the chosen path $dy_1 = 0$ and y_1 should be set to y. In the limit of $\Delta x \to 0$ the point $\zeta \to x$, hence we have that $\partial U/\partial x = P(x,y)$.

Similarly, one can show that $\partial U/\partial y = Q(x,y)$ (do it!). Therefore, we find that the differential of the function $U(x,y)$ is

$$dU = \frac{\partial U}{\partial x}dx + \frac{\partial U}{\partial y}dy = Pdx + Qdy.$$

We see that it is indeed exactly equal to $Pdx + Qdy$, as required.

• 3→4. This has already been proven for exact differentials in Sect. 5.5 (see Eq. (5.22)), but can easily be repeated here. Since dU is an exact differential, then

$$\frac{\partial}{\partial y}\left(\frac{\partial U}{\partial x}\right) = \frac{\partial P}{\partial y} \quad \text{must be equal to} \quad \frac{\partial}{\partial x}\left(\frac{\partial U}{\partial y}\right) = \frac{\partial Q}{\partial x},$$

as required.

• 4→1. Since the region S is simply connected, we can use the Green's formula (6.34) for any closed loop L that lies inside S:

$$\oint_L Q(x,y)dx + P(x,y)dy = \int\int_{S_L}\left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right)dxdy,$$

where S_L is the region for which L is the bounding curve. However, due to condition 4, Eq. (6.38), the double integral in the right-hand side above is zero for any choice of L, i.e. the closed loop integral for any L is zero, i.e. the statement 1 is true. The theorem has been completely proven. **Q.E.D.**

The condition imposed on region S (simply connected) is very important here as otherwise we will have to use a more general Green's formula (see the comment just after the proof of the Green's formula in Sect. 6.3.3). For instance, region S cannot have holes. A classical example is the vector field

$$\mathbf{A} = -\frac{y}{x^2+y^2}\mathbf{i} + \frac{x}{x^2+y^2}\mathbf{j}.$$

It is seen that

$$\frac{\partial A_x}{\partial y} = \frac{y^2-x^2}{\left(x^2+y^2\right)^2} = \frac{\partial A_y}{\partial x},$$

so that $dU = A_x dx + A_y dy$ is the exact differential and thus the closed loop integral appears to be zero according to the Theorem 6.5 above. Consider, however, a closed loop integral around a circle of a unit radius *centred at zero* using polar coordinates $x = \cos\phi$ and $y = \sin\phi$:

$$\oint A_x dx + A_y dy = \oint -\frac{y}{x^2+y^2}dx + \frac{x}{x^2+y^2}dy$$
$$= \int_0^{2\pi} \left[-\frac{\sin\phi}{1}(-\sin\phi) + \frac{\cos\phi}{1}(\cos\phi) \right] d\phi = \int_0^{2\pi} d\phi = 2\pi.$$

It is not equal to zero in contrast to what the theorem would have suggested! The reason is that the functions $A_x(x,y)$ and $A_y(x,y)$ have a singularity at $x = y = 0$, so that the region S has a hole at the origin and hence is not simply connected. The theorem is not valid since the contour (the unit radius circle) has the origin inside it. If we calculated the line integral avoiding the origin, the integral would be equal to zero as all the conditions of the theorem would be satisfied.

We have already discussed in Sect. 5.5 a method of restoring the function $U(x,y)$ from its exact (total) differential $dU = P(x,y)dx + Q(x,y)dy$. Formula (6.39) suggests a much simpler method of doing the same job. Moreover, as the integral can be taken along an arbitrary path connecting points (x_0, y_0) and (x, y), we can find the simplest route to make the calculation the easiest. For instance, one may perform the integration along the following path consisting of two straight lines parallel to the Cartesian axes: $(x_0, y_0) \to (x, y_0) \to (x, y)$ as shown in Fig. 6.20. Since along the first part $y_1 = y_0$ and is constant, while along the second part $x_1 = x$ is fixed, the integral is transformed into a sum of two ordinary definite integrals:

$$U(x,y) = \int_{x_0}^{x} P(x_1, y_0)dx_1 + \int_{y_0}^{y} Q(x, y_1)dy_1. \tag{6.40}$$

Note that $U(x,y)$ is defined up to an arbitrary constant since the choice of the initial point (x_0, y_0) is arbitrary. The information about the initial point should combine into a single term to serve as such a constant.

Consider calculation of $U(x,y)$ from its differential $dU = 2xydx + x^2dy$ as an example. First, we check that condition (6.38) is satisfied:

$$\frac{\partial P}{\partial y} = \frac{\partial}{\partial y}(2xy) = 2x \text{ and } \frac{\partial Q}{\partial x} = \frac{\partial}{\partial x}\left(x^2\right) = 2x \text{ as well.}$$

This means that dU is indeed the exact differential. Then, using Eq. (6.40), we obtain

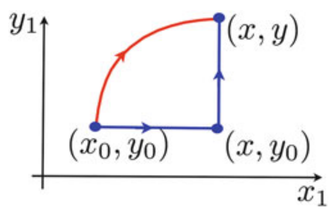


Fig. 6.20 For the calculation of the exact differential

$$U(x,y) = \int_{x_0}^{x} 2x_1 y_0 dx_1 + \int_{y_0}^{y} x^2 dy_1 = 2y_0 \left(\frac{x^2}{2} - \frac{x_0^2}{2} \right) + x^2 (y - y_0) = x^2 y - x_0^2 y_0.$$

Since $x_0^2 y_0$ is arbitrary, it serves as an arbitrary constant C, i.e. we conclude that any function $U(x,y) = x^2 y + C$ is a solution. This is also easily checked by the direct calculation: $\partial U / \partial x = 2xy$ and $\partial U / \partial y = x^2$, exactly as required.

Problem 6.31. *Verify that $x^2 y^3 dx + x^3 y^2 dy$ is the exact differential of some function $U(x,y)$ and hence explain why the integral $\oint_L x^2 y^3 dx + x^3 y^2 dy = 0$ for any contour L.*

Problem 6.32. *Find the function $U(x,y)$ that gives rise to the exact differential $dU = x^2 y^3 dx + x^3 y^2 dy$.* [*Answer*: $U = x^3 y^3 / 3 + C$.]

Problem 6.33. *Calculate the differential of $U(x,y) = x^3 y^3 / 3 + C$ and compare it with the one given in the previous problem.*

Problem 6.34. *Explain why the line integral*

$$U(x,y) = \int_{A(x_0,y_0)}^{B(x,y)} - \sin x_1 \sin y_1 dx_1 + \cos x_1 \cos y_1 dy_1$$

does not depend on the particular path that connects two points $A(x_0, y_0)$ and $B(x,y)$. Show that the function $U = \cos x \sin y + C$.

Problem 6.35. *Verify the Green's formula for $P(x,y) = x^2 + y^2$ and $Q(x,y) = 2x^2 y$ and the region of a square of side 2 shown in Fig. 6.15.*

6.4 Surface Integrals

So far we have discussed double integrals that are calculated on planar regions. However, one may consider a more general problem of the integration over non-planar surfaces. This is the subject of this section. We shall start, however, from a general discussion of how to define a surface and how to obtain its normal vector (or simply a *normal*).

6.4.1 Surfaces

Generally, a line in space can be specified parametrically as $x = x(u)$, $y = y(u)$ and $z = z(u)$, where u is a real parameter. A surface S can also be specified in the same way, but using *two* parameters:

$$x = x(u, v), \quad y = y(u, v) \;\text{ and }\; z = z(u, v), \tag{6.41}$$

so that any point lying on the surface is represented by a vector

$$\mathbf{r}(u, v) = x(u, v)\mathbf{i} + y(u, v)\mathbf{j} + z(u, v)\mathbf{k}. \tag{6.42}$$

Note that we assume a direct correspondence $(u, v) \to (x, y, z)$ here, i.e. only one unique set of (x, y, z) for each set (u, v) (it is allowed that several different sets (u, v) give the same point (x, y, z)).

For instance, a spherical surface of radius R can be specified by equations:

$$x = R\sin\theta\cos\phi, \quad y = R\sin\theta\sin\phi, \quad z = R\cos\theta,$$
$$\text{where } 0 \le \theta \le \pi \quad \text{and} \quad 0 \le \phi < 2\pi.$$

Here θ and ϕ are spherical angles defined as in Fig. 1.19(c). They play the role of the parameters that specify the surface. For instance, if we restrict θ to the interval $0 \le \theta \le \pi/2$ only, then we shall obtain the upper hemisphere, but to specify the whole sphere θ has to be between 0 and π. Note also that for $\theta = 0$ any value of ϕ results in the same point $x = y = 0$, $z = R$.

We say that $\mathbf{r}(u, v)$ describes a *smooth surface* if all three functions (6.41) and their first derivatives

$$\frac{\partial \mathbf{r}}{\partial u} = \frac{\partial x}{\partial u}\mathbf{i} + \frac{\partial y}{\partial u}\mathbf{j} + \frac{\partial z}{\partial u}\mathbf{k} \;\text{ and }\; \frac{\partial \mathbf{r}}{\partial v} = \frac{\partial x}{\partial v}\mathbf{i} + \frac{\partial y}{\partial v}\mathbf{j} + \frac{\partial z}{\partial v}\mathbf{k} \tag{6.43}$$

with respect to u and v are continuous functions in the domain of u and v, and are *nonparallel* everywhere there.

In order to understand what this means, let us calculate a normal to the surface at a point $M_0(x_0, y_0, z_0)$, corresponding to parameters u_0, v_0 as shown in Fig. 6.21. The line $CM_0D = \mathbf{r}(u, v_0)$ ($v = v_0$ is fixed, u alone is allowed to change!) passes through the point M_0, as does the other line $AM_0B = \mathbf{r}(u_0, v)$ obtained by setting $u = u_0$ and allowing only v to change; they both meet at the point M_0. The vectors

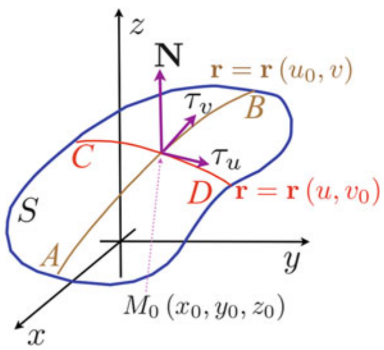


Fig. 6.21 Surface $\mathbf{r}(u, v)$. *Lines* across the surface corresponding to constant values of $u = u_0$ and $v = v_0$ are also indicated

$$\tau_u = \frac{\partial \mathbf{r}(u, v_0)}{\partial u} = \left(x'_u, y'_u, z'_u\right)_{v=v_0} \quad \text{and} \quad \tau_v = \frac{\partial \mathbf{r}(u_0, v)}{\partial v} = \left(x'_v, y'_v, z'_v\right)_{u=u_0}$$

are tangent to the two lines and lie in the *tangent plane* to our surface at point M_0 (see Sect. 5.4). The normal vector to the plane can be calculated as a vector product of the two tangent vectors:

$$\mathbf{N}(u_0, v_0) = \tau_u \times \tau_v = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ x'(u) & y'(u) & z'(u) \\ x'(v) & y'(v) & z'(v) \end{vmatrix}. \tag{6.44}$$

Expanding the determinant along the first row, we obtain:

$$\mathbf{N}(u_0, v_0) = \begin{vmatrix} y'(u) & z'(u) \\ y'(v) & z'(v) \end{vmatrix} \mathbf{i} + \begin{vmatrix} z'(u) & x'(u) \\ z'(v) & x'(v) \end{vmatrix} \mathbf{j} + \begin{vmatrix} x'(u) & y'(u) \\ x'(v) & y'(v) \end{vmatrix} \mathbf{k}$$
$$= \frac{\partial(y, z)}{\partial(u, v)}\mathbf{i} + \frac{\partial(z, x)}{\partial(u, v)}\mathbf{j} + \frac{\partial(x, y)}{\partial(u, v)}\mathbf{k} = J_x(u, v)\mathbf{i} + J_y(u, v)\mathbf{j} + J_z(u, v)\mathbf{k}. \tag{6.45}$$

It is seen that the normal is expressed via the Jacobians that form components of it in Eq. (6.45). Note that the Cartesian coordinates in the definition of the Jacobians together with the Jacobians' subscripts form the cyclic sequence $x \to y \to z \to x \to \ldots$, e.g.

$$J_x = \frac{\partial(y, z)}{\partial(u, v)}.$$

This helps to memorise this formula. Equivalently, one can use formula (6.44) for the whole vector $\mathbf{N}$ instead which also has a form simple enough to remember. Also note that the normal $\mathbf{N}$ may not (and most likely will never) be of unit length, although one can always construct the unit normal $\mathbf{n} = \mathbf{N}/\left|\mathbf{N}\right|$ from it if required.

Thus, we are now at the position to clarify the statement made above about smooth surfaces: the conditions formulated above actually mean that the normal to the smooth surface exists and is not zero. The latter is ensured by the condition that the derivatives (6.43) are non-parallel at each point on the surface (otherwise the normal as the vector product would be zero) and changes continuously as we move the point M_0 across the surface.

Let us now see how this technique can be used to obtain the normal of the side surface of a cylinder of radius R of infinite length running along the z axis. The surface of the cylinder can be described using the polar coordinates by means of the following equations:

$$x = R\cos\phi, \quad y = R\sin\phi \quad \text{and} \quad z = z,$$

i.e. the two parameters for the surface are (ϕ, z) with $0 \leq \phi < 2\pi$ and $-\infty < z < \infty$. To calculate the normal, we should calculate the three Jacobians from (6.45):

$$J_x = \frac{\partial(y,z)}{\partial(\phi,z)} = \begin{vmatrix} R\cos\phi & 0 \\ 0 & 1 \end{vmatrix} = R\cos\phi \ , \ J_y = \frac{\partial(z,x)}{\partial(\phi,z)} = \begin{vmatrix} 0 & -R\sin\phi \\ 1 & 0 \end{vmatrix} = R\sin\phi,$$

$$\text{and} \quad J_z = \frac{\partial(x,y)}{\partial(\phi,z)} = \begin{vmatrix} -R\sin\phi & R\cos\phi \\ 0 & 0 \end{vmatrix} = 0, \tag{6.46}$$

leading to the normal vector

$$\mathbf{N} = (R\cos\phi)\,\mathbf{i} + (R\sin\phi)\,\mathbf{j} + 0\mathbf{k} = x\mathbf{i} + y\mathbf{j} = \mathbf{r}_\perp .$$

It coincides with the vector perpendicular to the cylinder axis for it lies in the plane cutting the cylinder across. For instance, for $\phi = \pi/2$ we have $\mathbf{r}_\perp = R\mathbf{j}$ and for $\phi = \pi$ we obtain $\mathbf{r}_\perp = -R\mathbf{i}$.

In some applications the surface is specified by the equation $z = z(x,y)$. To accommodate this case, we can treat x and y themselves as the parameters u and v in Eq. (6.41), i.e. define the surface by the equations:

$$x = u, \ \ y = v, \ \ z = z(u,v),$$

which gives for the Jacobians:

$$J_x = \frac{\partial(y,z)}{\partial(u,v)} = \begin{vmatrix} 0 & z'_u \\ 1 & z'_v \end{vmatrix} = -z'_u = -\frac{\partial z}{\partial u} = -\frac{\partial z}{\partial x},$$

$$J_y = \frac{\partial(z,x)}{\partial(u,v)} = \begin{vmatrix} z'_u & 1 \\ z'_v & 0 \end{vmatrix} = -z'_v = -\frac{\partial z}{\partial v} = -\frac{\partial z}{\partial y} \ , \quad J_z = \frac{\partial(x,y)}{\partial(u,v)} = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = 1,$$

where we replaced back u, v by x, y, so that

$$\mathbf{N}(x,y) = -\frac{\partial z}{\partial x}\mathbf{i} - \frac{\partial z}{\partial y}\mathbf{j} + \mathbf{k}, \tag{6.47}$$

Recall that exactly the same expression for the surface normal was obtained by us earlier in Eq. (5.18), the result which was derived using a rather different consideration.

Problem 6.36. *Consider a surface specified by the equation* $z(x,y) = 3(x^4 + y^4)$. *Using* x *and* y *as the coordinates* u *and* v *specifying the surface, calculate the appropriate Jacobians and obtain an expression for the surface outward normal as a function of* x *and* y. [*Answer*: $\mathbf{N} = \left(-12x^3, -12y^3, 1\right)$.]

Problem 6.37. *A surface is specified parametrically by coordinates* (u, ϕ) *via the connection relations:* $x = u^2 - 1$, $y = usin\phi$ *and* $z = ucos\phi$. *Calculate the three Jacobians* J_x, J_y *and* J_z *for this surface and hence work out the expression for its normal.* [*Answer*: $\mathbf{N} = \left(-u, 2u^2 \sin\phi, 2u^2 \cos\phi\right)$.]

Some surfaces may consist of pieces of smooth surfaces, e.g. a cube has six such surfaces. The surfaces consisting of such smooth pieces are called *piecewise smooth* surfaces.

Also, the surfaces may have *sides* or *orientations*. Indeed, if we take a membrane, i.e. a surface that caps a closed line L, see Fig. 6.22(a), then two normals $\mathbf{n}_1$ and $\mathbf{n}_2 = -\mathbf{n}_1$ can be defined at each point. However, if we move along the surface, the normal on the upper side of the membrane, $\mathbf{n}_1$, will move continuously on the upper side; it will never come to the lower side. The same is true for the other (lower side) normal $\mathbf{n}_2$. Such surfaces are called *two-sided* or *orientable*. A classic example of a non-orientable (one-sided) surface is the "Möbius[8] strip" (or "leaf") shown in Fig. 6.23. In practically all physical problems the surfaces are two-sided so that we stick to those.

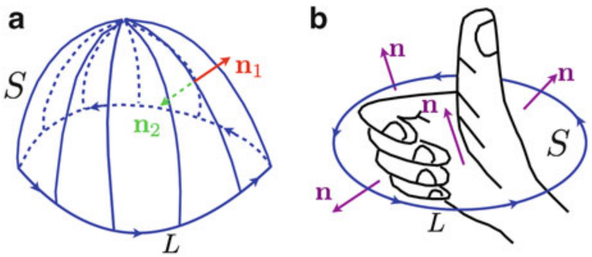
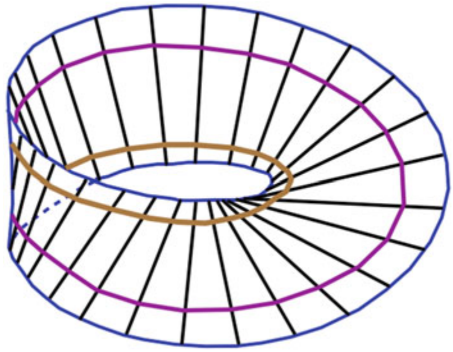


Fig. 6.22 (**a**) The surface S caps a closed loop line L that serves as its boundary. Two oppositely directed surface normal vectors $\mathbf{n}_1$ and $\mathbf{n}_2$ can be defined at each point M of S that specify the two sides (orientations) of the surface. The traverse direction along the line L corresponding to the upper side (the normal vector $\mathbf{n}_1$) is also shown: at each point along the path the surface is on the left which corresponds to the *right-hand rule* shown in (**b**)

[8]The strip was also discovered independently by Johann Benedict Listing.

Fig. 6.23 Möbius strip



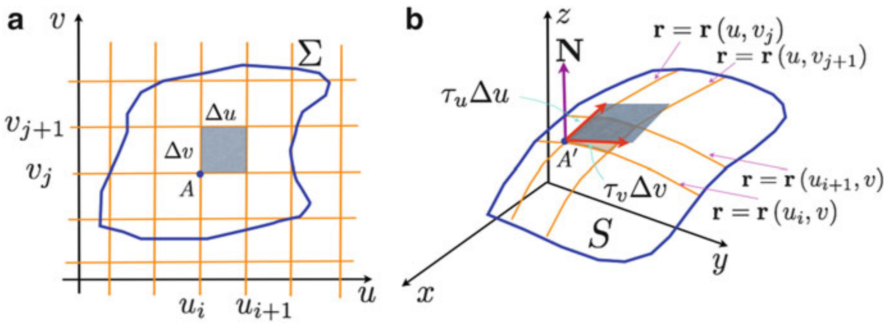


Fig. 6.24 The mapping of the (u, v) system into the (x, y, z) system: to the definition of the area of a general surface: the representation on the $u - v$ plane of the actual surface shown in (**b**) is given in (**a**)

6.4.2 *Area of a Surface*

Now we would like to calculate the area of a surface S specified parametrically as above, i.e. $\mathbf{r} = \mathbf{r}(u, v)$, using two parameters u and v. The problem here is that the surface may not be planar and we have to figure out how to calculate its area using some sort of integration. The driving idea to solve this problem is as follows. We consider the (u, v) coordinate system, in which the values of u and v form a plane surface region Σ, and draw horizontal and vertical lines $u = u_i$ and $v = v_j$ there to form a grid as shown in Fig. 6.24(a). The (u, v) grid lines will form coordinate lines $\mathbf{r}_i(v) = \mathbf{r}(u_i, v)$ and $\mathbf{r}_j(u) = \mathbf{r}(u, v_j)$ on the actual surface S in the (x, y, z) 3D space as schematically shown in Fig. 6.24(b).

Consider now a small surface element $\Delta u \Delta v$ in the (u, v) system around the point $A(u_i, v_j)$, the hatched area in Fig. 6.24(a). It will transform into a small surface element in the surface S; in particular, the point $A \rightarrow A'$, Fig. 6.24(b). For small Δu and Δv the actual surface element can be replaced by the corresponding surface element in the *tangent surface* built by the tangent vectors

$$\tau_u \Delta u = \frac{\partial \mathbf{r}(u, v_j)}{\partial u} \Delta u \quad \text{and} \quad \tau_v \Delta v = \frac{\partial \mathbf{r}(u_i, v)}{\partial v} \Delta v,$$

as shown in Fig. 6.24(b), and its area is

$$\begin{aligned}\Delta A_{ij} &= |\tau_u \Delta u \times \tau_v \Delta v| = |\tau_u \times \tau_v| \, \Delta u \Delta v = \left|\mathbf{N}(u_i, v_j)\right| \Delta u \Delta v \\ &= \sqrt{J_x^2(u_i, v_j) + J_y^2(u_i, v_j) + J_z^2(u_i, v_j)} \Delta u \Delta v,\end{aligned}$$

where we have used Eq. (6.45) for the normal vector to the surface. This way the whole surface S will be represented by small planar tiles smoothly going around it.

Summing areas of all these little sections (or tiles) up, we obtain an estimate for the surface area as the integral sum:

$$A \simeq \sum_{i,j} \Delta A_{ij} = \sum_{i,j} \sqrt{J_x^2(u_i, v_j) + J_y^2(u_i, v_j) + J_z^2(u_i, v_j)} \Delta u \Delta v,$$

which, in the limit of little sections tending to zero, gives the required *exact* formula for the surface area as a double integral over all allowed values of u and v:

$$A = \int\int_{\Sigma} \sqrt{J_x^2(u, v) + J_y^2(u, v) + J_z^2(u, v)} du dv = \int\int_{\Sigma} |\mathbf{N}(u, v)| \, du dv, \tag{6.48}$$

the final result sought for.

Let us illustrate the formula we have just derived by calculating the surface area of a hemisphere shown in Fig. 6.4(a). The hemisphere is specified parametrically by spherical coordinates

$$x = R \sin\theta \cos\phi, \quad y = R \sin\theta \sin\phi \quad \text{and} \quad z = R \cos\theta,$$

with the parameters $0 \leq \theta \leq \pi/2$ and $0 \leq \phi < 2\pi$. The required Jacobians in Eq. (6.45) are calculated to be:

$$J_x = \frac{\partial(y, z)}{\partial(\theta, \phi)} = \begin{vmatrix} R \cos\theta \sin\phi & -R \sin\theta \\ R \sin\theta \cos\phi & 0 \end{vmatrix} = R^2 \sin^2\theta \cos\phi,$$

$$J_y = \frac{\partial(z, x)}{\partial(\theta, \phi)} = \begin{vmatrix} -R \sin\theta & R \cos\theta \cos\phi \\ 0 & -R \sin\theta \sin\phi \end{vmatrix} = R^2 \sin^2\theta \sin\phi,$$

$$\begin{aligned}J_z &= \frac{\partial(x, y)}{\partial(\theta, \phi)} = \begin{vmatrix} R \cos\theta \cos\phi & R \cos\theta \sin\phi \\ -R \sin\theta \sin\phi & R \sin\theta \cos\phi \end{vmatrix} \\ &= R^2 \sin\theta \cos\theta \left(\sin^2\phi + \cos^2\phi\right) = R^2 \sin\theta \cos\theta.\end{aligned}$$

The normal is then given by[9]

$$\mathbf{N}(\theta,\phi) = R^2\left(\sin^2\theta\cos\phi\mathbf{i} + \sin^2\theta\sin\phi\mathbf{j} + \sin\theta\cos\theta\mathbf{k}\right) = (R\sin\theta)\,\mathbf{r}, \tag{6.49}$$

and thus its length is $|\mathbf{N}(\theta,\phi)| = (R\sin\theta)\,R = R^2\sin\theta$. Therefore, the surface of the hemisphere, according to Eq. (6.48), is

$$A = \int\int dA = \int\int\left(R^2\sin\theta\right)d\theta d\phi = R^2\int_0^{\pi/2}\sin\theta d\theta\int_0^{2\pi}d\phi$$
$$= 2\pi R^2\left(-\cos\theta\right)_0^{\pi/2} = 2\pi R^2,$$

as expected: the total surface area of the complete sphere must be two times larger, $4\pi R^2$.

Remember this expression for the differential surface area in the spherical coordinates,

$$dA = R^2\sin\theta d\theta d\phi. \tag{6.50}$$

This result can also be obtained using a rather simple geometrical consideration. In spherical coordinates (and assuming a constant radius R) we have $x = R\sin\theta\cos\phi$, $y = R\sin\theta\sin\phi$ and $z = R\cos\theta$. Consider a spherical region $ABCD$ in Fig. 6.25. It can approximately be considered rectangular with the sides $AD = rd\theta$ and $CD = FD\cdot d\phi = R\sin\theta d\phi$, so that the area becomes $dA = R^2\sin\theta d\theta d\phi$, which is exactly the same expression as obtained above using Jacobians.

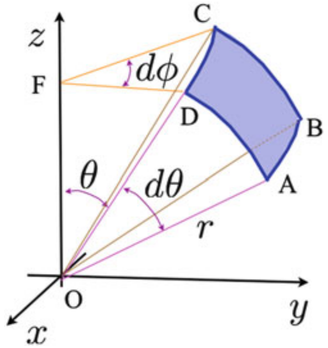


Fig. 6.25 For the geometrical derivation of the differential surface area in spherical coordinates

[9]Note that $\mathbf{N} \sim \mathbf{r}$ as one would intuitively expect for the normal to a sphere!

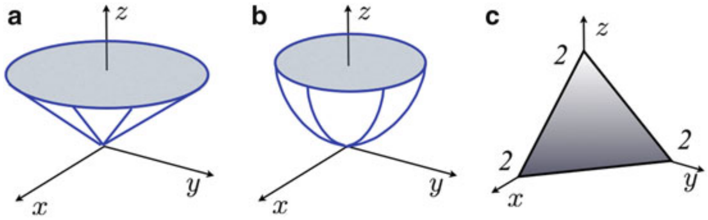


Fig. 6.26 (**a**) Elliptic cone bound from above by the plane $z = R$; (**b**) paraboloid of the height h^2; (**c**) triangle

Problem 6.38. *Show that if the surface is specified by the equation $z = z(x, y)$, then, the surface area (6.48) becomes:*

$$A = \int\int_{\Sigma} \sqrt{1 + \left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2}\, dxdy, \tag{6.51}$$

where Σ is the projection of the surface S on the $x - y$ plane.

As an illustration of application of Eq. (6.51), we shall calculate the surface area of the elliptic cone $z = \sqrt{x^2 + y^2} > 0$ cut by the horizontal plane $z = R$, see Fig. 6.26(a). We have:

$$\frac{\partial z}{\partial x} = \frac{x}{\sqrt{x^2 + y^2}} \quad \text{and} \quad \frac{\partial z}{\partial y} = \frac{y}{\sqrt{x^2 + y^2}},$$

and thus

$$\sqrt{1 + \left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2} = \sqrt{1 + \frac{x^2}{x^2 + y^2} + \frac{y^2}{x^2 + y^2}} = \sqrt{2}.$$

The projection of the surface S on the $x - y$ plane (serving as the $u - v$ plane) is a circle of radius R. Therefore, the surface area

$$A = \int\int \sqrt{2}dxdy = \sqrt{2}\int\int_{\text{circle}} dxdy = \sqrt{2}\pi R^2,$$

where we have used πR^2 for the area of a circle of radius R (the "spread" of x, y).

Problem 6.39. *The paraboloid in Fig. 6.26(b) is described by the equation* $z = x^2 + y^2$. *Write the equation for the surface in a parametric form* $x = x(u, v)$, $y = y(u, v)$ *and* $z = z(u, v)$ *using* x *and* y *as the simplest choice of the parameters* u *and* v. *Obtain the expression for the* outer *normal to the surface* $\mathbf{N}$. [*Hint*: to choose the correct sign for the outer normal, consider a particular point on the surface (choose the simplest one as the surface is one-sided).][*Answer*: $\mathbf{N} = (2x, 2y, -1)$.]

Problem 6.40. *Repeat the previous problem by considering the polar coordinates* (r, ϕ) *as* u *and* v. *Notice that the length of* $\mathbf{N}$ *will change, but not the direction.*

Problem 6.41. *For the paraboloid of Fig. 6.26(b), calculate the surface area for the part of the figure that is bound by the planes* $z = 0$ *and* $z = h^2$. [*Answer*: $\pi\left[\left(4h^2+1\right)^{3/2}-1\right]/4$.]

Problem 6.42. *Repeat the previous problem using the polar coordinates instead of* x *and* y *as* u *and* v. *Note that the same result is obtained although the length of the vector* $\mathbf{N}$ *changed with this choice of the parameters describing the surface.*

Problem 6.43. *The torus in Fig. 4.13 is specified by the parametric equations:*

$$x = (c + R\cos\phi)\cos\theta\ ,\ y = (c + R\cos\phi)\sin\theta \quad \textit{and} \quad z = R\sin\phi, \tag{6.52}$$

where $0 \leq \theta < 2\pi$ *and* $0 \leq \phi < 2\pi$. *Calculate (a) the normal to the torus* $\mathbf{N}(\theta, \phi)$ *and (b) the entire surface area.* [*Answer*: (a) $\mathbf{N} = R(c + R\cos\phi)\left(\cos\phi\cos\theta, -\cos\phi\sin\theta, \sin\phi\right)$; (b) $4\pi^2 cR$.]

6.4.3 Surface Integrals for Scalar Fields

If we are given a scalar field $f(\mathbf{r})$, defined on a surface S, one can define an integral sum

$$\sum_{ij} f(M_{ij})\Delta A_{ij}$$

with ΔA_{ij} being the surface area of the ij-th surface element. Then the expression above in the limit of all little surface areas tending to zero yields the *surface integral*:

$$\int\int_S f(\mathbf{r})dA. \tag{6.53}$$

Here points M_{ij} are on the surface. The area of the surface S we considered above is an example of the surface integral with the unit scalar field $f = 1$ everywhere on the surface. Therefore, the surface integral (6.53) is a natural generalisation of our previous definition. If the surface is specified parametrically via Eq. (6.42), the surface integral is calculated similarly to Eq. (6.48) as follows:

$$\begin{aligned}\int\int_S f(\mathbf{r})dA &= \int\int_\Sigma f(\mathbf{r}(u,v))\sqrt{J_x^2(u,v)+J_y^2(u,v)+J_z^2(u,v)}dudv \\ &= \int\int_\Sigma f(\mathbf{r}(u,v))\,|\mathbf{N}(u,v)|\,dudv. \end{aligned} \tag{6.54}$$

In particular, if the surface is defined via equation $z = z(x,y)$, we then have instead

$$\int\int_S f(\mathbf{r})dA = \int\int_\Sigma f(x,y,z(x,y))\sqrt{1+\left(\frac{\partial z}{\partial x}\right)^2+\left(\frac{\partial z}{\partial y}\right)^2}dxdy. \tag{6.55}$$

As an example, let us evaluate the surface integral $J = \int\int xdA$ over that part of the sphere $x^2+y^2+z^2 = R^2$ that lies in the region $x \geq 0$, $y \geq 0$ and $z \geq 0$ (the first octant). We use spherical coordinates to specify the surface, i.e. using the angles θ and ϕ in place of the two parameters u and v. Then the surface element in the spherical coordinates $dA = R^2 \sin\theta d\theta d\phi$, see Eq. (6.50). In the first octant $0 \leq \phi \leq \pi/2$ and $0 \leq \theta \leq \pi/2$, so that the integral becomes:

$$\begin{aligned}J &= \int\int (R\sin\theta\cos\phi)\left(R^2\sin\theta d\theta d\phi\right) = R^3\int_0^{\pi/2}\sin^2\theta d\theta\int_0^{\pi/2}\cos\phi d\phi \\ &= R^3\left(\frac{\theta}{2}-\frac{1}{4}\sin(2\theta)\right)_0^{\pi/2}(\sin\phi)_0^{\pi/2} = R^3\frac{\pi}{4}\ .\end{aligned}$$

Problem 6.44. *Consider a thin spherical surface layer of radius R which is uniformly charged with the surface charge density σ. Show by calculating the electrostatic potential*

$$\phi(\mathbf{r}) = \sigma\int\int_S \frac{dA'}{|\mathbf{r}-\mathbf{r}'|}$$

produced by the layer at a point $\mathbf{r}$, *that φ is constant anywhere* inside *the sphere (i.e. it does not depend on the position* $\mathbf{r} = (x,y,z)$ *there).* [*Hint*: place the point $\mathbf{r}$ on the z axis as shown by point A in Fig. 6.27 and use the spherical coordinates (θ,ϕ) when integrating over the surface of the sphere.]

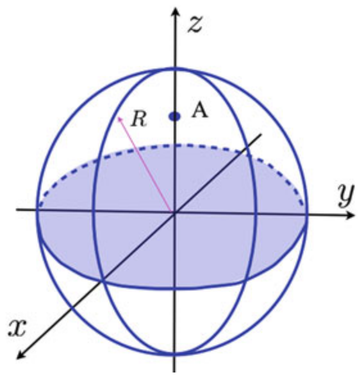


Fig. 6.27 For the calculation of the electrostatic potential from a thin spherical layer at point A of Problem 6.44

Problem 6.45. *Show that the potential of the sphere from the previous problem at the distance z outside the sphere (i.e. for $z > R$) is given by $\phi = Q/z$, where $Q = 4\pi R^2\sigma$ is the total charge of the sphere.*

Problem 6.46. *Consider now the potential*

$$\phi(\mathbf{r}) = \int\int_S \frac{\sigma(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} dA'$$

produced at point $\mathbf{r}$ by (generally) non-uniformly charged sphere S with the radius R and the surface charge density $\sigma(\mathbf{r}) = \alpha\left(x^2+y^2\right)$. In the equation above for the potential points $\mathbf{r}'$ lie on the surface S of the sphere and dA' is its differential surface area. Show that outside the sphere and at the distance D (where $D > R$) from its centre

$$\phi(D) = \frac{8\pi\alpha R^4}{3D}\left(1-\frac{R^2}{5D^2}\right).$$

Show also that the total charge on the sphere $Q = \int\int_S \sigma(\mathbf{r})\, dA = 8\pi\alpha R^4/3$, and hence the potential far away from the sphere $\phi \simeq Q/D$, as it would be for a point charge Q localised in the sphere centre.

6.4.4 Surface Integrals for Vector Fields

Similar to line integrals, which were generalised to vectors fields in Sect. 6.3.2, one can also consider surface integrals for vector fields.

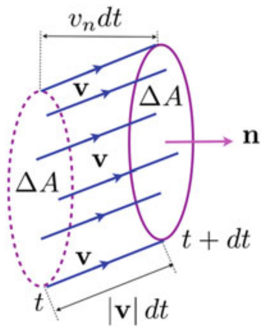


Fig. 6.28 Flow of a liquid through the surface ΔA during time dt

Let $\mathbf{F}(\mathbf{r})$ be a vector field defined on a surface S for which we have a well-defined *unit* normal $\mathbf{n}(M) = \mathbf{N}(M)/\,|\mathbf{N}(M)|$ at each point M belonging to S. Then, a *flux* of $\mathbf{F}$ through the surface S is defined as the following surface integral:

$$\Phi = \int\!\!\int (\mathbf{F}\cdot\mathbf{n})\, dA = \int\!\!\int F_n dA = \int\!\!\int \mathbf{F}\cdot d\mathbf{A}. \tag{6.56}$$

Here, as it is customarily done, we introduced a *directed* surface element $d\mathbf{A} = \mathbf{n}dA$ whose direction is chosen along the vector $\mathbf{n}$, and $F_n = \mathbf{F}\cdot\mathbf{n}$ is the projection of $\mathbf{F}$ on the unit normal vector $\mathbf{n}$. If the flux integral is taken over a *closed* surface S, the notation $\oiint$ for the surface integral is frequently used instead.

To see just one application where this type of the integral appears (many more will be given in Sects. 6.5 and 6.7), let us consider a flow of a liquid. If we take a plane surface ΔA and assume that the velocity distribution of the liquid particles $\mathbf{v}$ across the surface is uniform, then the volume of the liquid flown through the surface during the time dt, Fig. 6.28, is:

$$\frac{dV}{dt} = \frac{\Delta A v_n dt}{dt} = v_n \Delta A, \tag{6.57}$$

where $v_n = \mathbf{v}\cdot\mathbf{n}$ is the projection of the velocity vector on the normal $\mathbf{n}$ to the surface. Indeed, the amount of the liquid passed through ΔA over time dt would correspond to the volume dV of the cylinder in the figure whose base has the area of ΔA and the height $v_n dt$ (as $|\mathbf{v}|\, dt$ is the length of its side which may not necessarily be parallel to the normal $\mathbf{n}$). If, however, the surface was *not* planar and the velocity distribution *not* uniform, then the same result would have been obtained only for a small surface area ΔA which can approximately be assumed planar, and the velocity distribution across it uniform. Dividing the whole surface S into small surface elements ΔA and summing up contributions from each of them, $\sum v_n \Delta A$, we arrive at the flux integral $\int\int v_n dA = \int \mathbf{v}\cdot d\mathbf{A}$ in the limit. Hence, the flux integral of the velocity field through a surface gives a rate of flow of the liquid through that surface area, i.e. the volume of the liquid flown through it per unit time.

There is a very simple (and general!) formula for the calculation of the surface (flux) integrals for vector fields. If the surface S is specified parametrically as $x = x(u, v)$, $y = y(u, v)$ and $z = z(u, v)$, then, according to Eq. (6.48), $dA = |\mathbf{N}|\, dudv$. At the same time, the unit normal to the surface $\mathbf{n} = \mathbf{N}/|\mathbf{N}|$, hence we can write:

$$
\begin{aligned}
\int\int_S \mathbf{F} \cdot d\mathbf{A} &= \int\int (\mathbf{F} \cdot \mathbf{n})\, dA = \int\int \left(\mathbf{F} \cdot \frac{\mathbf{N}}{|\mathbf{N}|}\right) (|\mathbf{N}|\, dudv) = \int (\mathbf{F} \cdot \mathbf{N})\, dudv \\
&= \int\int [\mathbf{F}(x(u, v), y(u, v), z(u, v)) \cdot \mathbf{N}(u, v)]\, dudv \\
&= \int\int \left(F_x J_x + F_y J_y + F_z J_z\right) dudv. \qquad (6.58)
\end{aligned}
$$

To illustrate application of this powerful result, let us find the flux of the field $\mathbf{F} = (yz, xz, xy)$ through the outer surface of the cylinder $x^2 + y^2 = R^2$, $0 \le z \le h$. We use cylindrical coordinates: $r = R$ (fixed), ϕ and z which are defined via

$$
x = R\cos\phi, \quad y = R\sin\phi \quad \text{and} \quad z = z.
$$

The coordinates ϕ and z specify the surface of the cylinder. Thus, in this case $u = \phi$ and $v = z$, so that, see Eq. (6.46),

$$
J_x = R\cos\phi, \quad J_y = R\sin\phi \quad \text{and} \quad J_z = 0,
$$

and we obtain

$$
\begin{aligned}
\text{Flux} &= \int\int d\phi dz\, (yzR\cos\phi + xzR\sin\phi) \\
&= \int_0^h dz \int_0^{2\pi} d\phi \left(R^2 z\sin\phi\cos\phi + R^2 z\cos\phi\sin\phi\right) \\
&= 2R^2 \int_0^h zdz \int_0^{2\pi} \sin\phi\cos\phi d\phi = 2R^2 \left(\frac{h^2}{2}\right) \left(\frac{\sin^2\phi}{2}\right)_0^{2\pi} = 0\ .
\end{aligned}
$$

Problem 6.47. *Calculate the surface integral* $\int\int \mathbf{F} \cdot d\mathbf{A}$ *over the triangular surface bound by the points* $(2, 0, 0)$, $(0, 2, 0)$ *and* $(0, 0, 2)$ *as shown in Fig. 6.26(c), where* $F = (x, y, z)$ *and the outward normal to the surface is used.* [*Hint*: the general equation for the plane is $ax+by+cz = 1$, where the constants a, b and c can be found from the known points on the plane.][*Answer*: 4.]

There is also another way of calculating the surface integrals (6.56) for vector fields in which it is expressed via a sum of double integrals taken over Cartesian planes $x-y$, $y-z$ and $x-z$. This method is sometimes more convenient. Consider the vector field $\mathbf{F} = \left(F_x, F_y, F_z\right)$, then

$$\begin{aligned}\int\int \mathbf{F}\cdot d\mathbf{A} &= \int\int \left(F_x n_x + F_y n_y + F_z n_z\right) dA \\ &= \int\int F_x n_x dA + \int\int F_y n_y dA + \int\int F_z n_z dA,\end{aligned} \qquad (6.59)$$

since in the end of the day, the integral is an integral sum and the latter can always be split into the sum of the three separate integral sums leading to the expression above, with the three integrals in the right-hand side. Note that in each of them the integration is performed over the actual surface, i.e. x, y and z are related to each other by the equation of the surface. Above, as follows either from Eq. (6.45) or (6.47) (for the two ways the surface can be specified), we have:

$$n_x = \cos\alpha = \cos(\widehat{\mathbf{n},\mathbf{i}}) = \frac{N_x}{|\mathbf{N}|} = \frac{J_x(u,v)}{\sqrt{J_x^2(u,v)+J_y^2(u,v)+J_z^2(u,v)}}$$

$$\text{or } \frac{-\partial z/\partial x}{\sqrt{1+(\partial z/\partial x)^2+(\partial z/\partial y)^2}}, \qquad (6.60)$$

$$n_y = \cos\beta = \cos(\widehat{\mathbf{n},\mathbf{j}}) = \frac{N_y}{|\mathbf{N}|} = \frac{J_y(u,v)}{\sqrt{J_x^2(u,v)+J_y^2(u,v)+J_z^2(u,v)}}$$

$$\text{or } \frac{-\partial z/\partial y}{\sqrt{1+(\partial z/\partial x)^2+(\partial z/\partial y)^2}}, \qquad (6.61)$$

$$n_z = \cos\gamma = \cos(\widehat{\mathbf{n},\mathbf{k}}) = \frac{N_z}{|\mathbf{N}|} = \frac{J_z(u,v)}{\sqrt{J_x^2(u,v)+J_y^2(u,v)+J_z^2(u,v)}}$$

$$\text{or } \frac{1}{\sqrt{1+(\partial z/\partial x)^2+(\partial z/\partial y)^2}}. \qquad (6.62)$$

These are components of the normal vector $\mathbf{n}$ along the x, y and z axes, and $\alpha = (\widehat{\mathbf{n},\mathbf{i}})$, $\beta = (\widehat{\mathbf{n},\mathbf{j}})$ and $\gamma = (\widehat{\mathbf{n},\mathbf{k}})$ are the corresponding angles the normal makes with the Cartesian axes. Then, $n_z dA = dA\cos\gamma$ projects the surface area dA on the $x-y$ plane, Fig. 6.29, and thus can simply be replaced by $\pm dxdy$. Note the sign: depending on which side of the surface is considered (upper with n_z being positive or lower with n_z negative), a plus or minus sign should be attached to the surface element (since dA is always positive) as stressed in the figure. Similarly, $n_x dA = \pm dydz$ and $n_y dA = \pm dzdx$, so that the integrals (6.59) can be rewritten as

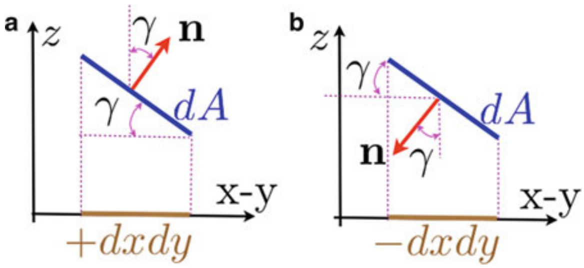


Fig. 6.29 Projection of the surface element dA on the $x-y$ plane for (**a**) $n_z > 0$ (*directed upwards*) and (**b**) $n_z < 0$ (*directed downwards*)

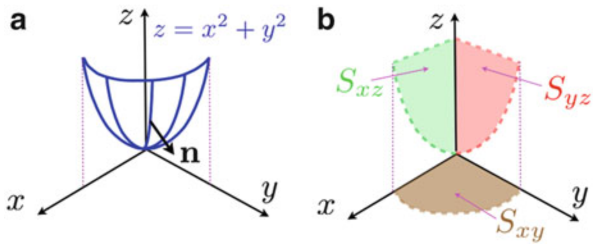


Fig. 6.30 (**a**) The surface $z = x^2 + y^2$ bounded by $0 \le z \le 1$, $x \ge 0$ and $y \ge 0$ (*blue*). Its projections on the three Cartesian planes are also shown in (**b**) by *light brown* (S_{xy}), *green* (S_{xz}) and *red* (S_{yz}) colours

$$\int\int_S \mathbf{F}\cdot d\mathbf{A} = \pm \int\int_{S_{yz}} F_x dydz \pm \int\int_{S_{xz}} F_y dzdx \pm \int\int_{S_{xy}} F_z dxdy, \tag{6.63}$$

where S_{xy}, S_{xz} and S_{yz} are projections of the surface S onto the corresponding coordinate planes.

This expression gives a practical way of calculating the flux integrals by essentially breaking them down into three double integrals with respect to the three projections on coordinate surfaces, and then calculating each one of them via

$$\int\int_{S_{yz}} F_x(x,y,z)dydz = \int\int F_x\left(x(y,z),y,z\right)dydz,$$

$$\int\int_{S_{xz}} F_y(x,y,z)dzdx = \int\int F_y\left(x,y(x,z),z\right)dzdx,$$

$$\int\int_{S_{xy}} F_z(x,y,z)dxdy = \int\int F_z\left(x,y,z(x,y)\right)dxdy.$$

As an example, we calculate the flux of the force field $\mathbf{F} = (x^2z, xy^2, z)$ through the outer surface of $z = x^2 + y^2$, $0 \le z \le 1$, $x \ge 0$ and $y \ge 0$, see Fig. 6.30. The integral

$$I = \int\int_S \mathbf{F}\cdot d\mathbf{A} = I_x + I_y + I_z = \int\int_{S_{yz}} x^2 z dydz + \int\int_{S_{xz}} xy^2 dxdz - \int\int_{S_{xy}} z dxdy,$$

where the surfaces S_{xy} (with $z = 0$), S_{xz} (with $y = 0$) and S_{yz} (with $x = 0$) are explicitly indicated in the figure. Note the minus sign before the $dxdy$ integral. This is because in the cases of I_x and I_y the normal is directed along the x and y axes, while in the case of I_z it is directed downwards (in the direction of $-z$), so that we have to take the negative sign for the $dxdy$ integral.

The individual integrals are calculated by remembering that the third variable is related to the other two used in the integration via the equation $z = x^2 + y^2$ of the surface, and hence can be easily expressed. We obtain:

$$\begin{aligned} I_x &= \int\int_{S_{yz}} x^2 z dydz = \int\int \left(z - y^2\right) z dydz \\ &= \int_0^1 dy \int_{y^2}^1 dz \left(z - y^2\right) z = \int_0^1 dy \left(\frac{1}{3} + \frac{y^6}{6} - \frac{y^2}{2}\right) = \frac{4}{21} , \\ I_y &= \int\int_{S_{xz}} xy^2 dxdz = \int\int x\left(z - x^2\right) dxdz \\ &= \int_0^1 dx \int_{x^2}^1 dz \left(z - x^2\right) x = \int_0^1 dx \left(\frac{x}{2} - x^3 + \frac{x^5}{2}\right) = \frac{1}{12}, \end{aligned}$$

and, finally, the last integral I_z is most easily calculated using polar coordinates (r, ϕ) (recall that the two-dimensional Jacobian is equal to r in this case):

$$\begin{aligned} I_z &= -\int\int_{S_{xy}} zdxdy = -\int\int \left(x^2 + y^2\right) dxdy \\ &= -\int_0^1 rdr \int_0^{\pi/2} r^2 d\phi = -\frac{\pi}{2} \frac{r^4}{4}\bigg|_0^1 = -\frac{\pi}{8}. \end{aligned}$$

Thus, $I = 4/21 + 1/12 - \pi/8$.

Problem 6.48. *Calculate the flux of the magnetic field* $\mathbf{B} = (2xy, yz, 1)$ *through the surface of the planar circle of radius* R *centred at the origin and located in the* $x - y$ *plane.* [*Answer*: πR^2.]

6.4.5 *Relationship Between Line and Surface Integrals: Stokes's Theorem*

We recall that there is a relationship between the line integral on the plane around some surface area and the double integral over that area (Green's formula), see Sect. 6.3.3. In fact, this, essentially two-dimensional result, can be generalised into all three dimensions and is called *Stokes's formula* (theorem).

Theorem 6.6 (Stokes's Theorem). *Let S be a simply connected non-planar piecewise smooth orientable (two-sided, see Fig. 6.22(a)) surface with the closed boundary line L. If three functions $F_x(\mathbf{r})$, $F_y(\mathbf{r})$ and $F_z(\mathbf{r})$ are all continuous together with their first derivatives everywhere in S including its boundary, then*

$$\oint_L \mathbf{F}\cdot d\mathbf{l} = \int\int_S \operatorname{curl}\mathbf{F}\cdot d\mathbf{A}, \tag{6.64}$$

where $\mathbf{F}(\mathbf{r}) = F_x(\mathbf{r})\mathbf{i}+F_y(\mathbf{r})\mathbf{j}+F_z(\mathbf{r})\mathbf{k}$ is the vector field, and we also defined another vector:

$$\operatorname{curl}\mathbf{F}(\mathbf{r}) = \left(\frac{\partial F_z}{\partial y}-\frac{\partial F_y}{\partial z}\right)\mathbf{i}+\left(\frac{\partial F_x}{\partial z}-\frac{\partial F_z}{\partial x}\right)\mathbf{j}+\left(\frac{\partial F_y}{\partial x}-\frac{\partial F_x}{\partial y}\right)\mathbf{k}. \tag{6.65}$$

This vector is called curl *of* $\mathbf{F}$*; it is composed of partial derivatives of the components of the vector field with respect to the Cartesian coordinates. Note that the surface boundary line L is traversed in a direction that corresponds to the chosen side of the surface to be always on the left (according to the* right-hand rule*, see Fig. 6.22(b)). Note that formula (6.64) can also be written differently as*

$$\oint_L F_x dx + F_y dy + F_z dz$$
$$= \int\int_S \left(\frac{\partial F_z}{\partial y}-\frac{\partial F_y}{\partial z}\right)dydz+\left(\frac{\partial F_x}{\partial z}-\frac{\partial F_z}{\partial x}\right)dxdz+\left(\frac{\partial F_y}{\partial x}-\frac{\partial F_x}{\partial y}\right)dxdy, \tag{6.66}$$

where in the right-hand side we employed an alternative form for the surface (flux) integral, see Sect. 6.4.4.

Proof. Consider first a surface S of a rather special shape which is specified by the equation $z = z(x,y)$ and let S_{xy} be its projection on the $x-y$ plane, Fig. 6.31. The surface boundary L will become the boundary L_{xy} in the projection S_{xy}. We can specify the boundary L by an additional equation relating y to x, i.e. $y = y(x)$, which we assume does exist and which, together with the equation for the surface $z = z(x,y)$, would give uniquely the boundary L. Then, consider the x-part of the line integral (see the text around Eq. (6.31)):

$$J_P = \oint_L F_x(x,y,z)dx = \oint_{L_{xy}} F_x\left(x, y(x), z\left(x, y(x)\right)\right)dx. \tag{6.67}$$

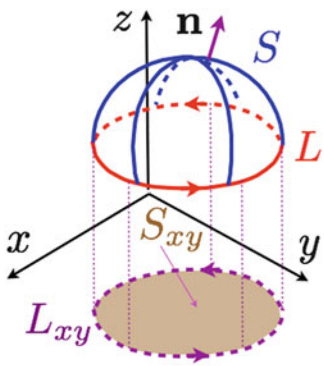


Fig. 6.31 To the proof of the Stokes's theorem: the surface S with the boundary L is projected on the $x-y$ plane forming a planar surface S_{xy} with the planar boundary L_{xy}

We have rewritten it as a line integral over the projection L_{xy} of the line L onto the $x-y$ plane, see again the figure. Using Green's formula (6.34) valid for planar regions, we obtain that the line integral J_P must be equal to the double integral over the planar region S_{xy} (note that $P = F_x$ and $Q = 0$ in our case):

$$J_P = \int\int_{S_{xy}} \left(-\frac{\partial F_x}{\partial y}\right)_{\text{tot}} dxdy = -\int\int_{S_{xy}} \left(\frac{\partial F_x}{\partial y} + \frac{\partial F_x}{\partial z}\frac{\partial z}{\partial y}\right) dxdy. \tag{6.68}$$

Note that the derivative $(\partial F_x/\partial y)_{\text{tot}}$ is the full derivative with respect to y, and, when calculating it, we must take account of the fact that $z = z(x, y)$ also depends on y. Next, consider the components of the normal vector $\mathbf{n}$ to the surface, Eqs. (6.60)–(6.62). Since

$$n_y = \cos(\widehat{\mathbf{n},\mathbf{j}}) = \frac{-\partial z/\partial y}{\sqrt{1 + (\partial z/\partial x)^2 + (\partial z/\partial y)^2}} \quad \text{and}$$

$$n_z = \cos(\widehat{\mathbf{n},\mathbf{k}}) = \frac{1}{\sqrt{1 + (\partial z/\partial x)^2 + (\partial z/\partial y)^2}},$$

then

$$\frac{n_y}{n_z} = -\frac{\partial z}{\partial y} \quad \Longrightarrow \quad \frac{\partial z}{\partial y} = -\frac{n_y}{n_z} = -\frac{\cos(\widehat{\mathbf{n},\mathbf{j}})}{\cos(\widehat{\mathbf{n},\mathbf{k}})},$$

and hence Eq. (6.68) can be rewritten as

$$J_P = -\int\int_{S_{xy}} \left(\frac{\partial F_x}{\partial y} - \frac{\partial F_x}{\partial z}\frac{\cos(\widehat{\mathbf{n},\mathbf{j}})}{\cos(\widehat{\mathbf{n},\mathbf{k}})}\right) dxdy. \tag{6.69}$$

On the other hand, consider the following surface integral over the surface S:

$$\int\int_S \left(-\frac{\partial F_x}{\partial y} \cos(\widehat{\mathbf{n},\mathbf{k}}) + \frac{\partial F_x}{\partial z} \cos(\widehat{\mathbf{n},\mathbf{j}}) \right) dA$$
$$= -\int\int_S \left(\frac{\partial F_x}{\partial y} - \frac{\partial F_x}{\partial z} \frac{\cos(\widehat{\mathbf{n},\mathbf{j}})}{\cos(\widehat{\mathbf{n},\mathbf{k}})} \right) \cos(\widehat{\mathbf{n},\mathbf{k}}) dA$$
$$= -\int\int_{S_{xy}} \left(\frac{\partial F_x}{\partial y} - \frac{\partial F_x}{\partial z} \frac{\cos(\widehat{\mathbf{n},\mathbf{j}})}{\cos(\widehat{\mathbf{n},\mathbf{k}})} \right) dxdy,$$

since n_z is positive, Fig. 6.31, and, therefore, $\cos(\widehat{\mathbf{n},\mathbf{k}})dA = +dxdy$. Thus, we notice that the same expression (6.69) has been obtained, i.e. we have just proven that

$$J_P = \oint_L F_x(x,y,z)dx = \int\int_S \left(\frac{\partial F_x}{\partial z} \cos(\widehat{\mathbf{n},\mathbf{j}}) - \frac{\partial F_x}{\partial y} \cos(\widehat{\mathbf{n},\mathbf{k}}) \right) dA. \qquad (6.70)$$

A similar consideration for the same surface S but specified via $y = y(x,z)$ and a function $F_z(\mathbf{r})$ gives

$$J_R = \oint_L F_z(x,y,z)dz = \int\int_S \left(\frac{\partial F_z}{\partial y} \cos(\widehat{\mathbf{n},\mathbf{i}}) - \frac{\partial F_z}{\partial x} \cos(\widehat{\mathbf{n},\mathbf{j}}) \right) dA, \qquad (6.71)$$

and for the surface S given via $x = x(y,z)$ and a function $F_y(\mathbf{r})$, one obtains instead:

$$J_Q = \oint_L F_y(x,y,z)dy = \int\int_S \left(\frac{\partial F_y}{\partial x} \cos(\widehat{\mathbf{n},\mathbf{k}}) - \frac{\partial F_y}{\partial z} \cos(\widehat{\mathbf{n},\mathbf{i}}) \right) dA. \qquad (6.72)$$

Any piecewise smooth simply connected surface can always be represented as a combination of elementary surfaces considered above. Therefore, the equations above are valid for any such S. Let us add these three identities (6.70)–(6.72) together. In the left-hand side we then get the line integral over the boundary L from all three functions (components of $\mathbf{F}$), while in the right-hand side—the sum of the surface integrals:

$$\oint_L F_x dx + F_y dy + F_z dz = \int\int_S \left(\frac{\partial F_x}{\partial z} \cos(\widehat{\mathbf{n},\mathbf{j}}) - \frac{\partial F_x}{\partial y} \cos(\widehat{\mathbf{n},\mathbf{k}}) \right) dA$$
$$+ \int\int_S \left(\frac{\partial F_z}{\partial y} \cos(\widehat{\mathbf{n},\mathbf{i}}) - \frac{\partial F_z}{\partial x} \cos(\widehat{\mathbf{n},\mathbf{j}}) \right) dA$$
$$+ \int\int_S \left(\frac{\partial F_y}{\partial x} \cos(\widehat{\mathbf{n},\mathbf{k}}) - \frac{\partial F_y}{\partial z} \cos(\widehat{\mathbf{n},\mathbf{i}}) \right) dA$$

$$
\begin{aligned}
&= \int\int_S \left[\left(\frac{\partial F_z}{\partial y} - \frac{\partial F_y}{\partial z} \right) \cos(\widehat{\mathbf{n}, \mathbf{i}}) + \left(\frac{\partial F_x}{\partial z} - \frac{\partial F_z}{\partial x} \right) \cos(\widehat{\mathbf{n}, \mathbf{j}}) \right. \\
&\quad \left. + \left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y} \right) \cos(\widehat{\mathbf{n}, \mathbf{k}}) \right] dA \\
&= \int\int_S \left[\left(\frac{\partial F_z}{\partial y} - \frac{\partial F_y}{\partial z} \right) n_x + \left(\frac{\partial F_x}{\partial z} - \frac{\partial F_z}{\partial x} \right) n_y + \left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y} \right) n_z \right] dA \\
&= \int\int_S (\mathrm{curl}\, \mathbf{F} \cdot \mathbf{n})\, dA = \int\int_S \mathrm{curl}\, \mathbf{F} \cdot d\mathbf{A}.
\end{aligned}
$$

If a more complex surface is considered, it can always be divided up into finite fragments for each of which the above identity is valid separately. Summing up such identities for all such fragments, we have in the right-hand side the surface integral over the entire surface, while in the left-hand side the line integral over all boundary lines including the internal lines used to divide the surface up. However, the internal boundaries will be contributing twice, each time traversed in the opposite directions, and hence there will be no contribution from them since the line integral changes sign if taken in the opposite direction (cf. our discussion of Green's theorem in Sect. 6.3.3 and especially Fig. 6.17(c)). Therefore, the formula is valid for any surface, which is the required general result. **Q.E.D.**

Note that the Stokes's theorem is also valid for more complicated surfaces as well, e.g. the ones that contain holes, Fig. 6.32. In this case the line integral $\oint_L$ is considered along all boundaries, including the inner ones, and the traverse directions should be chosen in accord with the right-hand rule as illustrated in the figure.

The curl of $\mathbf{F}$ can also be formally written as a determinant, and this form is easier to remember:

$$
\mathrm{curl}\, \mathbf{F}(\mathbf{r}) = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \partial/\partial x & \partial/\partial y & \partial/\partial z \\ F_x & F_y & F_z \end{vmatrix}. \tag{6.73}
$$

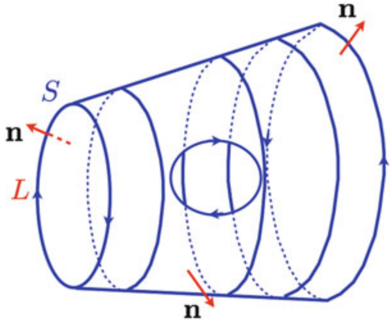


Fig. 6.32 Surface S contains a hole in it. The traverse directions, chosen with respect to the right-hand rule for the given direction of the normal **n** (*outward*) correspond to the surface being always on the left

This determinant notation is to be understood as follows: the determinant has to be opened along the first row with the operators in the second row to appear on the left from the components of the vector field $\mathbf{F}$ occupying the third row. Then, their "product" should be interpreted as the *action* of the operator on the corresponding component of $\mathbf{F}$. For instance, the x component of the curl of $\mathbf{F}$ is obtained as

$$(\operatorname{curl}\mathbf{F})_x = \begin{vmatrix} \partial/\partial y & \partial/\partial z \\ F_y & F_z \end{vmatrix} = \frac{\partial}{\partial y}F_z - \frac{\partial}{\partial z}F_y \quad \longmapsto \quad \frac{\partial F_z}{\partial y} - \frac{\partial F_y}{\partial z},$$

which is exactly the x component of the curl appearing in the original formula (6.65). It is easy to check that it works for the other two components as well.

As an example, we shall verify Stokes's theorem by considering the flux of the vector field $\mathbf{F} = -y\mathbf{i} + x\mathbf{j} + \mathbf{k}$ through the hemisphere $z = \sqrt{R^2 - x^2 - y^2} \geq 0$ centred at the origin. We shall first consider the line integral in the left-hand side of Stokes's theorem (6.64). Here the surface boundary L is a circle $x^2 + y^2 = R^2$ of radius R. Using polar coordinates, $x = R\cos\phi$ and $y = R\sin\phi$ with the angle ϕ as a parameter, we obtain

$$\begin{aligned} \oint_L \mathbf{F}\cdot d\mathbf{l} &= \oint_L F_x dx + F_y dy = \int_0^{2\pi} \left[(-R\sin\phi)(-R\sin\phi) + (R\cos\phi)(R\cos\phi)\right] d\phi \\ &= R^2 \int_0^{2\pi} d\phi = 2\pi R^2. \end{aligned}$$

On the other hand, let us now consider the surface integral in the right-hand side of Eq. (6.64). The curl of $\mathbf{F}$ is

$$\begin{aligned} \operatorname{curl}\mathbf{F} &= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \partial/\partial x & \partial/\partial y & \partial/\partial z \\ -y & x & 1 \end{vmatrix} \\ &= \left(\frac{\partial 1}{\partial y} - \frac{\partial x}{\partial z}\right)\mathbf{i} - \left(\frac{\partial 1}{\partial x} - \frac{\partial(-y)}{\partial z}\right)\mathbf{j} + \left(\frac{\partial x}{\partial x} - \frac{\partial(-y)}{\partial y}\right)\mathbf{k} \\ &= 0\mathbf{i} - 0\mathbf{j} + (1 - (-1))\,\mathbf{k} = 2\mathbf{k}, \end{aligned}$$

and the normal to the hemisphere is given by $\mathbf{N} = (R\sin\theta)\,\mathbf{r}$, see Eq. (6.49), so that, following Eq. (6.58), and using spherical coordinates,

$$\begin{aligned} \int\int_S \operatorname{curl}\mathbf{F}\cdot d\mathbf{A} &= \int\int (\operatorname{curl}\mathbf{F}\cdot\mathbf{N})\,d\theta d\phi = \int\int 2R\sin\theta\,(\mathbf{k}\cdot\mathbf{r})\,d\theta d\phi \\ &= 2R\int\int (\sin\theta)\,z d\theta d\phi = 2R\int\int (\sin\theta)(R\cos\theta)\,d\theta d\phi \\ &= 2R^2\int_0^{\pi/2} \sin\theta\cos\theta d\theta \int_0^{2\pi} d\phi = 2\pi R^2. \end{aligned}$$

The same result is obtained also using Eq. (6.59):

$$\int\int_S \operatorname{curl} \mathbf{F} \cdot d\mathbf{A} = \int\int 2\,(\mathbf{k} \cdot \mathbf{n})\,dA = 2\int\int n_z dA = 2\int\int_{\text{circle}} dxdy = 2\pi R^2 \ .$$

Note that, according to the Stokes's theorem, the surface (flux) integral of a vector field $\mathbf{V} = \operatorname{curl} \mathbf{F}$ does not depend on the actual surface, only on its bounding curve. Therefore, if a vector field can be written as $\mathbf{V} = \operatorname{curl} \mathbf{F}$, then the surface integral

$$\int\int_S (\mathbf{V} \cdot \mathbf{n})\,dA = \int\int_S (\operatorname{curl} \mathbf{F} \cdot \mathbf{n})\,dA = \oint_L \mathbf{F} \cdot d\mathbf{l}$$

will not depend on the actual surface, only on its bounding curve L. This property can be used to calculate some of the surface integrals by replacing them with simpler line integrals, as is illustrated by the surface integral $\int\int_S \operatorname{curl} \mathbf{F} \cdot d\mathbf{A}$ with $\mathbf{F} = (xyz, -x, z)$ and the surface S being a hemisphere $x^2 + y^2 + z^2 = R^2$ bounded below by the $z = 0$ plane. According to Stokes's theorem, the integral in question can be replaced by the line integral

$$\oint_{\text{circle}} F_x dx + F_y dy + F_z dz$$

over the bounding circle $x^2 + y^2 = R^2$. This is most easily calculated in polar coordinates ($x = R\cos\phi$, $y = R\sin\phi$ and $z = 0$):

$$\oint_{\text{circle}} = \int_0^{2\pi} [xyz\,(-R\sin\phi) - x\,(R\cos\phi)]\,d\phi = \int_0^{2\pi} [-R\cos\phi\,(R\cos\phi)]\,d\phi$$

$$= -R^2 \int_0^{2\pi} \cos^2\phi d\phi = -R^2 \left(\frac{\phi}{2} + \frac{1}{4}\sin(2\phi)\right)_0^{2\pi} = -\pi R^2 \ .$$

Problem 6.49. *Verify Stokes's theorem for* $\mathbf{F} = (x, 0, z)$ *and the surface* S *being the triangle in Fig. 6.26(c).*

Problem 6.50. *Verify Stokes's theorem for* $\mathbf{F} = (xy, yz, xz)$ *and the surface* S *being a square with corners at* $(\pm 1, \pm 1, 0)$. [*Answer*: either of the integrals is zero.]

Problem 6.51. *Consider a cylinder* $x^2 + y^2 = R^2$ *of height* h. *Its upper end at* $z = h$ *is terminating with the plane, while its other end at* $z = 0$ *is opened. Verify Stokes's theorem for the surface of the cylinder if* $\mathbf{F} = (0, -y, z)$. [*Answer*: either of the integrals is zero.]

Problem 6.52. *Prove that the* $\operatorname{curl}[f(r)\mathbf{r}] = 0$, *where* $f(r)$ *is an arbitrary function of* $r = |\mathbf{r}|$.

Problem 6.53. *Calculate the flux of the vector field* $\mathbf{F} = (0, 0, z)$ *through the external surface (i.e. in the outward direction) of the circular paraboloid* $z = x^2 + y^2$ *shown in Fig. 6.26(b), bound by the planes z = 0 and z = 1.* [*Answer*: $-\pi/2$.]

Problem 6.54. *Evaluate the flux integral over the cube* $0 < x < 1$, $0 < y < 1$ *and* $0 < z < 1$ *for the field* $\mathbf{F} = (x, y, 2)$ *and the outward direction for each face of the cube.* [*Answer*: 2.]

Problem 6.55. *Evaluate the surface (flux) integral over the outward surface of the unit sphere centred at the origin with* $\mathbf{F} = \mathbf{r}$. [*Answer*: 4π.]

Problem 6.56. *Calculate* $\operatorname{curl} \mathbf{F}$*, where* $\mathbf{F} = (yz, xz, xy)$. [*Answer:* $\mathbf{0}$.]

Problem 6.57. *Calculate* $\operatorname{curl} \mathbf{F}$ *at the point* $(1, 1, 1)$*, where* $\mathbf{F} = (2x^2 z, 2xy, xz)$. [*Answer:* $(0, 1, 2)$.]

Problem 6.58. *Use Stokes's theorem to evaluate the line integral* $\oint_L \mathbf{F} \cdot d\mathbf{l}$*, where* $\mathbf{F} = (-y, x, 2z)$ *and the contour L is the unit circle* $x^2 + y^2 = 1$ *in the* $z = 2$ *plane.* [*Hint*: cap the circle with the hemisphere.] [*Answer*: 2π.]

Problem 6.59. *Calculate, using Stokes's theorem, the surface integral* $\int\int_S \mathbf{F} \cdot d\mathbf{A}$*, where* $\mathbf{F} = \operatorname{curl} \mathbf{B}$ *with* $\mathbf{B} = (3y, -2x, xy)$*, and S is the hemispherical surface* $x^2 + y^2 + z^2 = R^2$ *with* $z \geq 0$. *Will the integral change if a different surface is chosen which caps* the same *bounding curve?* [*Answer*: $-5\pi R^2$; no.]

Problem 6.60. *Choose the planar surface S in the previous problem (i.e. the disk of radius R), that has the same bounding line* $x^2 + y^2 = R^2$. *Calculate* $\mathbf{F} = \operatorname{curl} \mathbf{B}$ *and show by the direct calculation of the surface integral,* $\int\int_S \mathbf{F} \cdot d\mathbf{A}$*, that it is equal to the same value as in the previous problem.*

Problem 6.61. *Verify Stokes's theorem if* $\mathbf{F} = (y^2, z^2, x^2)$ *and S is the surface of the circular ellipsoid described by* $z = 1 - x^2 - y^2$ *with* $z \geq 0$*, see Fig. 6.33.* [*Answer*: either of the integrals is zero.]

Problem 6.62. *Use Stokes's theorem to evaluate the line integral* $\oint_L \mathbf{F} \cdot d\mathbf{l}$ *over the unit circle* $x^2 + y^2 = 1$ *in the* $z = 5$ *plane, where* $\mathbf{F} = (x + y, z - 2x + y, y - z)$. [*Answer*: -3π.]

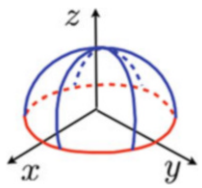


Fig. 6.33 Circular ellipsoid $z = 1 - x^2 - y^2, z \geq 0$

6.4.6 Three-Dimensional Case: Exact Differentials

As in the planar case, Sect. 6.3.4, there exists a very general theorem in the three-dimensional case as well, that states:

Theorem 6.7. *Let D be a simply connected region in the three-dimensional space, and there is a vector field $\mathbf{F}(\mathbf{r}) = (F_x, F_y, F_z)$ that is continuous in D together with its first derivatives. Then, the following four statements are equivalent to each other:*

1. *for any closed loop L inside D*

$$\oint_L F_x dx + F_y dy + F_z dz = \oint_L \mathbf{F} \cdot d\mathbf{l} = 0; \tag{6.74}$$

2. *the line integral $\int_C \mathbf{F} \cdot d\mathbf{l}$ does not depend on the actual integration curve C, only on its starting and ending points;*
3. *$dU(x, y, z) = F_x(x, y, z)dx + F_y(x, y, z)dy + F_z(x, y, z)dz$ is the exact differential, and*
4. *the following three conditions are satisfied:*

$$\frac{\partial F_x}{\partial y} = \frac{\partial F_y}{\partial x}, \quad \frac{\partial F_z}{\partial y} = \frac{\partial F_y}{\partial z} \quad \text{and} \quad \frac{\partial F_x}{\partial z} = \frac{\partial F_z}{\partial x}, \tag{6.75}$$

or, which is equivalent (see Eq. (6.65)), $\operatorname{curl} \mathbf{F} = 0$.

Proof. This theorem is proven similarly to the planar case of Sect. 6.3.4 by following the logic: 1→2, 2→3, 3→4 and then, finally, 4→1 again.

- 1→2 : the proof is identical to the planar case;
- 2→3: essentially the same as in the planar case, the function $U(x, y, z)$ giving rise to the exact differential dU is verified to be the line integral

$$U(x, y, z) = \int_{A(x_0, y_0, z_0)}^{B(x,y,z)} F_x dx + F_y dy + F_z dz, \tag{6.76}$$

  where $A(x_0, y_0, z_0)$ is a fixed point;
- 3→4: the same as in the planar case;
- 4→1: this is slightly different (although the idea is the same): we take a closed contour L and "dress" on it a surface S so that L would serve as its boundary. Then we use the Stokes's formula: $\oint_L \cdots = \int\int_S \cdots$. However, the curl of $\mathbf{F}$ is

equal to zero due to our assumption (property 4). Therefore, $\oint_L \cdots = 0$. Since the choice of L was arbitrary, this result is valid for any L in D. **Q.E.D.**

Note that Eq. (6.76) is indispensable for finding exact differentials, similarly to the two-dimensional case. Since the integral does not depend on the path, we can choose it to simplify the calculation. This is most easily done by taking the integration path $(x_0, y_0, z_0) \to (x, y_0, z_0) \to (x, y, z_0) \to (x, y, z)$ which contains straight lines going parallel to the Cartesian axis; it would split the line integral into three ordinary definite integrals:

$$U(x,y,z) = \int_{x_0}^{x} F_x(x_1, y_0, z_0) dx_1 + \int_{y_0}^{y} F_y(x, y_1, z_0) dy_1 + \int_{z_0}^{z} F_z(x, y, z_1) dz_1 . \tag{6.77}$$

An arbitrary constant appears naturally due to an arbitrary choice of the initial point $A\,(x_0, y_0, z_0)$. This results is a direct generalisation of the formula (6.40) from the two-dimensional case.

For instance, consider the expression $yzdx + zxdy + xydz$. First of all, we check that this expression is the exact differential. Here $F_x = yz$, $F_y = zx$ and $F_z = xy$. Therefore, if $\mathbf{F} = (F_x, F_y, F_z)$, then

$$\operatorname{curl} \mathbf{F} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \partial/\partial x & \partial/\partial y & \partial/\partial z \\ yz & zx & xy \end{vmatrix} = \mathbf{i}\,(x - x) - \mathbf{j}\,(y - y) + \mathbf{k}\,(z - z) = 0,$$

which means that we can apply the results of the theorem. Hence, we can use Eq. (6.77) to find $U(x, y, z)$ which would give rise to the differential $dU = yzdx + zxdy + xydz$:

$$\begin{aligned} U &= \int_{x_0}^{x} y_0 z_0 dx_1 + \int_{y_0}^{y} z_0 x dy_1 + \int_{z_0}^{z} xy dz_1 \\ &= y_0 z_0 (x - x_0) + z_0 x (y - y_0) + xy (z - z_0) = xyz - x_0 y_0 z_0 = xyz + C. \end{aligned}$$

It is readily checked that

$$dU = \frac{\partial U}{\partial x} dx + \frac{\partial U}{\partial y} dy + \frac{\partial U}{\partial z} dz = yzdx + xzdy + xydz,$$

which indeed is our expression.

6.4.7 Ostrogradsky–Gauss Theorem

We have seen above that there exists a simple relationship between the surface integral over some, generally non-planar, surface S and its bounding contour L (Stokes's theorem). Similarly, there is also a very general result relating the integral over some volume V and its bounding surface S which bears the names of Ostrogradsky and Gauss.

Theorem 6.8 (Divergence Theorem). *Let* $\mathbf{F} = (F_x, F_y, F_z)$ *be a vector field which is continuous (together with its first derivatives) in some three-dimensional region D. Then,*

$$\int\int\int_D \operatorname{div}\mathbf{F}\, dxdydz = \oiint_S \mathbf{F}\cdot d\mathbf{A}, \tag{6.78}$$

where S is the bounding surface to D, with the outward normal $\mathbf{n}$, *and the scalar*

$$\operatorname{div}\mathbf{F} = \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z} \tag{6.79}$$

is called divergence *of the vector field* $\mathbf{F}$.

Proof. Consider a region D that is cylindrical along the z axis and capped from below and above by surfaces $z = z_1(x, y)$ and $z = z_2(x, y)$, respectively, Fig. 6.34. Consider then the volume integral

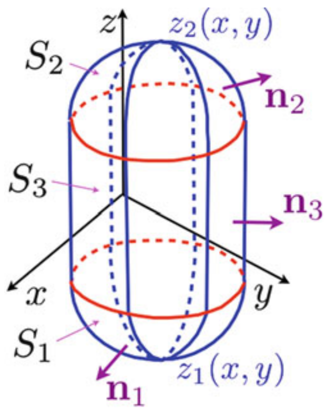


Fig. 6.34 A cylindrical space region bounded from below and above by surfaces $z = z_1(x, y)$ and $z = z_2(x, y)$, respectively. The *upper* cap forms a surface S_2, the *lower* one S_1 and the *cylindrical* part S_3

$$\int\int\int_D \frac{\partial F_z}{\partial z} dxdydz = \int\int dxdy \int_{z_1(x,y)}^{z_2(x,y)} \frac{\partial F_z}{\partial z} dz$$
$$= \int\int_{S_{xy}} \left[F_z\left(x, y, z_2(x, y)\right) - F_z\left(x, y, z_1(x, y)\right)\right] dxdy,$$

where S_{xy} is the projection of the surfaces S_1 or S_2 on the $x-y$ plane (both projections are the same due to the specific shape of D).

On the other hand, the surface integral over the whole surface of the cylinder $S = S_1 + S_2 + S_3$ (see the figure) for the vector field $\mathbf{G} = (0, 0, F_z)$ is

$$\oiint_S \mathbf{G} \cdot d\mathbf{A} = \int\int_S F_z n_z dA = \int\int_{S_1} F_z n_z dA + \int\int_{S_2} F_z n_z dA + \int\int_{S_3} F_z n_z dA.$$

The integral over S_3 is zero since $n_z = 0$ there; the other two integrals transform into:

$$\int\int_{S_1} F_z n_z dA = -\int\int_{S_{xy}} F_z\left(x, y, z_1(x, y)\right) dxdy \quad \text{and}$$
$$\int\int_{S_2} F_z n_z dA = +\int\int_{S_{xy}} F_z\left(x, y, z_2(x, y)\right) dxdy.$$

There is the minus sign chosen for the S_1 integral since $n_z < 0$ there, i.e. its normal leans towards the negative z direction. Therefore,

$$\oiint_S \mathbf{G} \cdot d\mathbf{A} = \int\int_{S_{xy}} \left[F_z\left(x, y, z_2(x, y)\right) - F_z\left(x, y, z_1(x, y)\right)\right] dxdy,$$

which is the same as the volume integral above. Hence:

$$\int\int\int_D \frac{\partial F_z}{\partial z} dxdydz = \oiint_S \begin{pmatrix} 0 \\ 0 \\ F_z \end{pmatrix} \cdot d\mathbf{A}, \tag{6.80}$$

where the vector $\mathbf{G}$ has been shown explicitly via its components. This result should be valid for any region D that can be broken down into cylindrical regions as above. This is because the volume integrals for these regions will just sum up (the left-hand side), while the surface integrals will be taken over all surfaces. However, for all *internal* surfaces (where different cylinders touch) the two integrals will be taken over different sides of the same surface (the normal vectors are opposite), and thus will cancel out as the surface integral changes sign if the other side of the surface is used.

Similarly, if the region D has a cylindrical shape with the axis along the x axis and the caps are specified by the equations $x = x_1(y, z)$ and $x = x_2(y, z)$, then we obtain

$$\int\int\int_D \frac{\partial F_x}{\partial x} dxdydz = \oiint_S \begin{pmatrix} F_x \\ 0 \\ 0 \end{pmatrix} \cdot d\mathbf{A}. \tag{6.81}$$

Finally, for cylindrical regions along the y axis with the two caps specified by $y = y_1(x, z)$ and $y = y_2(x, z)$, we obtain

$$\int\int\int_D \frac{\partial F_y}{\partial y} dxdydz = \oiint_S \begin{pmatrix} 0 \\ F_y \\ 0 \end{pmatrix} \cdot d\mathbf{A}. \tag{6.82}$$

As any volume can be broken down by means of either of these three ways, i.e. using touching cylindrical volumes oriented along either of the three coordinate axes, these three results are valid at the same time for any volume. Therefore, summing up all three expressions (6.80)–(6.82), we arrive at the final formula (6.78) sought for. **Q.E.D.**

Problem 6.63. *Verify the divergence theorem for* $\mathbf{F} = (xy, yz, zx)$ *and the volume region of a cube with vertices at the points* $(\pm 1, \pm 1, \pm 1)$. [*Answer*: each integral is zero.]

Problem 6.64. *Verify the divergence theorem if* $\mathbf{F} = (xy^2, x^2y, y)$ *and* S *is a circular cylinder* $x^2 + y^2 = 1$ *oriented with its axis along the* z *axis,* $-1 \leq z \leq 1$. [*Answer*: each of the integrals is equal to π.]

Problem 6.65. *Use the divergence theorem to calculate the surface flux integral* $\oiint_S \mathbf{F} \cdot d\mathbf{A}$ *over the entire surface of the cylinder* $x^2 + y^2 = R^2$, $-h/2 \leq z \leq h/2$, *where* $\mathbf{F} = (x^3, y^3, z^3)$. [*Answer*: $\pi h R^2 \left(6R^2 + h^2\right)/4$.]

There are also some variants of the Gauss theorem. We shall discuss only two of them. Consider three specific vector fields $\mathbf{F}_1 = (P, 0, 0)$, $\mathbf{F}_2 = (0, P, 0)$ and $\mathbf{F}_3 = (0, 0, P)$. Applying the divergence theorem to each of them separately, we obtain three integral identities:

$$\oiint_S P n_x dA = \int\int\int_V \frac{\partial P}{\partial x} dV\,, \quad \oiint_S P n_y dA = \int\int\int_V \frac{\partial P}{\partial y} dV \quad \text{and}$$
$$\oiint_S P n_z dA = \int\int\int_V \frac{\partial P}{\partial z} dV.$$

Multiplying each of these equations by the unit base vectors $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$, respectively, and then summing them up, we obtain a useful integral formula:

$$\oiint_S P\mathbf{n}dA = \int\int\int_V \left(\frac{\partial P}{\partial x}\mathbf{i} + \frac{\partial P}{\partial y}\mathbf{j} + \frac{\partial P}{\partial z}\mathbf{k}\right) dV = \int\int\int_V \mathrm{grad}P\, dV. \qquad (6.83)$$

We shall make use of this form of the Gauss's theorem later on in Sect. 6.7.3. Above we have introduced a vector field

$$\mathrm{grad}P = \frac{\partial P}{\partial x}\mathbf{i} + \frac{\partial P}{\partial y}\mathbf{j} + \frac{\partial P}{\partial z}\mathbf{k},$$

which is the gradient of the scalar field $P(\mathbf{r})$, see Sect. 5.8.

Problem 6.66. *Use a similar method to prove another integral identity:*

$$\oiint_S (\mathbf{a}\cdot\mathbf{n})\,\mathbf{b}\,dA = \int\int\int_V [\mathbf{b}\,\mathrm{grad}\,\mathbf{a} + (\mathbf{a}\cdot\mathrm{grad})\,\mathbf{b}]\,, \qquad (6.84)$$

where

$$\mathbf{a}\cdot\mathrm{grad}\,\mathbf{b} = (\mathbf{a}\cdot\mathrm{grad}\,b_x)\,\mathbf{i} + \left(\mathbf{a}\cdot\mathrm{grad}\,b_y\right)\mathbf{j} + (\mathbf{a}\cdot\mathrm{grad}\,b_z)\,\mathbf{k}$$

and

$$\mathbf{a}\cdot\mathrm{grad}f = a_x\frac{\partial f}{\partial x} + a_y\frac{\partial f}{\partial y} + a_z\frac{\partial f}{\partial z}$$

is a dot product of the vector $\mathbf{a}$ *and* $\mathrm{grad}f(\mathbf{r})$. [*Hint*: choose the vector fields $\mathbf{G}_1 = b_x\mathbf{a}$, $\mathbf{G}_2 = b_y\mathbf{a}$ and $\mathbf{G}_3 = b_z\mathbf{a}$.]

6.5 Application of Integral Theorems in Physics: Part I

The Gauss and Stokes's theorems are very powerful results that are frequently used in many areas of physics and engineering. Here we shall consider some of them. More applications are postponed until Sect. 6.7.

6.5.1 Continuity Equation

As our first example, we shall derive the so-called *continuity equation*

$$\frac{\partial \rho}{\partial t} + \mathrm{div}\,(\rho\mathbf{v}) = 0, \qquad (6.85)$$

that relates the time change of the density ρ of a fluid with its velocity distribution $\mathbf{v}(\mathbf{r})$.

The continuity equation corresponds to the conservation of mass in this case. Consider an arbitrary volume V of the fluid with the bounding surface S. The change of mass in the volume due to a flow of the liquid through the element dA of the bounding surface during time dt is

$$dM = \rho v_n dAdt = \rho\left(\mathbf{v}\cdot\mathbf{n}\right)dAdt = \rho\mathbf{v}\cdot d\mathbf{A}dt. \tag{6.86}$$

We have used here that the flow of the fluid through a surface area $d\mathbf{A}$ results in the change of mass per unit time equal to $\rho v_n dA = \rho\left(\mathbf{v}\cdot\mathbf{n}\right)dA = \rho\mathbf{v}\cdot d\mathbf{A}$, where $\mathbf{n}$ is the *outer* normal to the surface area dA, see Eq. (6.57). Thus, the mass flow is assumed to be positive if the fluid flows out of the volume. The total change (loss) of mass during time dt due to the flow of the fluid through the whole surface S that bounds the volume V is then the surface integral

$$dM = dt\oiint_S \rho\mathbf{v}\cdot d\mathbf{A}. \tag{6.87}$$

On the other hand, dM, i.e. the loss of mass in the volume, would correspond to the change of the density within it during this time, i.e. the same change of mass can be also calculated by integrating the density change over time dt across the whole volume V:

$$\begin{aligned} dM &= -\left[M(t+dt)-M(t)\right] = -\int\int\int_V\left[\rho(t+dt)-\rho(t)\right]dxdydz \\ &= -dt\int\int\int_V \frac{\partial\rho}{\partial t}dxdydz, \end{aligned}$$

where the minus sign reflects the fact that the mass is actually being lost in the volume (this is necessary from our assumption above when calculating the loss dM explicitly, that dM is positive if the mass is lost in the volume!). The two expressions should be equal to each other due to *mass conservation*:

$$\oiint_S \rho\mathbf{v}\cdot d\mathbf{A} = -\int\int\int_V \frac{\partial\rho}{\partial t}dxdydz.$$

Let us now transform the surface integral using the Gauss theorem of Eq. (6.78):

$$\begin{aligned} &\int\int\int_V \operatorname{div}\left(\rho\mathbf{v}\right)dxdydz = -\int\int\int_V \frac{\partial\rho}{\partial t}dxdydz \\ &\Longrightarrow \quad \int\int\int_V\left[\operatorname{div}\left(\rho\mathbf{v}\right)+\frac{\partial\rho}{\partial t}\right]dxdydz = 0. \end{aligned}$$

Since the 3D region V is arbitrary, this integral can only be equal to zero if the integrand is zero, i.e. we arrive at the continuity equation (6.85).

Note that the divergence term in the continuity equation (6.85) can also be written as:

$$\begin{aligned}\operatorname{div}(\rho\mathbf{v}) &= \frac{\partial(\rho v_x)}{\partial x} + \frac{\partial\left(\rho v_y\right)}{\partial y} + \frac{\partial(\rho v_z)}{\partial z} \\ &= \frac{\partial v_x}{\partial x}\rho + \frac{\partial\rho}{\partial x}v_x + \frac{\partial v_y}{\partial y}\rho + \frac{\partial\rho}{\partial y}v_y + \frac{\partial v_z}{\partial z}\rho + \frac{\partial\rho}{\partial z}v_z \\ &= \rho\left(\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}\right) + \left(\frac{\partial\rho}{\partial x}v_x + \frac{\partial\rho}{\partial y}v_y + \frac{\partial\rho}{\partial z}v_z\right) \\ &= \rho\operatorname{div}\mathbf{v} + \mathbf{v}\cdot\operatorname{grad}\rho,\end{aligned}$$

so that the continuity equation may also be written as follows:

$$\frac{\partial\rho}{\partial t} + \rho\operatorname{div}\mathbf{v} + \mathbf{v}\cdot\operatorname{grad}\rho = 0. \tag{6.88}$$

The continuity equation is also applicable in other situations. For instance, due to conservation of charge, the flux of charge in or out of a certain volume should be reflected by the change of the charge density inside the volume, which is precisely what the continuity equation (6.85) tells us. Indeed, consider a flow of particles each of charge q through an imaginable tube of the cross section dA (infinitesimally small). The total charge passed through dA per time dt will be then $dQ = \rho v_n dA dt = \rho(\mathbf{v}\cdot d\mathbf{A})\,dt$; this result is absolutely identical to the one above, Eq. (6.86), which we have derived for the mass of the fluid. Therefore, Eq. (6.85) must also be valid. However, in this particular case it is convenient to rewrite this equation in a slightly different form using the current density $\mathbf{j}$. It is defined in such a way that the current through a surface S can be written via the flux integral of the current density:

$$I = \int\int_S \mathbf{j}\cdot d\mathbf{A} = \int\int_S \mathbf{j}\cdot\mathbf{n}dA. \tag{6.89}$$

The current corresponds to the amount of charge dQ flown per unit time, i.e.

$$dQ = Idt = dt\int\int_S \mathbf{j}\cdot d\mathbf{A}.$$

Comparing this with expression (6.87) for the mass, we see that $\mathbf{j} = \rho\mathbf{v}$, i.e. the charge passed through a small area dA per unit time is in fact $dQ = \mathbf{j}\cdot d\mathbf{A}dt$ which makes perfect sense. Hence, the continuity equation can be rewritten as follows:

$$\frac{\partial\rho}{\partial t} + \operatorname{div}\mathbf{j} = 0. \tag{6.90}$$

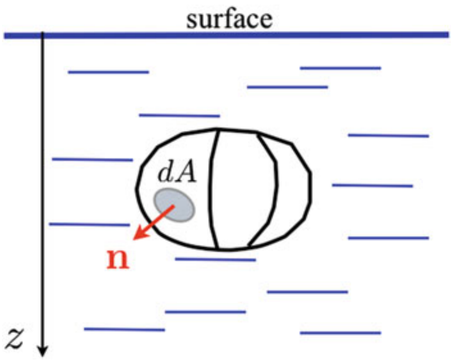


Fig. 6.35 An object is placed inside a uniform liquid. Note that the positive direction of the z axis is *down*

6.5.2 Archimedes Law

It is well known that the pulling force acting on an object of volume V in a uniform liquid of density ρ equals the weight of the liquid replaced by the object. This is in essence the law established by Archimedes. We shall now prove this statement for arbitrary shape of the volume V.

Indeed, the value of the force acting on a small area dA of the object surface that is located at the depth z (see Fig. 6.35) is $dF = \rho g z dA$, where g is the Earth gravity constant. Since the force acts in the direction opposite to the outer normal vector $\mathbf{n}$ of the surface, we can write in the vector form that the force $d\mathbf{F} = -\rho g z \mathbf{n} dA$. The components of the total force acting on the object from the liquid are obtained by summing up all contributions from all elements of the object surface:

$$F_x = -\int\int_S \rho g z n_x dA \ , \quad F_y = -\int\int_S \rho g z n_y dA \ \text{ and } \ F_z = -\int\int_S \rho g z n_z dA.$$

Now, consider the first integral. If we take $\mathbf{G} = (-\rho g z, 0, 0)$ as a vector field, then the force F_x can be written as the surface integral and hence we can use the divergence theorem (6.78):

$$F_x = \oiint_S G_x n_x dA = \oiint_S \mathbf{G} \cdot d\mathbf{A} = \int\int\int_V \operatorname{div} \mathbf{G} dxdydz = \int\int\int_V \frac{\partial G_x}{\partial x} dxdydz.$$

However, $\partial G_x / \partial x = 0$, since $G_x = -\rho g z$ depends only on z; hence, $F_x = 0$. Similarly, the force in the y direction is also zero, $F_y = 0$, so only the vertical force remains. If we introduce the vector field $\mathbf{G} = (0, 0, -\rho g z)$, then we can write:

$$\begin{aligned} F_z &= \oiint_S G_z n_z dA = \oiint_S \mathbf{G} \cdot d\mathbf{A} = \int\int\int_V \operatorname{div} \mathbf{G} dxdydz = \int\int\int_V \frac{\partial G_z}{\partial z} dxdydz \\ &= \int\int\int_V (-\rho g)\, dxdydz = -\rho g \int\int\int_V dxdydz = -\rho g V. \end{aligned}$$

Thus, the force is directed upwards (the positive direction of the z axis was chosen downwards in Fig. 6.35), and is equal exactly to the mass of the displaced liquid. This is a very powerful result since it is proven for an arbitrary shape of the object immersed in the liquid!

Problem 6.67. *An engineer is designing an air balloon with a specific ratio $\lambda = R/d < 1$ of the radius R of its horizontal cross section in the middle and of its half height, d. The shape of the balloon can be described by the equation $x^2 + y^2 + \lambda^2 z^2 = R^2$. Show that the Cartesian coordinates of the points on the balloon surface can be expressed in terms of the z and the polar angle ϕ as follows:*

$$x = \lambda\sqrt{d^2 - z^2}\cos\phi \ , \ \ y = \lambda\sqrt{d^2 - z^2}\sin\phi \ , \ \ z = z \ .$$

Then show that the corresponding Jacobians for the surface normal in terms of the coordinates (z, ϕ) are:

$$J_x = \lambda\sqrt{d^2 - z^2}\cos\phi \ , \ \ J_y = \lambda\sqrt{d^2 - z^2}\sin\phi \ , \ \ J_z = \lambda^2 z \ .$$

Hence, show that the surface area of the balloon is the following:

$$A = \frac{2\pi R d}{\sqrt{1-\lambda^2}}\left(\lambda\sqrt{1-\lambda^2} + \arcsin\sqrt{1-\lambda^2}\right),$$

and its volume $V = 4\pi\lambda^2 d^3/3$. If m is the maximum mass of a basket intended to be used in the flights, σ is the surface density of the material to be used for the balloon, and ρ is the air density, give advice to the engineer concerning the necessary conditions that the parameters λ and d should satisfy in order for the balloon to be able to take off.

6.6 Vector Calculus

6.6.1 Divergence of a Vector Field

We have already introduced the divergence of a vector field: if there is a vector field $\mathbf{F} = (F_x, F_y, F_z)$, then its divergence is given by Eq. (6.79). However, this specific definition is only useful in Cartesian coordinates. A more general definition of the divergence is needed in order to write the divergence in an alternate coordinate system, such as cylindrical or spherical coordinates. There are called *curvilinear coordinate* systems (we shall discuss them in detail in Book II, Chap. 7). So, if one would like to be able to write the divergence in any of these coordinate systems, a more general definition of the divergence is required.

This general definition can be formulated via a flux through a closed surface. Consider a point M, surround it by a closed surface S and calculate the ratio of the flux of the vector field $\mathbf{F}$ through that surface to the volume of the region V enclosed:

$$\frac{\text{flux}}{\text{volume}} = \frac{\oiint_S \mathbf{F} \cdot d\mathbf{A}}{V}.$$

Using the Gauss theorem (6.78), the surface integral is expressed via a volume integral:

$$\frac{\text{flux}}{\text{volume}} = \frac{\int\int\int_V \operatorname{div} \mathbf{F} dV}{V},$$

where $dV = dxdydz$ is a differential volume element. The average value theorem which we formulated in Eq. (6.2) for double integrals is also valid for volume integrals. Therefore,

$$\int\int\int \operatorname{div} \mathbf{F}(\mathbf{r}) dV = \operatorname{div} \mathbf{F}(\mathbf{p}) V \quad \Longrightarrow \quad \frac{\text{flux}}{\text{volume}} = \operatorname{div} \mathbf{F}(\mathbf{p}),$$

where the point $\mathbf{p}$ is somewhere inside the volume V. Now, take the limit of the volume V around the point M tending to zero. As the volume goes to zero, the point $\mathbf{p}$ remains inside the volume, and in the limit tends to the point M. Therefore, the ratio of the flux to the volume in the limit results exactly in $\operatorname{div} \mathbf{F}(M)$. Thus, we have just proven that

$$\lim_{V \to 0} \frac{1}{V} \oiint_S \mathbf{F} \cdot d\mathbf{A} = \operatorname{div} \mathbf{F}(M). \tag{6.91}$$

This definition of the divergence can be used to derive a working expression for it in general curvilinear coordinates (Volume II, Chap. 7).

Divergence can also be formally written via the del operator (5.71) as a dot product with the field:

$$\begin{aligned} \nabla \cdot \mathbf{F} &= \left(\frac{\partial}{\partial x}\mathbf{i} + \frac{\partial}{\partial y}\mathbf{j} + \frac{\partial}{\partial z}\mathbf{k} \right) \cdot \left(F_x \mathbf{i} + F_y \mathbf{j} + F_z \mathbf{k} \right) \\ &= \frac{\partial}{\partial x} F_x + \frac{\partial}{\partial y} F_y + \frac{\partial}{\partial z} F_z \quad \longmapsto \quad \operatorname{div} \mathbf{F}, \end{aligned} \tag{6.92}$$

which indeed is the divergence (6.79).

As an example of calculating the divergence, consider the vector field $\mathbf{F} = e^{-r^2} (yz\mathbf{i} + xz\mathbf{j} + xy\mathbf{k})$ at the point $A(1, 1, 1)$. Here, $r^2 = x^2 + y^2 + z^2$, so that the divergence is easily calculated from its definition using the necessary partial differentiation:

$$\nabla \cdot \mathbf{F} = \frac{\partial}{\partial x}\left(e^{-r^2}yz\right) + \frac{\partial}{\partial y}\left(e^{-r^2}xz\right) + \frac{\partial}{\partial z}\left(e^{-r^2}xy\right)$$
$$= e^{-r^2}(-2xyz) + e^{-r^2}(-2xyz) + e^{-r^2}(-2xyz) = -6xyze^{-r^2},$$

and at the point A we have: $\nabla \cdot \mathbf{F}(A) = -6e^{-3}$.

The divergence obeys some important properties. First of all, it is obviously additive:

$$\nabla \cdot (\mathbf{F}_1 + \mathbf{F}_2) = \nabla \cdot \mathbf{F}_1 + \nabla \cdot \mathbf{F}_2.$$

Other properties are given in the Problems below.

Problem 6.68. *Using the definition of the divergence in Cartesian coordinates, prove that*

$$\nabla \cdot (U\mathbf{F}) = U(\nabla \cdot \mathbf{F}) + \mathbf{F} \cdot (\nabla U), \tag{6.93}$$

where in the first term in the right-hand side there is a divergence of a vector field $\mathbf{F}$*, while in the second—gradient of the scalar field* U*. Using this result, calculate* $\nabla \cdot (xy\mathbf{F})$ *with* $\mathbf{F} = \left(y^2, 0, xz\right)$*. Compare your result with direct calculation of the divergence of the vector field* $\mathbf{G} = \left(xy^3, 0, x^2yz\right)$*.* [*Answer*: $y\left(y^2 + x^2\right)$.]

Problem 6.69. *One can also apply the divergence to the gradient of some scalar field* U*. Show that*

$$\operatorname{div}(\operatorname{grad} U) = \Delta U = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right)U = \frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} + \frac{\partial^2 U}{\partial z^2}. \tag{6.94}$$

Problem 6.70. *Prove that*

$$\operatorname{div}(\phi \operatorname{grad} \psi) = \phi \Delta \psi + \nabla \phi \cdot \nabla \psi,$$

where ϕ *and* ψ *are scalar fields.*

The construction (6.94) has a special name, *Laplacian*, and various notations for it are frequently used in the literature. Apart from the notation Δ for the *Laplace operator* which appears within the round brackets in (6.94), the notation $\nabla^2 = \nabla \cdot \nabla$

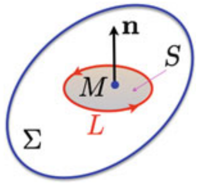


Fig. 6.36 A closed contour L around the point M on the surface Σ

is also often used since it can be easily recognised that the dot product of the del operator with itself is in fact Δ:

$$\nabla \cdot \nabla = \left(\frac{\partial}{\partial x}\mathbf{i} + \frac{\partial}{\partial y}\mathbf{j} + \frac{\partial}{\partial z}\mathbf{k}\right) \cdot \left(\frac{\partial}{\partial x}\mathbf{i} + \frac{\partial}{\partial y}\mathbf{j} + \frac{\partial}{\partial z}\mathbf{k}\right)$$
$$\longmapsto \quad \frac{\partial}{\partial x}\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\frac{\partial}{\partial y} + \frac{\partial}{\partial z}\frac{\partial}{\partial z} = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} = \Delta.$$

6.6.2 *Curl of a Vector Field*

We have already formally introduced the *curl* of a vector field $\mathbf{F}(\mathbf{r})$ by formula (6.65). Similarly to the divergence, a more general definition to the curl can also be given, and this is important as a tool to derive alternative expressions in other coordinate systems.

Consider a point M on a surface Σ, Fig. 6.36. Enclose M by a closed line L lying in the surface and consider the line integral of $\mathbf{F}$ along it divided by the area A of the closed loop, $\oint_L \mathbf{F} \cdot d\mathbf{l}/A$. Using Stokes's theorem (6.64), we can write:

$$\frac{1}{A}\oint_L \mathbf{F} \cdot d\mathbf{l} = \frac{1}{A}\int\int_S \operatorname{curl}\mathbf{F} \cdot d\mathbf{A} = \frac{1}{A}\int\int_S (\operatorname{curl}\mathbf{F} \cdot \mathbf{n})\, dA. \tag{6.95}$$

Above, S is the surface on Σ for which the line L serves as the boundary. Note the direction of the traverse along L is related to the normal of the surface Σ at M in the usual way (the right-hand rule). According to the average value theorem,

$$\frac{1}{A}\int\int_S (\operatorname{curl}\mathbf{F} \cdot \mathbf{n})\, dA = \frac{1}{A}(\operatorname{curl}\mathbf{F}(P) \cdot \mathbf{n})\, A = \operatorname{curl}\mathbf{F}(P) \cdot \mathbf{n},$$

where the point P lies somewhere inside surface S. It is seen then that in the limit of the area A tending to zero (assuming the point M is always inside S), the point $P \to M$, and we arrive at the dot product of the curl $\mathbf{F}(M)$ at the point M and the normal vector there:

$$\lim_{A\to 0}\frac{1}{A}\oint_L \mathbf{F} \cdot d\mathbf{l} = \operatorname{curl}\mathbf{F} \cdot \mathbf{n}. \tag{6.96}$$

This is the general expression for the curl we have been looking for, as it does not depend on the particular coordinate system. It can be used for the derivation of the curl in a general curvilinear coordinate system (Chap. 7, Volume II).

To understand better what does the curl mean, let us consider a fluid with the velocity distribution $\mathbf{v}(\mathbf{r})$. An *average* velocity *along* some circular contour L can be defined as

$$v_{av} = \frac{1}{2\pi R} \oint_L \mathbf{v} \cdot d\mathbf{l},$$

where R is the circle radius. The intensity of the rotation of fluid particles can be characterised by their angular velocity

$$\omega \sim \frac{v_{av}}{R} = \frac{1}{2\pi R^2} \oint_L \mathbf{v} \cdot d\mathbf{l} \sim \frac{1}{\pi R^2} \oint_L \mathbf{v} \cdot d\mathbf{l} = \frac{1}{A} \oint_L \mathbf{v} \cdot d\mathbf{l},$$

where $A = \pi R^2$ is the area of the circle. This result can now be compared with Eqs. (6.95) and (6.96). We see that $\mathrm{curl}\,\mathbf{v}(\mathbf{r})$ characterises the *intensity* of the fluid rotation around a given point $\mathbf{r}$ about the chosen direction given by the normal $\mathbf{n}$ to the closed loop L. The maximum intensity is achieved along the direction of the curl of $\mathbf{v}$.

Interestingly, curl of $\mathbf{F}$ can also be written via the del operator (5.71) as a *vector product* of ∇ and $\mathbf{F}$:

$$\mathrm{curl}\,\mathbf{F} = \nabla \times \mathbf{F}, \tag{6.97}$$

which is checked directly by looking at Eq. (6.73).

As an example, consider the curl of the vector field $\mathbf{F}(\mathbf{r}) = \mathbf{r}$. We have:

$$\mathrm{curl}\,\mathbf{F} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \partial/\partial x & \partial/\partial y & \partial/\partial z \\ x & y & z \end{vmatrix} = \left(\frac{\partial z}{\partial y} - \frac{\partial y}{\partial z}\right)\mathbf{i} - \left(\frac{\partial z}{\partial x} - \frac{\partial x}{\partial z}\right)\mathbf{j} + \left(\frac{\partial y}{\partial x} - \frac{\partial x}{\partial y}\right)\mathbf{k} = \mathbf{0},$$

which is consistent with the intuitive understanding of the curl, as the field $\mathbf{r}$ is radial (it corresponds to rays coming directly out of the centre of the coordinate system) and hence does not have rotations.

Now, let us look at another vector field

$$\mathbf{v}(\mathbf{r}) = \mathbf{w} \times \mathbf{r} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ w_x & w_y & w_z \\ x & y & z \end{vmatrix} = \left(w_y z - w_z y\right)\mathbf{i} + (w_z x - w_x z)\,\mathbf{j} + \left(w_x y - w_y x\right)\mathbf{k},$$

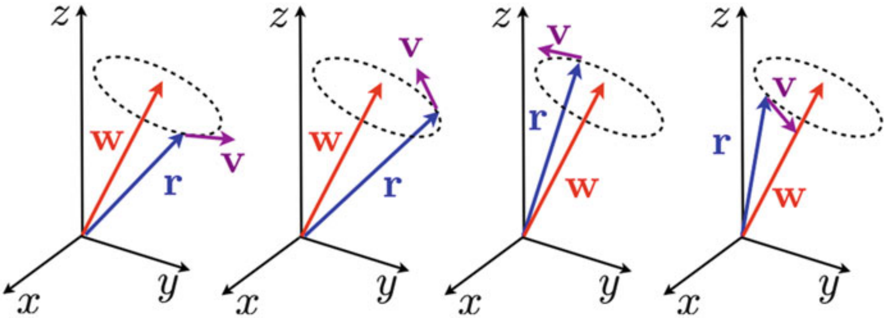


Fig. 6.37 The vector field $\mathbf{v}(\mathbf{r}) = \mathbf{w} \times \mathbf{r}$ corresponds to a vortex around the fixed vector $\mathbf{w}$ as the vector $\mathbf{r}$ samples the three-dimensional space

where $\mathbf{w}$ is a fixed vector. This may correspond, e.g., to a velocity distribution of a fluid in a vortex as shown in Fig. 6.37. Using again the general formula (6.73) for the curl, we write:

$$
\begin{aligned}
\operatorname{curl}\mathbf{v} &= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \partial/\partial x & \partial/\partial y & \partial/\partial z \\ w_y z - w_z y & w_z x - w_x z & w_x y - w_y x \end{vmatrix} \\
&= \left[\frac{\partial}{\partial y}\left(w_x y - w_y x\right) - \frac{\partial}{\partial z}\left(w_z x - w_x z\right)\right]\mathbf{i} \\
&\quad - \left[\frac{\partial}{\partial x}\left(w_x y - w_y x\right) - \frac{\partial}{\partial z}\left(w_y z - w_z y\right)\right]\mathbf{j} \\
&\quad + \left[\frac{\partial}{\partial x}\left(w_z x - w_x z\right) - \frac{\partial}{\partial y}\left(w_y z - w_z y\right)\right]\mathbf{k} \\
&= 2w_x\mathbf{i} + 2w_y\mathbf{j} + 2w_z\mathbf{k} = 2\mathbf{w}.
\end{aligned}
$$

This result is in complete agreement with what our intuitive understanding of the curl: it shows the direction (along the vector $\mathbf{w}$) and the strength (the velocity is proportional to the magnitude $w = |\mathbf{w}|$ of $\mathbf{w}$) of the rotation of the fluid around the direction given by the vector $\mathbf{w}$.

The curl possesses some important properties, e.g. additivity:

$$\operatorname{curl}(\mathbf{F}_1 + \mathbf{F}_2) = \operatorname{curl}\mathbf{F}_1 + \operatorname{curl}\mathbf{F}_2. \tag{6.98}$$

Another useful property is this:

$$\operatorname{div}\operatorname{curl}\mathbf{F} = 0. \tag{6.99}$$

This identity can be checked directly using an explicit expression (6.65) for the curl and the definition (6.79) of the div:

$$\nabla \cdot (\nabla \times \mathbf{F}) = \frac{\partial}{\partial x}\left(\frac{\partial F_z}{\partial y} - \frac{\partial F_y}{\partial z}\right) + \frac{\partial}{\partial y}\left(\frac{\partial F_x}{\partial z} - \frac{\partial F_z}{\partial x}\right) + \frac{\partial}{\partial z}\left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y}\right)$$
$$= \left(\frac{\partial^2 F_z}{\partial x \partial y} - \frac{\partial^2 F_y}{\partial x \partial z}\right) + \left(\frac{\partial^2 F_x}{\partial y \partial z} - \frac{\partial^2 F_z}{\partial y \partial x}\right) + \left(\frac{\partial^2 F_y}{\partial z \partial x} - \frac{\partial^2 F_x}{\partial z \partial y}\right) = 0 \ .$$

Problem 6.71. *Using the definition of the curl in Cartesian coordinates, prove that*

$$\text{curl}\,(U\mathbf{F}) = U\text{curl}\,\mathbf{F} - \mathbf{F} \times \text{grad}\,U, \tag{6.100}$$

where in the first term in the right-hand side there is a curl of a vector field $\mathbf{F}$*, while in the second gradient of a scalar field* U*.*

Problem 6.72. *Similarly, prove the identities:*

$$\text{curl grad}\,U = \mathbf{0}; \tag{6.101}$$

$$\text{curl curl}\,\boldsymbol{F} = \text{grad div}\,\mathbf{F} - \Delta\mathbf{F}, \tag{6.102}$$

where $\Delta\mathbf{F} = \Delta F_x \mathbf{i} + \Delta F_y \mathbf{j} + \Delta F_z \mathbf{k}$ *is the vector Laplacian.*

6.6.3 Vector Fields: Scalar and Vector Potentials

In applications related to vector fields, it is important to be able to characterise them. Here, we shall consider the so-called conservative and solenoidal vector fields since these possess certain essential properties. The field $\mathbf{F}(\mathbf{r})$ is called *conservative*, if one can introduce a *scalar potential* $\phi(\mathbf{r})$ such that $\mathbf{F} = \text{grad}\,\phi = \nabla\phi$. The field $\mathbf{F}(\mathbf{r})$ is called *solenoidal,* if there exists a *vector potential* $\mathbf{G}(\mathbf{r})$ such that $\mathbf{F} = \text{curl}\,\mathbf{G} = \nabla \times \mathbf{G}$. It appears that any vector field can always be represented as a sum of some conservative and solenoidal fields. This theorem which we shall discuss in more detail below plays an extremely important role in many applications in physics.

6.6.3.1 Conservative Fields

If the field $\mathbf{F}$ is conservative with the potential $\phi(\mathbf{r})$, i.e. $\mathbf{F} = \nabla\phi$, then the scalar function $\phi'(\mathbf{r}) = \phi(\mathbf{r}) + C$ is also a potential for any constant C. Thus, the conservative fields are determined by a single scalar function $\phi(\mathbf{r})$ which is defined up to a constant.

It is easy to see, for instance, that the electrostatic field $\mathbf{E}(\mathbf{r}) = -q\mathbf{r}/r^3$, where $r = |\mathbf{r}| = \sqrt{x^2 + y^2 + z^2}$, is conservative. Consider a function $\phi = q/r = q\left(x^2 + y^2 + z^2\right)^{-1/2}$ and let us show that it can be chosen as a potential for $\mathbf{E}$. We have $\partial\phi/\partial x = -qx/r^3$, etc., so we can write

$$\nabla\phi = \frac{\partial\phi}{\partial x}\mathbf{i} + \frac{\partial\phi}{\partial y}\mathbf{j} + \frac{\partial\phi}{\partial z}\mathbf{k} = \frac{-xq}{r^3}\mathbf{i} + \frac{-yq}{r^3}\mathbf{j} + \frac{-zq}{r^3}\mathbf{k}$$
$$= -\frac{q}{r^3}\left(x\mathbf{i} + y\mathbf{j} + z\mathbf{k}\right) = -\frac{q}{r^3}\mathbf{r} = \mathbf{E}(\mathbf{r}),$$

as required. So, the electrostatic potential in electrostatics (no moving charges) is basically, from the mathematical point of view, a potential of the conservative electric field (see Sect. 6.7.1 for more details).

The conservative field has a number of important properties that are provided by the following

Theorem 6.9. *Let the field* $\mathbf{F}(\mathbf{r})$ *be defined in a* simply connected *region* D*. Then the following three conditions are equivalent:*

1. *the field is conservative, i.e. there exists a scalar potential function* $\phi(\mathbf{r})$ *such that* $\mathbf{F} = \nabla\phi$*;*
2. *the field is* irrotational*, i.e.* $\operatorname{curl}\mathbf{F} = 0$*;*
3. *the line integral* $\oint_L \mathbf{F}\cdot d\mathbf{l} = 0$ *for any closed contour* L*.*

Proof. We shall proof the theorem using the familiar scheme: $1 \to 2 \to 3 \to 1$.

• $1 \to 2$: if $\mathbf{F} = \nabla\phi = \operatorname{grad}\phi$, then $\operatorname{curl}\mathbf{F} = \operatorname{curl}\operatorname{grad}\phi = 0$ according to Eq. (6.101).

• $2 \to 3$: take a closed contour L, then, using the Stokes's theorem (6.64),

$$\oint_L \mathbf{F}\cdot d\mathbf{l} = \int\int_S \operatorname{curl}\mathbf{F}\cdot d\mathbf{A} = 0,$$

since $\operatorname{curl}\mathbf{F} = \mathbf{0}$.

• $3 \to 1$: since $\oint_L \mathbf{F}\cdot d\mathbf{l} = 0$ for any contour L, then the line integral $\int_A^B \mathbf{F}\cdot d\mathbf{l}$ does not depend on the particular integration path $A\,(x_0, y_0, z_0) \to B(x, y, z)$, but only on the initial and final points A and B (see Sect. 6.4.6). Therefore, the function

$$\phi(x, y, z) = \int_{A(x_0,y_0,z_0)}^{B(x,y,z)} \mathbf{F}\cdot d\mathbf{l} \tag{6.103}$$

is the exact differential, and thus $\partial\phi/\partial x = F_x$, $\partial\phi/\partial y = F_y$ and $\partial\phi/\partial z = F_z$, i.e. $\mathbf{F} = \nabla\phi$, as required. **Q.E.D.**

Equation (6.103) gives a practical way of calculating the potential $\phi(\mathbf{r})$ for a conservative field from the field itself. As we know from the previous sections, this integral is only defined up to a constant, and this is how it should be for the potential, as was mentioned above.

It is essential that the region D is simply connected. The classical example is a magnetic field due to an infinite wire stretched along the z axis:

$$\mathbf{H}(\mathbf{r}) = \frac{\lambda\left(-y\hat{i} + x\hat{j}\right)}{\rho^2},$$

where $\rho = \sqrt{x^2 + y^2}$ is a distance from the wire. It is easy to see that

$$\begin{aligned}\operatorname{curl}\mathbf{H} &= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \partial/\partial x & \partial/\partial y & \partial/\partial z \\ -\lambda y/\rho^2 & \lambda x/\rho^2 & 0 \end{vmatrix} \\ &= \left(-\frac{\partial}{\partial z}\frac{\lambda x}{\rho^2}\right)\mathbf{i} - \left(\frac{\partial}{\partial z}\frac{\lambda y}{\rho^2}\right)\mathbf{j} + \left(\frac{\partial}{\partial x}\frac{\lambda x}{\rho^2} + \frac{\partial}{\partial y}\frac{\lambda y}{\rho^2}\right)\mathbf{k} = \mathbf{0}\end{aligned}$$

(perform differentiation and prove the result above!). However, the line integral $\oint_L \mathbf{H}\cdot d\mathbf{l}$ taken along the circle $x^2 + y^2 = 1$ (e.g. for $z = 0$) is not equal to zero. Using polar coordinates $x = \cos\phi$ and $y = \sin\phi$, we have:

$$\begin{aligned}\oint_L \mathbf{H}\cdot d\mathbf{l} &= \oint_L \frac{\lambda}{x^2 + y^2}\left(-ydx + xdy\right) = \int_0^{2\pi} \frac{\lambda}{1^2}\left[-\sin\phi\left(-\sin\phi\right)\right. \\ &\quad \left.+ \cos\phi\left(\cos\phi\right)\right] d\phi = \lambda\int_0^{2\pi} d\phi = 2\pi\lambda \neq 0.\end{aligned}$$

The obtained non-zero result for the line integral is because the region is not simply connected as the field $\mathbf{H}$ is infinite at any point along the wire $x = y = 0$. Hence, the whole z axis must be removed from D which makes the region not simply connected. Of course, any closed loop line integral which does not have the z axis inside the loop will be equal to zero as required by Theorem 6.9.

Problem 6.73. *Consider the vector field* $\mathbf{F} = (2x, 2y, 2z)$. *Prove that it is conservative. Then, using the line integral, show that its scalar potential* $\phi = x^2 + y^2 + z^2 + C$. *Check that indeed* $\mathbf{F} = \nabla\cdot\phi$.

Problem 6.74. *The same for* $\mathbf{F} = (yz, xz, xy)$. [*Answer*: $\phi = xyz + C$.]

Problem 6.75. *The same for* $\mathbf{F} = (y, x, -z)$. [*Answer*: $\phi = xy - z^2/2 + C$.]

Problem 6.76. *Given the vector force field* $\mathbf{F} = (2xz, 2yz, x^2 + y^2)$*, show that* $\mathbf{F}$ *is conservative. Explicitly calculate the line integral along the closed path* $x^2 + z^2 = 1$ *and* $y = 0$ *using polar coordinates. Verify that* $dU = F_x dx + F_y dy + F_z dz$ *is the exact differential, and next show that the function* $U(x, y, z) = \left(x^2 + y^2\right) z + C$*. State the relationship between the field* $\mathbf{F}$ *and the function* $U(x, y, z)$*.*

Problem 6.77. *Repeat all steps of the previous problem but for the vector field* $\mathbf{F} = (y + z, x + z, x + y)$*, showing that in this case* $U = xz + yz + xy + C$*.*

6.6.3.2 Solenoidal Fields

As was already mentioned, if a vector field $\mathbf{A}$ can be represented as a curl of another field, $\mathbf{B}$, called its *vector potential*, i.e. $\mathbf{A} = \operatorname{curl}\mathbf{B} = \nabla \times \mathbf{B}$, then the field $\mathbf{A}$ is called *solenoidal*. For this field, see Eq. (6.99),

$$\operatorname{div}\mathbf{A} = \operatorname{div}\operatorname{curl}\mathbf{B} = 0 \quad \text{or} \quad \nabla \cdot \mathbf{A} = 0. \tag{6.104}$$

This condition serves as another definition of the solenoidal field. However, it is not only the necessary condition as proven by the above manipulation. It is also a sufficient one as the reverse statement is also valid: if $\operatorname{div}\mathbf{A} = 0$, then there exists a vector potential $\mathbf{B}$, such that $\mathbf{A} = \operatorname{curl}\mathbf{B}$. To prove the existence, choose two points $M_0(x_0, y_0, z_0)$ and $M(x, y, z)$, and then consider a vector function $\mathbf{B} = (B_x, B_y, B_z)$ defined in the following way:

$$B_x = \int_{z_0}^{z} A_y(x, y, z_1)dz_1, \quad B_y = \int_{x_0}^{x} A_z(x_1, y, z_0)dx_1 - \int_{z_0}^{z} A_x(x, y, z_1)dz_1 \quad \text{and} \quad B_z = 0. \tag{6.105}$$

Consider the curl of this vector field. Starting from the x component of the curl, we write:

$$(\operatorname{curl}\mathbf{B})_x = \frac{\partial B_z}{\partial y} - \frac{\partial B_y}{\partial z} = -\frac{\partial B_y}{\partial z} = -\frac{\partial}{\partial z}\left(\int_{x_0}^{x} A_z(x_1, y, z_0)dx_1 - \int_{z_0}^{z} A_x(x, y, z_1)dz_1\right).$$

The first term does not contribute since it does not depend on z; by differentiating the second integral over the upper limit we get the integrand, see Sect. 4.3. Therefore,

$$(\operatorname{curl}\mathbf{B})_x = \frac{\partial}{\partial z}\int_{z_0}^{z} A_x(x, y, z_1)dz_1 = A_x(x, y, z).$$

Similarly,

$$(\mathrm{curl}\,\mathbf{B})_y = \frac{\partial B_x}{\partial z} - \frac{\partial B_z}{\partial x} = \frac{\partial B_x}{\partial z} = \frac{\partial}{\partial z}\left(\int_{z_0}^{z} A_y(x,y,z_1)dz_1\right) = A_y(x,y,z),$$

and

$$(\mathrm{curl}\,\mathbf{B})_z = \frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y} = \frac{\partial}{\partial x}\left(\int_{x_0}^{x} A_z(x_1,y,z_0)dx_1 - \int_{z_0}^{z} A_x(x,y,z_1)dz_1\right)$$
$$-\frac{\partial}{\partial y}\left(\int_{z_0}^{z} A_y(x,y,z_1)dz_1\right),$$

where differentiation of the first integral gives $A_z\,(x,y,z_0)$ in the same way as for the above two components, but in the two integrals taken over z_1 the differentiation is to be performed with respect to the variable which happens to be only in the integrand. According to Sect. 4.5.2, the operators $\partial/\partial x$ and $\partial/\partial y$ can then be inserted inside the integrals since the integration limits do not depend on the variables with respect to which the differentiation is performed. Hence,

$$\begin{aligned}(\mathrm{curl}\,\mathbf{B})_z &= A_z(x,y,z_0) - \int_{z_0}^{z} \frac{\partial A_x(x,y,z_1)}{\partial x}dz_1 - \int_{z_0}^{z} \frac{\partial A_y(x,y,z_1)}{\partial y}dz_1 \\ &= A_z(x,y,z_0) - \int_{z_0}^{z}\left[\frac{\partial A_x(x,y,z_1)}{\partial x} + \frac{\partial A_y(x,y,z_1)}{\partial y}\right]dz_1.\end{aligned}$$

Since we know that $\mathrm{div}\mathbf{A} = \partial A_x/\partial x + \partial A_x/\partial y + \partial A_x/\partial z = 0$, the expression in the square brackets above can be replaced by $-\partial A_z\,(x,y,z_1)\,/\partial z_1$, yielding finally

$$\begin{aligned}(\mathrm{curl}\,\mathbf{B})_z &= A_z(x,y,z_0) + \int_{z_0}^{z} \frac{\partial A_z(x,y,z_1)}{\partial z_1}dz_1 \\ &= A_z(x,y,z_0) + [A_z(x,y,z) - A_z(x,y,z_0)] = A_z(x,y,z).\end{aligned}$$

Thus, we have shown that indeed $\mathrm{curl}\mathbf{B} = \mathbf{A}$. This proves the made above statement that there *exists* a vector field $\mathbf{B}$ that can serve as a vector potential for the solenoidal field $\mathbf{A}$. Of course, the vector potential (6.105) is not the only one which can do this job (see below).

Thus, the two conditions for $\mathbf{A}$ to be a solenoidal field, namely that $\mathbf{A}$ is curl of some $\mathbf{B}$ and that $\mathrm{div}\mathbf{A} = 0$, are completely equivalent.

The choice of the vector potential $\mathbf{B}$ given above is not unique: any vector $\mathbf{B}' = \mathbf{B} + \mathrm{grad}\,U(x,y,z)$ with arbitrary scalar field $U(x,y,z)$ is also a vector potential since $\mathrm{curl}\,\mathrm{grad}\,U = \nabla\times\nabla U = \mathbf{0}$ for any U, see Eq. (6.101), yielding $\mathrm{curl}\,\mathbf{B} = \mathrm{curl}\,\mathbf{B}'$. Thus, the vector potential is defined only up to the gradient of an arbitrary scalar field.

Problem 6.78. *Show that the vector field* $\mathbf{F} = \left(-xz\cosh y, z\sinh y, x^2y\right)$ *is solenoidal.*

Problem 6.79. *Consider a point charge* q *at a position given by the vector* $\mathbf{R} = (X, Y, Z)$. *It creates an electric field*

$$\mathbf{E}(\mathbf{r}) = q\frac{\mathbf{r} - \mathbf{R}}{|\mathbf{r} - \mathbf{R}|^3}$$

at point $\mathbf{r}$, *where* $|\mathbf{r} - \mathbf{R}| = \sqrt{(x-X)^2 + (y-Y)^2 + (z-Z)^2}$. *Show that the field is solenoidal anywhere outside the position of the charge* q, *i.e. that*

$$\operatorname{div}\mathbf{E}(\mathbf{r}) = 0 \quad \text{if} \quad \mathbf{r} \neq \mathbf{R}. \tag{6.106}$$

Of course, the above statement is immediately generalised for the electrostatic field of arbitrary number of point charges.

Any solenoidal field has a number of properties as stated by the following:

Theorem 6.10. *If* $\mathbf{F}$ *is a solenoidal field, then its flux through any smooth closed surface is equal to zero.*

Proof. First, consider a simply connected region D and a closed smooth surface S anywhere inside it. Then, the flux through that surface

$$\oiint_S \mathbf{F}\cdot d\mathbf{A} = \int\int\int_V \operatorname{div}\mathbf{F}\, dxdydz = 0,$$

since the field is solenoidal, $\operatorname{div}\mathbf{F} = 0$. Here we have used the Gauss theorem, Eq. (6.78), and V is the region that is covered by S.

Next, consider a more general region D with the closed surface S, which contains a hole inside as schematically shown in Fig. 6.38. Let us divide it into two parts D_1 and D_2 by cutting through D with a plane P. The plane will cross S at the line L, and the surface S will also be broken down into two surfaces S_1 (left) and S_2 (right) that both cap the line L on both sides. Then, the surface integral over the entire outer surface of D,

$$\oiint_S \mathbf{F}\cdot d\mathbf{A} = \int\int_{S_1} \mathbf{F}\cdot d\mathbf{A} + \int\int_{S_2} \mathbf{F}\cdot d\mathbf{A},$$

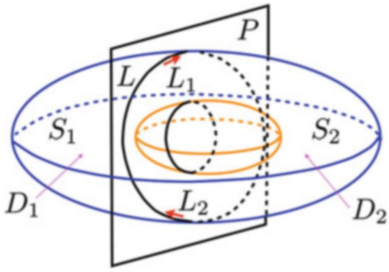


Fig. 6.38 Region D (*blue*) contains a hole inside (*orange*). The plane P cuts it into two regions D_1 and D_2, and crosses the surface S at the line L

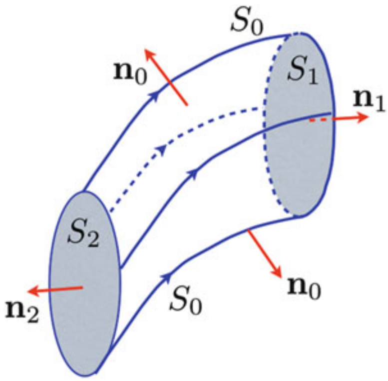


Fig. 6.39 Lines of flow of a liquid in a tube

can be represented as a sum of two flux integrals: over the surface S_1 and over S_2. Since div$\mathbf{F} = 0$, there exists a vector potential $\mathbf{B}$, so that $\mathbf{F} = \text{curl}\,\mathbf{B}$. Therefore, using the Stokes's theorem, Eq. (6.64), we have for each of the surfaces:

$$\int\int_{S_1} \mathbf{F} \cdot d\mathbf{A} = \int\int_{S_1} \text{curl}\,\mathbf{B} \cdot d\mathbf{A} = \oint_{L_1} \mathbf{B} \cdot d\mathbf{l},$$

where L_1 is the boundary of S_1 along the line L in the direction as indicated in Fig. 6.38. Similarly,

$$\int\int_{S_2} \mathbf{F} \cdot d\mathbf{A} = \int\int_{S_2} \text{curl}\,\mathbf{B} \cdot d\mathbf{A} = \oint_{L_2} \mathbf{B} \cdot d\mathbf{l},$$

L_2 is the boundary of the capping surface S_2. L_2 also goes along the line L but with the traverse direction opposite to that of L_1, Fig. 6.38. Since L_1 is opposite to L_2, the sum of the two line integrals is zero which proves that even in this case the flux is also zero. Above, we have only discussed the case with a single "hole"; obviously, the case of many "holes" can be considered along the same lines. **Q.E.D.**

Theorem 6.10 has a very simple consequence. Consider a flow of a fluid described by a vector field of velocity $\mathbf{v}$ that is solenoidal. Let us choose a surface S_0 around its lines of flow, Fig. 6.39, in such a way that the normal $\mathbf{n}_0$ to S_0 at every point is perpendicular to $\mathbf{v}$. It means that every line of flow does not cross S_0 and remains inside it, so that S_0 forms a sort of a tube. Now, consider two surfaces S_1 and S_2 on both sides of the tube with normals $\mathbf{n}_1$ and $\mathbf{n}_2$ directed outward as shown. A flux through the whole surface $S_0 + S_1 + S_2$ is zero for the solenoidal flow as was proven above. However, the flux through S_0 is also zero since at each point on the surface, $\mathbf{v}$ is perpendicular to $\mathbf{n}_0$. Therefore, the flux through $S_1 + S_2$ is zero:

$$\int\int_{S_1} \mathbf{v}\cdot\mathbf{n}_1 dA + \int\int_{S_2} \mathbf{v}\cdot\mathbf{n}_2 dA = \int\int_{S_1} \mathbf{v}\cdot\mathbf{n}_1 dA - \int\int_{S_2} \mathbf{v}\cdot(-\mathbf{n}_2) dA = 0$$

or

$$\int\int_{S_1} \mathbf{v}\cdot\mathbf{n}_1 dA = \int\int_{S_2} \mathbf{v}\cdot(-\mathbf{n}_2) dA,$$

where $-\mathbf{n}_2$ is the normal to S_2 directed along the line of flow similarly to $\mathbf{n}_1$. Thus, the flux through any surface cutting the tube is conserved along the tube. This also means that the lines of flow cannot converge into a point since in this case the flux through the tube will become zero which is impossible (as it should be conserved).

6.6.3.3 Expansion of a General Field into Solenoidal and Conservative Components

Consider a vector field $\mathbf{A}(\mathbf{r})$ defined in some *simply connected* region D which has non-zero divergence and curl, i.e. $\mathrm{div}\mathbf{A}(\mathbf{r}) = f(\mathbf{r}) \neq 0$ and $\mathrm{curl}\,\mathbf{A}(\mathbf{r}) = \mathbf{g}(\mathbf{r}) \neq 0$, where $f(\mathbf{r})$ and $\mathbf{g}(\mathbf{r})$ are some rather general scalar and vector functions, respectively.

Theorem 6.11. *Any vector field* $\mathbf{A}(\mathbf{r})$ *defined in a simply connected region D can always be represented as a sum of a conservative,* $\mathbf{A}_c(\mathbf{r})$*, and solenoidal,* $\mathbf{A}_s(\mathbf{r})$*, fields, i.e.*

$$\mathbf{A}(\mathbf{r}) = \mathbf{A}_c(\mathbf{r}) + \mathbf{A}_s(\mathbf{r}) \;\text{ with }\; \mathrm{curl}\,\mathbf{A}_c(\mathbf{r}) = 0 \;\text{ and }\; \mathrm{div}\mathbf{A}_s(\mathbf{r}) = 0. \qquad (6.107)$$

Proof. Let us first determine the conservative field $\mathbf{A}_c$ from the equations

$$\mathrm{curl}\,\mathbf{A}_c = 0 \quad \text{and} \quad \mathrm{div}\mathbf{A}_c = f(\mathbf{r}).$$

Since the field is conservative (due to our choice, $\mathrm{curl}\,\mathbf{A}_c = 0$), there exists a scalar potential U such that $\mathbf{A}_c = \mathrm{grad}\,U$. Therefore, for U we have an equation

$$\mathrm{div}\mathbf{A}_c = \mathrm{div}\,\mathrm{grad}\,U = \Delta U = f(\mathbf{r}). \qquad (6.108)$$

This is the so-called Poisson's equation, that always has a solution (see below), i.e. one can determine $\mathbf{A}_c$ by solving it.

Next, we consider the difference $\mathbf{A}-\mathbf{A}_c = \mathbf{A}_s$, and let us calculate its div and curl:

$$\mathrm{div}\mathbf{A}_s = \mathrm{div}\mathbf{A}-\mathrm{div}\mathbf{A}_c = f-f = 0 \quad \text{and} \quad \mathrm{curl}\,\mathbf{A}_s = \mathrm{curl}\,\mathbf{A}-\mathrm{curl}\,\mathbf{A}_c = \mathrm{curl}\,\mathbf{A} = \mathbf{g}.$$

We see that the difference, $\mathbf{A}-\mathbf{A}_c$ is, in fact, solenoidal. **Q.E.D.**

We have assumed above that the functions $\mathbf{g}(\mathbf{r})$ and $f(\mathbf{r})$ are not equal identically to zero in region D. If they were, then we would have $\mathrm{curl}\mathbf{A} = \mathbf{0}$ and $\mathrm{div}\mathbf{A} = 0$ anywhere in D, i.e. the field $\mathbf{A}(\mathbf{r})$ would be conservative and solenoidal at the same time. Since $\mathrm{curl}\,\mathbf{A} = \mathbf{0}$, then there exists a scalar potential U, such that $\mathbf{A} = \mathrm{grad}\,U$. However, since $\mathbf{A}$ is also solenoidal, $\mathrm{div}\mathbf{A} = \mathrm{div}\,\mathrm{grad}\,U = \Delta U = 0$, which means that the scalar potential U satisfies the so-called Laplace equation $\Delta U = 0$. Functions, that are solutions of the Laplace equation are called *harmonic functions*. Interestingly, the vector field satisfying $\mathrm{curl}\,\mathbf{A} = 0$ and $\mathrm{div}\mathbf{A} = 0$ is determined up to the gradient of a harmonic potential: $\mathbf{A}' = \mathbf{A} + \nabla U$. Indeed,

$$\mathrm{div}\mathbf{A}' - \mathrm{div}\mathbf{A} = \mathrm{div}\,\mathrm{grad}\,U = \Delta U = 0$$

for a harmonic function. The other condition,

$$\mathrm{curl}\,\mathbf{A}' - \mathrm{curl}\,\mathbf{A} = \mathrm{curl}\,\mathrm{grad}\,U = \mathbf{0},$$

is satisfied for any U since curl of grad is always equal to zero, Eq. (6.101).

It is possible to obtain the decomposition of a vector field $\mathbf{A} = \mathbf{A}_c + \mathbf{A}_s$ explicitly into a conservative and solenoidal components, given that

$$\mathrm{div}\mathbf{A} = f(\mathbf{r}) \quad \text{and} \quad \mathrm{curl}\,\mathbf{A} = \mathbf{g}(\mathbf{r}). \tag{6.109}$$

The method is based on a fact that a general solution of the Laplace equation $\Delta\phi = \rho$ can be written via a volume integral:

$$\phi(\mathbf{r}) = -\frac{1}{4\pi}\int\int\int_V \frac{\rho(\mathbf{r}')\,d\mathbf{r}'}{|\mathbf{r}-\mathbf{r}'|}. \tag{6.110}$$

Here the integration is performed over the volume of the spacial region V where $\rho(\mathbf{r})$ is defined. The notation for the volume differential, $d\mathbf{r}'$, used above for the volume integral is frequently used in physics literature; it simply means that $d\mathbf{r}' = dx'dy'dz' = dV'$; we use the prime here because $\mathbf{r}$ is already used in the left-hand side. We will not prove Eq. (6.110) here. It is done in Book II, see Sect. 5.3.2 there. However, this formula can be easily illustrated using known concepts of electrostatics discussed below in Sect. 6.7.1. It is discussed there, how the solution (6.110) of the Laplace equation $\Delta\phi = \rho$ can be justified, see specifically Eq. (6.119).[10]

[10]Note that the Laplace equations here and in electrostatics differ by an unimportant factor of -4π.

Since the conservative contribution $\mathbf{A}_c$ satisfies $\operatorname{curl}\mathbf{A}_c = 0$, one can introduce a potential $U(\mathbf{r})$ such that $\mathbf{A}_c = \operatorname{grad} U$. Also, since $\operatorname{div}\mathbf{A}_c = \operatorname{div}\mathbf{A} = f(\mathbf{r})$ (note that $\operatorname{div}\mathbf{A}_s = 0$ by construction), the potential U satisfies the Laplace equation,

$$\operatorname{div}\mathbf{A}_c = \operatorname{div}\operatorname{grad} U = \Delta U = f(\mathbf{r}),$$

with the general solution

$$U(\mathbf{r}) = -\frac{1}{4\pi}\int\int\int_V \frac{f(\mathbf{r}')\,d\mathbf{r}'}{|\mathbf{r}-\mathbf{r}'|}, \tag{6.111}$$

This formula solves for the $\mathbf{A}_c$ part, as U totally defines $\mathbf{A}_c$.

Now we need to find a general solution for the solenoidal component. The latter satisfies the equations:

$$\operatorname{div}\mathbf{A}_s = 0 \quad \text{and} \quad \operatorname{curl}\mathbf{A}_s = \operatorname{curl}\mathbf{A} = \mathbf{g}(\mathbf{r}), \tag{6.112}$$

since $\operatorname{curl}\mathbf{A}_c = 0$. Because the divergence of $\mathbf{A}_s$ is zero, there exists a vector potential $\mathbf{G}$ such that $\mathbf{A}_s = \operatorname{curl}\mathbf{G}$. Moreover, the choice of $\mathbf{G}$ is not unique, as $\mathbf{G}' = \mathbf{G} + \operatorname{grad}\Psi$ is also perfectly acceptable for any scalar field $\Psi(\mathbf{r})$ as we established above. On the other hand,

$$\operatorname{div}\mathbf{G}' = \operatorname{div}\mathbf{G} + \operatorname{div}\operatorname{grad}\Psi = \operatorname{div}\mathbf{G} + \Delta\Psi.$$

The field Ψ is arbitrary; we shall select it in such a way that $\Delta\Psi = -\operatorname{div}\mathbf{G}$. This is simply the Laplace equation with the right-hand side equal to $\operatorname{div}\mathbf{G}$. This means that we can always choose the vector potential $\mathbf{G}'$ being solenoidal, i.e. with $\operatorname{div}\mathbf{G}' = 0$. Hence, we have:

$$\operatorname{curl}\mathbf{A}_s = \mathbf{g}\,, \quad \operatorname{div}\mathbf{A}_s = 0 \quad \text{with} \quad \mathbf{A}_s = \operatorname{curl}\mathbf{G}' \quad \text{and} \quad \operatorname{div}\mathbf{G}' = 0.$$

From the first equation,

$$\operatorname{curl}\mathbf{A}_s = \operatorname{curl}\operatorname{curl}\mathbf{G}' = \operatorname{grad}\operatorname{div}\mathbf{G}' - \Delta\mathbf{G}' = -\Delta\mathbf{G}' = \mathbf{g},$$

where we used Eq. (6.102) and the fact that $\operatorname{div}\mathbf{G}' = 0$. It is seen that $\mathbf{G}'$ satisfies the vector Laplace equation, i.e. for each of its components we write a formula analogous to Eq. (6.111); in the vector form we have then:

$$\mathbf{G}'(\mathbf{r}) = \frac{1}{4\pi}\int\int\int_V \frac{\mathbf{g}(\mathbf{r}')\,d\mathbf{r}'}{|\mathbf{r}-\mathbf{r}'|}. \tag{6.113}$$

Equations (6.111) and (6.113) fully solve the decomposition problem of an arbitrary vector field $\mathbf{A}$ since the potentials U and $\mathbf{G}'$ define the conservative and solenoidal components of it, and hence we can write:

$$\mathbf{A}(\mathbf{r}) = \operatorname{grad} U(\mathbf{r}) + \operatorname{curl}\mathbf{G}'(\mathbf{r})\,. \tag{6.114}$$

6.7 Application of Integral Theorems in Physics: Part II

In this section more applications in physics of the calculus we developed in this chapter will be briefly outlined.

6.7.1 Maxwell's Equations

A nice application of integral theorems is provided by Maxwell's equations of electromagnetism. We shall denote the electric and magnetic fields in vacuum as $\mathbf{E}$ and $\mathbf{H}$ here.

Consider first electrostatics which is the simplest case. A point charge q creates around itself an electric field $\mathbf{E}(\mathbf{r}) = q\mathbf{r}/r^3$. A probe charge q_0 placed in this field experiences a force $\mathbf{F} = q_0\mathbf{E}$, where $\mathbf{E}$ is the electrostatic field due to q.

Problem 6.80. *Consider a sphere S_{sph} of radius R with the charge q placed at its centre. Show that the flux integral of* $\mathbf{E}$ *due to the charge q through the surface of the sphere is*

$$\oiint_{S_{\mathrm{sph}}} \mathbf{E} \cdot d\mathbf{A} = 4\pi q.$$

Note that the flux does not depend on the sphere radius. [*Hint*: choose the coordinate system with q in its centre and note that for any point on the sphere the normal $\mathbf{n}$ is proportional to $\mathbf{r}$.]

In fact, it is easy to see that the result is valid for any surface S, not necessarily the sphere. Indeed, as we have seen in the previous Sect. 6.6.3.2, the electrostatic field is solenoidal anywhere outside the charge. Therefore, if we surround the charge with a sphere S_{sph} which lies completely inside the surface S as shown in Fig. 6.40, then the flux through the sphere can be easily seen to be the same as the flux through S:

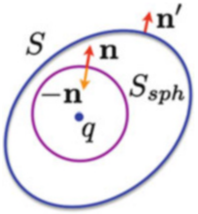


Fig. 6.40 Point charge q is surrounded by a sphere S_{sph} which in turn is surrounded by an arbitrary surface S

$$\oiint_S \mathbf{E}\cdot d\mathbf{A} = \left(\oiint_S \mathbf{E}\cdot d\mathbf{A} - \oiint_{S_{\text{sph}}} \mathbf{E}\cdot d\mathbf{A}\right) + \oiint_{S_{\text{sph}}} \mathbf{E}\cdot d\mathbf{A} = \oiint_{S_{\text{sph}}} \mathbf{E}\cdot d\mathbf{A} = 4\pi q,$$

since the expression in the brackets represents a flux through a closed surface $S+S_{\text{sph}}$ (with the outer normal $-\mathbf{n}$ used for the S_{sph} because of the minus sign before the flux integral) which is zero. This is because inside the region of space ΔV enclosed by the two surfaces S_{sph} and S the field $\mathbf{E}$ is solenoidal, and hence the flux is zero since

$$\oiint_S \mathbf{E}\cdot d\mathbf{A} - \oiint_{S_{\text{sph}}} \mathbf{E}\cdot d\mathbf{A} = \oiint_{S+S_{ph}} \mathbf{E}\cdot d\mathbf{A} = \int\int\int_{\Delta V} \text{div}\mathbf{E}\, dV = 0,$$

where the closed surface $S + S_{ph}$ encloses the region ΔV.

Next, if there are more than one charge inside some region, this result would be valid for each of the charges if the surface S contains all of them inside. But the fields due to every individual charge sum up (superposition principle). Therefore, generally, the flux of the total field $\mathbf{E}$ through a closed surface S containing charges q_1, q_2, etc. must satisfy the following equation:

$$\oiint_S \mathbf{E}\cdot d\mathbf{A} = 4\pi \sum_i q_i. \tag{6.115}$$

For a continuous distribution of the charge, we shall have instead in the right-hand side the total charge enclosed by the surface S, which is

$$Q = \int\int\int_V \rho dV,$$

where $dV = dxdydz$ is the differential volume element, V is the volume for which S is its boundary and ρ the charge density. Therefore,

$$\oiint_S \mathbf{E}\cdot d\mathbf{A} = 4\pi \int\int\int_V \rho dV.$$

The derived equation can in principle allow one to find the field $\mathbf{E}$ due to the charge distribution ρ. However, it is inconvenient in practical calculations as it contains integrals, it is much better to have an equation relating the field at each point to the charge density at that point. Such an equation can be obtained applying the Gauss (divergence) theorem to the flux integral in the left-hand side:

$$\oiint_S \mathbf{E}\cdot d\mathbf{A} = \int\int\int_V \text{div}\mathbf{E}\, dV \quad \Longrightarrow \quad \int\int\int (\text{div}\mathbf{E} - 4\pi\rho)\, dV = 0.$$

This result is valid for any volume V. Therefore, it must be that

$$\text{div}\mathbf{E} = 4\pi\rho. \tag{6.116}$$

This fundamental equation of electrostatics is the Maxwell's first equation.

The electrostatic field is also a conservative field for which one can choose[11] a potential ϕ, such that $\mathbf{E} = -\mathrm{grad}\,\phi$ (see Sect. 6.6.3.1; note that there only a single charge was considered; generalisation to many charges is, however, straightforward). Therefore, according to Theorem 6.9 of Sect. 6.6.3.1,

$$\mathrm{curl}\,\mathbf{E} = 0. \tag{6.117}$$

This is the Maxwell's second equation. The corresponding differential equation for the potential $\phi(\mathbf{r})$ is obtained by using $\mathbf{E} = -\mathrm{grad}\,\phi$ in Eq. (6.116):

$$\mathrm{div}\,\mathrm{grad}\,\phi = \Delta\phi = -4\pi\rho, \tag{6.118}$$

which is the familiar *Laplace equation*. It is straightforward to check explicitly that for a single point charge q located at $\mathbf{r}_q$, the potential at point $\mathbf{r}$ is $\phi(\mathbf{r}) = q/\left|\mathbf{r} - \mathbf{r}_q\right|$. Indeed, the electric field is easily calculated to be:

$$\begin{aligned}\mathbf{E} &= -\mathrm{grad}\,\frac{q}{\left|\mathbf{r} - \mathbf{r}_q\right|} \\ &= -\left(\mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y} + \mathbf{k}\frac{\partial}{\partial z}\right)\frac{q}{\sqrt{\left(x - x_q\right)^2 + \left(y - y_q\right)^2 + \left(z - z_q\right)^2}} = \frac{q\left(\mathbf{r} - \mathbf{r}_q\right)}{\left|\mathbf{r} - \mathbf{r}_q\right|^3},\end{aligned}$$

as required. Then, we know that the result for many charges is obtained simply by summing up contributions from all charges. If the charge is distributed continuously, then we divide up the space into small volumes $dV' = d\mathbf{r}'$ with the charge $dQ = \rho(\mathbf{r}')\,dV'$ in each, where the volume is positioned near the point $\mathbf{r}'$ with the charge density being $\rho(\mathbf{r}')$. Then the potential at point $\mathbf{r}$ due to this infinitesimal charge will be

$$d\phi = \frac{dQ}{|\mathbf{r} - \mathbf{r}'|} = \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}dV'.$$

The total potential due to the whole charge distribution is then given by the integral

$$\phi(\mathbf{r}) = \int\int\int_V \frac{\rho(\mathbf{r}')\,dV'}{|\mathbf{r} - \mathbf{r}'|}. \tag{6.119}$$

This formula can be considered as a general solution of the Laplace equation of Eq. (6.118).

[11] Note that it is customary in electrostatics to choose the minus sign when defining the potential.

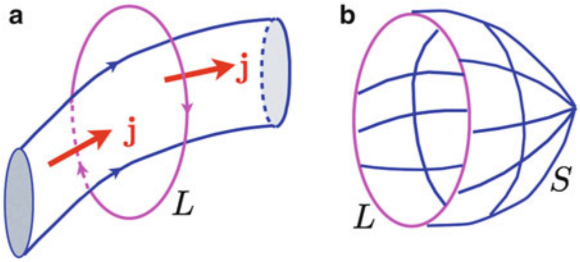


Fig. 6.41 (**a**) The line integral of the magnetic field **H** is taken around a wire with the current density **j**. (**b**) A surface S is chosen so that the line L is its boundary

Problem 6.81. *Consider a set of point charges* $\{q_i;\ i = 1, \ldots, n\}$ *positioned at* $\{\mathbf{r}_i;\ i = 1, \ldots, n\}$*, and prove by direct calculation that everywhere, including the positions of the charges themselves,* $\mathrm{curl}\,\mathbf{E} = 0$. [*Hint*: when considering the curl at the positions of the charges, consider a point close to one of their positions and then take the limit to approach that position.]

Moving charges create the magnetic field **H**, and the force acting on a probe charge q_0 moving with the velocity **v** in the magnetic field is given by the Lorenz formula: $\mathbf{F} = (q_0/c)\,[\mathbf{v} \times \mathbf{H}]$, where c is the speed of light. It was established experimentally by Ampère that a closed loop line integral of **H** around a wire with the current I is proportional to the current, see Fig. 6.41(a):

$$\oint_L \mathbf{H} \cdot d\mathbf{l} = \frac{4\pi}{c} I. \tag{6.120}$$

This is Ampère's circuital law which allows, e.g., calculating directly the magnetic field around an infinite wire.

Problem 6.82. *Show using Eq. (6.120) that the magnitude of the field* **H** *around an infinite wire of the current* I *at a distance* r *from the wire is* $H = 2I/cr$.

Maxwell generalised the result (6.120) for any flow of the current and postulated that the formula is valid for any closed-loop contour L surrounding the current I. Since the current can be expressed via the current density **j** using Eq. (6.89) with some surface S which has L as its boundary, see Fig. 6.41(b), then, upon using Stokes's theorem (6.64) for the line integral, we obtain:

$$\oint_L \mathbf{H} \cdot d\mathbf{l} = \int\!\!\int_S \mathrm{curl}\,\mathbf{H} \cdot d\mathbf{A} \quad \Longrightarrow \quad \int\!\!\int_S \mathrm{curl}\,\mathbf{H} \cdot d\mathbf{A} = \frac{4\pi}{c} \int\!\!\int_S \mathbf{j} \cdot d\mathbf{A},$$

and hence

$$\int\int_S \left(\operatorname{curl}\mathbf{H} - \frac{4\pi}{c}\mathbf{j}\right)\cdot d\mathbf{A} = 0 \quad \Longrightarrow \quad \operatorname{curl}\mathbf{H} = \frac{4\pi}{c}\mathbf{j}, \tag{6.121}$$

since the surface S (as well as the contour L) can be chosen arbitrarily. This is the next Maxwell's equation corresponding to stationary conditions (no time dependence).

Another experimental result states that the flux of the magnetic field through any closed surface S is zero.

Problem 6.83. *Prove from this that the magnetic field is solenoidal, i.e.*

$$\operatorname{div}\mathbf{H} = 0. \tag{6.122}$$

[*Hint*: use the divergence theorem.] *This is yet another Maxwell's equation.*

Problem 6.84. *Prove from Eq. (6.121) that* $\operatorname{div}\mathbf{j} = 0$. *This corresponds to the continuity equation (6.90) under stationary conditions.* [*Hint*: take the divergence of both sides.]

Equations (6.116), (6.117), (6.121) and (6.122) form the complete set of equations for the electric and magnetic fields in vacuum under stationary conditions. As we have seen in the last problem, they do not contradict the continuity equation.

Consider now the general case of the charge $\rho(\mathbf{r}, t)$ and current $\mathbf{j}(\mathbf{r}, t)$ densities depending not only on the spatial position $\mathbf{r}$, but also on time, t. This results in the fields $\mathbf{H}(\mathbf{r}, t)$ and $\mathbf{E}(\mathbf{r}, t)$ depending both on $\mathbf{r}$ and t. Faraday discovered the law of induction stating that the current will flow in a closed loop of wire L if the flux of the magnetic field through a surface S having the boundary L, see Fig. 6.41(b), changes in time. This observation can be formulated in the following way: the rate of change of the magnetic flux through S is equal to the electromotive force, i.e.

$$-\frac{1}{c}\frac{\partial}{\partial t}\int\int_S \mathbf{H}\cdot d\mathbf{A} = \oint_L \mathbf{E}\cdot d\mathbf{l}. \tag{6.123}$$

Problem 6.85. *Derive from this integral formulation of the induction law its differential analog:*

$$-\frac{1}{c}\frac{\partial\mathbf{H}}{\partial t} = \operatorname{curl}\mathbf{E}. \tag{6.124}$$

This equation replaces (6.117) for the case of non-stationary sources (charge and/or current densities).

Finally, Maxwell conceived the equation which was to replace Eq. (6.121) of the stationary case. He realised that a displacement current need to be added related to the rate of change of the electric flux through the surface S:

$$\oint_L \mathbf{H} \cdot d\mathbf{l} = \frac{4\pi}{c} I + \frac{1}{c} \frac{\partial}{\partial t} \int \int_S \mathbf{E} \cdot d\mathbf{A}. \tag{6.125}$$

Problem 6.86. *Derive from this integral equation its differential analog:*

$$\operatorname{curl} \mathbf{H} = \frac{4\pi}{c} \mathbf{j} + \frac{1}{c} \frac{\partial \mathbf{E}}{\partial t}. \tag{6.126}$$

Problem 6.87. *Show that Eq. (6.124) does not contradict Eq. (6.122).* [*Hint*: take divergence of both sides.]

Problem 6.88. *Show that Maxwell equations do not contradict the full continuity equation (6.90).* [*Hint*: apply divergence to both sides of Eq. (6.126).]

Equations (6.116), (6.122), (6.124) and (6.126) form the complete set of equations of classical electrodynamics. Introduction of the displacement current (the last term in the right-hand side of (6.126)) was necessary at least for two reasons. Firstly, without that term the continuity equation would not be satisfied. Secondly, this term is necessary for establishing electro-magnetic waves. Discussing now this latter point, let us consider Maxwell's equations in vacuum where there are no charges and currents, i.e. $\rho = 0$ and $\mathbf{j} = 0$. Then, take a partial time derivative of both sides of Eq. (6.126),

$$\operatorname{curl} \frac{\partial \mathbf{H}}{\partial t} = \frac{1}{c} \frac{\partial^2 \mathbf{E}}{\partial t^2}.$$

Then, we replace $\partial \mathbf{H} / \partial t$ using the other Maxwell's equation (6.124),

$$-c\,(\operatorname{curl} \operatorname{curl} \mathbf{E}) = \frac{1}{c} \frac{\partial^2 \mathbf{E}}{\partial t^2},$$

calculate the curl of the curl using Eq. (6.102),

$$\operatorname{curl} \operatorname{curl} \mathbf{E} = \operatorname{grad} \operatorname{div} \mathbf{E} - \Delta \mathbf{E},$$

in which only the second term remains in the right-hand side due to Eq. (6.116) (recall that there are no charges). This finally gives:

$$\frac{1}{c^2}\frac{\partial^2 \mathbf{E}}{\partial t^2} = \Delta \mathbf{E}. \tag{6.127}$$

This is a *wave equation* for the three components of the electric field **E**.

Problem 6.89. *Starting from Eq. (6.124), derive similarly the wave equation for the magnetic field:*

$$\frac{1}{c^2}\frac{\partial^2 \mathbf{H}}{\partial t^2} = \Delta \mathbf{H}. \tag{6.128}$$

Equations (6.127) and (6.128) describe oscillations of **E** and **H** which propagate in space with the speed c.

Another interesting point worth mentioning is the symmetry of the equations with respect to the two fields, **E** and **H** (compare Eq. (6.116) with (6.122), and also (6.124) with (6.126)); the only difference (apart from the sign in some places) comes from the charge density ρ and the current density $\mathbf{j}$. The absence of their magnetic analogues in Eqs. (6.122) and (6.124) corresponds to the experimental fact that magnetic charges do not exist, and hence there is also no magnetic current.

In practice, instead of the six quantities (three components of **E** and three of **H**) it is more convenient to work with only four fields: one scalar and one vector. These are introduced in the following way. Since **H** is solenoidal, Eq. (6.122), there exists a vector potential $\mathbf{A}(\mathbf{r}, t)$ such that $\mathbf{H} = \operatorname{curl} \mathbf{A}$. Replacing **H** in Eq. (6.124), we obtain:

$$\operatorname{curl}\left(\frac{1}{c}\frac{\partial \mathbf{A}}{\partial t} + \mathbf{E}\right) = 0,$$

i.e. the field inside the brackets is conservative. This in turn means that there exists a scalar field (potential) $-\phi\,(\mathbf{r}, t)$ (the minus sign is chosen due to historical reasons) such that

$$\frac{1}{c}\frac{\partial \mathbf{A}}{\partial t} + \mathbf{E} = -\operatorname{grad}\phi \quad \Longrightarrow \quad \mathbf{E} = -\operatorname{grad}\phi - \frac{1}{c}\frac{\partial \mathbf{A}}{\partial t}. \tag{6.129}$$

Problem 6.90. *Show now that replacing* **A** *and* ϕ *with*

$$\mathbf{A}' = \mathbf{A} + \operatorname{grad} U \quad \textit{and} \quad \phi' = \phi - \frac{1}{c}\frac{\partial U}{\partial t}, \tag{6.130}$$

respectively, does not change the fields **E** *and* **H**.

This shows that the choice of the scalar and vector potentials is not unique.

6.7.2 Diffusion and Heat Transport Equations

Consider diffusion of some particles of mass m in a media. This could be the case of solute molecules dissolved in a solvent (a solution), or some interstitial foreign atoms in a crystal lattice. If the concentration of solute molecules or foreign atoms is not homogeneous across the system, there will be a diffusion flux in the direction from higher to lower concentration until equilibrium is established (if there are no external fields, e.g. no electrostatic potential, then equilibrium would correspond to the homogeneous distribution of the particles). The flow of particles per unit time through a surface area $d\mathbf{A} = \mathbf{n}dA$ with the normal $\mathbf{n}$ is proportional to the directional derivative of the concentration $C(\mathbf{r}, t)$ of the particles, $\partial C/\partial n$, in the direction normal to the surface dA. That is, the total mass passed through dA is

$$dm = D\frac{\partial c}{\partial n}dAdt = D\left(\nabla C \cdot \mathbf{n}\right)dAdt = D\left(\nabla C \cdot d\mathbf{A}\right)dt, \tag{6.131}$$

where we have made use of the fact that the directional derivative can be related to the gradient, see Eq. (5.70). Above, D is the diffusion constant. The expression above is a well-known Fick's first law. The total mass passed through a closed surface S is then

$$dM = -dt \oiint_S D\nabla C \cdot d\mathbf{A},$$

and we used the minus sign to indicate that the diffusion out of the volume V enclosed by the surface S results in the *reduction* of mass (i.e. the direction *into* the volume is considered as positive).

On the other hand, the flow of the particles out of V results in the decrease of the total mass $M(t) = \int\int\int_V CdV$ in the volume:

$$dM = -\left[M\left(t+dt\right) - M\left(t\right)\right] = -\frac{\partial M}{\partial t}dt = -dt\int\int\int_V \frac{\partial C}{\partial t}dV.$$

The two expressions for the dM must be equal to each other. Using the divergence theorem for the surface integral, we then obtain the diffusion equation:

$$\operatorname{div}\left(D \operatorname{grad} C\right) = \frac{\partial C}{\partial t}. \tag{6.132}$$

This equation describes the time and space dependence of the concentration of the particles from the initial non-equilibrium situation towards equilibrium. If the diffusion coefficient is constant throughout the entire system, then it can be taken out of the divergence, and we obtain in this case

$$\Delta C = \frac{1}{D}\frac{\partial C}{\partial t}, \tag{6.133}$$

as div grad $C = \Delta C$, where Δ is the Laplace operator.

Problem 6.91. *The reader may have noticed that our derivation of the diffusion equation was very similar to that of the continuity equation in Sect. 6.5.1. Derive Eq. (6.132) directly from the continuity equation (6.85) or (6.90), extracting the needed flux* $\mathbf{j}$ *of mass from the Fick's law, Eq. (6.131).*

Problem 6.92. *Consider one-dimensional (along x) diffusion of particles, e.g., in a homogeneous solution (D is constant). Check that the function*

$$C(x,t) = \frac{1}{\sqrt{4\pi D t}} e^{-x^2/4Dt} \tag{6.134}$$

is a solution of the corresponding diffusion equation. Sketch this distribution as a function of x at different times and analyse your results. This solution corresponds to the spread of particles distribution with time when initially (at $t = 0$) the particles were all at $x = 0$.

The heat transport in a media can be considered along the same lines. This time, heat Q is transferred from regions of higher temperature $T(\mathbf{r}, t)$ to the regions of smaller temperature. We start by stating an experimental fact (Fourier's law) about the heat conduction: the amount of heat passing through a surface dA per unit time is proportional to the gradient of temperature there:

$$\frac{dQ}{dt} = \kappa \frac{\partial T}{\partial n} dA,$$

where κ is the thermal conductivity, $\partial T/\partial n = \mathbf{n} \cdot \operatorname{grad} T$ is the directional derivative showing the change of temperature in the direction perpendicular to the normal $\mathbf{n}$ to the surface dA.

Consider now a finite and arbitrary volume V within our system, with the surface S. If the energy flow goes across the boundary surface S outside in the direction along its normal $\mathbf{n}$ (directed outwards), the temperature inside the volume is reduced and hence obviously $\partial T/\partial n < 0$. We then define the amount of heat given away by the volume V to the environment around it,

$$dQ = -dt \oint\!\!\!\oint_S \kappa \frac{\partial T}{\partial n} dA = -dt \oint\!\!\!\oint_S \kappa \operatorname{grad} T \cdot d\mathbf{A} = -dt \int\!\!\int\!\!\int_V \operatorname{div}(\kappa \operatorname{grad} T)\, dV,$$

as *positive*. This is how we define dQ, and for this the minus sign is necessary; the divergence theorem was used in the last passage. Note that κ may depend on the spatial position in general and hence is kept inside the divergence.

On the other hand, the heat dQ is the one lost by the volume over time dt due to decrease of its temperature. If we take a small volume dV inside V, then this volume lost

$$-C_V\,(dT)\,\rho dV = -C_V \frac{\partial T}{\partial t}\rho dV dt$$

of heat (again, this quantity is positive as the time derivative of the temperature is negative). Here, C_V is the heat capacity per unit mass and ρ is the mass density. The total loss in the volume is then

$$dQ = -dt \int\int\int_V C_V \frac{\partial T}{\partial t}\rho dV.$$

The two expressions for dQ must be equal, which results in the heat transport equation:

$$C_V \rho \frac{\partial T}{\partial t} = \operatorname{div}\left(\kappa \operatorname{grad} T\right). \tag{6.135}$$

This heat transport equation, given the appropriate boundary and initial conditions, should provide us with the temperature distribution in the system $T(\mathbf{r}, t)$ over time. For a homogeneous system, the constant κ does not depend on the spatial variables and can be taken out of the divergence. Since divergence of the gradient is the Laplacian, we arrive at the more familiar form of the heat transport equation:

$$\frac{1}{D}\frac{\partial T}{\partial t} = \Delta T, \tag{6.136}$$

where $D = \kappa/\left(C_V \rho\right)$ is the thermal diffusivity. Comparing the two equations (6.132) and (6.135), we see that these are practically identical. Therefore, from the mathematical point of view, the solution of these equation would be the same.

6.7.3 Hydrodynamic Equations of Ideal Liquid (Gas)

Let us consider an *ideal liquid* (it could also be a gas), i.e. we assume that the forces applied to its any finite volume V with the boundary surface S, can be expressed via pressure P due to the external (with respect to the chosen volume) part of the liquid.

The latter exerts a force acting inside the volume, i.e. in the direction opposite to the surface normal $\mathbf{n}$ which is assumed to be directed out of the volume. In other words, the total force acting on the volume V is

$$-\oiint_S P\mathbf{n}dA = -\int\int\int_V \text{grad}P\, dV, \tag{6.137}$$

where we have used formula (6.83).

We can also have some *external* force density $\mathbf{F}$ acting on the liquid (i.e. there is the force $\mathbf{F}dm = \mathbf{F}\rho dV$ acting on the volume element dV of mass dm), so that the total force acting on the volume V will then be

$$\mathbf{f} = \int\int\int_V \mathbf{F}\rho dV - \int\int\int_V \text{grad}P\, dV = \int\int\int_V (\rho\mathbf{F} - \text{grad}P)\, dV.$$

According to Newton's second law, this force results in the acceleration of the liquid volume given by

$$\int\int\int_V \rho\frac{d\mathbf{v}}{dt}dV,$$

where $\mathbf{v}(\mathbf{r}, t)$ is the velocity of the liquid depending on time t and the spatial point $\mathbf{r}$. Therefore, one can write:

$$\int\int\int_V \rho\frac{d\mathbf{v}}{dt}dV = \int\int\int_V (\rho\mathbf{F} - \text{grad}P)\, dV \quad \Longrightarrow \quad \rho\frac{d\mathbf{v}}{dt} = \rho\mathbf{F} - \text{grad}P. \tag{6.138}$$

Note that here the acceleration in the left-hand side is given by the *total derivative* of the velocity. Indeed, $\mathbf{v}$ is the function of both time and the three coordinates, and the liquid particles move in space, so that their coordinates change as well. Therefore, similarly to our consideration of the derivative of the distribution function of a gas in Sect. 5.6 (see Eq. (5.26)), we can write:

$$\frac{d\mathbf{v}}{dt} \quad \longmapsto \quad \left(\frac{d\mathbf{v}}{dt}\right)_{tot} = \frac{\partial\mathbf{v}}{\partial t} + \frac{\partial\mathbf{v}}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial\mathbf{v}}{\partial y}\frac{\partial y}{\partial t} + \frac{\partial\mathbf{v}}{\partial z}\frac{\partial z}{\partial t} = \frac{\partial\mathbf{v}}{\partial t} + \frac{\partial\mathbf{v}}{\partial x}v_x + \frac{\partial\mathbf{v}}{\partial y}v_y + \frac{\partial\mathbf{v}}{\partial z}v_z.$$

The coordinate derivatives term above is normally written in the following short form as a dot product:

$$\frac{\partial\mathbf{v}}{\partial x}v_x + \frac{\partial\mathbf{v}}{\partial y}v_y + \frac{\partial\mathbf{v}}{\partial z}v_z = \mathbf{v} \cdot \text{grad}\, \mathbf{v},$$

where the gradient is understood to be taken separately for each component of the velocity field. Finally, we arrive at the following equations:

$$\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \operatorname{grad} \mathbf{v} = \mathbf{F} - \frac{1}{\rho} \operatorname{grad} P. \tag{6.139}$$

This equation is to be supplemented with the continuity equation (6.85) or (6.88) for the liquid which we derived before. Given the equation of states for the liquid, $P = f(\rho, T)$, i.e. how the pressure depends on the density and the temperature T, one can solve these equations to obtain the velocity field in the liquid under the applied external forces $\mathbf{F}$ and the temperature T.

Chapter 7
Infinite Numerical and Functional Series

In Sect. 1.9 we considered summation of finite series. In practice it is often necessary to deal with *infinite* series which contain an infinite number of terms a_n. Here a_n is a general (we say "the n-th") term of the series which is constructed via a some kind of formula depending on an integer $n = 1, 2, 3, \ldots$. For instance, the rule $a_n = a_0 q^n$ with $n = 0, 1, 2, 3, \ldots$ corresponds to an infinite geometric progression

$$a_0 + a_0 q + a_0 q^2 + \cdots = a_0 \left(1 + q + q^2 + \cdots\right).$$

Note that infinite numerical series were already mentioned briefly in Sect. 2.2.3. Here we shall consider this question in more detail. Our interest here is in understanding whether or not one can define a *sum* of such an infinite series,

$$S = \sum_{n=1}^{\infty} a_n, \tag{7.1}$$

i.e. whether the sum like this one converges to a well defined number S. It is also possible that each term of the series, $a_n(x)$, is a function of some variable x, and then we should discuss for which values of the x the series converges, and if it does, what are the properties of the function $S(x)$ of the sum. For instance, one may ask if it is possible to integrate or differentiate the series term-by-term, i.e. would the series generated that way converge for the same values of the x and, if it does, whether the result can simply be related to integrating or differentiating, respectively, the function $S(x)$ of the sum itself.

L. Kantorovich, *Mathematics for Natural Scientists*, Undergraduate Lecture Notes in Physics, DOI 10.1007/978-1-4939-2785-2_7

7.1 Infinite Numerical Series

We say that the series (7.1) converges if the limit of its partial sum containing N terms,

$$S_N = \sum_{n=1}^{N} a_n, \tag{7.2}$$

has a well-defined limit. The partial sums S_1, S_2, S_3, etc. form a numerical sequence which we considered in Sect. 2.2. Therefore, the numerical series (7.1) converges if the corresponding sequence of its partial sums, S_N, has a well-defined limit at $N \to \infty$.

An infinite geometrical progression, when $a_n = a_0 q^n$, has already been considered in Sect. 2.2.3, and can serve as a simple example. It converges when $|q| < 1$, and diverges otherwise. The series (1.70) converges to one since

$$\lim_{N\to\infty} S_N = \lim_{N\to\infty} \sum_{n=1}^{N} \frac{1}{n(n+1)} = \lim_{N\to\infty} \left(1 - \frac{1}{N+1}\right) = 1.$$

As an example of a diverging series, let us consider the so-called *harmonic* series

$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots = \sum_{n=1}^{\infty} \frac{1}{n}. \tag{7.3}$$

If S_N is its partial sum, then for any N:

$$S_{2N} - S_N = \sum_{n=1}^{2N} \frac{1}{n} - \sum_{n=1}^{N} \frac{1}{n} = \sum_{n=N+1}^{2N} \frac{1}{n} > \sum_{n=N+1}^{2N} \frac{1}{2N} = \frac{N}{2N} = \frac{1}{2},$$

since for any $n \leq 2N$ we have $1/n \geq 1/2N$ (the equal sign only when $n = 2N$). On the other hand, if we assume that the series converges to S, then $\lim_{N\to\infty} (S_{2N} - S_N) = S - S = 0$, which is impossible as it has been shown above that $S_{2N} - S_N > 1/2$ and hence cannot have the zero limit. This proves that the harmonic series actually diverges.

Problem 7.1. *Find the sum of the following infinite series:*

$$\sum_{n=1}^{\infty} \frac{1}{n^2 - 1}.$$

[*Hint*: the partial sum is given by Eq. (1.68). *Answer*: 3/4.]

Problem 7.2. *Prove that a combination of a geometric and arithmetic progressions, see Eq. (1.69), has a well-defined limit for* $|q| < 1$ *and any r:*

$$\sum_{i=0}^{\infty}(a_0 + ir)q^i = \frac{a_0}{1-q} + \frac{rq}{(1-q)^2}.$$

There are several simple theorems which establish essential properties of converging infinite numerical series.

Theorem 7.1. *If a series (7.1) converges to S, one can remove any first m terms in the series, and the rest of the series,*

$$\sum_{n=m+1}^{\infty} a_n,$$

would still converge.

Theorem 7.2. *If the series (7.1) of* a_n *converges to S, then the series of* $b_n = ca_n$ *converges to cS, where c is (generally complex) number.*

Theorem 7.3. *If two series* $\sum_n a_n$ *and* $\sum_n b_n$ *converge to* S_a *and* S_b*, respectively, then the series* $\sum_n (\alpha a_n + \beta b_n)$ *also converges to* $\alpha S_a + \beta S_b$*.*

Problem 7.3. Prove the above theorems. [*Hint*: consider the limits of the appropriate partial sums.]

Theorem 7.4 (Necessary Condition for Convergence). *If the series* $\sum_n a_n$ *converges, then* $a_n \to 0$ *as* $n \to \infty$*.*

Proof. if S_N is a partial sum of the series, then $a_N = S_N - S_{N-1}$. Taking the limit $N \to \infty$, we obtain the required result. **Q.E.D.**

This is not a sufficient condition, only a necessary one. Indeed, in the case of the harmonic series considered above, $a_n = 1/n$ and it does tend to zero when $n \to \infty$.

However, we know that the series diverges. Therefore, to establish convergence of a particular series encountered in practical problems, it is necessary to investigate the question of convergence in more detail; in particular, we need to develop sufficiency criteria as well.

7.1.1 Series with Positive Terms

It is convenient to start a more detailed analysis from a particular case of series which have all their terms positive, $a_n > 0$.

Theorem 7.5 (Necessary and Sufficient Condition for Convergence). *The series converges if and only if its numerical sequence of the partial sums $\{S_N\}$ is bounded from above.*

Proof. To prove sufficiency, we assume that the sequence of partial sums is bounded from above, i.e. $S_N \leq M$, where $M > 0$, and then we need to see if from this follows that the sequence converges. Indeed, because the sequence contains exclusively positive terms, the partial sums form an increasing sequence $0 < S_1 \leq S_2 \leq S_3 \leq \cdots$. It has to reach a well-defined limit since, according to our assumption, it is bounded from above (and hence cannot increase indefinitely). Therefore, once S_N has a limit, the numerical sequence converges to that limit.

Now, to prove the necessity, we first assume that the series converges to S, and then show that from this follows that it is bounded from above. Indeed, since the sequence of partial sums $\{S_N\}$ converges, it has a well-defined limit S. On the other hand, since the numerical sequence of the partial sums is increasing, it must be also bounded from above by S (otherwise, it would never converge, see Sect. 2.2.2). **Q.E.D.**

There are several sufficiency criteria to check whether the series converges or not.

Theorem 7.6 (Sufficient Criterion for Convergence). *Consider two series, both with all terms positive, $\sum_n a_n$ and $\sum_n b_n$, such that $a_n \leq b_n$ for any $n = 1, 2, 3, \ldots$. Then, if the second series $\sum_n b_n$ converges, then the first one, $\sum_n a_n$, does as well; if the first series diverges, so does the second.*

Proof. Let A_N and B_N be the corresponding partial sums for the two series, respectively; since for all n we have $a_n \leq b_n$, then obviously $A_N \leq B_N$ for any N. Let us now consider the first part of the theorem based on the assumption that the series $\sum_n b_n$ converges to some value B. Hence, $B_N \leq B$. Therefore, $A_N \leq B_N \leq B$, i.e. the series $\sum_n a_n$ is bounded from above and hence (according to Theorem 7.5) also converges. Now we turn to the second statement where the first series diverges. Let us assume that the second series converges. Then, according to the first part of the theorem, the first series should converge as well, which contradicts our assumption; therefore, if the first series diverges, then so does the second. **Q.E.D.**

As an example, consider the series with $a_n = 1/\sqrt{n}$. Since $1/n < 1/\sqrt{n}$ for any $n > 1$, and the series with $b_n = 1/n$ diverges, then the series $\sum_n n^{-1/2}$ diverges as well. In fact, the series $\sum_n n^{-\alpha}$ diverges for any $0 < \alpha \leq 1$.

Problem 7.4. *Prove that the series with* $a_n = \left(1 + 1/n^2\right)^n$ *converges.*

Problem 7.5. *Prove that the series with* $a_n = n^{-n}$ *converges.* [*Hint*: use the fact that $n^{-n} \leq 2^{-n}$ for any $n > 1$.]

Theorem 7.7 (The Ratio Test[1]). *If*

$$\lambda = \lim_{n\to\infty} \frac{a_{n+1}}{a_n} < 1,$$

the series converges; if $\lambda > 1$, it diverges; if $\lambda = 1$, then this test in inconclusive (the series may either converge or diverge, more investigations are required).

Proof. Let us first consider the case of $\lambda < 1$. Since the limit of a_{n+1}/a_n exists (and equal to λ), then for any $\epsilon > 0$ there exists an integer N such that for any $n > N$ we have $|a_{n+1}/a_n - \lambda| < \epsilon$, i.e. $\lambda - \epsilon < a_{n+1}/a_n < \lambda + \epsilon$. What is essential for us here is the second part: $a_{n+1}/a_n < \lambda + \epsilon$. Indeed, since $\lambda < 1$, one can always find ϵ such that $\lambda + \epsilon$ is still smaller than 1, i.e. for all $n > N$ we would have $a_{n+1}/a_n < \rho < 1$, or $a_{n+1} < \rho a_n$. Therefore, $a_{N+1} < \rho a_N$, $a_{N+2} < \rho a_{N+1} < \rho^2 a_N$, etc. In general $a_{N+r} < \rho^r a_N$, and hence,

$$\sum_{n=N+1}^{\infty} a_n = \sum_{r=1}^{\infty} a_{N+r} < \sum_{r=1}^{\infty} \rho^r a_N = a_N \sum_{r=1}^{\infty} \rho^r = \frac{a_N \rho}{1-\rho},$$

where we have made use of the sum of the geometrical progression (recall that $0 < \rho < 1$). Therefore, the series without the first N terms converges, and so does the whole series (Theorem 7.1).

[1] This test is due to Jean-Baptiste le Rond d'Alembert and normally bears his name.

Now consider the case of $\lambda > 1$. We can again write that $\lambda - \epsilon < a_{n+1}/a_n < \lambda + \epsilon$, or $\lambda - \epsilon < a_{n+1}/a_n$. Since $\lambda > 1$, one can always find such $\epsilon > 0$ that $\lambda - \epsilon > 1$. Therefore, $a_{n+1}/a_n > 1$ or $a_{n+1} > a_n$. This means that a_n cannot tend to zero as $n \to \infty$, which is the necessary criterion for convergence (Theorem 7.4), i.e. the original series indeed diverges. **Q.E.D.**

Nothing can be said about the case of $\lambda = 1$, the corresponding series may either diverge or converge, more powerful criteria must be used to establish convergence. For instance, consider the series with $a_n = 1/\sqrt{n}$ which, as we know from the analysis above, diverges. Using the ratio test, we have:

$$\lambda = \lim_{n\to\infty} \frac{a_{n+1}}{a_n} = \lim_{n\to\infty} \frac{1/\sqrt{n+1}}{1/\sqrt{n}} = \lim_{n\to\infty} \sqrt{\frac{n}{n+1}} = 1.$$

Therefore, the ratio test is not powerful enough to establish convergence or divergence of this particular series.

Problem 7.6. *Prove that $\lambda = 1$ as well for the series with $a_n = 1/n^2$ which (as we shall learn later on) converges. Therefore, the case of $\lambda = 1$ is also inconclusive in this case.*

Problem 7.7. *Prove using the ratio test that the series with $a_n = x^n/n!$ converges for all values of x (we shall see later on that this series gives e^x).*

Problem 7.8. *Prove using the ratio test that the series with $a_n = n^n/n!$ diverges.* [*Hint*: you will need the result (2.24).]

Theorem 7.8 (The Root Test[2]). *If a_n, starting from some $n \geq N \geq 1$, satisfies the inequality*

$$\sqrt[n]{a_n} \leq q < 1, \tag{7.4}$$

then the series converges; if, however, for all $n \geq M \geq 1$ we have $\sqrt[n]{a_n} > 1$, then the series diverges.

Proof. In the first case, $a_n \leq q^n$ with $q < 1$ for any $n \geq N$. The series with $b_n = q^n$ is a geometric progression with the q^N being its first term, and it converges

[2] Due to Augustin-Louis Cauchy.

for $q < 1$. Therefore, the series with a_n and $n \geq N$ converges as well because of Theorem 7.6. As a finite number of terms before the a_N term do not effect the convergence (Theorem 7.1), the whole series converges. In the second case, $a_n > 1$ and hence a_n does not tend to zero as $n \to \infty$ yielding divergence of the series (Theorem 7.4). **Q.E.D.**

The root test is stronger than the ratio test, i.e. in some cases when the latter test is inconclusive or cannot be applied, the root test may work. However, it may not be straightforward to apply the root test, and hence it is used less frequently than the ratio test. Both tests are inconclusive when $\sqrt[n]{a_n}$ and a_{n+1}/a_n tend to one when $n \to \infty$.

Problem 7.9. *Prove that the series with $a_n = e^{-\alpha n} \sin^2(\beta n)$ converges for any β and $\alpha > 0$. Try both the ratio and root tests. You will find that although the root test immediately gives the expected answer, application of the ratio test is inconclusive as the corresponding limit does not exist.*

Theorem 7.9 (Integral Test, also due to Cauchy). *Consider a series with positive non-increasing terms, i.e.*

$$a_1 \geq a_2 \geq a_3 \geq \cdots \geq a_n \geq a_{n+1} \geq \cdots > 0.$$

Next let us define a function $y = f(x)$ such that at integer values of the x it would be equal exactly to the terms of the series, i.e. $f(n) = a_n$ for any $n = 1, 2, 3 \ldots$. One can always define such a function, and it is clear that this function would be non-increasing. Then the convergence of the series can be concluded depending on whether the integral

$$A = \int_1^\infty f(x)dx \tag{7.5}$$

converges or not.

Proof. So, consider the partial sum of our series[3]:

$$S_{N-1} = f(1) + f(2) + f(3) + \cdots + f(N-1).$$

[3]Note that some ideas of the proof are similar to those we encountered when considering Darboux sums in Sect. 4.2 (specifically, see Fig. 4.2).

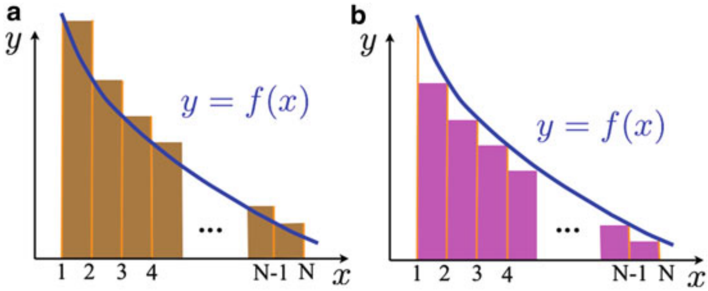


Fig. 7.1 To the proof of the integral convergence test: (**a**) the integral under the curve between 1 and infinity is smaller than the sum of functions at points 1, 2, 3, etc., while in (**b**) the integral is larger than the sum of functions at points 2, 3, 4, etc.

It can be thought of as a sum of $N-1$ rectangular areas shown in Fig. 7.1(a) since the corresponding width along the x axis of each rectangle is exactly equal to one. We also see from the figure that S_{N-1} will definitely be larger than the area A_N under the curve of the function $f(x)$ between the points 1 and N, i.e. $S_{N-1} > A_N$ (we can call A_N a partial integral). On the other hand, let us now consider the sum

$$S_N - f(1) = f(2) + f(3) + f(4) + \cdots + f(N).$$

This sum can also be interpreted as a total area of the coloured rectangles in Fig. 7.1(b), but this time it is obviously smaller than the area under the integral, i.e. $S_N - f(1) = S_N - a_1 < A_N$. Hence, we can write:

$$S_N - a_1 < \int_1^N f(x)dx < S_{N-1} = S_N - a_N, \tag{7.6}$$

where S_N is the partial sum of N terms of our series. If the integral (7.5) converges, then $A_N < A$ (since the function $f(x) > 0$, the partial integral is an increasing function, i.e. $A_{N+1} > A_N$, and A_N should be bounded from above[4]) and hence $S_N - a_1 < A_N < A$, i.e. $S_N < A + a_1$. This means that the partial sum is bounded from above, and then, according to Theorem 7.5, converges.

Now, let us assume that the integral (7.5) is equal to infinity (diverges). From the second part of the inequality (7.6) it then follows that $S_N > A_N + a_N$. However, the integral diverges and hence its partial sum A_N from some N can be made larger than any positive number. Therefore, S_N as well can be made arbitrarily big, i.e. it does not have a limit. The series diverges. **Q.E.D.**

As an example, let us consider the series with $a_n = 1/n^\alpha = n^{-\alpha}$, where $\alpha > 0$. This series was mentioned above in Problem 7.6 for $\alpha = 2$. Here $f(x) = x^{-\alpha}$ and therefore the convergence of our series depends on whether the integral $\int_1^\infty x^{-\alpha}dx$

[4]This follows from the general fact, see Sect. 4.5.3, that improper integrals are understood as limits.

converges or diverges. This latter question was answered in Sect. 4.5.3: the integral converges only for $\alpha > 1$ and diverges otherwise. Therefore, the series converges for $\alpha > 1$ and diverges otherwise. In particular, the harmonic series (7.3) corresponding to $\alpha = 1$ diverges, but the series

$$1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots \tag{7.7}$$

converges ($\alpha = 2$).

Problem 7.10. *Show that the root test is inconclusive for the series of inverse squares (7.7).* [*Hint*: in applying the root test, convert $1/n^2$ into an exponential and then consider the limit of the exponent as $n \to \infty$.]

Problem 7.11. *Using the integral test, prove that the series with $a_n = (n \ln n)^{-1}$ diverges.*

Problem 7.12. *Using the integral test, prove that the series with $a_n = ne^{-n}$ converges.*

7.1.2 Euler–Mascheroni Constant

We mentioned above that the harmonic series (7.3) diverges. Interestingly, the closely related numerical sequence

$$u_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} - \ln n \tag{7.8}$$

has a certain limit which is called Euler–Mascheroni constant:

$$\gamma = \lim_{n\to\infty} \left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} - \ln n\right) = 0.5772\ldots. \tag{7.9}$$

To show that the sequence u_1, u_2, etc. converges, let us consider the function $y = 1/x$ for $1 \le x \le n$. At integer values $x_k = k$, where $k = 1, \ldots, n$, the function $y(x_k) = 1/k$. Let us calculate the area under the curve between $x = 1$ and $x = n$ using two methods, similarly to what we did in the previous section. Indeed, the exact area under the curve is obviously

$$S_n = \int_1^n y(x)dx = \int_1^n \frac{dx}{x} = \ln n.$$

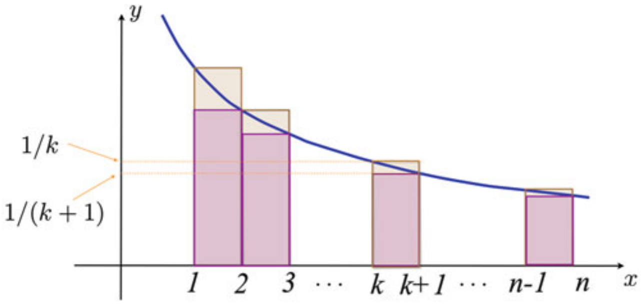


Fig. 7.2 The area under the curve $y = 1/x$ can be approximated either by rectangles which go above or below the curve (cf. Fig. 7.1)

By choosing rectangles which go above the curve, see Fig. 7.2, and summing up their areas (note that the base of each rectangular is equal exactly to one), we shall construct an approximation,

$$S_n^+ = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n-1},$$

which is larger than the actual area by

$$v_n = S_n^+ - S_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n-1} - \ln n \ = u_n - \frac{1}{n}.$$

Obviously, v_n is positive and forms an increasing sequence as positive extra areas are always added due to new rectangulars when n is increased.

On the other hand, the sequence $\{v_n\}$ is limited from above. Indeed, by choosing the lower rectangulars, we can construct an approximation to the area

$$S_n^- = \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} < S_n = \ln n \quad \Longrightarrow \quad v_n < 1 - \frac{1}{n} < 1.$$

Therefore, according to Theorem 7.5 the sequence v_n has a limit which is the required Euler–Mascheroni constant. Its numerical value is given above.

7.1.3 Alternating Series

So far, we have considered series with all terms positive; this consideration can of course be extended for the series with all terms negative. Before we touch upon general series where the sign of terms may change in a general way, it is instructive to consider another special case of numerical series in which the sign of terms alternates, i.e.

$$a_1 + a_2 + a_3 + \cdots + a_n + \cdots = |a_1| - |a_2| + |a_3| - \cdots \pm |a_n| \mp \cdots , \qquad (7.10)$$

i.e. $a_n = (-1)^{n+1} |a_n|$ (assuming, without loss of generality, that $a_1 > 0$).

For this type of series there exists a very simple sufficient convergence test which we formulate as the following theorem:

Theorem 7.10 (Due to Leibnitz). *For the series (7.10) to converge, it is sufficient that its terms do not increase in their absolute values, i.e.*

$$|a_1| \geq |a_2| \geq |a_3| \geq \cdots \geq |a_n| \geq \cdots > 0,$$

and also $|a_n| \to 0$ *when* $n \to \infty$.

Proof. Assume that $a_1 > 0$ and hence all odd terms in the series are positive; correspondingly, all even terms are negative. Then, consider a partial sum with an even number of terms:

$$S_{2N} = a_1 + a_2 + \cdots + a_{2N} = (a_1 - |a_2|) + (a_3 - |a_4|) + \cdots + (a_{2N-1} - |a_{2N}|) . \qquad (7.11)$$

Since terms in the series do not increase in their absolute value, each difference within the round parentheses is positive, and hence the partial sums S_{2N} form an increasing numerical sequence. On the other hand, one can also write S_{2N} as follows:

$$\begin{aligned} S_{2N} &= a_1 - (|a_2| - a_3) - (|a_4| - a_5) - \cdots - (|a_{2N-2}| - a_{2N-1}) - |a_{2N}| \\ &= a_1 - [(|a_2| - a_3) + (|a_4| - a_5) + \cdots + (|a_{2N-2}| - a_{2N-1}) + |a_{2N}|] < a_1, \end{aligned}$$

since the expression in the square brackets is definitely positive. Hence, S_{2N} is bounded from above by its first element, and, therefore, the numerical sequence of S_{2N} has a limit.

We still need to consider the partial sum of an odd number of terms:

$$S_{2N+1} = S_{2N} + a_{2N+1} \quad \Longrightarrow \quad \lim_{N\to\infty} S_{2N+1} = \lim_{N\to\infty} S_{2N} + \lim_{N\to\infty} a_{2N+1} = \lim_{N\to\infty} S_{2N},$$

since the terms in the series tend to zero as their number goes to infinity according to the second condition of the Theorem. Since both partial sums have their limits and these coincide, this must be the limit of the whole numerical series. **Q.E.D.**

It also follows from the above proof that the overall sign of the sum of the series $S = \lim_{N\to\infty} S_N$ is entirely determined by the sign of its first term, a_1. Indeed, assuming that $a_1 > 0$, we have from Eq. (7.11) that $S_{2N} > 0$ as the sum of positive terms. Since the sequences S_{2N} and S_{2N+1} converge to the same limit (if that limit exists), this proves the statement that was made. If $a_1 < 0$, then we similarly have that the sum $S < 0$.

Theorem 7.11. *Consider again the alternating series from the previous theorem. The absolute value of the sum of all dropped terms in the series does not exceed the first dropped term.*

Proof. Let us consider the first N terms of the series. The sum of all dropped terms is

$$S_{>N} = a_{N+1} + a_{N+2} + \cdots = |a_{N+1}| - (|a_{N+2}| - |a_{N+3}|) \\ - (|a_{N+4}| - |a_{N+5}|) - \cdots < |a_{N+1}| = a_{N+1},$$

where we have assumed that the first dropped term, a_{N+1}, is positive. The inequality holds since all expressions in the round brackets are positive. Since the first term, a_{N+1}, is positive, then $S_{>N} > 0$ as well, i.e. we obtain $0 < S_{>N} < a_{N+1}$.

Similarly, if the first dropped term is negative, we similarly write:

$$S_{>N} = a_{N+1} + a_{N+2} + \cdots = -|a_{N+1}| + (|a_{N+2}| - |a_{N+3}|) \\ + (|a_{N+4}| - |a_{N+5}|) + \cdots > -|a_{N+1}| = a_{N+1},$$

i.e. $S_{>N} < 0$ and, at the same time, $S_{>N} > a_{N+1} = -|a_{N+1}|$, i.e. $a_{N+1} < S_{>N} < 0$.

Combining both cases, we can write that $|S_{>N}| < |a_{N+1}|$, as required. **Q.E.D.**

As an example, we shall consider the series

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n}. \tag{7.12}$$

This series looks like the harmonic series (7.3), however, the signs of the terms in the series above alternate (it is sometimes called the alternating harmonic series). We recall that the harmonic series diverges. However, the series (7.12) actually converges as it fully satisfies the conditions of Theorem 7.10: $1/n > 1/(n+1)$ and $(-1)^{n+1}/n \to 0$ when $n \to \infty$. As will be shown in Sect. 7.3.3, the series (7.12) corresponds to the Taylor's expansion of $\ln(1+x)$ at $x = 1$ and hence converges to $\ln 2$ (see also discussion in Sect. 7.3.2).

7.1.4 General Series: Absolute and Conditional Convergence

Here we shall consider a general infinite series in which signs of its terms neither are the same nor they are alternating:

$$a_1 + a_2 + a_3 + \cdots = \sum_{n=1}^{\infty} a_n. \tag{7.13}$$

Theorem 7.12 (Sufficient Criterion). *If the series*

$$A = |a_1| + |a_2| + |a_3| + \cdots = \sum_{n=1}^{\infty} |a_n|, \tag{7.14}$$

constructed from absolute values of the terms of the original series, converges, then so does the original series.

Proof. Consider the first N terms of the series (7.13) and (7.14) with the corresponding partial sums being S_N and A_N. Terms of the series (7.13) may have different signs. Let us collect all positive terms into the partial sum $S'_N > 0$ and the *negative* values of all negative terms into the partial sum $S''_N > 0$. Obviously, $S_N = S'_N - S''_N$ and $A_N = S'_N + S''_N$. Positive terms form a subsequence of the whole sequence with the partial sum S'_N. Similarly, negative values of the negative terms also form a subsequence with the partial sum S''_N. Since we assume that the series (7.14) converges, i.e. A_N has a limit, then both partial sums S'_N and S''_N must be bounded from above (both contain positive terms): $S'_N < A_N < A$ and $S''_N < A_N < A$. Therefore, both partial sums have their limits: $S'_N \to S'$ and $S''_N \to S''$ (see also Theorem 2.4 on subsequences). Therefore, there is also a well-defined limit for the original series: $\lim_{N\to\infty} S_N = S' - S''$. **Q.E.D.**

Problem 7.13. *Prove that the series*

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

converges (absolutely) for any value of $\alpha > 1$.

Convergence tests considered in Sect. 7.1.1 can now be applied to general series with the only change that absolute values of the terms of the series are to be used. Indeed, according to the ratio test, the series of absolute values (7.14) converges if

$$\lim_{n\to\infty}\frac{|a_{n+1}|}{|a_n|}=\lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right|<1$$

and diverges if the limit of the ratio of absolute values is larger than one. Similarly for the root test: the series converges if $\sqrt[n]{|a_n|}<1$, and diverges if the root is larger than one. However, convergence of the series (7.14) guarantees the convergence of the original series (7.13), and this proves the statement made.

Problem 7.14. *Using the ratio test, show that the series $\sum_{n=1}^{\infty} x^n/n$ converges for $|x|<1$ and diverges for $|x|>1$.*

Problem 7.15. *Using the ratio test, show that the series with $a_n = x^n/\left(n^2+n\right)$ converges for $|x|<1$ and diverges for $|x|>1$.*

Theorem 7.12 provides only a sufficient condition for convergence. If a series converges together with the series of its absolute values, it is said that the series converges *absolutely*. However, the series may converge, but the corresponding series of absolute values of its terms may diverge. In this case it is said that the original series converges *conditionally*. For instance, the series (7.12) converges conditionally. This distinction has a very important meaning based on the following

Theorem 7.13. *If a series converges absolutely, terms in the series may be permuted arbitrarily without affecting the value of the sum.*

Proof. Consider first a series $\sum_n a_n$ of positive terms $a_n>0$ which partial sum, A_N, converges to A, i.e. $\lim_{N\to\infty}A_N=A$. Let us arbitrarily permute an *infinite* number of terms in the series.[5] The new series created in this way, $\sum_n b_n$, consists of elements $b_n>0$. We stress again that these are the same as the elements of the first series, but positioned in a different order.

Consider the first N terms in the second series, its partial sum we shall denote B_N. No matter how big N is, all first N terms b_n with $n\le N$ of the second series will be contained within the first M terms of the first series, where M is some integer. Since amongst the M terms there will most likely be some other terms as well, and

[5]Obviously, if a finite number of terms is permuted, these terms can be removed from the series without affecting its convergence.

all elements of the series are positive, one can write: $B_N \leq A_M$. But the first series converges to A, and hence $A_N \leq A$, i.e. we conclude that $B_N \leq A$, i.e. the partial sum of the second series is bounded from above and hence converges to some value $B \leq A$.

Conversely, let us consider the first N' elements of the first series; the second series would contain these elements within its first M' terms, and hence $A_{N'} \leq B_{M'}$. Since we already established that the second series converges to B, then $B_{M'} \leq B$, and hence $A_{N'} \leq B_{M'} \leq B$. This yields $A \leq B$. Consequently, $A = B$. So, we conclude that if all terms are positive,[6] then one can indeed permute terms in the series, and the new series created in this way would converge to the same sum.

It is also clear that if there are two series with non-negative elements, one can sum up or subtract them from each other term-by-term, i.e. $A \pm B = \sum_n (a_n \pm b_n)$, since this would simply correspond to a permutation of elements of the combined series $\sum_n a_n \pm \sum_n b_n = A \pm B$.

Now we consider an absolutely converging series $\sum_n a_n$ with terms a_n of arbitrary signs. We permute an infinite number of its elements arbitrarily and obtain a new series $\sum_n b_n$. Two auxiliary series, both containing positive terms, can then be constructed out of it:

$$S_1 = \sum_n \frac{1}{2} (b_n + |b_n|) \quad \text{and} \quad S_2 = \sum_n \frac{1}{2} (|b_n| - b_n) . \tag{7.15}$$

Since either of the series is built from only positive (more precisely, non-negative) elements, we can apply to them the statements made above, i.e. we can permute their elements arbitrarily without affecting their sums, and manipulate them. In particular, $S_1 - S_2 = \sum_n b_n$.

Next, we shall permute the elements of the b_n-series precisely in the opposite way to the one we have used initially when deriving the b_n-series from the a_n-series; then, the order of terms would correspond to the original a_n series. Two new series of non-negative terms can then be constructed,

$$S_1' = \sum_n \frac{1}{2} (a_n + |a_n|) \quad \text{and} \quad S_2' = \sum_n \frac{1}{2} (|a_n| - a_n) ,$$

which contain the same elements as above in Eq. (7.15), but in different order. Since the terms are non-negative, $S_1' = S_1$ and $S_2' = S_2$. Subtracting the two series term-by-term yields $S_1 - S_2 = \sum_n a_n$, which is the same as $\sum_n b_n$, **Q.E.D.**

A simple corollary from this theorem is that two absolutely converging series can be algebraically summed or subtracted from each other term-by-term. Indeed, this would simply correspond to permuting terms in the absolutely converging series:

[6] In fact, of the same sign.

$$A \pm B = \sum_n a_n \pm \sum_n b_n = \sum_n (a_n \pm b_n) .$$

Another simple corollary is that one can collect all positive and negative terms of an absolutely converging series into separate sums $A_> = \sum_n^> a_n$ and $A_< = \sum_n^< a_n = -\sum_n^< |a_n|$. Here $A_<$ includes all negative terms, while $A_>$ all positive; both sums converge independently.

One can also show that it is possible to multiply the two series term-by-term yielding a series converging to the product of the two sums, i.e.

$$\left(\sum_n a_n\right)\left(\sum_n b_n\right) = a_1b_1 + a_2b_1 + a_1b_2 + a_1b_3 + a_2b_2 + a_3b_1 + \cdots = AB, \tag{7.16}$$

as stated by the following theorem:

Theorem 7.14. *One can multiply terms of two absolutely converging series $A = \sum_n a_n$ and $B = \sum_n b_n$ term-by-term, and the resulting series would converge to the product of the two sums, AB.*

Proof. Let us start by assuming that both series consist of positive elements only. Consider a particular order in which two series are multiplied, e.g. the one in Eq. (7.16), and let us call the elements of the product series c_n, i.e. $c_n = a_kb_l$ with some indices k and l of the first and the second series, respectively. The new series converges to $C = AB$. Indeed, consider first M elements of the product series with the partial sum C_N. We can always find a number M such that all elements a_k and b_l we find in C_N are contained in the partial sums A_M and B_M. However, either of the latter sums may contain extra elements, i.e. $C_N \leq A_MB_M \leq AB$. The last inequality follows from the convergence of the two series, i.e. $A_N \leq A$ and $B_N \leq B$. Since C_N is bounded from above, it converges to $C \leq AB$.

Conversely, let us consider now two partial sums $A_{N'}$ and $B_{N'}$ with the first N' elements of the two original series, and construct the product $A_{N'}B_{N'}$ of all their elements. We can always find a number M' of terms in the product series (probably large) such that the latter contains all terms from the product $A_{N'}B_{N'}$. However, it would also contain other elements, which means that $A_{N'}B_{N'} \leq C_{M'} \leq C$. This yields $AB \leq C$. Combining the two inequalities we get $AB = C$.

Now consider two absolutely converging series $A = \sum_n a_n$ and $B = \sum_n b_n$ with general elements (i.e. of any signs), and let us construct a product series $\sum_n c_n$. According to the above proof, the sum of its absolute values converges to $|A| \cdot |B| = |AB|$, which means that the original product series converges absolutely. Therefore, we can collect all positive and negative terms in the three series into the sums $A_<$, $A_>$, $B_<$, $B_>$, $C_<$ and $C_>$. Since the product series converges absolutely, according to the previous Theorem 7.13, we are allowed to permute its terms arbitrarily without affecting the sum, i.e. the following manipulation is legitimate:

$$\sum_n c_n = \sum_k a_k \left(\sum_l b_l \right) = \sum_k a_k \left(B_> + B_< \right) = \sum_k a_k B = B \sum_n a_n = BA,$$

as required. **Q.E.D**

Therefore, we have proved that if two series converge absolutely to the values S_1 and S_2, then: (i) a series obtained by summing (or subtracting) the corresponding terms of the two series converges absolutely to $S_1 + S_2$ (or, correspondingly, $S_1 - S_2$); (ii) a series obtained by multiplication of the two series converges absolutely to $S_1 S_2$; (iii) the order of terms in a series can be changed arbitrarily without affecting its sum. In other words, absolutely converging series can be *manipulated algebraically*.

Interestingly, this is not the case for conditionally converging series. Moreover, as was proven by Riemann, by permuting terms in the series in a specific way one can obtain a series converging to any given value; the series may even become diverging. To illustrate the point of the dependence of the sum of a conditionally converging series on the order of terms in it, consider the series (7.12). Let us reorder the terms in it in such a way that after each positive term we shall put two negative ones picked up in the correct order:

$$\begin{aligned} S &= 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots = \left(1 - \frac{1}{2} - \frac{1}{4}\right) + \left(\frac{1}{3} - \frac{1}{6} - \frac{1}{8}\right) \\ &\quad + \left(\frac{1}{5} - \frac{1}{10} - \frac{1}{12}\right) + \left(\frac{1}{7} - \frac{1}{14} - \frac{1}{16}\right) + \cdots \\ &= \left(\frac{1}{2} - \frac{1}{4}\right) + \left(\frac{1}{6} - \frac{1}{8}\right) + \left(\frac{1}{10} - \frac{1}{12}\right) + \left(\frac{1}{14} - \frac{1}{16}\right) + \cdots \\ &= \frac{1}{2}\left(1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \frac{1}{7} - \frac{1}{8} + \cdots\right) = \frac{1}{2} S, \end{aligned}$$

i.e. S can be made equal to the half of itself. Therefore, conditionally converging series should be considered with great care.

The Coulomb potential of a lattice of point charges (the so-called Hartree potential) is an example of a conditionally converging series. Indeed, consider periodically arranged atoms with charges $Q > 0$ and $-Q$ along a single dimension as shown in Fig. 7.3. The Coulomb potential at some point x away from the charges is

$$V_{\mathrm{Coul}}(x) = \sum_{n=-\infty}^{+\infty} \left[\frac{Q}{|x - an - X_+|} - \frac{Q}{|x - an - X_-|} \right],$$

where we sum over all unit cells using the unit cell number n, a is the cell length and X_+ and X_- are positions of the positive and negative atoms, respectively, within the

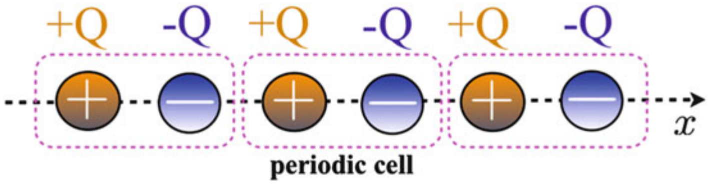


Fig. 7.3 A one-dimensional infinite periodic crystal of alternating point charges. Each unit cell (indicated) contains two oppositely charged atoms with charges $Q > 0$ and $-Q < 0$. The total charge of the unit cell is zero and there is an infinite number of such cells running along the positive and negative directions of the x axis

zero cell (with $n = 0$). This series does not converge absolutely. Indeed, the series with all terms positive would correspond to an infinitely long one-dimensional array of charges which are all positive, and hence the potential of such an infinite chain of positive atoms would be infinite. However, the potential of positive and negative charges does converge conditionally, although the result would depend on the order of the terms in the sum above.

7.2 Functional Series: General

Now we shall turn our attention to specific infinite series, called functional series, in which the general form of the n-th term a_n is a function $a_n(x)$ of a real variable x defined in some interval $a < x < b$. Correspondingly, as each term of the series is a function of x, i.e. $a_n(x)$, the sum of the series will also become a function of x:

$$S(x) = a_0(x) + a_1(x) + a_2(x) + \cdots = \sum_{n=0}^{\infty} a_n(x). \tag{7.17}$$

For a given value of x the series above is just an infinite numerical series we studied in the previous section. The difference now is that we deal with not just a single value of the x or even its finite set of values, but with x defined continuously in the interval $a < x < b$. Therefore, we shall be interested in the obvious questions related to the series being a function, such as: at which values of x the series converges and, if it does, whether the sum of the series is a continuous function. But the range of questions is much wider than that in fact. Can one take a limit inside the sign of the sum, or, in other words, would the limit of the sum be equal to the sum of the limits? Can one integrate or differentiate the series term-by-term, i.e. would the new series generated that way converge to a function obtained from the sum $S(x)$ after integration or differentiation? Finally, we shall consider a special and very important class of functional series called power series and show how functions can be expanded into such series, the so-called Taylor series. This point would conclude our discussion of the Taylor theorem started in Sect. 3.7. We shall only consider here the series which are functions of a single variable. The treatment of this simple case can be extended to functions of several variables almost without change.

7.2.1 *Uniform Convergence*

Because functional series have a parameter, x, their convergence may be different for different values of it; for some x the series may converge faster, for some other values slower. It is also possible that convergence within some interval of x is uniform, i.e. it does not depend on x (cf. Sect. 6.1.3). This class of series has a very special significance as for these series all the questions asked above can be answered satisfactorily. Therefore, we shall only be considering this case in what follows.

But before, let us recall the definition of the convergence of a numerical series (see Sect. 2.2.3): it is said that the series $\sum_n a_n$ converges to S if for any positive ϵ one can always find a number $\mathcal{N}$ such that for any $N > \mathcal{N}$ the partial sum S_N would be different from S by no more than ϵ, i.e. $|S_N - S| < \epsilon$. As in our case the terms of the series depend on x, for the given ϵ the number $\mathcal{N}$ may be different for different values of x; in other words, if we would like to approximate our series by its partial sum, different minimum number of terms $\mathcal{N}$ need to be taken for different values of x to achieve the same "precision" ϵ, i.e. $\mathcal{N} \Longrightarrow \mathcal{N}(x)$. If the convergence is *uniform*, however, then the same $\mathcal{N}$ can be used for *all values* of x, i.e. $\mathcal{N}$ does not depend on x. This is the idea of this essential new notion.

To illustrate this important concept, let us consider an alternating series

$$S(x) = \sum_{n=1}^{\infty} \frac{(-1)^n}{x+n}$$

for $x > 0$. It is known from Theorem 7.10 that this series converges; moreover, if we consider the first $\mathcal{N}$ terms, the sum of all dropped terms would not exceed the first dropped term (Theorem 7.11), i.e.

$$|S(x) - S_{\mathcal{N}}(x)| \leq \left| \frac{1}{x + \mathcal{N} + 1} \right| \leq \frac{1}{\mathcal{N}+1} < \frac{1}{\mathcal{N}},$$

where we have made use of the fact that $x > 0$. The above means that if we take $\epsilon = 1/\mathcal{N}$, then for any $N > \mathcal{N}$ we should have $|S(x) - S_N(x)| \leq 1/(N+1) < 1/\mathcal{N} = \epsilon$. We see that the same value of $\mathcal{N}$ can be chosen for any $x > 0$. Therefore, this series converges uniformly in the interval $0 < x < \infty$.

As another example, we consider a geometric progression

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

which converges for all x between -1 and 1 (excluding the boundary points). Consider the first N terms of it. The residue (the sum of all dropped terms) is

$$S(x) - S_N(x) = \sum_{n=N+1}^{\infty} x^n = x^{N+1} \sum_{n=0}^{\infty} x^n = \frac{x^{N+1}}{1-x}.$$

If we now require that the residue is smaller than ϵ, it should be clear that this would not be possible to achieve for all values of x between -1 and 1 since near $x = 1$ the residue steeply increases to infinity. In this case, therefore, the series converges non-uniformly.

Uniform convergence is sufficient in establishing a number of important properties of the functional series. But before discussing this, we shall provide one simple test for the uniform convergence.

Theorem 7.15 (Due to Weierstrass). *If all terms $a_n(x)$ of a functional series for all values of x within some interval satisfy the inequality*

$$|a_n(x)| \le c_n,$$

where numbers $c_n > 0$ ($n = 1, 2, 3, \ldots$) form a converging numerical series, then the functional series converges uniformly for all values of x from this interval.

Proof. Since the series $\sum_n c_n$ converges, it means that for any $\epsilon > 0$ one can always find $\mathcal{N}$ such that for any $N > \mathcal{N}$ the residue of the series,

$$\Delta C_N = \sum_{n=N+1}^{\infty} c_n,$$

satisfies $|\Delta C_N| < \epsilon$. Consider now the residue of our functional series for an arbitrary x within the given interval:

$$|\Delta S_N(x)| = |S(x) - S_N(x)| = \left| \sum_{n=N+1}^{\infty} a_n(x) \right| \le \sum_{n=N+1}^{\infty} |a_n(x)| < \sum_{n=N+1}^{\infty} c_n = \Delta C_N < \epsilon,$$

where we have used the fact here that the terms of the numerical series are positive and hence so is their residue ΔC_N. Note that here we chose $\mathcal{N}$ by the numerical series $\sum_n c_n$, and hence it is the same for all values of x from the interval. Hence, we established that for the given ϵ the residue of the functional series is less than ϵ for any $N > \mathcal{N}$ with $\mathcal{N}$ not depending on x, and this, according to our definition of the uniform convergence, proves the Theorem. **Q.E.D.**

In the so-called Fourier series $a_n(x)$ is given as either $\alpha_n \cos \frac{\pi n x}{l}$ or $\beta_n \sin \frac{\pi n x}{l}$, where numbers α_n and β_n form some numerical sequences, and l is a positive constant. The above theorem then establishes the uniform convergence of the two series based on the convergence of the series composed of the pre-factors α_n and β_n, since

$$\left|\alpha_n \cos \frac{\pi n x}{l}\right| \leq |\alpha_n| \quad \text{and} \quad \left|\beta_n \sin \frac{\pi n x}{l}\right| \leq |\beta_n| .$$

It also follows from Theorem 7.15 that the series composed of the absolute values of the terms of the original series, $|a_n(x)|$, would also converge uniformly; moreover, the given series $\sum_n a_n(x)$ would converge absolutely.

7.2.2 Properties: Continuity

Here we shall consider an important question about the continuity of the sum $S(x)$ of a uniformly converging series $\sum_n a_n(x)$.

Theorem 7.16. *Consider a uniformly converging series $\sum_n a_n(x)$ in the interval $a < x < b$. If the terms $a_n(x)$ of the series are continuous functions at a point x_0 belonging to the interval, then the sum of the series $S(x)$ is also a continuous function at this point.*

Proof. Let $\Delta S_N(x) = S(x) - S_N(x)$ be the residue of the series. Then,

$$\begin{aligned} |S(x) - S(x_0)| &= |S_N(x) - S_N(x_0) + \Delta S_N(x) - \Delta S_N(x_0)| \\ &\leq |S_N(x) - S_N(x_0)| + |\Delta S_N(x)| + |\Delta S_N(x_0)| . \end{aligned} \tag{7.18}$$

Since the series converges uniformly in the interval of interest (which includes the point x_0), then one can always find such N for the given $\epsilon > 0$ that $|\Delta S_N(x)| < \epsilon/3$ for all x between a and b, including x_0. Each of the functions $a_n(x)$ is assumed to be continuous by the conditions of the theorem, and so is their partial sum. Therefore, for any $\epsilon/3$ one should be able to find such $\delta > 0$ so that from $|x - x_0| < \delta$ follows $|S_N(x) - S_N(x_0)| < \epsilon/3$. Combining all these estimates, we can manipulate the inequality (7.18) into:

$$|S(x) - S(x_0)| \leq |S_N(x) - S_N(x_0)| + |\Delta S_N(x)| + |\Delta S_N(x_0)| < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon,$$

which proves the continuity of the sum of the series at point x_0. **Q.E.D.**

It is worth noting that condition of uniform convergence of the series on the interval $a < x < b$ was essential for the proof of the theorem.

The other question, closely related to the notion of continuity of the functional series, is the question of the limit: can one interchange the limit of the series at $x \to x_0$ with the sign of the sum? In other words, we inquire when the limit of the sum $S(x)$ is equal to the sum of the limits taken term-by-term:

$$\lim_{x \to x_0} \sum_n a_n(x) = \sum_n \lim_{x \to x_0} a_n(x). \tag{7.19}$$

This question is addressed by the following theorem proof of which is very similar to the one we have just considered.

Theorem 7.17. *Suppose the series $\sum_n a_n(x)$ converges uniformly to a function $S(x)$ on the interval $a < x < b$, and the terms of the series, $a_n(x)$, have well-defined limits at $x \to x_0$,*

$$\lim_{x \to x_0} a_n(x) = b_n.$$

Then $\lim_{x \to x_0} S(x) = B$, where B is the sum of the converging series $\sum_n b_n$. Basically, Eq. (7.19) is true.

Proof. We shall first prove one statement of the theorem, namely that the series $\sum_n b_n$ converges. We shall weaken the proof a bit by assuming that the uniform convergence is guaranteed by an auxiliary series of numbers c_n such that $|a_n(x)| \leq c_n$ (Weierstrass test). Taking the limit $x \to x_0$ in this inequality and employing the fact that any function $a_n(x)$ has the limit, we get $|b_n| \leq c_n$. Hence, since the series $\sum_n c_n$ converges, the series $\sum_n b_n$ converges as well and converges absolutely.

Let $B = B_N + \Delta B_N$ be the sum of the latter series with the corresponding partial sum B_N. Then,

$$|S(x) - B| = |S_N(x) - B_N + \Delta S_N(x) - \Delta B_N| \leq |S_N(x) - B_N| + |\Delta S_N(x)| + |\Delta B_N| . \tag{7.20}$$

Because of the uniform convergence, we can state that $|\Delta S_N(x)| < \epsilon/3$ for all x; since the $\sum_n b_n$ series converges, then one can always find such N that $|\Delta B_N| < \epsilon/3$. Finally, because the functions $a_n(x)$ all have a well-defined limit at $x \to x_0$, then for any ϵ one can always find such $\delta > 0$ that from $|x - x_0| < \delta$ follows $|a_n(x) - b_n| < \epsilon/3N$. Therefore,

$$\begin{aligned} |S_N(x) - B_N| &= \left| \sum_{n=1}^{N} a_n(x) - \sum_{n=1}^{N} b_n \right| = \left| \sum_{n=1}^{N} [a_n(x) - b_n] \right| \\ &\leq \sum_{n=1}^{N} |a_n(x) - b_n| < \sum_{n=1}^{N} \frac{\epsilon}{3N} = \frac{\epsilon}{3}. \end{aligned}$$

Collecting all results in Eq. (7.20), we finally obtain that from $|x - x_0| < \delta$ follows

$$|S(x) - B| \leq |S_N(x) - B_N| + |\Delta S_N(x)| + |\Delta B_N| < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon. \tag{7.21}$$

meaning that the series converges to B at $x \to x_0$. **Q.E.D.**

7.2.3 *Properties: Integration and Differentiation*

Here we shall investigate operations of term-by-term integration and differentiation of the uniformly converging series and will show that these operations are in fact legitimate.

Theorem 7.18. *Consider a uniformly converging series $S(x) = \sum_n a_n(x)$ on the interval $a < x < b$. Then, the integral of the sum is equal to the sum of the integrals of each individual term,*

$$\int_c^d S(x)\, dx = \sum_n \int_c^d a_n(x) dx, \tag{7.22}$$

where c and $d > c$ lie within the interval between a and b.

Proof. Since the series converges, we can write:

$$S(x) = \sum_{n=1}^{N} a_n(x) + \Delta S_N(x).$$

This equality contains a finite number of terms and hence can be integrated between c and d term-by-term:

$$\int_c^d S(x) dx = \sum_{n=1}^{N} \int_c^d a_n(x) dx + \int_c^d \Delta S_N(x) dx. \tag{7.23}$$

On the other hand, since the series converges uniformly on the interval $a < x < b$ (which includes completely the integration interval), one can always find such N that $|\Delta S_N(x)| < \epsilon$ for all x. Therefore, we can estimate the last integral as follows:

$$\left| \int_c^d \Delta S_N(x) dx \right| \leq \int_c^d |\Delta S(x)|\, dx < \int_c^d \epsilon dx = \epsilon\,(d - c) = \epsilon'.$$

This means that the integral of the residue $\Delta S_N(x)$ tends to zero as $N \to \infty$ by the definition of the limit. This proves that at this limit the last term in Eq. (7.23) tends to zero and hence the integral of the sum is equal to the sum of the integrals, i.e. the integration and summation can be interchanged in the case of the uniformly converging series. **Q.E.D.**

Now we shall prove a similar statement for differentiation.

Theorem 7.19. *Consider a uniformly converging series $S(x) = \sum_n a_n(x)$ on the interval $a < x < b$, and let us assume that functions $a_n(x)$ have continuous derivatives $a_n'(x)$. If the series containing derivatives of the terms of the original series, $\sum_n a_n'(x)$, converges and converges also uniformly, then the sum $S(x)$ of the original series has a well-defined derivative*

$$S'(x) = \sum_n a_n'(x). \tag{7.24}$$

Proof. The series $\sum_n a_n'(x)$ converges uniformly and hence its sum, $S_1(x)$, will be a continuous function (Theorem 7.16). According to Theorem 7.18, we can integrate the series term-by-term between c and some $x > c$, both lying somewhere between a and b:

$$\int_c^x S_1\left(x'\right) dx' = \sum_n \int_c^x a_n'\left(x'\right) dx' \quad \Longrightarrow \quad \int_c^x S_1\left(x'\right) dx' = \sum_n \left[a_n(x) - a_n(c)\right],$$

where above we calculated the integral of $a_n'\left(x'\right)$ explicitly. Using the fact that the original series of $a_n(x)$ converges to $S(x)$, we can write for the right-hand side of the above equation:

$$\sum_n \left[a_n(x) - a_n(c)\right] = \sum_n a_n(x) - \sum_n a_n(c) = S(x) - S(c),$$

yielding

$$\int_c^x S_1\left(x'\right) dx' = S(x) - S(c). \tag{7.25}$$

Let us now differentiate both sides of Eq. (7.25) with respect to x (recall Eq. (4.34) or a more general Eq. (4.65)), and we obtain: $S_1(x) = S'(x)$, as required. **Q.E.D.**

So, uniformly converging functional series can be both integrated and differentiated term-by-term (in the latter case the series of derivatives must also converge uniformly).

Problem 7.16. *In quantum statistical mechanics a single harmonic oscillator of frequency ω has an infinite set of discrete energy levels $\epsilon_n = \hbar\omega\,(n + 1/2)$, where $\hbar = h/2\pi$ is the "h-bar" Planck constant, and the integer $n = 0, 1, 2, \ldots$. The probability for an oscillator to be in the n-th state is given by $\rho_n = Z^{-1}\exp(-\epsilon_n/k_BT)$, where k_B is the Boltzmann's constant and T absolute temperature. Show that the so-called partition function Z of the oscillator is given by:*

$$Z = \sum_{n=0}^{\infty} e^{-\xi(n+1/2)} = \frac{e^{\xi/2}}{e^{\xi} - 1}, \quad \text{where} \quad \xi = \hbar\omega/k_BT,$$

Argue that the series for $Z(\xi)$ converges uniformly with respect to ξ. Therefore, prove the following formula:

$$\sum_{n=0}^{\infty} n e^{-\xi n} = -\frac{d}{d\xi}\left(\sum_{n=0}^{\infty} e^{-\xi n}\right) = \frac{e^{\xi}}{\left(e^{\xi} - 1\right)^2},$$

so that the average number of oscillators in the state n is given by

$$\langle n \rangle = \sum_{n=0}^{\infty} n\rho_n = \left(e^{\xi} - 1\right)^{-1},$$

which is called the Bose–Einstein distribution (statistics).

7.3 Power Series

Now we shall turn our attention to an important special case of functional series, when $a_n(x) = \alpha_n x^n$ with the coefficients α_n forming an infinite numerical sequence. Correspondingly, we shall be considering the so-called *power series* of the form:

$$S(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots = \sum_{n=0}^{\infty} \alpha_n x^n. \tag{7.26}$$

This power series is said to be run around point $x = 0$. A more general power series can be considered around an arbitrary point $x = a$,

$$S(x) = \alpha_0 + \alpha_1 (x - a) + \alpha_2 (x - a)^2 + \cdots = \sum_{n=0}^{\infty} \alpha_n (x - a)^n. \tag{7.27}$$

However, this one can be constructed from the first one (7.26) by a simple shift and hence this generalisation is not really necessary for proving general statements and theorems.

7.3.1 Convergence of the Power Series

We shall start by proving an important theorem related to convergence of the power series.

Theorem 7.20 (Due to Abel). *If the power series (7.26) converges at some x_0 (which is not equal to zero), then it* absolutely *converges for any x satisfying $|x| < |x_0|$, i.e. for $-|x_0| < x < |x_0|$.*

Proof. Since we know that the series with $a_n = \alpha_n x_0^n$ converges, then $\alpha_n x_0^n \to 0$ when $n \to \infty$, and therefore, since the limit exists, a general term of the series is bounded by some real positive number M_0, i.e. $\left|\alpha_n x_0^n\right| < M_0$. Consider now the series constructed of absolute values of the terms of the original series taken at the value of x satisfying the inequality $|x| < |x_0|$:

$$\sum_{n=0}^{\infty} |\alpha_n x^n| = \sum_{n=0}^{\infty} |\alpha_n x_0^n| \left|\frac{x}{x_0}\right|^n < \sum_{n=0}^{\infty} M_0 \left|\frac{x}{x_0}\right|^n = M_0 \sum_{n=0}^{\infty} q^n = \frac{M_0}{1-q},$$

where $q = |x/x_0| < 1$ and the sum of its powers forms a converging geometrical progression which sum we know well, Eq. (2.7). This means that our series converges absolutely for any x between $-|x_0|$ and $|x_0|$, i.e. $-|x_0| < x < |x_0|$, as required. **Q.E.D.**

Theorem 7.21 (Due to Abel). *If, however, the series (7.26) diverges at x_0, it diverges for any x satisfying $|x| > |x_0|$, i.e. for $x < -|x_0|$ and $x > |x_0|$.*

Problem 7.17. *Prove the above theorem using the method of contradiction.*

Schematically the convergence and divergence intervals are shown in Fig. 7.4.

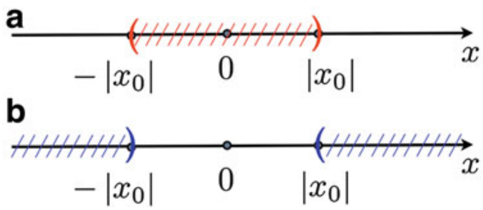


Fig. 7.4 (**a**) If a series converges at x_0, it also converges (and converges absolutely) within the interval shaded; (**b**) if the series diverges at x_0, it also diverges for any x in the two shaded intervals

Theorem 7.22. *If a series converges at some $x_0 \neq 0$ and also is known to diverge at some x_1 (obviously, $|x_1|$ must be larger than $|x_0|$, otherwise it would contradict the Abel's theorems), then there exists such positive R that the series converges for any $|x| < R$ and diverges for any $|x| > R$ (nothing can be said about $x = R$ though).*

Proof. Since the series converges at some x_0, it must converge at any x within the interval $-|x_0| < x < |x_0|$ according to Theorem 7.20; moreover, since it diverges at some x_1, it will also diverge for any x satisfying $x < -|x_0|$ and $x > |x_0|$ (Theorem 7.21). Assume then that the series converges at some point x_2 which absolute value is between $|x_0|$ and $|x_1|$, see Fig. 7.5(a). Then, according to the Abel's theorem, it will also converge for any x satisfying $-|x_2| < x < |x_2|$, i.e. in the interval which completely includes the previous convergence interval $-|x_0| < x < |x_0|$, see Fig. 7.5(b). Similarly, one can consider a point x_3 lying between $|x_2|$ and $|x_1|$. Assuming that the series diverges at x_3, we have to conclude from the Abel's theorem that it will diverge anywhere when $x > |x_3|$ and $x < -|x_3|$, i.e. the new divergence intervals would completely incorporate the previous ones, and the boundaries between the diverging and converging intervals would move towards each other, see Fig. 7.5(c). Continuing this process, we shall come to a point where the boundaries meet at point $R > 0$ and $-R < 0$, so that for any x between $-R$ and R the series converges, and for any $x < -R$ and $x > R$ it diverges. Since the intervals are exclusive of their boundary points, nothing can be said about the points $x = \pm R$. **Q.E.D.**

The positive number R is called the *radius of convergence*. If the series converges everywhere, it is said that the radius of convergence is infinite, $R = \infty$. If the series converges only at $x = 0$, then $R = 0$.

It is easy to see now that the radius of convergence can be calculated via

$$R = \lim_{n\to\infty} \left| \frac{\alpha_n}{\alpha_{n+1}} \right| . \tag{7.28}$$

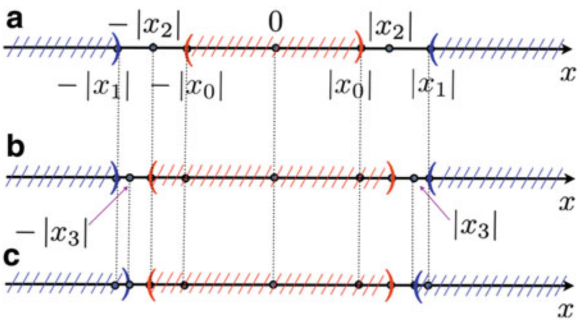


Fig. 7.5 For the proof of Theorem 7.22. (**a**) The series converges everywhere in the red-shaded intervals and diverges everywhere in the blue-shaded ones. (**b**) Assuming that it also converges at some x_2 which lies between $|x_0|$ and $|x_1|$, would inevitably widen the convergence interval according to the Abel's theorem. (**c**) Assuming that the series diverges at x_3, this widens the divergence intervals moving the boundaries of converging and diverging regions even closer to each other

If the limit above is infinite, then $R = \infty$, while for the zero limit we have $R = 0$.

Problem 7.18. *Prove the above formula by applying the ratio test.*

Problem 7.19. *Prove by applying the root test that R can also be written as*

$$\frac{1}{R} = Sup_{n\to\infty} \sqrt[n]{|\alpha_n|}, \tag{7.29}$$

where Sup means the maximum value achieved after taking the limit.

Problem 7.20. *Show that $R = \infty$ for the series with $\alpha_n = 1/n!$.*

Problem 7.21. *Show that $R = \infty$ for the series with $a_n = (-1)^n x^{2n+1} / (2n+1)!$.*

Problem 7.22. *Show that $R = \infty$ for the series with $a_n = (-1)^n x^{2n} / (2n)!$.*

Problem 7.23. *Show that $R = 1$ for the series with $\alpha_n = 1/n$.*

For the latter problem ($\alpha_n = 1/n$) nothing can be said about the points $x = \pm 1$ from the ratio test. However, we studied both series before: at $x = 1$, the power series becomes the harmonic series about which we already know that it diverges; at $x = -1$, the alternating harmonic series appears for which we also know that it converges (conditionally). The situation is opposite for the series with $\alpha_n = (-1)^n / n$.

7.3.2 *Uniform Convergence and Term-by-Term Differentiation and Integration of Power Series*

Power series present the simplest and very important example of the uniformly converging functional series and hence can be integrated and differentiated term-by-term. Indeed, the following theorem proves the uniform convergence.

Theorem 7.23. *If R is the radius of convergence of the power series (7.26), then the series converges uniformly within the radius of $0 < r < R$, i.e. for any x satisfying $-r \leq x \leq r$.*

Proof. Since $r < R$, then the series $\sum_n \alpha_n r^n$ converges absolutely, i.e. the series $\sum_n |\alpha_n r^n|$ converges. Also, for any $|x| < r$ we have

$$|a_n(x)| = |\alpha_n| \, |x|^n < |\alpha_n| \, r^n ,$$

which, by virtue of the uniform convergence test (Theorem 7.15), means that our series indeed converges uniformly. **Q.E.D.**

This in turn means that the power series represents a continuous function for any $-r \leq x \leq r$ with the positive $r < R$ (Theorem 7.16). Note that nothing can be said about the boundary points $x = \pm R$ where a special investigation is required. It can be shown, for instance, that if the series converges at $x = R$, it converges uniformly for all x satisfying $0 \leq x \leq R$. For instance, the series

$$x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n \tag{7.30}$$

at $x = 1$ becomes the alternating harmonic series (7.12) which, as we know from Sect. 7.1.3, converges (although not absolutely). This means that the series (7.30) converges uniformly for any $0 \leq x \leq 1$, and hence its sum is a continuous function for all these values of x. We shall see explicitly in the following that this series represents a power series of $\ln(1+x)$ and hence at $x = 1$ converges to $\ln 2$. Note that the statement made above comes from the fact that if $S(x) = \sum_n \alpha_n x^n$ uniformly converges for any x from the semi-open interval $0 \leq x < R$, where the function $S(x)$ is continuous, and since the series converges at the right boundary $x = R$, one can take the limit of $S(x)$ at $x \to R - 0$ (from the left) and then arrive at the converging series at $x = R$. So, the function $S(x)$ will be continuous in the closed interval $0 \leq x \leq R$.

If the power series converges uniformly in the open, semi-open or closed interval between $-R$ and R, then, according to Theorems 7.18 and 7.19 of Sect. 7.2.3 it can be integrated and differentiated term-by-term.

One point still needs clarifying though: it is that, according to Theorem 7.19, the series containing derivatives of the terms of the original series should also converge uniformly. According to Theorem 7.23, for this it is sufficient to demonstrate that the series of derivatives has the same radius of convergence as the original power series. To this end, we can apply the ratio test to the series of derivatives, which has the general term $b_n = na_n x^{n-1}$:

$$\lim_{n\to\infty}\left|\frac{b_{n+1}}{b_n}\right| = \lim_{n\to\infty}\left|\frac{(n+1)\,a_{n+1}x^n}{na_n x^{n-1}}\right| = \lim_{n\to\infty}\left|\left(1+\frac{1}{n}\right)\frac{a_{n+1}x}{a_n}\right| = \lim_{n\to\infty}\left|\frac{a_{n+1}x}{a_n}\right| .$$

But the original series of $\sum_n a_n x^n$ converges for any x between $-R$ and R and hence, according to the same ratio test, the last limit is smaller than one:

$$\lim_{n\to\infty}\left|\frac{a_{n+1}x^{n+1}}{a_n x^n}\right| = \lim_{n\to\infty}\left|\frac{a_{n+1}x}{a_n}\right| < 1 .$$

Hence, the ratio test establishes convergence of the series of derivatives within the same interval of the original power series.

Since the series of derivatives is also a power series, it can be also differentiated. It is seen from this discussion that the power series can be differentiated any number of times; each series thus obtained would have the same radius of convergence.

7.3.3 Taylor Series

The properties of the power series considered above allow us to formulate a very powerful method enabling one to expand a function $f(x)$ into a power series (cf. Sect. 3.7).

First of all, we prove that if $f(x)$ has a power series expansion,

$$f(x) = \sum_{n=0}^{\infty}\alpha_n x^n = \alpha_0 x^0 + \alpha_1 x^1 + \alpha_2 x^2 + \alpha_3 x^3 + \cdots , \tag{7.31}$$

then this expansion is unique. Indeed, suppose there exists another expansion of the same function,

$$f(x) = \sum_{n=0}^{\infty}\beta_n x^n = \beta_0 x^0 + \beta_1 x^1 + \beta_2 x^2 + \beta_3 x^3 + \cdots . \tag{7.32}$$

Since both series converge within some interval between $-R$ and R, they converge uniformly, and hence the series can be differentiated term-by-term. In fact, they can be differentiated many times:

$$f^{(1)}(x) = \sum_{n=1}^{\infty} n\alpha_n x^{n-1} = \sum_{n=1}^{\infty} \frac{n!}{(n-1)!}\alpha_n x^{n-1} = \alpha_1 x^0 + 2\alpha_2 x^1 + 3\alpha_3 x^2 + \cdots , \tag{7.33}$$

$$\begin{aligned} f^{(2)}(x) &= \sum_{n=2}^{\infty} n(n-1)\alpha_n x^{n-2} = \sum_{n=2}^{\infty} \frac{n!}{(n-2)!}\alpha_n x^{n-2} \\ &= 2\cdot 1\alpha_2 x^0 + 3\cdot 2\alpha_3 x^1 + 4\cdot 3\alpha_4 x^2 + \cdots , \end{aligned} \tag{7.34}$$

$$\begin{aligned} f^{(3)}(x) &= \sum_{n=3}^{\infty} n(n-1)(n-2)\alpha_n x^{n-3} = \sum_{n=3}^{\infty} \frac{n!}{(n-3)!}\alpha_n x^{n-3} \\ &= 3\cdot 2\cdot 1\alpha_3 x^0 + 4\cdot 3\cdot 2\alpha_4 x^1 + \cdots , \end{aligned} \tag{7.35}$$

and so on. The k-th term has the form:

$$\begin{aligned} f^{(k)}(x) &= \sum_{n=k}^{\infty} n(n-1)\cdots(n-k+1)\alpha_n x^{n-k} = \sum_{n=k}^{\infty} \frac{n!}{(n-k)!}\alpha_n x^{n-k} \\ &= k!\alpha_k x^0 + \frac{(k+1)!}{1!}\alpha_{k+1} x^1 + \cdots , \end{aligned} \tag{7.36}$$

and similarly for the other series (7.32). Notice that after the differentiation of the original series the constant term in the series (the term with x^0) disappears and the series starts from the next term; correspondingly, the series for the k-th derivative of $f(x)$ starts from the $n = k$ term. Let us now put $x = 0$ in Eqs. (7.31) and (7.32); we immediately obtain that $\alpha_0 = \beta_0$. If we now do the same in the two expressions for the first derivative of $f(x)$, we obtain $\alpha_1 = \beta_1$; continuing this process, we progressively have $\alpha_2 = \beta_2$, $\alpha_3 = \beta_3$, etc., i.e. $\alpha_k = \beta_k$ for any $k = 0, 1, 2$, etc. This proves the statement made above.

Next, we can relate the coefficients α_0, α_1, α_3, etc. to the function $f(x)$ itself and its derivatives. Indeed, putting $x = 0$ again in Eqs. (7.31), (7.33)–(7.35) yields, respectively:

$$f(0) = \alpha_0 , \quad f^{(1)}(0) = \alpha_1 , \quad f^{(2)}(0) = 2!\alpha_2 , \quad f^{(3)}(0) = 3!\alpha_3 , \ \ldots ,$$

and from (7.36) we generally have $f^{(k)}(0) = k!\alpha_k$, i.e. the expansion of $f(x)$, if it exists, should necessarily have the form:

$$f(x) = f(0) + f^{(1)}(0)x + \frac{f^{(2)}(0)}{2!}x^2 + \frac{f^{(3)}(0)}{3!}x^3 + \cdots = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!}x^n, \tag{7.37}$$

which is called Maclaurin series, and, of course, this is not accidental. After all, we can write the Maclaurin formula (3.62)

$$f(x) = f(0) + \frac{f'(0)}{1!}x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^2 + \cdots + \frac{f^{(n)}(0)}{n!}x^n + R_{n+1}(x), \tag{7.38}$$

when representing any n-times differentiable function $f(x)$, where

$$R_{n+1}(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}x^{n+1} = \frac{f^{(n+1)}(\vartheta x)}{(n+1)!}x^{n+1}, \tag{7.39}$$

is the remainder term with $0 < \vartheta < 1$. Comparing the two results, Eqs. (7.37) and (7.38), we immediately notice that the first n terms in both of them coincide. One may say then that the Maclaurin series is obtained from the Maclaurin formula of the infinite order ($n \to \infty$). This close relationship between the two results allows one to formulate the necessary and sufficient conditions at which the Maclaurin series of the function $f(x)$ converges exactly to the function $f(x)$ itself.

Theorem 7.24. *In order for the Maclaurin series (7.37) to converge to $f(x)$ within the interval $-R < x < R$, it is necessary and sufficient that the remainder term (7.39) of the Maclaurin formula would tend to zero as $n \to \infty$.*

Proof. We shall first prove sufficiency, i.e. we start by assuming that $\lim_{n\to\infty} R_{n+1}(x) = 0$. Then $f(x) = S_{n+1}(x) + R_{n+1}(x)$, where $S_{n+1}(x)$ contains the first $(n+1)$ terms of the Maclaurin formula, and, at the same time, coincides with the partial sum of the series with $(n+1)$ first terms. Since $\lim_{n\to\infty} R_{n+1}(x) = 0$ by our assumption, then

$$\lim_{n\to\infty} R_{n+1}(x) = \lim_{n\to\infty} [f(x) - S_{n+1}(x)] = f(x) - \lim_{n\to\infty} S_{n+1}(x) = 0$$
$$\Longrightarrow \quad \lim_{n\to\infty} S_{n+1}(x) = f(x),$$

which immediately proves the required statement of the convergence of the infinite power series to the function $f(x)$ at each point x between $-R$ and R.

Now let us prove the necessary condition. We start from the assumption that the series (7.37) converges to $f(x)$, and hence $\lim_{n\to\infty} S_{n+1}(x) = f(x)$ for $-R < x < R$. Correspondingly,

$$\lim_{n\to\infty} S_{n+1}(x) = \lim_{n\to\infty} [f(x) - R_{n+1}(x)] = f(x) - \lim_{n\to\infty} R_{n+1}(x) = f(x)$$
$$\Longrightarrow \quad \lim_{n\to\infty} R_{n+1}(x) = 0,$$

as required. **Q.E.D.**

Therefore, convergence of the series (7.37) requires in each case investigating the limiting behaviour of the remainder term (7.39). In this respect it is expedient to notice that an important n-dependent part of a general term in the Taylor's series tends to zero as $n \to \infty$, i.e.

$$\lim_{n\to\infty} \frac{x^n}{n!} = 0 \tag{7.40}$$

for any real x. This can be seen, e.g., by applying the ratio test to the sequence $a_n(x) = x^n/n!$ (which is the Maclaurin expansion of the e^x as we shall immediately see below). Since this series converges everywhere (the ratio test is a sufficient condition), then necessarily (Theorem 7.4) $a_n = x^n/n! \to 0$ as $n \to \infty$. Of course, condition (7.40) is only necessary and does not mean that for any $f(x)$ the series (7.37) converges for any real x since the behaviour of the pre-factors $f^{(n)}(0)$ may still affect the convergence region (the value of R).

As an example, let us consider the expansion of the function $f(x) = e^x$. The remainder term

$$R_{n+1}(x) = x^{n+1}\left[\frac{(e^x)^{(n+1)}}{(n+1)!}\right]_{x\to\vartheta x} = \frac{e^{\vartheta x}x^{n+1}}{(n+1)!} = e^{\vartheta x}\left[\frac{x^{n+1}}{(n+1)!}\right] \to 0$$

as $n \to \infty$ because of the necessary condition (7.40). Therefore, the expansion of the exponential function has the form:

$$e^x = 1 + x + \frac{x^2}{2!} + \cdots = \sum_{n=0}^{\infty} \frac{x^n}{n!}. \tag{7.41}$$

The series converges to e^x for any x from $-\infty < x < \infty$ (in other words, the radius of convergence $R = \infty$).

Problem 7.24. *Show that*

$$\lim_{x\to 0}\left(\frac{\partial}{\partial x}\right)^n e^{\sigma x^2/2} = \frac{n!}{2^{n/2}(n/2)!}\sigma^{n/2},$$

if n is even, and that it is equal to zero if n is odd. [*Hint*: expand first the exponential function into the Maclaurin series.]

Problem 7.25. *The so-called cumulants K_n and moments M_n of a random process are defined by the following connection relation:*

$$1+\sum_{n=1}^{\infty}\frac{u^n}{n!}M_n=\exp\left(\sum_{n=1}^{\infty}\frac{u^n}{n!}K_n\right). \tag{7.42}$$

By expanding the exponential function in the right-hand side and comparing the coefficients at the same powers of the parameter u in both sides, show that:

$$M_1=K_1\ ;\quad M_2=K_2+K_1^2\ ;\quad M_3=K_3+3K_1K_2+K_1^3.$$

Similarly one can consider expansion of other elementary functions into the Maclaurin series and investigate their convergence. The form of the series themselves can be borrowed from the corresponding Taylor's formulae (3.66)–(3.69):

$$\ln(1+x)=x-\frac{x^2}{2}+\frac{x^3}{3}-\cdots+(-1)^{n+1}\frac{x^n}{n}+\cdots=\sum_{n=1}^{\infty}(-1)^{n+1}\frac{x^n}{n}; \tag{7.43}$$

$$\sin x=x-\frac{x^3}{3!}+\frac{x^5}{5!}-\frac{x^7}{7!}+\cdots+\frac{(-1)^{n+1}x^{2n-1}}{(2n-1)!}+\cdots=\sum_{n=1}^{\infty}\frac{(-1)^{n-1}x^{2n-1}}{(2n-1)!}; \tag{7.44}$$

$$\cos x=1-\frac{x^2}{2!}+\frac{x^4}{4!}-\cdots+\frac{(-1)^n x^{2n}}{(2n)!}+\cdots=\sum_{n=0}^{\infty}\frac{(-1)^n x^{2n}}{(2n)!}; \tag{7.45}$$

$$(1+x)^{\alpha}=1+\alpha x+\frac{\alpha(\alpha-1)}{2}x^2+\cdots+\frac{\alpha(\alpha-1)(\alpha-2)\cdots(\alpha-n+1)}{n!}x^n$$
$$+\cdots=\sum_{n=0}^{\infty}\binom{\alpha}{n}x^n, \tag{7.46}$$

where the generalised binomial coefficients $\binom{\alpha}{n}$ are given by Eq. (3.70); further,

$$\arcsin(x)=\sum_{n=0}^{\infty}\frac{(2n)!}{4^n(2n+1)(n!)^2}x^{2n+1}; \tag{7.47}$$

$$\arccos(x) = \frac{\pi}{2} - \sum_{n=0}^{\infty} \frac{(2n)!}{4^n (2n+1) (n!)^2} x^{2n+1}; \tag{7.48}$$

$$\arctan(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} x^{2n+1}. \tag{7.49}$$

Note that the series for the inverse sine, cosine and tangent functions also converge at $x = \pm 1$ and hence the formulae (7.47)–(7.49) are valid for these values as well (in the former two cases this can be shown with the ratio test, a more powerful test should be used for the arctangent).

Problem 7.26. *Consider an expression* $(1+x)^{\alpha}(1+x)^{\beta} = (1+x)^{\alpha+\beta}$. *By expanding the left- and the right-hand sides using Eq. (7.46) and then comparing the coefficients in both sides to the same power of x, prove the following identity:*

$$\sum_{k=0}^{n} \binom{\alpha}{k} \binom{\beta}{n-k} = \binom{\alpha+\beta}{n}. \tag{7.50}$$

Problem 7.27. *Show (using, e.g., the ratio test) that the radius of convergence of the* $\ln(1+x)$ *expansion is* $R = 1$. *Explain why the expansion also converges (to* $\ln 2$*) at* $x = 1$, *but diverges at* $x = -1$.

Problem 7.28. *Show that* $R = \infty$ *for the sine and cosine functions.*

Problem 7.29. *Show that* $R = 1$ *for the expansion of* $(1+x)^{\alpha}$.

Problem 7.30. *Prove the Euler's formula* $e^{ix} = \cos x + i \sin x$ *by formally applying the Maclaurin expansions to all functions involved.*

Problem 7.31. *Show that* $R = 1$ *for the inverse sine and cosine functions.*

Problem 7.32. *Consider the geometric progression for* $(1+x)^{-1}$. *Obtain the Maclaurin expansion of* $\ln(1+x)$ *by integrating the progression term-by-term.*

Problem 7.33. *Consider the geometric progression for* $\left(1+x^2\right)^{-1}$. *Obtain the Maclaurin expansion of* $\arctan x$ *by integrating the progression term-by-term.*

Problem 7.34. *Consider the geometric progression for* $\left(1-x^2\right)^{-1/2}$. *Obtain the Maclaurin expansion of* $\arcsin x$ *by integrating the progression term-by-term.*

Problem 7.35. *Determine the first five terms of the Maclaurin expansion of the following functions:*

$$e^{2\cos x}\ ;\quad \sin\left(1+x^2\right)\ ;\quad \frac{\sin x^2}{x^2}\ ;\quad \frac{2x+1}{\left(x^2-4x+1\right)^3}.$$

Perform the calculation using two methods: (1) use directly formula (7.37) which would require differentiating the corresponding functions several times and then setting $x=0$*; (2) recognise that the above functions can be obtained from elementary functions with a more complex argument; correspondingly, use the known expansions and replace the x there with the actual argument and keep the terms up to the 5-th power of the x. In the case of the first function the first terms of the expansion for* $\cos x$ *will also have to be used.*

Problem 7.36. *Determine the first five terms of the Maclaurin series for the following functions of t:*

$$\frac{1}{\sqrt{1-2tx+t^2}};\quad e^{-2xt-t^2}\ ;\quad \frac{1}{1-t}\exp\left(-\frac{xt}{1-t}\right);\quad \frac{1-tx}{1-2xt+t^2}.$$

The corresponding coefficients of the expansion are polynomials in x, and are related, respectively, to the so-called Legendre, Hermite, Laguerre and Chebyshev polynomials.

Problem 7.37. *Prove for all the cases above in Eqs. (7.43)–(7.49) that the corresponding remainder* $R_{n+1}\to 0$ *and* $n\to\infty$*. Explicit expressions for the remainder term in each case are given at the end of Sect. 3.7.*

Problem 7.38. *Take the logarithm of both sides of Eq. (7.42), expand the logarithm and show that the low order cumulants* K_n *are related to the moments* M_n *via*

$$K_1=M_1\ ;\quad K_2=M_2-M_1^2\ ;\quad K_3=M_3-3M_1M_2+2M_1^3.$$

Maclaurin series is frequently used for approximate calculations. In particular, one can obtain expansions into a power series of some integrals which cannot be calculated analytically. This is illustrated by the following problems.

Problem 7.39. *Using the expansion of* $\sin x/x$*, prove the following formula:*

$$\int_0^x \frac{\sin t}{t}dt = x-\frac{x^3}{3!\cdot 3}+\frac{x^5}{5!\cdot 5}+\cdots=\sum_{n=0}^{\infty}\frac{(-1)^n x^{2n+1}}{(2n+1)!\,(2n+1)}. \tag{7.51}$$

Problem 7.40. *Using the expansion of e^{-x^2}, derive the following expansion of the integral of it (called the error function):*

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt = \frac{2}{\sqrt{\pi}} \left(x - \frac{x^3}{3 \cdot 1!} + \frac{x^5}{5 \cdot 2!} - \cdots \right)$$
$$= \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)n!}. \tag{7.52}$$

Problem 7.41. *Using the expansion of the sine function, obtain the following power expansion for the Fresnel integrals:*

$$C(x) = \int_0^x \cos\left(t^2\right) dt = \sum_{n=0}^{\infty} \frac{(-1)^n}{(4n+1)\,(2n)!} x^{4n+1}, \tag{7.53}$$

$$S(x) = \int_0^x \sin\left(t^2\right) dt = \sum_{n=0}^{\infty} \frac{(-1)^n}{(4n+3)\,(2n+1)!} x^{4n+3}. \tag{7.54}$$

Chapter 8
Ordinary Differential Equations

The ordinary differential equations (ODE's in short), or simply differential equations (DE), are the equations of the type

$$F\left(x, y, y', y'', \ldots, y^{(n)}\right) = 0,$$

relating the variable x, a function $y(x)$ of x, and its derivatives $\frac{dy}{dx} = y'$, $\frac{d^2y}{dx^2} = y''$, etc. with each other. The fundamental task is to find a function $y(x)$ that satisfies this equation. It may be that not one, but many functions form different solutions.

ODEs can be characterised by:

- *order*—order n of the highest derivative $y^{(n)}$ of the unknown function;
- *degree*—the power of its highest order derivative when the equation is written in a polynomial form (e.g. without roots).

A specific class of ODEs are *linear* ODEs if the unknown function and its derivatives are there in the first power without any cross terms. We shall see in this chapter that a significant progress can be made for this particular class of ODEs. Linear ODEs are also very often met in practical applications in physics, biology, chemistry and engineering.

Similar to integration, a general method for solving an arbitrary ODE does not exist, and whether the particular equation can be solved depends in many cases on fortune to find the right trick. There are certain cases, however, for which the necessary tricks have been developed and we shall consider some of them in this chapter.

An ODE may have a *general solution* that contains one or more *arbitrary constants* (that we shall write as C, C_1, C_2, etc. in what follows). The value(s) of the constant(s) can be fixed by specifying some information about the unknown function, e.g. its value $y(x_0)$ at some point x_0, and/or the value of its derivative(s). If the same point x_0 is used, these additional conditions are called *initial conditions*. If different points x_1, x_2, etc. are used, such conditions are usually called *boundary*

L. Kantorovich, *Mathematics for Natural Scientists*, Undergraduate Lecture Notes in Physics, DOI 10.1007/978-1-4939-2785-2_8

conditions since these points normally correspond to the boundary of the system in question. In this chapter we shall be only concerned with the former case; examples of the latter are met, e.g. when solving partial differential equations of mathematical physics which we do not consider here (see Book II, Chap. 8). The solution in which no arbitrary constant(s) is (are) present (and hence some specific information about the unknown function was already applied) is called a *particular solution*.

8.1 First Order First Degree Differential Equations

Consider an equation $F(x, y, y') = 0$, in which only the first derivative of an unknown function $y(x)$ is present. Suppose one can solve for the derivative y' in this (algebraic) equation. Then we have to study the ODE of the following simplest general form:

$$\frac{dy}{dx} = f(x, y). \tag{8.1}$$

In this section several methods will be considered which enable one to find the function(s) $y(x)$ satisfying this equation.

8.1.1 Separable Differential Equations

If $f(x, y) = g(x)/h(y)$, then Eq. (8.1) is solved first by rearranging the both sides:

$$\frac{dy}{dx} = \frac{g(x)}{h(y)} \quad \Longrightarrow \quad h(y)dy = g(x)dx. \tag{8.2}$$

The next step needs some explanations. Recall that differentials can not only be written for a single variable, like dx, but also for a function $f(x)$, like $df = \frac{df}{dx}dx = f'(x)dx$. Here $df = f(x + dx) - f(x)$ corresponds to the change of $f(x)$ associated with the variable x changing between x and $x + dx$. This means that an expression like $g(x)dx$ in the second step in Eq. (8.2) can be written as the differential $dG(x) = g(x)dx$ of some function $G(x) = \int g(x)dx$ of x, while the expression $h(y)dy$ there can also be written as a differential $dH(y) = h(y)dy$ of some other function $H(y) = \int h(y)dy$, with respect to the y. Therefore, the line $h(y)dy = g(x)dx$ is equivalent to saying that these two differentials are equal, $dH(y) = dG(x)$, which means that the functions themselves differ only by an arbitrary constant C, i.e. $H(y) = G(x) + C$. In other words, we obtain the final result which formally solves the ODE in question:

$$\int h(y)dy = \int g(x)dx + C,$$

where the two integrals are understood as indefinite ones. It is seen that the above solution can formally be obtained by integrating both sides of the equation for differentials:

$$h(y)dy = g(x)dx \quad \Longrightarrow \quad \int h(y)dy = \int g(x)dx,$$

and introducing a constant C after the integration. This gives the general solution of the ODE. If the initial condition is known, $y(x_0) = y_0$, then integration can be performed from y_0 to y in the left, and from x_0 to x in the right integrals, respectively, in which case a particular integral of the ODE is obtained.

As an example, consider equation

$$xy' - xy = y,$$

subject to the initial conditions $y(1) = 1$. Regrouping the terms, we can write:

$$x\frac{dy}{dx} = y + xy \quad \Longrightarrow \quad x\frac{dy}{dx} = y(1+x) \quad \Longrightarrow \quad \frac{dy}{y} = \frac{1+x}{x}dx$$
$$\Longrightarrow \quad \int \frac{dy}{y} = \int \left(\frac{1}{x} + 1\right) dx,$$

which gives the general solution as

$$\ln y = \ln x + x + C_1 \quad \Longrightarrow \quad \frac{y}{x} = \exp(x + C_1) = e^{C_1}e^x \quad \Longrightarrow \quad y = Cxe^x.$$

Here we have come across a typical situation related to the arbitrary constant: after integration, we introduced a constant C_1 which upon rearrangement turned into an expression e^{C_1} in the formula for the solution $y(x)$. Since C_1 is a constant, so is e^{C_1}. There is no need to keep this complicated form for an unknown constant, and we simply replaced the whole expression e^{C_1} with C in the final form. Applying the initial condition $y(1) = 1$, we write: $1 = Ce$ yielding $C = e^{-1}$. Hence, the particular integral reads $y = xe^{x-1}$.

Problem 8.1. *Obtain the general solutions of the following separable ODE:*

$$x\sqrt{1-y^2}dx + y\sqrt{1-x^2}dy = 0,$$

subject to the initial condition $y(1) = 0$. [*Answer*: $\sqrt{1-x^2} + \sqrt{1-y^2} = 1$.]

Problem 8.2. *Obtain the general solutions of*

$$x^2\frac{dy}{dx} = y^2\frac{dx}{dy}.$$

What its particular solution would be if $y(1) = 1$*?* [*Answer*: $y = x$ and $y = 1/x$.]

8.1.2 "Exact" Differential Equations

If the ODE can be written in the form

$$A(x,y)dx + B(x,y)dy = 0, \tag{8.3}$$

in which the left-hand side is the exact differential (see Sect. 5.5) of a function $F(x,y)$ of two variables x and y, i.e.

$$dF = \frac{\partial F}{\partial x}dx + \frac{\partial F}{\partial y}dy = A(x,y)dx + B(x,y)dy, \tag{8.4}$$

then this equation can be integrated and presented in an algebraic form $F(x,y) = C$ that contains no derivatives. As such, it has to be considered as a *solution* of the ODE even though this equation may be a transcendental one with respect to y. Of course, if the equation $F(x,y) = C$ can be solved for $y(x)$, then the solution is obtained explicitly. However, in many cases this may not be possible or such a representation is cumbersome or inconvenient, as illustrated by the examples below.

In order for the left-hand side of Eq. (8.3) to be the exact differential, we should obviously have:

$$\frac{\partial F}{\partial x} = A(x,y) \quad \text{and} \quad \frac{\partial F}{\partial y} = B(x,y), \tag{8.5}$$

which inevitably results in the necessary condition (see Sect. 5.5)

$$\frac{\partial B}{\partial x} = \frac{\partial A}{\partial y} \tag{8.6}$$

for the two functions $A(x,y)$ and $B(x,y)$ to satisfy. If this equation is satisfied, then the solution is obtained by integrating either of Eq. (8.5) by, e.g., the method explained in detail in Sect. 5.5. Briefly, consider equation $\partial F/\partial x = A(x,y)$ in which the variable y is fixed. It can be written as $dF = Adx$ (keeping in mind that y is fixed, we can temporarily use d instead of ∂ here), which is separable and hence can be integrated::

$$F(x,y) = \int A(x,y)dx + C_1(y), \tag{8.7}$$

where the "arbitrary constant" $C_1(y)$ may still (and most probably will) be a function of the other variable y. This unknown function $C_1(y)$ is then obtained from the other condition $\partial F/\partial y = B(x,y)$. Once the function $F(x,y)$ is obtained, one has to set it to a constant since $dF(x,y) = 0$ according to the DE (8.3).

We also note that the other method of solving for $F(x,y)$ is based on calculating the appropriate line integral, $\int Adx + Bdy$, see Sect. 6.3.4. Of course, both methods give identical results.

As an example of using the first method, consider the equation:

$$\frac{dy}{dx} = -\frac{2xy}{x^2+y^2}.$$

First of all, we rewrite this ODE in the form of Eq. (8.3):

$$2xydx + \left(x^2+y^2\right)dy = 0.$$

Next, we should check if the condition (8.6) is satisfied:

$$A(x,y) = 2xy, \;\; B(x,y) = x^2+y^2 \;\; \Longrightarrow \;\; \frac{\partial A}{\partial y} = 2x \;\text{ and }\; \frac{\partial B}{\partial x} = 2x,$$

i.e. the condition is indeed satisfied, and thus we can integrate the equation $\partial F/\partial x = A(x,y) = 2xy$ with respect to x to get

$$F = \int 2xydx = 2y\frac{x^2}{2} + C_1(y) = x^2y + C_1(y)$$

(note that y is a constant in this integral and hence must be considered as a parameter!). The unknown function $C_1(y)$ is obtained from $\partial F/\partial y = B(x,y) = x^2 + y^2$ as follows:

$$\frac{\partial F}{\partial y} = \frac{\partial}{\partial y}\left(x^2y + C_1(y)\right) = x^2 + \frac{dC_1}{dy} \;\; \Longrightarrow \;\; x^2 + \frac{dC_1}{dy} = x^2 + y^2 \;\; \Longrightarrow \;\; \frac{dC_1}{dy} = y^2,$$

that is easily integrated (since it is a separable equation) to give:

$$dC_1 = y^2dy \;\; \Longrightarrow \;\; C_1(y) = \int y^2dy = \frac{y^3}{3} + C,$$

with C being an arbitrary constant. Thus, the function $F(x, y)$ is now completely defined (up to the constant C, of course!) as

$$F(x, y) = x^2 y + \frac{y^3}{3} + C.$$

Since the total differential of $F(x, y)$, according to Eq. (8.3), is zero, the function F should be equal to a constant, say C_1. Therefore, we obtain:

$$F(x, y) = x^2 y + \frac{y^3}{3} + C = C_1,$$

i.e. the solution $y(x)$ is obtained by solving the following algebraic equation in y:

$$x^2 y + \frac{y^3}{3} = C_2,$$

where $C_2 = C_1 - C$ is an arbitrary constant to be determined from initial conditions or some other additional information available for $y(x)$. Note that the solution obtained can be checked by differentiating both sides of the equation above with respect to x (remember to treat y as a function of x):

$$\frac{d}{dx}\left(x^2 y + \frac{y^3}{3}\right) = 2xy + x^2\frac{dy}{dx} + \frac{3y^2}{3}\frac{dy}{dx} = 2xy + \left(x^2 + y^2\right)\frac{dy}{dx} = 0,$$

which is the original equation.

Problem 8.3. *Show that the following ODE can be solved using the method of exact differentials:*

$$\frac{dy}{dx} = -\frac{y + \sin(x)}{x - 2\cos(y)}$$

and then solve it. [*Answer*: $yx - \cos x - 2\sin y = C$.]

8.1.3 Method of an Integrating Factor

Sometimes, if an ODE (8.3) does not satisfy the condition (8.6), it can be multiplied by a function $I(x, y)$, called an *integrating factor*, so that with the replacements $A(x, y) \to I(x, y)A(x, y)$ and $B(x, y) \to I(x, y)B(x, y)$, the necessary condition

$$\frac{\partial}{\partial x}\left(I(x,y)B(x,y)\right) = \frac{\partial}{\partial y}\left(I(x,y)A(x,y)\right)$$

is satisfied and the original ODE is turned into an equation with the exact differential that can be solved using the method developed above. Note that the original ODE (8.3) can always be multiplied by any function in the left-hand side as we have zero in the right-hand side.

As an example of using this method, let us try to solve the following equation:

$$\frac{dy}{dx} = \frac{y}{x}.$$

This is a separable equation for which the solution is easily obtained:

$$\frac{dy}{y} = \frac{dx}{x} \implies \int \frac{dy}{y} = \int \frac{dx}{x} \implies \ln|y| = \ln|x| + \ln|C|$$
$$\implies \ln\left|\frac{y}{Cx}\right| = 0 \implies y = Cx.$$

However, it is instructive here to solve this equation with the method of the integrating factor. We write the equation in the unfolded form first as

$$ydx - xdy = 0,$$

with $A = y$ and $B = -x$, that obviously do not satisfy the necessary condition (8.6): $\partial A/\partial y = 1$, but $\partial B/\partial x = -1$. Let us try to find an integrating factor as $I(x,y) = x^{\alpha}y^{\beta}$. The new functions A and B become: $A(x,y) = x^{\alpha}y^{\beta+1}$ and $B(x,y) = -x^{\alpha+1}y^{\beta}$. The idea is to find such α and β that the newly defined functions A and B would satisfy the necessary condition:

$$\frac{\partial A}{\partial y} = (\beta+1)\,x^{\alpha}y^{\beta} \quad \text{should be equal to} \quad \frac{\partial B}{\partial x} = -\,(\alpha+1)\,x^{\alpha}y^{\beta}.$$

Comparing the two expressions, we get $\beta + 1 = -\alpha - 1$ or $\beta = -\alpha - 2$, a single condition for two constants to be determined, α and β; for instance, one can take $\alpha = -2$ and then getting $\beta = 0$, so that the required integrating factor becomes $I(x,y) = x^{-2}$. Now the equation becomes the one containing the exact differential:

$$x^{-2}ydx - x^{-1}dy = 0 \quad \text{since} \quad \frac{\partial}{\partial y}\left(x^{-2}y\right) = \frac{\partial}{\partial x}\left(-x^{-1}\right),$$

and hence can be integrated using the method of the previous subsection:

$$F(x,y) = \int x^{-2}ydx = y\frac{x^{-2+1}}{-2+1} + C_1(y) = -yx^{-1} + C_1(y)$$

and

$$\frac{\partial F}{\partial y} = \frac{\partial}{\partial y}\left(-yx^{-1} + C_1(y)\right) = -x^{-1} + \frac{dC_1}{dy} = -x^{-1},$$

so that $dC_1/dy = 0$ yielding $C_1(y) = C$, which is a constant. Therefore, $F(x, y) = -yx^{-1} + C$, and, since it must be a constant itself, this yields $-yx^{-1} = C_1$ or $y = -C_1 x$ with C_1 being another arbitrary constant. We have obtained the same general solution as above using this method.

The choice of an integrating factor in every case is a matter of some luck, but special methods have been developed for certain types of differential equations. For instance, the factor $I(x, y) = x^\alpha y^\beta$ used above may work if both original functions $A(x, y)$ and $B(x, y)$ have a polynomial form in x and y.

Problem 8.4. *Find the integrating factor for the DE*

$$-x\frac{dy}{dx} + y = 2x.$$

[*Hint*: assume that the integrating factor is a function of x only and obtain a separable ODE for it. *Answer*: the integrating factor may be chosen as $\mu(x) = x^{-2}$.]

Problem 8.5. *Find the integrating factor $I(x, y)$ to turn the ODE*

$$(1 - xy)\frac{dy}{dx} + y^2 + 3xy^3 = 0$$

into an exact DE. [*Hint*: Try $I(x, y) = x^\alpha y^\beta$ and determine α and β. There is only one possibility here. *Answer*: $\alpha = 0$, $\beta = -3$.]

Problem 8.6. *Solve the DE of the previous problem using the integrating factor $I(x, y) = y^{-3}$.* [*Answer*: $2xy + 3x^2y^2 - 1 = 2Cy^2$.]

8.1.4 Homogeneous Differential Equations

Some ODEs can be solved by a simple substitution. Specifically, consider ODEs of the form of Eq. (8.3) (not necessarily the exact differential equation!) in which functions $A(x, y)$ and $B(x, y)$ are *homogeneous functions* of the same degree. This means that if we scale x and y in both A and B by some λ, we will get the same factor of λ^n to come out:

$$A(\lambda x, \lambda y) = \lambda^n A(x, y) \quad \text{and} \quad B(\lambda x, \lambda y) = \lambda^n B(x, y). \tag{8.8}$$

If this is the case for the given equation, it can be solved by the substitution $y(x) \Rightarrow z(x) = y(x)/x$. Indeed, consider an ODE:

$$A(x,y)dx + B(x,y)dy = 0,$$

and let us make the substitution $y(x) = z(x)x$ in it:

$$dy = zdx + xdz, \; A(x,y) = A(x,zx) = x^n A(1,z) \text{ and } B(x,y) = B(x,zx) = x^n B(1,z),$$

where the role of λ was played by x in both $A(x,zx)$ and $B(x,zx)$. Using these results, we are now able to perform simple manipulations in the original ODE to obtain an equation which is separable:

$$\begin{aligned} & x^n A(1,z)dx + x^n B(1,z)(zdx + xdz) = 0 \\ \Longrightarrow \quad & A(1,z)dx + B(1,z)(zdx + xdz) = 0 \\ \Longrightarrow \quad & (B(1,z)z + A(1,z))\,dx + xB(1,z)dz = 0 \\ \Longrightarrow \quad & \frac{dx}{x} = -\frac{B(1,z)}{zB(1,z) + A(1,z)}dz, \end{aligned}$$

i.e. it is indeed separable now and hence can be integrated, at least, in principle.

As an example, consider equation

$$\frac{dy}{dx} = \frac{x^2 y}{x^3 + y^3}.$$

In the unfolded form the equation becomes

$$-x^2 y dx + \left(x^3 + y^3\right) dy = 0.$$

It is seen that both functions $A(x,y) = -x^2 y$ and $B(x,y) = x^3 + y^3$ are homogeneous functions of the order 3:

$$\begin{aligned} A(\lambda x, \lambda y) &= -(\lambda x)^2 (\lambda y) = \lambda^3 \left(-x^2 y\right) = \lambda^3 A(x,y), \\ B(\lambda x, \lambda y) &= (\lambda x)^3 + (\lambda y)^3 = \lambda^3 \left(x^3 + y^3\right) = \lambda^3 B(x,y), \end{aligned}$$

and thus the substitution $z = y/x$ should be appropriate: $y = zx$ and $dy/dx = (dz/dx)\,x + z$, so that

$$\frac{dy}{dx} = \frac{x^2 y}{x^3 + y^3} \quad \Longrightarrow \quad \frac{dz}{dx}x + z = \frac{x^3 z}{x^3 + z^3 x^3} \quad \Longrightarrow \quad \frac{dz}{dx}x + z = \frac{z}{1+z^3},$$
$$\Longrightarrow \quad \frac{dz}{dx}x = \frac{z}{1+z^3} - z \quad \Longrightarrow \quad \frac{1+z^3}{z^4}dz = -\frac{dx}{x},$$

which is integrated to give

$$\int \frac{1+z^3}{z^4} dz = -\int \frac{dx}{x} \Longrightarrow -\frac{1}{3z^3} + \ln|z| = -\ln|x| - \ln|C| = \ln\left|\frac{1}{Cx}\right|$$
$$\Longrightarrow -\frac{1}{3z^3} = \ln\left|\frac{1}{Czx}\right| .$$

Substituting back $z = y/x$, we obtain:

$$-\frac{x^3}{3y^3} = \ln\left|\frac{1}{Cy}\right| \quad \text{or} \quad -\frac{x^3}{3y^3} = -\ln|y| + C_1 ,$$

where C_1 is an arbitrary constant. The obtained solution is a transcendental algebraic equation with respect to y. In spite of the fact that this equation cannot be solved analytically with respect to y (it contains both the logarithm of y and its cube), it serves as the solution of the ODE since it does not contain anymore derivatives of y. Note that after applying the corresponding initial conditions to determine the constant C_1, this equation can easily be solved numerically by plotting for each value of x the functions of y which appear in the left- and right-hand sides and looking for the point y where they intersect.

Problem 8.7. *Solve the following homogeneous DE:*

$$x^2 dy + (y^2 - xy)dx = 0.$$

[*Answer*: $y = x/ (C + \ln|x|)$.]

Problem 8.8. *Show that the ODE*

$$\frac{dy}{dx} = \frac{xy}{x^2 - y^2}$$

is of the homogeneous type and then solve it. [*Answer*: $y(x)$ is determined from the equation $-x^2/2y^2 = \ln|y| + C$.]

8.1.5 Linear First Order Differential Equations

Consider the following first order first degree differential equation

$$\frac{dy}{dx} + p(x)y = q(x) \ , \quad a < x < b, \tag{8.9}$$

where $p(x)$ and $q(x)$ are some functions of x. We shall show that this equation can be solved for any functions p and q using the method of an integrating factor of Sect. 8.1.3.

Indeed, our ODE, if rewritten in the required (unfolded) form (8.3) as

$$(-q(x) + p(x)y)\, dx + dy = 0,$$

obviously does not satisfy the condition (8.6) for arbitrary $p(x)$ and $q(x)$. Consider then an integrating factor $I(x, y)$ that turns our equation into:

$$I(x, y)\, (-q(x) + p(x)y)\, dx + I(x, y)dy = 0,$$

and let us choose $I(x, y)$ in such a way that the condition (8.6) is satisfied:

$$\frac{\partial}{\partial y}\left[I(x, y)\, (-q(x) + p(x)y)\right] = \frac{\partial}{\partial x} I(x, y)$$
$$\Longrightarrow \quad \frac{\partial I(x, y)}{\partial y}\, (-q(x) + p(x)y) + p(x)I(x, y) = \frac{\partial I(x, y)}{\partial x}.$$

It is seen that if the function $I(x, y)$ depends only on x, then the first term in the left-hand side disappears completely and we get a closed equation for $I(x, y) = I(x)$:

$$p(x)I(x) = \frac{dI(x)}{dx} \Longrightarrow p(x)dx = \frac{dI}{I} \Longrightarrow \int p(x)dx = \int \frac{dI}{I}$$
$$\Longrightarrow \int p(x)dx = \ln |I|\,,$$

so that

$$I(x) = \exp\left[\int p(x)dx\right]. \tag{8.10}$$

The integral here is an *indefinite* integral which is considered as a function of x.

Once the integrating factor has been determined, we can solve the original equation as explained in Sect. 8.1.2. We have for the new functions A and B:

$$A(x, y) = I(x)\, (-q(x) + p(x)y) \quad \text{and} \quad B(x, y) = I(x),$$

so that

$$F(x, y) = \int B(x, y)dy = I(x)y + C_1(x),$$
$$\frac{\partial F(x, y)}{\partial x} = \frac{dI}{dx} y + \frac{dC_1}{dx} = A(x, y) = I(x)\, (-q(x) + p(x)y)\,.$$

However, the function $I(x)$ satisfies $p(x)I(x) = dI/dx$, so that, after simplification, we get:

$$\frac{dC_1}{dx} = -I(x)q(x),$$

which is in a separable form and hence accepts the solution (again, via an indefinite integral):

$$C_1(x) = -\int I(x)q(x)dx + C_2.$$

Thus, we have found the function completely:

$$F(x, y) = I(x)y + C_1(x) = I(x)y - \int I(x)q(x)dx + C_2.$$

Since it must be a constant anyway, we can write:

$$I(x)y - \int I(x)q(x)dx + C_2 = C_3 \quad \Longrightarrow \quad I(x)y - \int I(x)q(x)dx = C,$$

which can be solved for y to get (C is the final arbitrary constant):

$$y = I(x)^{-1}\left[C + \int I(x)q(x)dx\right]. \tag{8.11}$$

Problem 8.9. *Check that if each integral in the above formula is understood as a definite integral with the upper limit being the corresponding variable, the bottom limits do not matter.*

As an example, let us solve the following ODE

$$\frac{dy}{dx} + y = e^x$$

subject to the condition $y(0) = 1$. Here $p(x) = 1$ and $q(x) = e^x$. Therefore, the function $I(x)$ from Eq. (8.10) is simply

$$I(x) = e^{\int dx} = e^x,$$

and thus the solution is obtained as

$$y(x) = e^{-x}\left[C + \int e^x e^x dx\right] = Ce^{-x} + e^{-x}\int e^{2x}dx = Ce^{-x}$$
$$+ e^{-x}\frac{1}{2}e^{2x} = Ce^{-x} + \frac{1}{2}e^x,$$

which is the required solution (can be checked by a direct substitution).

Problem 8.10. *Obtain the general solutions of the following linear ODEs:*

$$(a)\; y' + y\cos(x) = \frac{1}{2}\sin(2x)\,; \quad (b)\; \frac{dy}{dx} + \frac{2}{x}y = 2 + x^2.$$

[*Answer*: (a) $y = C\exp(-\sin x) + \sin x - 1$; (b) $y(x) = 2x/3 + x^3/5 + C/x^2$.]

The solution of the general linear first order ODE (8.9) can also be obtained using the *method of variation of parameters* which we shall find especially useful in a more general case of linear second order ODE to be considered in Sect. 8.2.2.2. The idea of this method is to first solve the *homogeneous* equation (without the right-hand side):

$$\frac{dy}{dx} + p(x)y = 0 \implies \frac{dy}{y} = -pdx \implies y(x) = C\exp\left(-\int p(x)dx\right) = C\,I(x)^{-1}, \tag{8.12}$$

where C is an arbitrary constant. To solve the *inhomogeneous* equation, i.e. with the right-hand side, i.e. Eq. (8.9), we seek the solution in the form (8.12), but with the constant C replaced by a function $C(x)$:

$$y(x) = C(x)\exp\left(-\int p(x)dx\right) = C(x)I(x)^{-1}.$$

We need to substitute this trial solution into the ODE (8.9) to find an equation for the unknown function $C(x)$. This gives:

$$\left[C'(x)I(x)^{-1} - C(x)I(x)^{-1}p(x)\right] + p(x)\left[C(x)I(x)^{-1}\right] = q(x),$$

where we made use of the fact that $I(x)^{-1}$ satisfies the equation

$$\frac{d}{dx}I(x)^{-1} = -p(x)I(x)^{-1},$$

which follows from Eq. (8.10). Note that when differentiating the exponential term, we treated the integral as $\int^x p(x')\,dx'$, i.e. with the x appearing in the upper limit, and hence its derivative is $p(x)$. After simplification, we obtain:

$$C'(x)I(x)^{-1} = q(x) \implies C'(x) = q(x)I(x) \implies C(x) = \int q(x)I(x)dx + C_1,$$

where C_1 is an arbitrary constant. It is easy to see now that $y(x) = C(x)/I(x)$ is exactly the same as in Eq. (8.11).

8.2 Linear Second Order Differential Equations

Consider a general n-th order linear differential equation:

$$a_n(x)\frac{d^n y}{dx^n} + a_{n-1}(x)\frac{d^{n-1}y}{dx^{n-1}} + \cdots + a_1(x)\frac{dy}{dx} + a_0(x)y = f(x). \tag{8.13}$$

This equation contains a function $f(x)$ in the right-hand side depending only on x, and is called *inhomogeneous*. *Homogeneous* equations have zero instead.

A very general property of this equation is that it is *linear*. This has a very profound effect on the solution. Indeed, if we know two solutions $y_1(x)$ and $y_2(x)$ of this equation, then any of their linear combinations with *arbitrary* constants C_1 and C_2, i.e.

$$y(x) = C_1 y_1(x) + C_2 y_2(x),$$

will also be a solution. Generally, any n-th order ODE like (8.13) has n linearly independent solutions $y_1(x), \ldots, y_n(x)$, and a *general solution*

$$y(x) = C_1 y_1(x) + C_2 y_2(x) + \cdots + C_n y_n(x) = \sum_{i=1}^{n} C_i y_i(x) \tag{8.14}$$

is constructed as their linear combination with arbitrary coefficients C_1, C_2, etc. These are determined using additional information about the solution, e.g. *initial conditions,* i.e. known values of $y(0), y^{(1)}(0), \ldots, y^{(n-1)}(0)$, where $y^{(i)}(x) = d^i y/dx^i$.

Problem 8.11. *Prove that the linear combination (8.14) of the solutions of the n-th order ODE (8.13) is also its solution.*

Problem 8.12. *Consider a general linear second order ODE*

$$a_2(x)y'' + a_1(x)y' + a_0(x)y = 0. \tag{8.15}$$

Show that the function $f(x)$ which brings this ODE into the form $(a_2 f y')' + a_0 f y = 0$ satisfies the ODE $f'/f = a_1/a_2 - a_2'/a_2$, and obtain its solution. [*Answer*: $f(x) = \exp\left(\int (a_1/a_2)\, dx\right)/a_2$.]

From now on we shall only be concerned with linear *second order* ODEs. There must be two linearly independent solutions of the homogenous linear second order DE. One important general result that corresponds to any of these is that the second solution of the equation can be always found if the first solution is known.

Theorem 8.1. *The second solution $y_2(x)$ of a linear second order ODE (8.15) can always be found if the first solution $y_1(x)$ is known.*

Proof. Since both y_1 and y_2 satisfy the ODE (8.15), we have

$$a_2 y_1'' + a_1 y_1' + a_0 y_1 = 0 \text{ and } a_2 y_2'' + a_1 y_2' + a_0 y_2 = 0.$$

Next, we multiply the first equation by y_2 and the second by y_1, and subtract the resulting equations from each other:

$$\begin{aligned} y_2\left(a_2 y_1'' + a_1 y_1' + a_0 y_1\right) - y_1\left(a_2 y_2'' + a_1 y_2' + a_0 y_2\right) &= 0, \\ \Longrightarrow \quad a_2\left(y_2 y_1'' - y_1 y_2''\right) + a_1\left(y_2 y_1' - y_1 y_2'\right) &= 0. \end{aligned} \tag{8.16}$$

Consider now the determinant (called *Wronskian*)

$$W = \begin{vmatrix} y_1 & y_2 \\ y_1' & y_2' \end{vmatrix} = y_1 y_2' - y_2 y_1'. \tag{8.17}$$

Its derivative

$$W' = \left(y_1' y_2' + y_1 y_2''\right) - \left(y_2' y_1' + y_2 y_1''\right) = y_1 y_2'' - y_2 y_1''.$$

Therefore, Eq. (8.16) can now be rewritten as the first order ODE with respect to the function $W(x)$:

$$a_2(x)\frac{dW(x)}{dx} + a_1(x)W(x) = 0\ ,$$

which is separable and thus can be easily integrated:

$$\frac{dW}{W} = -\frac{a_1(x)}{a_2(x)}dx \Longrightarrow \int \frac{dW}{W} = -\int \frac{a_1(x)}{a_2(x)}dx \Longrightarrow \ln|W(x)| = -\int \frac{a_1(x)}{a_2(x)}dx,$$

where an indefinite integral has been used. Hence, the Wronskian becomes

$$W(x) = \exp\left(-\int \frac{a_1(x)}{a_2(x)}dx\right). \tag{8.18}$$

This result is called *Abel's formula*.

Once the function $W(x)$ has been calculated, it is now possible to consider Eq. (8.17) as a first order ODE for $y_2(x)$:

$$y_2'(x) - \frac{y_1'(x)}{y_1(x)} y_2(x) = \frac{W(x)}{y_1(x)}.$$

This ODE is exactly of the type (8.9) we considered before in Sect. 8.1.5, and its solution is given by Eq. (8.11) with $p(x) = -y_1'(x)/y_1(x)$ and $q(x) = W(x)/y_1(x)$.

This way of finding y_2, however, appears a bit tricky as one has to deal carefully with the modulus of $y_1(x)$. A much simpler and straightforward way of obtaining $y_2(x)$ is to try a substitution $y_2(x) = u(x)y_1(x)$ directly in the above equation:

$$\left(u'y_1 + uy_1'\right) - \frac{y_1'}{y_1}(uy_1) = \frac{W}{y_1} \quad \Longrightarrow \quad u'y_1 = \frac{W}{y_1} \quad \Longrightarrow \quad u'(x) = \frac{W(x)}{y_1^2(x)},$$

which, after integration, gives immediately

$$u(x) = \int \frac{W(x)}{y_1^2(x)} dx. \tag{8.19}$$

This should be understood as an indefinite integral without an arbitrary constant: this is because $y_2 = uy_1$ and a constant C in $u(x)$ would simply give an addition of Cy_1 in y_2 that will be absorbed in the general solution $y = C_1 y_1 + C_2 y_2$. Hence,

$$y_2(x) = u(x)y_1(x) = y_1(x) \int \frac{W(x)}{y_1^2(x)} dx = y_1(x) \int \frac{1}{y_1^2(x)} \exp\left(-\int \frac{a_1(x)}{a_2(x)} dx\right) dx. \tag{8.20}$$

Thus, if one solution of the second order linear ODE is known, the other one can always be obtained using the above formula. **Q.E.D.**

As an example, let us find the second solution of the equation

$$y'' + y = 0,$$

if it is known that the first one is $y_1(x) = \sin x$. Here $a_2 = 1$, $a_1 = 0$ and $a_0 = 1$, so that, according to Eq. (8.18), $W(x) = 1$ and, using Eq. (8.20), we get:

$$y_2 = \sin x \int \frac{dx}{\sin^2 x} = \sin x \,(-\cot x) = -\cos x,$$

which, as can be easily checked, is the correct solution.[1]

[1]The minus sign does not matter as in the general solution y_2 will be multiplied by an arbitrary constant anyway.

Problem 8.13. *Consider the following linear second order ODE:*

$$y'' + 4y' + 4y = 0, \tag{8.21}$$

which first solution is $y_1 = e^{-2x}$. *Show that the second solution is* $y_2 = xy_1 = xe^{-2x}$.

Problem 8.14. *Consider a one-particle one-dimensional Schrödinger equation*

$$-\frac{\hbar^2}{2m}\psi''(x) + V(x)\psi(x) = E\psi(x),$$

where $\psi(x)$ *is the particle wavefunction,* $V(x)$ *external potential,* m *particle mass, and* $\hbar$ *Planck's constant. Consider a discrete energy spectrum* E *for which* $\psi(\pm\infty) = 0$, *meaning that the particle is localised in some region of space (as opposite to the continuum spectrum when there is a non-zero probability to find the particle anywhere in space). Show that the energies* E *in the discrete spectrum cannot be degenerate, i.e. there can only be a single state* $\psi(x)$ *for each energy* E. [*Hint*: prove by contradiction, assuming there are two such solutions, ψ_1 and ψ_2; then, argue that ψ''/ψ for both should be the same; then use integration by parts.]

8.2.1 Homogeneous Linear Differential Equations with Constant Coefficients

We have learned in the previous section that the first order linear ODEs always allow for a general solution. However, it is not possible to obtain such a solution in the cases of higher-order ODEs containing variable coefficients. One specific case that allows for a general solution is that of a higher-order ODE with *constant coefficients*. We shall again limit ourselves with the ODEs of the second order; higher order ODEs can be considered along the same lines.

We shall consider equations of the form:

$$a_2\frac{d^2y}{dx^2} + a_1\frac{dy}{dx} + a_0y = 0, \tag{8.22}$$

where a_0, a_1 and a_2 are some constant coefficients. This equation is generally solved using an exponential trial solution e^{px}. Indeed, substituting this trial function into the ODE above, we have:

$$\left(a_2p^2 + a_1p + a_0\right)e^{px} = 0.$$

Since we want this equation to be satisfied for any x, we must have p to satisfy the following *characteristic equation*:

$$a_2 p^2 + a_1 p + a_0 = 0, \tag{8.23}$$

that yields either one or two solutions for p.

If there are *two (different)* solutions p_1 and p_2, then two functions $e^{p_1 x}$ and $e^{p_2 x}$ satisfy the same equation; note that the two functions are linearly independent in this case. Since the equation is linear, their linear combination would also satisfy it:

$$y = C_1 e^{p_1 x} + C_2 e^{p_2 x}.$$

Therefore, this must be the required general solution of the homogeneous equation (8.22).

Let us now consider some simple examples. To solve the equation

$$\frac{d^2 y}{dx^2} - 8\frac{dy}{dx} + 12y = 0,$$

the trial solution e^{px} is substituted into the equation to give $p^2 - 8p + 12 = 0$. Hence, two solutions, $p_1 = 2$ and $p_2 = 6$, are obtained for p. Hence the general solution is

$$y(x) = C_1 e^{2x} + C_2 e^{6x}.$$

To solve the equation

$$\frac{d^2 y}{dx^2} + 4y = 0,$$

we substitute the trial solution e^{px} into the equation to get $p^2 + 4 = 0$, which has two complex solutions: $p_1 = 2i$ and $p_2 = -2i$. Therefore, a general solution is

$$y(x) = C_1 e^{i2x} + C_2 e^{-i2x}.$$

This can actually be rewritten via sine and cosine functions using the Euler's identities:

$$\begin{aligned} y(x) &= C_1 \left[\cos(2x) + i\,\sin(2x)\right] + C_2 \left[\cos(2x) - i\,\sin(2x)\right] \\ &= A_1 \cos(2x) + A_2 \sin(2x), \end{aligned}$$

where $A_1 = C_1 + C_2$ and $A_2 = i\,(C_1 - C_2)$ are two new arbitrary (and in general complex) constants.

Let us now solve the equation

$$\frac{d^2 y}{dx^2} - 4\frac{dy}{dx} + 8y = 0.$$

The corresponding equation for p is $p^2-4p+8=0$, that has two complex solutions: $p_1=2+2i$ and $p_2=2-2i$. Therefore, a general solution is

$$\begin{aligned} y(x) &= C_1 e^{(2+2i)x} + C_2 e^{(2-2i)x} = e^{2x}\left(C_1 e^{i2x} + C_2 e^{-i2x}\right) \\ &= e^{2x}\left[A_1\cos(2x) + A_2\sin(2x)\right]. \end{aligned}$$

If, however, there is only *one* root in Eq. (8.23), i.e. $p_1 = p_2$ (the case of repeated roots), then only one solution of the ODE can be immediately written as an exponential function, $y_1 = \exp(p_1 x)$. To find the second solution, we recall that, as was found in Theorem 8.1, the second solution can always be found from the first via Eq. (8.20) in the form of $y_2(x) = u(x)y_1(x)$, where the function $u(x)$ is to be determined from Eq. (8.19). The Wronskian (8.18) in this case is

$$W(x) = \exp\left(-\frac{a_1}{a_2}\int dx\right) = \exp\left(-\frac{a_1}{a_2}x\right).$$

Since the roots of the characteristic equation (8.23) are the same, we must have $a_1^2 = 4a_0a_2$ and hence $p_1 = -a_1/2a_2$, i.e. $W(x) = \exp(2p_1x)$. Using this now in Eq. (8.19), we easily obtain:

$$u(x) = \int \frac{W}{y_1^2}dx = \int \frac{e^{2p_1x}}{e^{2p_1x}}dx = x,$$

and hence the second solution is to be taken in the form: $y_2 = xy_1 = x\exp(p_1x)$.

Problem 8.15. *Consider equation*

$$y'' - 4y' + 4y = 0,$$

which has repeated roots $p=2$ in the characteristic equation $p^2-4p+4=0$, i.e. $y_1(x) = e^{2x}$. To obtain the second solution, try directly the substitution $y = u(x)e^{2x}$ in the original DE. Show that $u''=0$, and hence $u(x) = A_1x + A_2$. Finally, argue that it is sufficient to take $y_2(x) = xe^{2x}$ as the second independent solution.

Problem 8.16. *Consider the case of two distinct but very close roots of the characteristic equation: p_1 and $p_2 = p_1 + \delta$. Prove that the same result for the second solution y_2 is obtained in the limit of $\delta \to 0$.*

Problem 8.17. *Consider the case of two distinct roots $p_1 \neq p_2$ of the characteristic equation. Show explicitly using Theorem 8.1 that from $y_1 = e^{p_1x}$ follows $y_2 = e^{p_2x}$.*

Problem 8.18. *Obtain the general solutions of:*

$$(a)\;\; y'' - 4y = 0; \quad (b)\;\; y'' - 6y' + 9y = 0.$$

[*Answer*: (a) $y = C_1 e^{2x} + C_2 e^{-2x}$; (b) $y = (C_1 x + C_2)\, e^{3x}$.]

8.2.2 Inhomogeneous Linear Differential Equations

Here we shall consider inhomogeneous equation (8.13), i.e. with non-zero right-hand side when the function $f(x) \neq 0$. Whatever the order of the ODE and the form of the functions-coefficients $a_n(x), \ldots, a_0(x)$, the construction of the general solution of the equation is based on the following theorem:

Theorem 8.2. *The general solution of a linear inhomogeneous equation is given by*

$$y(x) = y_c(x) + y_p(x), \tag{8.24}$$

where

$$y_c(x) = C_1 y_1(x) + \cdots + C_n y_n(x)$$

is the general solution of the corresponding homogeneous equation, called the complementary solution*, and* $y_p(x)$ *is the so-called* particular solution *(integral) of the inhomogeneous equation.*

Proof. Let us start by introducing a shorthand notation for the left-hand side of Eq. (8.13) by defining an *operator*

$$\mathcal{L} = a_n(x)\frac{d^n}{dx^n} + a_{n-1}(x)\frac{d^{n-1}}{dx^{n-1}} + \cdots + a_1(x)\frac{d}{dx} + a_0(x), \tag{8.25}$$

which, when acting on a function $g(x)$ standing on the *right* of it, produces the following action:

$$\mathcal{L}g(x) = a_n(x)\frac{d^n g}{dx^n} + a_{n-1}(x)\frac{d^{n-1} g}{dx^{n-1}} + \cdots + a_1(x)\frac{dg}{dx} + a_0(x)g.$$

The operator $\mathcal{L}$ is linear, i.e.

$$\mathcal{L}\left(c_1 g_1 + c_2 g_2\right) = c_1 \mathcal{L} g_1 + c_2 \mathcal{L} g_2, \tag{8.26}$$

where $g_1(x)$ and $g_2(x)$ are two arbitrary functions. Then, Eq. (8.13) can be rewritten in the following short form:

$$\mathcal{L} y = f(x). \tag{8.27}$$

After these notations, we see that the complementary and particular solutions are to satisfy two different equations:

$$\mathcal{L} y_c = 0 \ \text{ and } \ \mathcal{L} y_p = f(x). \tag{8.28}$$

If y is any other solution of the inhomogeneous ODE, then

$$\mathcal{L}\left(y - y_p\right) = \mathcal{L} y - \mathcal{L} y_p = f(x) - f(x) = 0, \tag{8.29}$$

i.e. $y - y_p$ must be the general solution y_c of the homogeneous equation $\mathcal{L} y_c = 0$, i.e. any solution has the form $y = y_c + y_p$. Note that in the first passage in Eq. (8.29) we have made use of the fact that the operator $\mathcal{L}$ is linear. **Q.E.D.**

Let us find a general solution of the equation

$$y'' + y = e^x.$$

According to the theorem proven above, we need first to obtain the complementary solution $y_c(x)$ which is the general solution of the homogeneous equation $y'' + y = 0$. This task must be trivial and yields $y_c = C_1 \cos x + C_2 \sin x$. Trying a particular solution in the form $y_p = Ae^x$ with some unknown constant A, we get after substitution in the equation:

$$Ae^x + Ae^x = e^x \implies 2A = 1 \implies A = \frac{1}{2} \implies y_p = \frac{1}{2}e^x,$$

so that the general solution is

$$y = y_c + y_p = C_1 \cos x + C_2 \sin x + \frac{1}{2}e^x.$$

Note that if we take two specific solutions, say, $2\sin x + \frac{1}{2}e^x$ and $-3\cos x + \frac{1}{2}e^x$, obtained by choosing particular values for the arbitrary constants C_1 and C_2, then the difference $2\sin x + 3\cos x$ of the two solutions is in fact a complementary solution of the homogeneous equation with the particular choice of the constants $C_1 = 3$ and $C_2 = 2$, as expected.

How can one choose the particular solution? We can see from the above example that a wise choice might be to try the form similar to that of the function $f(x)$ itself. In the case of a linear equation with constant coefficients this so-called *method of undetermined coefficients* may be successful in some cases and will be considered next.

8.2.2.1 Method of Undetermined Coefficients

Consider a second order linear inhomogeneous ODE with constant coefficients:

$$a_2y'' + a_1y' + a_0y = f(x).$$

Several cases of $f(x)$ can be considered:

- If $f(x)$ is a polynomial $P^{(n)}(x) = b_nx^n + \cdots + b_1x + b_0$ of the order n, then the particular solution should be sought in the form of a polynomial of the same order with unknown coefficients that must be determined by substituting it into the equation and comparing terms with the like powers of x on both sides.
- If $f(x) = A\cos(\omega x) + B\sin(\omega x)$ is a linear combination of trigonometric functions, try $y_p(x) = A_1\cos(\omega x) + B_1\sin(\omega x)$ in the same form and with identical ω, but with some unknown coefficients A_1 and B_1. These unknown coefficients are obtained by substitution into the ODE and comparing coefficients to $\sin(\omega x)$ and $\cos(\omega x)$ on both sides. Note that even if $A = 0$ or $B = 0$ in $f(x)$, still both sine and cosine terms must be present in $y_p(x)$.
- If $f(x)$ is an exponential $e^{\lambda x}$, then one has to use $y_p(x) = Ae^{\lambda x}$ as the particular solution with some constant A to be determined by substitution into the equation.
- If, more generally, $f(x)$ is given as a combination of all these functions, i.e.

  $$f(x) = P^{(n)}(x)e^{\lambda x}\left[A\cos(\omega x) + B\sin(\omega x)\right],$$

  then the particular solution is sought in a similar form keeping the same λ and ω, but with different coefficients in place of A, B and those in the polynomial; the coefficients are then obtained by substituting into the differential equation and comparing coefficients to the like powers of x and sine and cosine (the exponential function will cancel out).

Also note that if $f(x)$ is a sum of functions, particular integrals could be found individually for each of them; the final particular integral will be given as a sum of all these individual ones.

Examples below should illustrate the idea of the method.

Example 8.1. ► Solve the equation

$$y'' - 5y' + 6y = 1 + x.$$

Solution. The complementary solution reads $y_c = C_1 e^{2x} + C_2 e^{3x}$. Since the right-hand side is the first order polynomial, the particular integral is sought as a first order polynomial as well, $y_p = Ax + B$. Substituting it into the equation and comparing coefficients to the same powers of x, gives:

$$0 - 5A + 6(Ax + B) = 1 + x \implies \begin{cases} 6A = 1 \\ -5A + 6B = 1 \end{cases} \implies \begin{cases} A = 1/6 \\ B = 11/36 \end{cases},$$

so that the general solution

$$y = C_1 e^{2x} + C_2 e^{3x} + \frac{1}{6}\left(x + \frac{11}{6}\right). \blacktriangleleft$$

Example 8.2. ► Find the particular solution of the equation

$$y'' - 5y' + 6y = \sin(4x) + 3\cos x.$$

Solution. We must find the particular solutions individually for each of the terms in the right-hand side since they have different frequencies. For $\sin(4x)$ we try the function $A\sin(4x) + B\cos(4x)$, that gives upon substitution:

$$\begin{aligned} &[-16A\sin(4x) - 16B\cos(4x)] - 5[4A\cos(4x) - 4B\sin(4x)] \\ &+6[A\sin(4x) + B\cos(4x)] = \sin(4x), \end{aligned}$$

that, after comparing pre-factors to the sine and cosine functions, yields

$$\begin{cases} -16A + 20B + 6A = 1 \\ -16B - 20A + 6B = 0 \end{cases} \implies \begin{cases} A = -1/50 \\ B = 2/50 \end{cases},$$

so that the first part of the particular solution is

$$y_{p1} = -\frac{1}{50}\sin(4x) + \frac{2}{50}\cos(4x).$$

The second part is obtained by trying $C\sin x + D\cos x$ that similarly gives:

$$[-C\sin x - D\cos x] - 5[C\cos x - D\sin x] + 6[C\sin x + D\cos x] = 3\cos x,$$

$$\begin{cases} -C + 5D + 6C = 0 \\ -D - 5C + 6D = 3 \end{cases} \implies \begin{cases} C = -3/10 \\ D = 3/10 \end{cases} \implies y_{2p} = -\frac{3}{10}\sin x + \frac{3}{10}\cos x.$$

Therefore, the particular solution of the whole equation is

$$y_p = -\frac{1}{50}\sin(4x) + \frac{2}{50}\cos(4x) - \frac{3}{10}\sin x + \frac{3}{10}\cos x,$$

which can be checked by direct substitution. ◄

Example 8.3. ► Find the particular solution of the equation:

$$y'' - 5y' + 6y = 5xe^x.$$

Solution. Since the right-hand side is a product of the first order polynomial and the exponential function, we try $y_p=(Ax+B)\,e^x$. Substituting into the equation, this gives:

$$\begin{aligned}
&(Ax + 2A + B)\,e^x - 5\,(Ax + A + B)\,e^x + 6\,(Ax + B)\,e^x = 5xe^x \\
&\Longrightarrow \quad (Ax + 2A + B) - 5\,(Ax + A + B) + 6\,(Ax + B) = 5x \\
&\Longrightarrow \quad \begin{cases} A - 5A + 6A = 5 \\ 2A + B - 5\,(A + B) + 6B = 0 \end{cases} \Longrightarrow \begin{cases} A = 5/2 \\ B = 15/4 \end{cases} \\
&\Longrightarrow \quad y_p = \left(\frac{5}{2}x + \frac{15}{4}\right)e^x \quad ◄
\end{aligned}$$

There is a special case to be also considered when $f(x)$ happens to be *related* to one (or even both) solutions of the homogeneous equation (single or double "resonance"). In these cases the method should be modified.

Example 8.4. ► Find the particular solution of the equation:

$$y'' - 5y' + 6y = e^{2x}.$$

Solution. If we try $y_p = Ae^{2x}$ following the function type in the right-hand side, then we get zero in the left-hand side since e^{2x}, as can easily be checked, is a solution of the homogeneous equation. Obviously, this trial solution does not work. However, the required modification is actually very simple: if we multiply y_p tried above by x, it will work. Indeed, try $y_p = Axe^{2x}$, then, after substitution into the equation, we obtain:

$$\begin{aligned}
&(4A + 4Ax)\,e^{2x} - 5\,(A + 2Ax)\,e^{2x} + 6Axe^{2x} = e^{2x}, \\
&\qquad 4A + 4Ax - 5A - 10Ax + 6Ax = 1 \quad \text{or } A = -1,
\end{aligned}$$

so that $y_p = -xe^{2x}$ in this case. ◄

The considered case corresponds to a single "resonance" as the exponent in the right-hand side coincides with that of only *one* of the solutions of the homogeneous equation. If however *both* complementary exponents are the same, $p_1 = p_2$, and the same exponent happens to be present in the right-hand side, we have the case of a double "resonance". In that case the y_p constructed according to the rules stated above should be additionally multiplied by x^2 instead.

Example 8.5. ► Find the particular solution of the equation:

$$y'' - 4y' + 4y = xe^{2x}.$$

Solution. The complementary solution $y_c = C_1 e^{2x} + C_2 x e^{2x}$, and we see that the same exponent is used in the right-hand side. Substituting $y_p = (Ax + B)\, e^{2x}$ constructed according to usual rules gives zero in the left-hand side as it is basically the same as y_c; not good. Let us now try $y_p = x\,(Ax + B)\, e^{2x}$ instead, which gives:

$$\begin{aligned}&\left[2A + 4\,(2Ax + B) + 4\left(Ax^2 + Bx\right)\right] e^{2x} - 4\left[(2Ax + B) + 2\left(Ax^2 + Bx\right)\right] e^{2x}\\ &\quad + 4\left(Ax^2 + Bx\right) e^{2x} = xe^{2x} \quad \Longrightarrow \quad \left[2A + 4\,(2Ax + B) + 4\left(Ax^2 + Bx\right)\right] e^{2x}\\ &\quad - 4\left[(2Ax + B) + 2\left(Ax^2 + Bx\right)\right] e^{2x} + 4\left(Ax^2 + Bx\right) e^{2x} = xe^{2x}.\end{aligned}$$

Comparing coefficients to x^0, we get $A = 0$, good; however, coefficients to x^1 all cancel out in the left-hand side leading to an impossible $0 = 1$, i.e. we obtained a contradiction. Now let us try $y_p = x^2\,(Ax + B)\, e^{2x}$, which results in (after cancelling on e^{2x}):

$$\begin{aligned}&\left[4Ax^3 + 4\,(3A + B)\, x^2 + 2\,(3A + 4B)\, x + 2B\right]\\ &\quad -4\left[2Ax^3 + (3A + 2B)\, x^2 + 2Bx\right] + 4\left(Ax^3 + Bx^2\right) = x.\end{aligned}$$

The terms containing x^3 and x^2 in the left-hand side cancel out; the terms with x yield $A = 1/6$, while the terms with x^0 result in $B = 0$, i.e. $y_p = \left(x^3/6\right) e^{2x}$, which as easily checked by the direct substitution is the correct particular integral of this ODE. ◄

Problem 8.19. *Find the particular integral solutions of the ODE* $y'' + 3y' + 2y = f(x)$ *with the following right-hand side function* $f(x)$*:*

$$(a)\;\; 6 + x^2; \quad (b)\;\; 2e^{3x}; \quad (c)\;\; 6\sin x; \quad (d)\;\; 6e^{-3x}\sin x.$$

[*Answer*: (a) $y_p = 19/4 - 3x/2 + x^2/2$; (b) $y_p = e^{3x}/10$; (c) $y_p = (3/5)\sin x - (9/5)\cos x$; (d) $y_p = (3/5)\, e^{-3x}\,(\sin x + 3\cos x)$.]

Problem 8.20. *Obtain the general solution of the previous problem with* $f(x) = xe^{-x}$. [*Answer*: $y = C_1 e^{-2x} + C_2 e^{-x} + x\,(x/2 - 1)\, e^{-x}$.]

Problem 8.21. *Find the general solutions of the ODE $y'' - 4y' + 3y = f(x)$ for*

$$(a) \quad f(x) = 3e^{2x}; \quad (b) \quad f(x) = 5e^{x}.$$

[*Answer*: (a) $y = C_1 e^{3x} + C_2 e^{x} - 3e^{2x}$; (b) $y = C_1 e^{3x} + C_2 e^{x} - (5/2)\, xe^{x}$.]

Problem 8.22. *Find the general solutions of the equations:*

$$(a)\ y'' - 6y' + 9y = \left(3x^2 + 1\right) e^{3x}; \quad (b)\ y'' + 6y' + 9y = x\left(e^{-3x} - 1\right).$$

[*Answer*: (a) $y = \left(C_1 + C_2 x + x^2/2 + x^4/4\right) e^{3x}$; (b) $y = \left(C_1 + C_2 x + x^3/6\right) e^{-3x} + 2/27 - x/9$.]

Problem 8.23. *Solve the following ODEs:*

$$(a) \quad y'' + 9y = A\sin(3x); \quad (b) \quad y'' + 9y = 3\cos x + xe^{x}.$$

[*Answer*: (a) $y = C_1 \cos(3x) + C_2 \sin(3x) - (Ax/6)\cos(3x)$; (b) $y = C_1 \cos(3x) + C_2 \sin(3x) + (3/8)\cos x + (5x - 1)/50 e^{x}$.]

Problem 8.24. *Show that the solution of*

$$y'' - 4y' + 4y = e^{2x}\left(2x^2 + 3x + 1\right)$$

is $y = e^{2x}\left(C_1 + C_2 x + x^2/2 + x^3/2 + x^4/6\right)$.

8.2.2.2 Method of Variation of Parameters

Above we considered a rather simple case when the coefficients in the second order ODE are constants. Here we shall discuss a general case of the ODE with variable coefficients,

$$a_2(x)y'' + a_1(x)y' + a_0(x)y = f(x), \tag{8.30}$$

and will show that there is a technique for obtaining the particular integral of this inhomogeneous ODE for any right-hand side $f(x)$ provided the two solutions $y_1(x)$ and $y_2(x)$ of the corresponding homogeneous equation are known. In other words, the method works if we can solve the homogeneous equation (with $f(x) = 0$) and hence obtain the corresponding complementary solution:

$$y_c(x) = C_1 y_1(x) + C_2 y_2(x), \tag{8.31}$$

where C_1 and C_2 are arbitrary constants. The idea of the method is to search for the particular solution of the original equation in the form (cf. the end of Sect. 8.1.5)

$$y_p(x) = u_1(x)y_1(x) + u_2(x)y_2(x), \tag{8.32}$$

which is basically obtained from (8.31) by replacing the constants C_1 and C_2 with unknown functions $u_1(x)$ and $u_2(x)$. Substituting the trial solution (8.32) in Eq. (8.30), we obtain:

$$\begin{aligned} &a_2\left(u_1''y_1 + 2u_1'y_1' + u_1y_1'' + u_2''y_2 + 2u_2'y_2' + u_2y_2''\right) \\ &\quad + a_1\left(u_1'y_1 + u_1y_1' + u_2'y_2 + u_2y_2'\right) + a_0\left(u_1y_1 + u_2y_2\right) = f, \end{aligned}$$

or after some trivial manipulation:

$$\begin{aligned} &u_1\left(a_2y_1'' + a_1y_1' + a_0y_1\right) + u_2\left(a_2y_2'' + a_1y_2' + a_0y_2\right) \\ &\quad + a_2\left(u_1''y_1 + u_1'y_1' + u_2''y_2 + u_2'y_2'\right) + a_2\left(u_1'y_1' + u_2'y_2'\right) + a_1\left(u_1'y_1 + u_2'y_2\right) = f. \end{aligned}$$

The first two terms vanish since y_1 and y_2 satisfy the homogeneous equation by construction. The expression in the parentheses in the 3rd term can be written as

$$u_1''y_1 + u_1'y_1' + u_2''y_2 + u_2'y_2' = \frac{d}{dx}\left(u_1'y_1 + u_2'y_2\right) .$$

Therefore, we obtain:

$$a_2\frac{d}{dx}\left(u_1'y_1 + u_2'y_2\right) + a_2\left(u_1'y_1' + u_2'y_2'\right) + a_1\left(u_1'y_1 + u_2'y_2\right) = f. \tag{8.33}$$

Both functions $u_1(x)$ and $u_2(x)$ are to be determined. However, we have to find just one particular solution. This means that we are free to constrain the two functions $u_1(x)$ and $u_2(x)$ in some way, and this can be done to simplify Eq. (8.33), so that it can be solved. The simplest choice seems to be

$$u_1'y_1 + u_2'y_2 = 0, \tag{8.34}$$

because it would remove two terms from Eq. (8.33) yielding simply

$$a_2\left(u_1'y_1' + u_2'y_2'\right) = f. \tag{8.35}$$

Equations (8.34) and (8.35) are linear algebraic equations with respect to two unknown functions u_1' and u_2', and can easily be solved to give:

$$u_1' = -\frac{y_2 f}{a_2 W} \quad \text{and} \quad u_2' = \frac{y_1 f}{a_2 W}, \tag{8.36}$$

where $W = y_1 y_2' - y_1' y_2$ is the Wronskian of y_1 and y_2, see Eq. (8.17). It can be shown that $W \neq 0$ if y_1 and y_2 are linearly independent solutions of the homogeneous equation (see Problem 8.25 below). Thus, integrating the above equations, necessary functions u_1 and u_2 are found and the particular solution is obtained. Note that when integrating these equations, arbitrary constants can be dropped due to the specific form of Eq. (8.32) in which the particular solution is sought: the terms to be originated from those constants will simply be absorbed in the complementary solution (8.31).

> **Problem 8.25.** *Prove by contradiction that the Wronskian $W = y_1 y_2' - y_1' y_2 \neq 0$ if the two functions $y_1(x)$ and $y_2(x)$ are linearly independent.*

As an example, let us find the general solution of the equation:

$$y'' - 2y' + y = x^{-2} e^x .$$

Note that the right-hand side here has a form which is different from any type we could handle using the familiar method of undetermined coefficients. The complementary solution is:

$$y_c = C_1 e^x + C_2 x e^x ,$$

since the characteristic equation gives repeated roots. The Wronskian of $y_1 = e^x$ and $y_2 = xe^x$ is trivially found to be

$$W = y_1 y_2' - y_1' y_2 = e^x \left(e^x + xe^x\right) - e^x xe^x = e^{2x} .$$

Then, Eq. (8.36) can be written explicitly as

$$u_1' = -\frac{\left(xe^x\right)\left(x^{-2}e^x\right)}{e^{2x}} = -x^{-1} \implies u_1(x) = -\int \frac{dx}{x} = -\ln|x| ,$$

$$u_2' = \frac{e^x\left(x^{-2}e^x\right)}{e^{2x}} = x^{-2} \implies u_2(x) = -x^{-1} .$$

Therefore, the particular solution

$$y_p = -\left(\ln|x|\right) e^x + \left(-x^{-1}\right) xe^x = -e^x \ln|x| - e^x .$$

In fact, the second term, e^x, will be absorbed by the complementary solution, so that only the first term may be kept. Thus, the general solution is

$$y = y_c + y_p = C_1 e^x + C_2 x e^x - e^x \ln|x| ,$$

as can be checked by direct substitution.

8.3 Non-linear Second Order Differential Equations

Here we shall discuss a few simple tricks which in some cases are useful in solving non-linear second order ODEs. In all these cases it is possible to lower the order of the ODE and hence obtain a simpler equation which may eventually be solved.

Consider first an ODE of the form $y'' = f(x)$. It does not contain y and y' and is solved by simply noticing that $y'' = (y')'$. Therefore, by introducing a new function $z(x) = y'(x)$, the first order ODE $z' = f(x)$ with respect to the function $z(x)$ is obtained which is integrated immediately to give

$$z(x) = \int f(x)dx + C_1 .$$

To obtain $y(x)$, we have to integrate $y' = z(x)$ which is another first order ODE, solved similarly, finally yielding the solution

$$y(x) = \int z(x)dx + C_2 = \int^x dx_1 \int^{x_1} dx_2 f(x_2) + C_1 x + C_2 .$$

Two arbitrary constants appear naturally.

As an example, consider Newton's equation of motion for a particle moving under an external force changing sinusoidally with time:

$$m\ddot{x} = A \sin(\omega t) .$$

Let the particle be initially at rest at the centre of the coordinate system, i.e. $x(0) = \dot{x}(0) = 0$. To solve the equation above, we introduce the velocity $v = \dot{x}$, and then the equation of motion would read simply $\dot{v} = (A/m)\sin(\omega t)$. It is integrated to give

$$v(t) = \frac{A}{m}\int_0^t \sin(\omega\tau)\, d\tau = \frac{A}{m\omega}\left[1 - \cos(\omega t)\right] .$$

Integrating the velocity again, we obtain the particle position as a function of time:

$$x(t) = \int_0^t v(\tau)\, d\tau = \frac{A}{m\omega}\left[t - \frac{1}{\omega}\sin(\omega t)\right] .$$

Another class of equations which can be simplified is of the form $y'' = F(y, y')$, i.e. the ODE lacks direct dependence on x. In this case y can be considered instead of x as a variable. Indeed, we denote $z = y'$ and then notice that

$$y'' = \frac{d}{dx}\left(y'\right) = \frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx} = \frac{dz}{dy}z ,$$

which yields

$$z\frac{dz}{dy} = F(y, z),$$

which is now a first order ODE with respect to the function $z(y)$. If it can be solved, we then write $y' = z(y)$ which may be integrated without difficulty via

$$y' = z(y) \quad \Longrightarrow \quad \int \frac{dy}{z(y)} = \int dx \quad \Longrightarrow \quad \int \frac{dy}{z(y)} = x + C_2.$$

Note that the first arbitrary constant C_1 is contained in the function $z(y)$ after the first integration.

To illustrate this case, let us derive a trajectory $x(t)$ of a particle subject to an external potential $V(x)$. Let x_0 and v_0 be the initial ($t = 0$) position and velocity of the particle, respectively. The Newton's equation of motion in this case reads

$$m\ddot{x} = -\frac{dV}{dx} = f(x),$$

where $f(x) = -V'(x)$ is the external force due to the potential $V(x)$. Using the method described above, we introduce the velocity $v = \dot{x}$ and then write the acceleration as $a = \ddot{x} = \dot{v} = v\,(dv/dx)$, which gives

$$mv\frac{dv}{dx} = f(x) \quad \Longrightarrow \quad m\int_0^t v dv = \int_0^x f\left(x'\right) dx'$$
$$\Longrightarrow \quad \frac{mv^2}{2} - \frac{mv_0^2}{2} = \int_0^x f\left(x'\right) dx',$$

which is nothing but the expression of conservation of energy: in the left-hand side we have the change of the particle kinetic energy due to the work by the force given in the right-hand side. Alternatively, in the right-hand side one may recognise the corresponding difference

$$\int_0^x f\left(x'\right) dx' = -\int_0^x \frac{dV}{dx'} dx' = V(0) - V(x)$$

of the potential energies at two times leading to the energy conservation. Solving this equation with respect to the velocity, we find

$$v(x) = \sqrt{v_0^2 + \frac{2}{m}\int_0^x f(x')\, dx'},$$

which is also dx/dt. Therefore, integrating again, we obtain an equation for the trajectory in a rather general form:

$$\int_{x_0}^{x} \frac{dx''}{\sqrt{v_0^2 + \frac{2}{m}\int_0^{x''} f\left(x'\right) dx'}} = t.$$

Finally, if the ODE $f\left(y'', y', y, x\right) = 0$ is a homogeneous function with respect to its variables y, y' and y'' (treated together, see an example below), then one can introduce instead of $y(x)$ a new function $z(x)$ via

$$y(x) = \exp\left(\int z(x)dx\right) \tag{8.37}$$

to reduce the order of the ODE. Indeed,

$$y' = \exp\left(\int z(x)dx\right) z(x) = yz \quad \text{and} \quad y'' = (yz)' = y'z + yz' = y\left(z^2 + z'\right),$$

so that each derivative of y factorises into y and some linear combination of z and its first derivative; this allows cancelling out $y(x)$ in the equation as demonstrated by the following example. Let us consider

$$yy'' - y'\left(y + y'\right) = 0,$$

which has each of its terms being homogeneous of degree two. Using the substitution (8.37), we obtain:

$$y^2\left(z^2 + z'\right) - yz\left(y + zy\right) = 0 \quad \Longrightarrow \quad z^2 + z' - z\left(1 + z\right) = 0 \quad \Longrightarrow \quad z' = z,$$

which is easily integrated giving $z(x) = C_1 e^x$. Therefore, the function $y(x)$ sought for is found as

$$y(x) = \exp\left(\int C_1 e^x dx\right) = \exp\left(C_1 e^x + C_3\right) = C_2 \exp\left(C_1 e^x\right),$$

where C_1 and C_2 (or C_3, depending of the form of the solution taken) are two arbitrary constants.

Problem 8.26. *Solve the following ODE using one of the methods considered above:*

$$(a)\ y'' = e^x; \quad (b)\ y\left(y'' + y'\right) = \left(y'\right)^2.$$

[*Answer*: (a) $y = C_1 x + C_2 + e^x$; (b) $y = C_1 \exp\left(C_2 e^{-x}\right)$.]

8.4 Series Solution of Linear ODEs

In many applications, e.g. when solving partial differential equations of mathematical physics, one obtains ordinary second-order differential equations of the type

$$y''(x) + p(x)y'(x) + q(x)y(x) = 0, \tag{8.38}$$

where $p(x)$ and $q(x)$ are known variable functions of x. This is the so-called *canonical* form of the ODE. Since one can always ensure that the coefficient to the second derivative of y is equal to one by dividing on it both sides of the equation, it is sufficient to study any ODE in its canonical form only.

We know that when the coefficients $p(x)$ and $q(x)$ are both *constants,* it is possible to solve Eq. (8.38) and express the solution in terms of elementary functions (exponentials, sine and cosine). However, if the coefficients $p(x)$ and $q(x)$ are *variable* functions, then the solution of (8.38) in terms of elementary functions is possible only in special cases; in most cases, the solution cannot be expressed in terms of a finite number of elementary functions. It can be shown that it is generally possible to devise a method whereby the solution is expressed as an infinite power series.[2] This can be used either as an analytical tool or as a means of calculating the solution numerically when the series is terminated at a finite number of terms.

It is absolutely essential first to choose a point $x = a$ around which the series is to be developed. The point $x = a$ is said to be an *ordinary* point of the differential equation (8.38) if *both* $p(x)$ and $q(x)$ are "well-behaved" functions around this point, i.e. the functions are single-valued, finite and continuous there possessing derivatives $f^{(k)}(a)$ of *all* orders. If the point $x = a$ is *not* an *ordinary* point, then we shall call it a *singular* point.

For instance, the differential equation

$$x^2(x+3)y''(x) - (x+3)y'(x) + 3xy(x) = 0, \tag{8.39}$$

if written in the canonical form, yields the two functions

$$p(x) = -\frac{1}{x^2} \quad \text{and} \quad q(x) = \frac{3}{x(x+3)}, \tag{8.40}$$

which are singular at points $x = 0$ and -3. All other points in the interval $-\infty < x < \infty$ are therefore *ordinary* points.

Before discussing the most general case, we shall first introduce main ideas of the method for the case when the series is constructed around an ordinary point.

[2]In some cases the series may be finite.

8.4.1 Series Solutions About an Ordinary Point

Suppose that the initial conditions $y(a)$ and $y'(a)$ are given at the point $x = a$ which we assume is an *ordinary* point of Eq. (8.38). Then we can, *at least in principle*, calculate all the derivatives $y^{(k)}(a)$ $(k = 1, 2, 3 \ldots)$ in terms of $y(a)$ and $y'(a)$, by repeated differentiation of Eq. (8.38) with respect to x. Indeed, $y''(a)$ follows immediately from the equation itself. Differentiating the equation once and setting $x = a$, we can express the third derivative of $y(x)$ via $y(a)$, $y'(a)$ and $y''(a)$. Repeating this process, one can calculate any derivative $y^{(r)}(a)$ of $y(x)$ at $x = a$, where $r = 1, 2, \ldots$. Hence, we can *formally* construct a Taylor *series* about $x = a$ for our function $y(x)$,

$$y(x) = \sum_{r=0}^{\infty} c_r (x-a)^r \quad \text{with} \quad c_r = \frac{y^{(r)}(a)}{r!}, \tag{8.41}$$

which forms a solution of the equation; note that the initial conditions are satisfied if $c_0 = y(a)$ and $c_1 = y'(a)$. This series gives us a *general* solution of Eq. (8.38) in the neighbourhood of $x = a$ if $c_0 = y(a) = C_1$ and $c_1 = y'(a) = C_2$ are considered as arbitrary constants.

This solution would converge within an interval $|x-a| < R$, where R is the *radius of convergence* of the power series, Eq. (8.41). Usually, the radius of convergence R is equal to the distance from the point $x = a$ to the nearest singular point of Eq. (8.38). This point follows from a more general discussion based on the theory of functions of complex variables (see Sect. 2.8, Book II).

The method considered above is a proof that the solution can always be found and is *unique*. However, this method is not convenient for actual applications. In *practice* the coefficients c_r are most easily found by substituting the series (8.41) into the differential equation (8.38), collecting terms with the like powers of x and setting their coefficients equal to zero. Indeed, $p(x)$ and $q(x)$ can be expanded into the Taylor series around $x = a$ (recall that these functions have no singularities at $x = a$):

$$p(x) = p_0 + p_1 (x-a) + p_2 (x-a)^2 + \cdots ,$$
$$q(x) = q_0 + q_1 (x-a) + q_2 (x-a)^2 + \cdots ,$$

which, when substituted into the differential equation (8.38), results in the following infinite power series being set to zero:

$$(2c_2 + p_0c_1 + q_0c_0)\, u^0 + (3 \cdot 2c_3 + 2p_0c_2 + p_1c_1 + q_0c_1 + q_1c_0)\, u^1$$
$$+ (4 \cdot 3c_4 + 3p_0c_3 + 2p_1c_2 + p_2c_1 + q_0c_2 + q_1c_1 + q_2c_0)\, u^2 + \cdots = 0,$$

where $u = x - a$. Since this equation is expected to be satisfied for a continuous set of x values in some vicinity of the point $x = a$ (or for u around zero), and because the functions u^0, u^1, u^2, etc. form a linearly independent set, this can only be possible if all coefficients of the powers of u are equal to zero at the same time, which yields an infinite set of algebraic equations for the unknown coefficients c_2, c_3, c_4, etc.:

$$\begin{aligned} 2c_2 + p_0c_1 + q_0c_0 &= 0, \\ 3 \cdot 2c_3 + 2p_0c_2 + p_1c_1 + q_0c_1 + q_1c_0 &= 0, \\ 4 \cdot 3c_4 + 3p_0c_3 + 2p_1c_2 + p_2c_1 + q_0c_2 + q_1c_1 + q_2c_0 &= 0, \end{aligned}$$

and so on. The first equation gives c_2, the second equation gives c_3 via c_2, the third c_4 via c_2 and c_3, etc. This way one can obtain recurrence relations for the coefficients which express c_r ($r \geq 3$) via the previous ones: $c_2, c_3, \ldots, c_{r-1}$. So, given the values of the initial coefficients c_0 and c_1, it is possible to generate all the coefficients c_r up to any desired maximum value of the index r.

Although at first sight the procedure looks cumbersome, it is actually very convenient in actual applications, especially when the functions $p(x)$ and $q(x)$ are polynomials.

Example 8.6. ► Determine the series solution of the differential equation

$$y'' - xy = 0, \tag{8.42}$$

satisfying the boundary conditions $y(0) = 1$ and $y'(0) = 0$.

Solution. Expanding around an ordinary point $x = 0$ (i.e. for $a = 0$), we can write

$$y(x) = \sum_{r=0}^{\infty} c_r x^r. \tag{8.43}$$

Differentiating $y(x)$, we find:

$$y'(x) = \sum_{r=1}^{\infty} rc_r x^{r-1} \quad \text{and} \quad y''(x) = \sum_{r=2}^{\infty} r(r-1)c_r x^{r-2}. \tag{8.44}$$

Note that in $y'(x)$ the first term ($r = 0$) in the sum disappears, while the first two terms ($r = 0, 1$) disappear in the sum in the expression for $y''(x)$. Substituting the series (8.43) and (8.44) into the differential equation (8.42) gives

$$\sum_{r=2}^{\infty} r(r-1)c_r x^{r-2} - \sum_{r=0}^{\infty} c_r x^{r+1} = 0.$$

Here the first sum contains all powers of x starting from x^0, while the second sum has all powers of x starting from x^1. To collect terms with like powers of x, we need to combine the two sums into one. To this end, it is convenient to separate out the x^0

term in the first sum (corresponding to $r = 2$), so that the two sums would contain identical powers of x:

$$2c_2 + \sum_{r=3}^{\infty} r(r-1)c_r x^{r-2} - \sum_{r=0}^{\infty} c_r x^{r+1} = 0. \tag{8.45}$$

Indeed, now the first sum contains terms x^1, x^2, etc. and one can easily see that the same powers of x appear in the second sum. Once we have both sums starting from the same powers of x, we can now make them look similar which would eventually enable us to combine them together into a single sum. For instance, this can be done by replacing the summation index r in the first sum with a different index m selected in such a way that the power of x in the first sum would look the same as in the second. Making the replacement $m + 1 = r - 2$ for the summation index in the first sum of Eq. (8.45) does the trick, and we obtain:

$$2c_2 + \sum_{m=0}^{\infty} (m+3)(m+2)c_{m+3} x^{m+1} - \sum_{r=0}^{\infty} c_r x^{r+1} = 0.$$

Note that $m = r - 3$ here, and hence it starts from the zero value corresponding to $r = 3$ of the original first sum. Now the two sums look indeed the same: they have the same powers of x and the summation indices start from the same (zero) value. To stress the similarity even further, we can use in the first sum the same letter r for the summation index instead of m (the summation index is a "dummy index" and any symbol can be used for it!), in which case the two sums are readily combined:

$$2c_2 + \sum_{r=0}^{\infty} (r+3)(r+2)c_{r+3} x^{r+1} - \sum_{r=0}^{\infty} c_r x^{r+1} = 0$$

$$\Longrightarrow \quad 2c_2 + \sum_{r=0}^{\infty} \left[(r+3)(r+2)\, c_{r+3} - c_r\right] x^{r+1} = 0. \tag{8.46}$$

The differential equation is satisfied for all values of x provided that

$$c_2 = 0 \tag{8.47}$$

and, at the same time,

$$(r+3)(r+2)\, c_{r+3} = c_r \quad \Longrightarrow \quad c_{r+3} = \frac{c_r}{(r+2)(r+3)}\,, \quad r = 0, 1, 2, \ldots. \tag{8.48}$$

From this recurrence relation, we can generate the coefficients c_r:

$$
\begin{aligned}
r = 0 \quad &\Longrightarrow \quad c_3 = \frac{c_0}{2 \cdot 3}, \\
r = 1 \quad &\Longrightarrow \quad c_4 = \frac{c_1}{3 \cdot 4}, \\
r = 2 \quad &\Longrightarrow \quad c_5 = \frac{c_2}{4 \cdot 5} = 0, \\
r = 3 \quad &\Longrightarrow \quad c_6 = \frac{c_3}{5 \cdot 6} = \frac{c_0}{2 \cdot 3 \cdot 5 \cdot 6}, \\
r = 4 \quad &\Longrightarrow \quad c_7 = \frac{c_4}{6 \cdot 7} = \frac{c_1}{3 \cdot 4 \cdot 6 \cdot 7}, \\
r = 5 \quad &\Longrightarrow \quad c_8 = \frac{c_5}{7 \cdot 8} = 0,
\end{aligned}
$$

and so on. One can see that three families of coefficients are generated. If we start from c_0, we generate c_3, c_6, etc.; if we start from c_1, we generate c_4, c_7, etc; finally, starting from c_2 (which is zero), we generate c_5, c_8, etc. which are all equal to zero. Since nothing can be said about the coefficients c_0 and c_1, we have to accept that these can be arbitrary. Therefore, if we recall the general form of the solution,

$$
y(x) = c_0 x^0 + c_1 x^1 + c_2 x^2 + \cdots ,
$$

then two solutions are obtained, one starting from c_0, and another one from c_1:

$$
\begin{aligned}
y_1(x) &= c_0 + c_3 x^3 + c_6 x^6 + \cdots \\
&= c_0 \left[1 + \frac{x^3}{2 \cdot 3} + \frac{x^6}{(2 \cdot 5)(3 \cdot 6)} + \frac{x^9}{(2 \cdot 5 \cdot 8)(3 \cdot 6 \cdot 9)} + \cdots \right], \\
y_2(x) &= c_1 x + c_4 x^4 + c_7 x^7 + \cdots \\
&= c_1 \left[x + \frac{x^4}{3 \cdot 4} + \frac{x^7}{(3 \cdot 6)(4 \cdot 7)} + \frac{x^{10}}{(3 \cdot 6 \cdot 9)(4 \cdot 7 \cdot 10)} + \cdots \right].
\end{aligned}
$$

As expected, the coefficients c_0 and c_1 serve as arbitrary constants, so that expressions in the square brackets can in fact be considered as linearly independent solutions

$$
\begin{aligned}
y_1(x) &= 1 + \frac{x^3}{2 \cdot 3} + \frac{x^6}{(2 \cdot 5)(3 \cdot 6)} + \frac{x^9}{(2 \cdot 5 \cdot 8)(3 \cdot 6 \cdot 9)} + \cdots , \\
y_2(x) &= x + \frac{x^4}{3 \cdot 4} + \frac{x^7}{(3 \cdot 6)(4 \cdot 7)} + \frac{x^{10}}{(3 \cdot 6 \cdot 9)(4 \cdot 7 \cdot 10)} + \cdots .
\end{aligned}
$$

while the general solution is

$$y(x) = C_1 y_1(x) + C_2 y_2(x),$$

with arbitrary constants C_1 and C_2. Note that $y_1(0) = 1$ and $y_2(0) = 0$, while $y_1'(0) = 0$ and $y_2'(0) = 1$.

The application of the initial conditions $y(0) = C_1 = 1$, and $y'(0) = C_2 = 0$ gives the constants and hence the required *particular* solution (integral) is:

$$y(x) = y_1(x) = 1 + \frac{x^3}{2 \cdot 3} + \frac{x^6}{(2 \cdot 5)(3 \cdot 6)} + \frac{x^9}{(2 \cdot 5 \cdot 8)(3 \cdot 6 \cdot 9)} + \cdots . \quad (8.49)$$

The convergence of the series solution (8.49) can be investigated directly using the ratio test with the help of the recurrence relation (8.48). Recall that an infinite series $\sum_{k=0}^{\infty} a_k$ is *convergent* if $\lim_{k\to\infty} |a_{k+1}/a_k| < 1$. In our case of the series (8.49) the two adjacent terms in the series are $c_{3k}x^{3k}$ and $c_{3k+3}x^{3k+3}$, which gives

$$\lim_{k\to\infty} \left| \frac{c_{3k+3}x^{3k+3}}{c_{3k}x^{3k}} \right| = \lim_{k\to\infty} \frac{c_{3k+3}}{c_{3k}} |x|^3 = |x|^3 \lim_{k\to\infty} \frac{1}{(3k+2)(3k+3)} = 0 \quad (8.50)$$

for any value of x. We see from this result that the series solution (8.49) converges for *all* finite values of x. ◀

8.4.2 Series Solutions About a Regular Singular Point

Let us now suppose that the differential equation (8.38) has a *singular point* at $x = a$. It can be shown (Sect. 2.8, Book II) that one has to distinguish two cases: (i) if the functions $(x-a)p(x)$ and $(x-a)^2 q(x)$ are *both "well-behaved"* at $x = a$, then we say that $x = a$ is a *regular singular point* (RSP) of Eq. (8.38). Otherwise, a *singular point* is called an *irregular singular point* (ISP) of the differential equation (8.38). This *classification scheme* for singular points is *important* because it is always possible to obtain two independent solutions of Eq. (8.38) in the neighbourhood of a regular singular point $x = a$ by using a *generalised* series method to be introduced below; no such expansion is possible around an irregular singular point.

The introduced classification of singular points we shall illustrate for the differential equation (8.39) for which the functions $p(x)$ and $q(x)$ are given by Eq. (8.40). It has two singular points: at $x = 0$ and $x = -3$. For the singular point $a = -3$ we have

$$(x-a)\,p(x) \quad \Longrightarrow \quad (x-(-3))p(x) = -\frac{(x+3)}{x^2}$$

and

$$(x-a)^2\, q(x) \quad \Longrightarrow \quad (x-(-3))^2 q(x) = 3\frac{(x+3)}{x}.$$

Both these functions are "well-behaved" at $x = -3$ which is then a *regular singular point*. For the singular point, $a = 0$, we similarly find that

$$(x-0)p(x) = -\frac{1}{x} \quad \text{and} \quad (x-0)^2 q(x) = \frac{3x}{(x+3)}.$$

It is clear that the function $(x-0)p(x) = -1/x$ is *not* "well-behaved" at $x = 0$, and hence the point $x = 0$ is an *irregular singular point*.

If $x = a$ is an RSP of Eq. (8.38), then the Taylor series solution (8.41) will *not be possible* because the functions $p(a)$ and/or $q(a)$ do not exist and hence the derivatives $y^{(r)}(a)$ for $r = 2, 3, 4, \ldots$ are also not defined. However, it can be shown that Eq. (8.38) always has at least *one* solution of the *Frobenius* type (Sect. 2.8, Book II):

$$y(x) = (x-a)^s \sum_{r=0}^{\infty} c_r (x-a)^r = \sum_{r=0}^{\infty} c_r (x-a)^{r+s}, \tag{8.51}$$

where the exponent s can be *negative* and/or even *non-integral*. The radius of convergence R of the series (8.51) is at least as large as the distance to the nearest singular point of the differential equation (8.38).

In *practice* the exponent s and the coefficients c_r are found by exactly the same method as for an ordinary point discussed above: by substituting the series (8.51) into the differential equation, collecting terms with the like powers of x and equating the corresponding coefficients to zero. This procedure leads to an *indicial equation* for s and a *recurrence relation* for the coefficients c_r. It is more transparent to illustrate this method considering an example.

Example 8.7. ► Determine the general series solution of the differential equation

$$2xy'' + y' + y = 0 \tag{8.52}$$

about $x = 0$.

Solution. In this differential equation $p(x) = q(x) = 1/2x$, and hence $x = 0$ is a singular point. We should first check if it is regular or irregular. Since the functions $xp(x) = 1/2$ and $x^2 q(x) = x/2$ are both "well-behaved" at $x = 0$, it follows, therefore, that $x = 0$ is an RSP. We can therefore assume a series solution of the Frobenius type around $x = 0$:

$$y(x) = \sum_{r=0}^{\infty} c_r x^{r+s} \tag{8.53}$$

with some yet to be determined number s. From (8.53) we find that

$$y'(x) = \sum_{r=0}^{\infty} c_r (r+s) x^{r+s-1} \quad \text{and} \quad y''(x) = \sum_{r=0}^{\infty} c_r (r+s)(r+s-1) x^{r+s-2}. \tag{8.54}$$

Note that, contrary to Example 8.6 considered above for the case of an ordinary point, in this case the first terms in the sums do *not* disappear after differentiation as s may be non-integer in general, and hence in both cases we have to keep all values of the summation index r starting from zero. The substitution of these results in Eq. (8.52) and combining the sums originating from y'' and y' gives:

$$\sum_{r=0}^{\infty} c_r (r+s)(2r+2s-1) x^{r+s-1} + \sum_{r=0}^{\infty} c_r x^{r+s} = 0,$$

$$c_0 s(2s-1)x^{s-1} + \sum_{r=1}^{\infty} c_r (r+s)(2r+2s-1) x^{r+s-1} + \sum_{r=0}^{\infty} c_r x^{r+s} = 0, \tag{8.55}$$

where the first ($r = 0$) term in the first sum has been separated out as the second sum does not have a term with x^{s-1}. Next, we make the index shift $r \mapsto r+1$ in the first summation in Eq. (8.55), i.e. we introduce a new index $r' = r - 1$ in the sum and then write r instead of r' for convenience, so that the first sum undergoes the transformation:

$$\begin{aligned} \sum_{r=1}^{\infty} c_r (r+s)(2r+2s-1) x^{r+s-1} &= \sum_{r'=0}^{\infty} c_{r'+1} (r'+s+1)(2r'+2s+1) x^{r'+s} \\ &= \sum_{r=0}^{\infty} c_{r+1} (r+s+1)(2r+2s+1) x^{r+s}. \end{aligned}$$

At the last step we used r instead of r' for the dump index for convenience. The two sums in (8.55) can now be combined into one, which leads us to the equation:

$$c_0 s\,(2s-1)\,x^{s-1} + \sum_{r=0}^{\infty} \left[(r+s+1)\,(2r+2s+1)\,c_{r+1} + c_r \right] x^{r+s} = 0. \tag{8.56}$$

This is an infinite sum of different powers of x which is to be equal to zero for all values of the x around $x = 0$. This means that in each and every term in the expansion above the coefficients to the powers of x must vanish, i.e. we should have $c_0 s\,(2s-1) = 0$ and, at the same time,

$$\begin{aligned} &(r+s+1)(2r+2s+1)c_{r+1} + c_r = 0 \\ \Longrightarrow \quad & c_{r+1} = -\frac{c_r}{(r+s+1)(2r+2s+1)} \end{aligned} \tag{8.57}$$

for every $r = 0, 1, 2, \ldots$.

Consider first the recurrence relation (8.57). One can see that c_{r+1} is directly proportional to c_r. More explicitly: c_1 is proportional to c_0, c_2 is proportional to c_1 and hence eventually also to c_0. Discussing along similar lines it becomes apparent that any coefficient c_r for $r \geq 1$ will be proportional to c_0 in the end. Next, let us consider the first equation, $c_0 s(2s-1) = 0$. It accepts as a solution $c_0 = 0$. However, it is clear from what was said above that in this case all other coefficients will also be equal to zero and hence this solution leads to the trivial solution $y(x) = 0$ of the differential equation and is thus of no interest. Hence, we have to assume that $c_0 \neq 0$.

Hence, as $a_0 \neq 0$, we arrive at an algebraic equation $s(2s-1) = 0$, which is a quadratic equation for s. It is called *indicial* equation and gives two solutions for s, namely: $s = 1/2$ and $s = 0$. Using these two particular values of s, we can now generate the values of the coefficients c_r via the recurrence relation (8.57) and hence build completely two linearly independent solutions.

Consider first the case of $s = 1/2$. From (8.57) the coefficients c_r satisfy the recurrence relation

$$c_{r+1} = -\frac{c_r}{(r+3/2)(2r+2)} = -\frac{c_r}{(r+1)\,(2r+3)} \quad \text{for} \quad \mathrm{r} = 0, 1, 2, \ldots .$$

It is clear that all coefficients will be proportional to c_0, the latter will therefore eventually become an arbitrary constant. Therefore, for constructing a linearly independent elementary solution, it is sufficient (and convenient) to set $c_0 = 1$. It follows then that the first solution corresponding to $s = 1/2$ becomes:

$$y_1(x) = \sum_{r=0}^{\infty} c_r x^{r+1/2} = \sqrt{x}\left(1 - \frac{x}{1\cdot 3} + \frac{x^2}{1\cdot 2\cdot 3\cdot 5} - \frac{x^3}{1\cdot 2\cdot 3^2\cdot 5\cdot 7} + \cdots\right). \tag{8.58}$$

The value of $s = 0$ leads to the recurrence relation

$$c_{r+1} = -\frac{c_r}{(r+1)\,(2r+1)},$$

and correspondingly to the second elementary solution (we set $c_0 = 1$ again):

$$y_2(x) = 1 - x + \frac{x^2}{2\cdot 3} - \frac{x^3}{2\cdot 3^2\cdot 5} + \cdots . \tag{8.59}$$

The general solution is given as a linear combination of the two constructed elementary solutions. ◄

Problem 8.27. *By using the recurrence relations for both values of* $s = 1/2$ *and* $s = 0$*, work out explicitly the first four terms in the expansions of* $y_1(x)$ *and* $y_2(x)$ *and hence confirm expressions (8.58) and (8.59) given above.*

The next example shows that the Frobenius method can also be used even in the case of expanding around an ordinary point.

Example 8.8. ► Solve the harmonic oscillator equation

$$y''(x) + \omega^2 y(x) = 0,$$

using the generalised series expansion method.

Solution. This equation does not have any singular points and hence we can expand around $x = 0$ which is an ordinary point using an ordinary Taylor's series. However, to illustrate the power of the Frobenius method, we shall solve the equation using the generalised series expansion assuming some s:

$$y = \sum_{r=0}^{\infty} c_r x^{r+s}.$$

After substituting into the DE, we get:

$$\sum_{r=0}^{\infty} c_r (r+s)(r+s-1)x^{r+s-2} + \omega^2 \sum_{r=0}^{\infty} c_r x^{r+s} = 0.$$

The following powers of x are contained in the first sum: x^{s-2}, x^{s-1}, x^s, x^{s+1}, etc., corresponding to the summation index $r = 0, 1, 2, 3$, etc., respectively. At the same time, the second sum contains the terms with x^s, x^{s+1}, etc. Anticipating that we will have to combine the two sums together later on, we separate out the first two "foreign" terms from the first sum:

$$c_0 s(s-1)x^{s-2} + c_1(s+1)sx^{s-1} + \sum_{r=2}^{\infty} c_r (r+s)(r+s-1)x^{r+s-2} + \omega^2 \sum_{r=0}^{\infty} c_r x^{r+s} = 0.$$

In the first sum we shift the summation index $r - 2 \to r$, so that the two sums can be combined into one:

$$c_0 s(s-1)x^{s-2} + c_1 s(s+1)x^{s-1} + \sum_{r=0}^{\infty} \left[(r+s+2)(r+s+1)c_{r+2} + \omega^2 c_r \right] x^{r+s} = 0. \tag{8.60}$$

It is convenient to start by assuming that $c_0 \neq 0$. Then, from the first term we conclude that $s(s-1) = 0$ (the indicial equation), which gives us two possible values of s, namely 0 and 1. Therefore, we need to consider two cases.

In the case of $s = 0$ we see that the second term containing x^{s-1} is zero automatically, so that c_1 is arbitrary at this stage. From the last term in (8.60) we obtain the recurrence relation

$$c_{r+2} = -\frac{\omega^2}{(r+2)(r+1)} c_r \ , \ \text{ where } \ r = 0, 1, 2, \ldots ,$$

which (assuming $c_0 = 1$) generates the coefficients with even indices: $c_2 = -\omega^2/2$, $c_4 = \omega^4/4!$, etc. It is easy to see that generally

$$c_{2n} = (-1)^n \frac{\omega^{2n}}{(2n)!} \ , \ n = 1, 2, \ldots ,$$

which can be proven, e.g., by the method of mathematical induction. Indeed, this is true for $n = 1$. We then assume that it is also true for some value of n. Using the recurrence relation, we get for the next coefficient:

$$\begin{aligned} c_{2(n+1)} = c_{2n+2} &= -\frac{\omega^2}{(2n+2)(2n+1)} c_{2n} \\ &= -\frac{\omega^2}{(2n+2)(2n+1)} \cdot (-1)^n \frac{\omega^{2n}}{(2n)!} = (-1)^{n+1} \frac{\omega^{2n+2}}{(2n+2)!} , \end{aligned}$$

as required. Combining all terms with even coefficients (and thus even powers of x), we obtain the first elementary solution of the equation:

$$y_1(x) = \sum_{n=0}^{\infty} (-1)^n \frac{\omega^{2n}}{(2n)!} x^{2n} ,$$

which as can easily be seen is the Taylor's expansion of the cosine function $\cos(\omega x)$.

Let us now consider the case of $s = 1$. In this case c_0 is arbitrary as $s(s-1) = 0$ in the first term in Eq. (8.60); however, the second term there gives $c_1 = 0$ as $s(s+1) = 2 \neq 0$. Then the corresponding recurrence relation

$$c_{r+2} = -\frac{\omega^2}{(r+3)(r+2)} c_r \ , \ r = 0, 1, 2, \ldots$$

leads to

$$y_2(x) = \sum_{n=1}^{\infty} (-1)^n \frac{\omega^{2n}}{(2n+1)!} x^{2n+1} = \frac{1}{\omega} \sum_{n=1}^{\infty} (-1)^n \frac{(\omega x)^{2n+1}}{(2n+1)!} = \frac{1}{\omega} \sin(\omega x) ,$$

which is the Taylor's expansion of the corresponding sine function. Note that since $c_1 = 0$ there will be no odd terms at all: $c_3 = c_5 = \cdots = 0$. Thus, by using the second value of s, the expected second solution is generated. Concluding, the final general solution of the harmonic oscillator equation is

$$y(x) = C_1 \sin(\omega x) + C_2 \cos(\omega x) ,$$

as expected, where C_1 and C_2 are arbitrary constants. ◄

Problem 8.28. *Above, both solutions were obtained by assuming that $c_0 \neq 0$. Consider now another possible proposition which is assuming that $c_1 \neq 0$. Show that in this case s is equal to either 0 or -1 which generate the same solutions.*

Example 8.9. ► Solve the Legendre equation

$$\left(1 - x^2\right) y''(x) - 2xy'(x) + l(l+1)y(x) = 0, \tag{8.61}$$

using the generalised series method. Here $l > 0$.

Solution. First of all, we consider $p(x)$ and $q(x)$ to find all singular points of the DE and characterise them:

$$p(x) = -\frac{2x}{1-x^2} = -\frac{2x}{(1+x)(1-x)} \quad \text{and} \quad q(x) = \frac{l(l+1)}{(1+x)(1-x)}.$$

Points $x = \pm 1$ are regular singular points since:

$$(1+x)p(x) = -\frac{2x}{1-x} \quad \text{and} \quad (1+x)^2 q(x) = (1+x)\frac{l(l+1)}{1-x}$$

are both regular at $x = -1$; similarly for the $x = 1$:

$$(1-x)p(x) = -\frac{2x}{1+x} \quad \text{and} \quad (1-x)^2 q(x) = (1-x)\frac{l(l+1)}{1+x}$$

are both regular at $x = 1$ and hence this point is also an RSP.

Correspondingly, we seek the series expansion around $x = 0$ exactly in the middle of the interval $-1 < x < 1$ since in this case the series is supposed to converge for any x within this interval. This follows from the fact, mentioned above, that the interval of convergence of the generalised series expansion is determined by the distance from the point a (which is 0 in our case) to the nearest singular point (which is either $+1$ or -1).

Substituting the expansion

$$y = \sum_{r=0}^{\infty} c_r x^{r+s}$$

into the differential equation, we get:

$$\sum_{r=0}^{\infty} c_r(r+s)(r+s-1)\left(1-x^2\right)x^{r+s-2} - 2\sum_{r=0}^{\infty} c_r(r+s)x^{r+s} + l(l+1)\sum_{r=0}^{\infty} c_r x^{r+s} = 0. \tag{8.62}$$

The first term leads to two sums due to $\left(1-x^2\right)$. The first one,

$$
\begin{aligned}
1\mathrm{st} &= \sum_{r=0}^{\infty} c_r(r+s)(r+s-1)x^{r+s-2} = c_0 s(s-1)x^{s-2} + c_1 s(s+1)x^{s-1} \\
&+ \sum_{r=2}^{\infty} c_r(r+s)(r+s-1)x^{r+s-2},
\end{aligned}
$$

after changing the summation index $r-2 \to r' \to r$, is worked out into

$$
1\mathrm{st} = c_0 s(s-1)x^{s-2} + c_1 s(s+1)x^{s-1} + \sum_{r=0}^{\infty} c_{r+2}(r+s+2)(r+s+1)x^{r+s},
$$

while the second sum is

$$
2\mathrm{nd} = -\sum_{r=0}^{\infty} c_r(r+s)(r+s-1)x^{r+s}.
$$

We separated out the first two terms in the first sum above in order to be able to combine the two sums together into a single sum. Moreover, all other sums in Eq. (8.62) have now the same structure and can be combined all together:

$$
\begin{aligned}
& c_0 s(s-1)x^{s-2} + c_1 s(s+1)x^{s-1} + \sum_{r=0}^{\infty} \left[(r+s+2)(r+s+1)c_{r+2}\right. \\
& \quad \left. -(r+s)(r+s-1)c_r - 2(r+s)c_r + l(l+1)c_r\right] x^{r+s} = 0,
\end{aligned}
$$

which after rearranging yields:

$$
\begin{aligned}
& c_0 s(s-1)x^{s-2} + c_1 s(s+1)x^{s-1} + \sum_{r=0}^{\infty} \left\{(r+s+2)(r+s+1)c_{r+2}\right. \\
& \quad \left. - \left[(r+s)(r+s+1) - l(l+1)\right] c_r\right\} x^{r+s} = 0.
\end{aligned} \tag{8.63}
$$

The coefficients to x in the first two terms should be zero, so that, assuming for definiteness that c_0 is arbitrary, we have the indicial equation $s(s-1)=0$, resulting in two values of s. We must consider both cases. If $s=0$, then c_1 must be arbitrary as well because of the second term in Eq. (8.63) which is zero for any c_1. Next, the recurrence relation (from the last term with the sum sign) reads:

$$
c_{r+2} = \frac{r(r+1) - l(l+1)}{(r+2)(r+1)} c_r. \tag{8.64}
$$

We can start either from c_0 or c_1. If we first start from $c_0 = 1$, then we can generate c_{2n} coefficients with even indices (below we use $k = l(l+1)$):

$$c_2 = \frac{-l(l+1)}{2 \cdot 1} = -\frac{k}{2!}\ , \quad c_4 = \frac{2 \cdot 3 - k}{4 \cdot 3} c_2 = -\frac{(6-k)k}{4!},$$

and so on, leading to a series solution

$$y_1(x) = 1 - \frac{k}{2!}x^2 - \frac{(6-k)k}{4!}x^4 - \cdots\ , \tag{8.65}$$

containing even powers of x.

Starting from c_1, we similarly obtain a series expansion containing only odd powers of x:

$$y_2(x) = x + \frac{2-k}{3!}x^3 + \frac{(12-k)(2-k)}{5!}x^5 + \cdots\ . \tag{8.66}$$

We now consider the other case of $s = 1$. The recurrence relation in this case is:

$$c_{r+2} = \frac{(r+1)(r+2) - k}{(r+3)(r+2)} c_r.$$

Since the second term in Eq. (8.63) should be zero, we have to set $c_1 = 0$ since $s(s+1) \neq 0$ when $s = 1$. Then, starting from c_0, we obtain coefficients c_{2n} with even indices and it can easily be checked that this way we arrive at $y_2(x)$ again (note that in the case of $s = 1$ all terms in the expansion $\sum_r c_{2r} x^{2r+1}$ have odd powers, as required). Thus, the second value of s does not lead to any new solutions.

Let us now investigate the convergence of the first series ($s = 0$) using the ratio test and employing the recurrence relation (8.64):

$$\left|\frac{c_{r+2}x^{r+2}}{c_r x^r}\right| = \frac{r(r+1) - k}{(r+2)(r+1)}x^2 = \left[\frac{r}{r+2} - \frac{k}{(r+1)(r+2)}\right]x^2 \to x^2 \ \text{ as } r \to \infty.$$

Therefore, the series $y_1(x)$ converges for $-1 < x < 1$, as expected. Similar analysis is performed for $y_2(x)$:

$$\left|\frac{c_{r+2}x^{r+2}}{c_r x^r}\right| = \frac{r(r+1) - k}{(r+1)(r+2)}x^2 \to x^2$$

as $r \to \infty$, leading therefore to exactly the same interval.

In general, either of the solutions diverges at $x = \pm 1$. However, when l is a positive integer, then one of the solutions contains a finite number of terms becoming a polynomial, and therefore it obviously converges at the boundary points $x = \pm 1$ (as well as at any other x between $\pm\infty$). Indeed, consider first l being even. In the expansion of $y_1(x)$ only coefficients c_r with even indices r are present. Consider then the recurrence relation for the coefficients:

$$c_{r+2} = \frac{r(r+1)-k}{(r+1)(r+2)}c_r = \frac{r(r+1)-l\,(l+1)}{(r+1)(r+2)}c_r.$$

It is readily seen that when $r = l$ (which is possible as both r and l are even integers) we have $c_{l+2} = 0$ because of the numerator in the equation above. As a result, all other coefficients c_{l+4}, c_{l+6}, etc. will also be equal to zero, i.e. the solution $y_1(x)$ contains only a finite number of terms and is a polynomial of degree l. For, example, if $l = 2$, then $k = 2 \cdot 3 = 6$ and

$$c_2 = \frac{-6}{1 \cdot 2} = -3\,, \quad c_4 = -\frac{2 \cdot 3 - 6}{3 \cdot 4}c_2 = 0\,, \quad c_6 = c_8 = \cdots = 0,$$

and thus $y_1(x)$ becomes simply $[y_1(x)]_{l=2} = 1 - 3x^2$. At the same time, it can be seen that the other solution, $y_2(x)$, will not terminate for any even l since in the recurrence relation the numerator reads $r\,(r+1) - l\,(l+1)$ and is not equal to zero for any odd r.

If l is odd, however, then $y_2(x)$ is of a polynomial form, while $y_1(x)$ will be an infinite series. The two polynomial solutions are directly proportional to the so-called *Legendre polynomials*. ◀

Problem 8.29. *Derive a few more solutions proportional to Legendre polynomials:*

$$[y_1(x)]_{l=4} = 1 - 10x^2 + \frac{35}{3}x^4; \quad [y_2(x)]_{l=3} = x - \frac{5}{3}x^3.$$

Problem 8.30. *Show that the two independent solutions of the DE*

$$2xy'' - y' + 2y = 0$$

are:

$$y_1 = 1 + 2x - \sum_{n=2}^{\infty} \frac{(-2)^n}{n!\,(2n-3)!!}x^n\,, \quad y_2 = x^{3/2}\sum_{n=0}^{\infty} \frac{3\,(-2)^n}{n!\,(2n+3)!!}x^n,$$

where $k!! = 1 \cdot 3 \cdot 5 \cdot \ldots \cdot k$ (k is odd) is a product of all odd integers from 1 to k. Using the ratio test, prove that either series converges for any x.

Problem 8.31. *Show that the series solutions of the DE*

$$8x^2y'' + 10xy' - (1+x)\,y = 0$$

are

$$y_1 = x^{1/4}\left(1 + \frac{x}{14} + \frac{x^2}{616} + \cdots\right)\,, \quad y_2 = x^{-1/2}\left(1 + \frac{x}{2} + \frac{x^2}{40} + \cdots\right).$$

Problem 8.32. *Show that the series solutions of the DE*

$$x^2y'' + 2x^2y' - 2y = 0$$

are

$$y_1 = -1 + x^{-1}\,, \quad y_2 = x^2\left(1 - x + \frac{3x^2}{5} - \frac{4x^3}{15} + \cdots\right).$$

Problem 8.33. *Show that the series solutions of the DE*

$$3xy'' + (3x+1)\,y' + y = 0$$

are

$$y_1 = \sum_{n=0}^{\infty}\frac{(-x)^n}{n!} = e^{-x}\,, \; y_2 = x^{2/3}\left(1 - \frac{3x}{5} + \frac{(-3x)^2}{5\cdot 8} + \frac{(-3x)^3}{5\cdot 8\cdot 11} + \cdots\right).$$

Problem 8.34. *Use the generalised series method to obtain general solutions of the following DEs:*

(a) $x^2y'' + xy' - 9y = 0$; *(b)* $x^2y'' - 6y = 0$; *(c)* $x^2y'' + xy' - 16y = 0$.

[*Answer*: (a) $y = C_1x^3 + C_2x^{-3}$; (b) $y = C_1x^{-2} + C_2x^3$; (c) $y = C_1x^4 + C_2x^{-4}$.]

8.4.3 Special Cases

The procedure described above always gives the general series solution of the differential equation (8.38) about an RSP when the roots of the indicial equation differ by more than an integer. If $s_2 = s_1 + m$, where m is a positive integer, including zero, only one series solution $y_1(x)$ corresponding to s_1 can be found; to find the other solution the method must be modified. Here we shall consider this case in some detail.

Adopting the most general expansion of the functions $p(x)$ and $q(x)$ around the point x_0, we can write:

$$p(x) = \frac{p_{-1}}{x - x_0} + p_0 + p_1 (x - x_0) + p_2 (x - x_0)^2 + \cdots , \tag{8.67}$$

$$q(x) = \frac{q_{-2}}{(x - x_0)^2} + \frac{q_{-1}}{x - x_0} + q_0 + q_1 (x - x_0) + q_2 (x - x_0)^2 + \cdots . \tag{8.68}$$

These expressions take account of the fact that the singular point is regular and that the limits

$$\lim_{x \to x_0} (x - x_0)\, p(x) \quad \text{and} \quad \lim_{x \to x_0} (x - x_0)^2\, q(x)$$

both exist, as otherwise the criteria at the beginning of Sect. 8.4.2 would not be satisfied. Correspondingly we shall consider the differential equation

$$\begin{aligned}(x - x_0)^2 y'' &+ \left[p_{-1} (x - x_0) + p_0 (x - x_0)^2 + \cdots\right] y' \\ &+ \left[q_{-2} + q_{-1} (x - x_0) + q_0 (x - x_0)^2 + \cdots\right] y = 0,\end{aligned} \tag{8.69}$$

which we obtained from (8.38) upon multiplication on $(x - x_0)^2$ after employing explicit expansions of $p(x)$ and $q(x)$ given above.

Problem 8.35. *Substitute*

$$y(x) = (x - x_0)^s \left[1 + c_1 (x - x_0) + c_2 (x - x_0)^2 + \cdots\right]$$

into Eq. (8.69). Then, collecting terms with like powers of x and setting the coefficients to zero, derive the following equation for the number s:

$$f(s) = s(s-1) + p_{-1}s + q_{-2} = s(s - 1 + p_{-1}) + q_{-2} = 0. \tag{8.70}$$

Note that above we have defined a function $f(s)$. Correspondingly, show that the two roots of this quadratic equation, s_1 and s_2, are related via:

$$s_1 + s_2 = 1 - p_{-1}. \tag{8.71}$$

(continued)

Problem 8.35 (continued)

Next, derive equations for the coefficients c_1, c_2*:*

$$c_1 f(s+1) + p_0 s + q_{-1} = 0,$$

$$c_2 f(s+2) + c_1 \left[(s+1) p_0 + q_{-1}\right] + p_1 s + q_0 = 0.$$

In fact, show that an equation for c_n $(n = 1, 2, 3, \ldots)$ *generally has the form:*

$$c_n f(s+n) + F_n(s) = 0\ , \quad n = 1, 2, 3, \ldots, \tag{8.72}$$

where $F_n(s) = \alpha_n s + \beta_n$ *is a linear polynomial in* s *with the coefficients* α_n *and* β_n *which depend on the expansion coefficients of* $p(x)$ *and* $q(x)$*, as well as on coefficients* $c_1, \ldots, c_{n-1}$.

It is readily seen from Eq. (8.72) that it is only possible to determine the coefficients $\{c_n\}$ if the function $f(s+n) \neq 0$ for *any* integer $n = 1, 2, 3, \ldots$ assuming either $s = s_1$ or $s = s_2$. We shall now see that this condition means that the two roots s_1 and s_2 cannot differ by an integer. Indeed, assume that the smallest root is s_2 and the largest $s_1 = s_2 + m$, where m is a positive integer. Since both roots satisfy (8.70), we have $f(s_1) = 0$ and $f(s_2) = f(s_1 - m) = 0$. When considering Eq. (8.72) for the coefficients c_n corresponding to the solution associated with s_1, we need $f(s_1 + n) \neq 0$ to be satisfied for any $n = 1, 2, 3, \ldots$. This condition is valid since there could only be two roots of the quadratic equation (8.70) and $s_1 + n > s_2$ for any $n = 1, 2, 3, \ldots$. Consider now if it is possible to construct in the same way a solution corresponding to $s_2 = s_1 - m$. In this case we will have to consider Eq. (8.72) with the coefficient to c_n being $f(s_2 + n) = f(s_1 - m + n)$. It is seen that for $n = m$ the function $f(s_2 + m) = f(s_1) = 0$ and hence the corresponding coefficient c_n cannot be determined, i.e. the method fails.

Thus, the method needs to be modified when $s_1 = s_2 + m$, where m is a positive integer. We can also include the case of $m = 0$ here as only one value of s is available (as $s_1 = s_2$) and hence only one generalised series solution can be constructed.

The idea of the modified method is based on Theorem 8.1 which states that the second linearly independent solution can always be found using the formula

$$y_2(x) = y_1(x) \int \frac{1}{y_1^2(x)} \exp\left(-\int p(x)dx\right) dx. \tag{8.73}$$

We shall use this result to find a general form of $y_2(x)$. What we need to do is to use the form (8.67) for $p(x)$ and a general form

$$y_1(x) = (x - x_0)^{s_1} P_1(x - x_0) = u^{s_1} P_1(u) \tag{8.74}$$

for the first solution associated with s_1, to investigate the integral in Eq. (8.73) and hence find the general form of the second solution. In Eq. (8.74) $u = x - x_0$ and $P_1(u)$ is a Taylor expansion of some function of u. From this point on we shall be using u instead of x in both integrals in Eq. (8.73). We can write:

$$\int p(u)du = p_{-1}\ln u + p_0 u + \frac{1}{2}p_1 u^2 + \cdots = p_{-1}\ln u + P_2(u),$$

where $P_2(u)$ is another well-behaved function which hence also accepts a Taylor's series expansion. Several other such functions will be introduced below, we shall distinguish them with a different subscript. Therefore, the exponential function in (8.73) now reads:

$$\exp\left(-\int p(u)du\right) = u^{-p_{-1}}\exp\left(-P_2(u)\right) = u^{-p_{-1}}P_3(u),$$

so that the integrand in Eq. (8.73) becomes:

$$\begin{aligned}\frac{1}{y_1^2}\exp\left(-\int p(u)du\right) &= \frac{u^{-p_{-1}}P_3(u)}{u^{2s_1}P_1^2(u)} = u^{-(2s_1+p_{-1})}P_4(u)\\ &= u^{-(2s_1+p_{-1})}\left(\gamma_0 + \gamma_1 u + \gamma_2 u^2 + \cdots\right).\end{aligned}$$

But, because of Eq. (8.71), we have

$$2s_1 + p_{-1} = 2s_1 + (1 - s_1 - s_2) = 1 + s_1 - s_2 = 1 + m,$$

so that the integrand in Eq. (8.73) can be finally written as:

$$\frac{1}{y_1^2}\exp\left(-\int p(u)du\right) = \frac{P_4(u)}{u^{m+1}} = \frac{\gamma_0}{u^{m+1}} + \frac{\gamma_1}{u^m} + \cdots + \frac{\gamma_m}{u} + \gamma_{m+1} + \gamma_{m+2}u + \cdots,$$

which, when integrated, gives

$$\begin{aligned}\int \frac{1}{y_1^2}\exp\left(-\int p(u)du\right)du &= -\frac{\gamma_0/m}{u^m} - \frac{\gamma_1/(m-1)}{u^{m-1}}\\ &+ \cdots - \frac{\gamma_{m-1}}{u} + \gamma_m \ln u + \gamma_{m+1}u + \frac{1}{2}\gamma_{m+2}u^2 + \cdots\\ &= \gamma_m \ln u + u^{-m}\left[-\frac{\gamma_0}{m} - \frac{\gamma_1}{m-1}u + \cdots\right] = \gamma_m \ln u + u^{-m}P_5(u).\end{aligned}$$

Therefore, the second solution should have the following general form:

$$y_2(x) = u^{s_1} P_1(u) \left[\gamma_m \ln u + u^{-m} P_5(u)\right] = \gamma_m u^{s_1} P_1(u) \ln u + u^{s_1 - m} P_6(u),$$

and hence we can finally write returning back to x:

$$y_2(x) = \gamma_m y_1(x) \ln(x - x_0) + (x - x_0)^{s_2} P_6(x - x_0). \quad (8.75)$$

We see that the second solution should have, apart from the usual second term with s_2, also an additional term containing the product of the first solution and the logarithm function. In fact, the pre-factor γ_m can always be chosen as equal to one as this would simply rescale the function $y_2(x)$.

Problem 8.36. *Obtain both solutions of the zero order Bessel DE*

$$xy'' + y' + xy = 0$$

by employing the generalised series method for $x_0 = 0$. (i) First of all, show that only one value of $s = 0$ exists. (ii) Then demonstrate that the corresponding solution is

$$J_0(x) = 1 - \frac{x^2}{2^2} + \frac{x^4}{2^2 4^2} - \frac{x^6}{2^2 4^2 6^2} + \cdots = \sum_{r=0}^{\infty} \frac{(-1)^r x^{2r}}{2^{2r} (r!)^2}.$$

(iii) Write the second solution in the form $K_0(x) = J_0(x) \ln x + P(x)$, substitute into the DE, use the fact that $J_0(x)$ already satisfies it, and hence derive the following inhomogeneous equation for the function $P(x)$:

$$xP'' + P' + xP = -2J_0'(x).$$

(iv) Finally, obtain a Taylor series expansion for $P(x)$:

$$P(x) = \sum_{r=1}^{\infty} p_r x^{2r} = \frac{x^2}{4} - \frac{3x^4}{128} + \frac{11x^6}{13824} - \cdots \quad \textit{with} \quad p_r = -\frac{p_{r-1}}{4r^2} + \frac{(-1)^{r+1}}{2^{2r} (r!)^2 r},$$

where p_0 can be set to zero. Note that $P(x)$ contains only even powers of x.

8.5 Examples in Physics

8.5.1 Harmonic Oscillator

In many physical problems it is necessary to find a general solution of the following second order linear ODE

$$y'' + \gamma y' + \omega_0^2 y = F_d \sin(\omega t)\,, \tag{8.76}$$

with respect to the function of time $y(t)$.

Depending on the physical problem, the actual meaning of $y(t)$ may be different. A natural example of a physical problem where this kind of equation plays the principal role is, for instance, a harmonic oscillator with friction subjected to a sinusoidal external force; we shall be calling this a forced harmonic oscillator. A more complex variant of this is encountered e.g. in modelling oscillations of an atomic force microscopy (AFM) tip in non-contact AFM experiments (Book II, Sect. 4.8.4). Another example is related to electric circuits, for instance, the one shown in Fig. 8.1. It contains a resistance R, an inductance L and a capacitance C, and a voltage applied to the circuit is some sinusoidal function of time, e.g. $V(t) = V_0 \cos(\omega t)$. Using the second Kirchoff's law, we obtain for the voltage drop along the circuit:

$$L\frac{di}{dt} + Ri + \frac{1}{C}\int_{t_0}^{t} i(t_1)dt_1 = V(t) = V_0\cos(\omega t)\,,$$

where $i(t)$ is the current in the circuit. Differentiating both sides of the equation above, we get:

$$L\frac{d^2 i}{dt^2} + R\frac{di}{dt} + \frac{1}{C}i = -\omega V_0 \sin(\omega t)\,,$$

which is the forced harmonic oscillator equation (8.76) with $\omega_0 = 1/\sqrt{LC}$, $\gamma = R/L$ and $F_d = -\omega V_0/L$. The term with resistance (the one containing the first derivative of the current) corresponds to losses in the circuit (dissipation), while the inductance and capacitance terms work towards current oscillations. The applied voltage induces oscillations with a different frequency ω. It is sometimes called an excitation signal.

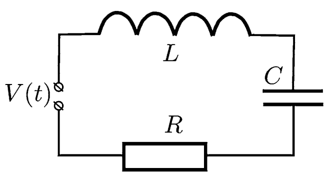


Fig. 8.1 Electrical circuit with a resistance R, an inductance L and a capacitance C connected to a voltage source $V(t)$

8.5.1.1 Harmonic Motion

Consider first the equation without resistance and excitation:

$$y'' + \omega_0^2 y = 0. \tag{8.77}$$

Its solution,

$$y(t) = A\sin(\omega_0 t) + B\cos(\omega_0 t),$$

corresponds to a sinusoidal oscillations with frequency ω_0. Indeed, after some rearrangements,

$$\begin{aligned} y(t) &= \sqrt{A^2+B^2}\left[\frac{A}{\sqrt{A^2+B^2}}\sin(\omega_0 t) + \frac{B}{\sqrt{A^2+B^2}}\cos(\omega_0 t)\right] \\ &= \sqrt{A^2+B^2}\left[\cos(\phi)\sin(\omega_0 t) + \sin(\phi)\cos(\omega_0 t)\right] = \sqrt{A^2+B^2}\sin(\omega_0 t+\phi), \end{aligned} \tag{8.78}$$

where $\tan\phi = B/A$. Therefore, the solution of the homogeneous equation (8.77) can always be written as

$$y(t) = D\sin(\omega_0 t + \phi) \tag{8.79}$$

with two arbitrary constants: D is called the amplitude and ϕ phase. This solution corresponds to, e.g., a *self-oscillation* in a circuit.

8.5.1.2 Driven Oscillator with No Friction

An interesting effect happens when we add an excitation signal with $\omega \neq \omega_0$:

$$y'' + \omega_0^2 y = F_d \sin(\omega t). \tag{8.80}$$

This is a harmonically driven oscillator without friction (or a circuit without resistance). In this case the harmonic oscillation (8.79) represents only a complementary solution as there must be a particular solution as well.

Problem 8.37. *Using the trial function* $y = B_1\sin(\omega t) + B_2\cos(\omega t)$ *for Eq. (8.80), show that* $B_1 = -F_d/\left(\omega^2-\omega_0^2\right)$ *and* $B_2 = 0$*, so that the general solution becomes*

$$y(t) = D\sin(\omega_0 t+\phi) - \frac{F_d}{\omega^2-\omega_0^2}\sin(\omega t). \tag{8.81}$$

We see from the above problem that there are two harmonic motions: one due to self-oscillation (angular frequency ω_0) and another due to applied excitation that happens with the same frequency ω as the excitation signal itself; its amplitude B_1 depends crucially on the difference between the squares of the two frequencies. This dependence is such that when the two frequencies are close, the amplitude of the particular integral solution becomes increasingly large, i.e. the excited oscillation dominates for excitation frequencies ω which are close to ω_0.

If the two frequencies are very close, but are not exactly the same, say $\omega = \omega_0 + \epsilon$ with small ϵ, then the motion overall represents *beats*. Indeed, assume that initially there was no motion in our oscillator, i.e. $y(0) = y'(0) = 0$. Then, from Eq. (8.81) it follows that

$$D\sin(\phi) = 0 \quad \text{and} \quad D\omega_0\cos(\phi) - \frac{F_d\omega}{\omega^2 - \omega_0^2} = 0.$$

Note that D cannot be equal zero as in this case the second equation above would not be satisfied. Thus, from the first equation we get $\phi = 0$, and hence from the second equation we obtain

$$D = \frac{F_d\omega}{\omega_0\left(\omega^2 - \omega_0^2\right)},$$

and thus

$$\begin{aligned}
y(t) &= \frac{F_d}{\omega_0\left(\omega^2 - \omega_0^2\right)}\left[\omega\sin(\omega_0 t) - \omega_0\sin(\omega t)\right] \\
&= \frac{F_d}{\omega_0\left((\omega_0+\epsilon)^2 - \omega_0^2\right)}\left[(\omega_0+\epsilon)\sin(\omega_0 t) - \omega_0\sin(\omega_0 t + \epsilon t)\right] \\
&\simeq \frac{F_d}{2\omega_0^2\epsilon}\left[\omega_0\sin(\omega_0 t) - \omega_0\sin(\omega_0 t + \epsilon t)\right] = \frac{F_d}{2\omega_0\epsilon}\left[\sin(\omega_0 t) - \sin(\omega_0 t + \epsilon t)\right] \\
&= -\frac{F_d}{\omega_0\epsilon}\sin\left(\frac{\epsilon t}{2}\right)\cos\left(\omega_0 t + \frac{\epsilon t}{2}\right) \simeq -\frac{F_d}{\omega_0\epsilon}\sin\left(\frac{\epsilon t}{2}\right)\cos(\omega_0 t)\,.
\end{aligned} \tag{8.82}$$

To obtain the first expression in the last line we have used the trigonometric identity (2.49). The signal $y(t)$ in Eq. (8.82) can be considered as a harmonic signal $B(t)\cos(\omega_0 t)$ of frequency ω_0 and the amplitude $B(t) \sim \sin(\epsilon t/2)$ that also changes in time harmonically with a much smaller frequency (larger period). The effect of this is demonstrated in Fig. 8.2 and is called *amplitude modulation*.

The above consideration is only valid if $\omega \neq \omega_0$. When the excitation frequency is the same as the self-oscillation one, the trial function for the particular solution has to be modified (called a "resonance", which we talked about in Sect. 8.2.2.1).

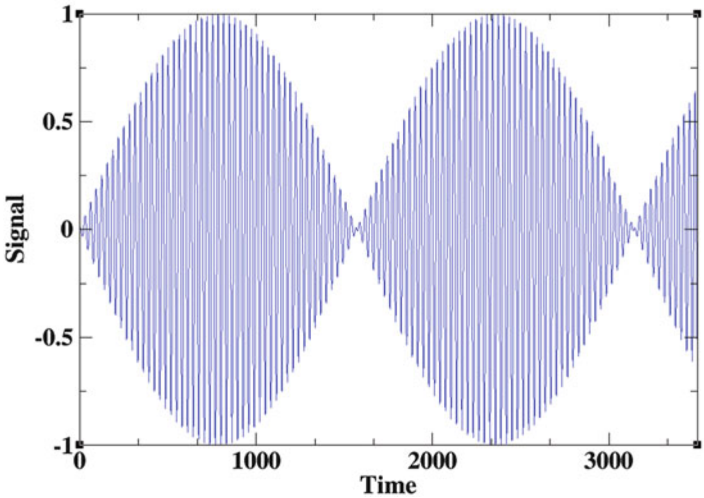


Fig. 8.2 Amplitude modulation: Eq. (8.82) plotted with $-F_d/\omega_0\epsilon = 1$, $\epsilon = 0.004$ and $\omega_0 = 0.2$

Problem 8.38. *Using the trial function*

$$y_p(t) = t\,[B_1 \sin(\omega_0 t) + B_2 \cos(\omega_0 t)]$$

in Eq. (8.80) with $\omega = \omega_0$, show that in this case $B_1 = 0$ and $B_2 = -F_d/2\omega_0$, so that the whole solution reads

$$y(t) = D \sin(\omega_0 t + \phi) - \frac{F_d t}{2\omega_0} \cos(\omega_0 t)\,. \tag{8.83}$$

We see from Eq. (8.83) that the amplitude of the contribution due to the excitation signal increases indefinitely with time, independently of the initial conditions (that are responsible for the values of the arbitrary constants D and ϕ), then the system is said to be at *resonance*. The frequency ω_0 is sometimes called the *resonance angular frequency*.

Actually, the infinite increase of the amplitude at resonance frequency can be anticipated from Eq. (8.82) as well since

$$\frac{1}{\epsilon} \sin\left(\frac{\epsilon t}{2}\right) = \frac{t}{2} \frac{\sin(\epsilon t/2)}{\epsilon t/2} \to \frac{t}{2} \quad \text{when} \quad \epsilon \to 0.$$

Note also that at resonance the second term in the solution (8.83) is shifted by $-\pi/2$ with respect to the driving force itself, compare $F_d \sin(\omega t)$ in Eq. (8.80) and $-\cos(\omega_0 t)$ in Eq. (8.83). Therefore, at resonance there is a shift of $-\pi/2$ in phase between the driving force and the oscillating signal following it.

8.5.1.3 Harmonic Oscillator with Friction

Now let us consider a self-oscillating system with friction (e.g. the resistance R is present in the circuit in Fig. 8.1):

$$y'' + \gamma y' + \omega_0^2 y = 0. \tag{8.84}$$

The characteristic equation $p^2 + \gamma p + \omega_0^2 = 0$ in this case has two solutions:

$$p_{1,2} = \frac{1}{2}\left(-\gamma \pm \sqrt{\gamma^2 - 4\omega_0^2}\right), \tag{8.85}$$

and hence there are three cases to consider.

Under-Damping: If $\gamma^2 - 4\omega_0^2 < 0$, i.e. if the friction is small, $\gamma < 2\omega_0$, then $p_{1.2} = -\gamma/2 \pm iw$ are complex conjugates, where $w = \sqrt{\omega_0^2 - (\gamma/2)^2}$, and the general solution is then

$$\begin{aligned} y(t) &= C_1 e^{(-\gamma/2+i\omega)t} + C_2 e^{(-\gamma/2-i\omega)t} = e^{-\gamma t/2}\left[C_1 e^{i\omega t} + C_2 e^{-i\omega t}\right] \\ &= A e^{-\gamma t/2} \sin(\omega t + \phi), \end{aligned} \tag{8.86}$$

where A and ϕ is a set of arbitrary constants replacing the set C_1 and C_2. In this case the motion is easily recognisable as an oscillation with frequency w (which may be very close to ω_0 if $\gamma \ll \omega_0$); however, its amplitude is exponentially decreasing as shown in Fig. 8.3(a) providing an exponentially decaying envelope for the sinusoidal motion.

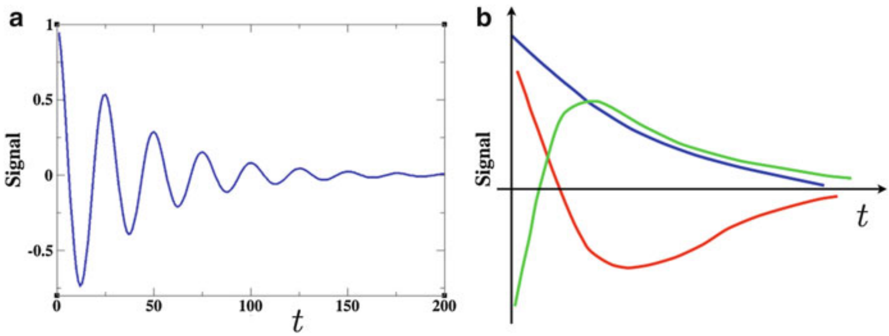


Fig. 8.3 (**a**) Under-dumped motion with $A = 1$, $\phi = 0$, $\gamma = 0.1\,\mathrm{s}^{-1}$ and $w = 0.5\ \mathrm{s}^{-1}$. (**b**) Over-dumping for various initial conditions and values of the parameters

Over-Dumping: If $\gamma^2 - 4\omega_0^2 > 0$, i.e. if the friction is large, $\gamma > 2\omega_0$, then both $p_{1.2}$ are real and negative:

$$y(t) = C_1 e^{p_1 t} + C_2 e^{p_2 t}.$$

The oscillations are *not* observed, the motion comes to a halt very quickly. Typical behaviours of the solution in this case are shown in Fig. 8.3(b) by different curves.

Critical Dumping: When $\gamma^2 - 4\omega_0^2 = 0$, i.e. if the friction value is transitional between the two regimes, the two roots $p_{1,2}$ coincide and the solution has the form

$$y(t) = (C_1 + C_2 t)\, e^{-\gamma t/2},$$

and is similar in shape to the over-dumped case.

8.5.1.4 Driven Harmonic Oscillator with Friction

Finally, we shall consider the most common case of the sinusoidally driven dumped harmonic oscillator described by Eq. (8.76). Its complementary solution $y_c(t)$ was considered in the previous subsection and corresponds to either under-dumped, over-dumped or critically dumped case. However, in either case the *transient* motion described by $y_c(t)$ dies away either slowly or quickly (depending on the value of the friction γ). Therefore, if one is interested in a solution corresponding to long times, the complementary solution could be disregarded and one has to consider the particular integral only. The particular solution will survive until, of course, the driving force ceases to be applied.

Problem 8.39. *Using the trial function*

$$y_p = A \sin(\omega t) + B \cos(\omega t)$$

in Eq. (8.76), show that

$$A = \frac{F_d \Delta}{\Delta^2 + (\gamma\omega)^2} \quad \text{and} \quad B = -\frac{F_d \gamma\omega}{\Delta^2 + (\gamma\omega)^2}, \quad \text{where} \quad \Delta = \omega_0^2 - \omega^2.$$

We know from Sect. 8.5.1.1 (see Eq. (8.78)) that the sum of sine and cosine functions in y_p can be rewritten using a single sine function as:

$$y_p(t) = D \sin(\omega t + \phi)\,, \tag{8.87}$$

where the amplitude is

$$D = \sqrt{A^2 + B^2} = \frac{F_d}{\sqrt{\Delta^2 + (\gamma\omega)^2}} = \frac{F_d}{\sqrt{\left(\omega^2 - \omega_0^2\right)^2 + (\gamma\omega)^2}} \tag{8.88}$$

and the phase ϕ is defined via

$$\tan\phi = \frac{B}{A} = -\frac{\gamma\omega}{\Delta} = \frac{\gamma\omega}{\omega^2 - \omega_0^2}. \tag{8.89}$$

Thus, the signal remains sinusoidal with the frequency of the driving force; it is called the *steady-state solution* since it continues to hold without change as long as the driving force is acting.

The situation is similar to the driven harmonic oscillator without friction considered in Sect. 8.5.1.2. Therefore, we may anticipate that the amplitude of the steady-state motion will depend critically on the interplay between the two frequencies ω and ω_0. And, indeed, as shown in Fig. 8.4(a), the amplitude demonstrates a maximum (resonance) at some ω that is close to ω_0 at small γ.

Problem 8.40. *More precisely, show that D is maximum when ω is equal to the* resonance frequency

$$\omega_{res} = \sqrt{\omega_0^2 - \frac{\gamma^2}{2}}. \tag{8.90}$$

We see that the resonance frequency is very close to the fundamental frequency ω_0 of the harmonic oscillator when the friction is small ($\gamma \ll \omega_0$).

The phase as a function of ω is shown in Fig. 8.4(b). It approaches the value of $-\pi/2$ at resonance (note that $\omega_{res} < \omega_0$ so that $\tan\phi$ is negative and in fact very close to $-\infty$ for small γ).

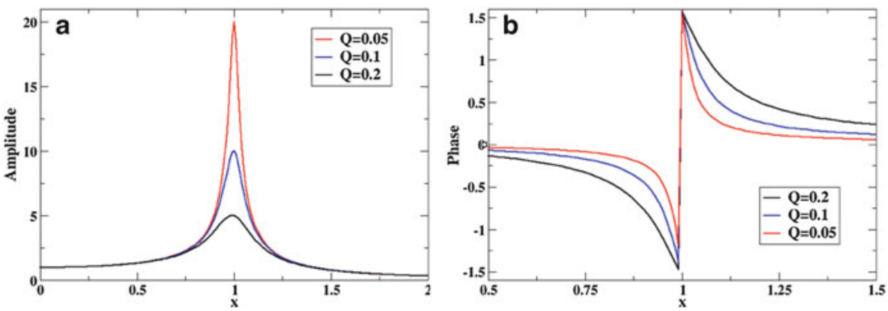


Fig. 8.4 Amplitude D (**a**) and the phase ϕ (**b**) of the steady-state solution (8.87) of the damped driven harmonic oscillator as functions of the relative driving frequency $x = \omega/\omega_0$. Here $Q = \gamma\omega_0$

8.5.2 Falling Water Drop

Consider a drop of water of mass m falling down from a cloud under gravity. The equation of motion for the drop is

$$m\frac{dv}{dt} = mg - \gamma v,$$

where γ is a friction coefficient (due to air resistance), g the gravity constant (acceleration in the gravity field of the Earth), v the particle vertical velocity (positive if directed downwards). We shall obtain its general solution and also its particular solution assuming that the initial velocity of the particle was $v_0 = 0$.

This ODE can be solved by rearranging terms and integrating both sides:

$$\frac{dv}{g - \frac{\gamma}{m}v} = dt \implies \int \frac{dv}{g - \frac{\gamma}{m}v} = \int dt \implies -\frac{m}{\gamma}\int \frac{dv}{v - \frac{gm}{\gamma}} = t - C_1$$
$$\implies \frac{m}{\gamma}\ln\left(v - \frac{gm}{\gamma}\right) = -t + C_1 \implies v - \frac{gm}{\gamma} = Ce^{-\gamma t/m},$$

where $C = e^{\gamma C_1/m}$ is another constant.[3] Thus, a general solution for the velocity of the drop is

$$v(t) = Ce^{-\gamma t/m} + \frac{gm}{\gamma}.$$

The particular solution is now obtained if we find the constant C from the additional information given about the velocity, i.e that we know its initial value: $v(0) = v_0$. This gives:

$$v(0) = v_0 = C + \frac{gm}{\gamma} \implies C = v_0 - \frac{gm}{\gamma}.$$

If $v_0 = 0$, then $C = -gm/\gamma$, so that we finally obtain:

$$v(t) = \frac{gm}{\gamma}\left(1 - e^{-\gamma t/m}\right).$$

Thus, initially the drop has the velocity $v_0 = 0$. Then the velocity is increased; however, after some time the acceleration slows down (due to air resistance) and the velocity approaches a constant value of $v_\infty = gm/\gamma$, called terminal velocity. In fact, the exponential term becomes very small rather quickly, so that almost all the way down the drop will fall approximately with the constant velocity v_∞. Note that it is positive since it is directed downward.

[3] As usual, there is no need to keep the old arbitrary constant if it comes inside some function. This will be just another arbitrary constant!

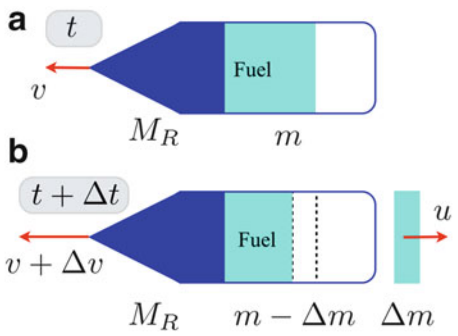


Fig. 8.5 For the derivation of the Tsiolkovsky equation. (**a**) At time t the rocket has $m(t)$ of fuel, while (**b**) after time Δt when the mass Δm of the fuel is burned off, the mass of the fuel left will be $m - \Delta m$. The rocket gets a bust Δv in its velocity, while Δm mass of the fuel escapes from the rocket in the opposite direction with the relative velocity u

8.5.3 Tsiolkovsky's Formula

Consider the ideal rocket equation first derived by Tsiolkovsky. We consider an idealised rocket of initial mass $M_0 = M_R + m_0$, where M_R is the mass of the rocket without fuel and m_0 is the initial mass of all fuel (in fact, normally the fuel determines the initial mass at the launch, i.e. $m_0 \gg M_R$). We assume that the fuel escapes with a constant velocity u with respect to the rocket (called the "exhaust velocity"). If $v(t)$ and $M(t) = M_R + m(t)$ are the velocity of the rocket and its total mass at time t ($m(t)$ is the mass of the fuel left at this time), then the equation of motion of the rocket will be given by the second Newton's law $\Delta P/\Delta t = F_{ext}$, where ΔP is the change of the whole system momentum over time Δt and F_{ext} is an external force. Considering system's momentum at two times, t and $t + \Delta t$, we can write for ΔP (see Fig. 8.5):

$$\Delta P = P(t + \Delta t) - P(t) = [(M - \Delta m)(v + \Delta v) - (u - v)\Delta m] - Mv,$$

where $u - v$ stands for the actual velocity of the fuel in a fixed coordinate frame of the Earth (note that u is the *relative* velocity of the escaping fuel with respect to the rocket). Neglecting the second order term $\Delta m \Delta v$ and keeping only the first order terms, we get

$$\Delta P = M\Delta v - u\Delta m = M\Delta v + u\Delta M,$$

where ΔM (negative) is the change of mass of the whole rocket over time Δt. Balancing this change of the momentum with the external force of Earth gravity, $-Mg$, acting in the opposite direction to the (positive) direction of the rocket movement, we obtain:

$$M\frac{\Delta P}{\Delta t}=-Mg \quad \Longrightarrow \quad M\frac{\Delta v}{\Delta t}+u\frac{\Delta M}{\Delta t}=-Mg \quad \Longrightarrow \quad M\frac{dv}{dt}=-u\frac{dM}{dt}-Mg, \tag{8.91}$$

which is the equation of motion sought for.

This equation can be solved if we know how the fuel is burned off. Assuming, for instance, that it is burned with the constant rate proportional to the initial mass of the rocket, $dM/dt = -\gamma M_0$ with some positive γ, we write $dv/dt = (dv/dM)\,(dM/dt) = -\gamma M_0\,(dv/dM)$, which allows us to obtain the equation we can solve:

$$-\gamma M_0 M\frac{dv}{dM}=u\gamma M_0 - Mg.$$

Upon integration between $t=0$ (when $v(0)=0$ and $M(0)=M_0$) and t, we obtain:

$$\begin{aligned} v &= \int_{M_0}^{M}\left(-\frac{u}{M}+\frac{g}{M_0\gamma}\right)dM = -u\ln\frac{M}{M_0}+\frac{g}{\gamma M_0}\left(M-M_0\right) \\ &= u\ln\frac{M_0}{M}-\frac{g}{\gamma M_0}\left(M_0-M\right). \end{aligned}$$

This formula is valid until the whole fuel is burned off. The maximum velocity is achieved when the whole fuel is exhausted, in which case $M = M_R$, giving

$$v_{\max}=u\ln\frac{M_R+m_0}{M_R}-\frac{g}{\gamma}\frac{m_0}{M_R+m_0}\simeq u\ln\frac{m_0}{M_R}-\frac{g}{\gamma},$$

where in the last passage we used the fact that most of the mass of the rocket is due to the fuel. The last equation also allows one to obtain the amount of fuel,

$$m_0 = M_R\exp\left(\frac{v+g/\gamma}{u}\right),$$

required for the rocket to reach the required velocity v on the orbit (e.g. the escape velocity v_{esc} needed to overcome the Earth gravity).

8.5.4 Distribution of Particles

Let us consider a random distribution of particles in a volume as shown in Fig. 8.6(a). We would like to calculate the probability $dp = p(r)dr$ of finding a closest particle to a given particle. Let us draw a sphere of radius r around our particle. Then, the probability to find a particle anywhere within that sphere is given

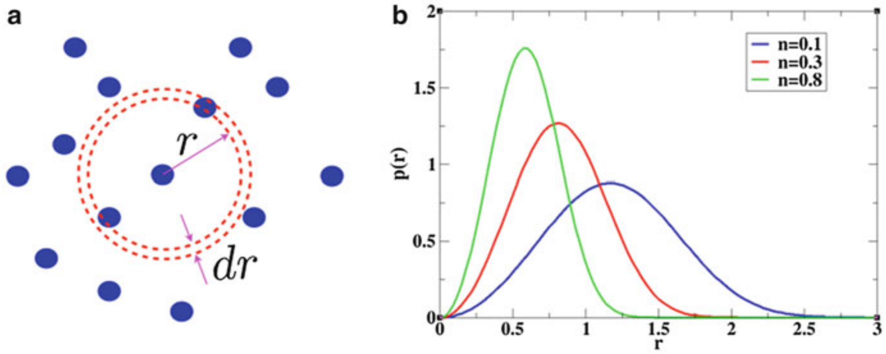


Fig. 8.6 (**a**) To the derivation of the probability to find a closest neighbour around a chosen particle in a volume: we draw a spherical shell of radius r and width dr around the particle. (**b**) Probabilities $p(r)$ for three values of the particles concentration $n = 0.1, 0.3, 0.8$

by summing up all such probabilities for all radii from zero to r:

$$P(r) = \int_0^r p\left(r'\right) dr'.$$

Now, the probability to find the closest particle within the spherical shell of radius r and width dr will be given by a product of two probabilities: the probability that no particles are found inside the sphere, $Q(r) = 1 - P(r)$, and the probability ndV that there is a particle within the spherical shell. Here $dV = 4\pi r^2 dr$ if the shell volume and n is the particle concentration. On the other hand, this is exactly the definition of the probability $p(r)dr$. Therefore, we can write:

$$p(r)dr = [1 - P(r)]\, 4\pi r^2 n dr \quad \Longrightarrow \quad \frac{p(r)}{r^2} = \left[1 - \int_0^r p\left(r'\right) dr'\right] 4\pi n. \tag{8.92}$$

To obtain an equation for the probability $p(r)$ we now differentiate Eq. (8.92) with respect to r which yields:

$$\frac{d}{dr}\left(\frac{p}{r^2}\right) = 4\pi n \frac{d}{dr}\left[1 - \int_0^r p\left(r'\right) dr'\right] \quad \Longrightarrow \quad \frac{1}{r^2}\frac{dp}{dr} - \frac{2p}{r^3} = -4\pi n p,$$

which can be rewritten as

$$\frac{dp}{dr} = \left(\frac{2}{r} - 4\pi r^2 n\right) p. \tag{8.93}$$

This ODE is separable and can be easily integrated:

$$\int \frac{dp}{p} = \int \left(\frac{2}{r} - 4\pi r^2 n \right) dr \quad \Longrightarrow \quad \ln p = 2\ln r - \frac{4\pi}{3} r^3 n + C$$

$$\Longrightarrow \quad p = Cr^2 \exp\left(-\frac{4\pi}{3} r^3 n \right).$$

The arbitrary constant C is determined by normalising the distribution as there will definitely be a particle found somewhere around the given particle:

$$\int_0^\infty p(r)dr = 1 \quad \Longrightarrow \quad C = 4\pi n \quad \Longrightarrow \quad p(r) = 4\pi n r^2 \exp\left(-\frac{4\pi}{3} r^3 n \right). \tag{8.94}$$

This is the final result which is shown in Fig. 8.6(b) for three values of the particles concentration n. It shows that initially $p(r)$ peaks up, but then decays to zero at large r, i.e. there is always the most probable distance between particles (the maximum of $p(r)$) which is reduced with the increase of the concentration. This distribution is used in various fields of physics including nucleation theory, cosmology (relative distribution of starts in a galaxy), etc.

Problem 8.41. *Show that the two-dimensional analogue of formula (8.94) is:*

$$p(r) = 2\pi n r \exp\left(-\pi n r^2\right).$$

8.5.5 Residence Probability

Here we shall derive the *residence kinetics*, i.e. we shall consider a system, which can change its current state (i.e. make a transition to another state) with a certain rate R. For instance, this could be a molecule on a crystal surface that can be in either of two states (conformations), *cis* and *trans* (differing by the orientation of a certain group of atoms). The states are separated by an energy barrier, and hence both are stable; the molecule may change its state, however, by jumping from one state to another, but these events are not deterministic, i.e. they happen with a certain probability depending on the transition rates between the two states and the temperature. Another example is an atomic diffusion on a surface of a crystal. Suppose the lowest energy state of the adatom is above a surface atom; since there are many surface atoms, the adatom may occupy many adsorption positions. Since each of them is stable, there must be an energy barrier for the adatom to overcome in order to jump from one position to the other. The rate of such jumps depends on the energy barrier and temperature.

Our goal is to derive a probability $P(t)$ for a system to remain in the current state by the time t. The formula we are about to derive lies in the foundation of a very powerful and popular simulation method called *kinetic Monte Carlo* (KMC).

If R is the rate (probability per unit time) for the system to move away (escape) from the given state it is currently in, then the probability for this to happen over time dt is Rdt. The probability for this *not to happen* over the same time is $1 - Rdt$. Therefore, the probability $P(t + dt)$ for the system to remain in the current state up to the time $t + dt$ is a product of two probabilities corresponding to two events: (i) the system is at this state by the time t with the probability $P(t)$, and (ii) the system remains in the same state over the consecutive time dt, the probability of this is $1 - Rdt$. Therefore,

$$P(t + dt) = P(t)\,(1 - Rdt) \quad \Longrightarrow \quad P(t + dt) - P(t) = -RP(t)dt$$
$$\Longrightarrow \quad \frac{dP}{dt} = -RP.$$

This ODE is easily solved giving $P(t) = C\exp(-Rt)$. Since at $t = 0$ our system is in the current state with probability one, then $C = 1$. Therefore, the required result is:

$$P(t) = \exp(-Rt)\,.$$

This formula has a very simple meaning: the probability to remain in the current state decreases exponentially with time since the longer we wait the higher the probability for the system to escape from the current state. Eventually the probability to remain in the current state approaches zero, i.e. the system will definitely escape at some point in time.

Problem 8.42. *Generalise the previous treatment to the case when the escape rate depends on time, $R = R(t)$. Show that in this case*

$$P(t) = \exp\left(-\int_0^t R(\tau)d\tau\right).$$

8.5.6 Defects in a Crystal

Consider defects in a crystal that can be found in two states A and B; this could be, e.g., a metastable atom. The state A lies higher in energy than B (i.e. the state A is less energetically favourable than B, so that there should be on average more of defects B than A), but it is relatively stable. The energy barriers for a defect to overcome jumping between states, $A \to B$ and $B \to A$, are Δ_{AB} and Δ_{BA}, respectively, so that

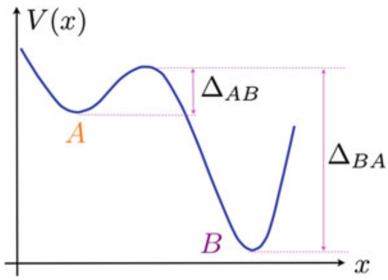


Fig. 8.7 Double well potential energy surface with two minima: A and B. The barriers Δ_{AB} and Δ_{BA} separating these states are also indicated

transition probabilities per unit time (the so-called *transition rates* w_{AB} or w_{BA}) for the two transitions can be assumed to be of the Arrhenius type*:*

$$w_{AB} = \nu e^{-\Delta_{AB}/k_B T} \text{ and } w_{BA} = \nu e^{-\Delta_{BA}/k_B T},$$

where w_{AB} is the transition rate to jump from state A to B, while w_{BA} corresponds to the reverse transition, T is temperature and ν the so-called attempt frequency (a pre-factor). The exponential term gives a probability to overcome the barrier during a single jump (attempt) that is increased with an increase of temperature T or a decrease of the energy barrier (Δ_{AB} or Δ_{BA}), while ν gives the number of attempts per unit time to jump.

The corresponding double well potential energy surface is shown in Fig. 8.7. There will be certain concentrations n_A and n_B of defects in the two states at equilibrium that should be established over some time. Assuming that initially all defects were created in a metastable state A, we can determine their concentration at time t in both states. Indeed, the time dependence of the defects concentrations in the two states should satisfy the following "equations of motion":

$$\frac{dn_A}{dt} = -w_{AB}n_A + w_{BA}n_B \text{ and } \frac{dn_B}{dt} = w_{AB}n_A - w_{BA}n_B. \tag{8.95}$$

The first equation states that the change of the defects concentration in state A is due to two competing processes. Firstly, there is an incoming flux of defects $w_{BA}n_B$ jumping from the state B which is proportional to the existing concentration n_B of defects in state B; this term works to increase n_A (the "gain" term). Secondly, there is also an outgoing flux $-w_{AB}n_A$ from the state A, which is proportional to n_A, that works to reduce their concentration in this state (the "loss" term). The meaning of the second equation is similar, but this time the balance is written for the defects in the other state.

Actually, since the total concentration of the defects $n = n_A + n_B$ is constant (and indeed, summing up the two equations we get that $dn/dt = 0$), the two equations are completely equivalent, so that only one need to be used, e.g. the first:

$$\frac{dn_A}{dt} = -w_{AB}n_A + w_{BA}\left(n - n_A\right). \tag{8.96}$$

It is easily seen that this equation is exactly of the form (8.9):

$$\frac{dn_A}{dt} + wn_A = w_{BA}n \quad \text{with} \quad w = w_{AB} + w_{BA}.$$

Therefore, using the general method of Sect. 8.1.5, we obtain

$$I(t) = \exp\left(\int wdt\right) = e^{wt}$$

and

$$n_A(t) = e^{-wt}\left(C + \int w_{BA}ne^{wt}dt\right) = Ce^{-wt} + e^{-wt}\frac{w_{BA}n}{w}e^{wt} = \frac{w_{BA}n}{w} + Ce^{-wt}.$$

Since initially $n_A(0) = n$ (all defects were in the A state), then, applying this initial condition, we obtain:

$$n = \frac{w_{BA}n}{w} + C \implies C = n - \frac{w_{BA}n}{w},$$

and thus

$$n_A(t) = \frac{w_{BA}n}{w} + \left(n - \frac{w_{BA}n}{w}\right)e^{-wt}.$$

This is the final result. It shows that the concentration of defects in state A is reduced over time approaching the equilibrium value (at $t = \infty$) of $n_A(\infty) = n\omega_{BA}/\omega$. Consequently, the concentration of defects in state B

$$n_B(t) = n - n_A(t) = \left[n - n_A(\infty)\right]\left(1 - e^{-wt}\right)$$

is initially zero, but then it is increased reaching at equilibrium the value of $n - n_A(\infty)$. Note that equilibrium concentrations can be obtained immediately from Eq. (8.95) by setting the time derivatives to zero.

Index

L. Kantorovich, *Mathematics for Natural Scientists*, Undergraduate Lecture Notes in Physics, DOI 10.1007/978-1-4939-2785-2

W