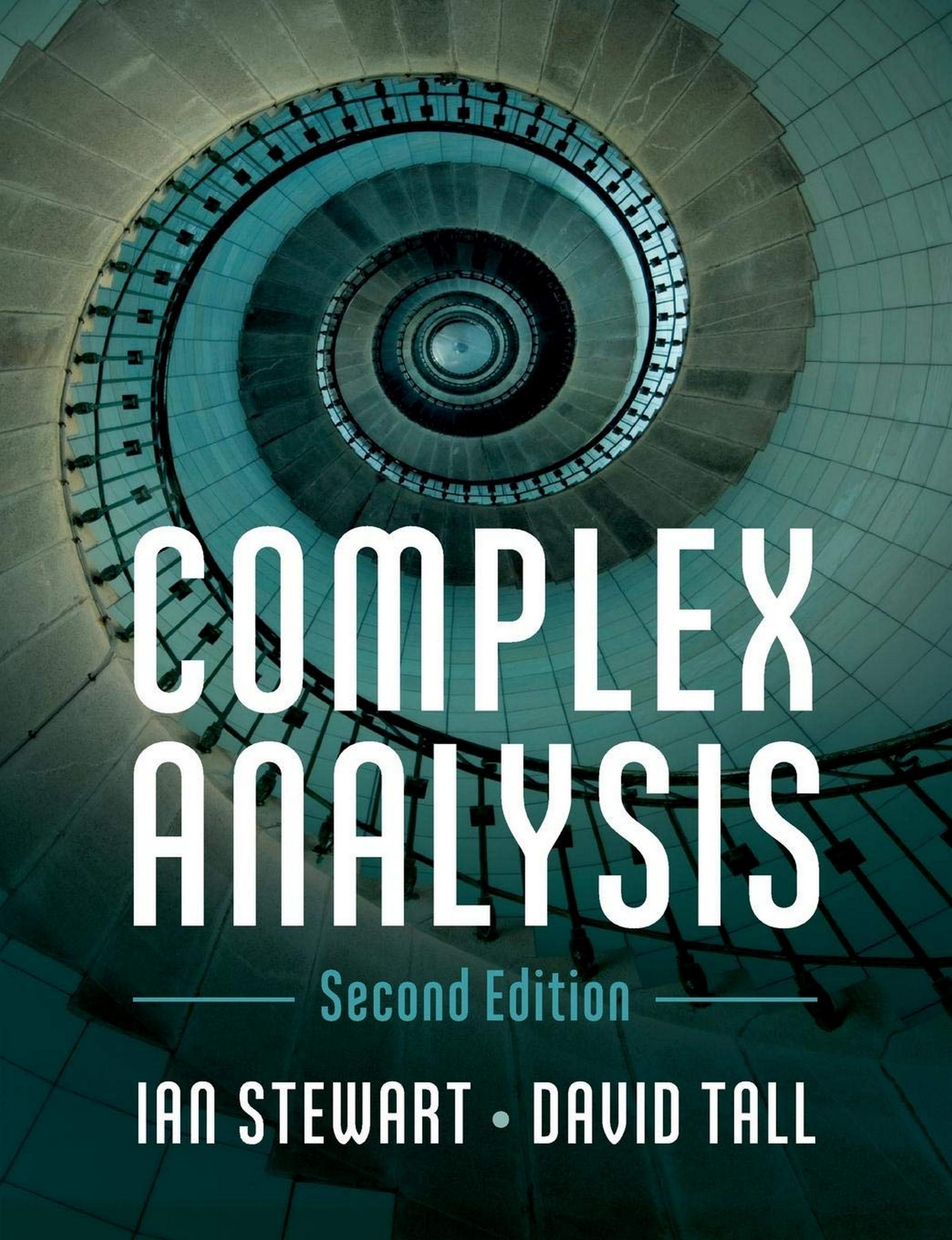

COMPLEX ANALYSIS

Second Edition

IAN STEWART • DAVID TALL

Complex Analysis

Second edition

This new edition of a classic textbook develops complex analysis from the established theory of real analysis by emphasising the differences that arise as a result of the richer geometry of the complex plane. Key features of the authors' approach are to use simple topological ideas to translate visual intuition into rigorous proof, and, in this edition, to address the conceptual conflicts between pure and applied approaches head-on.

Beyond the material of the clarified and corrected original edition, there are three new chapters: Chapter 15 on infinitesimals in real and complex analysis; Chapter 16 on homology versions of Cauchy's Theorem and Cauchy's Residue Theorem, linking back to geometric intuition; and Chapter 17 outlines some more advanced directions in which complex analysis has developed, and continues to evolve into the future.

With numerous worked examples and exercises, clear and direct proofs, and a view to the future of the subject, this is an invaluable companion for any modern complex analysis course.

Ian Stewart, FRS is Emeritus Professor of Mathematics at the University of Warwick. He is author or coauthor of over 190 research papers and is the bestselling author of over 120 books, from research monographs and textbooks to popular science and science fiction. His awards include the Royal Society's Faraday Medal, the IMA Gold Medal, the AAAS Public Understanding of Science Award, the LMS/IMA Zeeman Medal, the Lewis Thomas Prize, and the Euler Book Prize. He is an honorary wizard of the Discworld's Unseen University.

David Tall is Emeritus Professor of Mathematical Thinking at the University of Warwick and is known internationally for his contributions to mathematics education. He is author or coauthor of over 200 papers and 40 books and educational computer software, covering all levels from early childhood to research mathematics.

Complex Analysis

(The Hitch Hiker's Guide to the Plane)

Second edition

IAN STEWART

DAVID TALL

University of Warwick

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108436793
DOI: 10.1017/9781108505468

First published 1983
Thirteenth reprint 2004
Second edition 2018

Printed in the United Kingdom by TJ International Ltd, Padstow, Cornwall 2018

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data
Names: Stewart, Ian, 1945– author. | Tall, David Orme, author.
Title: Complex analysis : the hitch hiker's guide to the plane / Ian Stewart (University of Warwick), David Tall (University of Warwick).
Other titles: Hitch hiker's guide to the plane
Description: Second edition. | Cambridge : Cambridge University Press, 2018. | Includes bibliographical references and index.
Identifiers: LCCN 2018007009 | ISBN 9781108436793 (alk. paper)
Subjects: LCSH: Functions of complex variables. | Numbers, Complex. | Geometry, Analytic–Plane.
Classification: LCC QA331 .S85 2018 | DDC 515/.9–dc23
LC record available at https://lccn.loc.gov/2018007009

ISBN 978-1-108-43679-3 Paperback

Additional resources for this publication at www.cambridge.org/Stewart&Tall2ed

Contents

Preface to the Second Edition

The first edition of *Complex Analysis* focused on generalising concepts from real analysis to the complex case. Where there were differences, we looked at the geometric picture to see why they were happening. This second edition does the same, but it also focuses on the increasing sophistication of mathematical ideas as we build from intuition to rigour, in a manner where greater understanding leads to more sophisticated intuitions and ways of working. New concepts and methods often start out in a technical way, with problematic aspects that conflict with intuition. As well as generalising real analysis, we move beyond it by addressing these conceptual conflicts, resolving them, and providing more sophisticated concepts and methods appropriate to complex analytic functions.

This approach is used throughout the book. So, for example, the text now includes a short (but complete) discussion of the construction of a space-filling curve, to challenge our intuition about continuity and to explain why we have had to be careful with topological assertions that appear obvious. The treatment here is simpler than most of the literature on space-filling curves. We have spent some time examining different notions of a path, especially the role of smoothness.

We have added three new chapters. Chapter 15 introduces ideas about infinitesimals in real and complex analysis, thought of as variables that tend to zero, and formulated as elements of extensions of the real and complex fields. Chapter 16 gives a formal link from analysis back to geometric intuition, formulating and proving homology versions of Cauchy's Theorem and Cauchy's Residue Theorem. Chapter 17 outlines a few of the more advanced directions in which complex analysis has developed, and continues to evolve into the future.

Chapter 15 has been added for the following reasons. Since the first edition appeared in 1983, the ways in which we operate mathematically have changed dramatically. Not only are there computers that perform numerical and symbolic operations at a speed way beyond that previously available to the individual mind; there are also interactive graphics drawn on high-resolution screens that let us visualise mathematical ideas in completely new ways. In particular, we can dynamically magnify pictures to see tiny detail that lets us represent 'arbitrarily small' quantities.

This second edition therefore includes an extra chapter to introduce formally defined infinitesimals that lie in an ordered extension field K of the real numbers, which can be manipulated algebraically and visualised formally on an extended number line. This approach generalises to the complex case using the field $K(\mathrm{i})$ where $\mathrm{i}^2 = -1$, which can be visualised in the extended complex plane. This construction offers a meaningful bridge between the epsilon-delta rigour of pure mathematics and the intuitive use of infinitesimals in applications.

It can easily be shown that any proper ordered field extension K of the reals must contain infinitesimal elements x: that is, elements that are not zero yet satisfy $|x| < r$ for all positive real numbers r. Using the completeness of the real numbers, we prove a simple theorem that any finite element of K has the form $k = c + h$, where c is real and h is infinitesimal or zero. A transformation in the form $m(x) = (x - c)/\varepsilon$, where ε is a positive infinitesimal, then lets us magnify infinitesimal detail near c and see it with our unaided human eyes in a real picture. This technique extends to the complex case in the field $K(\mathrm{i})$.

We can now illustrate why complex analysis is so different from real analysis. A differentiable complex function defined on an open set is locally expressible as a power series, and we may take K to be the smallest ordered extension field generated by a single infinitesimal ε. The elements are power series $\sum_{r\geq n} a^r \varepsilon^r$ in ε with possibly a finite number of terms in $1/\varepsilon$, and each non-zero element has an order of infinitesimality n related to the first non-zero coefficient a_n (where the element may be infinite if n is negative). Meanwhile a differentiable real function may be differentiable once but not twice, and this requires a much more sophisticated extension field K such as that given by the logical theory of non-standard analysis. While Gottfried Leibniz imagined infinitesimals of different orders, non-standard analysis fails to have this property and requires a much more sophisticated construction. At the end of the chapter we compare and contrast the various theories within a single framework.

Chapter 16 on homology complements Chapter 9 on homotopy versions of Cauchy's Theorem, and logically it could have been placed immediately after that. We postpone it to the penultimate chapter because we do not wish to delay the more practical payoff from Cauchy's Theorem – Taylor and Laurent series, residues, evaluation of integrals, summation of series, and so on.

Homology can be thought of as a way of characterising 'holes' in a topological space, which here is the domain of a complex function f. Singularities, where f is not differentiable, create such holes, and homology helps to describe the topological effect of singularities; for example, in the homology version of Cauchy's Residue Theorem. To avoid including big chunks of algebraic topology, our approach to homology is based on step paths in open subsets of the plane, one of the main simplifying tools in this book. The proof is 'bare hands' and exploits the simple geometry of step paths and the abelian group structure of homology.

Chapter 17 has been included to make it clear that complex analysis is still a major area of mathematical research. Complete though the classical theory may seem to be, there are numerous generalisations and new questions. The main topics mentioned are the Riemann Hypothesis, modular functions, several complex variables, complex manifolds, and complex dynamics – leading to the fractal geometry of Julia sets and the famous Mandelbrot set.

In this new edition of *Complex Analysis* we have corrected all known typographical errors, simplifed some proofs, and reorganised the material in mostly harmless ways to improve readability. We have brought the text and layout into line with current practice, and redrawn all the figures. Proofs, definitions, and examples are terminated with the ‘end of proof’ symbol □. The same symbol indicates the absence of a proof when the result is clear or has already been proved. Contrary to the prevailing wisdom, we do not insert punctuation marks at the end of displayed formulas. (Your tutors may object to this. Tradition is on their side. If they do, they can set you an extra exercise: *insert all missing punctuation.*) But it is now the twenty-first century. No one puts full stops (US: periods) at the end of book titles, or chapter or section headings. So why do this in displayed formulas, where it may cause confusion because punctuation marks are also often part of the symbolism? We suggest that clean typography should override pedantic punctuation.

Formulas in the main text are another matter; here the *absence* of punctuation can cause confusion. We have followed tradition here.

Online Supplementary Material

Supplementary material including a concordance showing in more detail the changes between the previous edition and this one, and links to *GeoGebra*, can be found on the Cambridge University Press website: www.cambridge.org/Stewart&Tall2ed.

Preface to the First Edition

Students faced with a course on 'Complex Analysis' often find it to be just that – complex. In the sense of 'complicated'.

It's true, of course, that the proofs of some of the major theorems in the subject can demand a certain technical versatility. But in many ways, on a conceptual level, complex analysis is actually *easier* than real analysis; it just isn't always taught that way.

This book is intended for use at the level of second or third year undergraduates, and it is based on experience accumulated from teaching such courses over the past decade. To exhibit the inherent simplicity of complex analysis we have organised the material around two basic principles: (1) generalise from the real case, and (2) when that reveals new phenomena, use the rich geometry of the plane to understand them. Our aim throughout is to encourage geometric thinking, with the proviso that it must be adequately backed by analytic rigour.

The opening chapter sets the work in its historical context, and the history is often alluded to later as partial motivation. However, we feel that cultural changes often affect the status of conceptual problems: what was once an important difficulty can become a triviality when viewed with hindsight. It is not always necessary to drag today's students through yesterday's hang-ups. We argue the point at greater length below: it is fundamental to our entire approach.

0 The Origins of Complex Analysis, and Its Challenge to Intuition

In a lecture in 1886, Leopold Kronecker asserted that the integers are made by God and all the rest is the work of Man (Gray [7]). If so, complex numbers are certainly one of humanity's most intriguing mathematical artefacts. For centuries they have been a wonder to mathematicians and philosophers alike. It took nearly 300 years from their first appearance in Girolamo Cardano's *Ars Magna* (The Great Art) to the publication of a formal definition that satisfies modern standards of rigour. Building on such foundations, the initiated reader might be forgiven for thinking that complex analysis must be an incredibly complicated theory. Yet here we come to a historical puzzle. Although it took nearly three centuries to obtain a satisfactory treatment of complex *numbers*, it then took less than a tenth of that time to complete a major part of complex *analysis*, which is far more sophisticated and extensive.

Obviously the numbers must come first, or there is nothing to do analysis with, but the timescale is surprising. A possible explanation is that setting up the foundations adequately involved deep problems of a philosophical nature: it took a long time to come to grips with them, but once the 'breakthrough' had occurred, the further development was easy by comparison.

History suggests otherwise.

0.1 The Origins of Complex Numbers

Cardano's celebrated *Ars Magna* of 1545 is one of the most important early algebra texts. Diophantus's *Arithmetica* of about 250 discussed the solution of equations and introduced a rudimentary form of algebraic notation. Muhammad al-Khwarizmi's *Al-kitab al-mukhtasar fi hisab al-gabr wa'l-muqabala* (The Compendious Book on Calculation by Completion and Balancing) appeared around 820. Its translation into Latin as *Liber Algebrae et Almucabola* gave us the word 'algebra'. Al-Khwarizmi's discussion was verbal, with no symbols but occasional diagrams.

Cardano introduced a systematic algebraic notation, very different from what we use today. He used this to present the newly discovered solutions of cubic and quartic equations. His book contained the solution of cubics discovered by Scipione del Ferro around 1500, and independently by Niccolo Fontana (nicknamed 'Tartaglia', the stammerer) around 1535. The high point of the text is the solution of quartic equations found by Cardano's student Lodovico Ferrari. The tangled tale of alleged duplicity and public controversy that accompanied these discoveries can be found in Stewart [19, 20] and other historical sources.

Ars Magna also discussed the simultaneous equations

$$x + y = 10$$
$$xy = 40$$

and obtained a solution (in modern notation) of the form

$$x = 5 + \sqrt{-15} \qquad y = 5 - \sqrt{-15}$$

Cardano gave no interpretation for the square root of a negative number, but he did observe that, on the assumption that the quantities obey the usual algebraic rules, we can check that they satisfy the equations. His attitude to the discovery was dismissive: 'So progresses arithmetic subtlety, the end of which . . . is as refined as it is useless.'

In the same book he observed that applying Tartaglia's formula to the cubic equation

$$x^3 = 15x + 4 \tag{0.1}$$

leads to the solution

$$x = \sqrt[3]{2 + \sqrt{-121}} + \sqrt[3]{2 - \sqrt{-121}}$$

in contrast to the obvious answer $x = 4$.

In both instances there was a conflict between the intuition about numbers that mathematicians had built up over the years, and the formal behaviour of the symbolic manipulations that Cardano was carrying out. It took centuries for mathematicians to extend the number concept and develop a refined intuition in which Cardano's observations make sense. The first step happened not long after, however. Raphael Bombelli (1526–73) suggested a way to reconcile the two solutions of (0.1) by manipulating the 'impossible' roots *as if they are ordinary numbers*. Since

$$(2 \pm \sqrt{-1})^3 = 2 \pm \sqrt{-121}$$

Cardano's expression becomes

$$x = (2 + \sqrt{-1}) + (2 - \sqrt{-1}) = 4$$

and the 'impossible' root is just the familiar root in a complex disguise. Bombelli's work was the first hint that complex numbers can prove useful in solving real mathematical problems. But the message took a long time to sink in.

In *La Géometrie* (1637), René Descartes made the distinction between 'real' and 'imaginary' numbers, interpreting the occurrence of imaginaries as a sign that the problem concerned is insoluble, an opinion shared by Isaac Newton at a later date. However, this view sits uneasily with Bombelli's realisation that a formula involving complex numbers sometimes leads to a real solution, suggesting that the issue is not that simple.

John Wallis [25] represented a complex number geometrically in his *Algebra* of 1685. On a fixed line the real part of the number was measured off (in the direction given by its sign); then the imaginary part was measured off at right angles, Figure 1. But this idea was largely forgotten.

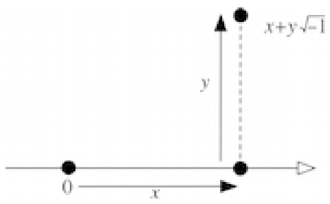

Figure 1 Wallis's geometric representation of a complex number.

In 1702 John Bernoulli was evaluating integrals of the form

$$\int \frac{\mathrm{d}x}{ax^2 + bx + c}$$

by partial fractions. Using the philosophy that complex numbers can be manipulated like real ones, he wrote the integrand as

$$\frac{1}{ax^2 + bx + c} = \frac{A}{x - \alpha} + \frac{B}{x - \beta}$$

(using modern notation) where α, β are the roots of the quadratic denominator, and obtained the integral in the form

$$A \log(x - \alpha) + B \log(x - \beta)$$

His bold decision to use the same method when the quadratic had no real solutions led to logarithms of complex numbers. But what were they? Both Bernoulli and Leibniz used the method, and by 1712 they were engaged in controversy. Leibniz asserted that the logarithm of a negative number is complex, while Bernoulli insisted it is real. Bernoulli argued that, since

$$\frac{\mathrm{d}(-x)}{-x} = \frac{\mathrm{d}x}{x}$$

it follows by integration that $\log(-x) = \log(x)$. Leibniz, on the other hand, insisted that the integration was correct only for positive x. Once again, formal calculations that seemed sensible were in conflict with intuition.

Leonhard Euler resolved the controversy in favour of Leibniz in 1749, pointing out that integration requires an arbitrary constant

$$\log(-x) = \log(x) + c$$

a point that Bernoulli had ignored. By formally manipulating expressions involving complex numbers, Euler derived a host of theoretical relations, including the famous formula of 1748:

$$\mathrm{e}^{\mathrm{i}\theta} = \cos\theta + \mathrm{i}\sin\theta \tag{0.2}$$

Putting $\theta = \pi$ we find

$$\mathrm{e}^{\mathrm{i}\pi} = -1 \tag{0.3}$$

a fantastic relation that blends the three mathematical symbols e, i, and π in one surprising equation. The formula (0.3) is widely referred to as Euler's formula, although he never published it explicitly. He did publish (0.2), of which it is a simple corollary, and this is also known as Euler's formula. However, a formula equivalent to (0.2) had been found earlier by Roger Cotes in 1714.

Extending the theory of logarithms to the complex case by defining

$$\log z = w \text{ if and only if } e^w = z$$

we obtain other intriguing results. Formal manipulation gives

$$e^{\log z + m\pi i} = e^{\log z}(e^{\pi i})^m = z \cdot (-1)^m$$

For an even integer $m = 2n$ this gives

$$e^{\log z + 2n\pi i} = z$$

So $\log z + 2n\pi i$ is also a logarithm of z: the complex logarithm is *many-valued.* For an odd integer $m = 2n + 1$ we have

$$e^{\log z + (2n+1)\pi i} = -z$$

whence

$$\log(-z) = \log z + (2n + 1)\pi i$$

This resolves the Leibniz–Bernoulli controversy: if x is real and positive, then $\log(-x)$ must be complex.

As mathematicians refined their intuition to encompass complex numbers, everything started to fit together and make sense. The theory of complex numbers grew ever more fascinating. What was lacking was an interpretation that explained precisely what these entities are – a formal counterpart to the newly extended intuitions.

In 1797 Caspar Wessel published a paper in Danish describing the representation of a complex number as a point in the plane. It went almost totally unnoticed until a French translation was published a hundred years later. Meanwhile the idea was attributed to Jean-Robert Argand, who wrote it up independently in 1806. Since that time the geometric interpretation of complex numbers has commonly become known as the Argand diagram.

Another pioneer of the theory of complex numbers was Carl Friedrich Gauss. In his doctoral dissertation of 1799 he addressed a problem that had concerned mathematicians since the early eighteenth century. Initially it had been widely believed that, just as the solutions of real quadratic equations could lead to new 'complex' numbers, so would solutions of equations with complex coefficients lead to even more kinds of new numbers. But Jean d'Alembert (1717–83) conjectured that complex numbers alone suffice. Gauss confirmed this in the 'fundamental theorem of algebra' – every polynomial equation has a complex root. At first he proved it in the purely real form that any real polynomial factorises into linear and quadratic factors, avoiding explicit use of imaginaries; later he treated the general case. By 1811 he viewed the complex numbers as points in the plane, saying so in a letter to Friedrich Bessel. In 1831 he published full details of his representation of complex numbers, which had begun to acquire an air of respectability.

In 1837, nearly three centuries after Cardano's use of 'imaginary numbers', William Rowan Hamilton published the definition of complex numbers as ordered pairs of real numbers subject to certain explicit rules of manipulation. (In the same year Gauss wrote to Wolfgang Bolyai that he had developed the same idea in 1831.) At last this placed the complex numbers on a firm *algebraic* basis.

0.2 The Origins of Complex Analysis

Unlike the gradual emergence of the complex *number* concept, the development of complex *analysis* seems to have been the direct result of the mathematician's urge to generalise. It was sought deliberately, by analogy with real analysis. However, the mathematicians of the period tended to assume that everything in real analysis must automatically be meaningful in the complex case, so the main question must be how 'the' complex version behaves. That there might not *be* a complex version, or several alternatives, was seldom appreciated, as the controversy over $\log(-x)$ illustrates.

As noted above, there are early traces of analytic operations on complex functions in the work of Bernoulli, Leibniz, Euler, and their contemporaries.

In his 1811 letter to Bessel, Gauss shows that he knew the basic theorem on complex integration around which complex analysis was subsequently built. In real analysis, when we integrate a function f between limits a and b, to get

$$\int_a^b f(x)\mathrm{d}x$$

the limits fully specify the integral. But in the complex case, where a and b represent points in the plane, it is also necessary to specify a definite path from a to b, and to 'integrate along the path'. The question is: to what extent does the value of the integral depend on the chosen path?

Gauss says:

> I affirm now that the integral $\int f(x)\mathrm{d}x$ has only one value even if taken over different paths, provided $f(x)$ . . . does not become infinite in the space enclosed by the two paths. This is a very beautiful theorem whose proof . . . I shall give on a convenient occasion.

It seems the occasion never arose. The crucial step of publishing a proof of this result was taken in 1825 by the man who was to occupy centre stage during the first flowering of complex analysis: Augustin-Louis Cauchy. After him, this result is called 'Cauchy's Theorem'. In Cauchy's hands the basic ideas of complex analysis rapidly emerged. For a complex function to be differentiable, it must have a very specialised nature: its real and imaginary parts must satisfy certain properties called the Cauchy–Riemann Equations. He showed that contour integrals of differentiable functions have the property noted privately by Gauss. Further, if an integral is computed along a path that winds round points where the function becomes infinite, Cauchy showed how to compute this integral using the 'theory of residues'. The latter requires no more than the calculation of a constant, called the 'residue' of the function, at each exceptional point, and knowing how many times the paths winds around that point. The precise route of the path does not matter at all – only how it winds round these exceptional points.

Power series turned out to be important in the theory, and other workers extended these ideas. Pierre-Alphonse Laurent introduced 'Laurent series' involving negative powers in 1843. In this formulation, near an exceptional point z_0, a differentiable function is expressed as a sum of two series

$$f(x) = [a_0 + a_1(z - z_0) + \cdots + a_n(z - z_0)^n + \cdots]$$
$$+[b_1(z - z_0)^{-1} + \cdots + a_n(z - z_0)^{-n} + \cdots]$$

The residue of $f(z)$ at $z = z_0$ is then just the coefficient b_1. Using the theory of residues, the computation of complex integrals often proved to be far simpler than could ever have been dreamed.

Cauchy's definition of analytic ideas such as continuity, limits, derivatives, and so on, were not the same as those we use today. He based them on infinitesimal notions, which fell into disrepute in the late nineteenth century – though recent developments in 'non-standard analysis', and a new theory we present in Chapter 15, show that we may have been over-hasty in judging Cauchy's ideas. Moreover, Cauchy's concept of 'infinitesimal' was a variable quantity that approaches zero as closely as we please, not a fixed quantity. See Tall and Katz [24] for detailed discussion and educational implications.

A rigorous treatment was devised by Karl Weierstrass (1815–97) using definitions which are still regarded as fundamental, the 'epsilon-delta' formulation. Weierstrass founded his whole approach on power series. However, the geometric viewpoint was sorely lacking in his work (at least as published). This deficiency was remedied by far-reaching ideas introduced by Bernhard Riemann (1826–66). In particular, the concept of a 'Riemann surface', which dates from 1851, treats many-valued functions by splitting the complex plane into multiple layers, on each of which the function is single-valued. The crucial feature is how the layers join up topologically.

From the mid-nineteenth century onwards, the progress of complex analysis has been strong and steady, with many far-reaching developments. The fundamental ideas of Cauchy remain, now refined and clothed in more recent topological language. The abstruse invention of complex numbers, once described by our mathematical forebears as 'impossible' and 'useless', has become part of an aesthetically satisfying theory with eminently practical applications in aerodynamics, fluid mechanics, electronics, control theory, and many other areas.

Since the first edition of this book, formal theory has also evolved so that Cauchy's ideas of infinitesimals can be visualised as points on an extended number line, which we describe in our new Chapter 15.

0.3 The Puzzle

We return to our historical puzzle. *Why was the development of complex numbers so laboured and hesitant, whereas that of complex analysis was explosive?* We suggest a possible answer (only personal opinion and thus open to dispute). It is somewhat different from the 'foundations + breakthrough' explanation offered earlier.

Looking at the early history of complex numbers, the overall impression is of countless generations of mathematicians beating out their brains against a brick wall in search of – what? A triviality. The definition of complex numbers as ordered pairs of points (x, y), or as points in the plane, was obtained over and over and over again. It is even implicit in Bombelli's work; it is there for all to see in Wallis's; it crops up again by way of Wessel, Argand, and Gauss. Morris Kline remarks on page 629 of [11]:

> That many men – Cotes, de Moivre, Euler, and Vandermonde – really thought of complex numbers as points in the plane follows from the fact that all, in attempting to solve $x^n - 1 = 0$, thought of solutions . . . as the vertices of a regular polygon.

If the problem has such a simple solution, why was this not recognised sooner?

Perhaps the early mathematicians were not so much seeking a *construction* for complex numbers as a *meaning*, in the philosophical sense: 'what *are* complex numbers?' However, the development of complex *analysis* showed that the complex number concept was so useful that no mathematician in his right mind could possibly ignore it. The unspoken question became 'what can we *do* with complex numbers?', and once that had been given a satisfactory answer, the original philosophical question evaporated. There was no jubilation at Hamilton's incisive answer to the 300-year old foundational problem – it was 'old hat'. Once mathematicians had woven the notion of complex numbers into a powerful coherent theory, the fears that they had concerning the existence of complex numbers became unimportant, because mathematicians lost interest in that issue.

There are other cases of this nature in the history of mathematics, but perhaps none is more clear-cut. As time passes, the cultural world-view changes. What one generation sees as a problem or a solution is not interpreted in the same way by a later generation. It is worth bearing this in mind when thinking about the historical development of mathematics. To interpret history solely from the viewpoint of the current generation may easily lead to distortion and misinterpretation.

What this explanation omits is any discussion of *why* mathematicians lost interest in the meaning of complex numbers. And that leads to a question that sheds a different light on the historical development, which we now discuss.

0.4 Is Mathematics Discovered or Invented?

Students trying to understand new concepts are in a similar position to the pioneers who first investigated them. At any stage in our education, we build not just on our current knowledge, but on a variety of beliefs and intuitions that are often vague, and may not be consciously recognised. As a trivial example, children familiar with counting numbers may find it hard to adapt their thinking to negative numbers, or rational numbers. When faced with questions like 'what is 3 minus 7?' or 'what is 3 divided by 7', intuition based solely on whole numbers leads to the answer 'can't be done'. That makes it hard to understand -4 or $3/7$. In fact, these is not really trivial examples, because the world's top mathematicians, centuries ago, were just as confused by the question 'what is the square root of minus one?' Even their terminology – 'imaginary' – reveals how puzzled they were. Intuitively they considered numbers to be 'real' – not in the sense we now use to distinguish real from complex, but as direct representations of real measurements. The new objects behaved like numbers in many ways, but they seemed not to correspond directly to reality.

In such circumstances, it can be tempting to discard existing intuition completely. But it is more sensible to adapt the intuition to fit the new circumstances. It is much easier to do arithmetic with negative numbers or fractions if you remember how to do it with whole numbers; it is much easier to do algebra with complex numbers if you bear in mind how to do it with real numbers. So the trick is to sort out which aspects of existing intuition remain valid, and which need to be refined into a broader kind of understanding.

One way to approach this issue is to take seriously a question that is often asked but seldom answered satisfactorily: is mathematics discovered or invented? One answer is to dismiss the question, and agree that neither word is entirely appropriate; moreover, they are not mutually exclusive. Most discoveries have elements of invention, most inventions have elements of discovery. Galileo would not have discovered the moons of Jupiter without the invention of the telescope. The telescope could not have been invented without discovering that sand could be melted to make glass.

But leaving such quibbles aside, we can make a rough distinction between discovery, which is finding something that is *already there* but has not hitherto been noticed, and invention, which is a creative act that brings into being something that has not previously existed. There is a case to be made that in this sense, mathematicians invent new concepts but then discover their properties. For example, complex integration is all about 'paths' in the complex plane. Intuitively, a path is a line drawn by moving the hand so that the pencil remains in contact with the paper – no jumps. We might choose to formalise this notion as a continuous curve – the image of a continuous map from a real interval to the complex plane. We might be interested in how the pencil point moves along this curve, which requires the map itself, not just its image. Sometimes we might wish the path to be smooth – to have a well-defined tangent.

As it happens, we need all of these notions. Intuitively, they are all based on the same mental image. Formally, they are all very different. They have different definitions, different meanings, and different properties. A smooth path always has a meaningful length, for instance; a continuous path may not. The definitions we settle on in this book fit conveniently into the standard ideas of analysis, but they are not built into the fabric of the universe. We chose them, and by so doing we invent concepts such as 'path', 'curve', and 'smooth'.

On the other hand, once a concept has been invented, we cannot invent its *properties*. When we also invent the concept 'length', we discover that every smooth path has finite length. We cannot 'invent' a theorem that the length of a smooth path can be infinite. If we weaken 'smooth' to 'continuous', however, we can discover that infinite lengths are possible; indeed, 'length' need not have a sensible meaning at all. In short: invention opens up new mathematical territory, but exploring it leads to discoveries. We may not know what things are present in the territory, but we do not get to choose them.

Sometimes – in fact, very often – we discover that our inventions have features that we neither expected nor intended them to have. We discover, perhaps to our dismay, that the image of a smooth path can have a right-angled corner, see Section 6.7. We did not expect that: a corner does not feel 'smooth'. But its possibility is a direct consequence of the definition we invented.

When this kind of thing happens, we have two choices. Accept the surprises as the price for having a nice, tidy definition; or rule them out by changing the definition – inventing a more comfortable alternative. In practice we often do both, by giving the alternative a different name. Here we could (and do) define a 'regular path' to be a smooth path $\gamma : [a, b] \to \mathbb{C}$ for which $\gamma'(t) \neq 0$ whenever $t \in [a, b]$. Now the image cannot have a sharp corner. On the other hand, every theorem about regular paths must now take account of the consequences of that extra condition. We also have to remember that some theorems may be valid for regular paths but not for smooth paths, and so on.

As we move from intuitive ideas to formal ones, we also refine our intuition so that it matches the formal theory better. Formal calculations start to make sense, not just as strings of symbols that follow from previous strings, but as meaningful statements that agree with our new intuitions. From this point of view, the history of complex analysis is the story of intuition co-evolving with an increasingly formal approach. This suggests that mathematicians lost interest in the meaning of complex numbers when they incorporated them into their intuitive assumptions and beliefs. With the apparent conflicts resolved by these refined intuitions, they were free to push the subject forward, no longer worried that it did not make logical sense.

When a mathematical area 'settles down' into a mature theory, there is a broad consensus that certain concepts provide the most convenient route through the material. These concepts then become standard – things like 'continuous', 'connected', and so on. They get taught in lecture courses and printed in books. We may start to feel that the standard definitions are the only reasonable ones. Even so, we are always free to work with different concepts if that seems sensible, or even to modify definitions while retaining the same name – though that can be dangerous. Today's concept of continuity is quite different from what it was in the time of Euler, but we use the same word; we just bear in mind that it now has a specific technical meaning. A historian reading Euler would need to be on their guard.

It is also worth remarking that many mathematical concepts seem more natural to us than others. Counting numbers are very natural (we even call them the 'natural numbers'). The number i was baffling for centuries (and was called 'imaginary' as a result). Our culture, our society, and even our senses, predispose us towards certain concepts. Euclid's points and lines correspond to early stages of the processing of images sent from the retina to the visual cortex. Newton's concept of acceleration being related to an applied force reflects the way our ears sense accelerations and make us 'feel' a push – a force.

It then becomes easy to imagine that mathematics somehow already exists in a realm outside the natural world. Even if humans invented numbers, in retrospect they seem such a natural idea that surely they were just hanging around waiting to be invented. If so, that is more like discovery. This view is often called Platonism: the idea that mathematical concepts already exist in some ideal form in some kind of world outside the physical universe, and mathematicians merely discover how these ideal forms work. The contrary view is that mathematics is a shared human construct, *but* that construct is by no means arbitrary, because every new invention is made in the context of existing knowledge, and every new discovery must be logically valid.

A major theme of this book is that many apparently puzzling aspects of complex analysis can be made more intuitive by paying attention to the geometry of the complex plane (in a broad sense, including its topology). This brings one of the human brain's most powerful abilities, visual intuition, into play. For this reason, we draw a lot of pictures. However, a picture, and our visual intuition, can be misleading unless we examine the unstated assumptions that they involve. By doing so, we can refine out intuition and make it more reliable. For this reason, we do not just introduce important definitions and then deduce theorems that refer to them. We try to relate those definitions to intuition, to make the proofs easier to understand. Then we exhibit some of the positive results that arise, to convince you that the new concept is worth considering. And then . . . we show you that sometimes the formally defined concept does *not* behave the way intuition might suggest. Sometimes it turns out to be useful to strengthen the definition so that it matches intuition more closely. Sometimes we refine our intuition so that it matches the formal definition. Sometimes we can even do both, in which case we have to make some careful but useful distinctions.

The historical events sketched earlier in this chapter offer many examples of this process. The square root of minus one went from being a puzzling idea that seemed to have no meaning to one of the most important concepts in the whole of mathematics. Along the way, mathematicians' intuition for 'number' underwent a revolution. We can now to some extent short-circuit the historical debates – what were hang-ups then need not be hang-ups now – but when a new idea puzzles us, and doesn't seem to make sense until we finally sort it out, it is helpful to remember that the mathematical pioneers often experienced exactly the same feelings, for much the same reasons.

0.5 Overview of the Book

It is often useful to set the development of a mathematical theory in its historical context, but it is not always necessary to fight the historical battles again. In this text we give honour where we can to those pioneers who carved their way through uncharted mathematical territory. But more recent developments let us see the theory itself in a new light. To the modern ear the very *name* 'complex analysis' carries misleading overtones: it suggests complexity in the sense of complication. The older meaning, 'composite', was perhaps appropriate when the 'real part' of a complex number had a quite different status from that of the 'imaginary part'. But nowadays a complex number is a perfectly integrated whole. To think of complex analysis as if it were, so to speak, two copies of real analysis, is to place undue emphasis on the algebra at the expense of the geometry, which in the long run has been far more influential. And in fact complex numbers are *not* more complicated than reals: in some ways, they are simpler. For instance, polynomials always have roots. Likewise, complex analysis is often simpler than real analysis: for example, every differentiable function is differentiable as often as we please, and has a power series expansion.

In preparing our approach to the subject we have adopted two basic organising principles. The first is the direct generalisation of real analysis to the complex case. Definitions, of limits, continuity, differentiation, and integration are natural extensions of the corresponding real notions. Since nowadays any student taking a course in complex analysis may be assumed to have made a study of the real counterpart, *many battles have already been won*. We can refer students to their accumulated knowledge, pausing only to phrase it appropriately. This saves time and energy, allowing us to proceed straight to the heart of the subject, where the interesting differences occur. Invariably this happens because the plane has a richer geometry than the line, and this leads to our second major organising principle: geometric insight is valuable and should be cultivated. Of course this insight must be translated into sound formal arguments; this can often be done using modern topological notions.

From these two principles, a straightforward approach to complex analysis emerges. First, complex numbers are defined formally as ordered pairs of real numbers, giving them a geometric interpretation as points in the plane. The topology of complex numbers is then a natural consequence of plane topology. In quick succession it is possible to derive complex generalisations of the notions of continuity, limits, and differentiation, with particular emphasis on power series, which play a central role later. A study of the complex exponential function, defined by the usual power series, reveals the intimate connection between this function and the trigonometric functions (also considered as power series). After generalising the notion of integration, the logarithm can be viewed either as the inverse function of the exponential, or as the integral

$$\log z = \int \frac{\mathrm{d}z}{z}$$

suitably interpreted. Either approach has to deal with the multivalued nature of the complex logarithm. This arises because the complex exponential has period $2\pi\mathrm{i}$, so cannot be one-one. Resolving these issues involves close links between geometric intuition and formal analysis.

At this stage Cauchy's Theorem is presented in various guises, and the use of integration leads to a proof that every differentiable function can be expressed as a power series. More generally, Laurent series (using positive and negative powers) take care of isolated points where functions become infinite, and lead to the powerful 'theory of residues' for calculating complex integrals, summing series, and counting zeros of equations in a given region of the complex plane.

Returning to geometric ideas, complex analysis has many practical applications. Today it is widely used by physicists and engineers, in many different contexts. In particular, it has proved invaluable in two-dimensional potential theory. The geometric ideas of Riemann can be viewed in terms of modern topology, to give a global insight into 'many-valued' functions (such as the logarithm) and open up new areas of progress.

In this second edition of the book, we continue by presenting a formal set-theoretic approach to infinitesimals that has evolved since the first edition was published 35 years ago. It offers a new vision of complex analysis that includes both the analytic epsilon-delta approach of Riemann and the infinitesimal ideas of Cauchy in a broader overall theory.

Next, we revisit Cauchy's Theorem in the context of homology theory. Homology is a topological property of the domain of the function, and it detected the presence of holes. These holes are obstacles that cause integrals of complex functions to depend on the chosen path. Using step paths, we reformulate complex integration over 'cycles' in a domain. These are formal integer combinations of closed loops, so they form an abelian group. The subgroup of 'boundaries' has the property that the integral of any continuous function over a boundary is zero. So the difference between cycles and boundaries controls how integrals depend on the choice of path. The corresponding algebraic object is the quotient group of the group of cycles modulo the subgroup of boundaries, and this is the (first) homology group of the domain. It provides a formal algebraic interpretation of how integrals depend on the choice of path. This chapter provides a gentle introduction to homology in its simplest (old-fashioned) form, though even this approach requires some mathematical sophistication. The topological ideas shed light on the general area surrounding Cauchy's Theorem.

Finally, to show that complex analysis is still alive and kicking in the modern era, Chapter 17 provides a simplified overview of a few more recent developments. These include the still-unsolved Riemann Hypothesis, modular functions, generalising complex analysis to several variables (where strange new phenomena occur), to complex manifolds (multidimensional 'surfaces' with a complex structure, generalising Riemann surfaces), and the iteration of complex maps, or complex dynamics, which leads to remarkable fractal structures such as Julia sets and the Mandelbrot set.

1 Algebra of the Complex Plane

'The Divine Spirit found a sublime outlet in that wonder of analysis, that portent of the ideal world, that amphibian between being and not-being, which we call the imaginary root of negative unity.' So said Leibniz in 1702 – though he may have let his eloquence run away with him. The current view of $\sqrt{-1}$ is a little more prosaic, though the uses made of it are at least as inspiring. The logical status of complex numbers, which caused so much distress during the eighteenth century, is now seen to be very much on a par with that of the 'real' numbers. What puzzled the ancients was the obvious artificiality and abstraction of the complex number system, in contrast to the apparently natural and concrete *real* number system. But the mathematician of today sees even real numbers as possessing a similar artificiality and abstraction.

In this chapter we discuss the construction of a system of numbers that contains the familiar real numbers and permits the solution of the equation $x^2 = -1$. This system is known as the *complex numbers*. Many readers will already know the contents of this chapter: they should read it through rapidly to check such items as notation, and pass on at once to the next.

There is a natural geometric representation of complex numbers as a plane, analogous to that of the reals as a line. The extra freedom inherent in the plane gives the whole subject a very geometric flavour, which it is our intention to keep to the fore in the development of the theory.

1.1 Construction of the Complex Numbers

We begin with the definition that emerged from the insights of Wallis, Wessel, Argand, Gauss, and Hamilton:

DEFINITION 1.1. A *complex number* is an ordered pair (x, y) of real numbers. Addition and multiplication of complex numbers are defined by:

$$(x_1, y_1) + (x_2, y_2) = (x_1 + x_2, y_1 + y_2) \tag{1.1}$$

$$(x_1, y_1)(x_2, y_2) = (x_1x_2 - y_1y_2, x_1y_2 + x_2y_1) \tag{1.2}$$

□

For example,

$$(3, 5)(2, 7) = (3 \cdot 2 - 5 \cdot 7, 3 \cdot 7 + 5 \cdot 2) = (-29, 31)$$

This definition is the culmination of several centuries of struggle to understand complex numbers, and it shows how elusive a simple idea can be. Before we see what these pairs have to do with $\sqrt{-1}$, however, let us establish some of their properties.

THEOREM 1.2. *The set of complex numbers, with the operations defined by* (1.1, 1.2), *is a field. That is, the following axioms hold: if* $z_1 = (x_1, y_1)$, $z_2 = (x_2, y_2)$, *and* $z_3 = (x_3, y_3)$ *are complex numbers, then*

(a) *Addition and multiplication are commutative:*

$$\begin{aligned} z_1 + z_2 &= z_2 + z_1 \\ z_1 z_2 &= z_2 z_1 \end{aligned} \tag{1.3}$$

(b) *Addition and multiplication are associative:*

$$\begin{aligned} (z_1 + z_2) + z_3 &= z_1 + (z_2 + z_3) \\ (z_1 z_2) z_3 &= z_1 (z_2 z_3) \end{aligned} \tag{1.4}$$

(c) *There is an additive identity* $(0,0)$*:*

$$z_1 + (0,0) = z_1 \tag{1.5}$$

(d) *There is a multiplicative identity* $(1,0)$*:*

$$z_1(1,0) = z_1 \tag{1.6}$$

(e) *Each element has an additive inverse:*

$$(x,y) + (-x,-y) = (0,0) \tag{1.7}$$

(f) *Each element other than* $(0,0)$ *has a multiplicative inverse:*

$$(x,y)\left(\frac{x}{x^2+y^2}, \frac{-y}{x^2+y^2}\right) = (1,0) \tag{1.8}$$

(g) *Multiplication distributes over addition:*

$$z_1(z_2 + z_3) = z_1 z_2 + z_1 z_3 \tag{1.9}$$

Proof. All assertions (a)–(g) are direct consequences of (1.1) and (1.2), using only the field properties of the set $\mathbb{R}$ of real numbers. For example, (1.9) holds because

$$\begin{aligned} z_1(z_2 + z_3) &= (x_1, y_1)(x_2 + x_3, y_2 + y_3) \\ &= (x_1(x_2 + x_3) - y_1(y_2 + y_3), x_1(y_2 + y_3) + y_1(x_2 + x_3)) \\ &= (x_1x_2 + x_1x_3 - y_1y_2 - y_1y_3, x_1y_2 + x_1y_3 + y_1x_2 + y_1x_3) \end{aligned}$$

and

$$\begin{aligned} z_1z_2 + z_1z_3 &= (x_1, y_1)(x_2, y_2) + (x_1, y_1)(x_3, y_3) \\ &= (x_1x_2 - y_1y_2, x_1y_2 + y_1x_2) + (x_1x_3 - y_1y_3, x_1y_3 + y_1x_3) \\ &= (x_1x_2 - y_1y_2 + x_1x_3 - y_1y_3, x_1y_2 + y_1x_2 + x_1y_3 + y_1x_3) \end{aligned}$$

which, by real algebra, is the same ordered pair.

The reader should supply similar proofs for the remaining assertions. □

The symbol $\mathbb{C}$ is used for the field of complex numbers.

1.2 The $x + iy$ Notation

The symbol commonly used for a complex number is not (x, y) but $x + \mathrm{i}y$. This notation goes back to Euler, who used i to denote $\sqrt{-1}$ in 1777, though the notation was first used consistently by Gauss.

To recover this notation, we proceed as follows. First note that since

$$\begin{aligned}(x_1, 0) + (x_2, 0) &= (x_1 + x_2, 0)\\ (x_1, 0)(x_2, 0) &= (x_1x_2, 0)\end{aligned}$$

we may identify a complex number $(x_1, 0)$ with the real number x_1. More pedantically the map $(x_1, 0) \mapsto x_1$ defines an isomorphism between the set of complex numbers of the form $(x_1, 0)$ and the field $\mathbb{R}$ of real numbers. Now define

$$\mathrm{i} = (0, 1)$$

Then

$$\begin{aligned}x + \mathrm{i}y &= (x, 0) + (0, 1)(y, 0)\\ &= (x, y) \quad \text{by (1.1) and (1.2)}\end{aligned}$$

Finally, observe that

$$\begin{aligned}\mathrm{i}^2 &= (0, 1)(0, 1)\\ &= (0 \cdot 0 - 1 \cdot 1, 0 \cdot 1 + 1 \cdot 0)\\ &= (-1, 0)\\ &= -1\end{aligned}$$

In this sense, we may say that $\mathrm{i} = \sqrt{-1}$.

The $x + \mathrm{i}y$ notation is more convenient, and will be used from now on. (Sometimes we use $x + y\mathrm{i}$ instead. By (1.3) this represents the same number. In particular, we use this form when x, y are specific real numbers, because $1 + 2\mathrm{i}$ looks more sensible than $1 + \mathrm{i}2$.)

Algebraic computations in this notation are easy. They use all the normal algebraic rules, *plus* the rule $\mathrm{i}^2 = -1$. So to multiply, we work out

$$\begin{aligned}(x_1 + \mathrm{i}y_1)(x_2 + \mathrm{i}y_2) &= x_1x_2 + x_1\mathrm{i}y_2 + \mathrm{i}y_1x_2 + \mathrm{i}y_1\mathrm{i}y_2\\ &= x_1x_2 + \mathrm{i}(x_1y_2 + y_1x_2) + \mathrm{i}^2y_1y_2\end{aligned}$$

But $\mathrm{i}^2 = -1$, so this becomes

$$x_1x_2 - y_1y_2 + \mathrm{i}(x_1y_2 + y_1x_2)$$

This computation, of course, explains the choice of the multiplication formula (1.2). The addition formula (1.1) comes the same way but is easier. The definition by (1.1) and (1.2) is thus a very sneaky piece of hindsight.

The formula (1.8) for inverses may also be derived as follows:

$$\frac{1}{x + \mathrm{i}y} = \frac{1}{x + \mathrm{i}y}\frac{x - \mathrm{i}y}{x - \mathrm{i}y} = \frac{x - \mathrm{i}y}{x^2 + y^2}$$

Example 1.3. Express $\frac{2+3\mathrm{i}}{1+2\mathrm{i}}$ in the form $x + \mathrm{i}y$.

We have

$$\frac{2+3\mathrm{i}}{1+2\mathrm{i}} = \frac{2+3\mathrm{i}}{1+2\mathrm{i}}\frac{1-2\mathrm{i}}{1-2\mathrm{i}} = \frac{2+6+\mathrm{i}(-4+3)}{5} = -\frac{8}{5} - \frac{\mathrm{i}}{5} \qquad \square$$

1.3 A Geometric Interpretation

Since ordered pairs (x, y) provide coordinates in the plane $\mathbb{R}^2$, we can visualise $\mathbb{C}$ as a plane, with the number $x + \mathrm{i}y$ corresponding to the point (x, y) as in Figure 1.1.

The identification of $(x, 0)$ with $x \in \mathbb{R}$ then amounts to considering the real numbers as forming the *real axis* in the plane, as in Figure 1.2.

The y-axis, at right angles to this, is the *imaginary axis*.

This geometric representation is often called the Argand diagram or the Gauss plane. Since so many other mathematicians (especially Wessel) have justifiable claims to it, we avoid the danger of giving undue credit to any of them by referring to it as the *complex plane*. In purely geometric terms, of course, it is just the real plane $\mathbb{R}^2$, but interpreted as $\mathbb{C}$ it has the additional algebraic structure of a field, not just a vector space over $\mathbb{R}$. It is this extra structure that gives the complex plane its special qualities.

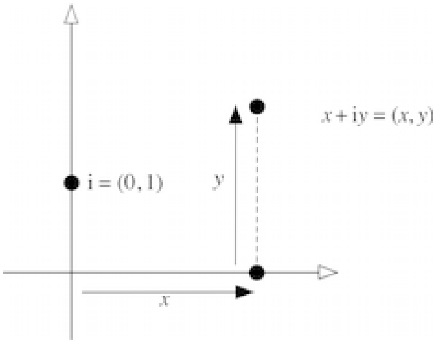

Figure 1.1 Visualising a complex number as a point in the plane $\mathbb{R}^2$.

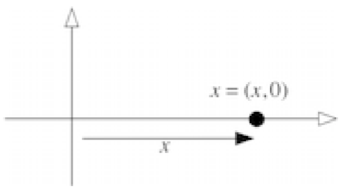

Figure 1.2 Identifying $x + 0\mathrm{i} \in \mathbb{C}$ with $x \in \mathbb{R}$.

1.4 Real and Imaginary Parts

Given a complex number $z = x + \mathrm{i}y$, we call x the *real part* of z and y the *imaginary part*, using the notation

$$x = \mathrm{re}(z)$$
$$y = \mathrm{im}(z)$$

Both are *real* numbers: the coordinates of z in the complex plane.

1.5 The Modulus

The modulus, or absolute value, of a real number x is defined to be

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

As it stands, there is no obvious generalisation to complex numbers, because (see Section 1.8 below) there is no useful ordering on $\mathbb{C}$. However, we can interpret $|x|$ geometrically as the distance from x to the origin of the real number line. This translates directly to the complex plane, leading to the definition

$$|z| = \sqrt{x^2 + y^2}$$

for the *modulus*, or *absolute value*, of a complex number $z = x + \mathrm{i}y$. Here we mean the positive square root: since $x^2 + y^2$ is always a positive real number, the formula defines $|z|$ as a real number.

THEOREM 1.4. *The modulus has the following properties:*

$$|z_1 + z_2| \leq |z_1| + |z_2| \tag{1.10}$$
$$|z_1 z_2| = |z_1||z_2| \tag{1.11}$$
$$||z_1| - |z_2|| \leq |z_1 - z_2| \tag{1.12}$$

Proof. Property (1.11) follows at once from the definitions. The *triangle inequality* (1.10) is a little harder to prove directly, although its geometric interpretation (Figure 1.3) is the obvious fact that one side of a triangle is no longer than the sum of the lengths of the other two sides.

To prove (1.10) note that, since both sides are positive, it is equivalent to

$$(|z_1 + z_2|)^2 \leq (|z_1| + |z_2|)^2$$

which takes the form

$$(x_1 + x_2)^2 + (y_1 + y_2)^2 \leq |z_1|^2 + 2|z_1||z_2| + |z_2|^2$$

where $z_1 = x_1 + \mathrm{i}y_1$, $z_2 = x_2 + \mathrm{i}y_2$. Simplifying, this holds if and only if

$$x_1 x_2 + y_1 y_2 \leq |z_1||z_2|$$

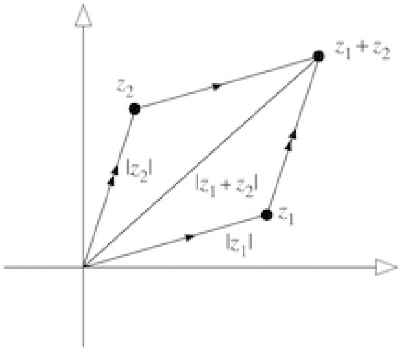

Figure 1.3 Geometry for the triangle inequality.

Since the right-hand side is positive, we may square again, and the desired inequality follows from

$$(x_1x_2 + y_1y_2)^2 \leq |z_1|^2|z_2|^2$$

But

$$\begin{aligned} |z_1|^2|z_2|^2 - (x_1x_2 + y_1y_2)^2 &= (x_1^2 + y_1^2)(x_2^2 + y_2^2) - (x_1x_2 + y_1y_2)^2 \\ &= (x_1y_2 - x_2y_1)^2 \end{aligned}$$

which is positive.

Property (1.12) is a consequence of (1.10). This implies that

$$|z_1 - z_2| + |z_2| \geq |z_1|$$

so

$$|z_1 - z_2| \geq |z_1| - |z_2|$$

Swapping z_1 and z_2 we also have

$$|z_1 - z_2| = |z_2 - z_1| \geq |z_2| - |z_1|$$

Combining the two inequalities yields

$$||z_1| - |z_2|| \leq |z_1 - z_2| \qquad \square$$

1.6 The Complex Conjugate

If $z = x + \mathrm{i}y$, its *complex conjugate* is

$$\bar{z} = x - \mathrm{i}y$$

Geometrically, this is obtained by reflecting z in the x-axis, Figure 1.4.

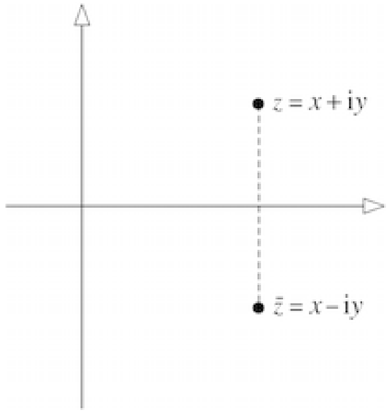

Figure 1.4 Geometry for the complex conjugate.

The following properties are easy to verify directly:

$$\overline{z_1 + z_2} = \bar{z_1} + \bar{z_2} \tag{1.13}$$
$$\overline{z_1 z_2} = \bar{z_1}\bar{z_2} \tag{1.14}$$
$$\text{re}(z) = \tfrac{1}{2}(z + \bar{z}) \tag{1.15}$$
$$\text{im}(z) = \tfrac{1}{2\text{i}}(z - \bar{z}) \tag{1.16}$$
$$|z|^2 = z\bar{z} \tag{1.17}$$
$$\bar{z} \in \mathbb{R} \text{ if and only if } z = \bar{z} \tag{1.18}$$

Properties (1.13, 1.14) have the important implication that the complex conjugate of any polynomial expression in complex numbers $z_1, z_2, \ldots, z_n$ can be obtained by writing a bar over each individual coefficient or variable in the expression. This is easily proved by induction. For example,

$$\begin{aligned}\overline{5z_1z_2 - z_3{}^7 + 2\text{i}z_1} &= \bar{5}\bar{z_1}\bar{z_2} - \bar{z_3}{}^7 + \bar{2}\bar{\text{i}}\bar{z_1} \\ &= 5\bar{z_1}\bar{z_2} - \bar{z_3}{}^7 - 2\text{i}\bar{z_1}\end{aligned}$$

since 5, 2 are real and i is imaginary, so $\bar{5} = 5, \bar{2} = 2, \bar{\text{i}} = -\text{i}$.

1.7 Polar Coordinates

The expression $x+\text{i}y$ for a complex number is intimately related to *Cartesian* coordinates (x, y) in the plane. It turns out often to be useful to work with *polar* coordinates (r, θ), which we recall correspond to a point distance r from the origin making an angle θ measured from the positive x-axis in an anticlockwise direction, Figure 1.5. Of course we measure θ in radians. These coordinate systems are related as follows:

$$\begin{aligned} x &= r\cos\theta \\ y &= r\sin\theta \end{aligned} \tag{1.19}$$

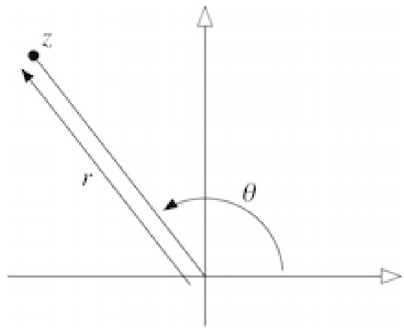

Figure 1.5 Polar coordinates.

Therefore

$$r = \sqrt{x^2 + y^2} = |z|$$

where $z = x + \mathrm{i}y$.

Finding θ is slightly trickier because it is not unique. Any value of θ for which (1.19) holds is called an *argument* of z. The article 'an' is used to reflect the lack of uniqueness: if θ is an argument then so is $\theta + 2k\pi$ for any integer k. With the understanding that θ is unique only up to multiples of 2π, we may use the notation

$$\theta = \arg z$$

Often the choice of θ is rendered unique by imposing some convention: for example, we may insist that θ is chosen in the interval $[0, 2\pi)$, or in $(-\pi, \pi]$. The unique value of θ in the interval $(-\pi, \pi]$ is known as the *principal value* of the argument. (We follow standard practice in taking this particular interval. Its main advantage is that θ then behaves nicely near the positive real axis, where $\theta = 0$. But this is a technical point that only acquires importance much later. The non-uniqueness of θ is a phenomenon with tremendous ramifications in the theory, as we shall see.)

With r, θ defined as above,

$$z = x + \mathrm{i}y = r(\cos\theta + \mathrm{i}\sin\theta)$$

The expression $\cos\theta + \mathrm{i}\sin\theta$ is of considerable importance in complex analysis. In Chapter 5 we relate it to the complex exponential function.

1.8 The Complex Numbers Cannot be Ordered

The real numbers may be given an ordering (the usual one, $>$) which has among its properties the following:

$$\text{If } x \neq 0 \text{ then either } x > 0 \text{ or } -x > 0, \text{ but not both} \tag{1.20}$$

$$\text{If } x, y > 0 \text{ then } x + y > 0, xy > 0 \tag{1.21}$$

No such ordering can be defined on the complex numbers. Suppose for a contradiction that one can. Since $\mathrm{i} \neq 0$, (1.20) implies that either $\mathrm{i} > 0$ or $-\mathrm{i} > 0$. Then (1.21) implies that either $-1 = \mathrm{i} \cdot \mathrm{i} > 0$ or $-1 = (-\mathrm{i}) \cdot (-\mathrm{i}) > 0$. At the same time, $1 = (-1)^2 > 0$. But then both 1 and -1 are greater than 0, contrary to (1.20).

It is therefore not possible to use inequalities, analogous to those for reals, when discussing complex numbers. Any inequality that occurs must involve only *real* numbers, possibly related to the given complex numbers. For example, if $z \in \mathbb{C}$ then

$$z > 1$$

makes no sense, but either of

$$|z| > 1$$

or

$$\mathrm{re}(z) > 1$$

is acceptable. (They do not mean the same thing!) As a convention, if we write a statement such as

$$\varepsilon > 0$$

this will automatically imply that ε is assumed to be a real number.

1.9 Exercises

1. Check in full detail that the complex numbers $\mathbb{C}$ form a field under the operations of addition and multiplication defined in (1.1, 1.2).
2. In Figure 1.6 the black dots represent three complex numbers u, v, w (as marked). The circle is the unit circle $|z| = 1$. The open dots a, b, c, d, e, f, g, h represent (in some order) the numbers $u + v, u + w, v + w, u + v + w, uv, uw, vw, uvw$. Which is which?

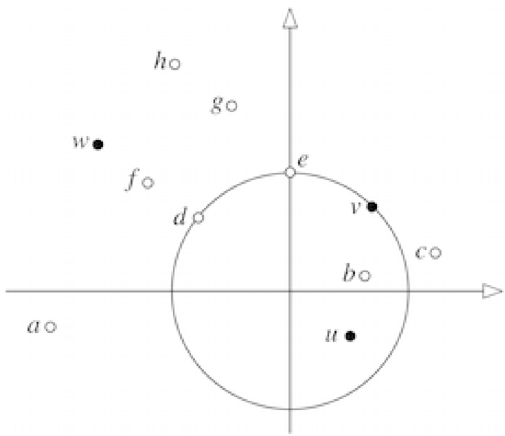

Figure 1.6 Data for Exercise 2.

3. By writing z in the form $z = a + b\mathrm{i}$, find all solutions z of the following equations:

(i) $z^2 = -5 + 12\mathrm{i}$
(ii) $z^2 = 2 + \mathrm{i}$
(iii) $(7 + 24\mathrm{i})z = 375$
(iv) $z^2 - (3 + \mathrm{i})z + (2 + 2\mathrm{i}) = 0$
(v) $z^2 - 3z + 1 + \mathrm{i} = 0$

4. If λ is a positive real number, show that

$$\{z \in \mathbb{C} : |z| = \lambda|z - 1|\}$$

is a circle, *unless* λ takes one particular value (which?)

5. Draw the set of points

$$\{z \in \mathbb{C} : \mathrm{re}(z + 1) = |z - 1|\}$$

by substituting $z = x + \mathrm{i}y$ and computing the real equation relating x and y.

Now note that $\mathrm{re}(z+1)$ is the distance from z to the line $y = -1$, and $|z-1|$ is the distance between z and 1. Compare with the classical 'focus-directrix' definition of a parabola: the locus of a point equidistant from a fixed line (here $y = -1$) and a fixed point (here $(x, y) = (1, 0)$).

6. Draw the set of all $z \in \mathbb{C}$ satisfying the following conditions:

(i) $\mathrm{re}(z) > 2$
(ii) $1 < \mathrm{im}(z) < 2$
(iii) $1 < \mathrm{im}(z - \mathrm{i}) < 2$
(iv) $|z| < 2$
(v) $|z| > 1$
(vi) $1 < |z| < 2$
(vii) $|z - 1| < 1$
(viii) $|z - 1| < |z + 1|$

7. Draw the set of all $z \in \mathbb{C}$ satisfying the following conditions:

(i) $z\bar{z} = 1$
(ii) $z + \mathrm{i}\bar{z} + 1 + \mathrm{i} = 0$
(iii) $z + \bar{z} + 2 = 0$
(iv) $z + \bar{z} + 2\mathrm{i} = 0$

8. Let r, s, θ, ϕ be real. Let

$$z = r(\cos\theta + \mathrm{i}\sin\theta)$$
$$w = s(\cos\phi + \mathrm{i}\sin\phi)$$

Form the product zw and use the standard formulas for $\cos(\theta + \phi), \sin(\theta + \phi)$ to show that $\arg(zw) = \arg(z) + \arg(w)$ (for any values of arg on the right, and *some* value of arg on the left).

By induction on n, derive *De Moivre's Theorem*

$$(\cos\theta + \mathrm{i}\sin\theta)^n = \cos n\theta + \mathrm{i}\sin n\theta$$

for all natural numbers n.

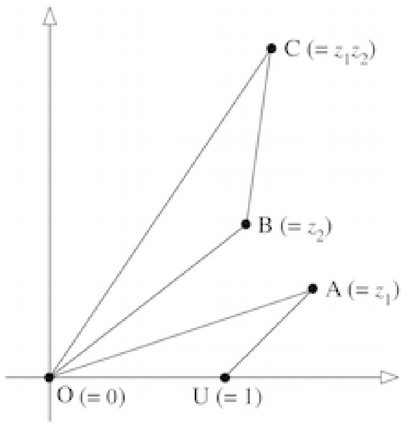

Figure 1.7 Data for Exercise 10.

Specialise to the case $n = 3$ and recover the usual formulas for $\cos 3\theta$ and $\sin 3\theta$ in terms of $\cos\theta$ and $\sin\theta$.

9. Use De Moivre's Theorem (Exercise 8) and the substitution $z = r(\cos\theta + \mathrm{i}\sin\theta)$ to show that the equation $z^3 = 1$ has three distinct complex roots. Find them.

Compute the square roots of $1 + \mathrm{i}\sqrt{3}$, $\sqrt{3} - \mathrm{i}$, and $1 + \mathrm{i}$, and the cube roots of $\sqrt{3} + \mathrm{i}$, $1 - \mathrm{i}$, i. Sketch these points in the complex plane.

10. In earlier textbooks, multiplication of complex numbers is often defined as follows. Given two complex numbers z_1, z_2, represent them by points A and B in the complex plane; and let O, U be the points $z = 0, 1$ respectively, Figure 1.7.

Draw triangle OBC similar to triangle OUA (where ∠BOC = ∠UOA, ∠OBC = ∠OUA). Then z_1z_2 is represented by the point C so constructed.

Using the fact that $|z_1z_2| = |z_1||z_2|$, and the result of Exercise 5, show that this construction agrees with our definition (1.2).

11. Define a square root $\sqrt{z}$ of a complex number z to be any complex number w such that $w^2 = z$. Prove that every non-zero complex number has exactly two square roots, and give formulas for them in terms of $\mathrm{re}\, z$ and $\mathrm{im}\, z$.

If $a, b, c \in \mathbb{C}$ with $a \neq 0$, show that the solutions of the quadratic equation

$$az^2 + bz + c = 0$$

are precisely

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

12. Use De Moivre's Theorem (Exercise 8) to compute $\cos 5\theta$ and $\sin 5\theta$ in terms of $\cos\theta$ and $\sin\theta$.

13. Prove that De Moivre's Theorem remains true if n is a negative integer.

14. Define a kth root $\sqrt[k]{z}$ to be any w such that $w^k = z$. Use De Moivre's Theorem to find an expression for $\sqrt[k]{r(\cos\theta + \mathrm{i}\sin\theta)}$.

2 Topology of the Complex Plane

In this chapter we collect together the basic topological ideas required for our study of complex analysis. The list is not very demanding. Some items are needed to handle differentiation neatly, and some are needed for integration. Differentiation is naturally set against a background of limits and continuity, and these are best dealt with on open sets. On the other hand, an interval from one complex number to another is computed along a specified path between them. A set in which any two points can be joined by a path is said to be path-connected. To be able to cope with both integration and differentiation in the simplest possible manner later on, we restrict our complex functions to be those defined on open path-connected sets. Such a set is called a domain.

When the set is open we often abbreviate 'path-connected' to 'connected'. The term 'connected' is used in point-set topology with a specific technical meaning, but for open sets in $\mathbb{C}$ it is equivalent to being path-connected. So this abbreviation does no harm.

Domains can have exotic shapes and paths can wiggle around a great deal. To be able to appeal to geometric intuition without our imagination having to work overtime thinking about such complications, we use a carefully conceived technical device called the *Paving Lemma*. We show in this lemma that a path in an open set (in particular, a domain) can be subdivided into a finite number of smaller pieces in such a way that each piece is contained in a disc within the open set, thus 'paving' the path with discs, see Figure 2.1. A disc is the interior of a circle, which is geometrically very simple: for instance, any two points in it can be joined by a straight line. Joining the end points of pieces of the original path in each disc paving it, we obtain a new path made up of straight line segments, still lying in the open set and joining the ends of the original path. We see, therefore, that given any path whatsoever between two points in an open set, no matter how much the path twists and turns, there is an alternative path in the open set, between the same points, that is made up of a finite number of straight line segments. We can even insist that the segments are parallel to the real or imaginary axis, giving a *step path* in the open set. To do so, take a suitable step path inside each paving disc and join them together, see Figure 2.2.

With techniques such as this we can use the Paving Lemma to illuminate complex analysis, yielding fully rigorous proofs linked firmly to geometric intuition.

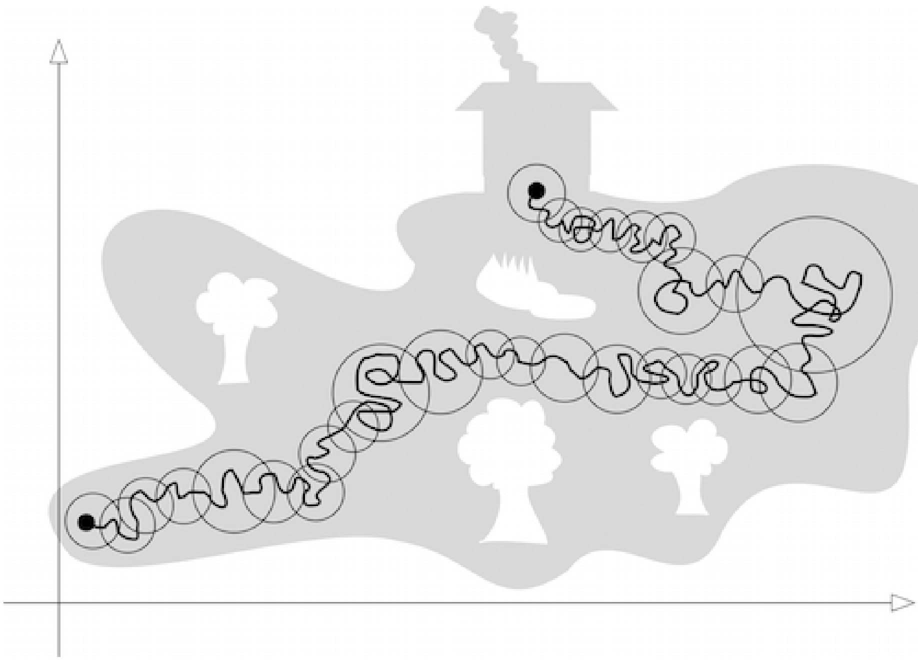

Figure 2.1 A domain (shaded) and a path in the domain (solid curve). Circles show a finite set of discs inside the domain, whose union contains the path.

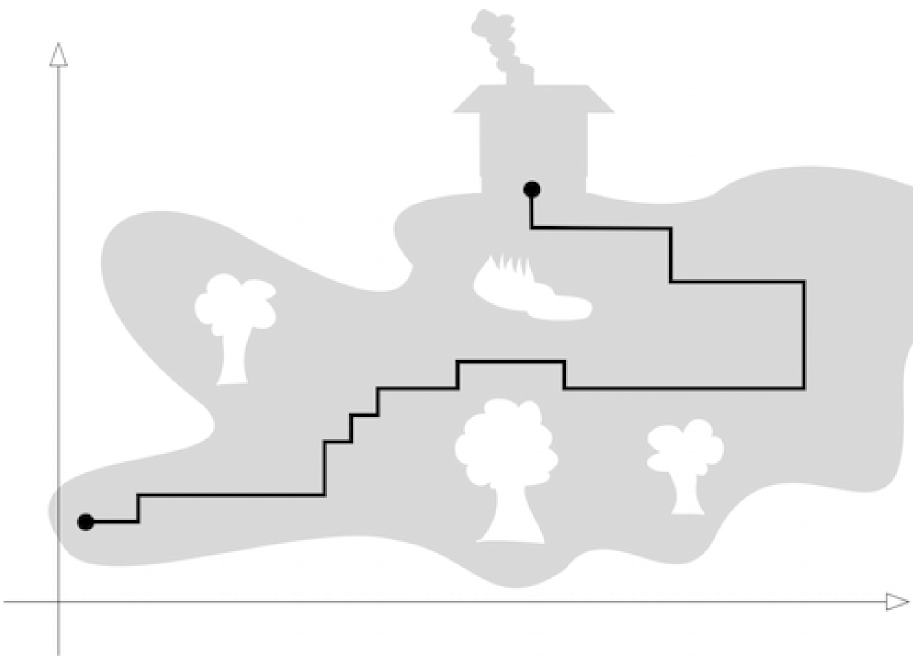

Figure 2.2 Replacing the path in Figure 2.1 by a step path.

2.1 Open and Closed Sets

DEFINITION 2.1. For a complex number z_0 and a positive real number ε, the *ε-neighbourhood* of z_0 is

$$N_\varepsilon(z_0) = \{z \in \mathbb{C} : |z - z_0| < \varepsilon\}$$

This is an open disc of radius ε.

Geometrically, $N_\varepsilon(z_0)$ is the disc centre z_0 of radius ε, Figure 2.3.

A subset $S \subseteq \mathbb{C}$ is said to be *open* if for every $z_0 \in S$ there is a real number ε such that $N_\varepsilon(z_0) \subseteq S$. We emphasise that ε may depend on z_0. □

Example 2.2. The disc $N_\varepsilon(z_0)$ is itself open, for if $z_1 \in N_\varepsilon(z_0)$ then $|z_1 - z_0| < \varepsilon$. Choose $\delta > 0$ such that $\delta < \varepsilon - |z_1 - z_0|$. By the triangle inequality, $N_\delta(z_1) \subseteq N_\varepsilon(z_0)$, Figure 2.4. □

DEFINITION 2.3. The *complement* of a subset $S \subseteq \mathbb{C}$ is

$$\mathbb{C} \setminus S = \{z \in \mathbb{C} : z \notin S\}$$

A subset S is *closed* if $\mathbb{C} \setminus S$ is open. □

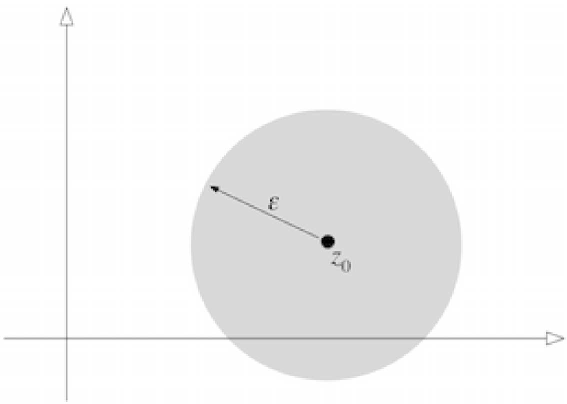

Figure 2.3 The *ε-neighbourhood* of a point z_0.

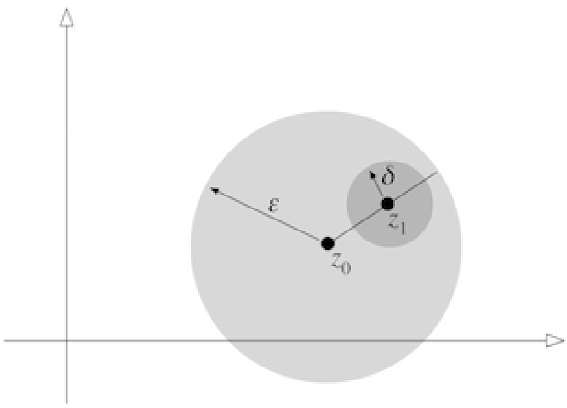

Figure 2.4 The *ε-neighbourhood* of z_0 is open.

There is another way to characterise closed sets, using the notion of a *limit point* of a subset S. A complex number z_0 is a limit point of S if every neighbourhood $N_\varepsilon(z_0)$ contains a point of S not equal to z_0. In this definition, z_0 does not itself have to belong to S, though it may do. The essential feature of a limit point of S is that it has points of S arbitrarily close to it. In fact, each $N_\varepsilon(z_0)$ must contain an *infinite* number of points of S – for if some $N_\varepsilon(z_0)$ contains only finitely many points $z_1, \ldots, z_n$ of S, distinct from z_0, we can take ε_1 to be the smallest of the distances $|z_0 - z_r|$. Then $N_{\varepsilon_1}(z_0)$ contains no points of S, a contradiction.

An alternative characterisation of a closed set is:

PROPOSITION 2.4. *A subset $S \subseteq \mathbb{C}$ is closed if and only if S contains all its limit points.*

Proof. Suppose S is closed and z_0 is a limit point. If $z_0 \in \mathbb{C} \setminus S$, which is open, then $N_\varepsilon(z_0) \subseteq \mathbb{C} \setminus S$ for some $\varepsilon > 0$, so $N_\varepsilon(z_0)$ contains no point of S, contradicting z_0 being a limit point. Hence $z_0 \in S$.

Conversely, suppose that S contains all its limit points. Then any $z_0 \in \mathbb{C} \setminus S$ is not a limit point of S, so there exists $\varepsilon > 0$ such that $N_\varepsilon(z_0)$ contains no point of S. Therefore $N_\varepsilon(z_0) \subseteq \mathbb{C} \setminus S$, so $\mathbb{C} \setminus S$ is open, and S is closed. □

Not every point of a closed set need be a limit point of that set. For instance, if

$$T = \{z \in \mathbb{C} : z = 0 \text{ or } z = 1/n \text{ for a positive integer } n\}$$

then the only limit point of T is 0. Since this is in T, the set T is closed.

Points in S that are not limit points of S are said to be *isolated points* of S. All points of T except 0 are isolated points of T.

In general, an isolated point z_0 of S has a neighbourhood $N_\varepsilon(z_0)$ that contains no points of S other than z_0. For instance, in the set T, if $\varepsilon = \frac{1}{n} - \frac{1}{n+1}$ then $N_\varepsilon(1/n)$ contains no other points of T. On the other hand, it is clear from the definitions that *every* element of an *open* set is a limit point of that set.

We will also need:

DEFINITION 2.5. The *closure* of a subset $S \subseteq \mathbb{C}$ is the intersection of all closed subsets that contain S. This is closed, and is the smallest closed subset containing S. It consists of S together with all limit points of S. □

2.2 Limits of Functions

The notion of a limit

$$\lim_{z \to z_0} f(z)$$

is analogous to the real case, and its properties follow by similar arguments.

DEFINITION 2.6. If $f : S \to \mathbb{C}$ is an arbitrary complex function and z_0 is a limit point of S, then $\lim_{z \to z_0} f(z) = l$ if, given any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\text{for all } z \in S, 0 < |z - z_0| < \delta \text{ implies } |f(z) - l| < \varepsilon \qquad (2.1)$$

□

Two points should be made about this definition:

(a) The point z_0 need not belong to S, so $f(z_0)$ need not be defined. Even if $z_0 \in S$, $f(z_0)$ may not equal l.

For example, if

$$f(z) = \begin{cases} 0 & (z \neq 0) \\ 1 & (z = 0) \end{cases}$$

then $\lim_{z \to 0} f(z) = 0 \neq f(0)$

(b) It is essential that z_0 is a limit point of S, or else there would exist $\delta > 0$ such that $\{z : |z - z_0| < \delta\}$ contains no point of S. In this case condition (2.1) would be vacuously true for *any* $l \in \mathbb{C}$.

As in the real case, we have:

PROPOSITION 2.7. *If z_0 is a limit point of S and $\lim_{z \to z_0} f(z) = l$, the limit is unique.*

Proof. Suppose that $l' \neq l$ is another candidate for the limit. Take $\varepsilon = \frac{1}{2}|l - l'|$ to find $\delta_1 > 0, \delta_2 > 0$ such that

$$z \in S, 0 < |z - z_0| \text{ implies } |f(z) - l| < \varepsilon$$
$$z \in S, 0 < |z - z_0| \text{ implies } |f(z) - l'| < \varepsilon$$

Because z_0 is a limit point, there exists $z^* \in S$, where $0 < |z_0 - z^*| < \min(\delta_1, \delta_2)$. Then

$$\begin{aligned} |l - l'| &= |l - f(z^*) + f(z^*) - l'| \\ &\leq |l - f(z^*)| + |f(z^*) - l'| \\ &< \varepsilon + \varepsilon \\ &= 2\varepsilon \end{aligned}$$

contradicting the choice of ε. □

Standard properties of complex limits may be proved using methods analogous to the real case:

PROPOSITION 2.8. *If $\lim_{z \to z_0} f(z) = l$, $\lim_{z \to z_0} g(z) = k$, then*

(i) $\lim_{z \to z_0} f(z) + g(z) = l + k$
(ii) $\lim_{z \to z_0} f(z) - g(z) = l - k$
(iii) $\lim_{z \to z_0} f(z)g(z) = lk$
(iv) $\lim_{z \to z_0} f(z)/g(z) = l/k$ *(for $l \neq 0$)*

Proof. Part (i) is routine. Part (ii) follows immediately if we first show that $\lim_{z \to z_0}(-g(z)) = -k$, which is trivial.

Part (iii) is a little trickier. Bear in mind that $f(z_0), g(z_0)$ may not be defined since z_0 need not belong to S. Write

$$|f(z)g(z) - lk| = |f(z)g(z) - lg(z) + lg(z) - lk| \tag{2.2}$$
$$\leq |f(z) - l||g(z)| + |l||g(z) - k| \tag{2.3}$$

Since $\lim_{z\to z_0} g(z) = k$, there exists $\delta_0 > 0$ such that

$$z \in S \text{ and } 0 < |z - z_0| < \delta \quad \text{implies } |g(z) - k| < \varepsilon$$

Therefore $|g(z)| < |k| + \varepsilon = M$, say, whenever $0 < |z - z_0|$. Therefore $|g(z)|$ is bounded above near (but not necessarily at) z_0 by M.

Since $\lim_{z\to z_0} f(z) = l$, there exists $\delta_1 > 0$ such that

$$z \in S \text{ and } 0 < |z - z_0| < \delta_1 \quad \text{implies } |f(z) - l| < \varepsilon/(2M)$$

Let $N \in \mathbb{R}$ such that $N > |l|$, so in particular $N > 0$. Since $\lim_{z\to z_0} g(z) = k$, there exists $\delta_2 > 0$ such that

$$z \in S \text{ and } 0 < |z - z_0| < \delta_2 \quad \text{implies } |g(z) - k| < \varepsilon/(2N)$$

For $\delta = \min(\delta_0, \delta_1, \delta_2)$ we then have

$$\begin{aligned} &z \in S \text{ and } 0 < |z - z_0| < \delta_1 \quad \text{implies} \\ &\quad |f(z) - l||g(z)| + |l||g(z) - k| < \frac{\varepsilon}{2M}M + N\frac{\varepsilon}{2N} = \varepsilon \end{aligned}$$

By (2.3), for $z \in S$ and $0 < |z - z_0| < \delta_2$, we have

$$|f(z)g(z) - lk| < \varepsilon$$

proving (iii).

To prove (iv), it is enough to prove that $\lim_{z\to z_0} 1/g(z) = 1/k$, and then to appeal to (iii) for the functions $f, 1/g$.

For this to make sense we must show that for some $\delta_1 > 0$, $g(z)$ is not zero when $z \in S$ and $0 < |z - z_0| < \delta_1$. Then $1/g(z)$ is defined.

This follows since $\lim_{z\to z_0} g(z) = k$ and $k \neq 0$. Indeed, there exists δ_1 such that

$$z \in S \text{ and } 0 < |z - z_0| < \delta_1 \quad \text{implies } |g(z) - k| < \tfrac{1}{2}|k|$$

which in turn tells us that

$$z \in S \text{ and } 0 < |z - z_0| < \delta_1 \quad \text{implies } |g(z)| > \tfrac{1}{2}|k|$$

We now find δ_2 such that

$$z \in S \text{ and } 0 < |z - z_0| < \delta_2 \quad \text{implies } |g(z) - k| < \tfrac{1}{2}|k|^2\varepsilon$$

Now, if $\delta = \min(\delta_1, \delta_2)$,

$$\begin{aligned} &z \in S \text{ and } 0 < |z - z_0| < \delta \quad \text{implies} \\ &\quad \left|\frac{1}{g(z)} - \frac{1}{k}\right| < \frac{|k - g(z)|}{|g(z)k|} < \tfrac{1}{2}|k|^2\varepsilon/(\tfrac{1}{2}|k|^2) = \varepsilon \end{aligned}$$ □

The real and imaginary parts of a complex function,

$$f(z) = \operatorname{re} f(z) + \mathrm{i} \operatorname{im} f(z)$$

may be considered separately. If

$$\lim_{z\to z_0} f(z) = l = \alpha + \mathrm{i}\beta \quad (\alpha, \beta \in \mathbb{R}) \tag{2.4}$$

then, because

$$|\mathrm{re} f(z) - \alpha| = |\mathrm{re}(f(z) - l)| \leq |(f(z) - l)|$$

we deduce from the definition that

$$\lim_{z \to z_0} \mathrm{re} f(z) = \alpha \tag{2.5}$$

Proposition 2.8 (ii) now implies that

$$\lim_{z \to z_0} \mathrm{im} f(z) = \beta \tag{2.6}$$

Conversely, if (2.5,2.6) both hold, Proposition 2.8 (i) implies (2.5).

We can rephrase the preceding argument by recalling that if $S \subseteq \mathbb{R}^2$ the limit of a real function $\phi : S \to \mathbb{R}$ of two variables, $\phi(x, y)$ for $(x, y) \in S$ is defined as follows:

$$\phi(x, y) \to \lambda \text{ as } (x, y) \to (a, b) \text{ if, given } \varepsilon > 0, \text{there exists } \delta > 0 \text{ such that}$$
$$\text{for all } (x, y) \in S, 0 < \sqrt{(x-a)^2 + (y-b)^2} < \delta \text{ implies } |\phi(x, y) - \lambda| < \varepsilon$$

Identifying (x, y) with $z = x + \mathrm{i}y \in \mathbb{C}$, and setting $z_0 = a + \mathrm{i}b$, this may be written as

$$z \in S \text{ and } 0 < |z - z_0| < \delta \quad \text{implies } |\phi(x, y) - \lambda| < \varepsilon$$

If we now write

$$f(z) = u(x, y) + \mathrm{i}v(x, y)$$

where the real and imaginary parts of f are considered as real functions u and v of two real variables x, y, we have proved:

PROPOSITION 2.9. *Let* $z_0 = a + \mathrm{i}b$. *Then* $\lim_{z \to z_0} f(z) = l = \alpha + \mathrm{i}\beta$ *if and only if*

$$u(x, y) \to \alpha \quad v(x, y) \to \beta \quad \text{as } (x, y) \to (a, b)$$ □

In this way, the limit of a complex functions is equivalent to the limit of a pair of real functions of two real variables. However, the notation in the complex case is generally simpler.

2.3 Continuity

The definition of continuity for a complex function mimics that for a real function, using the complex version of the modulus:

DEFINITION 2.10. A function $f : S \to \mathbb{C}$ is *continuous at* $z_0 \in S$ if, given $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\text{for all } z \in S, |z - z_0| < \delta \text{ implies } |f(z) - f(z_0)| < \varepsilon$$

A function is *continuous* if it is continuous at every point $z_0 \in S$. □

If z_0 is a limit point of S, this is equivalent to saying that $\lim_{z \to z_0} f(z)$ exists and

$$\lim_{z \to z_0} f(z) = f(z_0)$$

If z_0 is an isolated point of S then there is a neighbourhood $N_\delta(z_0)$ that contains no other points of S apart from z_0, so

$$\text{for all } z \in S, |z - z_0| < \delta \text{ implies } z = z_0$$

which in turn implies

$$|f(z) - f(z_0)| = 0$$

So a complex function is always continuous at an isolated point, according to the definition. Actually, this will not be an issue for us, because we will consider only functions where all points of S are limit points. In fact, S is usually open. But it seemed worth tidying up a potential loose end.

We can rephrase the definition of continuity in terms of open discs: f is continuous at $z_0 \in S$ if, given $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\text{for all } z \in S, z \in N_\delta(z_0) \text{ implies } f(z) \in N_\varepsilon(f(z_0))$$

Or, more succinctly,

$$f(N_\delta(z_0)) \subseteq N_\varepsilon(f(z_0))$$

We can develop an alternative way to define continuous functions in terms of open sets. First we need a generalisation: a subset $V \subseteq S$ is said to be *relatively open in S*, or just *open in S* for short, if for every $z_0 \in V$ there exists $\sigma > 0$ such that $N_\sigma(z_0) \cap S \subseteq V$.

Example 2.11. A relatively open set need not be open. The interval $(a,b) = \{x \in \mathbb{R} : a < x < b\}$ is open in $\mathbb{R}$, but not in $\mathbb{C}$. In fact, when $S = \mathbb{C}$ let $x \in (a,b)$ and suppose that $\sigma > 0$. Then $x + i\sigma/2 \in N_\sigma(x) = N_\sigma(x) \cap S$, but $x + i\sigma/2 \notin V$. See Figure 2.5. □

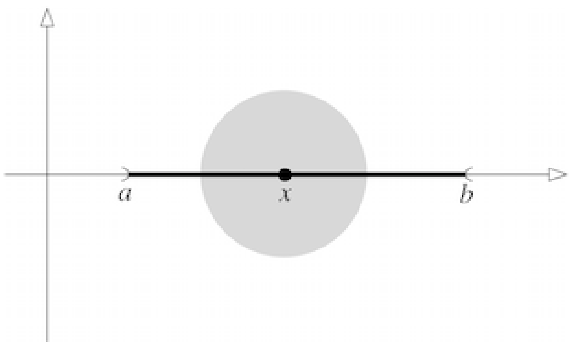

Figure 2.5 The interval (a, b) is open in $\mathbb{R}$ but not in $\mathbb{C}$.

Using the standard set-theoretic notation

$$f^{-1}(U) = \{z \in S : f(z) \in U\}$$

we get an alternative characterisation of a continuous function:

PROPOSITION 2.12. *A complex function $f : S \to \mathbb{C}$ is continuous if and only if, for every open set U in $\mathbb{C}$, the set $f^{-1}(U)$ is open in S.*

Proof. Suppose that f is continuous and U is open. Let $z_0 \in f^{-1}(U)$. Then $f(z_0) \in U$ so there exists $\varepsilon > 0$ such that $N_\varepsilon(f(z_0)) \subseteq U$. By continuity of f there exists $\delta > 0$ such that

$$f(N_\delta(z_0) \cap S) \subseteq N_\varepsilon(f(z_0)) \subseteq U$$

Hence

$$N_\delta(z_0) \cap S \subseteq f^{-1}(U)$$

and $f^{-1}(U)$ is open.

Conversely, suppose that $f^{-1}(U)$ is open in S for every open set U. Given $z_0 \in S$ and $\varepsilon > 0$, the set $N_\varepsilon(f(z_0))$ is open, so $f^{-1}(N_\varepsilon(f(z_0)))$ is open in S and there exists $\delta > 0$ such that

$$N_\delta(z_0) \cap S \subseteq f^{-1}(N_\varepsilon(f(z_0)))$$

Hence

$$f(N_\delta(z_0) \cap S) \subseteq N_\varepsilon(f(z_0))$$

so f is continuous. □

Proposition 2.12 is favoured by topologists as the *definition* of continuity. It is particularly useful when S is itself an open set. Then, given a subset $V \subseteq S$ that is open in S, and any $z_0 \in V$, there exist $\sigma, \varepsilon > 0$ such that

$$N_\sigma(z_0) \subseteq V \quad \text{(because } V \text{ is open in } S\text{)}$$
$$N_\varepsilon(z_0) \subseteq S \quad \text{(because } S \text{ is open)}$$

Taking $\delta = \min(\varepsilon, \sigma)$, we deduce that

$$N_\delta(z_0) \subseteq V$$

so V is an open set in $\mathbb{C}$. This proves:

COROLLARY 2.13. *If $S \subseteq \mathbb{C}$ is open, then $f : S \to \mathbb{C}$ is continuous if and only if, for every open set U, the inverse image $f^{-1}(U)$ is open.* □

Figure 2.6 illustrates this result.

Composition of two functions is defined in the usual manner: if $f : S \to \mathbb{C}$ and $g : T \to \mathbb{C}$ where $f(S) \subseteq T$, then $g \circ f : S \to \mathbb{C}$ is given by

$$g \circ f(z) = g(f(z))$$

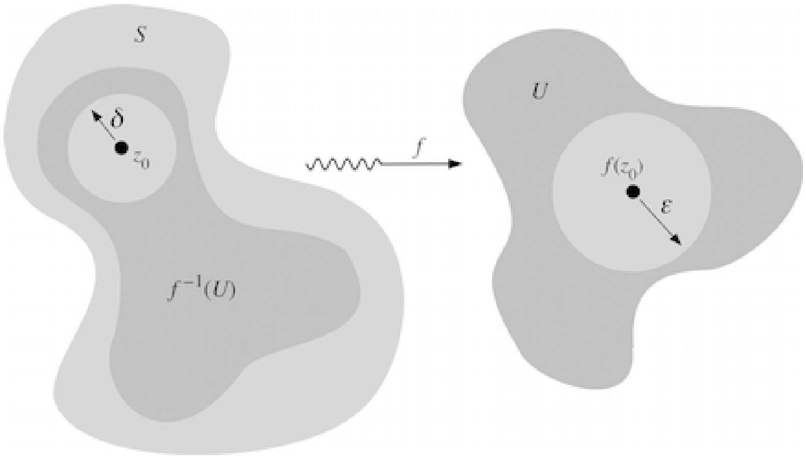

Figure 2.6 Definition of continuity when S is open in $\mathbb{C}$.

PROPOSITION 2.14. *If f is continuous at $z_0 \in S$ and g is continuous at $f(z_0)$, then $g \circ f$ is continuous at z_0.*

Proof. Easy exercise. □

Addition, subtraction, multiplication, and division of complex functions $f_1 : S \to \mathbb{C}$ and $f_2 : S \to \mathbb{C}$ are defined by

$$f_1 + f_2 : S \to \mathbb{C} \text{ where } (f_1 + f_2)(z) = f_1(z) + f_2(z) \quad (z \in S)$$
$$f_1 - f_2 : S \to \mathbb{C} \text{ where } (f_1 - f_2)(z) = f_1(z) - f_2(z) \quad (z \in S)$$
$$f_1 f_2 : S \to \mathbb{C} \text{ where } (f_1 f_2)(z) = f_1(z) f_2(z) \quad (z \in S)$$
$$f_1/f_2 : S \to \mathbb{C} \text{ where } (f_1/f_2)(z) = f_1(z)/f_2(z) \quad (z \in S')$$

Here $S' = \{z \in S : f_2(z) \neq 0\}$.

Warning: the composition $g \circ f$ is often abbreviated to gf when it is clear that the product is not intended.

PROPOSITION 2.15. *If f_1 and f_2 are continuous, then so are $f_1 + f_2, f_1 - f_2, f_1 f_2$, and f_1/f_2 (when $f_2(z) \neq 0$).*

Proof. This is a direct consequence of Proposition 2.8. □

This result lets us show very quickly that certain functions built up from functions known to be continuous are themselves continuous. For instance, the constant function $k(z) = c$ (where $c \in \mathbb{C}$) and the identity function $I(z) = z$ are clearly continuous. We give the easy proofs. Let $\varepsilon > 0$. For k take any $\delta > 0$; then

$$|z - z_0| < \delta \text{ implies } |k(z) - k(z_0)| = |c - c| = 0 < \varepsilon$$

For I take $\delta = \varepsilon$; then

$$|z - z_0| < \varepsilon \text{ implies } |I(z) - I(z_0)| = |z - z_0| < \varepsilon$$

On that basis, Proposition 2.15 shows immediately that $f(z) = cz$ is continuous. By induction on n, it follows that $f(z) = cz^n$ is continuous for any $n \in \mathbb{N}$. A further induction proves that any polynomial function

$$f(z) = a_n z^n + a_{n-1}z^{n-1} + \cdots + a_0$$

is continuous. Then any rational function $p(z)/q(z)$, where p, q are polynomials, is continuous at any point z_0 where $q(z_0) \neq 0$. Using these ideas we seldom have to perform any intricate epsilon-delta calculations.

Example 2.16. The modulus function $m(z) = |z|$ is continuous for all z. Given $\varepsilon > 0$, take $\delta = \varepsilon$. Then

$$|m(z) - m(z_0)| < \varepsilon \text{ implies } ||z| - |z_0|| \leq |z - z_0| < \varepsilon$$

using (1.12). □

Example 2.17. $f(z) = \frac{|z|^2+17z^3+1066z}{1+z}$ is continuous for $z \neq -1$. Proving this with an explicit choice of δ for given ε is tedious and messy. A quicker way is to observe that $|z|^2 = |z||z|$ and is therefore continuous, $17z^3 + 1066z$ is a polynomial and so is also continuous, and the same goes for $z + 1$. Since we are dividing by this, we must have $z \neq -1$. □

Writing $f(z) = u(x, y)+\mathrm{i}v(x, y)$ for $z = x+\mathrm{i}y \in S$, where u, v are real-valued functions of the real variables x, y, Proposition 2.9 gives:

PROPOSITION 2.18. *The function $f(z) = u(x, y)+\mathrm{i}v(x, y)$ is continuous at $z_0 = x_0+\mathrm{i}y_0$ if and only if u, v are continuous at (x_0, y_0).* □

Example 2.19. If $f(z) = z^2$, then $f(z) = (x+\mathrm{i}y)^2 = x^2-y^2+2\mathrm{i}xy$. Hence $u(x, y) = x^2-y^2$ and $v(x, y) = 2xy$. The function f is continuous for all $z \in \mathbb{C}$, just as u, v are continuous functions of x, y for all $(x, y) \in \mathbb{R}^2$. □

An interesting case occurs when S is the real interval

$$S = [a, b] = \{x \in \mathbb{R} : a \leq x \leq b\}$$

considered as a subset of $\mathbb{C}$. Here $z = x + 0\mathrm{i}$, so we can simplify notation and write $f(z) = u(x) + \mathrm{i}v(x)$. Now the function $f : [a, b] \to \mathbb{C}$ is continuous if and only if both u and v are continuous.

Example 2.20. $f : [0, 1] \to \mathbb{C}$, where $f(x) = x^2 + \mathrm{i}x^3$ is continuous, since both $u(x) = x^2$ and $v(x) = x^3$ are continuous. □

Functions defined on a real interval play a central role because they define paths, as we now discuss.

2.4 Paths

In the early development of complex analysis, mathematicians worked out how to integrate complex functions along paths in $\mathbb{C}$, without worrying too much about the meaning of 'path'. This happened back in the days when notions of continuity and limits were still under development. The intuitive idea of a path (or curve) was something that could be drawn by moving a pencil (with an infinitely fine point) without making any jumps. Such paths were assumed to have various nice properties, such as having a continuously varying tangent, or not crossing themselves. These properties were not made explicit until the foundational issues in analysis were sorted out, in particular by Bolzano and Weierstrass.

Later it turned out that some of these assumptions are not necessary to develop a satisfactory theory of integration, while others must be made precise to avoid running into contradictions. We have no wish to make readers repeat all the twists and turns of history; as we have said, some battles have already been won. But from time to time we find it useful to point out some of the pitfalls that can arise, and to emphasise some subtle distinctions – for example, the definition of a path as a map γ from a real interval $[a,b]$ into the complex plane, and the distinction between this map and its image. Our pictures generally show the image, leaving the reader to infer the map. We often add an arrow as a reminder that as a parameter t runs through the interval $[a,b]$, the point $\gamma(t)$ moves along the image in a specific direction.

One of the pleasant features of complex analysis is that fairly often that information is all we need. At various points in the book we also find it useful to impose further conditions on paths, to ensure that they have appropriate properties.

The upshot of more than a century of deep contemplation and debate is the following definition:

DEFINITION 2.21. A *path* in the complex plane is a continuous function

$$\gamma : [a,b] \to \mathbb{C} \quad (a \leq b \in \mathbb{R})$$

Its *initial point* is $\gamma(a)$ and its *final point* (or *end point*) is $\gamma(b)$. □

Sometimes we speak of γ as 'a path in $\mathbb{C}$ from z_1 to z_2' when $z_1 = \gamma(a)$ is the initial point and $z_2 = \gamma(b)$ is the final point. When $t \in [a,b]$ we also refer to $z = \gamma(t)$ as a 'point on the path γ', although strictly speaking z is on the *image* of the function γ. If we think of t as 'time' and imagine t increasing from a to b, the point $\gamma(t)$ traces a curve in the plane from $\gamma(a)$ to $\gamma(b)$. When drawing a diagram we often indicate the direction of motion (increasing t) by an arrow, as in Figure 2.7. However, it must be emphasised that this is a makeshift device since the curve may cross itself or turn back on itself, so the picture may get quite complicated. We pick up this point again in Section 6.7.

2.4.1 Standard Paths

To simplify matters, when we speak of line segments and circles, considered as paths, we assume they are specified by the following standard functions:

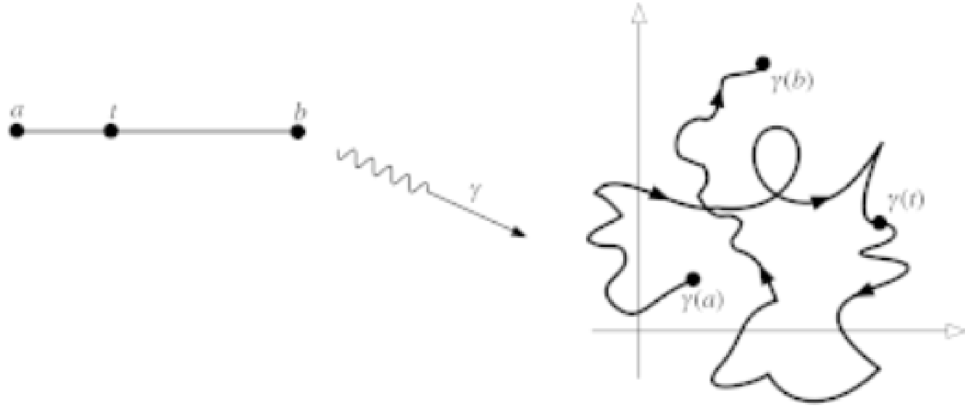

Figure 2.7 A path in $\mathbb{C}$, showing its parametrisation by $t \in [a, b]$.

(i) The *line segment* $L = [z_1, z_2]$ from z_1 to z_2 is

$$L(t) = (1 - t)z_1 + tz_2 \quad (t \in [0, 1])$$

When $z_1 = a, z_2 = b \in \mathbb{R}$ then the image of the path is the closed interval $[a, b] \subseteq \mathbb{R} \subseteq \mathbb{C}$.

(ii) The *unit circle* C:

$$C(t) = \cos t + \mathrm{i} \sin t \quad (t \in [0, 2\pi])$$

(iii) The circle S with centre z_0 and radius $r \geq 0$:

$$S(t) = z_0 + r(\cos t + \mathrm{i} \sin t) \quad (t \in [0, 2\pi])$$

See Figure 2.8.

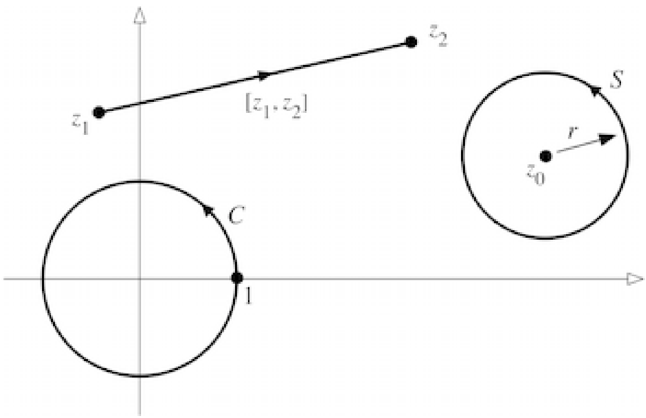

Figure 2.8 Three standard paths.

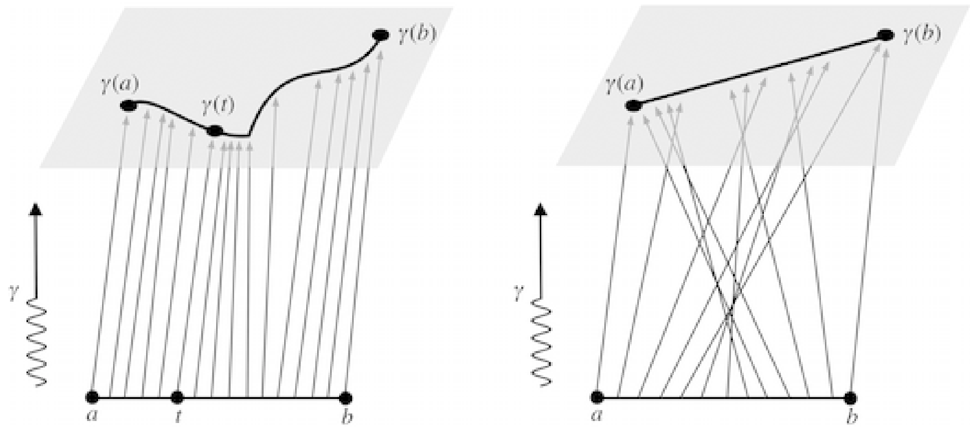

Figure 2.9 ‘Graphs’ of paths considered as functions $[a, b] \to \mathbb{C}$. *Left*: A path that travels once along the image in a fixed direction. *Right*: A path that travels once along the image, turns round and travels back, then turns again and travels along the image for a third time.

2.4.2 Visualising Paths

We are used to visualising a function $f : \mathbb{R} \to \mathbb{R}$ using the corresponding graph $y = f(x)$ in the plane. A similar idea can be exploited to visualise paths in a more explicit way. Figure 2.9 shows analogues of graphs for two maps $\gamma : [a,b] \to \mathbb{C}$. The arrows join representative points $t \in [a,b]$ to their images $\gamma(t) \in \mathbb{C}$. The left-hand figure illustrates a path where $\gamma(t)$ travels along the image without turning back on itself. The right-hand figure illustrates a path where $\gamma(t)$ travels along the image to the far end, turns back on itself to revisit the start, then turns back again to travel along the image for a third time. Although paths like this are seldom encountered in practice, *they are allowed by the definition*, so we have to take that possibility into account. Example 6.17 below shows that simple formulas can produce this type of behaviour.

An alternative would be to change the definition to rule out such behaviour, but most of the theory works in the general case with no extra effort, so it turns out to be simpler to leave the definition as it stands, while bearing in mind that simple pictures might be misleading in some respects.

2.4.3 The Image of a Path

To emphasise the distinction between a path and its image we introduce a further definition:

DEFINITION 2.22. A *curve* C from z_1 to z_2 in the complex plane is a subset of $\mathbb{C}$ that is equal to the image of a path $\gamma : [a, b] \to \mathbb{C}$, where $\gamma(a) = z_1$ and $\gamma(b) = z_2$.

Given a curve C, a *parametrisation* of C is a path $\sigma : [c, d] \to \mathbb{C}$ such that C is the image of σ, where $\sigma(a) = z_1$ and $\sigma(b) = z_2$.

The corresponding *parameter* is a variable $t \in [a, b]$.

We sometimes call $[a, b]$ the *parametric interval* of σ. □

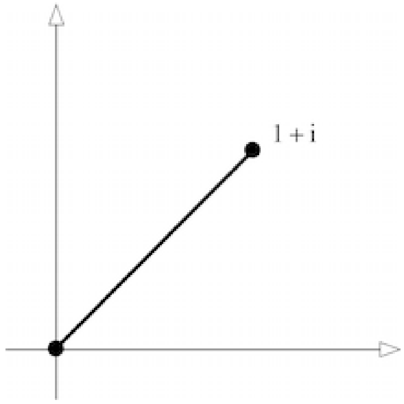

Figure 2.10 The two distinct paths γ_1, γ_2 have the same image, so they define the same curve.

If we think of t as 'time', and imagine a particular parametrisation $\gamma : [a,b] \to \mathbb{C}$, then as t increases from a to b the point $\gamma(t)$ traces the corresponding curve in the plane from $\gamma(a)$ to $\gamma(b)$, moving continuously with t.

Example 2.23. The same curve may be traced by many different paths. For example, in Figure 2.10 the paths

$$\gamma_1(t) = 2(t + \mathrm{i}t) \quad (0 \leq t \leq \tfrac{1}{2})$$
$$\gamma_2(t) = t^2 + \mathrm{i}t^2 \quad (0 \leq t \leq 1)$$

traverse the same curve

$$\{x + \mathrm{i}y \in \mathbb{C} : x = y, 0 \leq x \leq 1\} \qquad \square$$

2.5 Change of Parameter

Definition 2.21 equips every path $\gamma : [a,b] \to \mathbb{C}$ with a parameter $t \in [a,b]$. It is often convenient to change the parameter and its parametric interval, because suitable choices can simplify calculations and proofs. In this section we discuss such changes.

We begin with Example 2.23, which distinguishes a path from its image by exhibiting two different paths that have the same image curve. Dynamically, these two paths trace the same curve in the same direction, but at different speeds. They are related by the change of parameter $\rho : [0, \frac{1}{2}] \to [0,1]$, where $\rho(t) = \frac{1}{2}t^2$. This has an inverse map $\rho^{-1} : [0,1] \to [0, \frac{1}{2}]$ given by $\rho^{-1}(s) = \sqrt{2s}$.

We now discuss general changes of parameter, defined as follows:

DEFINITION 2.24. Let $\gamma : [a,b] \to \mathbb{C}$ be a path, and suppose that $\rho : [c,d] \to [a,b]$ is continuous and satisfies

$$\rho(c) = a \qquad \rho(d) = b$$

so in particular ρ is onto $[a, b]$. Then the composition $\gamma \circ \rho : [c, d] \to \mathbb{C}$ is also a path. We call $\gamma \circ \rho$ a *reparametrisation* of γ.

Note that the parametric interval changes when $[c, d] \neq [a, b]$. □

2.5.1 Preserving Direction

If we impose extra conditions on ρ, a change of parameter can preserve extra properties. The image of the path is important, but another property is also vital: the direction in which the path is traced. Now our mental image involves the dynamics of a point moving along the path, as well as the image that it traces out.

At this stage we are working with general continuous paths, so we adopt the following approach. Recall that a map $\sigma : [a, b] \to [c, d]$ is *strictly increasing* if $a \le t_1 < t_2 \le b$ implies $\sigma(t_1) < \sigma(t_2)$. Both of the above maps ρ and ρ^{-1} are strictly increasing.

In real analysis it is proved that any strictly increasing continuous map has a strictly increasing continuous inverse. We give the easy proof for completeness:

LEMMA 2.25. *If $\rho : [a, b] \to [c, d]$ is continuous and strictly increasing with $\rho(a) = c$ and $\rho(b) = d$, then ρ has a strictly increasing continuous inverse $\rho^{-1} : [c, d] \to [a, b]$.*

Proof. If $s \in [c, d]$ then $\rho(a) \le s \le \rho(b)$. By the Mean Value Theorem there exists $t \in [a, b]$ such that $\rho(t) = s$. This determines a strictly increasing inverse function $\rho^{-1} : [c, d] \to [a, b]$, where ρ and ρ^{-1} both map open intervals to open intervals, so ρ^{-1} is also continuous. □

We can now state:

DEFINITION 2.26. Let C be a curve, parametrised by two maps $\gamma : [a, b] \to \mathbb{C}$ and $\lambda : [c, d] \to \mathbb{C}$. The maps have the *same direction* if there is a strictly increasing function $\rho : [a, b] \to [c, d]$ such that

$$\gamma = \lambda \circ \rho$$

Then ρ is a *direction preserving* change of parameter. □

If the change in parameter ρ is not increasing, it can lead to part of the curve being traced back and forth as t increases. In Figure 2.11, as t increases from p to q, the value of $\rho(t)$ increases, then decreases, then increases again. As this happens, the points on the curve from P to Q are traversed first in the direction from P to Q, then back to P, then from P to Q once more. The curve is the same, but the distance travelled increases. This means that when we calculate the length of a curve, defined in Section 6.3, we must prescribe how the curve is traced.

2.6 Subpaths and Sums of Paths

In the general theory, and in applications, it is useful to be able to chop paths into pieces, or join pieces together.

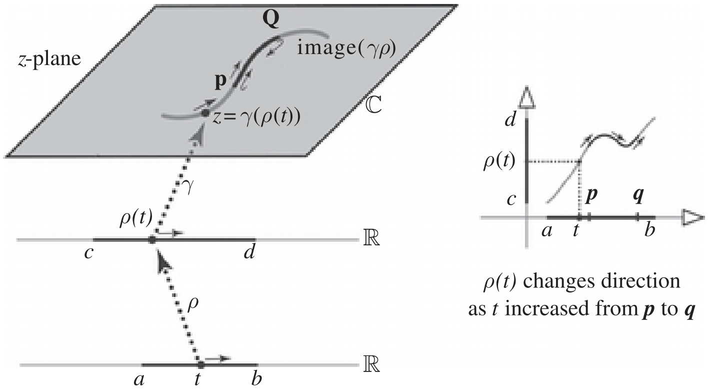

Figure 2.11 *Left*: A change in parameter. *Right*: If ρ has a graph like this then the parameter t changes direction along the image curve.

DEFINITION 2.27. For an arbitrary path $\gamma : [a,b] \to \mathbb{C}$, a *subpath* is obtained by restricting γ to a subinterval $[c,d]$, where $a \leq c \leq d \leq b$. If

$$a = x_0 < x_1 < \cdots < x_n = b$$

and γ_r is the subpath obtained by restricting γ to $[x_{r-1}, x_r]$, we write

$$\gamma = \gamma_1 + \cdots + \gamma_n$$

In so doing, we think of γ as being made up by tracing along the paths $\gamma_1, \ldots, \gamma_n$ taken in order, see Figure 2.12. □

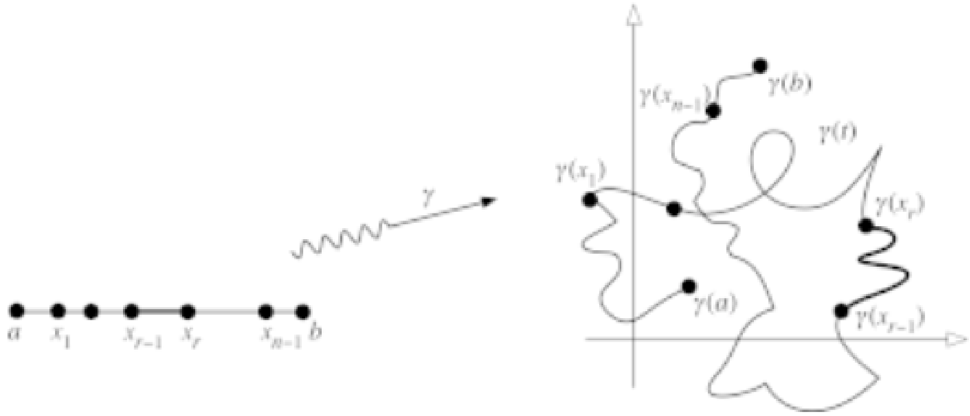

Figure 2.12 Decomposing a path into subpaths.

Similarly, if the final point of a path γ_1 coincides with the initial point of a path γ_2, it is useful on occasion to create a combined path by first tracing γ_1 and then γ_2. There is a minor technicality here, because $\gamma_1(b)$ can be the same as $\gamma_2(c)$ even though γ_1 is defined on $[a,b]$ and γ_2 defined on $[c,d]$ where $b \neq c$. The next example shows how to get round this difficulty.

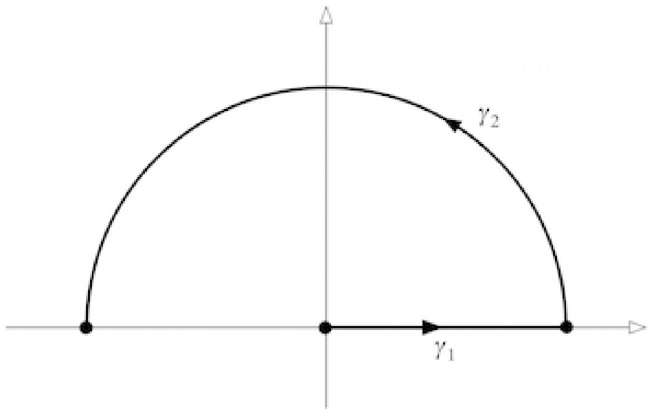

Figure 2.13 Joining two paths by shifting parameters.

Example 2.28. Suppose that

$$\gamma_1(t) = t \quad (t \in [0, 1])$$
$$\gamma_2(t) = \cos t + \mathrm{i} \sin t \quad (t \in [0, \pi])$$

as in Figure 2.13. Here $\gamma_1(1) = \gamma_2(0)$ but $1 \neq 0$. □

In such a case we extend our notation a little. Suppose that $\gamma_1 : [a, b] \to \mathbb{C}$, and $\gamma_2 : [c, d] \to \mathbb{C}$, with $\gamma_1(b) = \gamma_2(c)$. Then we define the combination $\gamma = \gamma_1 + \gamma_2 : [a, b + d - c] \to \mathbb{C}$ by

$$\gamma(t) = \begin{cases} \gamma_1(t) & (t \in [a, b]) \\ \gamma_2(t + c - b) & (t \in [b, b + d - c]) \end{cases}$$

In effect, what we have done is to shift the parametric interval of the second path from $[c, d]$ to $[b, b + d - c]$, by adding $b - c$ to all points in $[c, d]$.

In Example 2.28, for instance,

$$\gamma(t) = \begin{cases} t & (t \in [0, 1]) \\ \cos(t - 1) + \mathrm{i} \sin(t - 1) & (t \in [1, 1 + \pi]) \end{cases}$$

and γ consists of the line segment γ_1 followed by the semicircle γ_2.

The path $\gamma_1 + \gamma_2$ is defined only when the final point of γ_1 is the same as the initial point of γ_2, so perhaps this is not a fully appropriate use of the + sign. However, we do have

$$(\gamma_1 + \gamma_2) + \gamma_3 = \gamma_1 + (\gamma_2 + \gamma_3) \tag{2.7}$$

whenever the appropriate end points coincide, so we can omit the parentheses in such a 'sum', and more generally write $\gamma_1 + \cdots + \gamma_n$ to indicate the path obtained by successively tracing $\gamma_1, \ldots, \gamma_n$, whenever the final point of each γ_{r-1} coincides with the initial point of $\gamma_r (1 \leq r \leq n)$.

Equation (2.7) states that addition of paths is associative, but in general it is not commutative. Exercises 7 and 8 provide simple examples to show that when $\gamma_1 + \gamma_2$ is defined, $\gamma_2 + \gamma_1$ may not be, and even when it is, $\gamma_1 + \gamma_2$ need not equal $\gamma_2 + \gamma_1$.

For a path $\gamma : [a,b] \to \mathbb{C}$, another useful concept is the *opposite* path $-\gamma : [a,b] \to \mathbb{C}$ defined by

$$-\gamma(t) = \gamma(a+b-t) \quad (t \in [a,b])$$

As t increases from a to b, so $-\gamma$ describes the same curve as γ, but it does so in the reverse direction. If γ is a path from z_1 to z_2, then $-\gamma$ is a path from z_2 to z_1.

Example 2.29. Suppose that L is the line segment $[z_1, z_2]$ given by

$$L(t) = (1-t)z_1 + tz_2 \quad (t \in [0,1])$$

Then $-L$ is given by

$$-L(t) = tz_1 + (1-t)z_2 \quad (t \in [0,1])$$

which is, of course, $[z_2, z_1]$.

If S is the circle centre z_0 and radius $r \geq 0$, then

$$S(t) = r(\cos t + \mathrm{i}\sin t) \quad (t \in [0,2\pi])$$

describes the circle once, anticlockwise, while the opposite path

$$\begin{aligned} -S(t) &= z_0 + r(\cos(2\pi - t) + \mathrm{i}\sin(2\pi - t)) \quad (t \in [0,2\pi]) \\ &= z_0 + r(\cos t - \mathrm{i}\sin t) \quad (t \in [0,2\pi]) \end{aligned}$$

describes it once, clockwise. □

In a sum such as $\gamma_1 + (-\gamma_2) + \gamma_3$ we omit the parentheses and write $\gamma_1 - \gamma_2 + \gamma_3$. This is the path that traces first along γ_1, then back along the opposite path to γ_2, and then along γ_3. As always, the appropriate final and initial points must agree. See Figure 2.14.

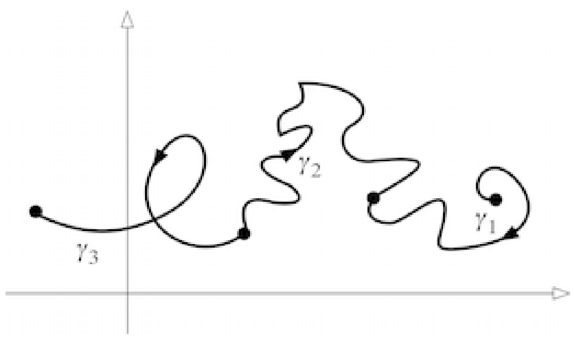

Figure 2.14 The path $\gamma_1 - \gamma_2 + \gamma_3$.

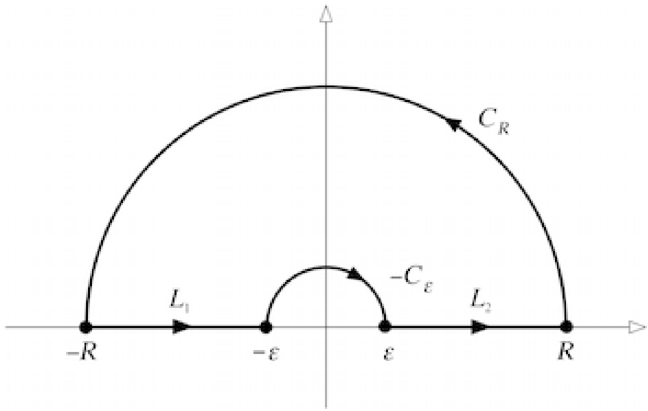

Figure 2.15 The path $C_R + L_1 - C_\varepsilon + L_2$.

Example 2.30. Suppose that $0 < \varepsilon < R$ and

$$\begin{aligned} C_R(t) &= R(\cos t + \mathrm{i}\sin t) \quad (t \in [0, \pi]) \\ C_\varepsilon(t) &= \varepsilon(\cos t + \mathrm{i}\sin t) \quad (t \in [0, \pi]) \\ L_1(t) &= t \quad (t \in [-R, -\varepsilon]) \\ L_2(t) &= t \quad (t \in [\varepsilon, R]) \end{aligned}$$

Then $C_R + L_1 - C_\varepsilon + L_2$ is the path describing the curve in Figure 2.15 once, anticlockwise. □

2.7 The Paving Lemma

We now come to a technical lemma that plays a vital role in our approach to complex integration, which we call the Paving Lemma. This result is so important that we present two different proofs of it. To set it up, we first need another piece of terminology.

A path $\gamma : [a, b] \to \mathbb{C}$ is said to be a *path in S* if the image

$$\{\gamma(t) : t \in [a, b]\} \subseteq S$$

We signify this by writing $\gamma : [a, b] \to S$.

Example 2.31. The unit circle $C(t) = \cos t + \mathrm{i}\sin t$ $(t \in [0, 2\pi])$ is a path in the region $S = \{z \in \mathbb{C} : \frac{1}{2} \le |z| \le 2\}$, Figure 2.16.

A region of this form, contained between two concentric circles, is called an *annulus*. □

This example is very simple. It must not lull us into a false sense of security: far more intricate paths are possible.

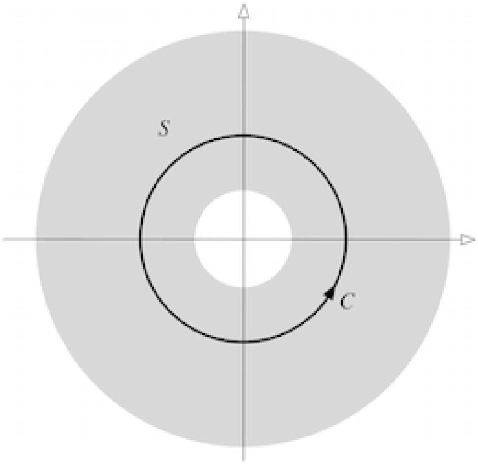

Figure 2.16 Path in an annulus.

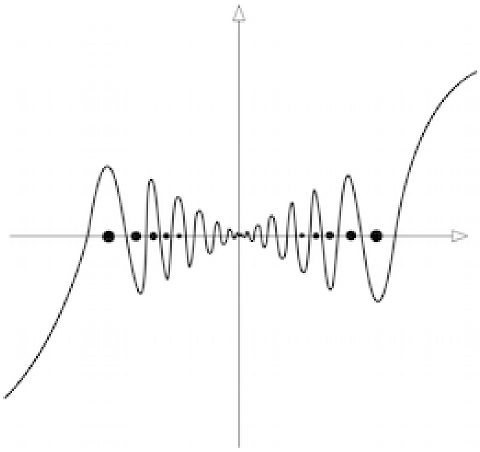

Figure 2.17 A more complicated path.

Example 2.32. Let $V = \{z \in \mathbb{C} : z \neq 1/n \text{ for a non-zero integer } n\}$ and define

$$\sigma(t) = \begin{cases} t + \mathrm{i}t\cos(\pi/t) & (t \in [-1,1] \setminus \{0\}) \\ 0 & (t = 0) \end{cases}$$

Then σ is a path in V, Figure 2.17. Here σ winds in between the points $1/n$ where n is a non-zero integer, crossing the real axis when $t = 1/(n + \frac{1}{2})$. □

It does not take much imagination to realise that an open set S can be very complicated, and a path γ can be intertwined in S in a very intricate manner. The space-filling curve described in Section 2.9 is given by a path whose image is the unit square, including its interior. Much more complicated images are possible. Rather than trying to imagine all possible intricacies, we sidestep the issue, as follows. If S is an open set,

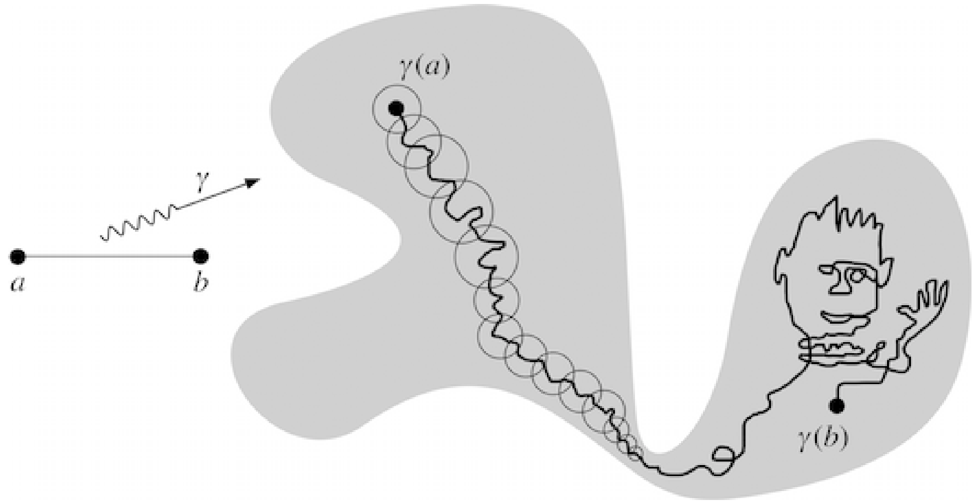

Figure 2.18 Paving a path with a sequence of discs. But could this process get stuck?

we can cover γ with a finite number of open discs, all contained in S. Then we use the discs to simplify the path, as the next result explains.

LEMMA 2.33 (Paving Lemma). *Let $\gamma : [a, b] \to S$ be a path in an open set S. Then there exists a subdivision $a = t_0 < t_1 < \cdots < t_n = b$ such that each subpath γ_r obtained by restricting γ to $[t_{r-1}, t_r]$ lies inside an open disc $D_r \subseteq S$.*

Proof. Because S is open, there is an open disc D_1 such that $\gamma(a) \in D_1 \subseteq S$.

The idea of the proof is that if the path has been paved up to some point, we can always add another disc to take the paving further, unless we have reached the far end. Making this argument rigorous takes a little care, however, as we now indicate.

The disc D_1 is the first in a sequence of discs, obtained by moving along the path, covering it bit by bit with discs as in Figure 2.18. The existence of D_1 shows we can get started. Indeed, since γ is continuous, there exists $\delta > 0$ such that

$$\gamma(t) \in D_1 \text{ whenever } a \leq t < a + \delta \tag{2.8}$$

Our only fear is that we might not progress all the way; for instance, the discs might decrease dramatically in size and we might never reach the end. Some real analysis dispels this fear. If $a \leq x \leq b$, say that γ restricted to $[a, x]$ 'can be paved' if it can be subdivided into a *finite* number of subpaths $\gamma_1, \ldots, \gamma_m$, where each γ_r lies inside an open disc $D_r \subseteq S$. Let P be the set of all $x \in [a, b]$ such that $[a, x]$ can be paved. We know that $a \in P$ so P is non-empty.

Also P is bounded above by b. Therefore, by real analysis, P has a least upper bound c, where $a \leq c \leq b$. We claim that $c = b$, which will prove the lemma.

For a contradiction, suppose that $c < b$. Since S is open, there exists an open disc D such that $c \in D \subseteq S$. By continuity of γ, there exists $\varepsilon > 0$ such that if $c - \varepsilon < t < c + \varepsilon$ and $t \in [a, b]$, the image $\gamma(t) \in D$.

By (2.8), $a + \delta/2 \in P$, so $c > a$. Since c is the least upper bound of P, there exists $x \in P$ with $c - \varepsilon < x < c$. Therefore the segment $\{\gamma(t) : a \leq t \leq x\}$ can be paved by open discs $D_1, \ldots, D_m$ all lying inside S. Let $D_{m+1} = D$. Then the segment $\{\gamma(t) : a \le t \le c\}$ can be paved by open discs $D_1, \ldots, D_m, D_{m+1}$ all lying inside S.

Finally, we claim that $c = b$. If $c < b$, continuity of γ implies that $\gamma(c + t) \in D$ for all t such that $0 < t < \tau$, therefore $[a, c + \tau/2]$ can be paved, contradicting c being an upper bound for P. Therefore $c = b$ so $[a, b]$ can be paved as required. □

For future reference, and because this result is used repeatedly, we now give a second proof, which in some respects is simpler. A generalisation to two dimensions will also prove useful in Proposition 9.5.

Alternative Proof of Lemma 2.33. We show more: the subdivision can be obtained by repeatedly bisecting $[a, b]$ a finite number of times. For the purposes of this proof *only*, it is useful to define and interval to be *pavable* if it can be bisected repeatedly a finite number of times into closed intervals, so that the image of every interval obtained in this manner lies inside a disc in D. Otherwise, the interval is *unpavable*.

Suppose the result is false. Then $[a, b]$ is unpavable. Bisect $[a, b]$ to get two closed intervals, intersecting at the midpoint. At least one these intervals, call it I_1, is unpavable. Otherwise we could pave a finitely repeated bisection of each interval by finitely many discs, whose union paves $[a, b]$.

Bisect I_1. The same argument applies to the two halves, so we get an unpavable closed interval $I_2 \subseteq I_1$ of half the size. Because we are assuming the result false, this sequence of bisections continues indefinitely to give a nested sequence of unpavable closed intervals

$$[a, b] \supseteq I_1 \supseteq I_2 \supseteq \cdots$$

each half the size of the previous one.

The intersection of these intervals is easily seen to be a unique point $z_0 \in D$ (Cauchy sequences of real numbers converge). However, D is open, so there is a disc $N_\varepsilon(\gamma(z_0)) \subseteq D$. Since γ is continuous, $\gamma^{-1}(N_\varepsilon(\gamma(z_0)))$ is open in $[a, b]$ and contains z_0. Therefore it contains I_n for sufficiently large n. Now $I_n \subseteq N_\varepsilon(\gamma(z_0))$, so it is pavable (with one disc), a contradiction. □

In Example 2.31, for any point on the unit circle C, we have $N_{1/2}(z) \subseteq S$. Clearly a finite number of such discs will pave C within S, Figure 2.19.

However, the path in Example 2.32 cannot be paved by a finite number of discs in V. Any disc containing the origin (which lies on σ includes points of the form $1/n$ lying outside V. But this also means that V is *not open*. So the conditions required for the Paving Lemma are not satisfied.

Readers familiar with basic point-set topology will recognise that the proof is 'really' about the compactness of the closed interval $[a, b]$. We prefer not to develop the abstract machinery of compactness here; the Paving Lemma performs the same task.

2.8 Connectedness

We now define an important topological property, which is central to the theory of complex functions.

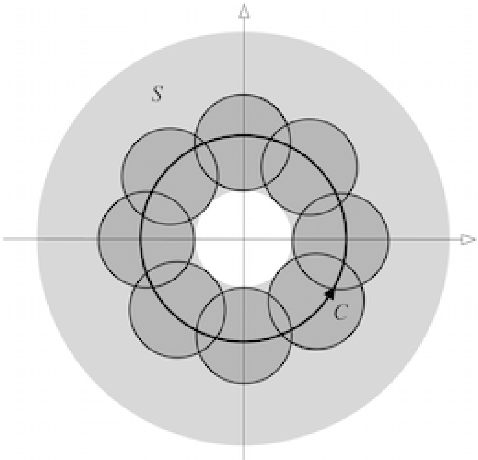

Figure 2.19 Paving a circular path inside an annulus is easy.

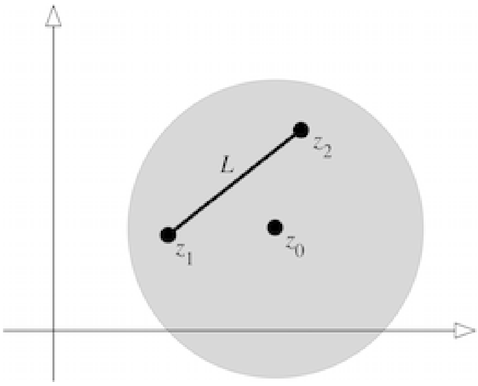

Figure 2.20 A disc is path-connected.

DEFINITION 2.34. A subset $S \subseteq \mathbb{C}$ is *path-connected* if, given $z_1, z_2 \in S$, there exists a path γ in S from z_1 to z_2. □

Example 2.35. Any open disc $N_r(z_0)$ is path-connected. For, given $z_1, z_2 \in N_r(z_0)$, let $L(t) = (1-t)z_1 + tz_2, (t \in [0, 1])$. Then

$$\begin{aligned} |L(t) - z_0)| &= |(1-t)z_1 + tz_2 - z_0| \\ &= |(1-t)z_1 - (1-t)z_0 + tz_2 - tz_0| \\ &\leq (1-t)|z_1 - z_0| + t|z_2 - z_0| \\ &< (1-t)r + tr = r \end{aligned}$$

so the line segment L lies in $N_r(z_0)$. See Figure 2.20. □

Example 2.36. The set $S = \{z \in \mathbb{C} : z = t + \mathrm{i}t^2, t \in \mathbb{R}\}$ is path-connected. For if $t_1 + \mathrm{i}t_1^2$ and $t_2 + \mathrm{i}t_2^2 \in S$ where $t_1 \leq t_2$, then $\gamma(t) = t + \mathrm{i}t^2$ $(t \in [t_1, t_2])$ is a path in S between those points. See Figure 2.21. □

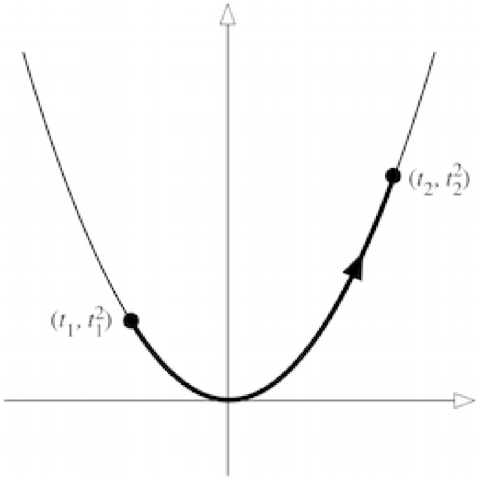

Figure 2.21 The set $\{z \in \mathbb{C} : z = t + \mathrm{i}t^2, t \in \mathbb{R}\}$ is path-connected.

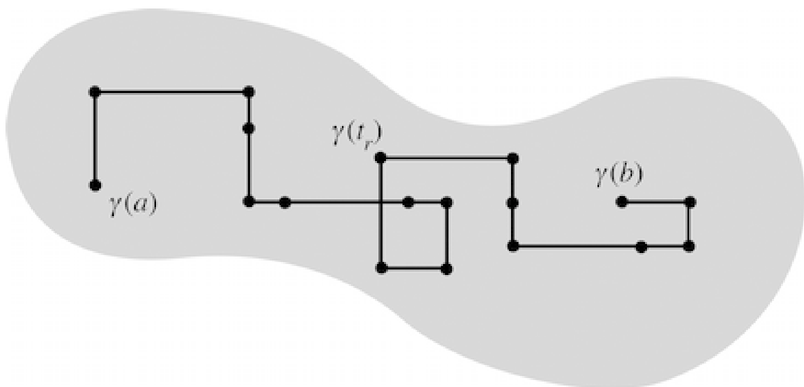

Figure 2.22 A step path.

It is often convenient to use a very simple type of path.

DEFINITION 2.37. A *step path* in S is a path $\gamma : [a, b] \to S$ together with a subdivision $a = t_0 < t_1 < \cdots < t_n = b$ such that on each subinterval $[t_{r-1}, t_r]$ either re γ or im γ is constant. This means that the image of γ consists of a finite number of straight line segments, each parallel to the real or imaginary axis, see Figure 2.22.

In other words, γ is a step path if $\gamma = \gamma_1 + \cdots + \gamma_n$ where each γ_r is a line segment parallel to the real or imaginary axis.

A subset $S \subseteq \mathbb{C}$ is *step-connected* if, given $z_1, z_2 \in S$, there exists a step path γ in S from z_1 to z_2 . □

Evidently a step-connected path is path-connected. The converse, however, is not always true: Example 2.36 is path-connected but not step-connected. A path as simple as the diagonal line segment $[0, 1 + \mathrm{i}]$ is path-connected but not step-connected. The main issue here is that unlike a general continuous path, a step path consists of finitely many straight line segments. But the requirement that these segments should be horizontal or vertical also comes into play.

These considerations are mentioned here to help develop intuition, but they do not cause problems for the theory, because every path-connected *open* set is step-connected. This constitutes the first success – albeit a minor one – for the Paving Lemma. To prove this, we first observe a simple fact:

LEMMA 2.38. *Any open disc is step-connected.*

Proof. Consider an open disc $S = N_r(z_0)$. Changing coordinates to $z - z_0$ (translating the origin) we may assume without loss of generality that $z_0 = 0$. Let $z_1 = x_1 + iy_1 \in S$. We show that there is a step path in S from z_1 to the centre 0. Let L_1 be the (vertical) line segment from z_1 to x_1, and L_2 be the (horizontal) line segment from x_1 to 0. Then simple inequalities show that L_1 and L_2 are inside S, and $L_1 + L_2$ is a step path from z_1 to 0.

By the same argument, if $z_2 \in S$ there is a step path $L_3 + L_4$ in S from z_2 to 0. But now $L_1 + L_2 - L_4 - L_3$ is a step path in S from z_1 to z_2. □

We can now prove:

PROPOSITION 2.39. *A path-connected open set is step-connected.*

Proof. Let z_1, z_2 in S, which is open, and let γ be a path from z_1 to z_2 in S.

By the Paving Lemma there is a finite sequence of open discs $D_1, \ldots, D_r \subseteq S$ and a path γ_r in each D_r such that $\gamma = \gamma_1 + \cdots + \gamma_r$.

By Lemma 2.38 there is a step path σ_r in each D_r with the same initial and final points as γ_r. Now $\sigma = \sigma_1 + \cdots + \sigma_r$ is a step path in S from z_1 to z_2. □

On the other hand, the parabola S in Example 2.36 is path-connected, but not step-connected. This does not conflict with the proposition above, because S is not open in $\mathbb{C}$.

Arbitrary sets need not be path-connected. However, for any set S we can define the relation $z_1 \sim z_2$ $(z_1, z_2 \in S)$ to mean that there is a path in S from z_1 to z_2. It is easy to see that $\sim$ is an equivalence relation (Exercise 7 below) and that each equivalence class is path-connected. These equivalence classes are called the *path-connected components*, or for simplicity the *connected components* or just *components* of S. (We always take S open in the sequel, so here is no conflict with the more general notions of connectedness and connected components that are standard in point-set topology, but irrelevant here.)

For instance, the connected components of

$$A = \{z \in \mathbb{C} : |z| \neq 1\}$$

are clearly

$$A_1 = \{z \in \mathbb{C} : |z| < 1\} \qquad A_2 = \{z \in \mathbb{C} : |z| > 1\}$$

If S is *open* then its connected components are all open: given z_0 in a connected component S_0 of S, since some $N_r(z_0) \subseteq S$ and $N_r(z_0)$ is path-connected, we must have $N_r(z_0) \subseteq S_0$, so S_0 is open.

A particularly interesting case is the complement of the image of a path $\gamma : [a, b] \to \mathbb{C}$, or more briefly the *complement of a path*, by which we mean

$$\{z \in \mathbb{C} : z \neq \gamma(t) \text{ for any } t \in [a, b]\}$$

The complement may be connected, for instance when

$$\gamma(t) = t \quad (t \in [0, 1])$$

On the other hand, it may have two or more components, Figure 2.23.

Indeed, the complement of $\gamma = \gamma_1 - \gamma_2$, where

$$\gamma_1(0) = 0, \gamma_1(t) = t + \mathrm{i}t \sin(\pi/t) \quad (0 < t \leq 1)$$
$$\gamma_2(0) = 0, \gamma_2(t) = t - \mathrm{i}t \sin(\pi/t) \quad (0 < t \leq 1)$$

has infinitely many components, Figure 2.24.

DEFINITION 2.40. A subset $S \subseteq \mathbb{C}$ is *bounded* if there exists $K \geq 0$ such that $|z| \leq K$ for all $z \in S$.

A subset $S \subseteq \mathbb{C}$ is *unbounded* if it is not bounded. □

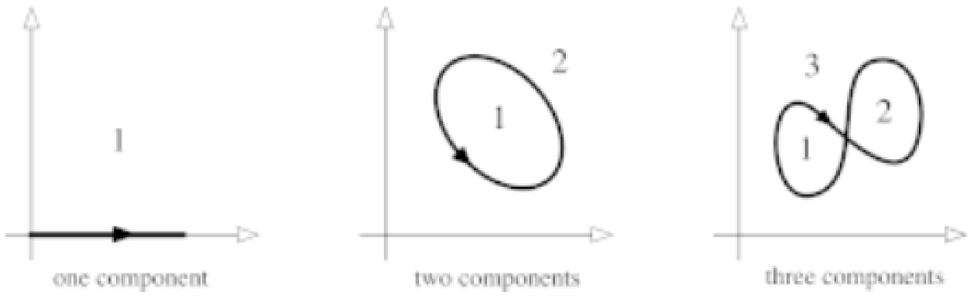

Figure 2.23 Paths whose complements have one, two, or three components.

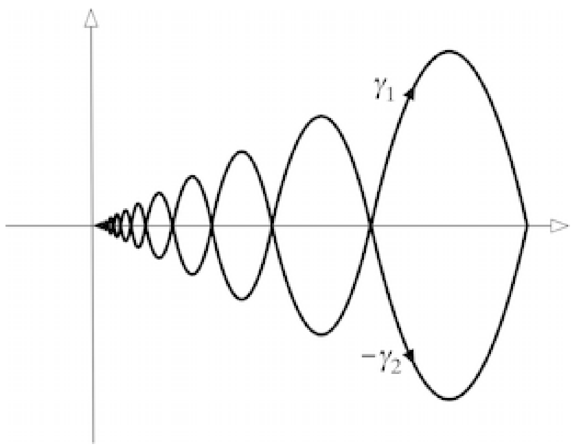

Figure 2.24 A path whose complement has infinitely many components.

However complicated a path may be, we can prove:

PROPOSITION 2.41. (i) *The image of a path in* $\mathbb{C}$ *is closed and bounded.*
(ii) *The complement of a path in* $\mathbb{C}$ *is open, and precisely one component is unbounded.*

Proof. For the second part of (i), let the path be $\gamma : [a,b] \to \mathbb{C}$, and let $m(t) = |\gamma(t)|$. Since it is the composition of two continuous functions, $m : [a,b] \to \mathbb{R}$ is continuous. By real analysis it is bounded, say $m(t) \leq K$, whence $|\gamma(t)| \leq K$ for all $t \in [a,b]$ and the image of $[a,b]$ lies inside the disc centre 0 of radius K.

The first part of (i) follows from (ii).

To prove (ii), suppose that z_0 lies in the complement T of γ. Then

$$\mu(t) = |\gamma(t) - z_0|$$

is a positive real number for all $t \in [a,b]$. Further, μ is continuous, so it is bounded and attains its bounds. That is, there exists $k \geq 0$ such that $\mu(t) \geq k$ for all $t \in [a,b]$, and there exists $t_0 \in [a,b]$ such that $\mu(t_0) = k$.

If $k = 0$ then $z_0 \in \gamma([a,b])$, but it is in T. So $k > 0$. Then $N_k(z_0) \subseteq T$, so T is open.

Finally, the image of γ lies inside

$$A = \{z \in \mathbb{C} : |z| \leq K\}$$

and the complement

$$B = \{z \in \mathbb{C} : |z| > K\}$$

is clearly connected. Therefore any unbounded component of T intersects B, and since components are disjoint, at most one of them can do so. Further, B lies inside some component, so there is exactly one unbounded component. □

REMARK 2.42. A closed, bounded set in $\mathbb{R}^n$, in particular one in $\mathbb{C} = \mathbb{R}^2$, is said to be *compact*. Compactness is a very important property in point-set topology (where the definition is given in a more general form). In this text we do not need to set up the full machinery of compactness, because the simple cases that arise can be tackled bare-handedly using real analysis – as we did in proving Proposition 2.41. □

We now define a special type of set that is fundamental to the theory in this text.

DEFINITION 2.43. A *domain* is a non-empty, path-connected, open subset of the complex plane. □

By Proposition 2.39, a domain is also step-connected, and this property will be useful later. This means that we can refer to a domain as being 'connected', meaning either path- or step-connected as appropriate.

As the theory of complex analysis unfolds, the reader will see the immense importance of this definition. We restrict the concept of a complex function f to those of the form $f : D \to \mathbb{C}$ where D is a domain. Openness of D lets us deal neatly with limits, continuity, and differentiability, because $z_0 \in D$ implies that $N_\varepsilon(z_0) \subseteq D$ for some $\varepsilon > 0$, so $f(z)$ is defined for all z near z_0. Connectedness of D guarantees there is a (step-) path between any two points in D, which lets us define the integral of f along such a path.

However, restricting complex functions to those defined on domains has far subtler consequences than merely providing a platform for the appropriate definition. Those who have studied the intricacies of real analysis in depth will find untold riches in complex analysis, quite unlike the real case. For instance, if two differentiable complex functions are defined on the same domain D and they happen to be equal on a small disc in D, they are equal throughout D. No such result holds for general differentiable functions in the real case. This result is just one of many that illustrate the beauty and simplicity of complex analysis. We shall not prove it until Chapter 10, but we mention it here to underline the fundamental importance of establishing appropriate topological foundations for the subject.

2.9 Space-filling Curves

A space-filling curve is defined by a path whose image fills a two-dimensional region, typically the unit square. Such curves show that our default mental image of a continuous path can be distinctly misleading. To explain why care is needed, we digress to describe one simple construction of such a path.

Let $\mathbb{U}^2$ be the unit square $\{x + iy : 0 \leq x \leq 1, 0 \leq y \leq 1\}$.

DEFINITION 2.44. A *space-filling curve* is a continuous path $\gamma : [0, 1] \to \mathbb{U}^2$ such that the image of γ is the whole of $\mathbb{U}^2$. □

Giuseppe Peano constructed the first space-filling curve in 1890, motivated by Georg Cantor's proof of 1878 that $[0, 1]$ and $\mathbb{U}^2$ have the same transfinite cardinal. In 1879 Eugen Netto had proved that there is no *continuous* bijection between these sets; Peano showed that such a map exists if we do not require it to be bijective. In 1891 Hilbert published another example of a space-filling curve, and included a picture. Both constructions are fairly complicated: Figure 2.25 shows early stages in their construction. Many others have since been found. Details can easily be found on the Internet, and are given in Sagan [18] and Bader [2].

We give a simpler construction, which at each stage produces a curve whose image is a square grid.

Peano's result was unexpected because it challenges the naive notion of 'dimension'. A curve has dimension 1, but $\mathbb{U}^2$ has dimension 2. It was counterintuitive that a continuous map can *increase* the dimension of a space. Exercises 15–17 show that there is no bound on the size of this increase. This phenomenon led to more careful theories of the concept of dimension for topological spaces, and the realisation that stronger conditions than continuity are needed for some intuitively plausible properties to hold.

The idea behind the construction of such curves is purely topological. It relies on simple properties of uniform convergence, as studied in real analysis courses. It requires the definition and properties of the closure of a subset $S \subseteq \mathbb{C}$, stated in Definition 2.5.

PROPOSITION 2.45. *Let $\gamma_n : [0, 1] \to \mathbb{U}^2$ be a sequence of continuous paths, for $n \in \mathbb{N}$. Suppose that:*

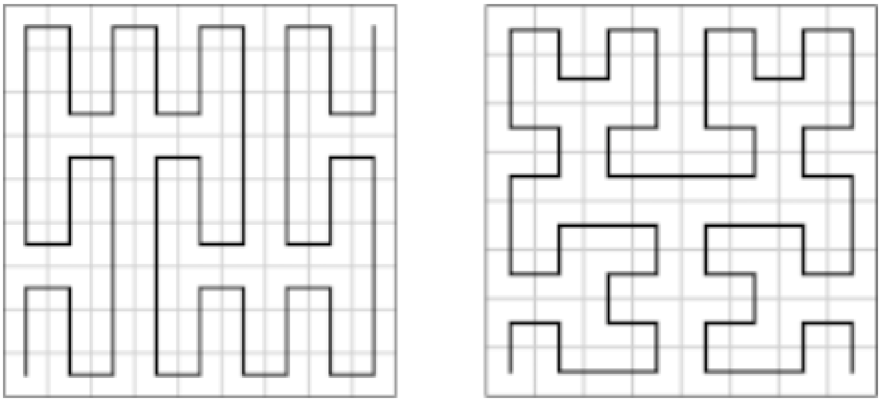

Figure 2.25 *Left*: An early stage in the construction of Peano's space-filling curve. *Right*: An early stage in the construction of Hilbert's space-filling curve.

(i) *The sequence of functions* (γ_n) *is uniformly convergent.*
(ii) *The closure of the union of the images of all* γ_n *is the whole of* $\mathbb{U}^2$.

Then the limit

$$\gamma = \lim_{n \to \infty} \gamma_n$$

is continuous, and its image is $\mathbb{U}^2$.

Before giving the proof, it must be made clear that we are not just asserting that γ comes *close* to every point of $\mathbb{U}^2$. It actually passes through every point of $\mathbb{U}^2$ in the sense that if $z \in \mathbb{U}^2$ there exists $t \in [0,1]$ such that $\gamma(t) = z$. In contrast, the union of the images of the γ_n is not the whole of $\mathbb{U}^2$. It is just dense in $\mathbb{U}^2$: that is, its closure is $\mathbb{U}^2$.

Proof. By condition (i), a basic theorem in real analysis implies that the limit γ exists and is continuous.

Let $z \in \mathbb{U}^2$. By condition (ii), for each $m > 0, m \in \mathbb{N}$, there exists $t_m \in [0,1]$ and $n_m \in \mathbb{N}$ such that

$$|\gamma_{n_m}(t_m) - z| < \frac{1}{m}$$

The subsequence γ_{n_m} also tends uniformly to γ. The sequence (t_m) lies in $[0,1]$ which is closed and bounded, so it has a convergent subsequence, with limit t_0. We claim that $\gamma(t_0) = z$. By uniform continuity, if m is large enough,

$$|\gamma_{n_m}(t_0) - \gamma(t_0)| < \frac{1}{m}$$

So

$$|\gamma(t_0) - z| < \frac{2}{m}$$

for all m, so $\gamma(t_0) = z$ as claimed. □

We now construct such a sequence of maps $\gamma_n : [0,1] \to \mathbb{U}^2$. We prescribe it geometrically, but making the construction precise is routine.

Define γ_0 so that its image is the boundary of $\mathbb{U}^2$ described anticlockwise, parameterised proportionally to arc-length; see top left of Figure 2.26.

To obtain γ_1, replace each segment of γ_0 by a T-shaped polygon, as shown at the bottom of Figure 2.26. The extra part of the T always extends to the left of the segment, as oriented by the arrow. The path γ_1 goes along the original path to its midpoint, turns left, turns back on itself to return to the midpoint, and then continues along the original segment to its end point. If we do this for each of the four segments of γ_0 and add the T-shaped paths (whose ends and starts fit together correctly) we obtain the path γ_1 at top right of Figure 2.26. For clarity, the grey polygon shows the direction in which t traverses γ_1.

This path visits the centre of the square four times, but all segments are treated in the same manner, making it easier to define the γ_n inductively. We scale the arc-length parametrisation so that each γ_n has the same parametric interval $[0, 1]$.

We now repeat the same construction on every straight segment of γ_1 to obtain γ_2, and continue inductively to define γ_n for all $n \in \mathbb{N}$.

The image of γ_n is a square grid of lines separated by distance 2^{-n}. Clearly the union of these images is dense in $\mathbb{U}^2$: it consists of all $z = x + iy$ for which either x or y is of the form $m2^{-n}$ for $0 \leq m \leq 2^n$, where $m \in \mathbb{N}$. So condition (i) holds, and we can define γ as the limit of the γ_n.

By considering the construction step at the bottom of Figure 2.26, it is easy to prove that

$$|\gamma_n(t) - \gamma_{n+1}(t)| \leq \sqrt{2} \cdot 2^{-n}$$

from which condition (ii) follows (by summing a geometric series).

Therefore γ is a space-filling curve by Proposition 2.45.

Although the image of γ_n is a grid, the actual structure of γ becomes increasingly complicated because of the way it zigzags back and forth. If we write

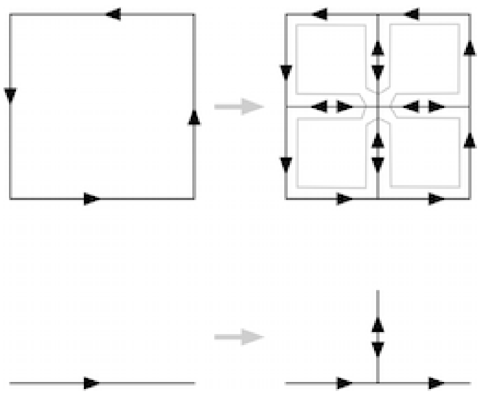

Figure 2.26 Inductive step used to construct a space-filling curve.

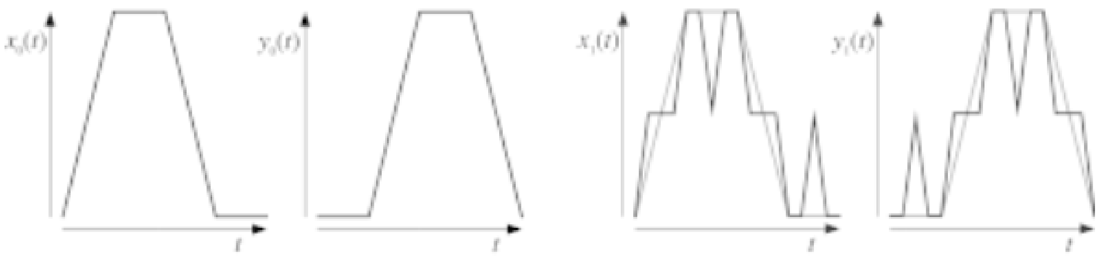

Figure 2.27 Graphs of real and imaginary parts of γ_0, γ_1. Grey lines included for comparison purposes.

$$\gamma_n(t) = x_n(t) + \mathrm{i}y_n(t)$$

so $x_n, y_n : [0, 1] \to [0, 1]$, then Figure 2.27 shows the graphs of x_0, y_0 and x_1, y_1.

The slopes of the parts of these graphs that are not horizontal become increasingly steep as n increases. Nevertheless, the limit is continuous – because of uniform convergence, and more intuitively, because the intervals on which such slopes occur become arbitrarily small (though there are increasingly many such intervals).

Exercise 15 shows that the unit hypercube in $\mathbb{R}^n$ is also the continuous image of an interval in $\mathbb{R}$. In fact, far more complicated sets can be the images of space-filling curves. The last word in such results is the Hahn–Mazurkiewicz Theorem: proved by Hans Hahn and Stefan Mazurkiewicz. This states that a set S is the image of a continuous map from $[0, 1]$ to a Hausdorff topological space if and only if S is compact, connected, locally connected, and metric. See Hocking and Young [9], page 129.

2.10 Exercises

1. Let $z_0 \in \mathbb{C}$ and $a, b \in \mathbb{R}$ be arbitrary. Draw the set consisting of all $z \in \mathbb{C}$ satisfying the following conditions. In each case, state whether the set is open, closed, or neither.

(i) $1 < |z| < 2$
(ii) $\mathrm{re}\, z \geq 0$ and $\mathrm{im}\, z \leq \mathrm{re}\, z$
(iii) $\mathrm{re}\, z \geq 0$ and $1 < |z| < 2$
(iv) $\mathrm{re}(zz_0) > 0$
(v) $a < \arg(z - z_0) < b$ where $-\pi < a < b \leq \pi$
(vi) $|z - \bar{z}_0| = |\bar{z} - z_0|$
(vii) $|z - \bar{z}_0| < |\bar{z} - z_0|$
(viii) $|z - \bar{z}_0| \leq |\bar{z} - z_0|$
(ix) $\mathrm{re}(z^2) > 0$
(x) $\mathrm{re}(z^2) \leq 0$
(xi) $\mathrm{re}(z^2) > 1$
(xii) $\mathrm{re}(z^2) \leq 1$

2. Prove Proposition 2.8 parts (i) and (ii).

3. Find the following limits, if they exist. If they do not exist, explain why.

(i) $\lim_{z\to 0} |z|/z$
(ii) $\lim_{z\to 0} |z|^2/z$
(iii) $\lim_{z\to 0} z/|z|^2$
(iv) $\lim_{z\to 0}(z - \operatorname{re} z)/\operatorname{im} z$

4. Prove from first principles that the following functions are continuous:

(i) $\operatorname{re} z$
(ii) $\operatorname{im} z$
(iii) $z + |z|$
(iv) $1/z \ (z \neq 0)$
(v) $|z|^2$

5. Prove Proposition 2.14.

6. For each of the following sets and pairs of points, define (if possible) (a) a path in the set between the two points, and (b) a step path in the set between the two points.

(i) $|z| < 2; 1 + \mathrm{i}, 1 - \mathrm{i}$
(ii) $|z| = 1; -\mathrm{i}, \mathrm{i}$
(iii) $1 \leq |z| \leq 2; \sqrt{2}, -\sqrt{2}$
(iv) $|\operatorname{re} z| > 5; -9 + 37\mathrm{i}, 1066 + \mathrm{i}(\pi + \sqrt{5}/17)$
(v) $|1 - |z|| > \frac{1}{2}; \mathrm{i}/3, 49(1 + \mathrm{i})$

7. Let $\gamma_1 = [\mathrm{i}, 0]$ and $\gamma_2 = [0, 1]$. Show that $\gamma_1 + \gamma_2$ is defined, but $\gamma_2 + \gamma_1$ is not.

8. Let $\gamma_1 = [0, 1]$ and $\gamma_2 = [1, 0]$. Show that $\gamma_1 + \gamma_2$ and $\gamma_2 + \gamma_1$ are both defined, but they are different paths.

9. Let S be a subset of $\mathbb{C}$. If $z, w \in S$, define $z \sim w$ if and only if there is a path from z to w. Show that $\sim$ is an equivalence relation. The equivalence classes are *components* of S. If S is open and non-empty, show that each component is a domain.

10. Let S be a path-connected subset of $\mathbb{C}$, and let $f : S \to \mathbb{C}$ be a continuous function. Prove that $f(S)$ is path-connected (even though it may not be open).

11. Give explicit functions for paths that describe the curves of Figure 2.28 in the direction indicated by the arrows. (All subpaths are parts of circles or line segments; $0 < \varepsilon < R$ and x_1, x_2, y_1, y_2 are positive reals.)

12. Let S be a subset of $\mathbb{C}$. A point $z \in \mathbb{C}$ is a *boundary point* of S if z is a limit point of S and also a limit point of the complement $\mathbb{C} \setminus S$. The *boundary* ∂S of S is the set of all boundary points of S. In the following cases, describe ∂S and state whether ∂S is path-connected. Draw a picture in each case.

(i) $S = \{z \in \mathbb{C} : 1 < |z| < 2\}$
(ii) $S = \{z \in \mathbb{C} : z \neq 0\}$
(iii) $S = \{z \in \mathbb{C} : z = x + \mathrm{i}y \text{ where } x, y \in \mathbb{Q}\}$
(iv) $S = \{z \in \mathbb{C} : 0 \leq \operatorname{re} z \leq 1,\ 0 \leq z \leq 1\}$
(v) $S =$ the intersection of the sets S in (iii) and (iv)
(vi) $S = \{z \in \mathbb{C} : z \neq \mathrm{i}y \text{ where } y \in \mathbb{R}, y \leq 0\}$
(vii) $S =$ the intersection of the sets S in (vi) and (ii)

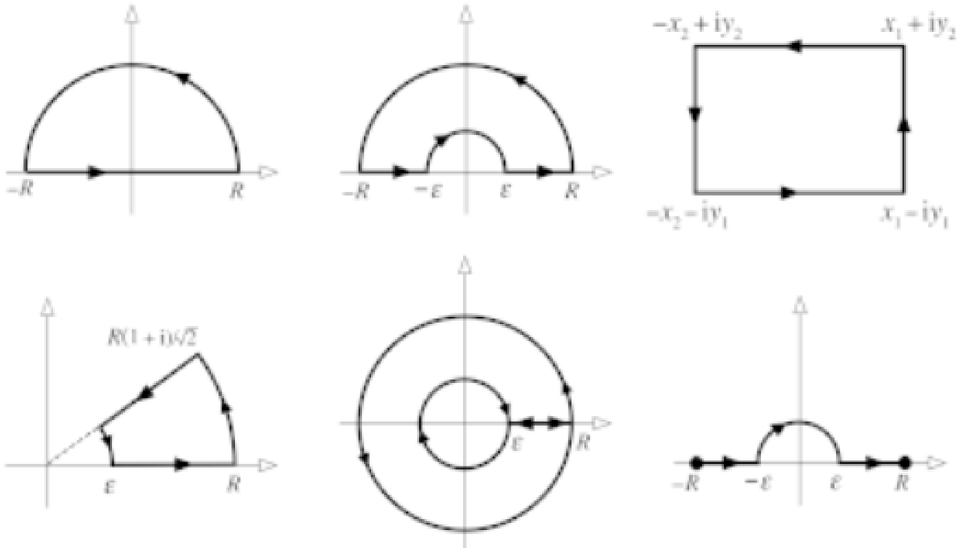

Figure 2.28 Six paths in $\mathbb{C}$: write down formulas for them.

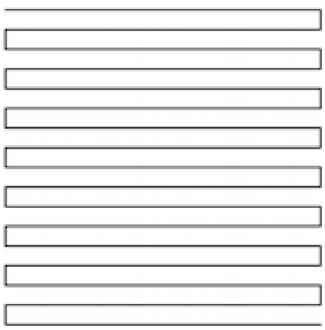

Figure 2.29 Why can paths like this not be used to construct a space-filling curve?

13. In the following cases the boundary of S (see Exercise 12) can be described as the image of a path. Draw a picture of S and specify a function γ giving such a path.

(i) $S = \{z \in \mathbb{C} : |z| \leq 1, \operatorname{im} z \geq 0\}$
(ii) $S = \{z \in \mathbb{C} : 1 \leq |z| \leq 2, \operatorname{im} z \geq 0\}$
(iii) $S = \{z \in \mathbb{C} : 0 \leq \operatorname{re} z \leq 1, 0 \leq \operatorname{im} z \leq 1\}$
(iv) $S = \{z \in \mathbb{C} : 1 \leq |z| \leq 2, 0 \leq \operatorname{im} z \leq \operatorname{re} z\}$

14. In the construction of a space-filling curve in Section 2.9, why can we not just take γ_n to be a path that zigzags n times across $\mathbb{U}^2$, as in Figure 2.29? Justify your answer.

15. Let $\mathbb{U}^3$ be the unit cube in $\mathbb{R}^3$. Show that there exists a continuous map from $[0, 1]$ onto $\mathbb{U}^3$. (Hint: the hard way is to construct paths with similar properties to those in Section 2.9. The easy way is to observe that $\mathbb{U}^3 = \mathbb{U}^2 \times [0, 1]$. Map $[0, 1] \times [0, 1]$ onto $\mathbb{U}^2 \times [0, 1]$; then compose with $\gamma : [0, 1] \to [0, 1] \times [0, 1]$.)

16. Let $\mathbb{U}^n$ be the unit hypercube in $\mathbb{R}^n$. Show that there exists a continuous map from $[0, 1]$ onto $\mathbb{U}^n$.

17. Let $m, n > 0$ be integers. Prove that there exists a continuous map from $\mathbb{U}^m$ onto $\mathbb{U}^n$, even when $m < n$.

18. If γ_n is a sequence of step paths in $\mathbb{U}^2$, can the union of their tracks equal $\mathbb{U}$ (rather than just being dense in $\mathbb{U}^2$)? Prove your answer correct.

19. Does there exist a continuous map $\gamma : [0, 1] \to \mathbb{U}^2$ that is one-one as well as onto? Justify your answer.

3 Power Series

Many of the more important functions studied in real analysis, such as the exponential and trigonometric functions, are most conveniently defined using power series. The familiar power series for the sine and cosine go back at least to around 1400 with the work of Madhava of Sangamagra, whose discovery of these series was reported in the 1501 *Yuktibhasa* of Jyesthadeva. This book also included Gregory's series

$$\tan^{-1} x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots$$

for the inverse tangent function:

> The first term is the product of the given sine and radius of the desired arc divided by the cosine of the arc. The succeeding terms are obtained by a process of iteration when the first term is repeatedly multiplied by the square of the sine and divided by the square of the cosine. All the terms are then divided by the odd numbers 1, 3, 5, ... The arc is obtained by adding and subtracting respectively the terms of odd rank and those of even rank.

Both Madhava and Jyesthadeva were members of the Kerala school of Indian mathematics. The series were rediscovered in the West in 1676, when Newton stated them in a letter to Henry Oldenburg, Secretary of the Royal Society. They were derived again by Abraham de Moivre in 1698, by James Bernoulli in 1702, and were widely used by Euler in the 1730s.

Power series are, if anything, even more important when we move from real to complex functions. In the 1820s Cauchy made considerable use of power series $\sum a_n z^n$ of a complex variable z. In particular, any real function having a power series development automatically gives rise to a complex function with the corresponding power series. This provides a natural method for extending real functions to the complex case. In the 1840s Weierstrass showed how to base the theory of complex functions on power series.

In this chapter we develop some elementary properties of sequences and series of complex numbers, mostly by direct analogy with the real case, and then specialise to a deeper study of power series.

3.1 Sequences

For our purposes, sequences are required only as a stepping-stone towards series. A (complex) *sequence* is a function

$$f : \mathbb{N} \to \mathbb{C}$$

where as usual $\mathbb{N}$ denotes the natural numbers $\{0, 1, 2, \dots\}$. It is convenient to write

$$z_n = f(n)$$

and to arrange the numbers z_n, called the *terms* of the sequence, in order as

$$z_0, z_1, z_2, \dots, z_n, \dots$$

Alternative notation

$$(z_n) \quad (n \geq 0)$$

or just

$$(z_n)$$

are often used for brevity, and it is sometimes helpful to start the sequence at 1 instead of 0, like this:

$$z_1, z_2, z_3, \dots, z_n, \dots$$

Now the corresponding f maps $\mathbb{N} \setminus \{0\}$ to $\mathbb{C}$.

DEFINITION 3.1. A sequence (z_n) *tends to the limit z as n tends to* ∞ if, given any real $\varepsilon > 0$, there exists a natural number $N(\varepsilon)$ such that

$$n > N(\varepsilon) \text{ implies } |z_n - z| < \varepsilon$$

A sequence that tends to a limit is *convergent*. □

This definition is identical to the usual one for real numbers, except that z, z_n may be complex, and the absolute value is as defined for $\mathbb{C}$.

We write

$$\lim_{n\to\infty} z_n = z$$

or

$$z_n \to z \text{ as } n \to \infty$$

The geometric content of this definition is that for sufficiently large n, all terms z_n lie inside arbitrarily small discs centred on z, as in Figure 3.1. This again is reminiscent of the real case.

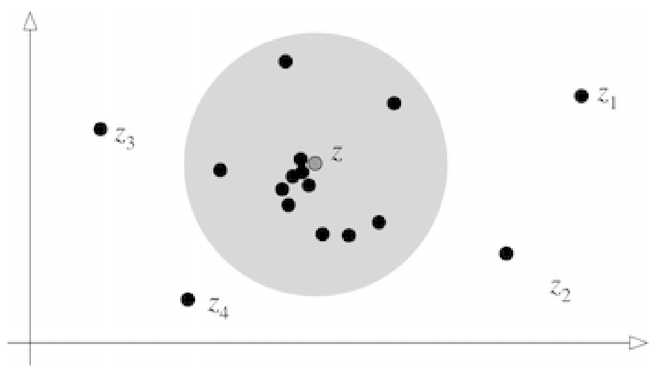

Figure 3.1 A complex sequence converging to a limit.

The problem of finding limits of sequences of complex numbers can be reduced directly to the real case:

LEMMA 3.2. *Let* (z_n) *be a sequence of complex numbers, with* $z_n = x_n + \mathrm{i}y_n$, $(x_n, y_n \in \mathbb{R})$. *Let* $z = x + \mathrm{i}y$ $(x, y, \in \mathbb{R})$. *Then*

$$\lim_{n\to\infty} z_n = z$$

if and only if

$$\lim_{n\to\infty} x_n = x \quad \text{and} \quad \lim_{n\to\infty} y_n = y$$

Proof. Suppose that $\lim_{n\to\infty} z_n = z$, and let $\varepsilon > 0$. There exists $N \in \mathbb{N}$ such that $|z_n - z| < \varepsilon$ if $n > N$. For any $w = u + \mathrm{i}v$ we have

$$|u| \le \sqrt{u^2 + v^2} = |w| \qquad |v| \le \sqrt{u^2 + v^2} = |w|$$

Therefore, taking $w = z_n - z$, for all $n > N$ we have

$$|x_n - x| \le |z_n - z| < \varepsilon$$
$$|y_n - y| \le |z_n - z| < \varepsilon$$

so $x_n \to x$ and $y_n \to y$.

Conversely, suppose that $x_n \to x$ and $y_n \to y$. Given $\varepsilon > 0$ there exist $M, N \in \mathbb{N}$ such that

$$|x_m - x| \le \varepsilon/2 \text{ for all } m \ge M$$
$$|y_n - y| \le \varepsilon/2 \text{ for all } n \ge N$$

Let $R = \max(M, N)$. If $m > R$ then

$$|z_n - z| \le |x_n - x| + |y_n - y| \le \varepsilon/2 + \varepsilon/2 = \varepsilon$$

so $z_n \to z$. □

Since (x_n) and (y_n) are real sequences, their limits x and y are also real and can be found using the techniques of real analysis.

Example 3.3. Let $z_n = (n + \mathrm{i}n^2 + 1)^{-1}$. Does z_n converge? If so, to what?

Separate z_n into real and imaginary parts, thus:

$$\begin{aligned} z_n &= (n + \mathrm{i}n^2 + 1)^{-1} \\ &= (n+1)[(n+1)^2 + n^4]^{-1} - \mathrm{i}n^2[(n+1)^2 + n^4] \end{aligned}$$

using (1.8) of Chapter 1. Then

$$x_n = \frac{n+1}{(n+1)^2 + n^4} \to 0 \text{ as } n \to \infty$$
$$y_n = \frac{-n^2}{(n+1)^2 + n^4} \to 0 \text{ as } n \to \infty$$

so

$$z_n \to 0 + \mathrm{i} \cdot 0 = 0$$

□

A similar idea gives the complex version of the *General Principle of Convergence*:

THEOREM 3.4 (General Principle of Convergence). *A sequence* (z_n) *tends to a limit* z *as* $n \to \infty$ *if and only if for all* $\varepsilon > 0$ *there exists* $N(\varepsilon) \in \mathbb{N}$ *such that*

$$m, n > N \text{ implies } |z_m - z_n| < \varepsilon \tag{3.1}$$

Proof. First, let $z_n \to z$. Then there exists $N = N(\varepsilon)$ such that $|z_n - z| < \varepsilon/2$ for all $n > N$. If $m, n > N$ then

$$|z_m - z_n| < |z_m - z| + |z - z_n| < \varepsilon/2 + \varepsilon/2 = \varepsilon$$

so (3.1) holds.

Conversely, assume (3.1) holds. Let $x_n = \text{re}(z_n), y_n = \text{im}(z_n)$. If $m, n > N$ then

$$|x_m - x_n| \leq |z_m - z_n| < \varepsilon$$
$$|y_m - y_n| \leq |z_m - z_n| < \varepsilon$$

By the General Principle of Convergence for real sequences (that is, the same statement for a real sequence using the real absolute value), it follows that $x_n \to x$ and $y_n \to y$ for some $x, y \in \mathbb{R}$. By Lemma 3.2, $z_n \to z = x + \text{i}y$. □

Example 3.5. Use the General Principle of Convergence to show that the sequence defined by

$$z_n = \text{i}\sqrt{2} + \left(\frac{3 - 4\text{i}}{6}\right)^n$$

converges.

Compute

$$\begin{aligned} |z_m - z_n| &= \left|\left(\frac{3 - 4\text{i}}{6}\right)^m - \left(\frac{3 - 4\text{i}}{6}\right)^n\right| \\ &\leq \left|\left(\frac{3 - 4\text{i}}{6}\right)^m\right| + \left|\left(\frac{3 - 4\text{i}}{6}\right)^n\right| \\ &= \left(\frac{5}{6}\right)^m + \left(\frac{5}{6}\right)^n \\ &\leq 2 \cdot \left(\frac{5}{6}\right)^r \end{aligned}$$

where $r = \min(m, n)$. Since $\frac{5}{6} < 1$ we can make this less than any given $\varepsilon > 0$ by taking r large enough.

Note that trying to tackle this question using Lemma 3.2 directly does not work out very easily, although the intrepid reader who converts $(3 - 4\text{i})/6$ into polar coordinates stands a better chance). It may of course be shown from the definition that $z_n \to \text{i}\sqrt{2}$. □

3.2 Series

Given a sequence (z_n), we can form another sequence of *partial sums* defined by

$$s_r = \sum_{r=0}^{n} z_r = z_0 + z_1 + \cdots + z_n$$

If s_n tends to a limit $s \in \mathbb{C}$, we define

$$\sum_{r=0}^{\infty} z_r = s = \lim_{n\to\infty} s_n$$

and say that the *series*

$$\sum_{r=0}^{\infty} z_r$$

converges to s.

A series that converges to some s is said to be *convergent*. We often also write

$$s = z_0 + z_1 + z_2 + \cdots$$

when convenient, but note that this is a *definition* of the $+\cdots$ notation, since infinite additions do not have any meaning until we give them one.

A series that is not convergent is called *divergent*.

By way of this definition, any question about the convergence of a series can be turned into one about the corresponding sequence (s_n) of partial sums. For example, Theorem 3.4 applied to (s_n) yields:

LEMMA 3.6. *A series $\sum_{r=0}^{\infty} z_r$ converges if and only if, for all $\varepsilon > 0$, there exists $N(\varepsilon) \in \mathbb{N}$ such that*

$$m, n > N(\varepsilon) \text{ implies } \left| \sum_{r=m+1}^{n} z_r \right| < \varepsilon$$

Proof. Observe that $\sum_{r=m+1}^{n} z_r = s_n - s_m$. Now use Theorem 3.4. □

We can replace $m+1$ by m here, by working with $N+1$ if necessary. This is usually more convenient.

COROLLARY 3.7. *If $\sum_{r=0}^{\infty} z_r$ converges then $z_r \to 0$ as $r \to \infty$.* □

Example 3.8. Let

$$z_n = \left(\frac{3-4\mathrm{i}}{6}\right)^n$$

Does $\sum_{r=0}^{\infty} z_r$ converge?

We have

$$\begin{aligned}\left|\sum_{r=m}^{n}\left(\frac{3-4\mathrm{i}}{6}\right)^r\right| &\le \sum_{r=m}^{n}\left|\left(\frac{3-4\mathrm{i}}{6}\right)^r\right| \\ &= \sum_{r=m}^{n}\left(\frac{5}{6}\right)^r\end{aligned}$$

This series is a (finite) geometric series, whose sum is known to be

$$\left[\left(\frac{5}{6}\right)^m - \left(\frac{5}{6}\right)^{n+1}\right]\left(1 - \frac{5}{6}\right)^{-1}$$

which obviously may be made less than any required $\varepsilon > 0$ by taking m, n large enough. □

WARNING 3.9. There is an ancient, venerable, and politically incorrect joke about two country yokels who each buy horses and want to be sure which is which. So they cut the tail off one. Next day, someone has cut the tail off the other. So they try again with an ear . . . with the same result. Finally, in desperation, one says to the other 'Tell you what: you take the black one and I'll take the white one.'

For some reason, many students become confused between sequences and series. There is no reason to fall into this trap, because:

$$\text{series are the ones that begin with } \sum_{r=0}^{\infty}$$

This does not eliminate *all* sources of confusion, but it's a good start. □

As with sequences, it is sometimes preferable to start at 1 rather than 0. The series then takes the form

$$\sum_{r=1}^{\infty} z_r$$

or

$$z_1 + z_2 + z_3 + \cdots$$

whose precise definition is left to the reader. Variations, such as

$$\sum_{r=3}^{\infty} z_r$$

are also feasible. However, this is no more than $\sum_{r=0}^{\infty} z_r$ where we set $z_0 = z_1 = z_2 = 0$.

To simplify notation, we often abbreviate such sums to

$$\sum z_r$$

when the limits of summation are clear.

The summation of complex series may, if desired, be reduced to equivalent real series. Applying Lemma 3.2 to the partial sums, we obtain an immediate proof of:

LEMMA 3.10. *Let $z_r = x_r + \mathrm{i}y_r$ where $x_r, y_r \in \mathbb{R}$. Then $\sum z_r$ converges if and only if both real series $\sum y_r$ and $\sum y_r$ converge. If so,*

$$\sum z_r = \sum x_r + \mathrm{i} \sum y_r$$

Example 3.11. 11 Let $z_n = (\mathrm{i})^n/n^2$. Does $\sum z_r$ converge?

We have

$$x_n = \begin{cases} 0 & \text{if } n \text{ is odd} \\ (-1)^{n/2}/n^2 & \text{if } n \text{ is even} \end{cases}$$

Ignoring the zero terms, $\sum x_n$ is an alternating series (its terms are alternately positive and negative) whose terms tend to zero, so it converges.

Similarly, $y_n = \mathrm{im}(z_n)$ yields a convergent alternating series. Hence $\sum z_r$ converges. □

In the same manner, considering real and imaginary parts, or by mimicking the usual proof for real series, we get:

LEMMA 3.12. *Let $\sum a_r$ and $\sum b_r$ be convergent series, and let c be a complex number. Then*

$$\sum(a_r + b_r) = \sum a_r + \sum b_r$$
$$\sum ca_r = c\sum a_r$$ □

When verifying that a complex series converges without computing its actual sum, it is often better to concentrate on the modulus rather than the real and imaginary parts. By analogy with the real case, we introduce:

DEFINITION 3.13. A series $\sum z_r$ is *absolutely convergent* if and only if the series $\sum |z_r|$ is convergent. □

THEOREM 3.14. *An absolutely convergent series is convergent.*

Proof. Let $\sum z_r$ be absolutely convergent and let $z_r = x_r + \mathrm{i}y_r$. Then $|x_r| \le |z_r|$ and $|y_r| \le |z_r|$. But $\sum |z_r|$ is convergent, so by the comparison test for real convergence, $\sum |x_r|$ and $\sum |y_r|$ are convergent. Since absolute convergence implores convergence in the real case, $\sum x_r$ and $\sum y_r$ are convergent. The result follows by Lemma 3.10. □

Example 3.15. $\sum z_n = \sum(\mathrm{i}^n/n^2)$ converges because $\sum |z_n| = \sum(1/n^2)$ converges. (Compare this proof with the previous example.) □

Theorem 3.14 is useful; because it lets us verify convergence of many *complex* series $\sum z_r$ by reference to the *real* series $\sum |z_r|$. Since the latter has *positive* terms there is a complex version of:

THEOREM 3.16 (Comparison Test). *Let $\sum a_r$ and $\sum b_r$ be complex series, with $\sum a_r$ absolutely convergent. If*

$$|b_r| < K|a_r| \quad (r > N)$$

for some positive K and integer N, then $\sum b_r$ is absolutely convergent, hence convergent.
□

There is also a complex version of:

THEOREM 3.17 (Ratio Test). *Let $\sum a_r$ be a complex series with non-zero terms such that*

$$\lim_{r\to\infty} \frac{|a_r|}{|a_r - 1|} = \lambda$$

If $\lambda < 1$, the series $\sum a_r$ is absolutely convergent. If $\lambda > 1$, the series is divergent. If $\lambda = 1$, it may be convergent or divergent.

Proof. For $\lambda < 1$, let $\rho = \frac{1}{2}(\lambda + 1)$. Then $\lambda < \rho < 1$ and there exists N such that

$$|a_r|/|a_{r-1}| < \rho \quad (r > N)$$

Hence

$$|a_r| < \rho|a_{r-1}| < \rho^2|a_{r-2}| < \cdots < \rho^{r-N}|a_N| \quad (r > N)$$

and $\sum a_r$ converges by comparison with $\sum \rho^r$ for real $\rho < 1$.

In the case $\lambda > 1$, we see that for some N

$$|a_r|/|a_{r-1}| > 1 \quad (r > N)$$

By Corollary 3.7, $\sum a_r$ cannot converge because the terms a_r do not tend to zero.

The final case $\lambda = 1$ applies to $\sum 1/r$ and $\sum 1/r^2$. The first diverges, the second converges. So both possibilities may occur. □

3.3 Power Series

We can now give a formal introduction to one of the central ideas in complex analysis:

DEFINITION 3.18. Let $z_0 \in \mathbb{C}$. A series of the form

$$\sum_{n=0}^{\infty} a_n(z - z_0)^n \tag{3.2}$$

with *coefficients* $a_n \in \mathbb{C}$ is a *power series about* z_0. □

By the change of variable $z' = z - z_0$ we can usually reduce properties of power series to the case $z_0 = 0$. Here, convergence is governed by the following results:

THEOREM 3.19. (i) *If a power series $\sum a_n z^n$ converges for $z = z_1 \neq 0$, then it converges absolutely for all z with $|z| < |z_1|$.*

(ii) *If a power series $\sum a_n z^n$ diverges for $z = z_2$, then it diverges for all z with $|z| > |z_2|$.*

Proof. (i) If a power series $\sum a_n z^n$ converges then $|a_n z^n| \to 0$ as $n \to \infty$, by Corollary 3.7. Thus there exists $K \in \mathbb{R}$ such that $|a_n z^n| < K$ for all n. If $|z| < |z_1|$ then $q = |z/z_1| < 1$. Now

$$|a_n z^n| = |a_n z_1^n||z/z_1|^n < Kq^n$$

so by the comparison test, $\sum |a_n z^n|$ converges.

(ii) If $|z| > |z_2|$ and $\sum a_n z^n$ converges then by (i), $\sum a_n z_2^n$ also converges, a contradiction. So $\sum a_n z^n$ diverges. □

These results lead to an important concept:

DEFINITION 3.20. Let

$$R = \sup\{|z| : \text{there exists } z \text{ such that } |a_n z^n| \text{ converges}\}$$

(allowing $R = \infty$ if no real supremum exists) then it follows at once that

$$\sum a_n z^n \begin{cases} \text{converges} & \text{for } |z| < R \\ \text{diverges} & \text{for } |z| > R \end{cases}$$

(We cannot yet say what happens for $|z| = R$.) We define R to be the *radius of convergence* of the series, and the set

$$\{z \in \mathbb{C} : |z| < R\}$$

is the *disc of convergence*, classically often called the *circle of convergence* because geometrically it is the interior of a circle centred at the origin. In extreme cases this may be just the origin when $R = 0$, or the whole of $\mathbb{C}$ when $R = \infty$. □

In the general case (3.2) where z_0 is arbitrary, we apply the same definition with z replaced by $z' = z - z_0$. Now the conditions

$$\sum a_n (z - z_0)^n \begin{cases} \text{converges} & \text{for } |z - z_0| < R \\ \text{diverges} & \text{for } |z - z_0| > R \end{cases}$$

determine the radius of convergence R, and the disc of convergence is

$$\{z \in \mathbb{C} : |z - z_0| < R\}$$

Example 3.21. The series $1 + z + z^2 + \cdots$ is convergent for $|z| < 1$, since it is absolutely convergent in that case. It diverges for $z = 1$ and hence for all z with $|z| > 1$. Therefore its radius of convergence is 1. □

When $|a_r|/|a_{r-1}|$ tends to a limit as r tends to infinity, the radius of convergence can be computed as follows. Let

$$\lim_{r\to\infty} \frac{|a_r|}{|a_{r-1}|} = l$$

(which is positive or zero). Then

$$\lim_{r\to\infty} \frac{|a_r z^r|}{|a_{r-1} z^{r-1}|} = l|z|$$

By the ratio test, for $l > 0$ the series is convergent for $l|z| < 1$ and divergent for $l|z| > 1$. Hence the radius of convergence is $1/l$. Put another way, the radius of convergence of $\sum a_r z^r$ is

$$\lim_{r\to\infty} \frac{|a_{r-1}|}{|a_r|}$$

provided this limit exists. If $l = 0$ then the radius of convergence is ∞.

Example 3.22. If $a_n = 1/n$ so $\sum a_n z^n = \sum z^n/n$, then

$$\lim_{r\to\infty} \frac{|a_{r-1}|}{|a_r|} = 1$$

so the radius of convergence is 1. □

In general, the ratio $|a_r|/|a_{r-1}|$ may not tend to a limit. Sometimes it is possible to spot the radius of convergence by native wit, but when all else fails we can use a powerful technique that works in *all* cases:

THEOREM 3.23. *The radius of convergence of $\sum a_n z^n$ is given by*

$$1/R = \limsup |a_n|^{1/n} \tag{3.3}$$

Proof. Define R using (3.3).

Suppose first that $|z| < R$. Then we can choose ρ such that $|z| < \rho < R$. By the definition of lim sup, we have $1/\rho > |a_n|^{1/n}$ for all n larger than some N. Hence $|a_n| < 1/\rho^n$, so

$$|a_n z^n| = |a_n \rho^n| \cdot |z/\rho|^n < |z/\rho|^n$$

Now $|z/\rho| < 1$ so $\sum_{n=N}^{\infty} |z/\rho|^n$ converges. By the comparison test, $\sum_{n=N}^{\infty} a_n z^n$ converges, so $\sum_{n=0}^{\infty} a_n z^n$ converges.

Next suppose that $|z| > R$. Choose ρ such that $|z| > \rho > R$. Then $1/\rho < |a_n|^{1/n}$ for all n larger than some N. By comparison with $\sum |z/\rho|^n$, $\sum_{n=N}^{\infty} a_n z^n$ diverges. □

Example 3.24. The series

$$\sum (-1)^n \frac{z^n}{n}$$

has radius of convergence R, where

$$1/R = \limsup (1/n)^{1/n} = 1$$

so $R = 1$. The series converges for $|z| < 1$ and diverges for $|z| > 1$. When $|z| = 1$ further analysis is required, for which we do not yet have the technique.

The series $\sum \frac{z^n}{n!}$ has radius of convergence R, where

$$1/R = \limsup (1/n!)^{1/n} = 0$$

Therefore $R = \infty$ and the series is absolutely convergent for all $z \in \mathbb{C}$.

The series

$$\sum (-1)^n \frac{z^{2n}}{2n!} \qquad \sum (-1)^n \frac{z^{2n+1}}{(2n+1)!}$$

similarly have $R = \infty$, so they are absolutely convergent for all $z \in \mathbb{C}$. □

In the next chapter we use the last three power series above to define the complex exponential, cosine, and sine functions. In order to derive some of their basic properties, we will use a theorem on the multiplication of power series. Roughly, this says that if we multiply them 'term by term' and collect terms in the right way, we get a series converging to the product. More precisely:

THEOREM 3.25. *Suppose that* $\sum a_n$ *and* $\sum b_n$ *are absolutely convergent, with sums* a, b *respectively. Let*

$$c_r = a_0 b_r + a_1 b_{r-1} + \cdots + a_r b_0 \tag{3.4}$$

Then $\sum c_n$ *is convergent, and its sum is ab.*

Proof. This is a direct extension of the corresponding result in real analysis. For readers unfamiliar with this theorem, we give details in Section 3.5. □

3.4 Manipulating Power Series

The results derived so far let us calculate with power series 'as if they are infinite polynomials', *provided they are absolutely convergent*. To see this, let $\sum a_n z^n$ and $\sum b_n z^n$ be power series with radii of convergence R_a, R_b respectively. Let $|z| < \min(R_a, R_b)$. By Lemma 3.12,

$$\sum a_n z^n + \sum b_n z^n = \sum (a_n + b_n) z^n$$

By Theorem 3.25,

$$\left(\sum a_n z^n\right)\left(\sum b_n z^n\right) = \sum c_n z^n$$

where c_r is defined by (3.4). If we replace the infinite sum by a finite one, $\sum_1^n$, these formulas become the usual ones for addition and multiplication of polynomials.

It is this feature that makes power series so useful: we can *calculate* with them relatively easily. In fact, we can use familiar algebraic methods.

We take advantage of this to exhibit some important features of the complex exponential and trigonometric functions.

DEFINITION 3.26. The complex exponential and trigonometric functions are defined by three power series:

$$\exp z = \sum \frac{z^n}{n!}$$

$$\cos z = \sum (-1)^n \frac{z^{2n}}{(2n)!}$$

$$\sin z = \sum (-1)^n \frac{z^{2n+1}}{(2n+1)!}$$

These series are motivated by the corresponding real series. They are absolutely convergent for all $z \in \mathbb{C}$ by Example 3.24. □

We begin with the expression $\cos\theta + \mathrm{i}\sin\theta$ of Section 1.7. Adding the power series, we obtain

$$\cos\theta + \mathrm{i}\sin\theta = \sum c_r\theta^r$$

where

$$\begin{aligned} c_r &= \frac{(-1)^{r/2}}{r!} && \text{if } r \text{ is even} \\ c_r &= \frac{\mathrm{i}(-1)^{(r-1)/2}}{r!} && \text{if } r \text{ is odd} \end{aligned}$$

Now $\mathrm{i}^2 = -1, \mathrm{i}^3 = -\mathrm{i}, \mathrm{i}^4 = 1$, so

$$\begin{aligned} \mathrm{i}^r &= (-1)^{r/2} && \text{if } r \text{ is even} \\ \mathrm{i}^r &= \mathrm{i}(-1)^{(r-1)/2} && \text{if } r \text{ is odd} \end{aligned}$$

Hence

$$\sum c_r\theta^r = \sum \frac{\mathrm{i}^r\theta^r}{r!} = \exp \mathrm{i}\theta$$

We thus have the important formula

$$\cos\theta + \mathrm{i}\sin\theta = \exp \mathrm{i}\theta$$

Therefore the polar coordinate form $z = r(\cos\theta + \mathrm{i}\sin\theta)$ for a complex number z can be written as $z = r \cdot \exp \mathrm{i}\theta$, or more briefly as $r\mathrm{e}^{\mathrm{i}\theta}$. (We define powers z^a of complex numbers in Section 7.3, and this will justify the notation $\exp \mathrm{i}\theta = \mathrm{e}^{\mathrm{i}\theta}$.)

Next, we apply the formula for the product of two power series to evaluate

$$\exp z \cdot \exp w$$

where $z, w \in \mathbb{C}$. We obtain

$$\begin{aligned} \exp z \cdot \exp w &= \left(\sum \frac{z^n}{n!}\right)\left(\sum \frac{w^n}{n!}\right) \\ &= \sum_{n=0}^{\infty}\left(\sum_{r=0}^{n} \frac{1}{r!}\frac{1}{(n-r!)} z^r w^{n-r}\right) \\ &= \sum_{n=0}^{\infty}\frac{1}{n!}\left(\sum_{r=0}^{n}\binom{n}{r} z^r w^{n-r}\right) \end{aligned}$$

By the Binomial Theorem this is

$$\sum_{n=0}^{\infty}\frac{1}{n!}(z+w)^n = \exp(z+w)$$

This proves

$$\exp z \cdot \exp w = \exp(z+w)$$

3.5 Products of Series

We sketch a proof of Theorem 3.25 above: if $\sum a_n$ and $\sum b_n$ are absolutely convergent, with sums a, b respectively, then $ab = \sum c_n$, where

$$c_r = a_0 b_r + a_1 b_{r-1} + \cdots + a_r b_0$$

The proof may be easier to follow using Figure 3.2, which represents all possible cross-products $a_s b_t$ on a square grid. The idea is that partial sum of $\sum c_n$ is the sum of terms in a triangular region like the one shown in light shading. In contrast, products of partial sums of $\sum a_n z^n$ and $\sum b_n z^n$ correspond to terms in a rectangle, such as the one shown in dark shading. Our main task is to estimate the sum of terms in suitable triangular regions by approximation them by rectangles. Here are the details.

Let $\sum |a_n| = A, \sum |b_n| = B$. Given $\varepsilon > 0$ we can choose N large enough so that all of the following conditions hold. (To do this, choose an N for each condition and take the maximum of the three Ns chosen.)

(i) $\left| \left(\sum_{n=0}^{N} a_n \right) \left(\sum_{n=0}^{N} b_n \right) - ab \right| < \dfrac{\varepsilon}{A+B+1}$

(ii) $\sum_{n=N+1}^{2N} |a_n| < \dfrac{\varepsilon}{A+B+1}$

(iii) $\sum_{n=N+1}^{2N} |b_n| < \dfrac{\varepsilon}{A+B+1}$

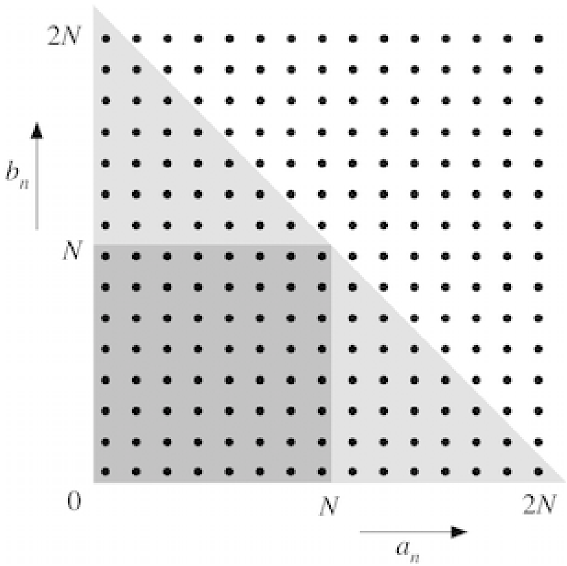

Figure 3.2 Terms of $\sum a_n z^n$, $\sum b_n z^n$, and their product.

Then

$$\left|\sum_{n=0}^{2N} c_n - ab\right| \leq \left|\sum_{n=0}^{2N} c_n - \left(\sum_{n=0}^{N} a_n\right)\left(\sum_{n=0}^{N} b_n\right)\right| + \left|\left(\sum_{n=0}^{N} a_n\right)\left(\sum_{n=0}^{N} b_n\right) - ab\right|$$

which from Figure 3.2 is less than or equal to

$$\left(\sum_{n=N+1}^{2N} |a_n|\right)\left(\sum_{n=0}^{N} |b_n|\right) + \left(\sum_{n=0}^{N} |a_n|\right)\left(\sum_{n=N+1}^{2N} |b_n|\right) + \frac{\varepsilon}{A+B+1}$$

$$< \frac{\varepsilon B}{A+B+1} + \frac{\varepsilon A}{A+B+1} + \frac{\varepsilon}{A+B+1} = \varepsilon$$

Therefore $\sum c_n$ converges to ab.

3.6 Exercises

1. Determine whether the following sequences converge, and find the limits of those that converge.

(i) $((1 + \mathrm{i})^n)$
(ii) $((1 + \mathrm{i})^n/n)$
(iii) $((1 + \mathrm{i})^n/n!)$
(iv) $(1/(1 + \mathrm{i})^n)$
(v) $(n/(1 + \mathrm{i})^n)$
(vi) $(n!/(1 + \mathrm{i})^n)$

2. For what values of $z \in \mathbb{C}$ does each of the following sequences converge?

(i) (z^n)
(ii) (z^n/n)
(iii) $(n!z^n)$
(iv) $(z^n/n!)$
(v) (z^n/n^k) where k is a positive integer.
(vi) $(a(a-1)\cdots(a-n+1)z^n/n!)$ where a is a fixed complex number.

3. Let $a \in \mathbb{R}$ have decimal expansion

$$a = a_0.a_1a_2\ldots a_n\ldots$$

where each a_n is an integer and $0 \leq a_i \leq 9$ for $n \geq 1$. Find all values of a for which the sequence (a_n) converges.

4. Let $z_n = \frac{1}{2}(\mathrm{i}^n + (-\mathrm{i})^n)$. Write down the first few terms of the sequence (z_n). Derive similar expressions for the nth terms of the sequences that begin as follows:

(i) $1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, \ldots$
(ii) $1, 0, 0, 0, -1, 0, 0, 0, 1, 0, 0, 0, -1, 0, 0, 0, \ldots$
(iii) $1, 0, -\frac{1}{2}, 0, \frac{1}{3}, 0, -\frac{1}{4}, 0, \ldots$
(iv) $1, 0, 1\frac{1}{2}, 0, \frac{2}{3}, 0, 1\frac{1}{4}, 0, \frac{4}{5}, 0, 1\frac{1}{6}, 0, \ldots$

5. Let (u_n) be a convergent sequence in $\mathbb{C}$. Let $v_n = (\sum_{r=1}^{n} u_r)/n$. By writing $v_n = v'_n + v''_n$ where

$$v'_n = \frac{1}{n}\left(\sum_{r\leq\sqrt{n}} u_r\right) \qquad v''_n = \frac{1}{n}\left(\sum_{\sqrt{n}<r\leq n} u_r\right)$$

show that (v_n) converges to the same limit as (u_n).

6. Find the radius of convergence of the following series:
   (i) $\sum z^n/n$
   (ii) $\sum z^n/n!$
   (iii) $\sum n!z^n$
   (iv) $\sum n^k z^n$ where k is a positive integer.
   (v) $\sum z^{n!}$

7. The 0th order Bessel function $J_0(z)$ is defined by the power series

$$J_0(z) = \sum_{n=0}^{\infty}(-1)^n \frac{1}{(n!)^2}\frac{z^n}{2n}$$

Find its radius of convergence.

8. Find the radius of convergence of the following series:
   (i) $z - \dfrac{z^3}{3!} + \dfrac{z^5}{5!} - \dfrac{z^7}{7!} + \cdots$
   (ii) $1 - \dfrac{z^2}{2!} + \dfrac{z^4}{4!} - \dfrac{z^6}{6!} + \cdots$
   (iii) $z - \dfrac{z^2}{2} + \dfrac{z^3}{3} - \dfrac{z^4}{4} + \cdots$
   (iv) $1 + az + \dfrac{a(a-1)}{2!}z^2 + \cdots + \dfrac{a(a-1)\cdots(a-n+1)}{n!}z^n + \cdots$ where $a \in \mathbb{C}$.
   (Note that in part (iv) the radius of convergence may differ for different values of a.)

9. Show that $\sum_{n=1}^{\infty} z^{n!}$ converges for $|z| < 1$, but diverges for infinitely many z with $|z| = 1$.

10. Suppose that $\sum a_n z^n$ has radius of convergence R and let C be the circle $\{z \in \mathbb{C} : |z| = R\}$. Prove or disprove the following (which may or may not be true):
    (i) If $\sum a_n z^n$ converges at some point on C then it converges everywhere on C.
    (ii) If $\sum a_n z^n$ converges absolutely at some point on C then it converges absolutely everywhere on C.
    (iii) If $\sum a_n z^n$ converges at every point on C, except possibly one, then it converges absolutely everywhere on C.
    (Hint: the series $\sum z^n/n$ could prove useful in this question.)

11. If $\sum a_n z^n$ has radius of convergence R, use the formula $1/R = \limsup |a_n|^{1/n}$ to find the radius of convergence of:
    (i) $\sum n^3 a_n z^n$
    (ii) $\sum a_n z^{3n}$
    (iii) $\sum a_n^3 z^n$

12. Prove that if each of the series $\sum a_n z^n, \sum b_n z^n$, and $\sum a_n b_n z^n$ has radius of convergence equal to 1, then so have the series $\sum a_n b_n^2 z^n$ and $\sum a_n^2 b_n z^n$.

13. For $|a_n| \le 1$, show that $\sum a_n z^n$ is absolute convergent for all $|z| < 1$. If $\sum a_n z^n = f(z)$ for $|a_n| < 1, |z| < 1$, show that

$$|f(z)| \le \frac{1}{1 - |z|}$$

14. Prove that, for $z \neq 1$,

$$\sum_{n=1}^{k} \frac{z^n}{n} = \frac{z}{1-z}\left(\sum_{n=1}^{k-1} \frac{1}{n(n+1)} - \sum_{n=1}^{k-1} \frac{z^n}{n(n+1)} + \frac{1-z^k}{k}\right)$$

Show that the series $\sum_{n=1}^{\infty} z^n/n$ and $\sum_{n=1}^{\infty} z^n/(n(n+1))$ have radius of convergence 1; that the latter series converges everywhere on $|z| = 1$, while the former series converges everywhere on $|z| = 1$ except $z = 1$.

15. Suppose that the power series $\sum_{n=0}^{\infty} a_n z^n$ has a recurring sequence of coefficients; that is, $a_{n+k} = a_n$ for some fixed positive integer k and all n. Prove that the series converges for $|z| < 1$ to a rational function $p(z)/q(z)$ where p, q are polynomials, and that the roots of q are all on the unit circle.
What happens if $a_{n+k} = a_n/k$ instead?

4 Differentiation

The derivative of a real function is defined by a limiting process, which generalises without difficulty to complex functions now that we have developed the necessary concept of a limit. There are few surprises in this chapter: the results on differentiation of sums, products, composites of functions, and power series parallel the real case – and even the proofs are essentially unchanged. One minor surprise is that the condition of differentiability implies certain relations between the real and imaginary parts of a complex function, specified by the *Cauchy–Riemann Equations*. We apply these equations immediately to prove that a function on a domain has zero derivative, it must be constant. While the *result* is analogous to the real case, the proof is not. The idea of differentiation can also be extended to *hybrid* functions $\mathbb{R} \to \mathbb{C}$ and $\mathbb{C} \to \mathbb{R}$, still without surprises.

To counter the impression that all the results of real analysis go through without change to the complex case, the final section previews a dramatic difference between the two theories: every differentiable complex function can be differentiated arbitrarily many times.

4.1 Basic Results

By analogy with the real case we state:

DEFINITION 4.1. (a) A complex function f defined on an open set S is *differentiable at* $z_0 \in S$ if

$$\lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}$$

exists.

(b) The value of this limit is then defined to be the *derivative* $f'(z_0)$ at that point.

(c) If f is differentiable at every $z_0 \in S$ then f is *differentiable*. In this case the derivative can also be considered as a function $f' : S \to \mathbb{C}$.

(d) If $f'(z)$ is also differentiable, we define the *second derivative* to be

$$f''(z) = \lim_{z \to z_0} \frac{f'(z) - f'(z_0)}{z - z_0}$$

(e) Repeating this process we obtain the usual notion of higher derivatives $f''(z_0)$, $f'''(z_0), \ldots, f^{(n)}(z_0)$ at z_0. □

Example 4.2. Suppose $f(z) = z^2$. Then

$$\lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0} = \lim_{z \to z_0} \frac{z^2 - z_0^2}{z - z_0} = 2z_0$$

Hence $f'(z_0) = 2z_0$ for all $z_0 \in \mathbb{C}$. Similarly $f''(z_0) = 2$ and $f^{(n)}(z_0) = 0$ for all $n \geq 3$. □

Several alternative notations are used of differentiation, the two most popular being $\mathrm{D}f(z)$ or $\mathrm{d}f(z)/\mathrm{d}z$ in place of $f'(z)$. The second derivatives are then denoted by $\mathrm{D}^2 f(z)$ and $\mathrm{d}^2 f(z)/\mathrm{d}z^2$ respectively. Sometimes (especially in classical texts) $f(z)$ is denoted by w and its derivative by $\mathrm{d}w/\mathrm{d}z$. In this text we use the notation $f'(z)$ because, when we think of the derivative as a function in its own right, we can then denote it by f'. The same goes for higher derivatives f'' and so on.

In many ways, results about complex differentiation follow naturally by analogy with the real case, as the next few results illustrate.

PROPOSITION 4.3. *If f is differentiable at z_0, then f is continuous at z_0.*

Proof.

$$\lim_{z \to z_0} f(z) - f(z_0) = \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}(z - z_0) = f'(z_0) \cdot 0 = 0$$

Thus differentiability at z_0 implies that

$$\lim_{z \to z_0} f(z) = f(z_0)$$

which is one formulation of continuity. □

Recall from Section 2.3 that the sum $f + g$, difference $f - g$, product $f \cdot g$, and quotient f/g of two functions $f : S \to \mathbb{C}, g : S \to \mathbb{C}$ are defined in the usual manner:

$$\begin{aligned}
(f + g)(z) &= f(z) + g(z) \quad (z \in S) \\
(f - g)(z) &= f(z) - g(z) \quad (z \in S) \\
(f \cdot g)(z) &= f(z)g(z) \quad (z \in S) \\
(f/g)(z) &= f(z)/g(z) \quad (z \in S, g(z) \neq 0)
\end{aligned}$$

We obtain the expected rules for differentiation of these combinations:

PROPOSITION 4.4. *If f and g are differentiable at z_0, then so are $f + g, f - g, f \cdot g$, and f/g (in the latter case, provided that $g(z_0) \neq 0$). The derivatives are:*

(i) $(f + g)' = f' + g'$
(ii) $(f - g)' = f' - g'$
(iii) $(f \cdot g)' = f \cdot g' + f' \cdot g$
(iv) $(f/g)' = (f'g - fg')/(g^2)$

Proof. The computations are analogous to the real case. As an example, we prove (iii):

$$
\begin{aligned}
(f \cdot g)'(z_0) &= \lim_{z \to z_0} \frac{f(z)g(z) - f(z_0)g(z_0)}{z - z_0} \\
&= \lim_{z \to z_0} \frac{f(z)g(z) - f(z)g(z_0) + f(z)g(z_0) - f(z_0)g(z_0)}{z - z_0} \\
&= \lim_{z \to z_0} f(z) \lim_{z \to z_0} \frac{g(z) - g(z_0)}{z - z_0} + g(z_0) \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}
\end{aligned}
$$

using the algebra of limits, which equals

$$f(z_0)g'(z_0) + f'(z_0)g(z_0)$$

because Proposition 4.3 implies that f is continuous at z_0.

Parts (i, ii, iv) can be proved by similar methods. □

COROLLARY 4.5. (i) *Let*

$$p(z) = a_0 + a_1 z + \cdots + a_n z^n$$

be a polynomial with complex coefficients a_r. *Then its derivative is*

$$p'(z) = a_1 + 2a_2 z + \cdots + n a_n z^{n-1}$$

(ii) *Let* $p(z), q(z)$ *be polynomials over* $\mathbb{C}$. *Then*

$$\frac{\mathrm{d}}{\mathrm{d}z} \frac{p(z)}{q(z)} = \frac{p'(z)q(z) - p(z)q'(z)}{q(z)^2}$$

whenever $q(z) \neq 0$.

Proof. We prove by induction that

$$\frac{\mathrm{d}}{\mathrm{d}z} z^n = n z^{n-1} \tag{4.1}$$

Everything else then follows from Proposition 4.4.

Clearly (4.1) is valid for $n = 0$. When $n = 1$, let $f(z) = z$. Then

$$f'(z_0) = \lim_{z \to z_0} \frac{z - z_0}{z - z_0} = 1$$

and again (4.1) holds. For the induction step, part (ii) of Proposition 4.4 yields:

$$\frac{\mathrm{d}}{\mathrm{d}z} z^{n+1} = \frac{\mathrm{d}}{\mathrm{d}z}(z^n \cdot z) = n z^{n-1} \cdot z + z^n \cdot 1 = (n+1)z^n$$ □

As usual, we denote the composition of $f : S \to \mathbb{C}$ and $g : T \to \mathbb{C}$, where $f(S) \subseteq T$, by $g \circ f$:

$$g \circ f(z) = g(f(z))$$

We then obtain:

PROPOSITION 4.6 (Chain Rule). *If* f *is differentiable at* z_0 *and* g *is differentiable at* $f(z_0)$, *then* $g \circ f$ *is differentiable at* z_0 *and*

$$(g \circ f)'(z_0) = g'(f(z_0))f'(z_0)$$

Before giving a proof, consider a common attempt to prove the chain rule, by writing

$$\frac{g(f(z)) - g(f(z_0))}{z - z_0} = \frac{g(f(z)) - g(f(z_0))}{f(z) - f(z_0)} \frac{f(z) - f(z_0)}{z - z_0} \tag{4.2}$$

provided that $f(z) \neq f(z_0)$. Since f is differentiable at z_0 it is continuous there, so $z \to z_0$ implies $f(z) \to f(z_0)$, giving

$$\lim_{z \to z_0} \frac{g(f(z)) - g(f(z_0))}{f(z) - f(z_0)} = g'(f(z_0))$$

Letting $z \to z_0$ in (3.1) then gives $(g \circ f)'(z_0) = g'(f(z_0))f'(z_0)$ as required.

Unfortunately this attempt has a nasty gap, because $f(z) - f(z_0)$ could be zero. If we knew it was non-zero in some neighbourhood of z_0, excluding $z = z_0$, we could work in this neighbourhood and the proof would then be valid. In fact Theorem 10.16 below shows that in the *complex* case such a neighbourhood always exists – an interesting example of the extra simplicity of complex analysis. In the real case, no such theorem holds.

However, the proof commonly used in the real case to patch up the gap in the above attempt also works in the complex case, and is much more elementary than Theorem 10.16. It goes like this:

Proof. Let $u = f(z_0)$. Define

$$h(w) = \frac{g(w) - g(u)}{w - u} - g'(u) \quad \text{if } w \neq u$$
$$h(u) = 0$$

Clearly h is continuous and defined for all w sufficiently near u. Also, as $z \to z_0$ we see that $h \circ f(z) \to h(f(z)) = h(u) = 0$. Now, the definition of h, applied when $w = f(z)$, can be written in the shape

$$g(f(z)) - g(u) = (h(f(z)) + g'(u))(f(z) - u)$$

when $f(z) \neq u$. But *it is also obviously true when* $f(z) = u$. Let $z \neq z_0$, divide by both sides by $z - z_0$, let $z \to z_0$ . . . Voilà! □

4.2 The Cauchy–Riemann Equations

Suppose that we write a complex function f in terms of two real functions u, v of two real variables

$$f(z) = u(x, y) + \mathrm{i}v(x, y)$$

where $x + \mathrm{i}y = z$. We now prove that differentiability of f imposes two conditions on the partial derivatives of u and v with respect to x and y. We use the notation

$$\frac{\partial u}{\partial x}(x, y) = \lim_{h \to 0} \frac{u(x + h, y) - u(x, y)}{h}$$
$$\frac{\partial u}{\partial y}(x, y) = \lim_{k \to 0} \frac{u(x, y + k) - u(x, y)}{k}$$

and abbreviate these to $\partial u/\partial x$ and $\partial u/\partial y$ when no confusion can occur. We can then prove:

THEOREM 4.7 (Cauchy–Riemann Equations). *If f is differentiable at $z = x + \mathrm{i}y$ then $\partial u/\partial x, \partial u/\partial y, \partial v/\partial x, \partial v/\partial y$ all exist at (x, y), and*

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \qquad \frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y} \tag{4.3}$$

Proof. We calculate $f'(z)$ in two different ways. First we take a point near z in the form $z + h = (x + h) + \mathrm{i}y$, where h is real, and compute

$$\begin{aligned} f'(z) &= \lim_{h\to 0} \frac{f(z+h) - f(z)}{h} \\ &= \lim_{h\to 0} \frac{u(x+h, y) + \mathrm{i}v(x+h, y) - u(x, y) - \mathrm{i}v(x, y)}{h} \\ &= \lim_{h\to 0} \frac{u(x+h, y) - u(x, y)}{h} + \mathrm{i} \lim_{h\to 0} \frac{v(x+h, y) - v(x, y)}{h} \\ &= \frac{\partial u}{\partial x} + \mathrm{i}\frac{\partial v}{\partial x} \end{aligned}$$

Next we take a point near z in the form $z = x + \mathrm{i}(y + k)$, where k is real, and compute

$$\begin{aligned} f'(z) &= \lim_{k\to 0} \frac{u(x, y+k) + \mathrm{i}v(x, y+k) - u(x, y) - \mathrm{i}v(x, y)}{h} \\ &= \frac{\partial v}{\partial y} - \mathrm{i}\frac{\partial u}{\partial y} \end{aligned}$$

Now equate real and imaginary parts in the two expressions. □

The equations (4.3) are called the *Cauchy–Riemann Equations* after Cauchy (1789–1852) and Riemann (1826–1866). However, they were known to d'Alembert in 1752. Observe that one equation has a minus sign but the other does not. This happens because $\mathrm{i}^2 = -1$ and it is an important feature of the Cauchy–Riemann Equations. We can check this using the simplest nonlinear complex function:

Example 4.8. Suppose that $f(z) = z^2 = (x^2 - y^2) + 2\mathrm{i}xy$, so that

$$u(x, y) = x^2 - y^2 \quad v(x, y) = 2xy$$

Then

$$\begin{aligned} \frac{\partial u}{\partial x} &= 2x \\ \frac{\partial u}{\partial y} &= -2y \\ \frac{\partial v}{\partial x} &= 2y \\ \frac{\partial v}{\partial y} &= 2x \end{aligned}$$

and the Cauchy–Riemann Equations are valid. □

Example 4.9. The function $f(z) = \mathrm{re}\, z$ is continuous everywhere but differentiable nowhere. In this case $u(x,y) = x, v(x,y) = 0$. So $\partial u/\partial x = 1, \partial v/\partial y = 0$, which are never equal. Analogous real functions are much harder to find, see Section 4.6. □

Example 4.10. The converse of Theorem 4.7 is false. Let

$$f(x+\mathrm{i}y) = \begin{cases} 0 & \text{if one or both of } x, y \text{ is zero} \\ 1 & \text{if neither of } x, y \text{ is zero} \end{cases}$$

For this function, the partial derivatives of u and v all exist at the origin and are zero, so the Cauchy–Riemann Equations are certainly satisfied. But f is not continuous at the origin, so it cannot be differentiable there. □

Once again the introduction of real analysis creates complications. However, in this case it is relatively easy to patch things up. It is rather like starting a well-tuned vintage car on an icy morning. Once it is going, it will run smoothly, but getting it started can take hard work. Complex analysis will prove to be a well-oiled machine, but taking real analysis as a point of departure requires paying careful attention to the starting conditions. If we do that, the machine runs well. But to verify this claim, we have to take a very close look at the mechanics. We begin with a technical lemma.

LEMMA 4.11. *If $\partial u/\partial x$ and $\partial u/\partial y$ exist at (x,y) and $\partial u/\partial x$ is continuous there, then*

$$u(x+h,y+k) - u(x,y) = h\left(\frac{\partial u}{\partial x}(x,y) + \varepsilon(h,k)\right) + k\left(\frac{\partial u}{\partial y}(x,y) + \eta(h,k)\right)$$

where $\varepsilon(h,k), \eta(h,k) \to 0$ as $h,k \to 0$.

Proof. Write $u(x+h,y+k) - u(x,y)$ as

$$u(x+h,y+k) - u(x,y+k) + u(x,y+k) - u(x,y)$$

By the Mean Value Theorem for one real variable applied to $\phi(t) = u(x+t,y+k)$, there exists θ with $0 < \theta < 1$ such that

$$u(x+h,y+k) - u(x,y+k) = h\frac{\partial u}{\partial x}(x+\theta h, y+k)$$

By continuity of $\partial u/\partial x$,

$$\frac{\partial u}{\partial x}(x+\theta h, y+k) - \frac{\partial u}{\partial x}(x,y) = \varepsilon(h,k)$$

where $\varepsilon(h,k) \to 0$ as $h,k \to 0$. Hence

$$u(x+h,y+k) - u(x,y+k) = h\left(\frac{\partial u}{\partial x}(x,y) + \varepsilon(h,k)\right) \tag{4.4}$$

More simply, $k \to 0$ implies

$$\frac{u(x,y+k) - u(x,y)}{k} \to \frac{\partial u}{\partial y}(x,y)$$

Therefore if

$$\eta(h,k) = \frac{u(x,y+k) - u(x,y)}{k} - \frac{\partial u}{\partial y}(x,y)$$

then

$$u(x, y+k) - u(x, y) = k\left(\frac{\partial u}{\partial y}(x, y) + \eta(h, k)\right) \tag{4.5}$$

where $\eta(h, k) \to 0$ as $h, k \to 0$.

Adding (4.4) and (4.5) gives the required result. □

By imposing extra continuity conditions on the partial derivatives of u and v we can prove a converse to Theorem 4.7:

THEOREM 4.12. *If $f(z) = u(x, y) + \mathrm{i}v(x, y)$ is a complex function defined on an open set S, and at some point $z_0 = x_0 + \mathrm{i}y_0$ the partial derivatives $\partial u/\partial x, \partial u/\partial y, \partial v/\partial x, \partial v/\partial y$ all exist, are continuous, and satisfy*

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \qquad \frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}$$

then f is differentiable at z_0.

Proof. Using Lemma 4.11, we can write

$$\begin{aligned} f(z) - f(z_0) &= u(x_0 + h, y_0 + k) + \mathrm{i}v(x_0 + h, y_0 + k) - u(x_0, y_0) - \mathrm{i}v(x_0, y_0) \\ &= h\left(\frac{\partial u}{\partial x} + \varepsilon_1\right) + k\left(\frac{\partial u}{\partial y} + \eta_1\right) + \mathrm{i}h\left(\frac{\partial v}{\partial x} + \varepsilon_2\right) + \mathrm{i}k\left(\frac{\partial v}{\partial y} + \eta_2\right) \end{aligned}$$

where $\varepsilon_1, \varepsilon_2, \eta_1, \eta_2 \to 0$ as $h, k \to 0$.

By the Cauchy–Riemann Equations,

$$\begin{aligned} f(z) - f(z_0) &= (h + \mathrm{i}k)\left(\frac{\partial u}{\partial x} + \mathrm{i}\frac{\partial v}{\partial x}\right) + h\varepsilon_1 + k\eta_1 + h\varepsilon_2 + k\eta_2 \\ &= (z - z_0)\left(\frac{\partial u}{\partial x} + \mathrm{i}\frac{\partial v}{\partial x}\right) + \rho \end{aligned}$$

where $\rho = h\varepsilon_1 + k\eta_1 + h\varepsilon_2 + k\eta_2$, so

$$\frac{f(z) - f(z_0)}{z - z_0} = \frac{\partial u}{\partial x} + \mathrm{i}\frac{\partial v}{\partial x} + \frac{\rho}{z - z_0}$$

when $z \neq z_0$. But

$$\begin{aligned} \left|\frac{\rho}{z - z_0}\right| &= \frac{|\rho|}{\sqrt{h^2 + k^2}} \leq \frac{|h||\varepsilon_1| + |k||\eta_1| + |h||\varepsilon_2| + |k||\eta_2|}{\sqrt{h^2 + k^2}} \\ &\leq |\varepsilon_1| + |\eta_1| + |\varepsilon_2| + |\eta_2| \end{aligned}$$

Let $h, k \to 0$. Then $|\rho|/(z - z_0)| \to 0$, so

$$\lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0} = \frac{\partial u}{\partial x} + \mathrm{i}\frac{\partial v}{\partial x}$$

as required. □

Example 4.13. The function $f(z) = |z|^2$ is differentiable at the origin and nowhere else, since

$$u(x, y) = x^2 + y^2 \qquad v(x, y) = 0$$

Hence $\partial u/\partial x = 2x, \partial u/\partial y = 2y, \partial v/\partial x = 0, \partial v/\partial y = 0$, and the Cauchy–Riemann Equations are satisfied only at $x = y = 0$ at which point the partial derivatives are all continuous. □

4.3 Connected Sets and Differentiability

If $f(z) =$ constant, then $f'(z) = 0$, but what of the converse? When the derivative is zero, does this imply that the function is constant? The answer is obviously 'no' for a function defined on a disconnected set, But it is 'yes' when f is defined on a *connected* set. Recall that a domain is defined to be a connected open set. We now prove:

THEOREM 4.14. *If f is differentiable in a domain D and $f'(z) = 0$ for all $x \in D$, then f is constant on D.*

Proof. By the proof of Theorem 4.7, $f'(z) = \partial u/\partial x + \mathrm{i}\partial v/\partial x = \partial v/\partial y = 0 - \mathrm{i}\partial u/\partial y$. So $f'(z) = 0$ implies that all four partial derivatives are zero.

From real analysis, if $\phi' = 0$ on a closed interval $[a, b]$, then ϕ is constant on $[a, b]$.

If $L = \{t + iy_0 : t \in [a, b]\}$ is a line segment in D, let $\phi(t) = u(t, y_0)$. Then $\partial u/\partial x = \phi' = 0$ so u is constant on L. By a similar argument, u, v are both constant on any horizontal or vertical line segment in D. Hence u, v are both constant on any step path in D. But any two points in D can be connected by a step path, so f is constant on D. □

The same technique proves:

PROPOSITION 4.15. *If f is differentiable in a domain D and any one of* $\mathrm{re}\, f, \mathrm{im} f$, *or $|f|$ is constant, then f is constant.*

Proof. If $f = u + \mathrm{i}v$ and $\mathrm{re} f = u$ is constant, then $\partial u/\partial x = \partial u/\partial y = 0$. The Cauchy–Riemann Equations then give $\partial v/\partial x = \partial v/\partial y = 0$, and by the argument of the previous proof, f is constant on D. The case when $\mathrm{im} f$ is constant is similar.

If $|f|$ is constant then $u^2 + v^2 = c$ for some $c \in \mathbb{C}$. If $c = 0$ then $f = 0$, so we may assume $c \neq 0$. Differentiating,

$$2u\frac{\partial u}{\partial x} + 2v\frac{\partial v}{\partial x} = 0$$
$$2u\frac{\partial u}{\partial y} + 2v\frac{\partial v}{\partial y} = 0$$

and the Cauchy–Riemann Equations give

$$u\frac{\partial u}{\partial x} - v\frac{\partial u}{\partial y} = 0$$
$$u\frac{\partial u}{\partial y} + v\frac{\partial u}{\partial x} = 0$$

Add u times the first equation to v times the second, to get

$$(u^2 + v^2)\frac{\partial u}{\partial x} = 0$$

Since $u^2 + v^2 = c \neq 0$ we have $\partial u/\partial x = 0$. Similarly, the other partial derivatives of u, v are zero, so u, v are real constants, and f is constant on D. □

4.4 Hybrid Functions

At this point we briefly consider hybrid functions, by which we mean real-valued functions of a complex variable or complex-valued functions of a real variable. There are evident notions of differentiation in both cases. For instance, a real-valued function of a complex variable $f : D \to \mathbb{R}$ where D is an open subset of $\mathbb{C}$ may be regarded merely as a complex function with imaginary part zero. Such a hybrid function is a very dull fellow, for if it is differentiable, its constant imaginary part implies that it must be constant, by Proposition 4.15.

We fare a little better with complex-valued functions of a real variable. The most interesting case is $f : [a, b] \to \mathbb{C}$, which (when continuous) is a path in the complex plane. Define the derivative at $t_0 \in [a, b]$ to be

$$\lim_{t \to t_0} \frac{f(t) - f(t_0)}{t - t_0}$$

in the obvious way (and allowing appropriate one-sided limits at the end points a and b) we obtain the expected generalisations of the usual properties of the derivative:

PROPOSITION 4.16. *If $f : [a, b] \to \mathbb{C}$, $g : [a, b] \to \mathbb{C}$ are differentiable at $t \in [a, b]$, then*

$$\begin{aligned}(f \pm g)'(t) &= f'(t) \pm g'(t)\\ (f \cdot g)'(t) &= f(t)g'(t) + f'(t)g(t)\\ (f/g)'(t) &= (f'(t)g(t) - f(t)g'(t))/(g(t))^2\end{aligned}$$

□

The chain rule involving a complex function f of a real variable appears in two guises. We may either precede it by a real function h, or follow it by a complex function g, to obtain:

PROPOSITION 4.17. *If $h : [a, b] \to [c, d], f : [a, b] \to D$, and $g : D \to \mathbb{C}$, then*

$$\begin{aligned}(f \circ h)'(s) &= f'(h(s))h'(s)\\ (g \circ f)'(t) &= g'(f(t))f'(t)\end{aligned}$$

wherever the derivatives on the right-hand side of the equations are defined. □

The proofs of Propositions 4.16 and 4.17 follow the same pattern as the real and complex cases.

We end this section by characterising the geometric meaning of the derivative γ' of a smooth path γ when γ' is non-zero.

PROPOSITION 4.18. *If $\gamma : [a, b] \to \mathbb{C}$ is a path and the derivative $\gamma'(t)$ exists and is non-zero for some $t \in [a, b]$ (including the possibility of a one-sided derivative at either end point) then the tangent to γ exists at $\gamma(t)$ and a point on the tangent is of the form $\gamma(t) + h\gamma'(t)$ for any $h \in \mathbb{R}$. (Figure 4.1.)*

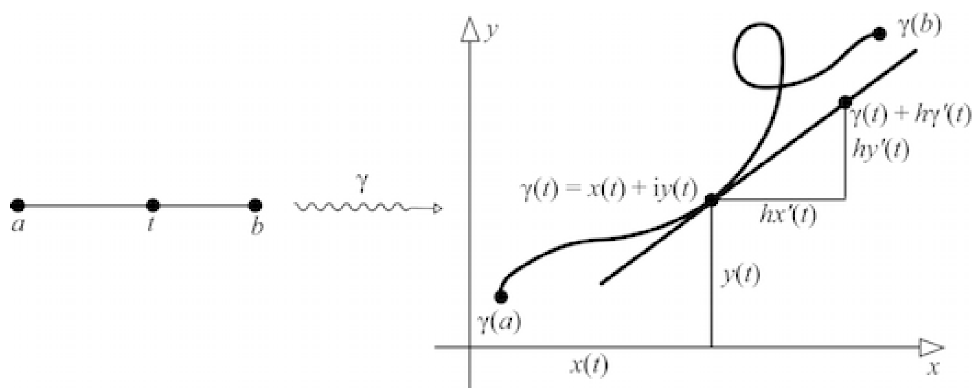

Figure 4.1 The tangent to a path $\gamma : [a,b] \to \mathbb{C}$.

Proof. Let $\gamma(t) = x(t) + iy(t)$, where x and y are real functions. Then $\gamma'(t) = x'(t) + iy'(\gamma(t))$. The parametric function $f : [a,b] \to \mathbb{R}^2$ given by $f(t) = (x(t), y(t))$ has a graph whose tangent is $(x(t) + hx'(t), y(t) + hy'(t))$ for any $h \in \mathbb{R}$. This is the complex number $x(t) + hx'(t) + i(y(t) + hy'(t)) = x(t) + hx'(t) + i(y(t) + hy'(t)) = \gamma(t) + h\gamma'(t)$. □

This proposition will be of value when we seek to visualise paths that are differentiable, particularly for complex integration in Chapter 6 and in later applications.

4.5 Power Series

More exciting creatures are power series, for they will later prove to be the foundation for all differentiable complex functions.

By Corollary 4.5, the derivative of a polynomial

$$p(z) = a_0 + a_1 z + \cdots + a_n z^n$$

is the polynomial

$$p'(z) = a_1 + 2a_2 z + \cdots + na_n z^{n-1}$$

This suggests that if f is a power series

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

we should have

$$f'(z) = \sum_{n=0}^{\infty} na_n z^{n-1}$$

If this is true, we say that f *can be differentiated term by term*. When is this the case? Certainly $f(z)$ must be convergent. The next two results show that this is almost sufficient: in fact, term-by-term differentiation is always valid within the disc of convergence of $f(z)$.

LEMMA 4.19. *Let* $f(z) = \sum a_n z^n$ *converge absolutely for* $|z| < R$. *Then* $g(z) = \sum n a_n z^{n-1}$ *converges for* $|z| < R$.

Proof. For $|z| < R$, choose r such that $|z| < r < R$. Then, by Theorem 3.25, $\sum a_n r^n$ converges absolutely and there exists $K \in \mathbb{R}$ such that $|a_n r^n| < K$ for all n. Now $q = |z|/r$ is less than 1, so

$$\begin{aligned} |n a_n z^{n-1}| &= n|a_n||z/r|^{n-1} r^{n-1} \\ &< \frac{nK}{r} q^{n-1} \end{aligned}$$

But for $0 \leq q < 1$ the real series $\sum nKq^{n-1}$ converges to $K(1-q)^{-2}$. By the comparison test, $\sum |n a_n z^{n-1}|$ converges, so by Theorem 3.14 the series $\sum n a_n z^{n-1}$ converges. □

THEOREM 4.20. *A power series* $f(z) = \sum a_n z^n$ *may be differentiated term by term within its disc of convergence. That is,*

$$f'(z) = \sum n a_n z^{n-1}$$

Proof. By Lemma 4.19, $g(z) = \sum n a_n z^{n-1}$ is absolutely convergent for $|z| < R$. We must show that for $|z_0| < R$,

$$f'(z_0) = \lim_{z \to z_0} \left(\frac{f(z) - f(z_0)}{z - z_0} \right) = g(z_0)$$

or equivalently that

$$\lim_{z \to z_0} \left(\frac{f(z) - f(z_0)}{z - z_0} - g(z_0) \right) = 0$$

Taking our courage in both hands we compute

$$\frac{f(z) - f(z_0)}{z - z_0} - g(z_0) = \sum_{n=1}^{\infty} \left(a_n \frac{z^n - z_0^n}{z - z_0} - n a_n z^{n-1} \right)$$

since power series may be added and subtracted term by term, by Lemma 3.12. But the above expression can be written as

$$\begin{aligned} &\sum_{n=1}^{\infty} a_n [z^{n-1} + z_0 z^{n-2} + \cdots + z_0^{n-1} - n z_0^{n-1}] \\ &= \sum_{n=1}^{N} a_n [z^{n-1} + z_0 z^{n-2} + \cdots + z_0^{n-1} - n z_0^{n-1}] \\ &\quad + \sum_{n=N+1}^{\infty} a_n [z^{n-1} + z_0 z^{n-2} + \cdots + z_0^{n-1} - n z_0^{n-1}] \\ &= \Sigma_1 + \Sigma_2, \text{say} \end{aligned}$$

Given any $\varepsilon > 0$, we first choose any r such that $|z_0| < r < R$. Then $\sum na_n r^{n-1}$ is convergent and, as in Lemma 3.6, there exists $N = N(\varepsilon)$ such that

$$\sum_{n=N+1}^{\infty} |na_n r^{n-1}| < \varepsilon/4$$

Since $|z_0| < r$, if z is close enough to z_0 to ensure that $|z| < r$ also, then

$$|\Sigma_2| \leq \sum_{n=N+1}^{\infty} 2n|a_n|r^{n-1} < \varepsilon/2 \tag{4.6}$$

Furthermore, Σ_1 is a polynomial in z so $\Sigma_1 \to 0$ as $z \to z_0$. We can therefore find $\delta > 0$ such that

$$|z - z_0| < \delta \text{ implies } |\Sigma_1| < \varepsilon/2 \tag{4.7}$$

We now choose z close enough to z_0 to ensure that both (4.6) and (4.7) hold, and then

$$\left|\frac{f(z) - f(z_0)}{z - z_0} - g(z_0)\right| = |\Sigma_1 + \Sigma_2| \leq |\Sigma_1| + |\Sigma_2| = \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

Therefore $f'(z_0) = g(z_0)$ as claimed. □

Theorem 4.20 is immensely important, for it tells us not only about the first derivative of a power series, but about all higher derivatives as well. We just repeat it as often as we want to get:

COROLLARY 4.21. *All the higher derivatives $f', f'', f''', \ldots, f^{(k)}, \ldots$ of a power series $f(z) = \sum a_n z^n$ exists for z within the disc of convergence, and*

$$\begin{aligned} f^{(k)}(z) &= \sum_{n=k}^{\infty} n(n-1)\cdots(n-k+1)a_n z^{n-k} \\ &= \sum_{n=k}^{\infty} \frac{n!}{(n-k)!} a_n z^{n-k} \end{aligned}$$

Proof. Use induction on k. □

Replacing z by $z - z_0$ in Corollary 4.21, we find that if a power series $f(z) = \sum a_n(z - z_0)^n$ has disc of convergence $|z - z_0| < R$, then inside this disc of convergence, all the higher derivatives of f exist, and

$$f^{(k)}(z) = \sum_{n=k}^{\infty} \frac{n!}{(n-k)!} a_n (z - z_0)^{n-k}$$

Putting $z = z_0$ in this series we get

$$f^{(k)}(z_0) = k! a_k$$

which gives yet another important corollary:

COROLLARY 4.22 (Taylor's Theorem). *If* $f(z) = \sum a_n(z-z_0)^n$ *for* $|z-z_0| < R$, *then*

$$a_k = \frac{f^{(k)}(z_0)}{k!}$$ □

We can thus express $f(z)$ as a convergent *Taylor series*

$$f(z) = \sum_{n=0}^{\infty} \frac{f^{(n)}(z_0)}{n!} z^n \quad (|z-z_0| < R)$$

Example 4.23. Suppose that $f(z) = 1/(1-z) = 1 + z + z^2 + \cdots$ for $|z| < 1$. Differentiating, we find that

$$f'(z) = \frac{1}{(1-z)^2} = 1 + 2z + 3z^2 + \cdots$$
$$f''(z) = \frac{2}{(1-z)^3} = 2 + 6z + 12z^2 + \cdots$$

and so on. We also have $f^{(n)}(0) = n!$ and $f(z) = \sum f^{(n)}(0)z^n/n$ as expected. □

4.6 A Glimpse Into the Future

We now set the scene for an area where complex analysis is much better behaved than real analysis: differentiability. A real function may be differentiable but have a non-differentiable derivative. By repeated integration and the Fundamental Theorem of Calculus, it follows that for any n there exist real functions that are differentiable n times but not $n+1$ times. There also exist continuous real functions that are differentiable nowhere, such as the 'blancmange function' described later in this section.

Complex functions are just as flexible as real ones when it comes to continuity, but the condition of differentiability is much more stringent in the complex case. Constructing a continuous complex function that is differentiable nowhere is trivial: a simple example is the real part $f(z) = \operatorname{re} z$, as we proved in Example 4.9. In compensation, we prove in Chapter 10 that if a complex function is differentiable in a domain D, then it is differentiable any number of times in D, and it even has a convergent power series expansion near any point in D.

4.6.1 Real Functions Differentiable Only Finitely Many Times

In the real case, for any $n \in \mathbb{N}$, there exist functions that are differentiable n times but not $n+1$ times. A simple example when $n = 1$ is

$$\phi(x) = \begin{cases} 0 & (x \leq 0) \\ x^2 & (x > 0) \end{cases}$$

Trivially,

$$\phi'(x) = \begin{cases} 0 & (x \le 0) \\ 2x & (x > 0) \end{cases}$$

and an easy calculation gives

$$\phi'(0) = \lim_{x \to 0} \frac{\phi(x) - \phi(0)}{x} = 0$$

Thus ϕ' exists, and is even continuous at 0. But $\phi''(0)$ does not exist because

$$\frac{\phi'(x) - \phi'(0)}{x} = \begin{cases} 0 & (x \le 0) \\ 2 & (x > 0) \end{cases}$$

More generally, the function

$$\phi(x) = \begin{cases} 0 & (x \le 0) \\ x^{n+1} & (x > 0) \end{cases}$$

is differentiable n times everywhere, but not $n + 1$ times at the origin.

We shall see later that there is no similar method for piecing together complex functions to obtain this type of behaviour. For a real function, there are only two ways to approach a limit point x_0: from the left (smaller values of x) and from the right (larger values of x). We can piece real functions together quite happily, provided we make sure that the left and right derivatives are the same to whatever extent we deem necessary.

4.6.2 Bad Behaviour of Real Taylor Series

Indeed, we can go further. Above, we pieced together 0 on the left and x^{n+1} on the right to obtain a function that can be differentiated n times but not $n + 1$ times. But we can even find functions like

$$\phi(x) = \begin{cases} 0 & (x \le 0) \\ e^{-1/x} & (x > 0) \end{cases} \tag{4.8}$$

which has *all* derivatives in agreement at the origin, when approached either from left or right. This ‘Frankenstein’s monster’ creation is patched up well, but there is something unnatural about it. Because all derivatives are zero at the origin, its Taylor series about the origin is

$$\sum_{n=0}^{\infty} \frac{\phi^{(n)}(0)}{n!} = 0 + 0x + 0x^2 + \cdots + 0x^n + \cdots = 0$$

This converges for all real x, but its sum is *not equal to* $\phi(x)$.

Thus in the real case we can find infinitely differentiable functions that are not equal to their Taylor series. Despite our reference to Frankenstein, there is nothing mysterious about this. It just means that the remainder term $R_n(x)$ in the Taylor series

$$\phi(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n + R_n(x)$$

does not tend to zero as n tends to ∞. Indeed, here $R_n(x) = \phi(x)$ for all n. (In the theory of differentiable manifolds, the ability to patch real functions together smoothly like this is actually very useful. It provides topological flexibility. In this context, the complex case really is more complicated.)

4.6.3 The Blancmange function

Real analysis has even more grey areas, inhabited by functions that are continuous everywhere yet differentiable nowhere. Bernard Bolzano found such a function around 1831, but it was not published until 1922. Charles Cellérier discovered one around 1860, which was published in 1890, shortly after his death. The first such function to be published was constructed by Weierstrass in 1872, as the sum

$$f(x) = \sum_{n=0}^{\infty} a^n \cos(b^n \pi x)$$

where $0 < a < 1, b$ is a positive odd integer, and $ab > 1 + 3\pi/2$. The smallest value of b satisfying these conditions is $b = 7$. Weierstrass's function is continuous by uniform continuity of the terms in the sum. Term-by-term differentiation leads to a series that does not converge, and Weierstrass gave a rigorous proof that the function is nowhere differentiable. Later Godfrey Harold Hardy improved the conditions to $0 < a < 1, ab > 1$. Cellérier's example is very similar to Weierstrass's. Figure 4.2 shows how irregular this function is.

We describe a simpler example that goes by the name of the *blancmange function*. Let $G(x)$ be the distance from $x \in \mathbb{R}$ to the nearest integer. This has a graph like the teeth of a saw:

$$G(x) = \begin{cases} x & (0 \le x \le \frac{1}{2}) \\ 1 - x & (\frac{1}{2} \le x \le 1) \end{cases}$$

and G is periodic with $G(x + n) = G(x)$ for any integer n. It is not differentiable at $x = \frac{1}{2}n$ for any integer n. Moreover, $|G(x) \le \frac{1}{2}|$ for all $x \in \mathbb{R}$. It follows that the function $G_n(x) = (\frac{1}{4})^n G(4^n x)$ is not differentiable at $x = \frac{1}{2}(\frac{1}{4})^m$ for any integer m, and satisfies $0 \le G_n(x) \le \frac{1}{2}(\frac{1}{4})^n$.

The sum

$$b(x) = \sum_{n=0}^{\infty} G_n(x)$$

is sketched in Figure 4.3, which shows why it is called the blancmange function.

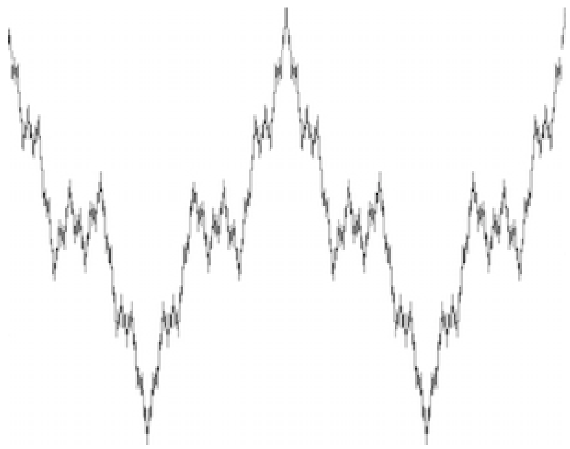

Figure 4.2 Weierstrass's continuous nowhere differentiable function.

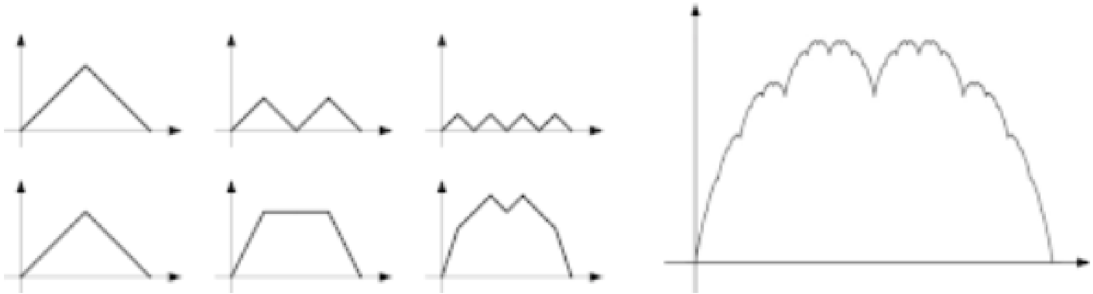

Figure 4.3 *Top left*: Successive halvings of the sawtooth function. *Bottom left*: Adding the sawtooth functions gives successive stages of the blancmange function. *Right*: An advanced stage of the construction. We show only the unit interval; the full function is periodic with period 1 and defined on the whole of $\mathbb{R}$.

Intuitively, the curve is a fractal, a set with detailed structure on all scales of magnification, Mandelbrot [13] and Falconer [4]. This is clear because the graph of the blancmange function has smaller blancmanges everywhere. Some are clearly recognisable, while others are disguised by being sheared vertically as they stand on straight line segments that lie at an angle. In general, there are two kinds of points on the graph: those at dyadic rationals $x = k2^{-n}$ for integers k, n with $n \geq 0$, where the graph comes down vertically from the left and goes up vertically to the right. Here, under high magnification the graph looks like a half-line pointing upwards. Everywhere else, the graph is locally a tiny blancmange, sheared on a small line segment. This is easy to see when the segment is not too steep, but not when it is nearly vertical.

A rigorous proof requires extra effort, and we sketch one now:

PROPOSITION 4.24. *The blancmange function is continuous everywhere and differentiable nowhere.*

Proof. The estimate

$$0 \leq b(x) \leq \sum_{n=0}^{\infty} \tfrac{1}{2}(\tfrac{1}{4})^n = \frac{\frac{1}{2}}{1-\frac{1}{4}} = \tfrac{2}{3}$$

holds, so by uniform continuity, as in Section 2.9 on space-filling curves, b is continuous everywhere.

To prove it is differentiable nowhere, observe that in order for b to be differentiable at α, the expression $\gamma(x) = (b(x) - b(\alpha))/(x - \alpha)$ must tend to a limit. We construct a sequence $(\alpha_n) \to \alpha$ such that $\gamma(\alpha_n)$ does not converge.

Under each line segment of G_n there are two identical teeth of G_{n+1} of length $(\frac{1}{4})^n$, so we can find $\alpha_n = \alpha \pm (\frac{1}{4})^n$ such that

$$G_m(\alpha_n) = G_m(\alpha) \quad (m \geq n+1)$$

The gradient of the straight bit of the tooth of G_n and larger teeth is

$$\frac{G_m(\alpha_n) - G_m(\alpha)}{\alpha_n - \alpha} = 1 \quad (m \leq n)$$

Hence

$$\gamma(\alpha_n) = \frac{b(\alpha_n) - b(\alpha)}{\alpha_n - \alpha} \leq \sum_{n=0}^{\infty} \frac{G_m(\alpha_n) - G_m(\alpha)}{\alpha_n - \alpha}$$

is a sum of $n + 1$ terms, each ± 1. Now $\gamma(\alpha_n)$ is an odd integer when n is even and an even integer when n is odd. Therefore $(\gamma(\alpha_n))$ cannot tend to a limit. □

Having created the bad function b, its antiderivative

$$b_1(x) = \int_0^x b(t)\mathrm{d}t$$

is differentiable once everywhere (with derivative b) but is twice differentiable nowhere. Repeating this process inductively we obtain a function b_n that is differentiable n times everywhere (with derivative b_{n-1}) but is $(n + 1)$ times differentiable nowhere.

Looking back at the idea of a space-filling curve, we begin to see why our intuitive ideas of continuity and differentiability need modifying to fit with the formal epsilon-delta definitions of continuity and limits. The space-filling curve is constructed as a formal limit of curves made up of straight line segments, and in the construction we take successively smaller line segments that change direction. In the case of the blancmange function, successive curves change direction at dyadic rationals of the form $k2^{-n}$ where k, n are integers and $n \geq 0$. The limit function is continuous everywhere and differentiable nowhere, with the slope at the dyadic rationals being vertically down to the left and vertically up to the right. The space-filling path constructed here is the limit of graphs whose real and imaginary parts also turn suddenly at dyadic rationals (Figure 2.26) yet the distance travelled in each step gets smaller and smaller (Figure 2.27). This is how it manages to go through every point in the unit square.

Our strategy to deal with this problem is simple. When we calculate integrals in complex analysis, we use ‘contours’ made up of a finite number of ‘smooth paths’, which have non-zero continuous derivatives to ensure that the formal approach resonates with our intuitive idea of tracing the contour in real time using a finger.

4.6.4 Complex Analysis is Better Behaved

Real analysis is a very hairy subject indeed. But what is the relevance of such bizarre functions in complex analysis? The answer is: NONE WHATSOEVER. They have been mentioned only to contrast the real case with the complex case.

There do, however, exist very simple complex functions that are continuous but differentiable nowhere. An example is $f(z) = |z|$, whose continuity is easily proved, but fails to satisfy the Cauchy–Riemann Equations at any point. There are also functions, such as $f(z) = |z|^2$, that are differentiable only at isolated points (here, the origin), and are thus not differentiable twice at those points. But that is the end of the line. If a complex function is differentiable once in a domain, then (as we prove in Chapter 10) it is differentiable any number of times, has a convergent Taylor series, and is equal to its Taylor series. The reason for this much nicer behaviour is that in computing the derivative

$$f'(z) = \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}$$

z can approach z_0 from any direction in the complex plane. The existence of the limit is therefore a much stronger condition than it is in the real case, and precludes patching together different functions at z_0.

The key to the whole theory of complex analysis is the fact that every differentiable complex function has a power series expansion near any point of its domain, so is equal to its own Taylor series as in Corollary 4.22. This will be established by a roundabout route in Chapter 10, but it is worth waiting for, and it underlines our emphasis on power series. They are not just good examples of differentiable functions: in a very genuine sense they are the *only* examples.

4.7 Exercises

1. From first principles, differentiate:
(i) $f(z) = z^2 + 2z$
(ii) $f(z) = 1/z\,(z \neq 0)$
(iii) $f(z) = z^3 + z^2$

2. Show that $f(z) = |z|$ is continuous everywhere and differentiable nowhere. Show that $f(z) = |z|^2$ is continuous everywhere and differentiable at the origin but nowhere else.

3. Differentiate:
(i) $3z^2 + 2z^3$
(ii) $(z^2 + 3\mathrm{i}z - 4)(z^3 - 7\mathrm{i})$
(iii) $(z^5 - 13\mathrm{i}z^2)^{99}$
(iv) $(z^3 + 3)^5(z^4 + (26 - 11\mathrm{i})z)^{12}$
(v) $(z^2 + 3)/(4z^3 + 5 - 3\mathrm{i})^2$
(vi) $\left(\dfrac{z+n}{z-n}\right)^n$, $n \in \mathbb{N}, z \neq n$

4. Let

$$f_n(z) = \left(1 + \frac{z}{n}\right)^n$$

Show that

$$f_n'(z) = f_{n-1}\left(\frac{(n-1)z}{n}\right)$$

What do you notice as $n \to \infty$?

5. Let $\mathbb{C}_\pi = \{z \in \mathbb{C} : z \neq x \in \mathbb{R}, x \leq 0$ be the 'cut plane' with the negative real axis removed. Prove that $\mathbb{C}_\pi$ is a domain. Define $r : \mathbb{C}_\pi \to \mathbb{C}$ by $(r(z))^2 = z$ and $\mathrm{re}\, r(z) > 0$. Prove that r is continuous on $\mathbb{C}_\pi$, and hence show from first principles that $r'(z) = 1/(2r(z))$.

6. Let $f(z)$ be a polynomial in $z \in \mathbb{C}$. Prove that the function $g(z) = \overline{f(\bar{z})}$ is differentiable everywhere, but that $h(z) = \overline{f(z)}$ is differentiable at 0 if and only if $f'(0) = 0$.

7. In each of the following cases, for f defined on the domain D, find explicit formulas for $u(x, y), v(x, y)$ where $f(z) = u(x, y) + \mathrm{i}v(x, y)$, where $z = x + \mathrm{i}y$ and all of x, y, u, v are real.
   (i) $f(z) = 1/z,\ D = \{z \in \mathbb{C} : z \neq 0\}$
   (ii) $f(z) = |z|,\ D = \mathbb{C}$
   (iii) $f(z) = \bar{z},\ D = \mathbb{C}$
   Show that u, v satisfy the Cauchy–Riemann Equations everywhere in (i) and nowhere in (ii), (iii).
8. Verify the Cauchy–Riemann Equations for the functions $u(x, y), v(x, y)$ defined in the given domains by
   (i) $u(x, y) = x^3 - 3xy^2, v(x, y) = 3x^2y - y^3\ (x, y \in \mathbb{R})$
   (ii) $u(x, y) = \sin x \cosh y, v(x, y) = \cos x \sinh y\ (x, y \in \mathbb{R})$
   (iii) $u(x, y) = x/(x^2 + y^2), v(x, y) = -y/(x^2 + y^2),\ (x^2 + y^2 \neq 0)$
   (iv) $u(x, y) = \frac{1}{2}\log(x^2 + y^2), v(x, y) = \sin^{-1}(y/\sqrt{x^2 + y^2})\ (x > 0)$
   In each case, state a complex function whose real and imaginary parts are $u(x, y)$ and $v(x, y)$.
9. For $z = x + \mathrm{i}y$, let

$$f(z) = \frac{x^3(1 + \mathrm{i}) - y^3(1 - \mathrm{i})}{x^2 + y^2} \quad (z \neq 0)$$
$$f(0) = 0$$

   Show that f is continuous at the origin, the Cauchy–Riemann Equations are satisfied there, yet $f'(0)$ does not exist. Why does this not contradict Theorem 4.12?
10. Let $f(z) = \sqrt{|xy|}$ for $z = x + \mathrm{i}y$. Show that the Cauchy–Riemann Equations are satisfied at the origin yet $f'(0)$ does not exist. Why does this not contradict Theorem 4.12?
11. Consider

$$f(z) = \frac{xy^2(x + \mathrm{i}y)}{x^2 + y^4} \quad (z = x + \mathrm{i}y \neq 0)$$
$$f(0) = 0$$

   Verify that $\lim_{z\to z_0}(f(z) - z_0)/(z - z_0) = 0$ as $z \to 0$ along any straight line, $z = (a + \mathrm{i}b)t, t \in \mathbb{R}$. This does not prove that $f'(0) = 0$, however. By considering $z \to 0$ along the path $z(t) = t^2 + \mathrm{i}t$, show that f is not differentiable at 0. (This shows that when computing f' it is not enough to consider a limit taken along certain specific paths or types of path. An entire neighbourhood of the point concerned must be considered.)
12. For each of the following, compute $f'(\gamma(t)), \gamma'(t), (f\gamma)'(t)$ and verify directly that

$$(f\gamma)'(t) = f(\gamma'(t))\gamma'(t)$$

   (i) $f(z) = z^2, \gamma(t) = t^3 + \mathrm{i}t^4\ (z \in \mathbb{C}, t \in [0, 1])$
   (ii) $f(z) = 1/z, \gamma(t) = \cos t + \mathrm{i}\sin t\ (z \neq 0, t \in [0, 2\pi])$
   (iii) $f(z) = 1 + z + z^2 + \cdots, \gamma(t) = t + \mathrm{i}t^2\ (|z| < 1, t \in [0, \frac{1}{2}])$

13. Suppose that $f(z) = \sum a_n z^n$ is convergent for all $z \in \mathbb{C}$ and satisfies $f' = f$ and $f(0) = 1$. Find a_n for all $n \geq 0$. Consider the derivative of g where

$$g(z) = f(c - z)f(z)$$

for $c \in \mathbb{C}$ and deduce that

$$f(a + b) = f(a) + f(b)$$

for all $a, b \in \mathbb{C}$. Compute $f(1)$ to five decimal places. (A calculator or computer may be used, but electronic assistance is not necessary.)

14. Show that for all $\alpha \in \mathbb{C}$ the power series

$$f_\alpha(z) = \sum_{n=1}^{\infty} \frac{\alpha(\alpha - 1)\cdots(\alpha - n + 1)}{n!} z^n$$

converges for $|z| < 1$, and that in this domain its derivative is $\alpha f_\alpha(z)/(1 + z)$.

Is the radius of convergence 1 for all $\alpha \in \mathbb{C}$?

By differentiating $f_\alpha(z)f_\beta(z)/f_{\alpha+\beta}(z)$ or otherwise, show that

$$f_{\alpha+\beta}(z) = f_\alpha(z)f_\beta(z)$$

Deduce that for $|z| < 1$,

$$f_n(z) = (1 + z)^n$$

for every integer n (positive or negative), and that

$$(f_{1/n}(z))^n = (1 + z)$$

for all positive integers n.

(This power series can be used to define $(1 + z)^\alpha$ for $\alpha \in \mathbb{C}$ and any z in the domain $|z| < 1$.)

15. Assume that the two power series $s(z) = \sum a_n z^n$ and $c(z) = \sum b_n z^n$ are convergent for all $z \in \mathbb{C}$, and that they satisfy the relations $s'(z) = c(z), c'(z) = -s(z)$. Deduce the identities

$$a_n = -a_{n-2}/(n(n-1)) \qquad b_n = -a_{n-2}/(n(n-1))$$

If further $s(0) = 0, c(0) = 1$, determine $s(z)$ and $c(z)$ completely. By differentiation, or otherwise, prove that

$$(c(z))^2 + (s(z))^2 = 1$$

16. For a positive integer n the Bessel function of order n is defined by

$$J_n(z) = \sum_{r=0}^{\infty} \frac{(-1)^r(\frac{1}{2}z)^{n+2r}}{r!(n+r)!}$$

Show that this converges for all complex z and satisfies the differential equation

$$z^2\frac{d^2y}{dz^2} + z\frac{dy}{dz} + (z^2 - n^2)y = 0$$

Verify the following:

(i) $J_{n-1}(z) + J_{n+1}(z) = \frac{2n}{z} J_n(z)$

(ii) $J_n'(z) = \frac{n}{z} J_n(z) - J_{n+1}(z)$

(iii) $J_n'(z) = \frac{1}{2}(J_{n-1}(z) - J_{n+1}(z))$

(iv) $J_n'(z) = J_{n-1}(z) - \frac{n}{z} J_n(z)$

(v) $\frac{\mathrm{d}}{\mathrm{d}z}(z^n J_n(z)) = z^n J_{n-1}(z)$

(vi) $J_2(z) - J_0(z) = 2J_0''(z)$

17. Define $f : \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \begin{cases} 0 & (x \leq 0) \\ \mathrm{e}^{-1/x} & (x > 0) \end{cases}$$

Show that f is differentiable arbitrarily many times, and that $f^{(n)}(0) = 0$ for all $n \in \mathbb{N}$.

5 The Exponential Function

An abstract theory of functions may be intellectually diverting, but it pays its way by its more concrete applications by defining specific functions of particular interest and establishing their properties. As a step in that direction we now discuss the complex versions of the usual (real) exponential and trigonometric functions exp, sin, and cos, defining them as power series. We also introduce complex versions of the usual related functions such as tan and cosec. From these definitions we develop some basic properties of these functions. Euler's famous formula

$$\exp(\mathrm{i}\theta) = \cos\theta + \mathrm{i}\sin\theta$$

follows at once from these definitions, and shows that the fundamental function here is exp. All of the others may be defined very simply in terms of it. We also discuss complex generalisations of the hyperbolic functions, such as sinh, cosh, tanh, which again have simple definitions in terms of exp.

Most of the material generalises directly from the real case, and will be presented in a compact form. In addition to deriving the standard formulas, we place some emphasis on convincing the reader that the power series do indeed represent the *usual* real functions, and that in particular the geometric interpretations of sin and cos in trigonometry are valid. We also derive further properties of these functions that are peculiar to the complex case, including their relation to hyperbolic functions.

The complex version of the logarithm requires deeper analysis, and is held over to Chapter 7.

5.1 The Exponential Function

Definition 3.26 defines the exponential function for complex numbers as the power series

$$\exp z = \sum_{n=0}^{\infty} \frac{z^n}{n!} \tag{5.1}$$

which is absolutely convergent for all $z \in \mathbb{C}$. We may therefore differentiate term by term. The result is the same series, proving that

$$\frac{\mathrm{d}}{\mathrm{d}z} \exp z = \exp z \tag{5.2}$$

With a little ingenuity we can use (5.2) to prove the formula

$$\exp(z_1 + z_2) = \exp(z_1)\exp(z_2) \tag{5.3}$$

derived in a cumbersome way in Chapter 3. Consider

$$f(z) = \exp(z)\exp(c - z)$$

for $c \in \mathbb{C}$. Differentiate:

$$\begin{aligned} f'(z) &= \exp'(z)\exp(c - z) + \exp(z)\exp'(c - z) \\ &= \exp(z)\exp(c - z) + \exp(z)\exp(c - z)(-1) \\ &= 0 \end{aligned}$$

By Theorem 4.14 f is constant; the constant must equal $f(0) = \exp(c)$. So

$$\exp(z)\exp(c - z) = \exp(c)$$

Put $z = z_1, c = z_1 + z_2$ to obtain (5.3).

We wish to use the customary notation e^z for $\exp z$. To avoid ambiguities we must show that when z is rational, say $z = m/n$, this agrees with the usual real exponential $\mathrm{e}^{m/n} = \sqrt[n]{\mathrm{e}^m}$. We do this as follows.

First, observe that $\exp(x) > 0$ for all $x \in \mathbb{R}$. This is obvious from the power series when $x \geq 0$. But when $x < 0$ we know that $\exp(-x)\exp(x) = \exp(0) = 1$, so $\exp(-x) = 1/\exp(x) > 0$. Therefore the nth root of $\exp(x)$ has an unambiguous meaning for any positive integer n.

Next, define the real number

$$\mathrm{e} = \exp(1) = 2.718\,281\ldots \tag{5.4}$$

by an easy computation using (5.1) with $z = 1$. Using (5.3) and induction on n we obtain

$$\exp(nz) = (\exp(z))^n$$

for any positive integer n. Therefore

$$\exp(n) = (\exp(1))^n = \mathrm{e}^n$$

Clearly

$$\exp(0) = 1$$

Then

$$\exp(n)\exp(-n) = \exp(n - n) = \exp(0) = 1$$

so that

$$\exp(-n) = (\exp(n))^{-1} = (\mathrm{e}^n)^{-1} = \mathrm{e}^{-n}$$

Now, for any rational m/n, $(n > 0)$ we have

$$(\exp(m/n))^n = \exp(nm/n) = \exp(m) = \mathrm{e}^m$$

so that

$$\exp(m/n) = (e^m)^{1/n} = e^{m/n}$$

Thus the notation

$$e^z = \exp(z)$$

does not conflict with the standard notation for powers of e, so we may (and do) use it from now on. Equation (5.3) becomes

$$e^{z_1} e^{z_2} = e^{z_1 + z_2} \tag{5.5}$$

Putting $z = x + iy$ this gives

$$e^{x+iy} = e^x e^{iy}$$

Here e^x is the usual real exponential. So we know how e^z behaves provided we also understand e^{iy}. We study both of these in the next two sections.

5.2 Real Exponentials and Logarithms

We briefly recall some standard properties of e^x when x is real.

Clearly $e^x > 1 + x$ for $x > 0$, which implies that $e^x > 0$ for $x \geq 0$, and $e^x \to \infty$ as $x \to \infty$. Also $e^{-x} = 1/e^x$ for $x < 0$, so $e^x > 0$ for all $x \in \mathbb{R}$ and $e^x \to 0$ as $x \to -\infty$. The derivative of e^x is $e^x > 0$ so e^x is monotonic increasing for all x.

That is, e^x defines a continuous strictly increasing function from $\mathbb{R}$ *onto* $\mathbb{R}^+ = \{x \in \mathbb{R} : x > 0\}$. By the Intermediate Value Theorem e^x has a continuous strictly increasing inverse function, defined to be the *natural logarithm*

$$\log : \mathbb{R}^+ \to \mathbb{R}$$

with the property

$$y = \log x \iff x = e^y$$

From (5.5) we obtain

$$\log(x_1 x_2) = \log x_1 + \log x_2 \quad (x_1, x_2 > 0) \tag{5.6}$$

Let $y = \log x,\ y_0 = \log x_0$ $(x, x_0 \in \mathbb{R}^+)$. Since log is continuous,

$$\lim_{x \to x_0} \frac{\log x - \log x_0}{x - x_0} = \lim_{y \to y_0} \frac{y - y_0}{e^y - e^{y_0}} = \frac{1}{e^{y_0}} = \frac{1}{x_0}$$

Therefore

$$\frac{d}{dx} \log x = \frac{1}{x}$$

5.3 Trigonometric Functions

Definition 3.26 also defines the complex sine and cosine functions by the power series

$$\cos z = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n}}{(2n)!} \tag{5.7}$$

$$\sin z = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n+1}}{(2n+1)!} \tag{5.8}$$

We know these are absolutely convergent for all $z \in \mathbb{C}$.

Putting $-z$ for z in (5.7, 5.8) we see that cos is an even function and sin is an odd function. That is,

$$\cos(-z) = \cos(z)$$
$$\sin(-z) = -\sin(z)$$

Also

$$\cos(0) = 1$$
$$\sin(0) = 0$$

Differentiating term by term,

$$\frac{\mathrm{d}}{\mathrm{d}z} \cos z = \sin z \tag{5.9}$$

$$\frac{\mathrm{d}}{\mathrm{d}z} \sin z = -\cos z \tag{5.10}$$

Term-by-term addition, as in Chapter 3, leads to *Euler's formula*

$$\mathrm{e}^{\mathrm{i}z} = \cos z + \mathrm{i} \sin z \tag{5.11}$$

Since $(\mathrm{e}^{\mathrm{i}z})^n = \mathrm{e}^{\mathrm{i}nz}$ for any integer n, equation (5.11) implies *De Moivre's Formula*

$$(\cos z + \mathrm{i} \sin z)^n = \cos nz + \mathrm{i} \sin nz \tag{5.12}$$

This may be used to obtain rapid derivations of formulas for $\cos n\theta$ and $\sin n\theta$ when $\theta \in \mathbb{R}$, by equating real and imaginary parts of both sides (Exercise 6).

Euler's Formula has the famous corollary

$$\mathrm{e}^{\mathrm{i}\pi} = -1 \tag{5.13}$$

Formulas (5.11) and (5.13) seem surprising when we first encounter them, because there is no obvious connection between e and π. Historically, these special numbers arose in very different areas of mathematics – natural logarithms and the circumference of a circle. Moreover, it is hard to see why i, which arises from yet a third area, polynomial equations, should link them together. But in fact, there is a simple reason why these results hold, based on yet another area of mathematics: differential equations, applied to uniform rotation in the plane. We discuss this in Section 5.6. If anything, this dynamic explanation adds to the beauty of these formulas by showing that they are at the core of a remarkable unification of distinct mathematical concepts.

Replacing z by $-z$ in (5.11) gives

$$e^{-iz} = \cos(-z) + i\sin(-z) = \cos z - i\sin z \tag{5.14}$$

From (5.11) and (5.14),

$$\cos z = \tfrac{1}{2}(e^{iz} + e^{-iz}) \tag{5.15}$$
$$\sin z = \tfrac{1}{2i}(e^{iz} - e^{-iz}) \tag{5.16}$$

From (5.15, 5.16), and (5.5), we obtain the usual addition formulas for sin and cos, now valid for all complex numbers z_1, z_2:

$$\sin(z_1 + z_2) = \sin z_1 \cos z_2 + \cos z_1 \sin z_2 \tag{5.17}$$
$$\sin(z_1 - z_2) = \sin z_1 \cos z_2 - \cos z_1 \sin z_2 \tag{5.18}$$
$$\cos(z_1 + z_2) = \cos z_1 \cos z_2 - \sin z_1 \sin z_2 \tag{5.19}$$
$$\cos(z_1 - z_2) = \cos z_1 \cos z_2 + \sin z_1 \sin z_2 \tag{5.20}$$

Putting $z_1 = z_2 = z$ in (5.20) gives

$$\cos^2 z + \sin^2 z = 1 \tag{5.21}$$

5.4 An Analytic Definition of π

Historically the real number π was defined as the ratio of the circumference of a circle to its diameter, and only later did its importance for trigonometric functions emerge. We reverse the process, defining π analytically and eventually showing in Section 7.1 that our definition agrees with the geometric one.

The idea is to define $\pi/2$ as the first positive real solution of the equation $\cos x = 0$. The problem is to show that there *is* such a thing.

We know that cos and sin are continuous functions. Also,

$$\begin{aligned}\cos(2) &= 1 - \frac{2^2}{4!} + \frac{2^6}{6!} - \cdots - \frac{2^{4n-2}}{(4n-2)!} + \frac{2^{4n}}{(4n)!} - \cdots \\ &= 1 - 2 + \frac{2}{3} - \cdots - \frac{2^{4n-2}}{(4n)!}[4n(4n-1) - 4] - \cdots \\ &< 1 - 2 + \frac{2}{3} < 0\end{aligned}$$

But $\cos(0) = 1$. By the Intermediate Value Theorem, $\cos t_0 = 0$ for some $t_0 \in (0, 2)$. Let k be the greatest lower bound of $\{t \in \mathbb{R} : t > 0, \cos t = 0\}$. By continuity,

$$\cos k = 0$$

By the definition of k, if $0 \leq x < k$ then $\cos x > 0$.

We define

$$\pi = 2k$$

Then the number π has been uniquely defined by the properties $\pi/2 > 0, \cos(\pi/2) = 0$ and $0 \le x < \pi/2$ implies $\cos x > 0$.

Since $\cos(2) < 0$, it follows that $0 < \pi < 4$. This is a crude estimate, and we improve it in Exercises 17 and 18 below.

5.5 The Behaviour of Real Trigonometric Functions

We know that $\cos x$ is positive for $0 \le x < \pi/2$. Since $\frac{\mathrm{d}}{\mathrm{d}x}\sin x = \cos x$, it follows that sin is strictly increasing on $[0, \pi/2]$. Since

$$\sin^2\frac{\pi}{2} + \cos^2\frac{\pi}{2} = 1$$

and $\cos \pi/2 = 0$, we must have $\sin \pi/2 = \pm 1$. But since sin is increasing on $[0, \pi/2]$, we must have

$$\sin\frac{\pi}{2} = 1$$

Now (5.18) implies that

$$\sin\left(\frac{\pi}{2} - x\right) = \cos x \tag{5.22}$$

Hence cos decreases monotonically from 1 to 0 in $[0, \pi/2]$. Using (5.17) and (5.19) repeatedly, we can deduce the behaviour of sin and cos in the intervals $[\pi/2, \pi], [\pi, 3\pi/2]$, and $[3\pi/2, 2\pi]$. We obtain

$$\begin{aligned}\cos\left(\frac{\pi}{2} + x\right) &= -\sin x\\ \sin\left(\frac{\pi}{2} + x\right) &= \cos x\\ \sin(\pi + x) &= -\sin x\end{aligned}$$

and so on. We tabulate the results in Table 5.1, where $a \nearrow b$ means ‘strictly increasing from a to b’, and $a \searrow b$ means ‘strictly decreasing from a to b’.

From the table,

$$\begin{aligned}\cos 2\pi &= 1\\ \sin 2\pi &= 0\end{aligned}$$

Therefore

$$\cos(x + 2\pi) = \cos x \cos 2\pi - \sin x \sin 2\pi = \cos x \tag{5.23}$$

$$\sin(x + 2\pi) = \sin x \cos 2\pi + \cos x \sin 2\pi = \sin x \tag{5.24}$$

leading to

$$\begin{aligned}\cos(x + 2n\pi) &= \cos x\\ \sin(x + 2n\pi) &= \sin x\end{aligned}$$

for all $n \in \mathbb{Z}$. So sin and cos are *periodic* with period 2π. In particular the behaviour in the table repeats on each interval $[2n\pi, (2n + 2)\pi]$.

Table 5.1 Behaviour of sin and cos in intervals of length $\pi/2$.

interval	cos	sin
$[0, \pi/2]$	$1 \searrow 0$	$0 \nearrow 1$
$[\pi/2, \pi]$	$0 \searrow -1$	$1 \searrow 0$
$[\pi, 3\pi/2]$	$-1 \nearrow 0$	$0 \searrow -1$
$[3\pi/2, 2\pi]$	$0 \nearrow 1$	$-1 \nearrow 0$

In this way, purely formal considerations show that $\sin x$ and $\cos x$ have their usual geometric properties for all real x, at least in outline. The final remaining step is to identify the point $\cos\theta + \mathrm{i}\sin\theta = \mathrm{e}^{\mathrm{i}\theta}$ on the unit circle $|z| = 1$ with the point θ radians from 1 in an anticlockwise direction. We indicate two different approaches in Section 5.6 and Exercise 8 in Chapter 6. We also discuss the topic in detail in Chapter 7.

More precise computations of the values of these functions may be performed to any desired degree of accuracy. By inspection, it may be seen that for real x (positive or negative) the terms of the power series for $\sin x$ and $\cos x$ always alternate in sign. From the standard theory for alternating real series, the sum of the first n terms alternately overestimates and underestimates the actual limit. This lets us make very precise estimates of the trigonometric functions.

Example 5.1. Compute e^{i} to four decimal places.

We have $\mathrm{e}^{\mathrm{i}} = \cos(1) + \mathrm{i}\sin(1)$. Now

$$\cos(1) = 1 - \frac{1}{2!} + \frac{1}{4!} - \frac{1}{6!} + \frac{1}{8!} - \cdots$$

Considering partial sums:

$$\begin{aligned}
\cos(1) &< 1 \\
\cos(1) &> 1 - \frac{1}{2!} = 0.5 \\
\cos(1) &< 1 - \frac{1}{2!} + \frac{1}{4!} = 0.541\,66\ldots \\
\cos(1) &> 1 - \frac{1}{2!} + \frac{1}{4!} - \frac{1}{6!} = 0.540\,27\ldots \\
\cos(1) &< 1 - \frac{1}{2!} + \frac{1}{4!} - \frac{1}{6!} + \frac{1}{8!} = 0.540\,30\ldots
\end{aligned}$$

Therefore

$$\cos(1) = 0.5403 \text{ (to 4 decimal places)}$$

Similarly

$$\sin(1) = 0.8415 \text{ (to 4 decimal places)}$$

Therefore

$$\mathrm{e}^{\mathrm{i}} = 0.5403 + 0.8415\mathrm{i} \text{ (to 4 decimal places)}$$ □

5.6 Dynamic Explanation of Euler's Formula

Euler's formula(s) linking $\mathrm{e}, \mathrm{i}, \pi$ seem surprising, but the existence of such a link, and even the details of how it goes, can be deduced very naturally from standard ideas in the theory of differential equations. This section is intended as motivation, and to help explain why such a connection exists. It can be made rigorous by setting up the necessary ideas formally. It is not used later in the book and can be skipped.

Consider the linear differential equation

$$\frac{\mathrm{d}z}{\mathrm{d}t} = \mathrm{i}z \quad (z \in \mathbb{C}) \tag{5.25}$$

on $\mathbb{C}$. Equation (5.2) and the chain rule show that a solution $z(t)$ with initial condition $z(0) = 1$ is

$$z(t) = \mathrm{e}^{\mathrm{i}t}$$

This solution is unique since the difference w between any two solutions satisfies

$$\frac{\mathrm{d}w}{\mathrm{d}t} = 0$$

so $w(t)$ is constant, and must equal $w(0) = 0$.

Geometrically, (5.25) corresponds to a point particle $z(t)$ moving in the plane $\mathbb{C}$, and it states that the velocity vector $z'(t)$ is at right angles to the position $z(t)$, and the speed is $|z(t)|$, Figure 5.1. (Here $'$ indicate the time-derivative. The right angle arises from the factor i in the equation.) That is: the particle always moves at right angles to the line joining it to the origin. Intuitively, the particle must move along the unit circle. To verify this, we prove that $z'(t)$ is always tangent to the unit circle $\mathbb{S}$, or equivalently that $z(t) \in \mathbb{S}$ for all t. The key point is that $|z(t)|^2$ is conserved; that is, it is constant. Compute:

$$\frac{\mathrm{d}}{\mathrm{d}t}|z(t)|^2 = \frac{\mathrm{d}}{\mathrm{d}t}z(t)\bar{z}(t) = z'(t)\bar{z}(t) + z(t)\bar{z}'(t) = (\mathrm{i}\mathrm{e}^{\mathrm{i}t})(\mathrm{e}^{-\mathrm{i}t}) + (\mathrm{e}^{\mathrm{i}t})(-\mathrm{i}\mathrm{e}^{-\mathrm{i}t}) = \mathrm{i} - \mathrm{i} = 0$$

so $|z(t)|^2$ is constant. Since it equals 1 at time 0, we know that $|z(t)|^2 = 1$ for all t. Therefore $z(t)$ always lies on the unit circle.

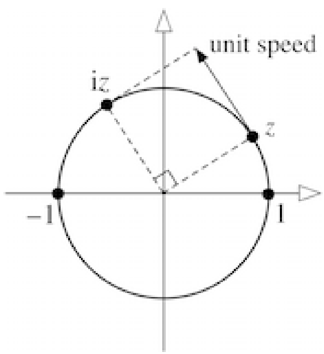

Figure 5.1 Uniform motion round the unit circle.

The speed of motion is

$$|z'(t)| = \sqrt{z'(t)\bar{z}'(t)} = \sqrt{(\mathrm{i}e^{\mathrm{i}t})(-\mathrm{i}e^{-\mathrm{i}t})} = 1$$

so the particle moves with unit speed (anticlockwise).

After time t, the particle therefore moves a distance t along the unit circle in the anticlockwise direction; that is, through an angle of t radians.

At time $t = \pi$ is has gone a distance π round the unit circle – that is, halfway round. So it must lie at the point diametrically opposite the starting point $z(0) = 1$. This point is -1. So $z(\pi) = -1$; that is, $e^{\mathrm{i}\pi} = -1$, the miraculous formula.

More generally, at time t its position on the unit circle makes angle t radians with the real axis at the origin, so its Cartesian coordinates are $(\cos t, \sin t)$. Interpreted as a complex number, this point is $\cos t + \mathrm{i} \sin t$. Therefore

$$e^{\mathrm{i}t} = \cos t + \mathrm{i} \sin t$$

and we recover Euler's formula (5.11). This approach is closely related to Section 7.1.

5.7 Complex Exponential and Trigonometric Functions are Periodic

We are used to the real sine and cosine functions being periodic, with period 2π, and we have proved that the same property holds for their complex generalisations. Euler's formula shows that their periodicity is inherited by the complex exponential function, as we now describe.

DEFINITION 5.2. A complex function $f : S \to \mathbb{C}$ has *period* $\rho \in \mathbb{C}$ if

$$f(z + \rho) = f(z) \qquad (z, z + \rho \in S)$$ □

Obviously, if ρ is a period of f, so is $n\rho$ for any positive integer n such that $m\rho \in S$ for $m \leq n$, and with similar conditions we can also take n negative.

For the complex exponential, $S = \mathbb{C}$, and

$$e^{2\pi\mathrm{i}} = \cos 2\pi + \mathrm{i} \sin 2\pi = 1$$

so

$$e^{z+2\pi\mathrm{i}} = e^z e^{2\pi\mathrm{i}} = e^z \cdot 1 = e^z$$

Therefore $2\pi\mathrm{i}$ is a period for exp. So is $2n\pi\mathrm{i}$ for any $n \in \mathbb{Z}$. There are no others:

PROPOSITION 5.3. *The complex number ρ is a period of* exp *if and only if $\rho = 2n\pi\mathrm{i}$ for some $n \in \mathbb{Z}$.*

Proof. If ρ is a period then $e^\rho = e^{z+\rho}/e^z = 1$. If $\rho - u + \mathrm{i}v$ then

$$1 = e^{u+\mathrm{i}v} = e^u e^{\mathrm{i}v} = e^u(\cos v + \mathrm{i} \sin v)$$

Taking the modulus,

$$1 = |e^u||\cos v + i \sin v| = e^u$$

By properties of the real exponential proved in Section 5.2, $u = 0$. So now

$$\cos v + i \sin v = 1$$

Therefore $\cos v = 1, \sin v = 0$. By Table 5.1, $v = 2n\pi, (n \in \mathbb{Z})$. □

We also have:

PROPOSITION 5.4. *The exponential function* e^z *is non-zero for any* $z \in \mathbb{C}$.

Proof. For any $z \in \mathbb{C}$, $e^z e^{-z} = e^0 = 1$. □

We now investigate sin and cos for periodicity in $\mathbb{C}$. Certainly $2n\pi$ is a period of both, because the proofs of (5.23) and (5.24) for real x also work for complex z.

PROPOSITION 5.5. *The complex number* ρ *is a period of* sin *or* cos *if and only if* $\rho = 2n\pi$ *for some* $n \in \mathbb{Z}$.

Proof. Since $\sin(z + \pi/2) = \cos z$ by (5.17), it follows that ρ is a period for cos if and only if it is a period for sin. Then

$$\cos(z + \rho) = \cos z$$
$$\sin(z + \rho) = \sin z$$

for all $z \in \mathbb{C}$. But then

$$e^{i(z+\rho)} = \cos(z + \rho) + i \sin(z + \rho) = \cos z + i \sin z = e^{iz}$$

so $i\rho$ is a period of exp. By Proposition 5.3, $i\rho = 2n\pi i$ for some $n \in \mathbb{Z}$, so $\rho = 2n\pi$. □

We can also find the zeros of sin and cos:

PROPOSITION 5.6. *Let* $z \in \mathbb{C}$. *Then*

$$\cos z = 0 \textit{ if and only if } z = (n + \tfrac{1}{2})\pi$$
$$\sin z = 0 \textit{ if and only if } z = n\pi$$

where $n \in \mathbb{Z}$.

Proof. Since $\sin(z + \pi/2) = \cos z$ by (5.17), the second equation implies the first. Now $\sin z = 0$ if and only if $(e^{iz} - e^{-iz})/2i = 0$. This holds if and only if $e^{iz} - e^{-iz} = 0$, that is, if and only if $e^{2iz} = 1$. So $2iz = 2n\pi i$. Therefore $z = n\pi$ as claimed. □

5.8 Other Trigonometric Functions

If $z \neq (n + \frac{1}{2})\pi$ then $\cos z \neq 0$, so we may define

$$\tan z = \frac{\sin z}{\cos z} \tag{5.26}$$

If $S = \{z \in \mathbb{C} : z \neq (n + \frac{1}{2})\pi, n \in \mathbb{Z}\}$ then S is a domain, and $\tan : S \to \mathbb{C}$ is a differentiable function. Its derivative is

$$\begin{aligned}\frac{d}{dz}\tan z &= \frac{\cos z\frac{d}{dz}\sin z - \sin z\frac{d}{dz}\cos z}{\cos^2 z}\\ &= \frac{\cos z\cdot\cos z - \sin z\cdot(-\sin z)}{\cos^2 z}\\ &= \frac{\cos^2 z+\sin^2 z}{\cos^2 z}\\ &= 1+\tan^2 z\end{aligned}$$

Similarly we define

$$\cot z = \frac{\cos z}{\sin z} \quad (z \neq n\pi) \tag{5.27}$$

$$\sec z = \frac{1}{\cos z} \quad (z \neq (n+\tfrac{1}{2})\pi) \tag{5.28}$$

$$\operatorname{cosec} z = \frac{1}{\sin z} \quad (z \neq n\pi) \quad \text{(often also written csc } z) \tag{5.29}$$

These are all differentiable functions (on the stated domains) whose derivatives may again be calculated in the usual manner. All of the standard formulas relating these trigonometric functions may be deduced from properties of sin and cos, hence also apply for complex values of the variables.

For example, using (5.17) and (5.19), we obtain

$$\tan(z_1+z_2) = \frac{\tan z_1 + \tan z_2}{1-\tan z_1 \tan z_2}$$

provided that $z_1, z_2, z_1+z_2 \in S$. This implies that $\tan(z+\pi) = \tan z$, so π is a period for tan. It is easy to see that the only periods of tan are $n\pi$ for $n \in \mathbb{Z}$.

The reader is encouraged to develop all of his or her favourite trigonometric formulas for the complex case, including the basic properties of cot, sec, and cosec.

5.9 Hyperbolic Functions

As in the real case, we define

$$\sinh z = \tfrac{1}{2}(e^z - e^{-z})$$
$$\cosh z = \tfrac{1}{2}(e^z + e^{-z})$$

for $z \in \mathbb{C}$. Differentiating,

$$\frac{d}{dz}\sinh z = \cosh z$$
$$\frac{d}{dz}\cosh z = \sinh z$$

Properties of the hyperbolic functions, analogous to those of trigonometric functions (such as addition formulas for $\sinh(z_1+z_2)$) follow either by direct computation, or by using the obvious identities

$$\sin iz = i\sinh z$$
$$\cos iz = \cosh z$$

For example,

$$\begin{aligned}\cosh^2 z - \sinh^2 z &= \cos^2 \mathrm{i}z - (-\mathrm{i}\sin^2 \mathrm{i}z)\\ &= \cos^2 \mathrm{i}z + \mathrm{i}\sin^2 \mathrm{i}z\\ &= 1\end{aligned}$$

The functions tanh, coth, sech, and cosech (or csch) are defined in the obvious way. We leave it to the reader to discover their properties, including zeros and periods.

Hyperbolic functions occur in expressions for the real and imaginary parts of $\sin z$ and $\cos z$. Thus let $z = x + \mathrm{i}y$. Then

$$\begin{aligned}\sin z &= \sin(x + \mathrm{i}y)\\ &= \sin x \cos \mathrm{i}y + \cos x \sin \mathrm{i}y\\ &= \sin x \cosh y + \mathrm{i}\cos x \sinh y \qquad (5.30)\end{aligned}$$

Similarly

$$\cos z = \cos x \cosh y - \mathrm{i}\sin x \sinh y \qquad (5.31)$$

5.10 Exercises

1. Express the following in the form $a + \mathrm{i}b$ for real a, b:
 (i) $\exp(\mathrm{i})$
 (ii) $\mathrm{e}^{2+\mathrm{i}\pi}$
 (iii) $1/\exp(2 + \mathrm{i}\pi)$

2. Express the following in the form $a + \mathrm{i}b$ for real a, b:
 (i) $\sin(\mathrm{i})$
 (ii) $\cos(\mathrm{i})$
 (iii) $\sinh(\mathrm{i})$
 (iv) $\cosh(\mathrm{i})$
 (v) $\cos(\pi/4 - \mathrm{i})$
 (vi) $\tan(1 + \mathrm{i})$

3. Differentiate the functions defined as follows:
 (i) $\exp(z^2 + 2z)$
 (ii) $1/\exp(z)$
 (iii) $\exp(z^2)/\exp(z + 1)$

4. Differentiate the functions defined as follows:
 (i) $\tan(z^2)$
 (ii) $\sinh(z + 2)\exp(z^3)$
 (iii) $\sin(z)\cosh(z)\exp(z)$

5. Use the identity $\mathrm{e}^{\mathrm{i}\theta}\mathrm{e}^{\mathrm{i}\phi} = \mathrm{e}^{\mathrm{i}(\theta+\phi)}$ to derive the usual formulas for $\cos(\theta + \phi)$ and $\sin(\theta + \phi)$. By a similar method, show that

$$1/(\cos\theta + \mathrm{i}\sin\theta) = \cos\theta - \mathrm{i}\sin\theta$$

6. Use the identity $\left(e^{i\theta}\right)^3 = e^{3i\theta}$ to give the usual formulas for $\cos(3\theta)$ and $\sin(3\theta)$. Derive similar formulas for $\cos(4\theta), \sin(4\theta), \cos(5\theta)$, and $\sin(5\theta)$.

7. Draw the following paths:
(i) $\gamma(t) = e^{-it}$ $(t \in [0, \pi])$
(ii) $\gamma(t) = 1 + i + 2e^{-it}$ $(t \in [0, 2\pi])$
(iii) $\gamma(t) = z_0 + re^{-it}$ $(t \in [0, 2\pi])$, where $z_0 \in \mathbb{C}$ and $r > 0$
(iv) $\gamma(t) = t + i\cosh t$ $(t \in [-1, 1])$
(v) $\gamma(t) = \cosh t + i\sinh t$

8. 'Osborne's rule' states that any formula involving sin and cos has an analogous formula involving sinh and cosh, which is the same in every way except that the product of two sines must be replaced by *minus* the product of two hyperbolic sines. For each of the following formulas, write down the corresponding formula using Osborne's rule and verify it from first principles:
(i) $\sin 2A + \cos A = 1$
(ii) $\sin(A - B) = \sin A \cos B - \cos A \sin B$
(iii) $\cos(A + B) = \cos A \cos B - \sin A \sin B$
Comment on Osborne's rule in the light of the formulas

$$\cos iz = \cosh z \qquad \sin iz = i\sinh z$$

9. Show that the complex conjugate of $\cos z$ is $\cos \bar{z}$ and that of $\sin z$ is $\sin \bar{z}$. Verify the identities

$$|\sin z|^2 = \tfrac{1}{2}(\cosh 2y - \cos 2x) = \sinh^2 y + \sin^2 x = \cosh^2 y - \cos^2 x$$
$$|\cos z|^2 = \tfrac{1}{2}(\cosh 2y + \cos 2x) = \sinh^2 y + \cos^2 x = \cosh^2 y - \sin^2 x$$

10. Show that $|\cos z|^2 + |\sin z|^2 = 1$ if and only if z is real, and that $\cos z$ is unbounded on $\mathbb{C}$ (that is, no $K > 0$ exists such that $|\cos z| \leq K$ for all $z \in \mathbb{C}$). This contrasts with the bound $|\cos x| \leq 1$ for all real x.

11. Derive formulas for the real and imaginary parts of the following functions of z, and check directly that they satisfy the Cauchy–Riemann Equations:
(i) $\exp z$
(ii) $\sin z$
(iii) $\cos z$
(iv) $\sinh z$
(v) $\cosh z$

12. Derive formulas for the real and imaginary parts of the following functions of z, specifying the largest domain on which they are defined, and check directly that they satisfy the Cauchy–Riemann Equations:
(i) $\tan z$
(ii) $\tanh z$
(iii) $\operatorname{cosec} z$
(iv) $\operatorname{cosech} z$
(v) $\cot z$
(vi) $\coth z$

13. Write $\tanh(x + iy)$ in real and imaginary parts and show that $\tanh(x + iy)$ is real if and only if $y = n\pi/2$ for $n \in \mathbb{Z}$.

14. For each of the functions exp, cos, sin, tan, cosh, sinh, tanh, find the set of points on which it assumes:
(i) real values; and
(ii) purely imaginary values.

15. By considering the real and imaginary parts of

$$1 + z + z^2 + \cdots + z^n = \frac{1 - z^{n+1}}{1 - z}$$

find explicit formulas for the sums:
(i) $1 + \cos x + \cos 2x + \cdots + \cos nx$
(ii) $\sin x + \sin 2x + \cdots + \sin nx$
By similar methods, find:
(iii) $\cos x + \cos 3x + \cdots + \cos(2n - 1)x$
(iv) $\sin x + \sin 3x + \cdots + \sin(2n - 1)x$
(v) $\sin x - \sin 2x + \cdots + (-1)^n \sin nx$
(vi) $\cos\theta + \cos(\theta + \phi) + \cdots + \cos(\theta + n\phi)$
(vii) $\sin\theta + \sin(\theta + \phi) + \cdots + \sin(\theta + n\phi)$

16. If $z(t) = x(t) + iy(t)$ is a solution of the differential equation

$$\frac{d^2z}{dt^2} + \lambda z = k_0 e^{i\omega t} \quad (\lambda, k_0, \omega \in \mathbb{R}) \tag{5.32}$$

show that $x = x(t) = \operatorname{re} z(t)$ is a solution of

$$\frac{d^2x}{dt^2} + \lambda x = k_0 \cos\omega t \tag{5.33}$$

By finding solutions of (5.32) of the form $z(t) = ke^{i\omega t}$, find a solution of (5.33).

If k_0 is *complex*, say $k_0 = k_1 e^{i\varepsilon}$ $(k, \varepsilon \in \mathbb{R})$, and λ, ω are real, write down the real part of (5.32) to obtain

$$\frac{d^2z}{dt^2} + \lambda z = k_0 \cos(\omega t + \varepsilon) \tag{5.34}$$

Show that there is a solution of (5.34) of the form

$$x(t) = \frac{k_1}{\lambda - \omega^2} \cos(\omega t + \varepsilon)$$

17. By using the sum of the geometric progression

$$1 + z + z^2 + \cdots + z^n = \frac{1 - z^{n+1}}{1 - z}$$

find $A_n(x)$ such that

$$\frac{1}{1 + x^2} = 1 - x^2 + x^4 - \cdots + (-1)^n x^{2n} + A_n(x)$$

Verify that the derivative of $\tan^{-1} x$ is $1/(1+x^2)$ where $\tan^{-1} : \mathbb{R} \to (-\pi/2, \pi/2)$ is the inverse function of $\tan : (-\pi/2, \pi/2) \to \mathbb{R}$. Hence deduce that

$$\tan^{-1} t = \int_0^t \frac{\mathrm{d}x}{1+x^2} = t - t^3/3 + t^5/5 - \cdots + (-1)^n t^{2n+1}/(2n+1) + \int_0^t A_n(x)\mathrm{d}x$$

By estimating the size of the second integral, show that the power series

$$t - t^3/3 + t^5/5 - \cdots + (-1)^n t^{2n+1}/(2n+1) + \cdots$$

converges to $\tan t$ for $|t| < 1$.

Deduce *Gregory's series*

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots$$

18. Gregory's series converges very slowly. Classically, better methods for calculating π were obtained using the related series

$$\frac{\pi}{4} = \tan^{-1}\frac{1}{2} + \tan^{-1}\frac{1}{3}$$
$$\frac{\pi}{4} = 4\tan^{-1}\frac{1}{5} + \tan^{-1}\frac{1}{239}$$

Verify these formulas using the addition formula for $\tan(\theta + \phi)$. Use the second to calculate π correct to 5 decimal places. (This requires only one term of $\tan^{-1}\frac{1}{239}$ and five terms of $\tan^{-1}\frac{1}{5}$.)

Series that converge far more rapidly to π are now known. You can find them easily on the Internet.

6 Integration

The next part of the grand plan is to define complex integration by analogy with the real case, and establish the inverse relation between differentiation and integration.

Consider a complex function $f : D \to \mathbb{C}$ on a domain D, and let $z_0, z_1 \in D$. The real integral $\int_a^b f(t)\mathrm{d}t$ does not generalise immediately to the complex integral $\int_{z_0}^{z_1} f(z)\mathrm{d}z$ because this expression does not specify how z goes from z_0 to z_1. To do that, we must specify a path γ between them. A sensible notation for the integral is then $\int_\gamma f(z)\mathrm{d}z$, or $\int_\gamma f$ for short.

In real analysis the value of a definite integral $\int_a^b f(t)\mathrm{d}t$ depends only on the limits a, b. Complex analysis is more complicated, and the integral along a path may depend on the path as well as on its end points. We give a simple example in Section 6.10 as soon as we have introduced the necessary concepts and techniques.

There are two approaches to defining $\int_\gamma f$. The first is to build up the theory of complex Riemann sums by mimicking the real case. We do this in Sections 6.1 and 6.2. This approach applies to any continuous function, provided it is integrated along a continuous path $\gamma : [a, b] \to \mathbb{C}$. This is one reason why we built the condition of continuity into the definition of a path in Chapter 2, by requiring γ to be a continuous map.

In the special case where the function γ is continuously differentiable, with derivative γ', we derive an explicit formula for the integral:

$$\int_\gamma f = \int_a^b (f\gamma(t))\gamma'(t)\mathrm{d}t \tag{6.1}$$

Here $z_0 = \gamma(a), z_1 = \gamma(b)$. This formula makes sense only when γ' is defined and the expression concerned is Riemann integrable; the requirement that γ should be continuously differentiable is the simplest way to ensure this. To simplify terminology, such a path is said to be smooth.

A key result, the Estimation Lemma 6.41, bounds the absolute value of such an integral by the product of the supremum of $|f(z)|$ on γ and the length of the path γ. It turns out that 'length' need not be a meaningful concept for a continuous path – another somewhat counterintuitive fact – but it is well-behaved for smooth paths. In Section 6.3 we define 'length' and show that for any smooth path it is given by the simple formula

$$L = \int_a^b |\gamma'(t)|\mathrm{d}t \tag{6.2}$$

Notice that the derivative of γ features explicitly in this formula, which is one reason for assuming smoothness.

These equations lead to a second approach, in which the reader is assumed to be familiar with real integration. Why not restrict the theory to smooth paths (or, more generally, piecewise smooth paths – that is, finite sums of smooth paths – which pose no new problems)? Then we can use (6.1) and (6.2) to *define* $\int_\gamma f$ and L. Admittedly, the integrand $(f\gamma(t))\gamma'(t)$ in (6.1) is complex, but we can reduce the formula to real integrals by writing it as $U(t) + \mathrm{i}V(t)$ and defining

$$\int_\gamma f = \int_a^b U(t)\mathrm{d}t + \mathrm{i}\int_a^b V(t)\mathrm{d}t$$

We adopt this method in Section 6.8 as an alternative route to complex integration, allowing Sections 6.1–6.3 to be omitted. This short cut has a price. It means that a couple of proofs later in the chapter must be given in a more technical and less intuitive manner. But this price is not very great, and it leaves the reader with a genuine choice: work through the theory of complex Riemann integration to experience the full analogy with the real case, using continuous paths, or bypass the next three sections and start at Section 6.4, restricting to smooth paths.

6.1 The Real Case

For the reader who has chosen to build up the analogy between the real and complex integrals, we begin by recalling the real case.

The Riemann integral $\int_a^b \phi(t)\mathrm{d}t$ of a real function $\phi : [a, b] \to \mathbb{R}$ is defined in stages. First, subdivide the interval $[a, b]$ to obtain a partition P of $[a, b]$ given by $a = t_0 < t_1 < \cdots < t_n = b$, and choose intermediate points s_r in each subinterval $t_{r-1} \leq s_r \leq t_r$. Then form the Riemann sum

$$S(P, \phi) = \sum_{r=1}^{n} \phi(s_r)(t_r - t_{r-1})$$

The points $t_0, t_1, \ldots, t_n$ are called the *division points* of P. Another partition Q is said to be *finer than* P if the division points of P are all included in those of Q.

The following result is well known from real analysis; we quote it without proof.

LEMMA 6.1. *Let $\phi : [a, b] \to \mathbb{R}$ be continuous. Then there exists a real number A such that for any $\varepsilon > 0$ there is a partition P_ε of $[a, b]$ such that, for every partition P finer than P_ε, we have*

$$|S(P, \phi) - A| < \varepsilon \qquad \square$$

This real number A is denoted by $\int_a^b \phi(t)\mathrm{d}t$, and is the *Riemann integral* of ϕ from a to b.

The actual computation of $\int_a^b \phi(t)\mathrm{d}t$ is usually performed by antidifferentiation:

THEOREM 6.2 (Fundamental Theorem of Calculus). (i) *If a real function ϕ is continuous on $[a, b]$ and $F' = \phi$, then*

$$\int_a^b \phi(t)\mathrm{d}t = F(b) - F(a)$$

(ii) *If ϕ is continuous on* $[a,b]$ *and*

$$I(x) = \int_a^x \phi(t)\mathrm{d}t \quad (x \in [a,b])$$

then $I' = \phi$. □

Example 6.3. To compute $\int_a^b t^5 \mathrm{d}t$ we do not need to calculate the sums $S(P,\phi)$ where $\phi(t) = t^5$. Since $F(t) = t^6/6$ satisfies $F' = \phi$, Theorem 6.2(i) immediately gives

$$\int_a^b t^5 \mathrm{d}t = \tfrac{1}{6}b^6 - \tfrac{1}{6}a^6$$ □

More generally, we can compute $\int_a^b \phi(t)\mathrm{d}t$ by seeking an antiderivative F and appealing to Theorem 6.2(i).

Before passing to the complex case, it is helpful to consider a slight generalisation: the Riemann–Stieltjes integral $\int_a^b \phi(t)\mathrm{d}\theta$, where θ is a second real function. Here, given a partition P of $[a,b]$ as above, we form the sum

$$S(P,\phi,\theta) = \sum_{r=1}^{n} \phi(s_r)(\theta(t_r) - \theta(t_{r-1}))$$

From real analysis we have a generalisation of Lemma 6.1:

LEMMA 6.4. *Let ϕ, θ be real functions defined on* $[a,b]$, *such that ϕ is continuous and θ has continuous derivative θ'. Then the real number* $B = \int_a^b \phi(t)\theta'(t)\mathrm{d}t$ *satisfies the following condition:*

Given any $\varepsilon > 0$ *there is a partition P_ε of* $[a,b]$ *such that, for every partition P finer than P_ε, we have*

$$|S(P,\phi,\theta) - B| < \varepsilon$$ □

The limit B of the sum $S(P,\phi,\theta)$ is also denoted by $\int_a^b \phi(t)\mathrm{d}\theta$. It is the *Riemann–Stieltjes integral* of ϕ from a to b with respect to θ. Lemma 6.4 tells us that when θ is differentiable, the Riemann–Stieltjes integral $\int_a^b \phi(t)\mathrm{d}\theta$ is equal to the Riemann integral $\int_a^b \phi(t)\theta'(t)\mathrm{d}t$. The conditions of the lemma ensure that the integrand $\phi(t)\theta'(t)$ exists and is continuous.

The Riemann and Riemann–Stieltjes integrals exist under far more general conditions on ϕ, θ than those we have mentioned, and the reduction from the Riemann–Stieltjes integral to the Riemann integral is then not always valid. The conditions in Lemma 6.4 are all that we require in this book.

6.2 Complex Integration Along a Smooth Path

Riemann sums define complex integrals for *continuous* functions and paths, but we can derive a simple formula for the integral, which often lets us compute it, if we work with a more special class of paths: those with continuous derivatives.

DEFINITION 6.5. A path $\gamma : [a, b] \to \mathbb{C}$ is *smooth* if γ' exists and is continuous throughout all of $[a, b]$. □

This means that if $\gamma(t) = x(t) + \mathrm{i}y(t)$, where x and y are real, then x' and y' exist and are continuous on the whole of $[a, b]$, including one-sided derivatives at the end points.

The definition is intended to formalise the physical sense of drawing a smooth path with a finger from the starting point to the final point. Invoking dynamical imagery, the path γ describes a point particle moving in the plane, whose position at any time $t \in [a, b]$ is $\gamma(t)$. Smoothness ensures that the point $\gamma(t)$ moves with a specific *velocity* at all times t, namely $\gamma'(t)$, and that this velocity changes continuously. Our intuition here is not just geometric: it is dynamic.

Recall that velocity is usually described as 'speed and direction', where speed is the magnitude of the velocity and direction is determined by the time-derivative of position. The speed is $|\gamma'(t)|$, a non-negative real number. When $\gamma'(t) \neq 0$, the coordinates $(x'(t), y'(t))$ of $|\gamma'(t)|$ determine a non-zero vector in the plane, which does indeed point in a specific direction. What is often ignored is the caveat that when $\gamma'(t) = 0$ this becomes the zero vector, which does *not* point in a specific direction. This distinction is important both geometrically and dynamically, and we discuss it further in Section 6.7.

We now consider integration of a complex function along a smooth path in $\mathbb{C}$. In this case, the Riemann integral of a complex function f can be approached by analogy with the limit of the sum $\sum \phi(s_r)(t_r - t_{r-1})$. We simply consider $\sum f(\zeta_r)(z_r - z_{r-1})$, where ζ_r, z_r are complex, and select ζ_r and z_r along a path γ in the domain of f as in Figure 6.1.

For this purpose we assume:

(i) $f : D \to \mathbb{C}$ is continuous, where D is a domain; and
(ii) $\gamma : [a, b] \to D$ is smooth.

For any partition P of $[a, b]$, given by $a = t_0 < t_1 < \cdots < t_n = b$, and intermediate points $t_{r-1} \le s_r \le t_r$, form the sum

$$S(P, f, \gamma) = \sum_{r=1}^{n} f(s_r)(\gamma(t_r) - \gamma(t_{r-1}))$$

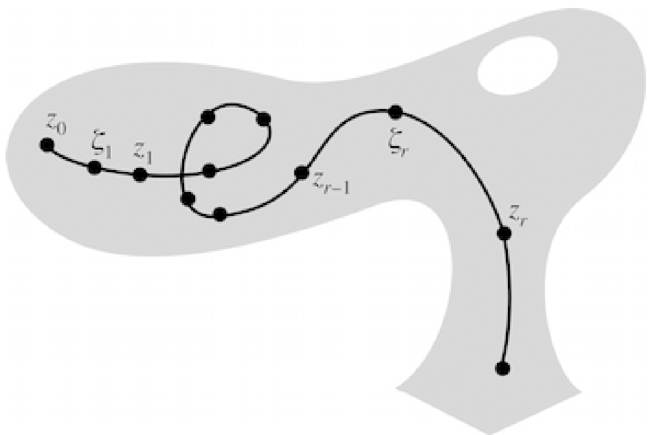

Figure 6.1 A partition of a path in $\mathbb{C}$.

Writing $z_r = \gamma(t_r), \zeta_r = \gamma(s_r)$, this Riemann–Stieltjes sum becomes

$$S(P,f,\gamma) = \sum_{r=1}^{n} f(\zeta_r)(z_r - z_{r-1})$$

which exhibits the direct analogy with the real case.

As in the real case, we rarely calculate an integral using this summation process. One method is to reduce the calculation to two real integrals by taking real and imaginary parts. If $\psi : [a,b] \to \mathbb{R}$, let

$$\psi(t) = U(t) + \mathrm{i}\,V(t) \quad (t \in [a,b])$$

and define

$$\int_\gamma \psi(t)\mathrm{d}t = \int_a^b U(t)\mathrm{d}t + \mathrm{i}\int_a^b V(t)\mathrm{d}t$$

With this convention we derive a complex version of Lemma 6.4:

THEOREM 6.6. *For a continuous complex function f defined on a domain D, and a smooth path $\gamma : [a,b] \to D$,*

$$\int_\gamma f = \int_a^b f(\gamma((t))\gamma'(t)\mathrm{d}t \tag{6.3}$$

Proof. Let $\gamma(t) = x(t) + \mathrm{i}y(t)$ $(t \in [a,b])$ and $f(z) = u(x,y) + \mathrm{i}v(x,y)$ $(z = x + \mathrm{i}y \in D)$. Denoting $u(x(t),y(t)), v(x(t),y(t)), x'(t), y'(t)$ by u, v, x', y' for short, our convention for complex integrals gives

$$\begin{aligned}\int_a^b f(\gamma((t))\gamma'(t)\mathrm{d}t &= \int_a^b (u + \mathrm{i}v)(x' + \mathrm{i}y')\mathrm{d}t \\ &= \int_a^b (ux' - vy') + \mathrm{i}(vx' + uy')\mathrm{d}t \\ &= \int_a^b ux'\mathrm{d}t - \int_a^b vy'\mathrm{d}t + \mathrm{i}\int_a^b vx'\mathrm{d}t + \mathrm{i}\int_a^b uy'\mathrm{d}t\end{aligned}$$

If we write $f(\gamma(s_r)) = u_r + \mathrm{i}v_r$ and $\gamma(t_r) = x_r + \mathrm{i}y_r$, then

$$\begin{aligned}S(P,f,\gamma) &= \sum_{r=1}^{n}(u_r + \mathrm{i}v_r)[(x_r + \mathrm{i}y_r) - (x_{r-1} + \mathrm{i}y_{r-1})] \\ &= \sum_{r=1}^{n} u_r(x_r - x_{r-1}) - \sum_{r=1}^{n} v_r(y_r - y_{r-1}) \\ &\quad + \mathrm{i}\sum_{r=1}^{n} v_r(x_r - x_{r-1}) + \mathrm{i}\sum_{r=1}^{n} u_r(y_r - y_{r-1})\end{aligned}$$

We now match the four integrals and four sums in pairs, and use Lemma 6.4. For instance,

$$\sum_{r=1}^{n} u_r(x_r - x_{r-1}) = \sum_{r=1}^{n} u(x(s_r), y(s_r))(x(t_r) - x(t_{r-1}))$$

where both $\phi(t) = u(x(t), y(t))$ and $x'(t)$ are continuous on $[a, b]$. So given $\varepsilon > 0$, we can find a partition $P_1(\varepsilon)$ such that for any partition P that is finer than $P_1(\varepsilon)$ we have

$$\left| \sum_{r=1}^{n} u_r(x_r - x_{r-1}) - \int_a^b ux' \mathrm{d}t \right| < \frac{\varepsilon}{4}$$

Similarly we find partitions $P_2(\varepsilon), P_3(\varepsilon), P_4(\varepsilon)$ such that all finer partitions P give similar inequalities between corresponding pairs of integrals and sums. Taking P_ε to have division points all those of $P_1(\varepsilon), P_2(\varepsilon), P_3(\varepsilon), P_4(\varepsilon)$, then whenever P is finer than P_ε all four inequalities hold. Therefore

$$\left| S(P, f, \gamma) - \int_a^b f(\gamma((t)))\gamma'(t)\mathrm{d}t \right| < \frac{\varepsilon}{4} + \frac{\varepsilon}{4} + \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \varepsilon$$

Hence

$$\int_\gamma f = \int_a^b f(\gamma((t)))\gamma'(t)\mathrm{d}t$$

□

Example 6.7. $f(z) = z^2, \gamma(t) = t^2 + \mathrm{i}t \ (t \in [0, 1])$. See Figure 6.2.

$$\begin{aligned}
\int_\gamma f &= \int_0^1 f(\gamma(t))\gamma'(t)\mathrm{d}t \\
&= \int_0^1 (t^2 + \mathrm{i}t)^2(2t + \mathrm{i})\mathrm{d}t \\
&= \int_0^1 (t^4 + 2\mathrm{i}t^3 - t^2)(2t + \mathrm{i})\mathrm{d}t \\
&= \int_0^1 (2t^5 - 4t^3)\mathrm{d}t + \mathrm{i}\int_0^1 (5t^4 - t^2)\mathrm{d}t \\
&= [\tfrac{1}{3}t^6 - t^4]_0^1 + \mathrm{i}[t^5 - \tfrac{1}{3}t^3]_0^1 \\
&= -\tfrac{2}{3} + \tfrac{2}{3}\mathrm{i}
\end{aligned}$$

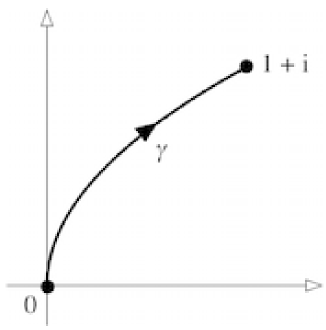

Figure 6.2 Path for Example 6.7.

6.3 The Length of a Path

The length of a path is defined in terms of approximating polygons, as follows.

DEFINITION 6.8. Given an arbitrary path $\gamma : [a, b] \to \mathbb{C}$, take a partition P of $[a, b]$ given by $a = t_0 < t_1 < \cdots < t_n = b$, and calculate the length

$$L(\gamma_P) = \sum_{r=1}^{n} |\gamma(t_r) - \gamma(t_{r-1})|$$

of the approximating polygonal curve γ_P with vertices $\gamma(t_0), \gamma(t_1), \ldots, \gamma(t_n)$, Figure 6.3.

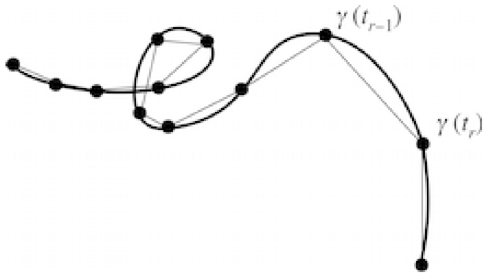

Figure 6.3 Length of a path as the supremum of the lengths of approximating polygons.

The *length* $L(\gamma)$ of γ is the supremum (least upper bound) of the lengths $L(\gamma_P)$ over all such approximating polygons. □

The length of an arbitrary path need not be finite:

Example 6.9. Suppose $\gamma(t)$ ($t \in [0, 1]$) is defined as follows. Let m_n be the midpoint of $[1/(n + 1), 1/n]$, so $m_n = (2n + 1)/(2n(n + 1))$. Introduce a function λ where the graph of $y = \lambda(t)$ consists of two straight line segments from $(\frac{1}{n+1}, 0)$ to $(m_n, \frac{1}{n})$ and then to $(\frac{1}{n}, 0)$. Define

$$\gamma(t) = \begin{cases} t + i\lambda(t) & (0 < t \leq 1) \\ 0 & (t = 0) \end{cases}$$

The image of γ is drawn in Figure 6.4.

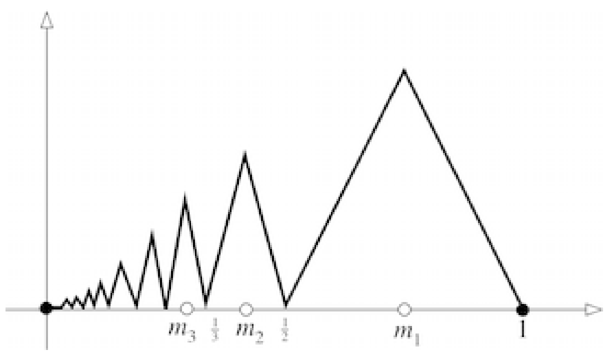

Figure 6.4 A path with infinite length.

The length of this path from $t = \frac{1}{n+1}$ to $t = \frac{1}{n}$ exceeds $\frac{2}{n}$. Let P be the partition $0 < \frac{1}{n+1} < m_n < \frac{1}{n} < \cdots < \frac{1}{2} < m_1 < 1$. Then the length of the polygonal path exceeds

$$\frac{2}{n} + \frac{2}{n-1} + \cdots + \frac{2}{1} = 2\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}\right)$$

The latter is twice the harmonic series, so it increases without limit as n increases. Therefore $L(\gamma)$ is infinite. □

In this example, the path is not smooth. The next example shows that even if the path is 'nearly smooth', its length can still be infinite.

Example 6.10. Consider the path

$$\gamma(t) = \begin{cases} t + \mathrm{i}t\sin(\pi/t) & (0 < t \leq 1) \\ 0 & (t = 0) \end{cases}$$

as in Figure 6.5.

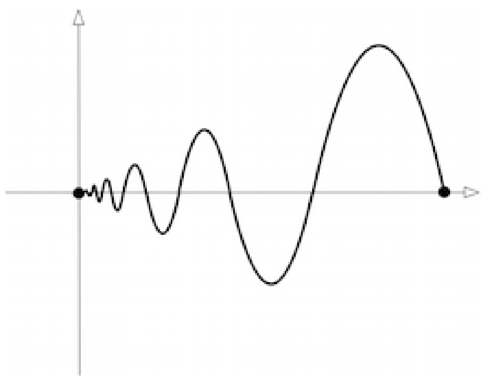

Figure 6.5 A nearly smooth path with infinite length.

A calculation shows that although

$$\gamma'(t) = 1 + \mathrm{i}(\sin(\pi/t) + t\cos(\pi/t))(-\pi/t^2)$$

is continuous for $0 < t \leq 1$, the limit of

$$\frac{\gamma(t) - \gamma(0)}{t} = 1 + \mathrm{i}\sin(\pi/t)$$

as t tends down to 0 does not exist. Hence γ' is not continuous on the closed interval $[0, 1]$, so is not smooth on $[0, 1]$ (though it *is* smooth on any subinterval $[k, 1]$ with $0 < k \leq 1$).

We claim this path has infinite length. Let P_n be the partition

$$0 < 1/n < 1/(n - \tfrac{1}{2}) < 1/(n-1) < \cdots < 1/2 < 1/(2 - \tfrac{1}{2}) < 1$$

Since

$$\gamma\left(\frac{1}{n}\right) = \frac{1}{n} + \mathrm{i}\frac{\sin n\pi}{n} = 1/n$$

and

$$\begin{aligned}\gamma\left(\frac{1}{n-\frac{1}{2}}\right) &= \frac{1}{n-\frac{1}{2}} + \mathrm{i}\frac{\sin(n-\frac{1}{2})\pi}{n-\frac{1}{2}} \\ &= \frac{1}{n-\frac{1}{2}} + \mathrm{i}\frac{(-1)^{n-1}}{n-\frac{1}{2}}\end{aligned}$$

the distance from $\gamma(1/n)$ to $\gamma(1/(n-\frac{1}{2}))$ exceeds $1/(n-\frac{1}{2})$, which exceeds $1/n$. We need only compute the lengths of alternate segments of the polygonal approximation γ_n to γ given by P_n to find that

$$L(\gamma_n) > \frac{1}{n} + \frac{1}{n-1} + \cdots + 1$$

which again increases without limit. □

6.3.1 Integral Formula for the Length of Smooth Paths and Contours

When a path is *smooth*, its length is finite, and can be calculated by an integral:

PROPOSITION 6.11. *The length of a smooth path $\gamma : [a, b] \to \mathbb{C}$ is*

$$L(\gamma) = \int_a^b |\gamma'(t)|\mathrm{d}t \tag{6.4}$$

and this is finite.

We prove this result below, but first we note that the integrand $|\gamma'(t)|$ is continuous on $[a, b]$, so the real integral

$$L = \int_a^b |\gamma'(t)|\mathrm{d}t$$

certainly exists, and is finite. We must show that L is the supremum of the lengths $L(\gamma_P)$ of approximating polygons γ_P.

Now L can be closely approximated by sums

$$S(P, \phi) = \sum_{r=1}^{n} |\gamma'(s_r)|(t_r - t_{r-1})$$

where P is the partition $a = t_0 < t_1 < \cdots < t_n = b$, $t_{r-1} < s_r < t_r$, and $\phi(t) = |\gamma'(t)|$.

The proof of Proposition 6.11 is greatly facilitated by:

LEMMA 6.12. *With the preceding notation, given any $\varepsilon > 0$, there exists a partition Q_ε such that for any partition P finer than Q_ε, the length $L(\gamma_P)$ of the approximating polygon corresponding to P satisfies*

$$|S(P,\phi) - L(\gamma_P)| < \varepsilon$$

Proof. By definition,

$$S(P,\phi) = \sum_{r=1}^{n} |\gamma'(s_r)|(t_r - t_{r-1})$$
$$L(\gamma_P) = \sum_{r=1}^{n} |\gamma(t_r) - \gamma(t_{r-1})|$$

Writing $\gamma(t) = x(t) + \mathrm{i}y(t)$, the Mean Value Theorem in real analysis gives

$$x(t_r) - x(t_{r-1}) = x'(\sigma_r)(t_r - t_{r-1}) \quad \text{for some } \sigma_r \in (t_{r-1}, t_r)$$
$$y(t_r) - y(t_{r-1}) = y'(\tau_r)(t_r - t_{r-1}) \quad \text{for some } \tau_r \in (t_{r-1}, t_r)$$

so

$$\gamma(t_r) - \gamma(t_{r-1}) = (x'(\sigma_r) + \mathrm{i}y'(\tau_r))(t_r - t_{r-1})$$

Now x', y' are continuous on $[a, b]$, and from real analysis they are *uniformly* continuous since $[a, b]$ is a closed interval. This means that for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$s,t \in [a,b] \quad \text{and} \quad |s-t| < \delta \quad \text{implies}$$
$$|x'(s) - x'(t)| < \varepsilon/2(b-a) \quad \text{and} \quad |y'(s) - y'(t)| < \varepsilon/2(b-a)$$

(where the same δ works for all s, t).

If we choose Q_ε to be *any* partition such that each subinterval $[t_{r-1}, t_r]$ has length less than δ, then any partition P finer than Q_ε has the same property. Because $s_r, \sigma_r, \tau_r \in [t_{r-1}, t_r]$, they are all within distance δ of each other, so

$$\begin{aligned}
&\big|\,|\gamma'(s_r)|(t_r - t_{r-1}) - |\gamma(t_r) - \gamma(t_{r-1})|\,\big| \\
&= \big|\,|\gamma'(s_r)| - |x'(\sigma_r) + \mathrm{i}y'(\tau_r)|\,\big|\,(t_r - t_{r-1}) \\
&\le |(x'(s_r) + \mathrm{i}y'(s_r)) - (x'(\sigma_r) + \mathrm{i}y'(\tau_r))|\,(t_r - t_{r-1}) \\
&\le (|x'(s_r) - x'(\sigma_r)| + |y'(s_r) - y'(\tau_r)|)(t_r - t_{r-1}) \\
&< (\frac{\varepsilon}{2(b-a} + \frac{\varepsilon}{2(b-a})(t_r - t_{r-1}) \\
&= \varepsilon(t_r - t_{r-1})/(b-a)
\end{aligned}$$

where in the third line we use property (1.12) of the modulus.

Thus

$$\begin{aligned}
|S(P,\phi) - L(\gamma_P)| &\le \sum_{r=1}^{n} \big|\,|\gamma'(s_r)|(t_r - t_{r-1}) - |\gamma(t_r) - \gamma(t_{r-1})|\,\big| \\
&< \sum_{r=1}^{n} \frac{\varepsilon}{b-a}(t_r - t_{r-1})
\end{aligned}$$

$$= \frac{\varepsilon}{b-a}\sum_{r=1}^{n}(t_r - t_{r-1})$$
$$= \varepsilon$$

as required. □

We are now prepared for the following:

Proof of Proposition 6.11. Let $L(\gamma)$ be as in (6.4). By the definition of the Riemann integral, there exists a partition P such that for any partition P finer than P_ε,

$$|S(P, \phi) - L| < \varepsilon \tag{6.5}$$

Adding more division points to P_ε as necessary to obtain a finer partition Q_ε with each subinterval of length less than δ, then for P finer than Q_ε, Lemma 6.12 gives

$$|S(P, \phi) - L(\gamma_P)| < \varepsilon \tag{6.6}$$

Combining this with (6.5) gives

$$L - 2\varepsilon < L(\gamma_P) < L + 2\varepsilon \tag{6.7}$$

Thus for any positive k we can find an approximating polygon γ_P with

$$L - k < L(\gamma_P) \tag{6.8}$$

Given any partition Q of $[a, b]$ with approximating polygon κ, the addition of further vertices to κ can only increase its length, by the triangle inequality. Thus if P is finer than Q then $L(\kappa) \le L(\gamma_P)$. We choose P finer than Q_ε so that (6.7) holds; then

$$L(\kappa) \le L(\gamma_P) < L + 2\varepsilon$$

But ε is any positive number, so for *any* approximating polygon κ,

$$L(\kappa) \le L \tag{6.9}$$

The inequalities (6.8) and (6.9) exhibit L as the supremum of the lengths of all approximating polygons, completing the proof. □

Example 6.13. The standard straight line path $\gamma = [z_1, z_2]$

$$\gamma(t) = z_1(1-t) + z_2 t \quad (t \in [0,1])$$

has length

$$L(\gamma) = \int_0^1 |\gamma'(t)|\mathrm{d}t = \int_0^1 |z_2 - z_1|\mathrm{d}t = |z_2 - z_1|$$

as expected. □

Example 6.14. The circle $S(t) = z_0 + r\mathrm{e}^{\mathrm{i}t}$ $(t \in [0, 2\pi])$, centre z_0 and radius $r > 0$, has length

$$L(S) = \int_0^{2\pi} |\mathrm{i}r\mathrm{e}^{\mathrm{i}t}| = \int_0^{2\pi} r\mathrm{d}t = 2\pi r$$ □

6.4 If You Took the Short Cut . . .

Some readers will have read the previous three sections, and some will not. For the latter, we now give some basic definitions that are motivated by the previous three sections:

DEFINITION 6.15. A path $\gamma : [a, b] \to \mathbb{C}$ is *smooth* if γ' exists and is continuous on the closed interval $[a, b]$.

If D is a domain, the *integral* of a continuous function $f : D \to \mathbb{C}$ along the path γ is

$$\int_\gamma f = \int_a^b f(\gamma(t))\gamma'(t)\mathrm{d}t \tag{6.10}$$

The *length* of γ is

$$L(\gamma) = \int_a^b |\gamma'(t)|\mathrm{d}t \tag{6.11}$$

□

A comment on (6.10) and (6.11) is in order to explain what they mean. The integrands $f(\gamma(t))\gamma'(t)$ and $|\gamma'(t)|$ are both continuous. The latter is real; the former may be written in real and imaginary parts as $f(\gamma(t))\gamma'(t) = U(t) + \mathrm{i}V(t)$ and the integral $\int_\gamma f$ can then be calculated using two real integrals:

$$\int_\gamma f = \int_a^b U(t)\mathrm{d}t + \mathrm{i}\int_a^b V(t)\mathrm{d}t$$

Since all integrands involved are continuous, the real integrals all exist.

At this stage, readers who made either choice are now equipped to proceed to the rest of the book. If you took the short cut, take a quick look at Examples 6.13 and 6.14 to convince yourself that the formula defining length makes sense for lines and circles.

6.5 Further Properties of Lengths

Recall from Definition 2.27 that paths can be added together if their start and end points match up correctly. Using approximating polygons, it is easy to prove that the length function L is additive:

PROPOSITION 6.16. *If $\gamma = \gamma_1 + \cdots + \gamma_n$ and $L(\gamma_r)$ exists for $1 \le r \le n$, then*

$$L(\gamma) = L(\gamma_1) + \cdots + L(\gamma_n)$$

□

It is important to understand that the length of a *path* need not be the same as the length of its image curve. In particular, this happens whenever the path 'doubles back on itself', as in the following example.

Example 6.17. Let $\sigma : [-2, 2] \to \mathbb{C}$ be any smooth path. Let $\rho : [-2, 2] \to [-2, 2]$ be given by $\rho(t) = t^3 - 3t$, which is a smooth map, and let $\gamma = \sigma \circ \rho$. Then it can be shown, using (6.4), that $L(\gamma) = 3L(\sigma)$. The geometric significance of this result is straightforward: as t increases from -2 to 2, the cubic $t^3 - 3t$ increases from -2 to 2, then decreases from 2 to -2, and finally increases from -2 to 2 again. So the *path* traces the *curve* three times. □

Next, we consider how a change of parameter affects the length of a smooth path.

DEFINITION 6.18. Smooth paths $\gamma : [a,b] \to \mathbb{C}$, $\lambda : [c,d] \to \mathbb{C}$ with the same image curve C are *smoothly equivalent parametrisations* of C if there is a smooth function $\rho : [a,b] \to [c,d]$ with non-zero derivative $\rho'(t) \neq 0$ for $t \in [a,b]$, where $\rho(a) = c$, $\rho(b) = d$, and $\gamma = \lambda \circ \rho$. □

This relationship is an equivalence relation, because ρ^{-1} exists, is smooth, and also has non-zero derivative, by the inverse function theorem. A smoothly equivalent parametrisation can be imagined as tracing the same curve in the same direction but at different speeds.

PROPOSITION 6.19. *If two smooth paths $\gamma : [a,b] \to \mathbb{C}$, $\lambda : [c,d] \to \mathbb{C}$ are smoothly equivalent parametrisations of the same curve C, then they have the same length: $L(\gamma) = L(\lambda)$.*

Proof. We have $\gamma = \lambda \circ \rho$ where $\rho : [a,b] \to [c,d]$ is a strictly increasing real function with continuous derivative ρ' on $[a,b]$. Therefore $\rho'(t) \geq 0$, so $\rho'(t) = |\rho'(t)|$. Let $s = \rho(t)$. Then as t increases from a to b, s increases from c to d and we can substitute $\mathrm{d}s = \rho'(t)\mathrm{d}t$ in the integral. The length of the path γ is therefore

$$\begin{aligned}
L(\gamma) &= \int_a^b |\gamma'(t)|\mathrm{d}t \\
&= \int_a^b |(\lambda \circ \rho)'(t)|\mathrm{d}t \\
&= \int_a^b |(\lambda'(\rho(t))||\rho'(t)|\mathrm{d}t \\
&= \int_a^b |(\lambda'(\rho(t))|\rho'(t)\mathrm{d}t \\
&= \int_a^b |\lambda'(s)|\mathrm{d}s = L(\lambda)
\end{aligned}$$

□

6.5.1 Lengths of More General Paths

You may be wondering why we require γ to be smooth in the definition of length in Definition 6.15? Proposition 6.16 applies to contours made from several smooth pieces, which need not fit together smoothly where they join, so paths that are not smooth can have meaningful lengths. Obviously, Definition 6.15 cannot be used for continuous paths, because it involves the derivative of γ. But the definition 'supremum of lengths of all approximating polygons', which occurs in the alternative treatment of Section 6.3, makes sense for any path, without assuming smoothness. However, this definition has its own awkward features. Sometimes it goes wrong in the mild sense that the entire path has infinite length, as in Examples 6.9 and 6.10. However, it gets much worse. Using the 'approximating polygon' definition for length, there are continuous paths $\gamma : [a, b] \to \mathbb{C}$ with the disturbing property that the length of *any segment* of the path, between points $c < d$ in $[a, b]$, is infinite.

In fact, we have already met just such a path: the graph of the blancmange function, Figure 4.3. Using dyadic rationals, it is straightforward to construct a sequence of partitions of $[c, d]$ such that the corresponding polygons have lengths tending to infinity. In fact, it is enough to do this for the graph of the blancmange function on the interval $[0, 1]$, because we have already pointed out that this graph contains arbitrarily small copies of the graph of the blancmange function, and a small number times infinity is still infinity. These copies are usually distorted by an affine transformation, but such a distortion keeps the length infinite. Similar remarks apply to space-filling curves, and to standard fractals such as the snowflake curve, Mandelbrot [13], and Falconer [4].

In other words, a path that is continuous but not smooth may not have a meaningful length. This is probably counterintuitive, because we have been trained from an early age to assume that every linear object does have a length. The ancient Greeks worried about the length of the circumference of a circle without ever defining what they were worrying about. It seemed obvious to them that any arc of a circle must have a length; their main aim was to find out what that length is. To do so, they tacitly assumed several plausible properties, such as finding the length by using polygons to approximate the circumference. Eventually mathematicians tightened up the logic, leading to Definition 6.8 above, and understood what can go wrong.

To complete the story, smoothness is not actually necessary for a path or curve to have a well-defined length. There is a more general notion of a rectifiable curve, for which the 'approximating polygon' definition is entirely satisfactory. However, the smooth case is all we need in this book.

6.6 Regular Paths and Curves

Just as we distinguish between a path γ and its image curve, we must distinguish between the derivative $\gamma'(t)$ and a tangent *line* to the curve. The derivative can be interpreted as the velocity vector at time t for a point $\gamma(t)$ moving along the curve. If $\gamma'(t) \neq 0$ it defines a tangent direction, hence a tangent line to the curve. When $\gamma'(t) = 0$ it does not defines a tangent direction, so the curve may not have a tangent line. Section 6.7 shows some of the things that can then happen.

First, some standard terminology:

DEFINITION 6.20. Let $\gamma : [a, b] \to \mathbb{C}$ be a smooth path.

If $t_0 \in [a, b]$ and $\gamma'(t_0) \neq 0$, then t_0 is a *regular point* of γ.

If $t_0 \in [a, b]$ and $\gamma'(t_0) = 0$, then t_0 is a *singular point* of γ. □

When the image curve has a well-defined tangent line, it *looks* smooth: see Proposition 6.22 below.

The above discussion leads naturally to a special type of path or curve that will be useful as we proceed, to relate the abstract theory to geometric intuition:

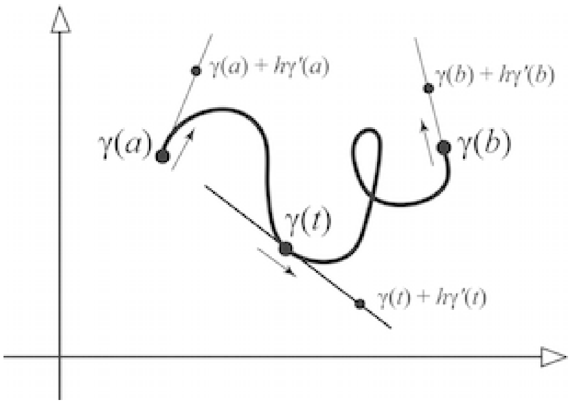

Figure 6.6 Tangents to a smooth path.

DEFINITION 6.21. A *regular path* is a smooth path $\gamma : [a,c] \to \mathbb{C}$ such that $\gamma'(t) \neq 0$ for all $t \in [a,b]$.

That is, every point on the path is a regular point.

A *regular curve* is the image of a regular path. □

If γ is regular, then by Proposition 4.18 a point on the tangent at $\gamma(t)$ is of the form $\gamma(t) + h\gamma'(t)$ for any $h \in \mathbb{R}$, Figure 6.6.

The standard paths $L(t)$ (line) and $C(t)$ (circle) in Section 2.4 are regular.

In Figure 6.6 the tangent line at $\gamma(t)$ is a good approximation to the curve given by the image of γ, near that point. To formalise this idea, we compare the path $\gamma(t)$ for t near some point $t_0 \in [a,b]$ with the corresponding tangent line. We can think of the tangent line as a path τ in its own right, defined by

$$\tau(t_0 + h) = \gamma(t_0) + h\gamma'(t_0) \quad (h \in \mathbb{R}) \tag{6.12}$$

and compare it with

$$\gamma(t_0 + h) \quad (t \text{ near } t_0)$$

We now show that when h is small, these two paths, treating h as a parameter, are very close together for any given choice of h:

PROPOSITION 6.22. *Let $\gamma : [a,b] \to \mathbb{C}$ be a regular path. As $h \to 0$, the expression*

$$\frac{1}{h}[\gamma(t_0 + h) - \tau(t_0 + h)]$$

tends to 0*.*

Before giving the simple proof, we interpret this result: it says that for sufficiently small h, the difference between the path γ and the tangent path τ tends to zero *faster than h does*.

Proof. Since γ is differentiable at t_0,

$$\lim_{h \to 0} \frac{\gamma(t_0 + h) - \gamma(t_0)}{h} = \gamma'(t_0)$$

Rewrite this as

$$\lim_{h\to 0}\frac{1}{h}[\gamma(t_0+h)-(\gamma(t_0)+h\gamma'(t_0))]=0$$

By (6.12) this is the same as

$$\lim_{h\to 0}\frac{1}{h}[\gamma(t_0+h)-\tau(t_0+h)]=0 \qquad \square$$

The proof does not use $\gamma'(t_0)\neq 0$; that is, the limit is still zero even when there is a singularity at t_0. However, in that case $\tau(t_0+h)=\gamma(t_0)$ for all $h\in\mathbb{R}$. That is, τ does not define a tangent line. So the result tells us nothing about the shape of the image of γ near t_0.

6.6.1 Parametrisation by Arc Length

Proposition 6.22 is a formal statement of the intuitive idea that a regular curve looks smooth near any point. It has a continuously turning tangent and a well-defined finite length. These properties are inherited by subpaths, leading to:

DEFINITION 6.23. Let $\gamma : [a,b]\to\mathbb{C}$ be a regular path, with image curve C. Let $t_0,t_1\in[a,b]$ with $t_0\le t_1$, and let $\gamma(t_0)=c,\gamma(t_1)=d$. Then the *arc length* $L_C(c,d)$ from c to d in C is the length of $\gamma|_{[t_0,t_1]}$; that is,

$$L_C(c,d)=\int_{t_0}^{t_1}|\gamma'(t)|\,\mathrm{d}t \qquad \square$$

We now prove that a regular curve can be smoothly reparametrised so that the parameter t is arc length, or a constant multiple of arc length if that is more convenient. Let the length of γ be L. Define $\lambda : [a,b]\to[0,L]$ by

$$\lambda(s)=L_C(a,s)=\int_a^s|\gamma'(t)|dt$$

Then

$$\lambda'(t)=|\gamma'(t)|\neq 0$$

so λ is a strictly increasing function on $[a,b]$ with a continuous derivative λ' on $[a,b]$, where $\lambda(a)=0$, $\lambda(b)=L$. It is regular since $\lambda'(t)\neq 0$. It therefore has a strictly increasing inverse function $\rho=\lambda^{-1}$. Now $\rho : [0,L]\to[a,b]$ and has continuous derivative

$$\rho'(t)=1/\lambda'(t)\neq 0$$

for $a<t<b$, so ρ is also regular.

The path

$$\tau=\gamma\circ\rho : [0,L]\to\mathbb{C}$$

is regular, and

$$\tau(\lambda(t))=\gamma\circ\rho\circ\lambda(t)=\gamma(t)$$

so τ is a reparametrisation of γ, with parameter $\lambda(t) = L_C(a,t)$. So τ reparametrises the image curve C of γ by arc length. Moreover,

$$|\tau'(t)| = |\gamma'(t)||\rho'(t)| = |\gamma'(t)|/|\lambda'(t)| = |\gamma'(t)|/|\gamma'(t)| = 1$$

So as t varies, $\tau(t)$ moves along the path with unit speed.

Multiplying the parameter by a non-zero constant scales the speed 1 to any non-zero constant speed, corresponding to a parameter that is a constant multiple of arc length. Any curve parametrised by arc length or such a multiple is regular.

6.7 Regular and Singular Points

Before tackling the intricacies of contour integration, we explain why the cases $\gamma'(t) \neq 0$ and $\gamma'(t) = 0$ differ significantly.

Near a regular point, the image of γ is a smooth curve in the geometric sense that it has a well-defined tangent direction (and this varies continuously). The curve may cross itself, but each separate segment near the crossing looks smooth.

Near a singular point, this may not be true. Sometimes there is a sensible, indeed visible, tangent direction, but sometimes there is not. The geometry of $\gamma(t)$ when t is near a singular point t_0 is highly sensitive to the precise behaviour of $\gamma(t)$ when t is near t_0. We give a series of simple examples to illustrate some of the possibilities.

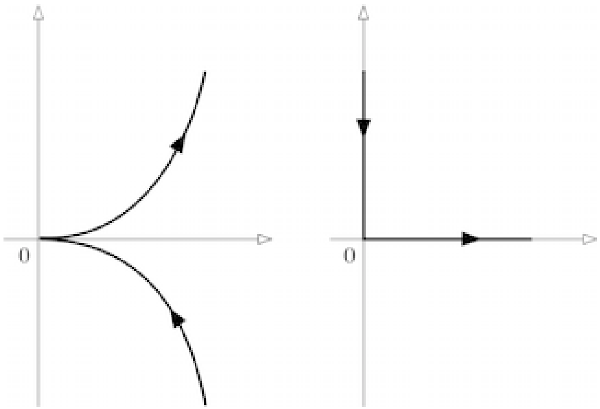

Figure 6.7 Examples of possible behaviour of a smooth path near a singular point. *Left*: Cusp. *Right*: Right-angled corner.

Example 6.24. Suppose that $\gamma : [-1,1] \to \mathbb{C}$ where $\gamma(t) = t^2 + it^3$. The derivative is $\gamma(t) = 2t + 3it^2$, which is non-zero except at $t = 0$. The image is a semicubical parabola with a cusp at the origin, Figure 6.7 (left). Everywhere else, the image is a smooth curve. The image has a natural tangent *line* at the cusp point, along the real axis, because this is the limiting direction of tangents near the origin. However, the velocity vector reverses direction as it passes through the origin. □

To make sense of what is happening at a singular point we compare it with the real case. In real analysis the graph of a function $f : \mathbb{R} \to \mathbb{R}$ consists of the points $(x, f(x)) \in \mathbb{R}^2$. If f is smooth, the graph has a well-defined tangent – even at a critical point where $f'(x) = 0$. However, if we imagine the point $y = f(x)$ moving as x increases steadily, y becomes stationary at any critical point. (Indeed, another term is *stationary point*.) The tangent at a critical point is horizontal. This is the same as the direction in which a point $(x, f(x))$ on the graph is projected to obtain the image $f(x)$. So the tangent projects to a single point.

The same kind of behaviour is happening at the cusp, but now γ is a hybrid function with a graph $\gamma : [-1, 1] \to \mathbb{C}$ in coordinate form $(t, t^2 + \mathrm{i}t^3) \in \mathbb{R} \times \mathbb{C}$. This is drawn in the upper left part of Figure 6.8, represented in three-dimensional real space as $(t, x, y) = (t, t^2, t^3)$, where t, x, y all lie between -1 and $+1$. Projection onto the vertical (x, y)-plane

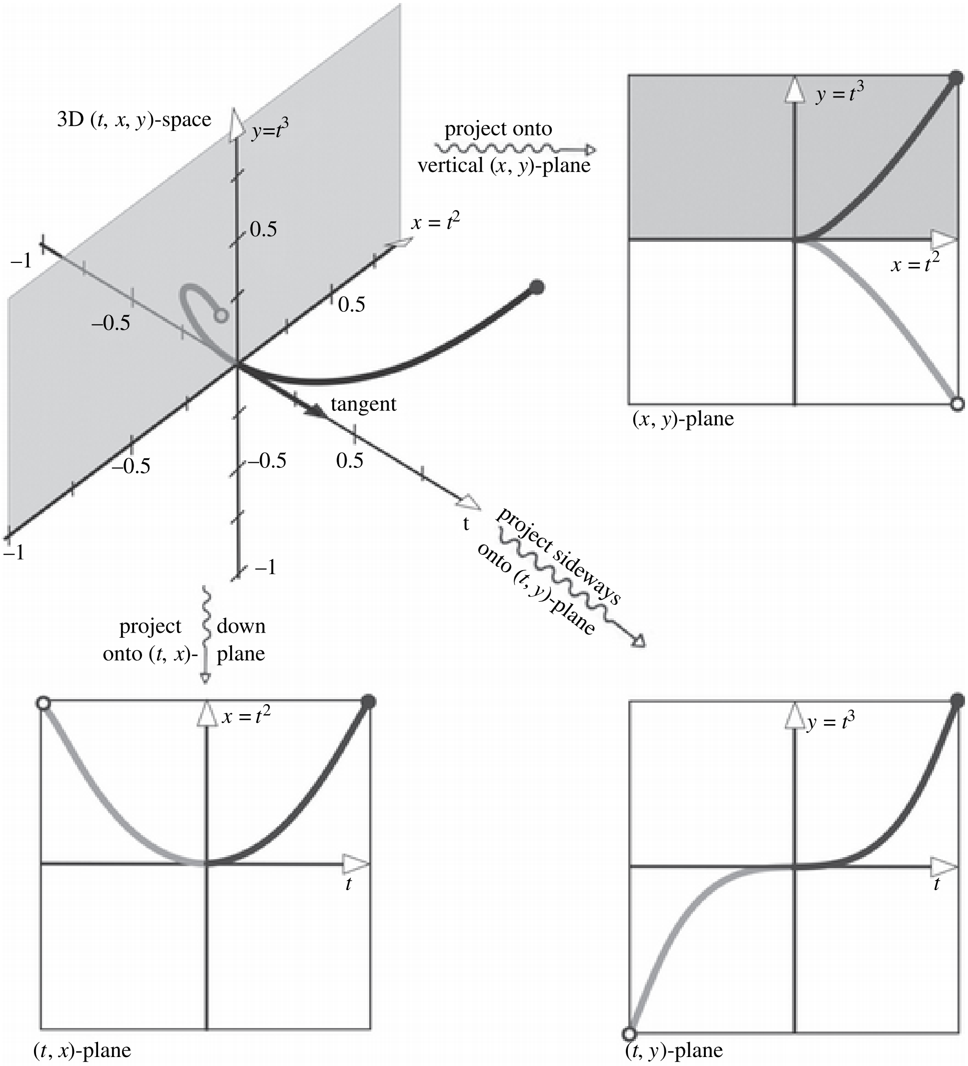

Figure 6.8 Example 6.24 represented in terms of the graph of the graph of the function $\gamma(t) = t^2 + \mathrm{i}t^3$ in $\mathbb{R} \times \mathbb{C}$ pictured in $\mathbb{R}^3$.

reveals the semicubical parabola (t^2, t^3); projection down onto the (t, x)-plane gives the parabola (t, x^2); and sideways projection onto the (t, y)-plane gives the cubic curve (t, x^3). The tangent at the origin is in the direction $(1, 2t, 3t^2) = (1, 0, 0)$, which points along the t-axis perpendicular to the complex plane and its projection onto the complex plane has length zero.

Example 6.25. More surprisingly, the image can have a sharp corner, contrary to the intuition that it ought to at least *look* smooth.

Define $\phi : \mathbb{R} \to \mathbb{R}$ by

$$\phi(t) = \begin{cases} 0 & (t \leq 0) \\ t^2 & (t > 0) \end{cases}$$

This function is continuously differentiable because both 0 and t^2 have the same derivative, mainly 0, when $t = 0$.

Now define $\gamma : [-1, 1] \to \mathbb{C}$ by

$$\gamma(t) = \phi(t) + \mathrm{i}\phi(-t)$$

The image is then as shown in Figure 6.7 (right).

The definition of ϕ can be modified to make γ infinitely differentiable, while retaining the right-angled corner. Just change t^2 to $\mathrm{e}^{-1/t}$ as in (4.8). □

Next we compare two different cases where the image of γ is a spiral.

Example 6.26. Suppose that

$$\gamma(t) = \begin{cases} 0 & (t = 0) \\ t\mathrm{e}^{\mathrm{i}/t} & (t > 0) \end{cases}$$

Then $\gamma : [0, 1] \to \mathbb{C}$ has a spiral image, Figure 6.9 (left). In this case, γ is not continuously differentiable. Indeed as $t \to 0$ from above, we have

$$\gamma'(t) = \mathrm{e}^{\mathrm{i}/t} + t.\frac{\mathrm{i}}{t}\mathrm{e}^{\mathrm{i}/t} = (1 + \mathrm{i})\mathrm{e}^{\mathrm{i}/t}$$

This goes round and round the circle of radius $|1 + \mathrm{i}| = \sqrt{2}$ at an ever-increasing rate, so it does not tend to a limit as $t \to 0$. The one-sided derivative at $t = 0$ is not defined, let alone continuous. □

Contrast this spiral with the following closely related one:

Example 6.27. Suppose that

$$\gamma(t) = \begin{cases} 0 & (t = 0) \\ t^2\mathrm{e}^{\mathrm{i}/t} & (t > 0) \end{cases}$$

Then $\gamma : [0, 1] \to \mathbb{C}$ has a spiral image, Figure 6.9 (right). The spiral is much tighter than the previous example, and now γ is continuously differentiable. Indeed as $t \to 0$

Figure 6.9 *Left*: A spiral path that is not smooth. *Right*: A smooth spiral path.

from above, we have

$$\gamma'(t) = (2 + \mathrm{i})t\mathrm{e}^{\mathrm{i}/t}$$

which tends to 0 as $t \to 0$. The one-sided derivative at $t = 0$ is defined, and the derivative is continuous there. As it happens, it is zero. □

We now consider the lengths of the spiral paths in Examples 6.26 and 6.27, which can be estimated by bounding the lengths of successive turns through 2π. In Example 6.26, turn n has length lying between two constant multiples of $1/n$. So the total length lies between the same constant multiples of $\sum_n 1/n$, the harmonic series, which is infinite. In Example 6.27, turn n has length lying between two constant multiples of $1/n^2$. Since $\sum_n 1/n^2$ is finite, this spiral has finite length.

These results are consistent with the lack of smoothness of Example 6.26 (indeed, since its length is infinite Proposition 6.11 implies that it cannot be smooth), and the smoothness of Example 6.27, which implies that its length must be finite.

6.8 Contour Integration

For readers who took the Riemann integral route, Section 6.2 defines the notion of a smooth path and deduces a formula for the integral along such a path. Those who opted for the short cut should refer to Definition 6.15. For all readers, we now generalise the notion of integration to allow paths made up of a finite number of smooth pieces:

DEFINITION 6.28. Using the notation of Section 2.4, a *contour* is a path of the form

$$\gamma = \gamma_1 + \cdots + \gamma_n$$

where $\gamma_1, \ldots, \gamma_n$ are smooth paths such that the final point of γ_r coincides with the initial point of γ_{r+1} for $r = 1, \ldots, n-1$. □

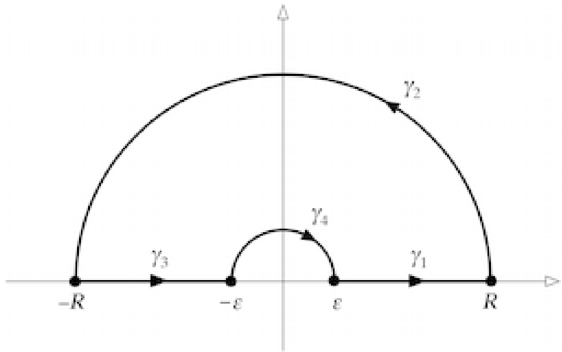

Figure 6.10 The contour defined in Example 6.29.

Example 6.29. Let

$$\begin{aligned}
\gamma_1(t) &= t && (t \in [\varepsilon, R]) \\
\gamma_2(t) &= R(\cos t + \mathrm{i} \sin t) && (t \in [0, \pi]) \\
\gamma_3(t) &= t && (t \in [-R, \varepsilon]) \\
\gamma_4(t) &= \varepsilon(-\cos t + \mathrm{i} \sin t) && (t \in [0, \pi])
\end{aligned}$$

Then $\gamma = \gamma_1 + \gamma_2 + \gamma_3 + \gamma_4$ is the (closed) contour drawn in Figure 6.10. □

6.8.1 Definition of Contour Integral

Integration along a contour is an easy extension of integration along a smooth path:

DEFINITION 6.30. If D is a domain, $f : D \to \mathbb{C}$ is continuous, and $\gamma = \gamma_1 + \cdots + \gamma_n$ is a contour (so all γ_r are smooth), then the *contour integral of f along γ* is

$$\int_\gamma f = \int_{\gamma_1} f + \cdots + \int_{\gamma_n} f$$

and

$$L(\gamma) = L(\gamma_1) + \cdots + L(\gamma_n)$$ □

It is obvious that if a smooth path σ is subdivided as

$$\sigma = \sigma_1 + \sigma_2$$

then

$$\int_\sigma f = \int_{\sigma_1} f + \int_{\sigma_2} f$$

so further subdivisions of the contours $\gamma_1, \ldots, \gamma_n$ in the above definitions will not affect the values of the integrals. The contour integrals are therefore well-defined.

The following standard properties hold, analogous to the real case:

PROPOSITION 6.31.

$$\int_{\gamma_1+\gamma_2} f = \int_{\gamma_1} f + \int_{\gamma_2} f \tag{6.13}$$

$$\int_{\gamma} (f_1 + f_2) = \int_{\gamma} f_1 + \int_{\gamma} f_2 \tag{6.14}$$

$$\int_{\gamma} cf = c\int_{\gamma} f \quad (c \in \mathbb{C}) \tag{6.15}$$

$$\int_{-\gamma} f = -\int_{\gamma} f \tag{6.16}$$

Proof. Equation (6.13) follows trivially from the definitions. The proofs of (6.14) and (6.15) depend on whether the reader has read Sections 6.1–6.3. If so, then Section 6.2 implies that

$$S(P, f_1 + f_2, \gamma) = S(P, f_1, \gamma) + S(P, f_2, \gamma)$$
$$S(P, cf, \gamma) = cS(P, f, \gamma)$$

for any partition P, which imply (6.14) and (6.15).

The reader who started from Section 6.8 can verify these formulas using known properties of real integrals, as follows.

First, it is sufficient to verify them for a single smooth path. To prove (6.14), observe that

$$\begin{aligned}\int_{\gamma}(f_1 + f_2) &= \int_a^b (f_1(\gamma(t)) + f_2(\gamma(t))\gamma'(t)\mathrm{d}t \\ &= \int_a^b f_1(\gamma(t))\gamma'(t)\mathrm{d}t + \int_a^b f_2(\gamma(t))\gamma'(t)\mathrm{d}t \\ &= \int_{\gamma_1} f + \int_{\gamma_2} f\end{aligned}$$

using the additivity property for real integrals on the real and imaginary parts.

The proof of (6.15) is slightly longer (a minor penalty for skipping Sections 6.1–6.3), because we must separate everything into real and imaginary parts. Let $c = \alpha + \mathrm{i}\beta$, $f(\gamma(t))\gamma'(t) = U(t) + \mathrm{i}V(t)$. Then

$$\begin{aligned}\int_{\gamma} cf &= \int_a^b (\alpha + \mathrm{i}\beta)(U(t) + \mathrm{i}V(t))\mathrm{d}t \\ &= \int_a^b ([\alpha U(T) - \beta V(t)] + \mathrm{i}[\alpha V(t) + \beta U(t)])\mathrm{d}t \\ &= \int_a^b [\alpha U(T) - \beta V(t)]\mathrm{d}t + \mathrm{i}\int_a^b [\alpha V(t) + \beta U(t)])\mathrm{d}t \\ &= \alpha\int_a^b U(T)\mathrm{d}t - \beta\int_a^b V(t)\mathrm{d}t + \mathrm{i}\alpha\int_a^b V(t)\mathrm{d}t + \mathrm{i}\beta\int_a^b U(t)\mathrm{d}t\end{aligned}$$

$$= (\alpha + \mathrm{i}\beta) \int_a^b (U(t) + \mathrm{i}V(t))\mathrm{d}t$$
$$= c \int_\gamma f$$

using standard properties of real integrals.

Finally, we prove (6.16). If $\gamma : [a, b] \to D$ is a contour, then the opposite path $-\gamma : [a, b] \to D$, defined by

$$-\gamma(t) = \gamma(a + b - t) \quad (t \in [a, b])$$

is also a contour. Now

$$\int_{-\gamma} f = \int_a^b f(\gamma(a + b - t)) \frac{\mathrm{d}}{\mathrm{d}t} \gamma(a + b - t)\mathrm{d}t$$
$$= -\int_a^b f(\gamma(a + b - t))\gamma'(a + b - t)\mathrm{d}t$$

Substituting $s = a + b - t$ this becomes

$$-\int_b^a f(\gamma(s))\gamma'(s)(-\mathrm{d}s)$$
$$= -\int_a^b f(\gamma(s))\gamma'(s)(\mathrm{d}s)$$
$$= -\int_\gamma f$$

Therefore

$$\int_{-\gamma} f = -\int_\gamma f$$

□

6.9 The Fundamental Theorem of Contour Integration

Integrals of complex functions are only occasionally computed by breaking them down into real and imaginary parts and calculating two real integrals as in the previous section. This technique is sometimes needed, but a far more efficient method is available for $\int_\gamma f$ if we can find an antiderivative of f. This concept is so important that we give it a formal definition:

DEFINITION 6.32. Let D be a domain. An *antiderivative* for a function $f : D \to \mathbb{C}$ is a function $F : D \to \mathbb{C}$ such that $F' = f$. □

An antiderivative, if it exists, is unique up to an added constant, for if $F' = G' = f$ in a domain, then $(F - G)' = 0$ so $F - G$ is constant by Theorem 4.14. If we can find an antiderivative, then the integral can be computed immediately using:

THEOREM 6.33 (Fundamental Theorem of Contour Integration). *If $f : D \to \mathbb{C}$ is continuous, $F : D \to \mathbb{C}$ is an antiderivative of f, and γ is a contour in D from z_0 to z_1, then*

$$\int_\gamma f = F(z_1) - F(z_0)$$

Proof. Let $w(t) = u(t) + \mathrm{i}v(t), W(t) = U(t) + \mathrm{i}V(t)$ for $t \in [a, b]$, where u, v, U, V are real. Then $W' = w$ if and only if $U' = u, V' = v$, and then

$$\begin{aligned}\int_a^b w(t)\mathrm{d}t &= \int_a^b u(t)\mathrm{d}t + \mathrm{i}\int_a^b v(t)\mathrm{d}t \\ &= U(b) - U(a) + \mathrm{i}V(b) - \mathrm{i}V(a) \\ &= W(b) - W(a)\end{aligned}$$

Let $w(t) = f(\gamma(t))\gamma'(t)$. Since $F' = f$,

$$w'(t) = F'(\gamma(t))\gamma'(t) = W'(t)$$

where $W(t) = F(\gamma(t))$. Therefore

$$\begin{aligned}\int_\gamma f = \int_a^b w(t)\mathrm{d}t &= W(b) - W(a) \\ &= F(\gamma(b)) - F(\gamma(a)) = F(z_1) - F(z_0)\end{aligned}$$ □

One remarkable feature of this result is that, in these circumstances, the integral does *not* depend on the path γ, only on its end points. In contrast, we show in Section 6.10 that the integral *may* depend on the path when the conditions of Theorem 6.33 fail to hold.

Example 6.34. If $f(z) = z^2$ and γ is any contour from $z_0 = 0$ to $z_1 = 1 + \mathrm{i}$, then $F(z) = \frac{1}{3}z^3$ is an antiderivative of f, and

$$\int_\gamma z^2\mathrm{d}z = \tfrac{1}{3}(z_1^3 - z_0^3) = \tfrac{1}{3}(1 + \mathrm{i})^3 = \tfrac{2}{3} + \tfrac{2}{3}\mathrm{i}$$ □

This example gives a visibly easier calculation than that performed at the end of Section 6.2 for the particular contour $\gamma(t) = t^2 + \mathrm{i}t$ ($t \in [0, 1]$).

However, any euphoria we feel over this phenomenon must be tempered with the realisation that, unlike the real case where Theorem 6.2 (iii) tells us that a continuous function always has an antiderivative, in the complex case there are continuous functions without antiderivatives. Indeed, as we prove in Chapter 10, any function that is differentiable once in a domain is differentiable as many times as we wish. Now, any antiderivative F must be differentiable once, so it is differentiable *twice*. Therefore its derivative $f = F'$ must be differentiable. So no non-differentiable function f has an antiderivative.

Example 6.35. We know that $f(z) = |z|^2$ is differentiable only at the origin, so it is useless to search for an antiderivative when integrating this f – for example along the path $\gamma(t) = t^2 + \mathrm{i}t$ ($t \in [0,1]$). In this case, we return to the basic formula

$$\begin{aligned}\int_\gamma |z|^2 \mathrm{d}z &= \int_0^1 (t^4 + t^2)(2t + \mathrm{i})\mathrm{d}t \\ &= \int_0^1 (2t^5 + 2t^3)\mathrm{d}t + \mathrm{i}\int_0^1 (t^4 + t^2)\mathrm{d}t \\ &= [\tfrac{1}{3}t^6 + \tfrac{1}{2}t^4]_0^1 + \mathrm{i}[\tfrac{1}{5}t^5 + \tfrac{1}{3}t^3]_0^1 \\ &= \tfrac{5}{6} + \tfrac{8}{15}\mathrm{i}\end{aligned}$$

□

Fortunately, many functions do have antiderivatives. For instance, by Corollary 4.5, any polynomial

$$p(z) = a_0 + a_1 z + \cdots + a_n z^n$$

has antiderivative

$$p(z) = a_0 z + \tfrac{1}{2}a_1 z + \cdots + \tfrac{1}{n+1}a_n z^{n+1}$$

More generally, a power series has an antiderivative everywhere inside its disc of convergence:

THEOREM 6.36. *If $f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n$ converges for $|z - z_0| < R$, then*

$$F(z) = \sum_{n=0}^{\infty} \frac{a_n}{n+1}(z - z_0)^{n+1}$$

also converges for $|z - z_0| < R$, and $F' = f$.

Proof. It is sufficient to show that $F(z)$ converges for $|z - z_0| < R$, for then we may differentiate term by term by Theorem 4.20 to obtain $F' = f$.

We know that the power series $\sum_{n=0}^{\infty} a_n (z - z_0)^n$ converges absolutely for $|z - z_0| < R$, by Theorem 3.19. For each n, either $a_n = 0$ or

$$\frac{|a_n (z - z_0)^n/(n+1)|}{|a_n (z - z_0)^n|} = \frac{|z - z_0|}{n+1}$$

tends to zero as $n \to \infty$, so by the comparison test $\sum_{n=0}^{\infty} \frac{a_n}{n+1}(z - z_0)^{n+1}$ converges. □

If $f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n$ has disc of convergence D, then for any contour γ in D from z_1 to z_2,

$$\int_\gamma f = \sum_{n=0}^{\infty} \frac{a_n (z_2 - z_0)^{n+1}}{n+1} - \sum_{n=0}^{\infty} \frac{a_n (z_1 - z_0)^{n+1}}{n+1}$$

In particular, for any contour γ in D from z_0 to z,

$$\int_\gamma f = \sum_{n=0}^{\infty} \frac{a_n (z - z_0)^{n+1}}{n+1}$$

6.10 An Integral that Depends on the Path

The Fundamental Theorem tells us that if a differentiable function f has an antiderivative throughout its domain D, then the integral $\int_\gamma f$ is independent of the choice of path. This applies to z^n for any integer $n \neq -1$, with antiderivative $z^{n+1}/(n+1)$ in the domain $\mathbb{C}\backslash\{0\}$. It applies to $1/z^2, 1/z^3, \ldots$ but fails for $1/z$, as we now show using formula (6.10).

Example 6.37. Consider the problem of integrating $1/z$ between -1 and 1. The path along the real axis passes through the origin, where $1/z$ is not defined, so we avoid this by choosing a semicircular path.

If the semicircle lies above the real axis (where $\operatorname{im} z \geq 0$) we can choose

$$\gamma_1(t) = \mathrm{e}^{-\mathrm{i}t} \quad (t \in [\pi, 2\pi])$$

noting that $\mathrm{e}^{-\mathrm{i}\pi} = -1, \mathrm{e}^{-\mathrm{i}2\pi} = 1$ and the $-t$ makes the path run clockwise. Now

$$\begin{aligned}
\int_{\gamma_1} \frac{1}{z}\mathrm{d}z &= \int_\pi^{2\pi} \frac{1}{\gamma_1(t)}\gamma_1'(t)\,\mathrm{d}t \\
&= \int_\pi^{2\pi} \frac{1}{\mathrm{e}^{-\mathrm{i}t}}(-\mathrm{i}\mathrm{e}^{-\mathrm{i}t})\,\mathrm{d}t \\
&= \int_\pi^{2\pi} -\mathrm{i}\,\mathrm{d}t \\
&= [-\mathrm{i}t]_\pi^{2\pi} \\
&= -\mathrm{i}(2\pi - \pi) \\
&= -\mathrm{i}\pi
\end{aligned}$$

All very well, but we could equally sensibly choose the semicircle that lies below the real axis (where $\operatorname{im} z \leq 0$). Now

$$\gamma_2(t) = \mathrm{e}^{\mathrm{i}t} \quad (t \in [\pi, 2\pi])$$

noting that $\mathrm{e}^{\mathrm{i}\pi} = -1, \mathrm{e}^{\mathrm{i}2\pi} = 1$ and the $+t$ makes the path run anticlockwise. A very similar calculation yields

$$\begin{aligned}
\int_{\gamma_2} \frac{1}{z}\mathrm{d}z &= \int_\pi^{2\pi} \frac{1}{\gamma_2(t)}\gamma_2'(t)\,\mathrm{d}t \\
&= \int_\pi^{2\pi} \frac{1}{\mathrm{e}^{\mathrm{i}t}}(\mathrm{i}\mathrm{e}^{\mathrm{i}t})\,\mathrm{d}t \\
&= \int_\pi^{2\pi} \mathrm{i}\,\mathrm{d}t \\
&= [\mathrm{i}t]_\pi^{2\pi} \\
&= \mathrm{i}(2\pi - \pi) \\
&= \mathrm{i}\pi
\end{aligned}$$

which is slightly different. □

At first it might appear that we made a mistake – it's easy to get a sign wrong – but the second integral is clearly the complex conjugate of the first, because everything is reflected in the real axis. Since neither is real, they cannot be equal.

This example does not contradict Theorem 6.33 because $1/z$ does not have an antiderivative in $\mathbb{C}\backslash\{0\}$. The obvious candidate is the logarithm, but by Section 5.7 exp is periodic over $\mathbb{C}$ hence not one-one. Therefore any inverse 'function' to exp must be multivalued, as explained in detail in Section 7.3. So the logarithm is not an antiderivative over the whole of the domain.

In Chapter 8 we show that we get a different answer here because the point $z = 0$, where $1/z$ is undefined, lies inside the region bounded by γ_1 and γ_2. This behaviour is closely connected to the multivalued nature of the complex logarithm. It has surprisingly profound implications, explored in Chapters 8–12.

At any rate, this example should counteract any feeling that complex integrals are always independent of the path, Theorem 6.33 notwithstanding.

6.11 The Gamma Function

So far the special complex functions that we have encountered are all extensions to $\mathbb{C}$ of familiar real functions: polynomials, powers, trigonometric functions, the exponential and the logarithm. In this section we briefly discuss a less familiar function, the *gamma function* $\Gamma(z)$. This also began as a real function and was later extended to the complex case. It has important applications, but we restrict attention to defining the function and establishing a few basic properties. It will play a key role in Chapter 17 when we sketch connections between complex analysis and number theory, culminating in the Riemann Hypothesis.

The definition of the gamma function goes back to Euler, who initially defined it as the infinite product

$$\Gamma(z) = \frac{1}{z}\prod_{n=1}^{\infty}\left[\left(1+\frac{1}{n}\right)^{z}\left(1+\frac{z}{n}\right)^{-1}\right]$$

This infinite product converges for all $z \in \mathbb{C}\setminus\{0, -1, -2, \ldots\}$. It defines a differentiable function on that domain, but we lack the techniques needed to prove this. Instead, we use another formula due to Euler, who showed that when $\operatorname{re} z > 0$ this is equal to the integral

$$\Gamma(z) = \int_0^{\infty} t^{z-1}\mathrm{e}^{-t}\,\mathrm{d}t \tag{6.17}$$

We take this as our definition, in order to illustrate complex integration. Here we define

$$\int_0^{\infty} f = \lim_{a\to\infty}\int_{[0,a]} f$$

so the interval $[0, \infty]$ is interpreted as the non-negative real axis. It is straightforward to verify that (6.17) converges, because the factor e^{-t} tends to zero rapidly. With this definition, the formula (6.17) defines $\Gamma(z)$ only for z in the positive half-plane

$$\mathbb{C}^+ = \{z \in \mathbb{C} : \operatorname{re} z > 0\}$$

We establish some basic properties of $\Gamma(z)$ using methods from complex integration.

First, we prove that the usual method of integration by parts generalises from real analysis to the complex case:

THEOREM 6.38 (Integration by Parts). *Let $f, g : D \to \mathbb{C}$ be differentiable functions on the domain D. Then for any path $\gamma : [a, b] \to D$ we have:*

$$\int_\gamma fg' = [fg]_a^b - \int_\gamma f'g \tag{6.18}$$

Proof. By Proposition 4.4 (iii) we have

$$(fg)' = f'g + fg'$$

so

$$fg' = (fg)' - f'g$$

Now integrate both sides over γ and use the Fundamental Theorem of Contour Integration, Theorem 6.33. □

Now we prove the basic functional equation for the gamma function:

THEOREM 6.39. *If $z \in \mathbb{C}^+$ then*

$$\Gamma(z+1) = z\Gamma(z) \qquad \Gamma(1) = 1 \tag{6.19}$$

Proof. By (6.17), noting that $z + 1 \in \mathbb{C}^+$, we have

$$\Gamma(z+1) = \int_0^\infty t^z \mathrm{e}^{-t} \mathrm{d}t$$

Now use integration by parts to rewrite this as

$$\begin{aligned}
\Gamma(z+1) &= [-t^z e^{-t}]_0^\infty + \int_0^\infty z t^{z-1} \mathrm{e}^{-t} \mathrm{d}t \\
&= \lim_{x\to\infty} (-x^z \mathrm{e}^{-x} - (0\mathrm{e}^0) + z \int_0^\infty t^{z-1} \mathrm{e}^{-t} \mathrm{d}t \\
&= 0 + z \int_0^\infty t^{z-1} \mathrm{e}^{-t} \mathrm{d}t \\
&= z\Gamma(z)
\end{aligned}$$

To calculate $\Gamma(1)$, observe that

$$\begin{aligned}
\Gamma(1) &= \int_0^\infty t^{1-1} \mathrm{e}^{-t} \mathrm{d}t \\
&= [-\mathrm{e}^{-t}]_0^\infty \\
&= 0 - (-1) \\
&= 1
\end{aligned}$$

□

By induction, this formula leads immediately to a curious result:

COROLLARY 6.40. *If $n \in \mathbb{N}$ then*

$$\Gamma(n) = (n-1)! \qquad \square$$

In effect, the gamma function is a sensible way to define factorials for complex numbers, in such a manner that the basic formula $(n+1)! = (n+1)n!$ continues to hold.

6.11.1 Known Properties of the Gamma Function

We record (but do not prove) several other properties of the gamma function. One is Euler's *reflection formula*

$$\Gamma(1-z)\Gamma(z) = \frac{\pi}{\sin \pi z}$$

which is a consequence of his infinite product definition and a standard infinite product for the sine. Setting $z = \frac{1}{2}$ we obtain

$$(\Gamma(\tfrac{1}{2}))^2 = \frac{\pi}{\sin \pi/2} = \pi$$

Therefore

$$\Gamma(\tfrac{1}{2}) = \sqrt{\pi}$$

The *duplication formula* states that

$$\Gamma(z)\Gamma(z+\tfrac{1}{2}) = 2^{1-2z}\sqrt{\pi}\,\Gamma(2z)$$

The derivative of the gamma function is obtained from the integral formula (6.17) by a technique called 'differentiating under the integral sign', and leads to

$$\Gamma'(z) = \int_0^\infty t^{z-1} \mathrm{e}^{-t} \log t \,\mathrm{d}t$$

(see Chapter 7 for the complex logarithm). This also proves that Γ is a differentiable function in $\mathbb{C}^+$.

Let γ be *Euler's constant*, defined by

$$\gamma = \lim_{n\to\infty}\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} - \log n\right) \sim 0.577\,216$$

Then

$$\Gamma'(1) = -\gamma$$

Weierstrass found another infinite product that works for all $z \in \mathbb{C} \setminus \{0, -1, -2, \ldots\}$:

$$\frac{1}{\Gamma(z)} = z\mathrm{e}^{\gamma z} \prod_{n=1}^{\infty} \left[\left(1 + \frac{z}{n}\right) \mathrm{e}^{-z/n}\right]$$

It is also possible to successively extend Euler's formula to $\operatorname{re} z > -1$, $\operatorname{re} z > -2$, and so on, using (6.19) in the form

$$\Gamma(z-1) = \frac{\Gamma(z)}{z-1}$$

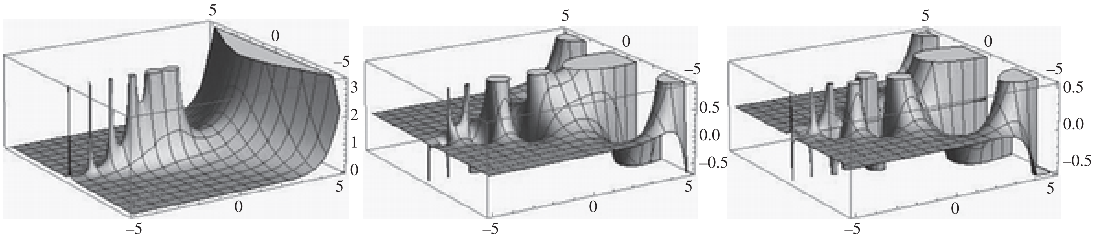

Figure 6.11 Graphs of (from left to right) the absolute value, real part, and imaginary part of the gamma function. Here x runs horizontally, y runs into the page, and the value of the function concerned runs vertically.

provided $z - 1 \neq 0$. This defines Γ on the whole of the complex plane, except for the points $\{0, -1, -2, \ldots\}$, and shows that the extended function is differentiable on its domain. It also explains why negative integers and zero must be excluded from the domain, because Γ would become infinite there; that is, have a pole. This procedure is an example of analytic continuation, which will be introduced in Chapter 14.

Graphs of $|\Gamma(z)|$, $\operatorname{re}\Gamma(z)$, and $\operatorname{im}\Gamma(z)$ are shown in Figure 6.11, for this extended domain.

The gamma function, and its poles, turned out to be an important ingredient in a radical new development in complex analysis: an unexpected link with number theory. In the hands of Riemann and his contemporaries, this led to major advances in number theory, and a conjecture that remains one of the biggest unsolved problems in mathematics, the Riemann Hypothesis (see Chapter 17).

6.12 The Estimation Lemma

We now return to more general questions about complex integration, and prove a very useful technical result. Some results in mathematics are servants of the theory, of little intrinsic interest in themselves, perhaps even a little dull, yet quietly figuring in important decisions all the time. 'Always a lemma, never a theorem'. One such result is the lemma we are about to prove. It is a simple idea, giving an upper bound for the size of an integral $|\int_\gamma f|$ in terms of an upper bound on $|f|$ and the length of γ, but it arises time and again in the theory, applying subtle pressure at critical points in the proofs of important theorems. Perhaps not a theorem of great stature, it is certainly worthy of a name. It is:

LEMMA 6.41 (Estimation Lemma). *If $f : D \to \mathbb{C}$ is continuous, γ is a contour of length L, and $|f(z)| \leq M$ for all z on γ, then*

$$\left|\int_\gamma f\right| \leq ML$$

Proof. It is poetic justice that we have to give two different proofs, depending on whether the reader has read Sections 6.1–6.3. Version A is the more natural, but it requires knowledge of Sections 6.2–6.3. For readers who took the short cut, Version B is provided.

Version A. It is sufficient to prove the result for a smooth path $\gamma : [a,b] \to D$. Let P be the partition $a = t_0 < t_1 < \cdots < t_n = b$, with $t_{r-1} \le s_r \le t_r$. Then

$$\begin{aligned}|S(P,f,\gamma)| &= \left|\sum_{r=1}^{n} f(\gamma(s_r))(\gamma(t_r) - \gamma(t_{r-1}))\right| \\ &\le \sum_{r=1}^{n} |f(\gamma(s_r))|\,|\gamma(t_r) - \gamma(t_{r-1})| \\ &\le ML(\gamma_P)\end{aligned}$$

where γ_P is the polygonal approximation to γ with vertices $\gamma(t_0), \ldots, \gamma(t_n)$. But $L(\gamma_P) \le L(\gamma)$, so

$$|S(P,f,\gamma)| \le ML(\gamma) \tag{6.20}$$

for all partitions P.

Given any $\varepsilon > 0$ we can find a partition P_ε such that

$$\left|S(P_\varepsilon,f,\gamma) - \int_\gamma f\right| < \varepsilon$$

so

$$\left|\int_\gamma f\right| < |S(P_\varepsilon,f,\gamma)| + \varepsilon$$

and (6.20) gives

$$\left|\int_\gamma f\right| < ML(\gamma) + \varepsilon$$

for all $\varepsilon > 0$. Hence

$$\left|\int_\gamma f\right| \le ML(\gamma)$$

as required.

Version B. First we establish

$$\left|\int_a^b (u(t) + \mathrm{i}v(t))\mathrm{d}t\right| \le \int_a^b |u(t) + \mathrm{i}v(t)|\mathrm{d}t \tag{6.21}$$

for continuous real functions u, v. Let $\int_a^b u(t)\mathrm{d}t = X$, $\int_a^b v(t)\mathrm{d}t = Y$. Then

$$\int_a^b (u(t) + \mathrm{i}v(t))\mathrm{d}t = X + \mathrm{i}Y$$

and

$$\begin{aligned}X^2 + Y^2 = (X - \mathrm{i}Y)(X + \mathrm{i}Y) &= \int_a^b (X - \mathrm{i}Y)(u(t) + \mathrm{i}v(t))\mathrm{d}t \\ &= \int_a^b (Xu(t) + Yv(t))\mathrm{d}t + \mathrm{i}\int_a^b (Xv(t) - Yu(t))\mathrm{d}t\end{aligned}$$

But $X^2 + Y^2 \in \mathbb{R}$, so

$$\int_a^b (Xv(t) - Yu(t))\mathrm{d}t = 0$$

and

$$X^2 + Y^2 = \int_a^b (Xu(t) + Yv(t))\mathrm{d}t$$

where the integrand $Xu(t) + Yv(t)$ is the real part of $(X - \mathrm{i}Y)(u(t) + \mathrm{i}v(t))$. So

$$\begin{aligned} Xu(t) + Yv(t) &\le |(X - \mathrm{i}Y)(u(t) + \mathrm{i}v(t))| \\ &= \sqrt{X^2 + Y^2}|u(t) + \mathrm{i}v(t)| \end{aligned}$$

From real analysis,

$$\int_a^b [Xu(t) + Yv(t)]\mathrm{d}t \le \int_a^b \sqrt{X^2 + Y^2}\,|u(t) + \mathrm{i}v(t)|\mathrm{d}t$$

which gives

$$X^2 + Y^2 \le \sqrt{X^2 + Y^2} \int_a^b |u(t) + \mathrm{i}v(t)|\mathrm{d}t$$

so that

$$\sqrt{X^2 + Y^2} \le \int_a^b |u(t) + \mathrm{i}v(t)|\mathrm{d}t \tag{6.22}$$

But

$$\sqrt{X^2 + Y^2} = |X + iY| = \left| \int_a^b (u(t) + \mathrm{i}v(t))\mathrm{d}t \right|$$

and with (6.22) this implies (6.21).

Now

$$\begin{aligned} \left| \int_\gamma f \right| &= \left| \int_a^b f(\gamma(t))\gamma'(t)\mathrm{d}t \right| \\ &\le \int_a^b |f(\gamma(t))||\gamma'(t)|\mathrm{d}t \quad \text{by (6.21)} \\ &= \int_a^b |M\gamma'(t)|\mathrm{d}t \quad \text{by real analysis} \\ &= ML \end{aligned}$$

□

The Estimation Lemma will be used in a variety of ways. We now state two applications.

PROPOSITION 6.42. *Let γ be a fixed contour and suppose that f varies.*

(i) *If the maximum value of $|f|$ on γ tends to zero, then $\int_\gamma f \to 0$.*
More generally, if $f \to f_0$, then $\int_\gamma f \to \int_\gamma f_0$.

(ii) *If the length of γ tends to zero, the image of γ tends to a point z_0, and $|f|$ remains bounded, then $\int_\gamma f \to 0$.*

In particular this is the case if f is continuous at z_0.

Proof. To prove (i), observe that as the maximum value of $|f|$ on γ tends to zero, $|\int_\gamma f| \leq ML$, which also tends to zero, so $\int_\gamma f \to 0$.

The second statements follows by considering $f - f_0$.

To prove (ii), first observe that if f is continuous at z_0, then $|f|$ is bounded in a neighbourhood of z_0. So we may assume $|f|$ is bounded on γ. Again, $|\int_\gamma f| \leq ML$, which tends to zero. □

6.13 Consequences of the Fundamental Theorem

If f is continuous and has an antiderivative F in a domain D, then we saw in Section 6.9 that, for any contour γ in D from z_0 to z_1,

$$\int_\gamma f = F(z_1) - F(z_0)$$

This result has interesting consequences. To explore them we need:

DEFINITION 6.43. A contour $\gamma : [a, b] \to \mathbb{C}$ is a *closed contour* if $\gamma(a) = \gamma(b)$. □

In real analysis, an integral $\int_a^a f$ is always zero. In the complex case, if f has an antiderivative in its domain D, then an integral $\int_\gamma f$ round a closed contour in D is always zero. (In the absence of an antiderivative, this need not be the case, as Example 6.37 shows.) Moreover, if f has an antiderivative in D and z_0, z_1 in D are different, it does not matter which contour γ from z_0 to z_1 we use to integrate f: the result is always the same.

These properties characterise the existence of an antiderivative:

THEOREM 6.44. *Let f be a continuous complex function defined on a domain D. Then the following conditions are equivalent:*

(i) *f has an antiderivative in D,*
(ii) *$\int_\gamma f = 0$ for every closed contour γ in D,*
(iii) *$\int_\gamma f$ depends only on the end points of γ for any contour γ in D.*

Proof. We have already established that (i) implies (ii) and (iii). To show that (ii) implies (iii), suppose that γ_1, γ_2 are contours in D from z_0 to z_1, as in Figure 6.12.

Then $\gamma_1 - \gamma_2$ is a closed contour, so by (ii) $\int_{\gamma_1-\gamma_2} f = 0$. But

$$\int_{\gamma_1-\gamma_2} f = \int_{\gamma_1} f - \int_{\gamma_2} f$$

so

$$\int_{\gamma_1} f = \int_{\gamma_2} f$$

which is (iii).

Finally, to show that (iii) implies (i), we fix $z_0 \in D$ and for any $z_1 \in D$ we choose a contour γ in D from z_0 to z_1 and define

$$F(z_1) = \int_{\gamma_1} f$$

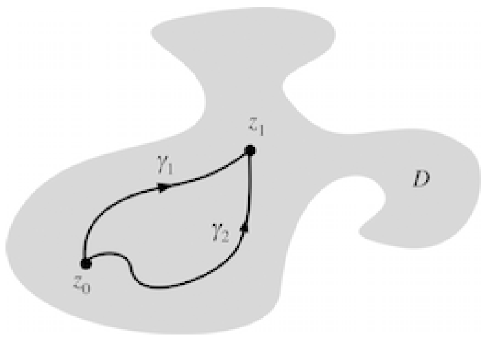

Figure 6.12 Two paths with the same end points determine a closed contour.

Because D is open, for some $\varepsilon_1 > 0$, if $|h| < \varepsilon_1$ then the line segment $\lambda(t) = z_1 + ht$ ($t \in [0, 1]$) lies in D, and

$$F(z_1 + h) = \int_{\gamma_1} f + \int_{\lambda} f$$

Thus

$$\frac{F(z_1 + h) - F(z_1)}{h} = \frac{1}{h}\int_{\lambda} f$$

For any constant c and contour γ from z_1 to z_2,

$$\int_{\gamma} c\,\mathrm{d}z = c(z_2 - z_1)$$

so

$$\int_{\lambda} \frac{f(z_1)}{h}\mathrm{d}z = f(z_1)$$

Therefore

$$\frac{F(z_1 + h) - F(z_1)}{h} - f(z_1) = \int_{\lambda} \frac{f(z) - f(z_1)}{h}\mathrm{d}z$$

Now we use the Estimation Lemma. Continuity of f implies that, given $\varepsilon > 0$, there exists $\delta > 0$ (which we may take to be less than ε_1) such that

$$|z - z_1| < \delta \text{ implies } |f(z) - f(z_1)| < \varepsilon$$

so that, when $0 < |h| < \delta$, for any z on the line segment λ the integrand satisfies

$$|(f(z) - f(z_1))/h| < \varepsilon/|h|$$

The length of λ is $|h|$, so whenever $|h| < \delta$,

$$\left|\int_{\lambda} \frac{f(z) - f(z_1)}{h}\mathrm{d}z\right| \leq \frac{\varepsilon}{|h|}|h| = \varepsilon$$

which implies

$$\left|\frac{F(z_1 + h) - F(z_1)}{h} - f(z_1)\right| \leq \varepsilon \quad (|h| < \delta)$$

But ε is arbitrary, so

$$\lim_{h\to 0}\frac{F(z_1+h)-F(z_1)}{h}=f(z_1)$$

and $F'(z_1)=f(z_1)$ for $z_1\in D$, as required. □

Example 6.45. The function $f(z) = |z|$ is not differentiable, so $\int_\gamma f$ may (and as we show, does) depend on the choice of contour γ between points. For instance, consider γ, σ from 0 to i given by

$$\gamma(t) = \mathrm{i}t \quad (t\in[0,1])$$

and $\sigma=\sigma_1+\sigma_2$ where

$$\sigma_1(t)=t \quad (t\in[0,1])$$
$$\sigma_2(t)=\mathrm{e}^{\mathrm{i}t} \quad (t\in[0,\pi/2])$$

as in Figure 6.13. Then

$$\int_\gamma |z|\mathrm{d}z=\int_0^1 |\mathrm{i}t|\cdot \mathrm{i}\mathrm{d}t=\left[\mathrm{i}t^2/2\right]_0^1=\tfrac{1}{2}\mathrm{i}$$

but

$$\int_\sigma |z|\mathrm{d}z=\int_0^1 |t|\cdot 1\ \mathrm{d}t+\int_0^{\pi/2}|\mathrm{e}^{\mathrm{i}t}|\cdot \mathrm{i}\mathrm{e}^{\mathrm{i}t}\ \mathrm{d}t=\left[t^2/2\right]_0^1+\left[\mathrm{e}^{\mathrm{i}t}\right]_0^{\pi/2}=\tfrac{1}{2}+\mathrm{i}-1=\mathrm{i}-\tfrac{1}{2}$$

□

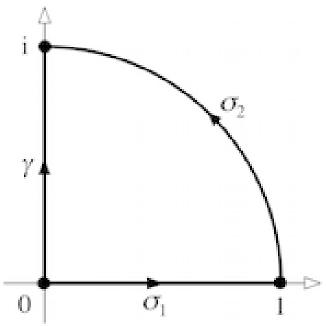

Figure 6.13 Two paths γ and $\sigma_1+\sigma_2$ that yield different values of the integral of $|z|$.

While complex functions that are merely continuous may be integrated using the basic formula, they offer little of interest. From now on, *differentiable* functions, and methods for integrating them, occupy centre stage. The resulting theory of integration has no natural real counterpart, because the real line lacks the flexible choice of path between points that occurs in the complex plane.

With minor exceptions, this chapter completes the natural analogies between the real and complex theories of differentiation and integration. From now on, new possibilities will unfold.

6.14 Exercises

1. Draw the contours $\gamma = [0, \mathrm{i}], \sigma = [0, 1] + [1, \mathrm{i}]$. Evaluate

$$\int_\gamma \operatorname{re} z \, \mathrm{d}z \qquad \int_\sigma \operatorname{re} z \, \mathrm{d}z$$

2. Draw the contours $\gamma = [-\mathrm{i}, \mathrm{i}]$ and σ where $\sigma(t) = \mathrm{e}^{\mathrm{i}t}$ $(t \in [-\pi/2, \pi/2])$. Evaluate

$$\int_\gamma |z| \mathrm{d}z \qquad \int_\sigma |z| \mathrm{d}z$$

3. Compute $\int_\gamma z^4 \mathrm{d}z$ for $\gamma(t) = (1 + \mathrm{i})t$ $(t \in [0, 1])$ using the formula

$$\int_\gamma f(z)\mathrm{d}z = \int_0^1 f(\gamma(t))\gamma'(t)\mathrm{d}t$$

and multiplying out the right-hand side and integrating the real and imaginary parts. If $\gamma = [0, 1] + [1, 1 + \mathrm{i}]$, what is $\int_\gamma z^4 \mathrm{d}z$?

4. Let n be a positive integer. If $\gamma(t) = z_0 + r\mathrm{e}^{\mathrm{i}t}$ $(t \in [0, 2n\pi])$ describe the contours γ and $-\gamma$ geometrically. Compute

$$\int_\gamma \frac{1}{z - z_0} \mathrm{d}z \quad \text{and} \quad \int_{-\gamma} \frac{1}{z - z_0} \mathrm{d}z$$

5. For $\gamma(t) = \mathrm{e}^{\mathrm{i}t}$ $(t \in [0, \pi])$ evaluate $\int_\gamma f$ for each of the following functions f:

(i) $1/z^2$
(ii) $1/z$
(iii) $\cos z$
(iv) $\sinh z$
(v) $\tan z$
(vi) $(\exp(z))^3$

6. Verify that the length function L satisfies

$$L(-\gamma) = L(\gamma) \qquad L(\gamma + \sigma) = L(\gamma) + L(\sigma)$$

where γ, σ are contours.

7. Let $\gamma(t) = z_0 + r\mathrm{e}^{\mathrm{i}t}$ $(t \in [0, \theta])$ be the arc centre z_0, radius r, subtending angle θ. Using Proposition 6.11, verify that

$$\theta = L(\gamma)/r$$

which is the usual definition of 'angle θ in radian measure'.

8. Show that the length of the parabolic arc $\gamma(t) = at^2 + 2a\mathrm{i}t$ $(t \in [0, 1])$, where $0 < a \in \mathbb{R}$, is

$$a\big(\sqrt{2} + \log\left(1 + \sqrt{2}\right)\big)$$

9. If γ is a closed contour in $\mathbb{C}$, the *signed area* enclosed by γ is defined to be

$$S = \frac{1}{2\mathrm{i}} \int_\gamma \bar{z} \mathrm{d}z$$

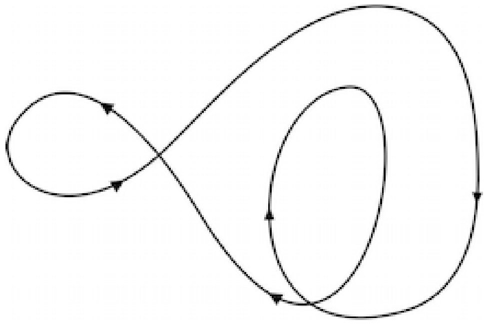

Figure 6.14 What does the signed area represent for this contour?

By writing the integral explicitly in the form $\int_a^b (u - \mathrm{i}v)(u' + \mathrm{i}v')$, or otherwise, show that S is real. Show that its value for circular and triangular contours is $\pm$ the usual area, and that both signs can occur. What is the geometric significance of the sign? What does the signed area represent for a contour such as Figure 6.14?

10. Show that the signed area in Exercise may also be given by

$$S = -\int_\gamma \operatorname{im} z \,\mathrm{d}z \qquad S = \frac{1}{\mathrm{i}}\int_\gamma \operatorname{re} z \,\mathrm{d}z$$

Write down the value of the integral $\int_\gamma \operatorname{im} z$, where γ is the square $\gamma = [0,1] + [1, 1+\mathrm{i}] + [1+\mathrm{i}, \mathrm{i}] + [\mathrm{i}, 0]$.

11. Let $\gamma(t) = \mathrm{e}^{\mathrm{i}t}$ $(t \in [0, \pi/2])$. Use integration by parts, Theorem 6.38, to compute

(i) $\int_\gamma z \sin z \,\mathrm{d}z$
(ii) $\int_\gamma z \cos z \,\mathrm{d}z$
(iii) $\int_\gamma z\mathrm{e}^{\mathrm{i}z} \,\mathrm{d}z$
(iv) $\int_\gamma z^2 \sin z \,\mathrm{d}z$

12. Let $c_r(t) = z_0 + r\mathrm{e}^{\mathrm{i}t}$ $(t \in [0, 2\pi])$ be the circle centre z_0, radius $r > 0$. If f is continuous in the domain D and $z_0 \in D$, use the Estimation Lemma to prove

(i) $\displaystyle\lim_{r\to 0} \int_{C_r} f(z)\mathrm{d}z = 0$
(ii) $\displaystyle\lim_{r\to 0} \int_{C_r} \frac{f(z)}{z - z_0}\mathrm{d}z = 2\pi \mathrm{i} f(z_0)$

13. For each of the following functions $f : D \to \mathbb{C}$, either state an antiderivative $F : D \to \mathbb{C}$ or explain why an antiderivative does not exist.

(i) $f(z) = z^2,\ D = \mathbb{C}$
(ii) $f(z) = 1/z^2,\ D = \mathbb{C} \setminus \{0\}$
(iii) $f(z) = 1/z,\ D = \mathbb{C} \setminus \{0\}$
(iv) $f(z) = z \sin z,\ D = \mathbb{C}$
(v) $f(z) = |z|^2,\ D = \mathbb{C}$
(vi) $f(z) = \bar{z},\ D = \mathbb{C}$

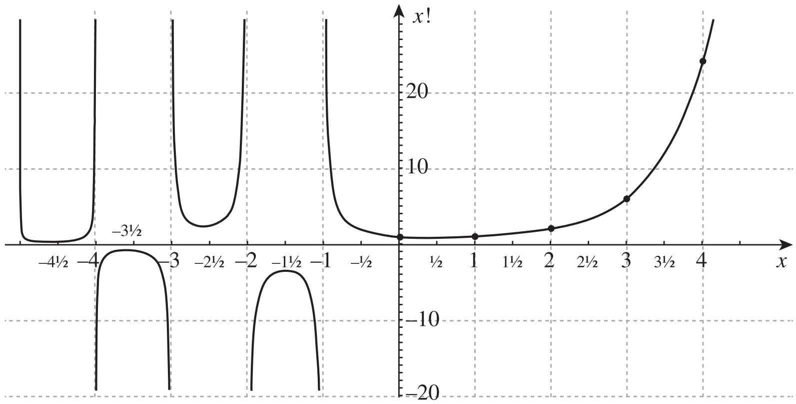

Figure 6.15 Graph of $x!$ for real x.

14. Use Euler's reflection formula

$$\Gamma(z)\Gamma(1-z) = \frac{\pi}{\sin(\pi z)}$$

to find $\Gamma(\frac{1}{2})$. Use the identity $\Gamma(z+1) = z\Gamma(z)$ to calculate $\Gamma(n+\frac{1}{2})$ for all $n \in \mathbb{N}$.

Gauss's pi function has the value $\Pi(z) = \Gamma(z+1)$ and is known to be analytic in $D = \mathbb{C} \setminus \{n \in \mathbb{Z} : n < 0\}$. By translating the identity $\Gamma(z+1) = z\Gamma(z)$ to $\Pi(z) = z\Pi(z-1)$ show that $\Pi(0) = 1$ and $\Pi(n) = n!$ for $n \in \mathbb{N}$.

Figure 6.15 shows the graph of $x!$ for real values of x. Calculate the numerical values of $(n+\frac{1}{2})!$ for $n = 0, 1, 2, 3, -1, -2, -3, -4$, and mark them on the graph.

7 Angles, Logarithms, and the Winding Number

So far, we have defined complex analogues of most of the basic special functions of real analysis – the exponential and trigonometric functions – and proved that many of their familiar properties extend to the complex case. Some new features also arise, such as the link between exponential and trigonometric functions and the periodicity of exp.

There is one glaring exception to our catalogue of complex analogues: the logarithm. In real analysis, we define this as the inverse function to $\exp : \mathbb{R} \to \mathbb{R}^+$, so $\log : \mathbb{R}^+ \to \mathbb{R}$. However, *because* exp is periodic in $\mathbb{C}$, it is not one-one (injective). Therefore it does not have an inverse function in the strict set-theoretic sense. This feature caused many of the historical disagreements about logarithms of negative and complex numbers, exacerbated by the general view that $\log z$ is automatically meaningful when $z \in \mathbb{C}$, and all of its properties ought to generalise to $\mathbb{C}$.

Resolving these issues involves being far more careful when defining complex logarithms. But the effort is worthwhile, because logarithms hold the key to many of the deeper and more powerful features of complex analysis.

The problem of finding inverses to functions that are not bijections arises in real analysis as well. Inverse trigonometric functions such as $\sin^{-1}$ are notorious for causing severe headaches, but the square root is the most obvious case. The square $f(x) = x^2$ is not injective as a function $f : \mathbb{R} \to \mathbb{R}$. It is also not surjective: negative real numbers are not squares of real numbers. The historical resolution was to define two square roots for every positive real number, $\pm\sqrt{x}$. Each is a one-sided inverse: $(\pm\sqrt{x})^2 = x$ when $x > 0$. But neither is an inverse on the domain of f, which is all of $\mathbb{R}$: their images are the positive reals and the negative reals, respectively, so they are not surjections onto $\mathbb{R}$. To obtain genuine inverse functions, we restrict the domain of f to the non-negative reals, when $\sqrt{x}$ is an inverse, or to the non-positive reals, when $-\sqrt{x}$ is an inverse. These two domains meet only at the origin.

It would be natural to try the same trick in the complex case. But now there is no very natural way to restrict the domain and codomain to make the logarithm a bijection. A variety of more or less artificial choices exist, such as the 'cut plane' $\mathbb{C}_\pi$ where the negative real axis is deleted, and these have their uses. To obtain a natural description of the complex logarithm, though, we use a more geometric approach. In classical terms, the complex logarithm is 'multivalued' – its value is not uniquely defined. The *way* in which its multiplicity of values fit together is closely analogous to the way that the measurement of an angle in radians gives not a single real number, but an infinite list differing only by multiples of 2π. Indeed, these two instances of non-uniqueness are closely related. In particular, the well-known complications when defining inverse trigonometric functions in real analysis also arise because these functions are 2π-periodic, hence not injective. The same issue has to be resolved to make sense of the complex logarithm.

We discuss these ideas below, and apply them to a topological invariant known as the *winding number* of a path relative to a point $z_0 \in \mathbb{C}$. In essence, this measures the total angle relative to z_0 traversed by a point moving along the path continuously from beginning to end. When divided by 2π, this angle tells us how many times the path winds round the point z_0. The winding number proves to be extremely useful in the deeper parts of the subsequent theory.

7.1 Radian Measure of Angles

We first relate the power series definitions of sine (5.7) and cosine (5.8) to the usual geometric ones, with the angle being measured in radians. This completes the proof that the power series represent the traditional trigonometric functions in the real case.

Let $0 \neq z \in \mathbb{C}$ and write z in polar coordinates as $z = r\mathrm{e}^{\mathrm{i}\theta}$, $(r > 0)$. Since $\mathrm{e}^{\mathrm{i}\theta} = \cos\theta + \mathrm{i}\sin\theta$, Table 5.1 in Section 5.5 implies that for fixed r, as θ increases from 0 to $\pi/2$, the real part of z decreases monotonically from r to 0 while the imaginary part increases monotonically from 0 to r. Since $|z|^2 = r^2(\cos^2\theta + \sin^2\theta) = r^2$, it follows that z traces out the first quadrant of a circle of radius r (in the anticlockwise direction) as θ increases from 0 to $\pi/2$, Figure 7.1 (left).

Similarly, as θ increases from $\pi/2$ to π, the point z traces out the second quadrant of a circle of radius r; from π to $3\pi/2$, it traces out the third quadrant; from π to $3\pi/2$, it traces out the fourth; and thereafter it continues to go round the circle in the anticlockwise direction, Figure 7.1 (middle).

We now compute the arc length from 1 to z along the circular path. For $\theta \geq 0$, let $\gamma(t) = r\mathrm{e}^{\mathrm{i}t}$, $(t \in [0, \theta])$. Then γ is a contour from 1 to z along the relevant arc. The length of γ is

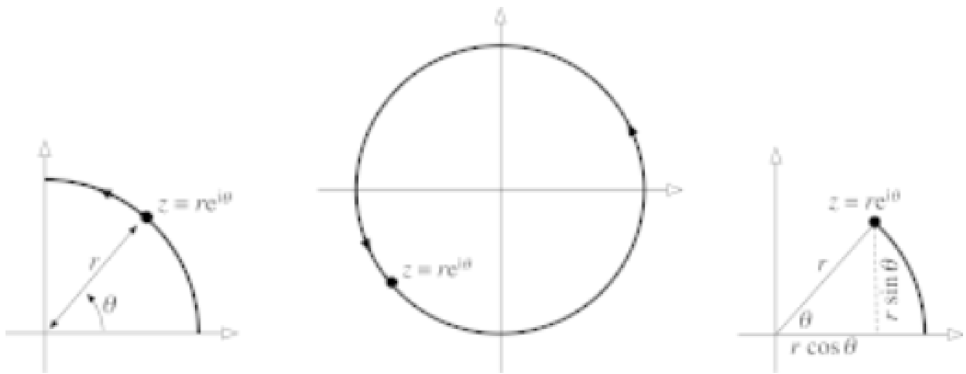

Figure 7.1 *Left*: $r\mathrm{e}^{\mathrm{i}\theta}$ traces out the first quadrant of a circle of radius r as θ increases from 0 to $\pi/2$. *Middle*: For $\theta \in \mathbb{R}$, $r\mathrm{e}^{\mathrm{i}\theta}$ repeatedly traces out the circle of radius r in the anticlockwise direction. *Right*: Traditional geometric definitions of $\sin\theta$ and $\cos\theta$ agree with the power series definitions.

$$L(\gamma) = \int_0^\theta |\gamma'(t)|\mathrm{d}t = \int_0^\theta |\mathrm{i}r\mathrm{e}^{\mathrm{i}t}|\mathrm{d}t = \int_0^\theta r\mathrm{d}t = r\theta$$

Thus $\theta = L(\gamma)/r$, which is the standard definition of 'angle measured in radians'. Figure 7.1 (right) then shows that the geometric definitions of $\sin\theta$ and $\cos\theta$, in terms of a right triangle with an angle θ radians, agrees with our analytic definition via power series.

The 2π-periodicity of sin and cos proved in Proposition 5.5 shows that this agreement extends to angles greater than 2π, and to negative angles, which of course are measured in the opposite sense round the circle: clockwise.

7.2 The Argument of a Complex Number

We now look in more detail at the expression of a complex number z in the form $r\mathrm{e}^{\mathrm{i}\theta}$. Here $r = |z|$, so it is unique. However, when $r = 0$ any value of θ leads to $z = 0$. And when $r > 0$ there are infinitely many possible values of θ, differing only by integer multiples of 2π. Recall that the unique value in the range $-\pi < \theta \leq \pi$ is called the *principal value of the argument of* z, and is denoted by

$$\arg z$$

This defines a function

$$\arg : \mathbb{C} \setminus \{0\} :\to \mathbb{R}$$

The function arg is *not* continuous on the negative real axis. This is a result of the need to choose θ uniquely: just above the axis θ is close to π, just below it is close to $-\pi$. We can sidestep this problem (inelegantly) by defining the *cut plane*

$$\mathbb{C}_\pi = \mathbb{C} \setminus \{x + \mathrm{i}y : y = 0, x \leq 0\}$$

as in Figure 7.2.

We claim that arg is continuous in the cut plane. This is plausible geometrically, but we must provide a rigorous proof. There is one technical difficulty: the bad behaviour of

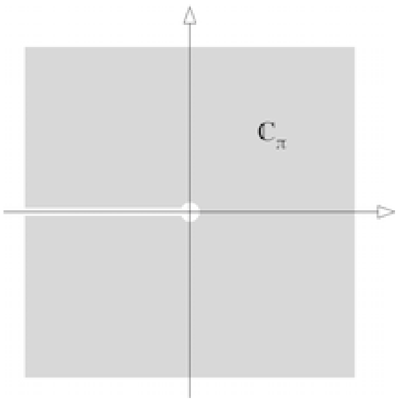

Figure 7.2 The cut plane $\mathbb{C}_\pi$.

inverse trigonometric functions (also caused by periodicity). We circumvent it by using several overlapping domains, on each of which the behaviour is sufficiently good. The proof that follows is an inelegant 'bare hands' reduction to properties of real functions: for a more elegant approach see Section 8.4.

PROPOSITION 7.1. *The function* arg *is continuous in the cut plane* $\mathbb{C}_\pi$.

Proof. Let

$$D_1 = \{x + iy \in \mathbb{C} : y > 0\}$$
$$D_2 = \{x + iy \in \mathbb{C} : x > 0\}$$
$$D_3 = \{x + iy \in \mathbb{C} : y < 0\}$$

Then

$$\mathbb{C} = D_1 \cup D_2 \cup D_3$$

(draw a picture!) and in each of these domains we have an easy way to find the value of $\arg z$.

To do so, observe that if

$$z = x + iy = e^{i\theta}$$

and we wish to solve for r, θ, then

$$r^2 = x^2 + y^2 \quad \text{so } r = \sqrt{x^2 + y^2}$$

and

$$\cos\theta = \frac{x}{\sqrt{x^2 + y^2}}$$
$$\sin\theta = \frac{y}{\sqrt{x^2 + y^2}}$$

From properties of sin and cos it follows that if $z \in D_1$ there is a unique solution for θ with $0 < \theta < \pi$. In this range cos is monotonic strictly decreasing and continuous, so its restriction to $(0, \pi)$ has a continuous inverse function

$$\cos^{-1} : (-1, 1) \to (0, \pi)$$

For $z \in D_1$ we have

$$\arg z = \cos^{-1} \frac{\operatorname{re} z}{|z|}$$

which, being a composition of continuous functions, is continuous.

Similarly if $x \in D_2$ then $-\pi/2 < \theta < \pi/2$. In this range, sin increases monotonically from -1 to 1, so it has a continuous inverse

$$\sin^{-1} : (-1, 1) \to \left(\frac{-\pi}{2}, \frac{\pi}{2}\right)$$

Now

$$\arg z = \sin^{-1} \frac{\operatorname{im} z}{|z|}$$

which is continuous on D_2.

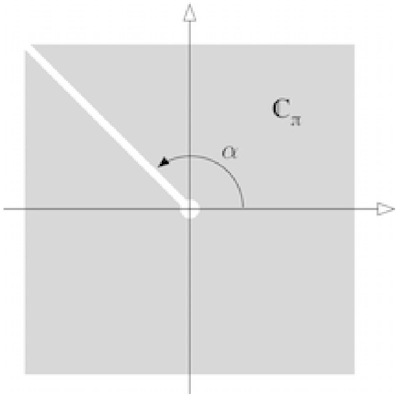

Figure 7.3 The cut plane $\mathbb{C}_\alpha$.

Finally, on D_3 we can use

$$\cos^{-1} : (-1, 1) \to (-\pi, 0)$$

which is a *different* choice of inverse cosine from that used in D_1 because we are inverting the restriction of cos to a different interval. Again arg is continuous on D_3.

Since D_1, D_2, D_3 are open, arg is continuous in $D_1 \cup D_2 \cup D_3 = \mathbb{C}_\pi$. □

It is sometimes convenient to proceed more generally. Let $\alpha \in \mathbb{R}$, and let R_α be the ray

$$R_\alpha = \{re^{i\alpha} : r \geq 0\} \subseteq \mathbb{C}$$

Let

$$\mathbb{C}_\alpha = \mathbb{C} \setminus \mathbb{R}_\alpha$$

as in Figure 7.3. Choose

$$\theta = \arg_\alpha z \quad (z \in \mathbb{C}_\alpha)$$

by the rule

$$z = re^{i\theta},\ r > 0,\ \alpha - 2\pi < \theta < \alpha$$

Then by a similar method it may be shown that arg is continuous in $\mathbb{C}_\alpha$.

7.3 The Complex Logarithm

The complex logarithm poses similar problems, because the complex exponential function is not one-one.

We wish to define $\log z$, for $0 \neq z \in \mathbb{C}$, by

$$w = \log z \quad \text{if} \quad z = e^w$$

(We exclude $z = 0$ because e^w is never zero, by Proposition 5.4.)

Let $z = re^{i\theta}, w = u + iv$. (The mixture of polar and Cartesian coordinates is deliberate!) We assume $r > 0,\ -\pi < \theta \leq \pi$, so $\theta = \arg z$. Then $z = e^w$ becomes

$$re^{i\theta} = e^{u+iv} \tag{7.1}$$

Taking moduli,

$$r = e^u \tag{7.2}$$

Since $r > 0$ and $r, u \in \mathbb{R}$ this has the unique solution

$$u = \log r$$

where log is the real natural logarithm. Then (7.1) and (7.2) imply that

$$e^{i\theta} = e^{iv}$$

so that

$$v = \theta + 2n\pi \quad (n \in \mathbb{Z})$$

Hence

$$\log z = w = \log r + i(\theta + 2n\pi)$$

or

$$\log z = \log |z| + i(\arg z + 2n\pi) \tag{7.3}$$

The complex logarithm is thus ‘multivalued’ and therefore not a function in the strict set-theoretic sense. To obtain a genuine (single-valued) function we define the *principal value* of the logarithm to be

$$\operatorname{Log} z = \log |z| + i \arg z \quad (0 \neq z \in \mathbb{C})$$

(Note the capital ‘L’ for the principal value.) This function is *not* continuous on the negative real axis. However, for $z \in \mathbb{C}_\pi$ the real and imaginary parts of Log are clearly continuous, so Log is continuous in the cut plane.

Now we can compute the derivative of the logarithm. We have

$$\begin{aligned}\frac{d}{dz}\operatorname{Log} z_0 &= \lim_{z \to z_0} \frac{\operatorname{Log} z - \operatorname{Log} z_0}{z - z_0} \\ &= \lim_{w \to w_0} \frac{w - w_0}{e^w - e^{w_0}}\end{aligned}$$

setting $z = e^w, z_0 = e^{w_0}$ and using continuity of Log. Thus

$$\frac{1}{\frac{d}{dz}\operatorname{Log} z_0} = \lim_{w \to w_0} \frac{e^w - e^{w_0}}{w - w_0} = \frac{d}{dz}\exp w_0 = e^{w_0}$$

and

$$\frac{d}{dz}\operatorname{Log} z_0 = \frac{1}{e^{w_0}} = \frac{1}{z_0}$$

and in general

$$\frac{d}{dz}\operatorname{Log} z = \frac{1}{z} \quad (z \in \mathbb{C}_\pi)$$

In the same way, in the cut plane $\mathbb{C}_\alpha$ we can define

$$\log_\alpha z = \operatorname{Log}|z| + \mathrm{i}\arg_\alpha z$$

and this is continuous, with derivative

$$\frac{\mathrm{d}}{\mathrm{d}z}\log_\alpha z = \frac{1}{z} \quad (z \in \mathbb{C}_\alpha)$$

Once the logarithm is defined, powers z^a for complex a can be defined, in a cut plane, by reducing them to exponentials. The *principal value* of z^a, when $z \neq 0$, is

$$z^a = \exp(a \operatorname{Log} z)$$

Exercises 9–14 of Section 7.9 explore this function. For a more global view of z^a see Chapter 14.

7.4 The Winding Number

Suppose that $\gamma : [a,b] \to \mathbb{C} \setminus \{0\}$ is a closed path. The choice of codomain here ensures that the path does not pass through the origin. If we imagine t, the parameter, to be time, increasing from a to b, and choose the argument $\theta(t)$ of $\gamma(t)$ to vary continuously with t, then as $\gamma(t)$ moves along the path, then intuitively the total change in argument is the number of times that γ winds round the origin (anticlockwise), multiplied by 2π. (If γ is not closed this can also be defined, but it need not be an integer.)

We wish, of course, to make this idea precise:

DEFINITION 7.2. A *continuous choice of argument* along a path $\gamma : [a,b] \to \mathbb{C} \setminus \{0\}$ (which for full generality we no longer require to be a closed path) is a *continuous* map $\theta : [a,b] \to \mathbb{R}$ such that

$$\mathrm{e}^{\mathrm{i}\theta(t)} = \frac{\gamma(t)}{|\gamma(t)|} \tag{7.4}$$

Condition (7.4) merely says that $\theta(t)$ is *one* of the possible values of the argument of $\gamma(t)$. □

Before proving that a continuous choice of argument exists, in Theorem 7.4 below, we consider a simple example.

Example 7.3. Let $\gamma(t) = r\mathrm{e}^{\mathrm{i}t}$ $(t \in [0,\pi])$, Figure 7.4.

A continuous choice of argument is $\theta(t) = t$, because obviously $r\mathrm{e}^{\mathrm{i}t}/|r\mathrm{e}^{\mathrm{i}t}| = \mathrm{e}^{\mathrm{i}t} = r\mathrm{e}^{\mathrm{i}\theta(t)}$. But this is not the only possible choice: $\theta(t) = t + 2\pi$ works equally well, or indeed $\theta(t) = t + 2n\pi$ for any fixed $n \in \mathbb{Z}$.

What we *cannot* do is 'change horses in midstream', say by choosing

$$\theta(t) = \begin{cases} t & (t \in [0,\pi/2]) \\ t + 2\pi & (t \in (\pi/2,\pi]) \end{cases}$$

If we do that, condition (7.4) holds, but θ is not continuous. □

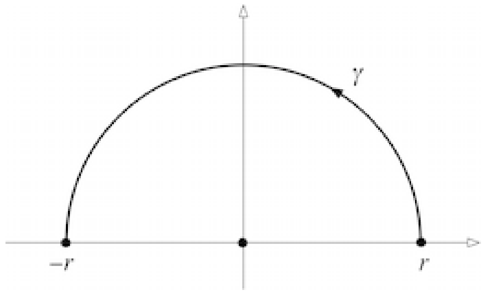

Figure 7.4 A semicircular path.

The difficulty (though it does not take much getting used to) in handling continuous choices of argument is that it does not in general suffice to insist on some simple recipe for choosing a value of the argument, such as keeping angles between 0 and 2π. Such a choice becomes *discontinuous* if the path winds round once anticlockwise, taking the argument to 2π, and then continues in the anticlockwise direction. Then the argument jumps discontinuously back to near 0 when we want it to continue increasing from 2π.

It turns out that we can choose the argument at the starting point to be any of the infinitely many possible values; but this choice then determines all of the rest uniquely – and the arguments that arise need not stay within any predetermined range. It is intuitively obvious that this kind of choice can be made, and that it is unique once we decide where to start. Its proof is a little tricky: it uses the Paving Lemma.

THEOREM 7.4. (i) *Let* $\gamma : [a,b] \to \mathbb{C} \setminus \{0\}$ *be a path not passing through the origin. Then there exists a continuous choice of argument for* γ.
(ii) *Any other continuous choice of argument differs from this by a constant integer multiple of* 2π.

Proof. By the Paving Lemma 2.33 we can subdivide γ into finitely many subpaths γ_r $(r = 1, \ldots, n)$ such that each γ_r lies inside a disc $D_r \subseteq \mathbb{C} \setminus \{0\}$, Figure 7.5. If the centre of d_r is at $\rho_r e^{i\theta_r}$, then taking $\alpha_r = \theta_r + \pi$, we find that $D_r \subseteq \mathbb{C}_\alpha$, so $\arg_\alpha$ is continuous in D_r.

This gives a continuous choice of argument across γ_r for each r, but of course these choices need not fit together continuously. However, *any* choice of argument can be obtained from any other by adding a suitable integer multiple of 2π. We can therefore adjust the $\arg_{\alpha_r}$ inductively so that they do fit together continuously, as follows.

As usual, let γ_r be defined on the parametric interval $[t_{r-1}, t_r]$. There is an integer n_2 such that

$$\arg_{\alpha_1}(t_1) = \arg_{\alpha_2}(t_1) + 2n_2\pi$$

Then there is an integer n_3 such that

$$\arg_{\alpha_2}(t_2) + 2n_2\pi = \arg_{\alpha_3}(t_2) + 2n_3\pi$$

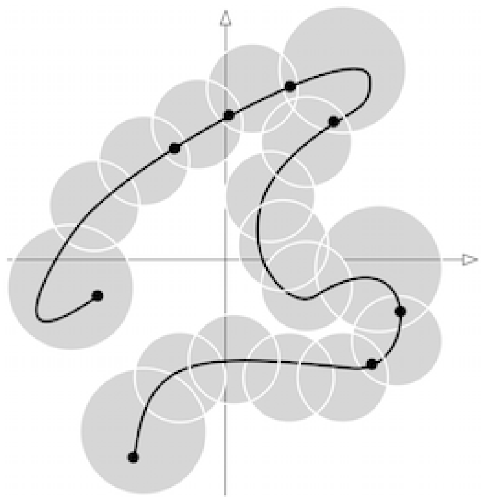

Figure 7.5 Paving a path.

and so on inductively, choosing n_{r+1} so that

$$\arg_{\alpha_r}(t_r) + 2n_r\pi = \arg_{\alpha_{r+1}}(t_r) + 2n_{r+1}\pi$$

Then we define

$$\theta(t) = \arg_{\alpha_r}(t_r) + 2n_r 2\pi \quad (t \in [t_{r-1}, t_r])$$

with n_1 conventionally defined to be 0. Then θ is continuous. This proves that a continuous choice of argument exists.

Next, suppose that θ^* is another continuous choice of argument. Then we must have

$$\theta^*(t) = \theta(t) + 2n(t)\pi$$

where $n(t)$ is an integer, possibly depending on t. But

$$n(t) = \frac{\theta^*(t) - \theta(t)}{2\pi}$$

is continuous, yet takes integer values. It is therefore constant. □

Note that a continuous choice of argument is defined on the parametric interval $[a, b]$, not on the image of γ. This means that if the path returns to the same point, with $\gamma(t_1) = \gamma(t_2)$ but $t_1 \neq t_2$, the arguments $\theta(t_1)$ and $\theta(t_2)$ may be different. Intuitively, it is clear that this will happen if the path winds round the origin a non-zero number of times between t_1 and t_2.

Example 7.5. Let $\gamma(t) = re^{4\pi i t}$ $(t \in [0, 1])$.

A continuous choice of argument is given by $\theta(t) = 4\pi t + 2n\pi$, for any choice of $n \in \mathbb{Z}$. Although $\gamma(t) = \gamma(t + \frac{1}{2})$ for $t \in [0, \frac{1}{2}]$, the choices of argument $\theta(t)$ and $\theta(t + \frac{1}{2})$ differ by 2π, because the path has travelled once round the origin (anticlockwise) between t and $t + \frac{1}{2}$.

More significantly, $\gamma(0) = \gamma(1)$, but the difference in arguments is 4π, because in total the path has travelled *twice* round the origin (anticlockwise) between $t = 0$ and $t = 1$. □

We can take advantage of this idea:

DEFINITION 7.6. The *winding number* of a path $\gamma : [a,b] \to \mathbb{C} \setminus \{0\}$ round the origin is

$$w(\gamma, 0) = \frac{\theta(b) - \theta(a)}{2\pi}$$

for a continuous choice of argument θ along γ. □

By Theorem 7.4 (ii) the winding number is well defined; that is, any continuous choice of argument gives the same value.

For any *closed* path, the winding number is an integer, because $\theta(b) - \theta(a)$ is an integer multiple of 2π.

The winding number takes the *sense* of the path into account: anticlockwise turns count as positive, clockwise turns count as negative.

Example 7.7. Let $\gamma(t) = r\mathrm{e}^{-\mathrm{i}t}$ $(t \in [0, 6\pi])$.

A continuous choice of argument is $\theta(t) = -t$. Then

$$w(\gamma, 0) = \frac{\theta(6\pi) - \theta(0)}{2\pi} = -3$$

and γ winds three times round the origin in the clockwise direction. □

The winding number is *additive*:

THEOREM 7.8. *Let γ_1 and γ_2 be two paths in $\mathbb{C} \setminus \{0\}$ such that the end point of γ_1 is the start of γ_2. Then*

$$w(\gamma_1 + \gamma_2, 0) = w(\gamma_1, 0) + w(\gamma_2, 0)$$

Proof. We may assume that γ_1, γ_2, and $\gamma_1 + \gamma_2$ have parametric intervals $[a,b], [b,c]$, and $[a,c]$ respectively. Let θ be a continuous choice of argument on $\gamma_1 + \gamma_2$. Then

$$\begin{aligned} w(\gamma_1 + \gamma_2, 0) &= \frac{\theta(c) - \theta(a)}{2\pi} \\ w(\gamma_1, 0) &= \frac{\theta(b) - \theta(a)}{2\pi} \\ w(\gamma_2, 0) &= \frac{\theta(c) - \theta(b)}{2\pi} \end{aligned}$$

□

This is an extremely useful result, because it lets us compute the winding number of a complicated path by breaking it up into nice pieces and adding their contributions. It extends easily to show that

$$w(\gamma_1 + \cdots + \gamma_n, 0) = w(\gamma_1, 0) + \cdots + w(\gamma_n, 0)$$

and that

$$w(-\gamma, 0) = -w(\gamma, 0)$$

7.5 The Winding Number as an Integral

If the path is smooth, we can specify its winding number as an integral along the path. Additivity of the winding number extends this formula to contours. First, consider a closed contour.

THEOREM 7.9. *Let γ be a closed contour. Then*

$$w(\gamma, 0) = \frac{1}{2\pi \mathrm{i}} \int_\gamma \frac{1}{z} \mathrm{d}z \tag{7.5}$$

Proof. Subdivide the path into subpaths $\gamma_1, \ldots, \gamma_n$ as in Theorem 7.4, whose notation we now use. Each γ_r lies in a cut plane $\mathcal{C}_{\alpha_r}$. If γ_r is defined on the interval $[t_{r-1}, t_r]$ then

$$\begin{aligned}
\int_{\gamma_r} \frac{1}{z} \mathrm{d}z &= \log_{\alpha_r} \gamma(t_r) - \log_{\alpha_r} \gamma(t_{r-1}) \\
&= \log|\gamma(t_r)| - \log|\gamma(t_{r-1})| \\
&= \mathrm{i}(\arg_{\alpha_r} |\gamma(t_r)| - \arg_{\alpha_r} |\gamma(t_{r-1})|)
\end{aligned}$$

As in Theorem 7.4 we make sure that $\arg_{\alpha_r} \gamma(t_r) = \arg_{\alpha_{r+1}} \gamma(t_r)$, that is, the choices of argument agree where the subpaths join together. Then, summing the integrals for $r = 1, \ldots, n$, the real parts cancel out because γ is a closed contour, and the imaginary parts add up to $2\pi w(\gamma, 0)$, as required. □

7.6 The Winding Number Round an Arbitrary Point

There is nothing very special about the origin as far as winding numbers are concerned. If $\gamma : [a, b] \to \mathbb{C}$ is a path and $z_0 \in \mathbb{C}$ does not lie on γ, we can define the winding number of γ around z_0. The easiest way to do this is to translate the origin, setting

$$\Gamma(t) = \gamma(t) - z_0 \quad (t \in [a, b])$$

and calculating

$$w(\gamma, z_0) = w(\Gamma, 0)$$

For a closed path γ, we get

$$\begin{aligned}
w(\gamma, z_0) &= \frac{1}{2\pi \mathrm{i}} \int_\Gamma \frac{1}{z} \mathrm{d}z \\
&= \frac{1}{2\pi \mathrm{i}} \int_a^b \frac{\Gamma'(t)}{\Gamma(t)} \mathrm{d}t \\
&= \frac{1}{2\pi \mathrm{i}} \int_a^b \frac{\gamma'(t)}{\gamma(t) - z_0} \mathrm{d}t \\
&= \frac{1}{2\pi \mathrm{i}} \int_\gamma \frac{1}{z - z_0} \mathrm{d}z
\end{aligned}$$

This leads to:

DEFINITION 7.10. If $\gamma : [a,b] \to \mathbb{C}$ is a closed path and $z_0 \in \mathbb{C}$ does not lie on γ, then the *winding number of* γ *around* z_0 is

$$w(\gamma, z_0) = \frac{1}{2\pi \mathrm{i}} \int_\gamma \frac{1}{z - z_0} \mathrm{d}z \tag{7.6}$$

□

The additivity of $w(\gamma, 0)$ in Theorem 7.8 extends easily to $w(\gamma, z_0)$ for fixed but arbitrary $z_0 \in \mathbb{C}$, using the same trick of translating the origin.

7.7 Components of the Complement of a Path

We consider how the winding number $w(\gamma, z_0)$ of a given closed path γ can vary as z_0 varies.

By Proposition 2.41, the complement S of the image of γ is open, and also each connected component of S is open. The winding number $w(\gamma, z_0)$ is defined for all $z_0 \in S$, and it is an *integer*-valued function on S. It is geometrically plausible that this function is constant on any connected component of S. We prove this analytically by showing that $w(\gamma, z_0)$ is a *continuous* function of $z_0 \in S$. The desired result then follows, since any integer-valued continuous function is constant on any connected set.

The proof that $w(\gamma, z_0)$ is continuous in z_0 is obtained by a direct estimate. Fix $z_0 \in S$. Since S is open there exists $k > 0$ such that $|z_1 - z_0| < k$ implies $z_1 \in S$. Therefore if z lies on the image of γ then $|z - z_0| \geq k$. Thus if $|z_1 - z_0| < k/2$ we have $|z - z_1| > k/2$. Now

$$\begin{aligned} |w(\gamma, z_0) - w(\gamma, z_1)| &= \frac{1}{2\pi} \left| \int_\gamma \left[\frac{1}{z - z_0} - \frac{1}{z - z_1} \right] \mathrm{d}z \right| \\ &= \frac{1}{2\pi} \left| \int_\gamma \frac{z_1 - z_0}{(z - z_0)(z - z_1)} \mathrm{d}z \right| \\ &\leq \frac{|z_1 - z_0|}{\pi k^2} \end{aligned}$$

by the Estimation Lemma 6.41.

Given any $\varepsilon > 0$, take $\delta = \min(k/2, \pi k^2 \varepsilon / 2L(\gamma))$. Then

$$|z_1 - z_0| < \delta \text{ implies } |w(\gamma, z_0) - w(\gamma, z_1)| < \varepsilon$$

Hence $w(\gamma, z_0)$ is continuous in z_0 on S.

Example 7.11. The path in Figure 7.6 has the winding numbers shown, around points z_0 in the components of S to which those numbers are assigned. The set S has a single unbounded component (Proposition 2.41) which we denote by $O(\gamma)$ (where the 'O' stands for 'outside').

As Figure 7.6 shows, if $z_0 \in O(\gamma)$ then $w(\gamma, z_0) = 0$. This is easily proved from the integral formula (7.6), as follows. Let z_0 be 'far from the image of γ', that is, assume $|z - z_0| \geq K$ for large K. Then the Estimation Lemma 6.41 shows that

$$|w(\gamma, z_0)| \leq L(\gamma)/2\pi K$$

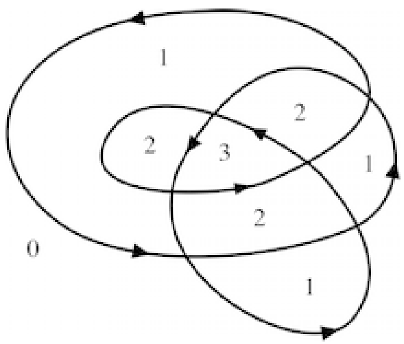

Figure 7.6 Winding numbers for components of the complement of a path.

which tends to zero for large K. But the left-hand side is a non-negative integer, so it must be *equal* to zero for large enough K. Since it is constant on $O(\gamma)$, it must be zero on the whole of $O(\gamma)$. □

7.8 Computing the Winding Number by Eye

The somewhat complicated definition of the winding number may give the impression that it is complicated to calculate. This is not so, at least for paths that are contours made up of smooth subpaths. For contours the calculation is simple: trace your finger along the contour and count the number of turns. The point of this section is that this process can easily be formalised so that, in this case, what is obvious is also true.

Example 7.12. We start with a path γ whose image is a rectangle with vertices $\pm 2 \pm \mathrm{i}$, Figure 7.7.

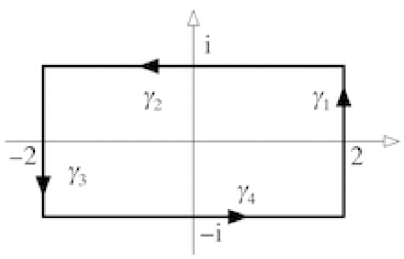

Figure 7.7 A rectangular path.

If you want a formula, it is not hard to give one; for example let

$$\gamma(t) = \begin{cases} 2 - \mathrm{i} + 2\mathrm{i}t & (t \in [0,1]) \\ 2 + \mathrm{i} - 4(t-1) & (t \in [1,2]) \\ -2 + \mathrm{i} - 2\mathrm{i}(t-2) & (t \in [2,3]) \\ -2 - \mathrm{i} + 4(t-3) & (t \in [3,4]) \end{cases}$$

This path is composed of four standard 'line segment' paths, one for each edge, Section 2.4. As the parameter t varies, the corresponding point moves along each edge at constant speed. (With our definition of a line $L(t)$, the speed along an edge depends on the length of that edge because t is taken in the interval $[0,1]$.)

We want to calculate $w(\gamma, 0)$.

First, here is how *not* to do the calculation.

Break γ up in the most natural way, into subpaths $\gamma_r = \gamma|_{[r-1,r]}$ for $r = 1, 2, 3, 4$. Then use the integral formula

$$w(\gamma, 0) = \frac{1}{2\pi \mathrm{i}} \int_\gamma \frac{1}{z} \mathrm{d}z = \frac{1}{2\pi \mathrm{i}} \sum_{r=1}^{4} \int_{\gamma_r} \frac{1}{z} \mathrm{d}z$$

Now (to take one subpath)

$$\begin{aligned} \int_{\gamma_1} \frac{1}{z} \mathrm{d}z &= \int_0^1 \frac{\gamma'(t)}{\gamma(t)} \mathrm{d}t \\ &= [\log(2 - \mathrm{i} + 2\mathrm{i})]_0^1 \\ &= \mathrm{Log}\,(2 + \mathrm{i}) - \mathrm{Log}\,(2 - \mathrm{i}) \end{aligned}$$

because γ_1 lies in $\mathbb{C}_\pi$, so the principal value Log is continuous on γ_1. Then

$$\begin{aligned} \mathrm{Log}\,(2 \pm \mathrm{i}) &= \log|2 \pm \mathrm{i}| + \mathrm{i}\arg(2 \pm \mathrm{i}) \\ &= \log\sqrt{5} \pm \mathrm{i}\sin^{-1}(1/\sqrt{5}) \end{aligned} \tag{7.7}$$

so

$$\int_{\gamma_1} \frac{1}{z} \mathrm{d}z = 2\mathrm{i}\sin^{-1}(1/\sqrt{5})$$

where the inverse sine is chosen between $-\pi/2$ and $\pi/2$.

You now have three similar integrals to evaluate. Add them, divide by $2\pi\mathrm{i}\ldots$, and do some *very* careful bookkeeping on the domains of the inverse trigonometric functions that occur. It *can* be done; in a sense it is not even difficult – but it is hardly to be recommended.

It is a little better (but not much) to work from the 'continuous choice of argument' definition. This starts you at stage (7.7) above for each subpath, with the same bookkeeping problems at the end.

Here is a more civilised method. Divide γ into subpaths δ_1, δ_2 as in Figure 7.8.

Now $w(\gamma, 0) = w(\delta_1, 0) + w(\delta_2, 0)$. Since δ_1 lies entirely within the cut plane $\mathbb{C}_\pi$, the principal value of arg is continuous on δ_1. Hence, dotting all i's and crossing all t's,

$$\begin{aligned} w(\delta_1, 0) &= [\arg(\mathrm{i}) - \arg(-\mathrm{i})]/2\pi \\ &= [\pi/2 - (-\pi/2)]/2\pi \\ &= \tfrac{1}{2} \end{aligned}$$

Similarly, δ_2 lies entirely within the cut plane $\mathbb{C}_0$, so $\arg_0$ is continuous on δ_2. Hence

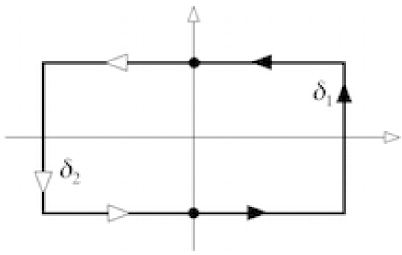

Figure 7.8 A more civilised method.

$$\begin{aligned} w(\delta_2, 0) &= [\arg_0(-\mathrm{i}) - \arg_0(\mathrm{i})]/2\pi \\ &= [-\pi/2 - (-3\pi/2)]/2\pi \\ &= \tfrac{1}{2} \end{aligned}$$

Adding, $w(\gamma, 0) = 1$. □

Clearly this process can be telescoped: the arg calculations merely confirm, in a predictable way, the obvious.

We summarise this method as follows:

(1) Break γ into convenient pieces, *each lying in some cut plane*.
(2) For each piece, compute the contribution to the winding number as the difference between the arguments of the end points, using the relevant continous arg – that is, $\arg_\alpha$ on $\mathbb{C}_\alpha$ – or, geometrically, find the angle subtended at the origin by the two end points of the subpath, with the appropriate sign.
(3) Add these contributions.

It helps if the dissection of γ into subpaths is performed by drawing a single straight line through the origin, because then the contributions are always $\pm\frac{1}{2}$. Thus in Figure 7.9, the line L cuts the path γ into four segments AB, BC, CD, and DA.

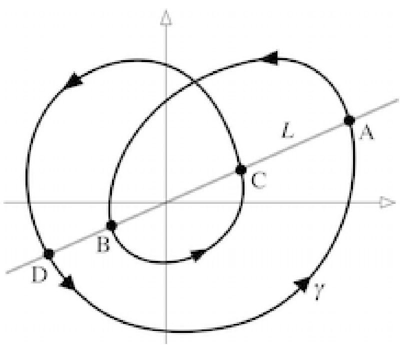

Figure 7.9 Cutting the path along line L makes the calculation of the winding number easy.

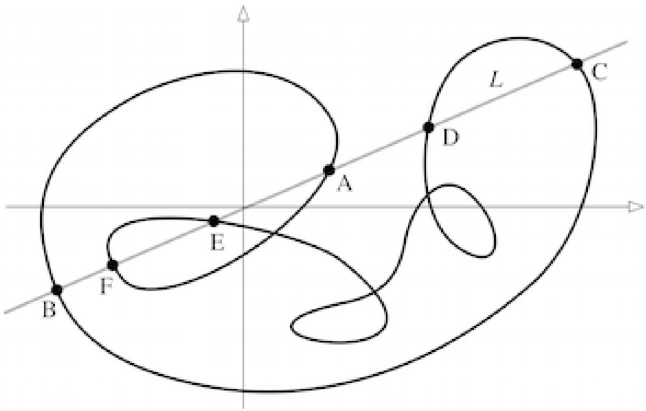

Figure 7.10 A more complicated example where cutting the path along a suitable line makes the calculation of the winding number easy.

Now

$$\begin{aligned} w(\gamma, 0) &= w(\text{AB}, 0) + w(\text{BC}, 0) + w(\text{CD}, 0) + w(\text{DA}, 0) \\ &= \tfrac{1}{2} + \tfrac{1}{2} + \tfrac{1}{2} + \tfrac{1}{2} \\ &= 2 \end{aligned}$$

Similarly, the path in Figure 7.10 gives

$$\begin{aligned} w(\gamma, 0) &= w(\text{AB}, 0) + w(\text{BC}, 0) + w(\text{CD}, 0) + w(\text{DE}, 0) + w(\text{EF}, 0) + w(\text{FA}, 0) \\ &= \tfrac{1}{2} + \tfrac{1}{2} + 0 + (-\tfrac{1}{2}) + 0 + \tfrac{1}{2} \\ &= 1 \end{aligned}$$

This method is essentially the mathematical equivalent of ‘run your finger along the path and count half-turns’. The ingredients needed to make it rigorous are *not* complicated evaluations of args. The real crunch comes in showing that the path really does cross the chosen line L at the points A, B, C, and so on; that it crosses the line only at those points; and that it passes through these points in the stated order; and that the sense (clockwise or anticlockwise) of each subpath is as in the diagram. It is also worth bearing in mind that a smooth path may cross a line infinitely many times, so this method is not foolproof. But whenever (as is always the case in the sequel) γ is specified in a straightforward way – such as by a simple formula, a polygon, or a series of straight lines and circular arcs – these ingredients are easily supplied. We then make no further fuss when we assert that a certain winding number takes a particular value.

Comparison of the first ‘bad’ method in Example 7.12 with the final ‘good’ one gives a striking illustration of the dangers of blind ‘formula-crunching’. Complex analysis is highly geometric, and the geometry should not be despised.

7.9 Exercises

1. Compute the principal values of the arguments of the following complex numbers: $1 + \mathrm{i}, \frac{1}{2} + \mathrm{i}\frac{\sqrt{3}}{2}, (1 + i)^3, (\frac{1}{2} + \mathrm{i}\frac{\sqrt{3}}{2})^{243}, (1 + \mathrm{i})^2(\frac{1}{2} + \mathrm{i}\frac{\sqrt{3}}{2})^3$.

2. Let $\arg z$ denote the principal value of the argument of $z \neq 0$, that is, $-\pi < \arg z \leq \pi$. For real x, y with $x < 0$ show:
(i) $\lim_{y\to 0} \arg(x + \mathrm{i}|y|) = \pi$
(ii) $\lim_{y\to 0} \arg(x - \mathrm{i}|y|) = -\pi$
Compute the corresponding limits when $x = 0$ and when $x > 0$.

3. Compute the following principal logarithms: Log (3i), Log (−2i), Log (1 + i), Log (−1), Log (z^{10}) where $z = 2\mathrm{e}^{\mathrm{i}\pi/3}$, Log ($x$) for real $x \neq 0$.

4. For $z_1, z_2 \neq 0$ show that

$$\mathrm{Log}\,(z_1 z_2) = \mathrm{Log}\,(z_1) + \mathrm{Log}\,(z_2) + 2n\pi\mathrm{i}$$

where n is an integer that need not be zero. Specify the values that n may take. Show that a logarithm of $z_1 z_2$ is of the form

$$\log z_1 + \log z_2$$

provided that appropriate values of the logarithms are taken.

5. The 'Bernoulli paradox' is as follows:

$$(-z)^2 = z^2$$

hence

$$2\log(-z) = 2\log(z)$$

so

$$\log(-z) = \log(z)$$

What is the fallacy?

6. Let $f : \mathbb{C} \to \mathbb{C}$ be given by $f(z) = \mathrm{e}^{\mathrm{i}\theta} z$ for constant real θ. Show that f rotates the complex plane through an angle θ. (Hint: prove that it is a rigid motion, that is, it satisfies $|f(z_1) - f(z_2)| = |z_1 - z_2|$ for all $z_1, z_2 \in \mathbb{C}$, and that it fixes the origin and only the origin. Then consider the value of $f(1)$.)

Show that the transformation $f(z) = \mathrm{i}z$ rotates the complex plane through a right angle, and describe the effects of the transformations $f(z) = -z, f(z) = -\mathrm{i}z$ as rotations.

For any complex number $\lambda = r\mathrm{e}^{\mathrm{i}\theta}$ describe the transformation $f(z) = \lambda z$ in geometric terms.

7. In each of the following cases, for $z = r\mathrm{e}^{\mathrm{i}\theta}$, draw the locations of z and $1/z$ in $\mathbb{C}$:
(i) $3\mathrm{e}^{\mathrm{i}\pi/2}$
(ii) $2\mathrm{e}^{\mathrm{i}\pi//4}$
(iii) $\frac{1}{2}\mathrm{e}^{\mathrm{i}\pi/3}$
(iv) $3\mathrm{e}^{-\mathrm{i}\pi}$
Describe the transformation $f(z) = 1/z$ in geometric terms.

8. Let n be a positive integer. A complex number ω is said to be an nth *root of unity* if $\omega^n = 1$.

(i) Find all nth roots of unity in polar coordinates and draw a picture.
(ii) For $n = 2, 3, 4$, express the nth roots of unity in the form $x + \mathrm{i}y$.
(iii) If ω_1, ω_2 are nth roots of unity, show that the following are also nth roots of unity:

$$\omega_1^m \quad \omega_1\omega_2 \quad \omega_1/\omega_2$$

(iv) For given $r, \theta \in \mathbb{R}$, $(r > 0)$, find all $z \in \mathbb{C}$ such that

$$z^n = re^{\mathrm{i}\theta}$$

(v) If $z_1^n = z_2^n$, show that $z_1 = \omega z_2$, where ω is an nth root of unity.

9. For $z, \beta \in \mathbb{C}$, $z \neq 0$, define the principal value of z^β to be

$$z^\beta = \exp(\beta \operatorname{Log} z)$$

Compute the principal values of the following powers:

$$1^{\sqrt{2}} \quad (-2)^{\sqrt{2}} \quad \mathrm{i}^{\mathrm{i}} \quad 2^{\mathrm{i}} \quad (3-4\mathrm{i})^{1+\mathrm{i}} \quad (3+4\mathrm{i})^5$$

10. Using the notation of Exercise 9, for $z \in \mathbb{C}_\pi$, show that $\mathrm{d}(z^\beta)/\mathrm{d}z = \beta z^{\beta-1}$. For fixed $a \in \mathbb{C}_\pi$, what is $\mathrm{d}(a^z)/\mathrm{d}z$?

11. Let $\alpha \in \mathbb{R}, \beta \in \mathbb{C}$. For $z \in \mathbb{C}_\alpha$, define

$$(z^\beta)_\alpha = \exp(\beta \log_\alpha z)$$

Compute all possible values of $(\mathrm{i}^{\mathrm{i}})_\alpha, (2^{\mathrm{i}})_\alpha, ((3-4\mathrm{i})^{1+\mathrm{i}})_\alpha$ for various values of α.

For fixed z, show that as α varies, $(z^\beta)_\alpha$ takes only finitely many values when β is a rational number.

If $m, n \in \mathbb{Z}, n > 0$, show that

$$((z^{m/n})_\alpha)^n = z^m$$

What is $((z^n)^{m/n})_\alpha$?

12. Using the notation of Exercise 11, for $z \in \mathbb{C}_\alpha$, compute $\mathrm{d}((z^\beta)_\alpha)/\mathrm{d}z$ and $\mathrm{d}((a^z)_\alpha))/\mathrm{d}z$.

13. Describe the image of the functions $f : \mathbb{C}_\pi \to \mathbb{C}$ geometrically where $f(z)$ is given by the principal values of the following:

$$z^{1/2} \quad z^{1/3} \quad z^{\mathrm{i}}$$

14. Let $\alpha \in \mathbb{R}, \beta \in \mathbb{C}, z \in \mathbb{C}_\pi$. By writing $\beta = u + \mathrm{i}v$, find

$$|(z^\beta)_\alpha| \quad \text{and} \quad \arg(z^\beta)_\alpha$$

Show that $|(z^\beta)_\alpha|$ is independent of α if and only if β is real.

For positive integers m, n let $f(z)$ be the principal value of $z^{m/n}$. Describe $f : \mathbb{C}_\pi \to \mathbb{C}$ geometrically.

15. Express Log $(1+z)$ as a power series in z of the form

$$\mathrm{Log}\,(1+z)=\sum_{n=0}^{\infty}a_n z^n \quad (|z|<R)$$

specifying the coefficients a_n and the radius of convergence R.

16. Show:
(i) $\cos(-\mathrm{i}\,\mathrm{Log}\,(z+\sqrt{z^2-1}))=z$
(ii) $\sin(-\mathrm{i}\,\mathrm{Log}\,(\mathrm{i}(z+\sqrt{z^2-1})))=z$
(iii) $\tan\left(\frac{\mathrm{i}}{2}\mathrm{Log}\left(\frac{\mathrm{i}+z}{\mathrm{i}-z}\right)\right)=z$
Use these properties to express $\cos^{-1}z, \sin^{-1}z, \tan^{-1}z$ in terms of the logarithm.

17. Show that all the values of $\cosh^{-1}z$ have the form

$$\cosh^{-1}z=\log(z+\sqrt{z^2-1})$$

for all possible values of the logarithm and square root. In the same sense, show that

$$\sinh^{-1}z=\log(z+\sqrt{z^2+1})$$
$$\tanh^{-1}z=\tfrac{1}{2}\log\left(\frac{1+z}{1-z}\right)$$

18. Let $D=\{z\in\mathbb{C}:\dfrac{\mathrm{i}+z}{\mathrm{i}-z}\in\mathbb{C}_\pi\}$. Describe D geometrically. Define the principal value of the inverse tangent $\tan^{-1}:D\to\mathbb{C}$ to be

$$\tan^{-1}z=\tfrac{1}{2}\mathrm{i}\,\mathrm{Log}\left(\frac{\mathrm{i}+z}{\mathrm{i}-z}\right)$$

(note the principal value of the logarithm). Show that $\tan^{-1}$ is differentiable in D with derivative $1/(1+z^2)$.

19. Draw the following paths and specify all the continuous choices of argument along them:
(i) $\gamma(t)=2\mathrm{e}^{-\mathrm{i}t}\ (t\in[0,4\pi])$
(ii) $\gamma(t)=t+\mathrm{i}(1-t)\ (t\in[0,1])$
(iii) $\gamma(t)=t-1+\mathrm{i}t^2\ (t\in[-1,1])$
(iv) $\gamma(t)=\begin{cases} t+\mathrm{i}(1-t) & (t\in[0,1])\\ 1+\mathrm{i}(t-1) & (t\in[1,2])\end{cases}$
In each case compute the winding number of the path round the origin.

20. Find the winding number $w(\gamma, z_0)$ for each of the following choices of γ, z_0:
(i) $\gamma(t)=2\mathrm{e}^{-\mathrm{i}t}\ (t\in[0,2\pi])$; $\quad z_0=1, 3\mathrm{i}$
(ii) $\gamma(t)=t+\mathrm{i}(1-t)\ (t\in[0,1])$; $\quad z_0=1+1, -\mathrm{i}, 10\mathrm{i}$

21. For each of the following, draw a picture of the path and use a sensible method to compute $\int_\gamma 1/(z-z_0)\mathrm{d}z$:
(i) $\gamma(t)=t\mathrm{e}^{-\mathrm{i}t}\ (t\in[\pi,5\pi])\quad z_0=0$
(ii) $\gamma(t)=-\mathrm{i}t\ (t\in[0,1])\quad z_0=1$
(iii) $\gamma(t)=\mathrm{i}t\ (t\in[-1,1])\quad z_0=1$

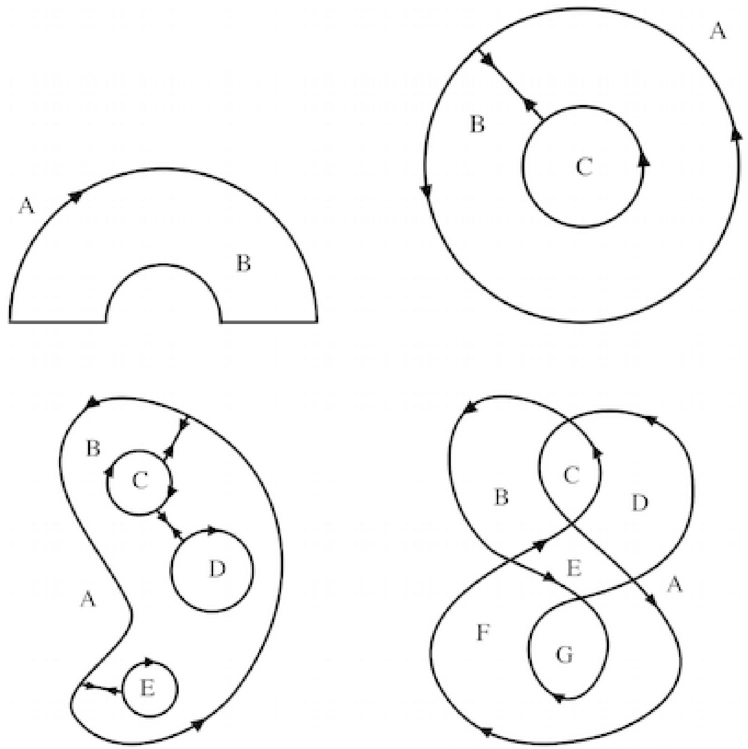

Figure 7.11 Compute (by eye) the winding numbers of these closed paths round a point in each component A, B, C, ...

(iv) $\gamma(t) = \sigma + [1, 2] + \rho + [-2, -1] \quad z_0 = 0$
where $\sigma(t) = e^{i(\pi - t)},\ (t \in [0, \pi])$ and $\rho(t) = 2e^{-it}$ $(t \in [0, 4\pi])$.

22. Compute (by eye) the winding number of the given closed paths round a point in each of the connected components A, B, C, . . . of the complement, drawn in Figure 7.11.

8 Cauchy's Theorem

The Fundamental Theorem of Contour Integration, Theorem 6.33, tells us that if D is a domain and $f : D \to \mathbb{C}$ has an antiderivative in D, then the integral of f along a path in D from z_0 to z_1 can be calculated using the formula

$$\int_\gamma f = F(z_1) - F(z_0)$$

In particular, if γ is closed, so $z_1 = z_2$, then

$$\int_\gamma f = 0$$

Cauchy's Theorem goes further. It states conditions under which $\int_\gamma f = 0$ when there is no initial reason for f to have an antiderivative. There are many different versions of Cauchy's Theorem – or, to be precise, many different theorems of this type. Cauchy was the first to publish such a result, announcing it in 1813 and getting it into print in 1825. Gauss was aware of the basic idea in 1811, but the accolade goes to Cauchy because he was the first to make it public.

Both Gauss and Cauchy realised the basic fact that if γ does not wind round points outside D (that is, the winding number round such points is zero) then $\int_\gamma f = 0$. For instance, $f(z) = 1/z$ has the single point 0 outside its domain, so if the closed contour γ does not wind round the origin, $\int_\gamma f = 0$, Figure 8.1.

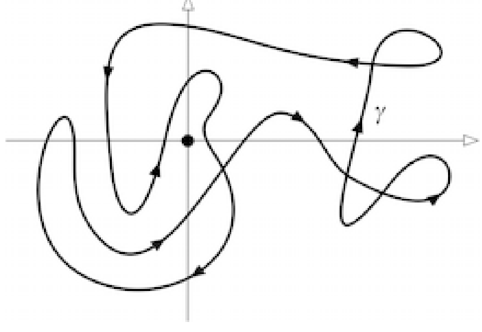

Figure 8.1 The integral of $1/z$ along this closed path is zero.

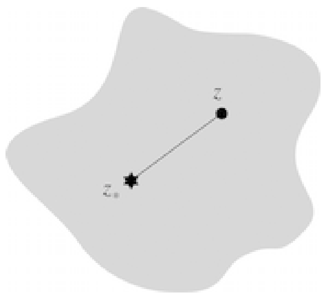

Figure 8.2 A star domain.

Our main aim in this chapter is to establish the following version of Cauchy's Theorem:

If f is differentiable in a domain D and $w(\gamma, z) = 0$ for all $z \notin D$, then $\displaystyle\int_\gamma f = 0.$

We start the theory rolling by proving the special case where the contour is a triangle.

Then we prove a theorem that requires a restriction on the domain D rather than the contour γ. we say that D is a *star domain* if it contains a point z_* such that for any other point $z \in D$ the line segment $[z_*, z] \subseteq D$, Figure 8.2. We then define $F(z) = \int_{[z_*,z]} f$ and use the triangle version of the theorem to show that F is an antiderivative of f. This means that $\int_\gamma f = 0$ for *any* closed contour in a star domain.

In particular, a disc is a star domain, and this leads to a very significant result. For a differentiable function f in a general domain D, we may not be able to find an antiderivative $F : D \to \mathbb{C}$. But if we restrict attention to any disc $\Delta \subseteq D$, there is an antiderivative $F : \Delta \to \mathbb{C}$. Thus an antiderivative may not exist *globally* throughout D, but it does exist locally in any neighbourhood $N_r(z_0) \subseteq D$ for any $z_0 \in D$.

Using the Paving Lemma 2.33, any contour γ in an arbitrary domain D can be written as $\gamma = \gamma_1 + \cdots + \gamma_n$ where each subcontour γ_r lies in a disc $D_r \subseteq D$. In the (star domain) D_r we can choose a step path σ_r with the same end points as γ_r and (by the existence of an antiderivative in D_r), $\int_{\sigma_r} f = \int_{\gamma_r} f$. If we set $\sigma = \sigma_1 + \cdots + \sigma_n$, then $\int_\sigma f = \int_\gamma f$, Figure 8.3.

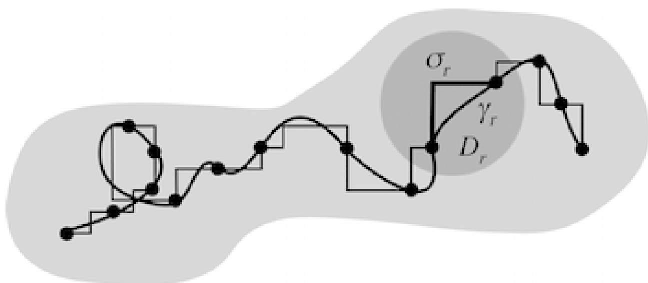

Figure 8.3 Reduction of a contour integral to an integral along a step path.

This reduces the investigation of $\int_\gamma f$ to the case of an integral along a step path σ, which can be attacked by geometrically inspired methods.

8.1 The Cauchy Theorem for a Triangle

At the end of the nineteenth century, amongst many different versions of Cauchy's Theorem, a most ingenious proof for a triangular contour was conceived by Eliakim Hastings Moore. Earlier proofs usually insisted that the function f should have a *continuous* derivative f'. By restricting the contour to a triangle, Moore's proof requires only that f' *exists* throughout D. It therefore provides a suitable basis for the development of the theory for all differentiable functions.

For $z_1, z_2, z_3 \in \mathbb{C}$, let $T(z_1, z_2, z_3)$ be the set of points inside and on the triangle with vertices z_1, z_2, z_3. Formally,

$$T(z_1, z_2, z_3) = \{z \in \mathbb{C} : z = \lambda_1 z_1 + \lambda_2 z_2 + \lambda_3 z_3, \lambda_j \in \mathbb{R}, \lambda_j \geq 0, \lambda_1 + \lambda_2 + \lambda_3 = 1\}$$

where $j = 1, 2, 3$.

The *boundary contour* of the triangle, composed of the three line segments that form its sides, is

$$\partial T(z_1, z_2, z_3) = [z_1, z_2] + [z_2, z_3] + [z_3, z_1]$$

Whenever there is no confusion we denote the triangle by T and its boundary by ∂T.

THEOREM 8.1 (Cauchy's Theorem for a Triangle). *Let f be a differentiable function in a domain D. If the triangle T lies in D, as in Figure 8.4, then $\int_{\partial T} f = 0$.*

Proof. Let $|\int_{\partial T} f| = c \geq 0$.

We prove that $c = 0$ by an indirect argument. First we subdivide T into four smaller triangles $T^{(1)}, T^{(2)}, T^{(3)}, T^{(4)}$ by joining the midpoints of the sides as in Figure 8.5.

We know that

$$\int_{\partial T} f = \sum_{r=1}^{4} \int_{\partial T^{(r)}} f$$

Therefore

$$c = \left| \int_{\partial T} f \right| \leq \sum_{r=1}^{4} \left| \int_{\partial T^{(r)}} f \right|$$

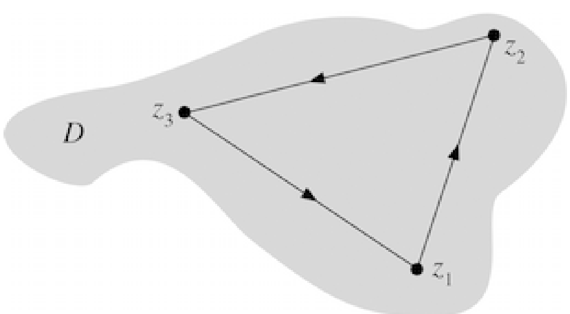

Figure 8.4 A triangular path in D.

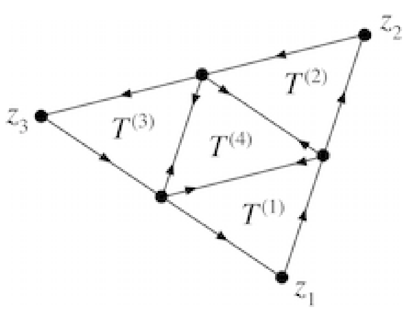

Figure 8.5 Subdividing the triangular path.

so we must be able to choose r such that

$$\left|\int_{\partial T^{(r)}} f\right| \geq \frac{c}{4}$$

(If more than one r satisfies this inequality, choose any of those – say the one with smallest r.) Define $T_1 = T^{(r)}$. Then

$$\left|\int_{\partial T_1} f\right| \geq \frac{c}{4} \quad \text{and} \quad L(\partial T_1) = \tfrac{1}{2}L(\partial T)$$

Repeat this process of subdivision to get a sequence of triangles

$$T \supseteq T_1 \supseteq T_2 \supseteq \cdots \supseteq T_n \cdots$$

satisfying

$$\left|\int_{\partial T_n} f\right| \geq (\tfrac{1}{4})^n c \quad \text{and} \quad L(\partial T_1) = (\tfrac{1}{2})^n L(\partial T) \tag{8.1}$$

Next we get another estimate for $|\int_{\partial T_n} f|$ using the fact that f is differentiable. Since triangles are bounded closed sets, the intersection of the nested sequence of triangles T_n contains a point $z_0 \in D$. Since f is differentiable at z_0, for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$0 < |z - z_0| < \delta \text{ implies } \left|\frac{f(z) - f(z_0)}{z - z_0}\right| < \varepsilon$$

Hence

$$0 < |z - z_0| < \delta \text{ implies } |f(z) - f(z_0) - f'(z_0)(z - z_0)| < \varepsilon|z - z_0|$$

For some integer N, every point of T_n is within δ of z_0 for all $n \geq N$. Thus

$$|f(z) - f(z_0) - f'(z_0)(z - z_0)| < \varepsilon|z - z_0| \quad \text{for all } z \in T_n, n \geq N$$

For $z \in T_n$, we trivially have $|z - z_0| \leq L(\partial T_n)$, so the Estimation Lemma 6.41 gives

$$\left|\int_{\partial T_n} [f(z) - f(z_0) - f'(z_0)(z - z_0)]\mathrm{d}z\right| \leq \varepsilon L(\partial T_n) \cdot L(\partial T_n) \tag{8.2}$$

But $-f(z_0) - f'(z_0)(z - z_0)$ is of the form $a + nz$, where a, b are constants. This has antiderivative $az + \frac{1}{2}bz^2$, so (8.2) reduces to

$$\left| \int_{\partial T_n} f \right| \le \varepsilon L(\partial T_n)^2$$

Comparing this with the earlier estimate (8.1) to get

$$(\tfrac{1}{4})^n c \le \left| \int_{\partial T_n} f \right| \le \varepsilon L(\partial T_n)^2 = (\tfrac{1}{4})^n \varepsilon L(\partial T_n)^2$$

giving

$$c \le \varepsilon L(\partial T)^2$$

But ε is arbitrary and $c \ge 0$, so we must have $c = 0$, that is

$$\int_{\partial T_n} f = 0 \qquad \square$$

This theorem deserves a commentary, because its analytic presentation obscures the fact that the basic ideas are very simple and very geometric. One is that the integral of f is *additive on contours.* That is, the contributions from the subdivided contours add up to that on the original one. The other is that a differentiable function is *approximately linear* (that is, of the form $a + bz$) near any point.

If it were possible to make it *exactly* linear, locally, then we could take a fine subdivision making it linear on each subcontour; get zero for the integral on each subcontour because there is an obvious antiderivative; then add all these zeros to get zero for the original integral.

Unfortunately this is not possible, and we are faced with adding a larger and larger number of contributions, each getting closer and closer to zero. But by *estimating* the rates of growth and shrinkage, we can show that errors introduced by assuming approximate linearity tend to zero fast enough to compensate for the increasing number of subcontours.

It is an interesting exercise to rewrite the proof in a way that turns this informal description of the argument into a formal proof that keeps the geometry to the fore.

8.2 Existence of an Antiderivative in a Star Domain

We begin with a formal definition, previewed in the introduction to this chapter:

DEFINITION 8.2. A domain D is a *star domain* if there exists $z_* \in D$, called a *star centre*, such that for all $z \in D$ the straight line segment $[z_*, z]$ lies in D.

(A star centre need not be unique. For example, a disc is a star domain and every point in the disc is a star centre.) $\square$

In a star domain there is an obvious candidate for an antiderivative of a function f, namely the integral $F(z) = \int_{[z_*, z]} f$. We now show that this is indeed an antiderivative, by applying Theorem 8.1.

THEOREM 8.3. *If f is differentiable in a star domain D with star centre z_*, then $F(z) = \int_{[z_*, z]} f$ is an antiderivative of f in D.*

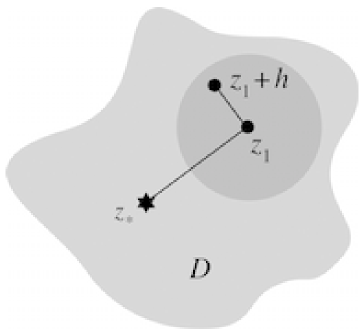

Figure 8.6 These three points in a star domain form a triangle inside the domain.

Proof. The domain D is open, so for any $z_1 \in D$ there exists $\varepsilon_1 > 0$ such that $N_{\varepsilon_1}(z_1) \subseteq D$. If $|h| < \varepsilon_1$, the triangle $T(z_*, z_1, z_1 + h)$ lies entirely in D, Figure 8.6.

Now Theorem 8.1 gives

$$\int_{[z_*,z_1]} f + \int_{[z_1,z_1+h]} f + \int_{[z_1+h,z_*]} f = 0$$

This can be written as

$$F(z_1) + \int_{[z_1,z_1+h]} f - F(z_1 + h) = 0$$

or

$$\frac{F(z_1) - F(z_1 + h)}{h} = \frac{1}{h} \int_{[z_1,z_1+h]} f$$

The proof now proceeds in the same manner as Theorem 6.44. For a constant $c \in \mathbb{C}$,

$$\int_{[z_1,z_1+h]} c \, \mathrm{d}z = ch$$

hence

$$\frac{F(z_1) - F(z_1 + h)}{h} - f(z_1) = \int_{[z_1,z_1+h]} \frac{f(z) - f(z_1)}{h} \mathrm{d}z \tag{8.3}$$

By continuity of f, given $\varepsilon > 0$ there exists $\delta > 0$ such that

$$|z - z_1| < \delta \text{ implies } |f(z) - f(z_1)| < \varepsilon$$

If z lies on the line segment $[z_1, z_1 + h]$,

$$|h| < \delta \text{ implies } |z - z_1| < \delta \quad \text{so } |f(z) - f(z_1)| < \varepsilon$$

The Estimation Lemma 6.41 gives

$$\left| \int_{[z_1,z_1+h]} \frac{f(z) - f(z_1)}{h} \mathrm{d}z \right| \le \frac{\varepsilon}{|h|} |h| = \varepsilon$$

and from (8.3), if $|h| < \delta$ then

$$\left| \frac{F(z_1) - F(z_1 + h)}{h} - f(z_1) \right| \le \varepsilon$$

Since ε is arbitrary,

$$\lim_{h\to 0} \frac{F(z_1) - F(z_1 + h)}{h} = f(z_1)$$

so $F' = f$. □

COROLLARY 8.4. *If f is differentiable in a star domain D, then $\int_\gamma f = 0$ for all closed contours γ in D, and the integral of f between any two points in D is independent of the choice of contour between the points.*

Proof. This is an immediate deduction from Theorems 8.3 and 6.44. □

8.3 An Example – the Logarithm

In Example 6.37 we showed that the integral of $1/z$ from -1 to 1 depends on the path between those points. We now examine how to restrict the domain to get round this problem. Later we describe a more elegant approach.

The function $1/z$ is differentiable in the domain $\mathbb{C} \setminus \{0\}$, but this is not a star domain; moreover, there is no antiderivative on this domain. However, if we restrict the function to any star domain $D \subseteq \mathbb{C} \setminus \{0\}$, the results of the previous section apply in D. It is easy to see that the cut plane $\mathbb{C}_\pi$ is a star domain, with star centre $z_* = 1$. Figure 8.7 makes this clear geometrically, and a rigorous proof is straightforward.

Because $1/z$ is differentiable in $\mathbb{C}_\pi$ it has an antiderivative there, the principal value of the logarithm, which satisfies

$$\operatorname{Log} z_1 = \int_{[1,z_1]} \frac{1}{z} \mathrm{d}z$$

We exploit the fact that this integral is independent of the path, provided the path is contained in $\mathbb{C}_\pi$, and integrate along a specially chosen contour that is not the line segment $[1, z_1]$. Instead we proceed as follows. Let $z_1 = re^{i\theta}$ where $r > 0, -\pi < \theta < \pi$,

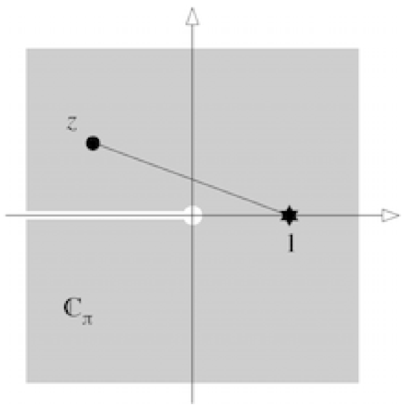

Figure 8.7 The cut plane $\mathbb{C}_\pi$ is a star domain with star centre $z_* = 1$.

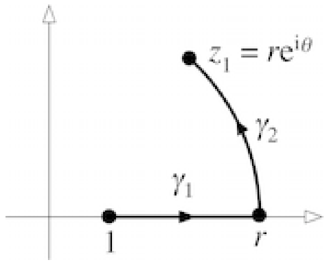

Figure 8.8 A more convenient path to integrate $1/z$ in $\mathbb{C}_\pi$.

and define $\gamma = \gamma_1 + \gamma_2$ where γ_1 is the line segment $[1, r]$ and $\gamma_2(t) = re^{it}$ $(t \in [0, \theta])$. See Figure 8.8.

Then

$$\begin{aligned}\operatorname{Log} re^{i\theta} &= \int_{\gamma_1} \frac{1}{z} dz + \int_{\gamma_2} \frac{1}{z} dz \\ &= \int_1^r \frac{1}{t} dt + \int_1^\theta \frac{1}{re^{it}} dt \\ &= \log r + i\theta\end{aligned}$$

This gives an alternative approach to the complex logarithm. In particular, it provides a far more satisfying proof of the continuity of the argument in the cut plane $\mathbb{C}_\pi$ than the prosaic version given in Section 7.2. Namely: the function Log is differentiable in $\mathbb{C}_\pi$, hence continuous, so its imaginary part (which is the argument of z) is also continuous there.

8.4 Local Existence of an Antiderivative

Let f be differentiable in an arbitrary domain D. We know that f may not have an antiderivative that works throughout D. But D is open, so for any $z_0 \in D$ there is an open disc $N_r(z_0) \subseteq D$. A function $F : N_r(z_0) \to \mathbb{C}$ such that $F'(z) = f(z)$ for all $z \in N_r(z_0)$ is called a *local antiderivative* of f. A disc is a star domain, so Theorem 8.3 tells us that f has a local antiderivative in every disc in D.

This immediately simplifies integration of a differentiable function along an arbitrary contour, because we can integrate along a step path instead:

LEMMA 8.5. *If γ is a contour from z_0 to z_1 in a domain D, then there exists a step path σ from z_0 to z_1 in D such that $\int_\gamma f = \int_\sigma f$ for every function f that is differentiable in D.*

Proof. By the Paving Lemma 2.33, $\gamma = \gamma_1 + \cdots + \gamma_n$ where each γ_r lies in a disc $D_r \subseteq D$. Let σ_r be a step path in D_r from the initial point of γ_r to its final point. Then by Corollary 8.4, in the star domain D_r, we have $\int_{\gamma_r} f = \int_{\sigma_r} f$. Now $\sigma_r = \sigma_1 + \cdots + \sigma_r$ is a step path from z_0 to z_1 in D, as in Figure 8.3, and

$$\int_\gamma f = \sum_{r=1}^{n} \int_{\gamma_r} f = \sum_{r=1}^{n} \int_{\sigma_r} f = \int_\sigma f \qquad \square$$

8.5 Cauchy's Theorem

We build up to Cauchy's Theorem in stages. First, consider a rectangle

$$R = \{x + iy \in \mathbb{C} : a \le x \le b, c \le y \le d\}$$

with boundary contour

$$\partial R = [z_1, z_2] + [z_2, z_3] + [z_3, z_4] + [z_4, z_1]$$

where $z_1 = a + ic, z_2 = b + ic, z_3 = b + id, z_4 = a + id$, as in Figure 8.9.

LEMMA 8.6. *If D is a domain, f is differentiable in D, and $R \subseteq D$, then $\int_{\partial R} f = 0$.*

Proof. Insert the opposite diagonal contours $[z_1, z_3]$ and $[z_3, z_1]$. Use Cauchy's Theorem for a triangle, Theorem 8.1, twice, and add. The integrals along the two diagonal paths cancel. See Figure 8.10. □

Now we take an arbitrary closed step path σ and insert extra line segments to make up a collection of rectangles. To do this, extend all horizontal and vertical line segments of σ to infinity, breaking the plane into a finite number of rectangles, some finite, $R_1, \ldots, R_k$, and some infinite, $R_{k+1}, \ldots, R_m$. (What we mean by 'rectangle' in

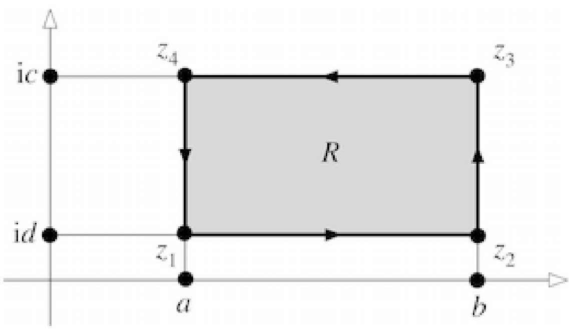

Figure 8.9 Contour defined by the boundary of a rectangle.

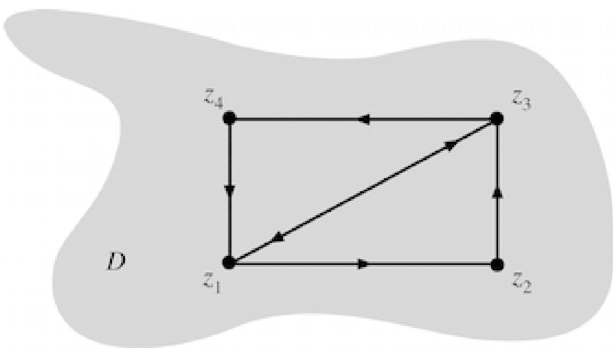

Figure 8.10 A rectangular contour is the sum of two triangular contours.

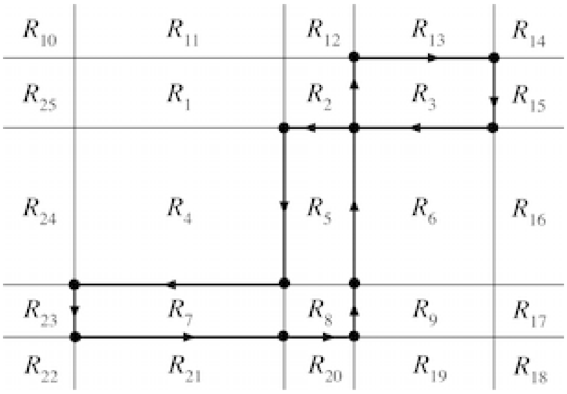

Figure 8.11 Extending the edges of a step path to tile $\mathbb{C}$ with finite or infinite rectangles.

the infinite case is clear from the figure. A 'side at infinity' is missing.) Figure 8.11 shows an example where $k = 9, m = 25$.

In the interior of each R_n choose a point z_n, and define

$$\nu_n = w(\sigma, z_n)$$

This is independent of the choice of z_n in R_n because the interior of R_n is connected.

Say that R_n is *relevant* if $\nu_n \neq 0$. Then R_n is relevant only when σ winds round it. In particular, all of the infinite rectangles $R_{k+1}, \dots, R_m$ are irrelevant, because they lie in the infinite component of the complement of σ. (In Figure 8.11 the only relevant rectangles are R_3, R_5, R_7, R_8.)

We now demonstrate that σ can be expressed in terms of the boundary contours of relevant rectangles, by taking ν_n copies of each boundary ∂R_n when $\nu_n > 0$, and $-\nu_n$ copies of each $-\partial R_n$ when $\nu_n < 0$.

For instance, in Figure 8.11 we take $-\partial R_3, \partial R_5, \partial R_7, \partial R_8$. Cancelling the opposite segments common to $\partial R_5, \partial R_8$ and to $\partial R_7, \partial R_8$ we finish up with the step contour σ.

To show this works with an arbitrary closed step path σ, it is convenient to use the notation $n\gamma$ to represent n copies of γ when $n \geq 0$ and $-n$ copies of $-\gamma$ when $n < 0$. The most straightforward procedure is to start with the set of contours

$$A = \{-\sigma, \nu_1 \partial R_1, \dots, \nu_k \partial R_k\}$$

and show that the cancellation of opposite line segments $L, -L$, wherever they occur, removes them all.

To prove this, suppose that when as many pairs as possible have been cancelled, There remain $q \neq 0$ copies of some line segment L. Then L is a side of at least one finite rectangle R_s, and by allowing q to be negative if necessary, we may suppose that L is traced in the same direction as ∂R_s. Let R_r be the rectangle on the other side of L (which may be a finite rectangle or an infinite one), as in Figure 8.12.

The set of closed contours

$$B = A \cup \{-q\partial R_s\}$$

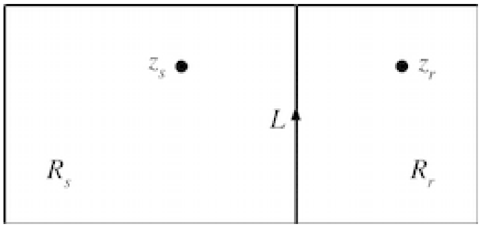

Figure 8.12 Rectangles either side of L.

then simplifies to have no copies of L. If we compute the winding numbers of the contours in B round $z_s \in R_s$ and $z_r \in R_r$, then the absence of L from the simplified set of contours tells us that the two winding numbers are the same. But the winding number round z_s is

$$\nu_1 w(\partial R_1, z_s) + \cdots + \nu_k w(\partial R_k, z_s) - w(\sigma, z_s) - qw(\partial R_s, z_r) = -q$$

and around z_r it is

$$\nu_1 w(\partial R_1, z_r) + \cdots + \nu_k w(\partial R_k, z_r) - w(\sigma, z_r) - qw(\partial R_s, z_r) = 0$$

Hence $q = 0$ as required.

This confirms that σ may be obtained by taking ν_n copies of each relevant contour ∂R_n and deleting opposite line segments wherever they occur.

LEMMA 8.7. *Let σ be a closed step path in a domain D such that $w(\sigma, z) = 0$ for all $z \notin D$. Then, for any function f differentiable in D, we have $\int_\sigma f = 0$.*

Proof. We express σ in terms of relevant rectangles, as above, and show that every relevant rectangle R_n must lie inside D. By definition, if z lies in the interior of R_n then $w(\sigma, z) = \nu_n \neq 0$ so $z \in D$. On the other hand, a point z on the boundary ∂R_n either lies on σ (and hence in D) or it is in the same component of the complement of σ as points in the interior of R_n, whence $w(\sigma, z) = \nu_n \neq 0$ and again $z \in D$.

By cancelling contributions along opposite contours,

$$\int_\sigma f = \sum_{n=1}^{k} \nu_n \int_{\partial R_n} f$$

Integrals on the right need be considered only when $\nu_n \neq 0$, in which case R_n is relevant so lies inside D. Then, by Lemma 8.6,

$$\int_{\partial R_n} f = 0$$

and $\int_\sigma f = 0$. □

We now reach the focal point of this chapter:

THEOREM 8.8 (Cauchy's Theorem). *Let f be differentiable in a domain D and let γ be a closed contour in D that does not wind round any points outside D (that is, $w(\gamma, z) = 0$ when $z \notin D$). Then $\int_\gamma f = 0$.*

Proof. By Lemma 8.5 there exists a step path σ in D such that $\int_\sigma \phi = \int_\gamma \phi$ for any function ϕ that is differentiable in D. In particular, $\int_\sigma f = \int_\gamma f$.

For $z_0 \notin D$ the function $\phi(z) = 1/(z - z_0)$ is also differentiable in D, so

$$w(\sigma, z_0) = \frac{1}{2\pi\mathrm{i}} \int_\sigma \phi = \frac{1}{2\pi\mathrm{i}} \int_\gamma \phi = w(\gamma, z_0) = 0$$

By Lemma 8.7, $\int_\sigma f = 0$. Hence $\int_\gamma f = \int_\sigma f = 0$. □

8.6 Applications of Cauchy's Theorem

The version of Cauchy's Theorem that we have just proved has far wider applications than simply showing that integrals round certain closed contours must be zero. It lets us calculate non-zero integrals as well. For example, suppose that γ_1 and γ_2 have the same winding number round all points outside D, so that $w(\gamma_1, z) = w(\gamma_2, z)$ when $z \notin D$. Let z_1 be the point where γ_1 begins and ends, and let z_2 be the point where γ_2 begins and ends. Take any contour σ from z_1 to z_2 in D, Figure 8.13.

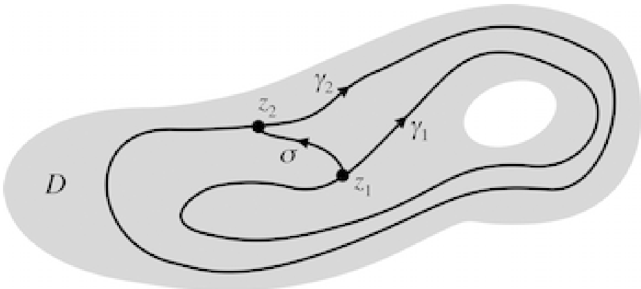

Figure 8.13 Creating a closed path from two paths.

Let $\gamma = \gamma_1 + \sigma - \gamma_2 - \sigma$. This is a closed contour in D, and $w(\gamma, z) = 0$ for $z \notin D$ because the winding number is additive. By Cauchy's Theorem, $\int_\gamma f = 0$. Therefore

$$\int_{\gamma_1} f + \int_\sigma f - \int_{\gamma_2} f - \int_\sigma f = 0$$

and

$$\int_{\gamma_1} f = \int_{\gamma_2} f$$

If we wish to compute $\int_{\gamma_1} f$, we may be able to find another contour γ_2 as above, for which $\int_{\gamma_2} f$ is simpler.

The technique of introducing opposite contours $\sigma, -\sigma$ whose contributions eventually cancel is also very useful. In particular, it lets us prove a much more powerful Cauchy-type theorem:

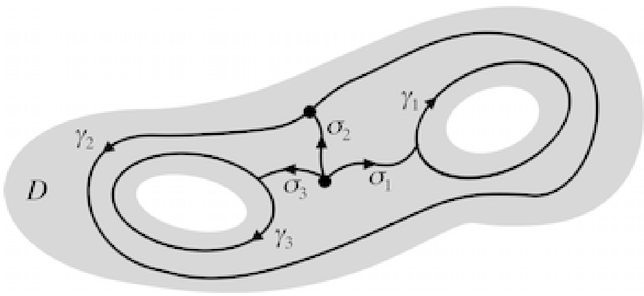

Figure 8.14 Joining a set of closed contours to form a single closed contour.

THEOREM 8.9 (Generalised Version of Cauchy's Theorem). *Suppose that $\gamma_1, \ldots, \gamma_n$ are closed contours in a domain D such that*

$$w(\gamma_1, z) + \cdots + w(\gamma_n, z) = 0 \quad \textit{for all } z \notin D$$

and let f be differentiable in D. Then

$$\int_{\gamma_1} f + \cdots + \int_{\gamma_n} f = 0$$

Proof. Suppose that γ_r begins and ends at z_r ($1 \leq r \leq n$). Choose any $z_0 \in D$ and contours $\sigma_1, \ldots, \sigma_n$ in D that join z_0 to $z_1, \ldots, z_n$ respectively, Figure 8.14.

Then

$$\gamma = \sigma_1 + \gamma_1 - \sigma_1 + \cdots + \sigma_n + \gamma_n - \sigma_n$$

is a closed contour beginning and ending at z_0, and

$$w(\gamma, z) = 0 \quad \text{for all } z \notin D$$

(again, by additivity of the winding number). By Cauchy's Theorem, $\int_\gamma f = 0$, so

$$\sum_{r=1}^{n} \int_{\sigma_r} f + \int_{\gamma_r} f - \int_{\sigma_r} f = 0$$

Therefore

$$\sum_{r=1}^{n} \int_{\gamma_r} f = 0$$

□

8.6.1 Cuts and Jordan Contours

Given two closed contours γ_1, γ_2 in D and a contour σ in D from a point on γ_1 to a point on γ_2, we call the pair of contours $\sigma, -\sigma$ a *cut* from γ_1 to γ_2. There is a historical reason for this name. Earlier versions of Cauchy's Theorem were invariably proved for Jordan contours. A *closed Jordan contour* is a closed contour $\gamma : [a, b] \to \mathbb{C}$ that does not cross itself. That is,

$$a \leq t_1 < t_2 < b \text{ implies } \gamma(t_1) \neq \gamma(t_2)$$

It is intuitively obvious, but difficult to prove analytically, that every closed Jordan contour γ separates the plane into two components, the points $O(\gamma)$ outside γ and the points $I(\gamma)$ inside γ, and that these are both connected sets, Figure 8.15.

Early versions of Cauchy's Theorem state that if γ and $I(\gamma)$ lie in D, then $\int_{\gamma} f = 0$. In applications it was then necessary to introduce 'cuts' to manufacture Jordan contours. For instance, suppose that f is differentiable everywhere except at z_0 and two Jordan contours γ_1, γ_2 both wind once round z_0 as in Figure 8.16 (left). Two cuts are made in Figure 8.16 (right) to create two new Jordan contours so that f is differentiable inside each of them. The integral round each new Jordan contour is then zero, and cancelling the contributions from the cuts we get $\int_{\gamma_1} f = \int_{\gamma_2} f$. Such methods usually relied on geometric intuition, sometime unsupported by analytic proof.

By introducing the winding number to link analysis to geometry, such pitfalls can be avoided, and there is no longer any necessity to restrict the theory to Jordan contours. Instead we can *define* the *inside* of any closed contour to be

$$I(\gamma) = \{z \in C : w(\gamma, z) \neq 0\}$$

and the *outside* to be

$$O(\gamma) = \{z \in C : w(\gamma, z) = 0\}$$

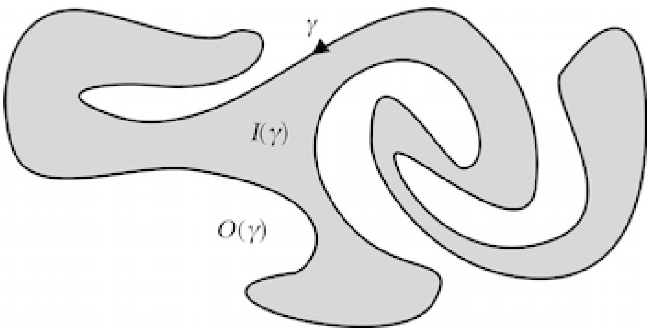

Figure 8.15 Closed Jordan contour and its inside and outside.

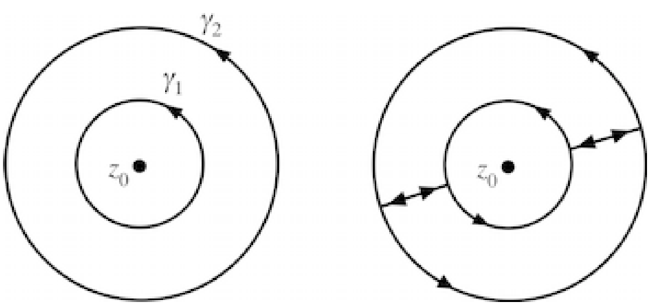

Figure 8.16 *Left*: Two Jordan contours. *Right*: 'Rewiring' the contours using cuts.

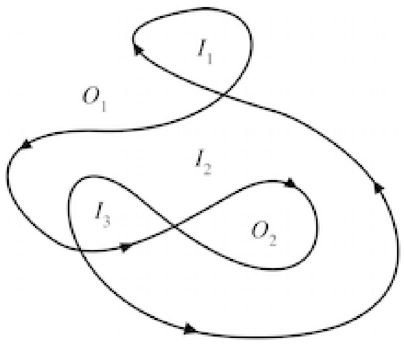

Figure 8.17 Inside and outside components of the complement of a closed contour.

In general, neither $I(\gamma)$ nor $O(\gamma)$ need be connected. For example, in Figure 8.17, $O(\gamma)$ has two components O_1 and O_2, and $I(\gamma)$ has three, I_1, I_2, and I_3.

The *Jordan Contour Theorem* (which we do not prove, but see Exercise 9 below for the step path case) then says that the outside and inside of a closed Jordan contour are both connected.

For an arbitrary closed contour we can rephrase Cauchy's Theorem as stated in Theorem 8.8 to get:

THEOREM 8.10. *Let f be differentiable in a domain D. If a closed contour γ and its inside $I(\gamma)$ lies in D, then $\int_\gamma f = 0$.* □

8.7 Simply Connected Domains

Theorem 6.44 states that if D is a domain and $f : D \to \mathbb{C}$ is differentiable in D, then $\int_\gamma f = 0$ for all closed contours γ in D if and only if f has an antiderivative in D. We can now state the precise conditions under which this happens for *all* functions f that are differentiable in D.

DEFINITION 8.11. A domain D is *simply connected* if $w(\gamma, z) = 0$ for every closed contour γ in D and every $z \notin D$.

Equivalently, $I(\gamma) \subseteq D$ for every closed contour γ in D. □

Intuitively, a domain is simply connected if it has no holes.

We now have:

THEOREM 8.12. *If D is a domain, then $\int_\gamma f = 0$ for all closed contours γ in D, if and only if D is simply connected.*

Proof. If D is simply connected then Cauchy's Theorem implies that $\int_\gamma f = 0$ for any closed contour γ in D and any $f : D \to \mathbb{C}$ that is differentiable in D. Conversely, if D is not simply connected, there exists a closed contour γ_0 in D and $z_0 \in D$ such that

$w(\gamma_0, z_0) \neq 0$. Let $\phi(z) = 1/(2\pi i(z - z_0))$. Then $\phi : D \to \mathbb{C}$ is differentiable in D and $\int_\gamma \phi = w(\gamma_0, z_0) \neq 0$. □

8.8 Exercises

1. State which of the following are star domains, proving the existence of a star centre for those that are, and justifying your answer for those that are not.

(i) $\{z \in \mathbb{C} : z \neq x + 0i \text{ where } |x| \geq 1\}$
(ii) $\{z \in \mathbb{C} : |z| > 1\}$
(iii) $\{z \in \mathbb{C} : z \neq e^{it} \text{ for } t \in [0, \pi]\}$
(iv) $\{z \in \mathbb{C} : |z| > 1 \text{ and either } \operatorname{im} z > 0 \text{ or } \operatorname{re} z > 0\}$

2. Let $D = \mathbb{C} \setminus \{0\}$. For $z_0 \in D$, specify a local antiderivative in some neighbourhood of z_0 for each of the following functions:

(i) $1/z$
(ii) $1/z^2$
(iii) $(z + 1)/z^2$
(iv) $(\cos z)/z$
(v) $(\sin z)/z$

3. Let

$$\gamma_1(t) = -1 + \tfrac{1}{2}e^{it} \ (t \in [0, 2\pi])$$
$$\gamma_2(t) = 1 + \tfrac{1}{2}e^{it} \ (t \in [0, 2\pi])$$
$$\gamma(t) = 2e^{-it} \ (t \in [0, 2\pi])$$

If $f(z) = 1/(z^2 - 1)$ use Theorem 8.9 to deduce that

$$\int_\gamma f = \int_{\gamma_1} f + \int_{\gamma_2} f$$

Interpret this statement in terms of the winding numbers of $\gamma, \gamma_1, \gamma_2$ round $1, -1$.

4. Show that $D = \{z \in \mathbb{C} : z \neq \pm 1\}$ is not simply connected. Let

$$L_1 = \{x + iy \in \mathbb{C} : y = 0, x \leq -1\}$$
$$L_2 = \{x + iy \in \mathbb{C} : y = 0, x \geq 1\}$$
$$D_0 = D \setminus \{L_1 \cup L_2)$$

Show that D_0 is simply connected. Is it a star domain? Does $f(z) = 1/(z^2 - 1)$ have an antiderivative in D_0? In each case justify your answer.

5. Prove that every quadrilateral whose edges do not cross is a star domain. What about pentagons? Hexagons?

6. Let $D = \{z \in \mathbb{C} : z \neq \pm i\}$ and let γ be a closed contour in D. Find all the possible values of $\int_\gamma 1/(z^2 + 1)dz$. If σ is a contour from 0 to 1, find all possible values of $\int_\sigma 1/(z^2 + 1)dz$.

7. Let $\gamma_1 = S_1 + L - S_2 - L, \gamma_2 = S_1 + L + S_2 - L$, where

$$S_1(t) = e^{it} \ (t \in [0, 2\pi])$$
$$S_2(t) = 2e^{it} \ (t \in [0, 2\pi])$$
$$L = [1, 2]$$

Describe the inside and outside of γ_1 and γ_2.

Let $f(z) = (\cos z)/z$, By writing $\cos z$ as a power series and considering $f(z) = (1/z) + g(z)$, or otherwise, compute $\int_{\gamma_1} f$ and $\int_{\gamma_2} f$. Compare the results with Theorem 8.10.

8. Let $D = \{z \in \mathbb{C} : z \neq z_1, \dots, z \neq z_k\}$ where $z_j \in \mathbb{C}$, and suppose that f is differentiable in D. Show that for any closed contour in D,

$$\int_\gamma f = \sum_{r-1}^{k} n_r \int_{S_r} f$$

where S_r is a sufficiently small circle centre z_r and n_r is an integer.

If $\lim_{z \to z_r}(z - z_r)f(z) = a_r \in \mathbb{C}$ for $r = 1, \dots, k$, show that

$$\int_\gamma f = \sum_{r-1}^{k} 2\pi i n_r a_r$$

9. *Jordan Contour Theorem for Step Paths.* Define a *crossing* of a closed step path $\sigma : [a, b] \to D$ to be a point $\sigma(t_1)$ in the image of σ, such that

$$\sigma(t_1) = \sigma(t_2) \quad (a \leq t_1 < t_2 < b)$$

Say that σ is *simple* if it has no crossings.

Prove the Jordan Contour Theorem for step paths: Let σ be a simple closed step path in $\mathbb{C}$. Then

(i) $\mathbb{C} \setminus \hat{\sigma}$ has precisely two connected components. One (call it $I(\sigma)$) is bounded, the other (call it $O(\sigma)$) is unbounded.
(ii) Either $w(\sigma, z) = 1$ for all $z \in I(\sigma)$, or $w(\sigma, z) = -1$ for all $z \in I(\sigma)$.
(iii) If $z \in O(\sigma)$, then $w(\sigma, z) = 0$.

10. *Cauchy's Theorem for a Simple Closed Step Path.* Prove Cauchy's Theorem for a simple closed step path σ: If σ does not wind round any point not in D, then $\int_\sigma f = 0$.

Hint: One way to do this is as follows:

(1) Construct a set of rectangles R_r as in Section 8.4.
(2) Prove that a rectangle R_r is relevant if and only if it lies in $I(\sigma)$.
(3) Prove that $I(\sigma) \subseteq D$, observing that by definition $I(\sigma)$ winds around any point in $I(\sigma)$.
(4) Prove that $\int_{R_r} f = 0$, by observing that R_r is a star domain.
(5) Define a closed path τ_r round the boundary of each rectangle, having the same winding number as σ round any point inside the rectangle. (This is either always 1 or always -1 by Exercise 9 (ii).)

(6) Show that

$$\sum_r \int_{R_r} f = \int_\sigma f$$

because edges of the R_r that are not contained in the image of σ cancel is pairs.

(7) Deduce Cauchy's Theorem for σ.

11. Suppose that σ_1, σ_2 are simple closed step paths in $\mathbb{C}$, and that the image of σ_1 is contained in $I(\sigma_2)$. Prove that $I(\sigma_1) \subseteq I(\sigma_2)$.

9 Homotopy Versions of Cauchy's Theorem

The results in this chapter are not essential for any later parts of the text, except for Chapter 16.

Cauchy's Theorem is one of the foundation stones of complex analysis. It resolves two apparently conflicting features of contour integrals $\int_\gamma f$ when f is fixed but γ can change. On the one hand, γ can be altered in fairly drastic ways with no effect on the integral – for instance, replacing a general path by a step path. On the other hand, changing a semicircular path in the upper half-plane of $\mathbb{C}$ into a semicircular path in the lower half-plane of $\mathbb{C}$ completely changes the integral of $1/z$ between -1 and 1, as in Section 6.10.

The version of Cauchy's Theorem in Theorem 8.8 explains these features: what really matters is the winding number of γ around points that lie outside the domain D of f. This result emphasises the topology of the domain, and how the path lies within it. To improve our understanding, we examine these topological issues in more detail. In fact, we do so in two ways. In this chapter we consider continuous deformations of γ, captured by the topological notion of homotopy. In Chapter 16 we discuss a related concept, homology. Both concepts formalise the idea that the domain D has 'holes', and the integral depends on how the path wanders around the domain, relative to these holes. However, they do this in two different, though related, ways. Homotopy is easier to visualise and geometrically quite natural. Homology is algebraically simpler, once some initial difficulties have been overcome, but it can appear artificial and contrived.

We reformulate Cauchy's Theorem from these two viewpoints. Each offers fresh insights. We also show that an arbitrary path γ in a domain D can be replaced by a suitable approximating polygon without changing the integral of any function that is differentiable in D. This generalises Cauchy's Theorem so that it applies to any closed path, removing the 'piecewise smooth' condition required of a closed contour.

9.1 Informal Description of Homotopy

Homotopy is one of the central ideas in topology. It describes topological features of spaces in terms of families of continuously varying paths. For example, consider the

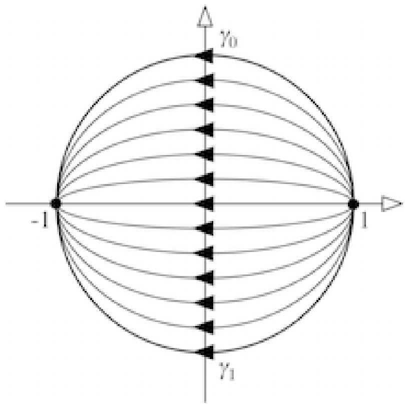

Figure 9.1 Deforming the semicircular path γ_0 continuously into γ_1.

anticlockwise semicircle γ_0 in $\mathbb{C}$ from 1 to -1 defined by

$$\gamma_0(t) = e^{it} \quad (t \in [0, \pi])$$

and the clockwise semicircle γ_1 in $\mathbb{C}$ from 1 to -1 defined by

$$\gamma_1(t) = e^{-it} \quad (t \in [0, \pi])$$

Figure 9.1 shows visually that we can deform γ_0 into γ_1 by displacing it vertically. In formulas, let

$$\gamma_s(t) = (1 - s)e^{it} + se^{-it}$$

Then γ_0 and γ_1 are the two semicircular paths, and the path γ_s varies continuously as s varies continuously from 0 to 1.

However, if the paths are not allowed to pass through the origin (for example, they may be paths in the domain of a function that has a singularity at the origin, such as $1/z$), this continuous deformation is not permitted, because $\gamma_{1/2}$ passes through the origin at $t = \pi/2$. A little experiment strongly suggests that no continuous deformation from γ_0 to γ_1 exists if no intermediate path meets the origin. The origin is an obstacle to deforming the path, and any attempt to do so causes the path to get 'hung up' at the origin. The origin creates a hole, and the path cannot cross the hole.

This suggestion is in fact true, but a proof requires some care: it can, for example, be based on properties of the winding number. Indeed, the winding number is a simple example of a homotopy invariant: a way to distinguish topologically different spaces using notions of homotopy.

In this chapter we apply homotopy ideas to complex integration by considering what happens to $\int_\gamma f$ when we allow the contour γ to vary continuously. We derive precise conditions under which γ can be deformed continuously without changing the integral. This can be done in two ways. One is to fix the end points z_0, z_1 and allow the contour between them to be continuously deformed. The other is to perform a continuous deformation of a closed contour. In both cases the deformations must keep the path inside a domain D on which f is differentiable. Figure 9.2 illustrates these two possibilities.

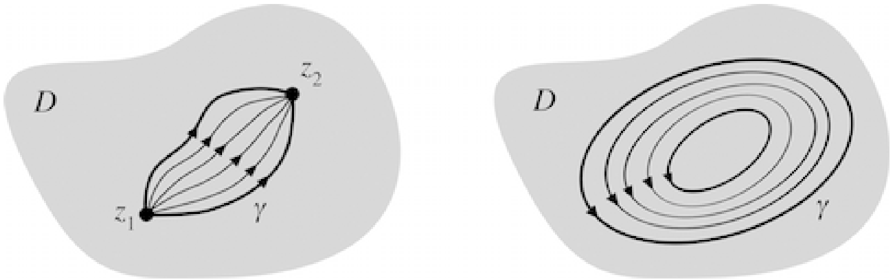

Figure 9.2 Two types of continuous deformation of a contour. *Left*: Keeping end points z_1, z_2 of γ fixed. *Right*: Deforming a closed contour γ.

Both of these methods are special cases of a single result, The Cauchy Theorem for a boundary, which is proved in Section 9.3. Before this, in Section 9.2 we show how the conditions on γ can be relaxed when f is differentiable to define $\int_\gamma f$ along any path (recall that a contour is piecewise smooth). This lets us vary γ freely without having to worry whether the intermediate paths involved are contours.

9.2 Integration Along Arbitrary Paths

Suppose that $f : D \to \mathbb{C}$ is differentiable in the domain D. Up to now we have usually integrated such a function along a *contour* in D – a piecewise smooth path. However, because differentiable functions, are fairly well behaved, the precise path taken turns out not to be so important. We saw this in Lemma 8.5, where we were able to replace a contour by a step path and get the same result. We can use the same technique to define an integral along an arbitrary path $\gamma : [a, b] \to D$. This definition agrees with the Riemann integral if that happens to be defined – an application of Cauchy's Theorem that we leave to the interested reader.

The Paving Lemma gives a partition $a = t_0 < t_1 < \cdots < t_n = b$ such that each subpath γ_r defined on $[t_{r-1}, t_r]$ lies in a disc $D_r \subseteq D$. Let λ be the approximating polygon $\lambda = \lambda_1 + \cdots + \lambda_n$, where λ_r is the line segment from $\gamma(t_{r-1})$ to $\gamma(t_r)$. Then λ_r lies inside D_r for all r, so λ lies inside D, and we can define the integral of f along γ to be

$$\int_\gamma f = \int_\lambda f$$

See Figure 9.3.

This definition can be proved to be independent of the choice of the partition that determines λ, provided these are chosen using the Paving Lemma as described above. For if extra division points are introduced between t_{r-1} and t_r, say

$$t_{r-1} < s_1 < \cdots < s_k < t_r$$

then $\gamma(s_1), \ldots, \gamma(s_k) \in D$ so the integral of f along λ_r has the same value as that along the polygon with vertices $\gamma(t_{r-1}), \gamma(s_1), \ldots, \gamma(s_k), \gamma(t_r)$, by Theorems 6.33 and 8.3. See Figure 9.4.

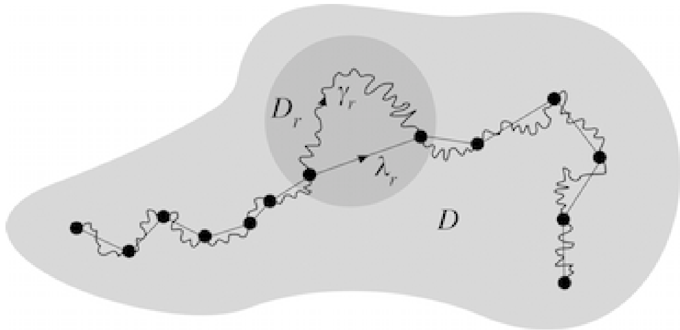

Figure 9.3 Replacing a continuous path by a polygonal one.

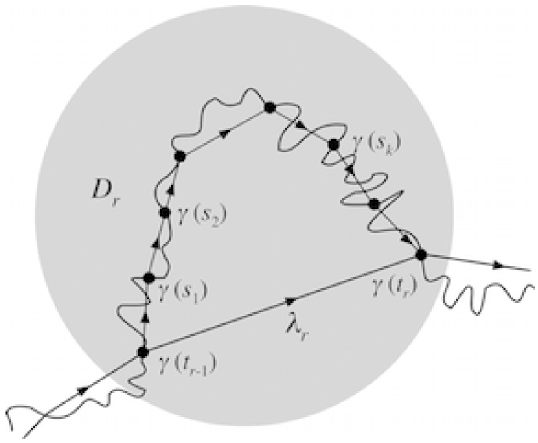

Figure 9.4 Subdividing a polygonal path inside a disc in D.

Thus a finer partition of $[a, b]$ does not change the integral. Given any two polygonal approximations λ, μ obtained using the Paving Lemma, let ν be the polygon whose vertices are those of λ, μ taken together. Then $\int_\lambda f = \int_\nu f = \int_\mu f$, so the definition of the integral is independent of the polygonal approximation – provided that is chosen using the Paving Lemma.

An arbitrary polygonal approximation to γ will not do. In Figure 9.5, suppose that $D = \mathbb{C} \setminus \{z_0\}$. The polygon λ with vertices $z_1, z_2, z_3, z_4, z_5, z_6, z_7$ is an approximation to γ found using the Paving Lemma, but the polygon with vertices z_1, z_6, z_7 is not. If $f = 1/(z - z_0)$ then $\int_\lambda f - \int_\mu f = 2\pi\mathrm{i}$, so $\int_\lambda f \neq \int_\mu f$.

It is straightforward to check that Theorems 8.8, 8.9, 8.10, and 8.12 of Chapter 8 also hold for arbitrary paths, however wild – even 'space-filling' curves, Section 2.9. For instance, if f is differentiable in D and does not wind round points outside D, then $\int_\gamma f = 0$. With such results at our disposal we can widen the scope of our ideas and introduce general notions from topology. These provide further insight into the various versions of Cauchy's Theorem and how they relate to each other. We return to the topological theme and develop it further in Chapter 16, using results from Section 9.3.

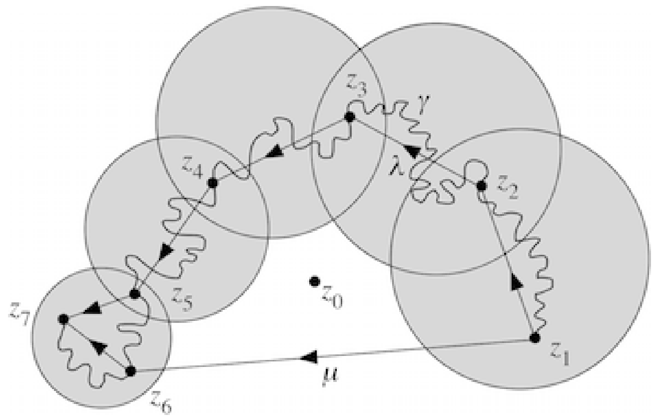

Figure 9.5 A different polygon can lead to a different value for the integral.

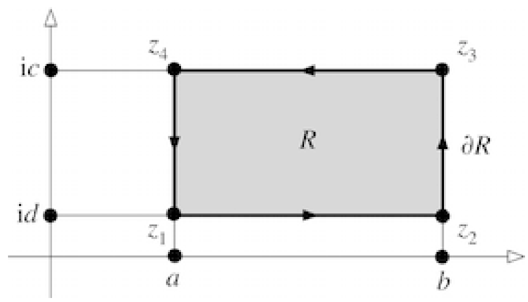

Figure 9.6 A rectangle and the path formed by its perimeter.

9.3 The Cauchy Theorem for a Boundary

The formal definition of homotopy (Definition 9.7 below) uses a map from a rectangle into $\mathbb{C}$. The top and bottom edges of the rectangle determine two paths, and parallel segments of its interior determine a continuous deformation of one path into the other. We therefore begin by considering maps defined on rectangles, and properties of their boundaries.

Let R be the rectangle

$$R = \{x + \mathrm{i}y \in \mathbb{C} : a \le x \le b, c \le y \le d\}$$

with boundary contour ∂R, Figure 9.6. (Here $a \le b$ and $c \le d$ are real.) Suppose that $\partial R : [0, p] \to \mathbb{C}$ is parametrised by arc length, measured anticlockwise, where $p = 2(b - a) + 2(d - c)$ is the perimeter of R, Figure 9.7.

Given a continuous map $\phi : R \to \mathbb{C}$, we define the *boundary* of ϕ to be

$$\partial\phi(t) = \phi(\partial R(t)) \quad (t \in [0, p])$$

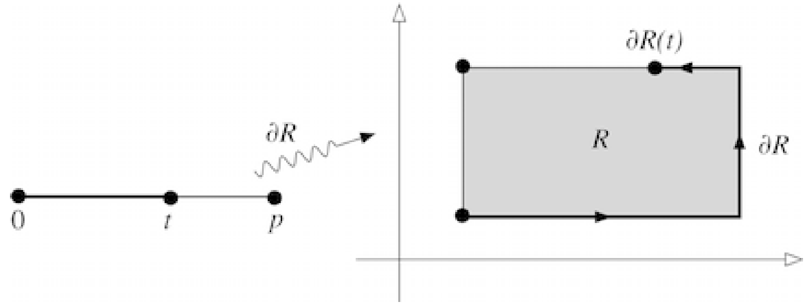

Figure 9.7 Parametrising the perimeter by arc length.

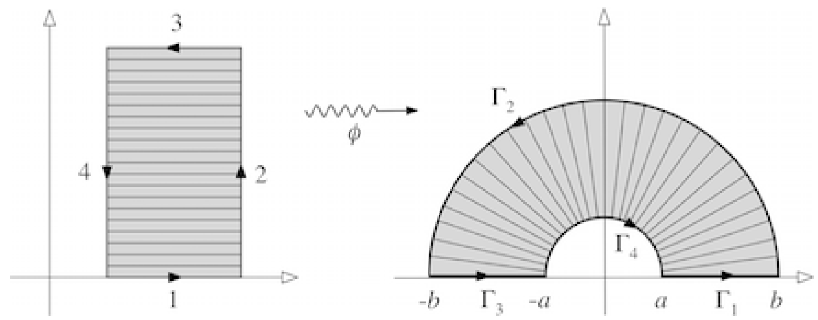

Figure 9.8 Boundary of the map ϕ in Example 9.2.

Example 9.1. If $\phi(x+\mathrm{i}y) = x\mathrm{e}^{\mathrm{i}\pi y}$ $(a \leq x \leq b, c \leq y \leq d)$ then $\partial\phi = \Gamma_1 + \Gamma_2 + \Gamma_3 + \Gamma_4$ as in Figure 9.8. □

REMARK 9.2. Here we are using notation in a flexible way by writing Γ_j for the image of Γ_j. Technically this is sloppy, but it avoids undue proliferation of 'image' symbols. Once we understand that a path differs from its image, we have earned the right to be sloppy. We exert that right repeatedly. □

Example 9.3. Let $R = \{x + \mathrm{i}y : 0 \leq x \leq 2, 0 \leq y \leq 2\}$. For $x + \mathrm{i}y \in R$ and $x + y \geq 1$, define $\phi(x + \mathrm{i}y) = x + \mathrm{i}y$, and for $x + \mathrm{i}y \in R$ and $x + y < 1$, define $\phi(x + \mathrm{i}y)$ to be the reflection of $x + \mathrm{i}y$ in the line $x + y = 1$. Then the effect of ϕ is to fold over the bottom left-hand corner of R as in Figure 9.9. This shows that the image of $\partial\phi$ need not be the boundary of the image $\phi(R)$. □

However, Figure 9.9 provides the clue that all points inside $\partial\phi$ lie in the image $\phi(R)$. We prove this is true in general:

LEMMA 9.4. *If $\phi : R \to \mathbb{C}$ is continuous, then $I(\partial\phi) \subseteq \phi(R)$.*

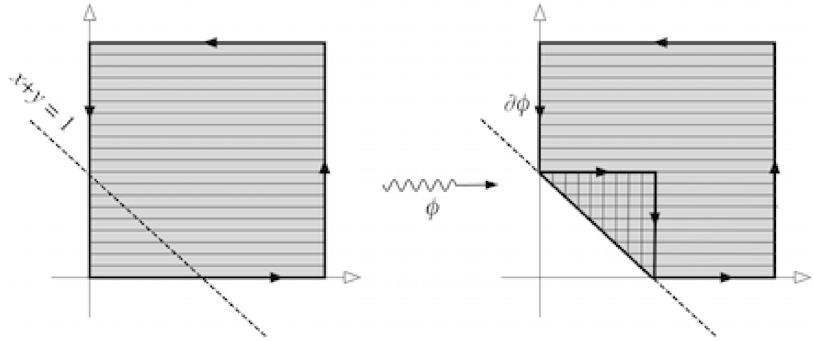

Figure 9.9 Boundary of the map ϕ in Example 9.3.

Proof. Suppose to the contrary that there exists $z_0 \in I(\partial\phi)$ but $z_0 \notin \phi(R)$. Let $D = \mathbb{C} \setminus \{z_0\}$. Then D is a domain and $\phi(R) \subseteq D$. If $f(z) = 1/(2\pi i(z - z_0))$ then f is differentiable in D and

$$\int_{\partial\phi} f = w(\partial\phi, z_0)$$

is a non-zero integer.

Subdivide the rectangle R by its vertical and horizontal bisectors into four equal rectangles $R^{(1)}, R^{(2)}, R^{(3)}, R^{(4)}$ and let $\phi^{(r)}$ be the restriction of ϕ to $R^{(r)}$. Then for $r = 1, 2, 3, 4$ the boundaries $\partial\phi^{(r)}$ are all closed paths in D. Because

$$\int_{\partial\phi} f = \sum_{r=1}^{4} \int_{\partial\phi^{(r)}} f \quad \text{and} \quad \int_{\partial\phi^{(r)}} f = w(\partial\phi^{(r)}, z_0)$$

at least one of the four integrals is a non-zero integer. Denote the corresponding rectangle $R^{(r)}$ by R_1 and the restriction of ϕ to R_1 by ϕ_1. Similarly dividing R_1 into four equal rectangles and repeating the process yields a nested sequence of rectangles

$$R \supseteq R_1 \supseteq \cdots \supseteq R_n \supseteq \cdots$$

where each $\int_{\partial\phi_n} f$ is a non-zero integer.

The sequence of rectangles contains a (unique) point $z_1 \in R$. For $\varepsilon = |\phi(z_1) - z_0|$, we have $N_\varepsilon(\phi(z_1)) \subseteq D$. By continuity of ϕ there exists $\delta > 0$ such that

$$\phi(N_\delta(\phi(z_1) \cap R) \subseteq N_\varepsilon(\phi(z_1))$$

For suitably large N we have $R_N \subseteq N_\delta(\phi(z_1))$, so $\partial\phi_N$ is a closed path in the disc $N_\varepsilon(\phi(z_1))$. But f is differentiable in $N_\varepsilon(\phi(z_1))$, a star domain,. So $\int_{\partial\phi_N} f = 0$ by Corollary 8.4, contradicting $\int_{\partial\phi_n} f$ being a non-zero integer. □

We can now prove an important variant of Cauchy's Theorem. But first (having decided not to rely on the topological notion of compactness) we need a two-dimensional analogue of the Paving Lemma.

PROPOSITION 9.5. *Let R be a rectangle and let $\phi : R \to D$ be a continuous map, with $D \subseteq \mathbb{C}$ a domain. Let R be divided into n^2 smaller rectangles, by subdividing its* *sides into n equal parts. Then for sufficiently large n, the image of every subrectangle is contained in some disc in D.*

Proof. The proof uses the same ideas as that of Lemma 2.33. Say that a rectangle is *pavable* if if it can be bisected repeatedly a finite number of times into closed subrectangles, so that the image of every subrectangle obtained in this manner lies inside a disc in D. Otherwise, the interval is *unpavable*.

Suppose the result is false. Then R is unpavable. Split it into four closed subrectangles by bisecting the sides. At least one these subrectangles, call it R_1, must be unpavable. Bisect R_1 to obtain an unpavable rectangle R_2, and so on. This sequence continues indefinitely because we are assuming the result false, so we get a nested sequence of rectangles

$$R \supseteq R_1 \supseteq R_2 \supseteq \cdots$$

each half the size of the previous one. The intersection of these rectangles is a unique point $z_0 \in R$.

However, D is open so there is a disc $N_\varepsilon(\phi(z_0)) \subseteq D$. If k is large enough, $R_k \subseteq N_\varepsilon(\phi(z_0))$, a contradiction.

Now take $n = 2^k$. □

THEOREM 9.6 (Cauchy's Theorem for a Boundary). *If $\phi : R \to D$ is a continuous map from a rectangle R into a domain D, as in Figure 9.10, and f is differentiable in D, then*

$$\int_{\partial\phi} f = 0$$

Proof. Cut R up into smaller rectangles R_{pq} by dividing it into n equal parts horizontally and vertically, where $1 \le p \le n, 1 \le q \le n$. By Proposition 9.5 this can be done so that each $\phi(R_{pq})$ is contained in a disc $D_{pq} \subseteq D$. See Figure 9.11.

Let ϕ_{pq} be the restriction of ϕ to R_{pq}. Then its boundary $\partial\phi_{pq}$ is a closed path in the disc $D_{pq} \subseteq D$, so

$$\int_{\partial\phi_{pq}} f = 0$$

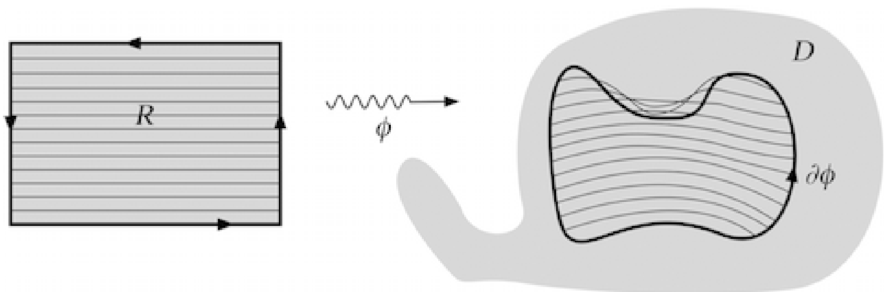

Figure 9.10 Effect of map ϕ on rectangle R.

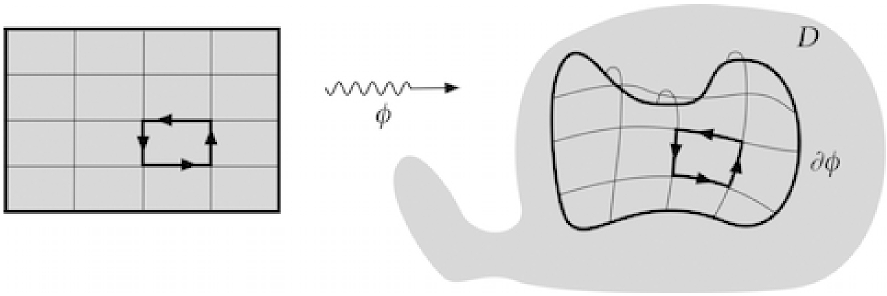

Figure 9.11 Subdividing the rectangle R, and the effect of ϕ.

Adding up all the integrals for $1 \le p \le n, 1 \le q \le m$, and cancelling integrals along opposite paths inside $\phi(R)$, we are left with

$$\int_{\partial\phi} f = 0$$

□

9.4 Formal Definition of Homotopy

A 'homotopy' between two paths $\gamma_0 : [a,b] \to D$ and $\gamma_1 : [a,b] \to D$ is, roughly speaking, a *continuously varying family* of paths $\gamma_s : [a,b] \to D$, where s runs over the interval $[0,1]$. At the start, $s = 0$ and $\gamma_s = \gamma_0$; at the end, $s = 1$ and $\gamma_s = \gamma_1$.

How do we make this precise? Notice that the whole family depends on two variables: the parameter, s, and the original variable t giving the position on the path. We can combine everything into a single function γ of (t, s) if we set

$$\gamma(t,s) = \gamma_s(t) \quad (t \in [a,b], s \in [0,1])$$

It is then natural to insist that γ, as a function of two real variables, should be continuous. We are thus led to:

DEFINITION 9.7. A *homotopy* in D between γ_0 and γ_1, as above, is a continuous map

$$\gamma : [a,b] \times [0,1] \to D$$

such that

$$\gamma(t,0) = \gamma_0(t) \quad \text{for all } t \in [a,b]$$
$$\gamma(t,1) = \gamma_1(t) \quad \text{for all } t \in [a,b]$$

□

We illustrate this idea in Figure 9.12. Note that we do not (yet) assume that γ_0 and γ_1 have the same start or end points.

It will prove convenient to think of $[a,b] \times [0,1]$ as a subset of $\mathbb{C}$, by identifying (t, s) with $t + \mathrm{i}s$.

Figure 9.12 captures well the continuous variation of γ_s, but it is misleadingly nice in that γ is one-one. There is no reason to require this in general, so a perfectly reasonable homotopy could resemble Figure 9.13.

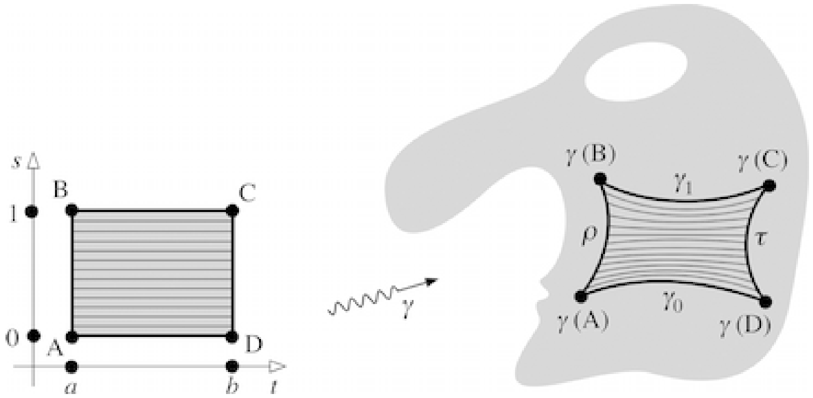

Figure 9.12 A homotopy between two paths γ_0 and γ_1.

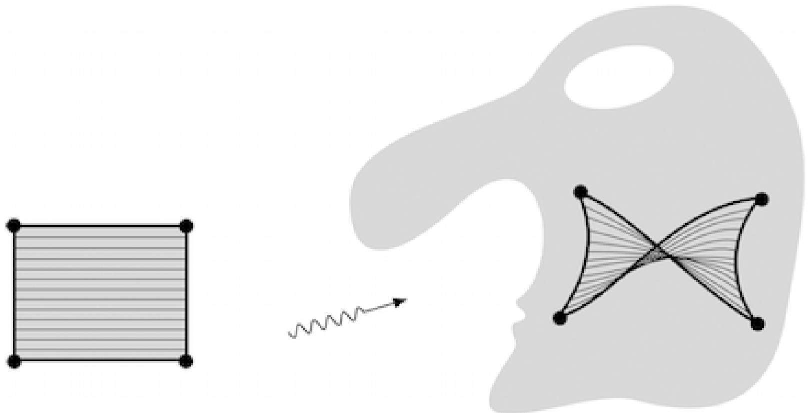

Figure 9.13 A homotopy that is not one-one between two paths γ_0 and γ_1.

Geometrically, of course, $[a, b] \times [0, 1]$ is a rectangle, and its boundary is a closed path, with corners. If we parametrise the four edges of the rectangle in some way, and then use γ to map the result into D, we obtain paths in D that join up to give a closed path, as for example in Figure 9.12.

Our aim is to apply Cauchy's Theorem for a boundary to this set of paths. Now, there is a problem: the two edges that give γ_0 and γ_1 are obviously useful, but the other two edges, marked ρ, τ in Figure 9.12, are going to be a nuisance. We therefore add conditions that eliminate them. There are two obvious ways to do this:

(a) Insist that each of ρ and τ goes to a single point in D.
(b) Insist that ρ and τ cancel each other out.

These two conditions yield two more restricted types of homotopy: fixed end point homotopy, and closed path homotopy. We describe these in detail in the next two sections.

9.5 Fixed End Point Homotopy

Consider the rectangle

$$R = \{t + \mathrm{i}s \in \mathbb{C} : t \in [a, b], s \in [0, 1]\}$$

DEFINITION 9.8. Two paths $\gamma_0 : [a, b] \to D$ and $\gamma_1 : [a, b] \to D$ are *fixed end point homotopic* in D if there is a continuous map

$$\phi : R \to D$$

and points $z_0, z_1 \in \mathbb{C}$ such that

$$\begin{aligned} \phi(t, 0) &= \gamma_0(t) && \text{for all } t \in [a, b] \\ \phi(t, 1) &= \gamma_1(t) && \text{for all } t \in [a, b] \\ \phi(s, 0) &= z_0 && \text{for all } s \in [0, 1] \\ \phi(s, 1) &= z_1 && \text{for all } s \in [0, 1] \end{aligned}$$

as in Figure 9.14. □

If we let $\gamma_s(t) = \phi(t, s)$ then γ_s is a path in D from z_0 to z_1 for all $s \in [0, 1]$, and as s increases continuously from 0 to 1, the path γ_s 'deforms continuously' from γ_0 to γ_1.

Example 9.9. Let $D = \{z \in \mathbb{C} : |z| < 2\}, \gamma_0(t) = t \ (t \in [-1, 1]), \gamma_1(t) = \mathrm{e}^{\frac{1}{2}\pi \mathrm{i}(t-1)}$ $(t \in [-1, 1])$. Then γ_0 and γ_1 are fixed end point homotopic in D, where

$$\phi(t, s) = (1 - s)\gamma_0(t) + s\gamma_1(t) \quad (t \in [-1, 1], s \in [0, 1])$$ □

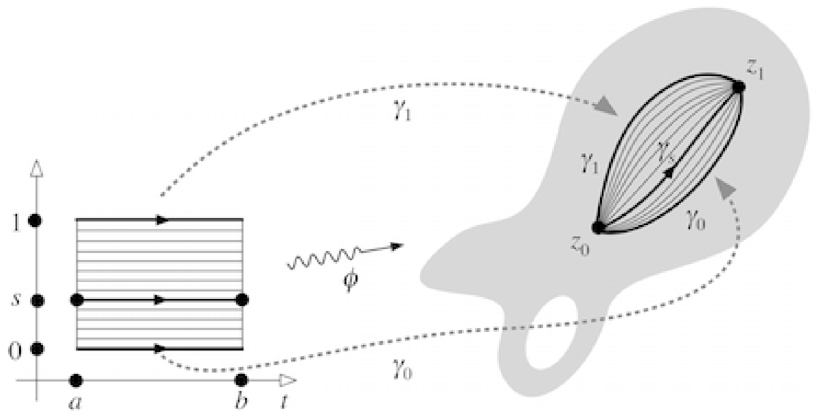

Figure 9.14 A fixed end point homotopy.

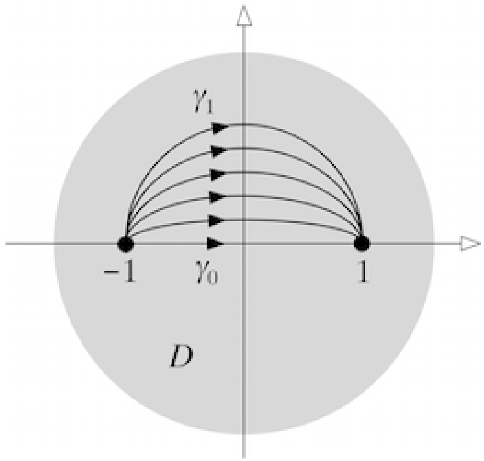

Figure 9.15 A fixed end point homotopy between γ_0 and γ_1.

As a corollary of Theorem 9.6 we deduce:

THEOREM 9.10. *If f is differentiable in a domain D and γ_0 is fixed end point homotopic in D to γ_1, then $\int_{\gamma_0} f = \int_{\gamma_1} f$.*

Proof. We have a continuous map $\phi : R \to D$ where

$$\begin{aligned} \phi(t,0) &= \gamma_0(t) && \text{for all } t \in [a,b] \\ \phi(t,1) &= \gamma_1(t) && \text{for all } t \in [a,b] \\ \phi(0,s) &= z_0 && \text{for all } s \in [0,1] \\ \phi(1,s) &= z_1 && \text{for all } s \in [0,1] \end{aligned}$$

as in Figure 9.15.

If $p_r : [0,1] \to D$ is the path $p_r(t) = z_r$ for $r = 0, 1$, whose image is a single point $\{z_r\}$, then $\int_{p_r} f = 0$ and $\partial\phi = \gamma_0 + p_1 - \gamma_1 + p_0$. By Cauchy's Theorem for a boundary,

$$\int_{\partial\phi} f = \int_{\gamma_0} f + \int_{p_1} f - \int_{\gamma_1} f + \int_{p_0} f = 0$$

so

$$\int_{\gamma_0} f = \int_{\gamma_1} f$$

□

9.6 Closed Path Homotopy

Once more we consider the rectangle

$$R = \{t + \mathrm{i}s \in \mathbb{C} : t \in [a,b], s \in [0,1]\}$$

but we impose different conditions on ϕ:

DEFINITION 9.11. Two (closed) paths $\gamma_0 : [a,b] \to D$ and $\gamma_1 : [a,b] \to D$ are *homotopic via closed paths* in D if there is a continuous map

$$\phi : R \to D$$

such that

$$\begin{aligned} \phi(t,0) &= \gamma_0(t) && \text{for all } t \in [a,b] \\ \phi(t,1) &= \gamma_1(t) && \text{for all } t \in [a,b] \\ \phi(a,s) &= \phi(b,s) && \text{for all } s \in [0,1] \end{aligned}$$

as in Figure 9.16. □

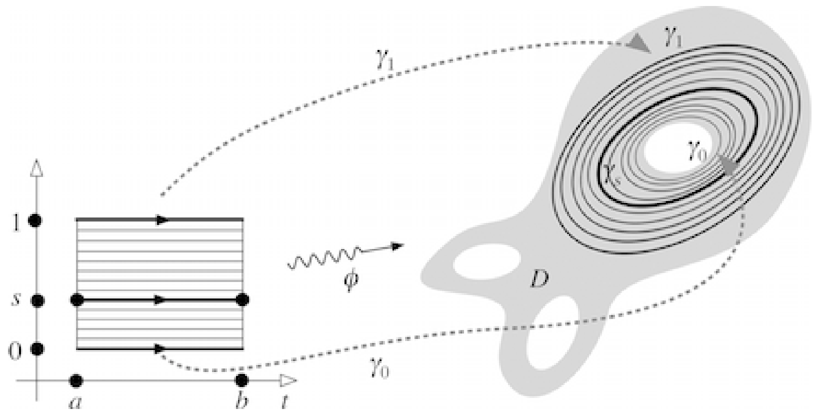

Figure 9.16 Homotopy via closed paths.

Again, if we define $\gamma_s(t) = \phi(t,s)$ then γ_s is a closed path in D for all $s \in [0,1]$, and as s increases continuously from 0 to 1, the path γ_s 'deforms continuously' from γ_0 to γ_1.

Example 9.12. For $|z_0| < K$, let $D = \{z \in \mathbb{C} : |z| < K, z \neq z_0\}$. For $|z_0| < \rho < K$, let $\gamma_0(t) = \rho e^{it}$ ($t \in [0,2\pi]$), and for $0 < \varepsilon < K - |z_0|$ let $\gamma_1(t) = z_0 + \varepsilon e^{it}$ ($t \in [0,2\pi]$), as in Figure 9.17. Then γ_0 is homotopic to γ_1 via closed paths in D, where

$$\phi(t,s) = (1-s)\gamma_0(t) + s\gamma_1(t) \quad (t \in [0,2\pi], s \in [0,1])$$

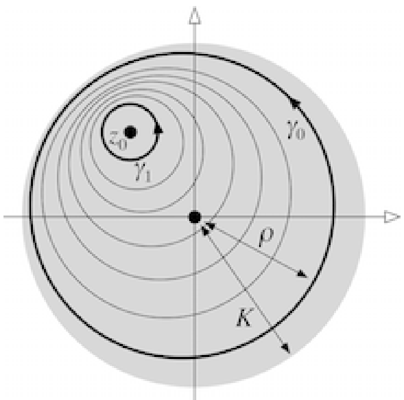

Figure 9.17 Homotopy between the closed paths in Example 9.12.

□

THEOREM 9.13. *If f is differentiable in a domain D and closed paths γ_0, γ_1 in D are homotopic via closed paths in D, then $\int_{\gamma_0} f = \int_{\gamma_1} f$.*

Proof. We have a continuous map $\phi : R \to D$ such that

$$\begin{aligned} \phi(t,0) &= \gamma_0(t) \quad \text{for all } t \in [a,b] \\ \phi(t,1) &= \gamma_1(t) \quad \text{for all } t \in [a,b] \\ \phi(a,s) &= \phi(b,s) \quad \text{for all } s \in [0,1] \end{aligned}$$

Let $\sigma(s) = \phi(a,s)$ ($s \in [0,1]$). Then $\partial\phi = \gamma_0 + \sigma - \gamma_1 - \sigma$ (so σ is a cut from γ_0 to γ_1 in the sense of Section 8.7). By Cauchy's Theorem for a boundary,

$$\int_{\partial\phi} f = \int_{\gamma_0} f + \int_{\sigma} f - \int_{\gamma_1} f - \int_{\sigma} f = 0$$

so

$$\int_{\gamma_0} f = \int_{\gamma_1} f \qquad \square$$

A closed path γ in D is *homotopic to zero* or *null-homotopic* if it is homotopic via closed paths in D to $\beta : [a,b] \to D$ where $\beta(t) = z_0$ for all $t \in [a,b]$, with $z_0 \in D$. We immediately deduce:

COROLLARY 9.14. *Let f be differentiable in a domain D and let γ be a closed path in D that is homotopic to zero. Then $\int_\gamma f = 0$.*

The geometric significance of being homotopic to zero is that γ can be continuously deformed to a single point – more precisely, a path whose image is a single point – as in Figure 9.18.

We can use a homotopy to prove that the interval of a differentiable complex function is very well behaved under reparametrisation, of the path of integration. Here we can work with merely continuous paths and parameter changes. For simplicity we do not change the parametric interval. It is a useful exercise to include that possibility (introduce an affine map $[a,b] \to [c,d]$ and examine how that affects the integral). In the smooth case it is also possible to use the formula for a change of variable in an integral.

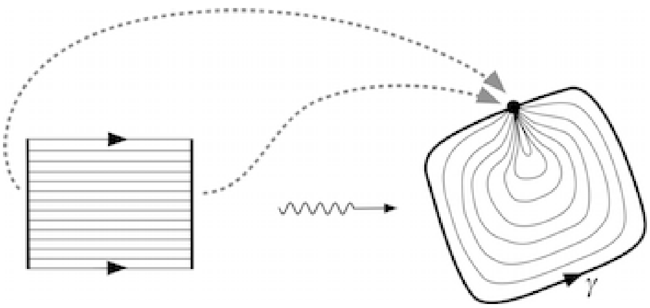

Figure 9.18 A homotopy to zero.

PROPOSITION 9.15. *Let f be differentiable on a domain D, and let $\lambda : [a,b] \to \mathbb{C}$ be a path in D. Let $\rho : [a,b] \to [a,b]$ be continuous, and let $\gamma = \lambda \circ \rho$. Then $\int_\gamma f = \int_\lambda f$.*

Proof. Define a homotopy $\sigma : [a,b] \times [0,1] \to D$ by

$$\sigma(t,s) = \lambda((1-s)t + s\rho(t))$$

Clearly $\sigma_s(t) \in D$ for $t \in [a,b]$. It is easy to check that

$$\sigma(t,0) = \gamma$$
$$\sigma(t,1) = \lambda$$

Therefore σ is a fixed end point homotopy from γ to λ. Now Theorem 9.10 applies. □

9.7 Converse to Cauchy's Theorem

Starting with Cauchy's Theorem for a boundary we have deduced first the homotopy-invariance of the integral for fixed end point or closed path homotopies, and then that the integral is zero for a path homotopic to zero. However, we can also start with Corollary 9.14 and argue the other way if we wish. First we need a simple result about changing the parameter of a path:

PROPOSITION 9.16. *Let $\gamma : [a,b] \to D$ be a path and let $\rho : [a,b] \to [a,b]$ be a continuous map such that $\rho(a) = a, \rho(b) = b$. Then γ is homotopic to $\gamma \circ \rho$.*

Proof. Define $\phi : [a,b] \times [0,1] \to D$ by

$$\phi(t,s) = \gamma((1-s)t + s\rho(t))$$

This is a homotopy, and $\phi(t,0) = \gamma(t), \phi(t,1) = \gamma(\rho(t))$. □

Note that the reparametrisation ρ is not required to be a bijection here. Also: all paths γ_s have the same image. In effect we are performing a homotopy on the parameter.

We can now prove:

PROPOSITION 9.17. *A closed path γ in D is a boundary $\partial\phi$, up to reparametrisation, if and only if γ is homotopic to zero.*

Proof. The need to reparametrise the interval on which γ is defined arises because of the way we have chosen a *specific* parametrisation for a boundary $\partial\phi$. By Proposition 9.16 we can adjust the parameter provided the images of γ and $\partial\phi$ coincide. This reduces the proof to a geometric argument. We give the essence of the proof in a series of pictures, leaving the reader the (routine) task of analytic definitions and verifications required to make it rigorous.

Define a map $H : R \to R$, where R is a rectangle, as in Figure 9.19. The definition proceeds in stages:

(1) Identify opposite vertical edges to get a cylinder.
(2) Squash the top rim to a point to get a cone.

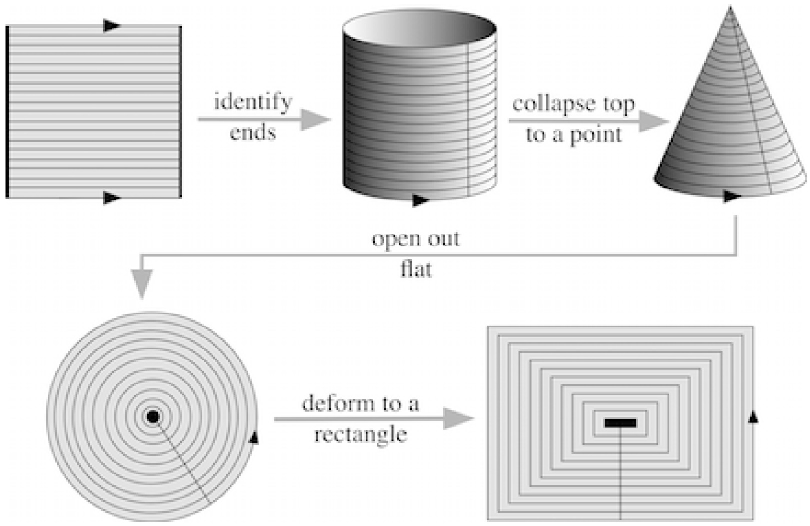

Figure 9.19 Topological proof of Proposition 9.17.

(3) Open the cone out flat to get a disc.
(4) Distort the disc to get a square.

Suppose that γ is a boundary, say $\gamma = \partial\phi$ where $\phi : R \to D$. Then $\phi \circ H : R \to D$ is a homotopy. The lower edge of R, marked by the heavy line in the first diagram, maps to (the image under ϕ of) $\partial\phi$. The top edge maps to a point. So γ is homotopic to zero.

Conversely, suppose that γ is a closed path homotopic to zero. Then we can define a map of the cone into D such that the base circle goes to (the image of) γ. Therefore (reversing the last two steps in the definition of H) we can map R into D so that its perimeter goes to γ. Hence, up to reparametrisation, $\gamma = \partial\phi$. □

9.8 The Cauchy Theorems Compared

A common cause of distress for students of complex analysis is the sudden appearance of a plague of Cauchy Theorems, having several variant hypotheses and a similar variety of conclusions, but all begin derivable from each other. At times like this it may be advisable to seek consolation elsewhere than mathematics: perhaps among the poets. Rudyard Kipling's *In the Neolithic Age* makes the point admirably:

There are nine and sixty way of constructing tribal lays
And – every – single – one – of – them – is – right!

It is much the same with the Cauchy Theorems: all of the different versions are essentially the same.

At the heart of all Cauchy-type theorems is the local existence of an antiderivative, Section 8.5. The theorems themselves supply, as hypotheses, conditions that permit this local result to be *globalised* in some way.

For example, in Chapter 8 the local result led to the central Cauchy Theorem 8.8, that if f is differentiable in D and a closed contour does not wind round any point outside D, then $\int_\gamma f = 0$. This 'non-winding' condition ensures that the local pieces of antiderivative fit together well on the global level. The generalised version, with several contours $\gamma_1, \dots, \gamma_n$, is a simple corollary obtained by making cuts between the contours.

In this chapter we studied what happens when a contour is deformed continuously. Again the local existence of an antiderivative is involved: it lets us define the integral of f along an arbitrary path, and it gives the main result of that chapter, that the integral of f along a boundary is zero. This is also a globalisation: associated with any boundary is a map of an entire rectangle, and the local antiderivatives all fit together properly across the image of this rectangle. Moreover, we saw that a path is a boundary if and only if it is homotopic to zero. Since homotopy leaves the integral round a closed path invariant, this makes the reason why $\int_\gamma f = 0$ transparent.

We therefore have *two* main variants of the Cauchy Theorem: one for paths that do not wind round points outside D, and another for paths that are homotopic to zero. But (more Kipling), like the Colonel's Lady and Judy O'Grady, they are 'sisters under their skins'. Obviously if γ is homotopic to zero and $z_0 \notin D$, then

$$w(\gamma, z_0) = \frac{1}{2\pi \mathrm{i}} \int_\gamma \frac{1}{z - z_0} \mathrm{d}z = 0$$

because $1/(z - z_0)$ is differentiable in D. So the 'non-winding' version of Cauchy's Theorem easily implies the 'homotopy' version.

It is in fact strictly stronger, in the following sense. The path in Figure 9.20 does not wind round any $z_0 \notin D$, but it is manifestly *not* homotopic to zero. (However, this is harder to prove than it might appear.) So the 'non-winding' hypothesis is *weaker*, hence applies to more cases.

Even this surface difference vanishes when we look a little more deeply, however. Every path that does not wind round any $z_0 \notin D$ can be transformed, using a series of cuts whose contributions to the integral cancel, into a closed path (or set of paths) homotopic to zero. For example, Figure 9.20 (left) is so transformed in Figure 9.20 (right). This fact is also non-trivial to prove, but it shows that the *practical* consequences

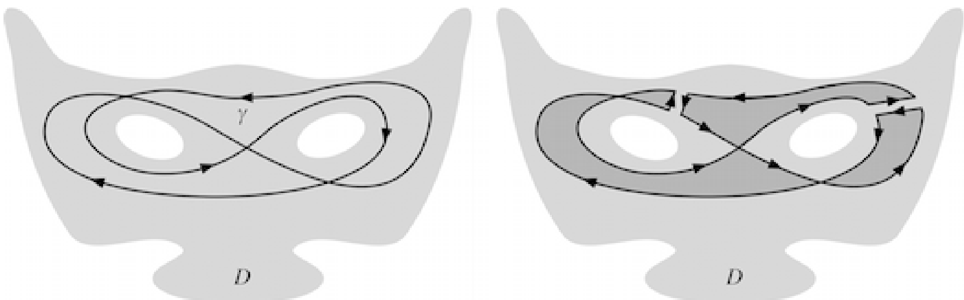

Figure 9.20 *Left*: A path that satisfies the non-winding condition but is not homotopic to zero. *Right*: Creating cuts in the path makes it homotopic to zero.

of the extra generality are largely spurious. (The theoretical consequences are more important: the 'non-winding' condition is part of 'homology theory', and what we have here is a relation between homotopy and homology. We postpone a homology version of Cauchy's Theorem to Chapter 16, so that we can move on to more practical issues in Chapters 10–14.)

9.9 Exercises

1. Let $D = \{z \in \mathbb{C} : z \neq 0\}$ and $S_r(t) = re^{\mathrm{i}t}$ $(t \in [0, 2\pi])$ for $r = 1, 2$. Define a continuous map $\phi : R \to D$, where R is a rectangle, such that
$$\int_{\partial\phi} f = \int_{S_1} f - \int_{S_2} f$$
for any f differentiable in D. Use Theorem 9.6 to deduce that
$$\int_{S_1} f = \int_{S_2} f$$
Describe a homotopy via closed paths in D from S_1 to S_2.

2. Let γ_1, γ_2 be closed paths in a domain D that are homotopic via closed paths in D. By making a suitable cut σ from γ_1 to γ_2, describe a fixed end point homotopy in D from γ_1 to $\sigma + \gamma_2 - \sigma$. Draw a picture to illustrate the continuous deformation.

3. Let the boundary $\partial\phi$ of a continuous map $\phi : R \to D$ be subdivided into two subpaths $\partial\phi = \gamma_1 + \gamma_2$. Describe a fixed end point homotopy in D from γ_1 to $-\gamma_2$. Draw a picture to illustrate the continuous deformation.

4. Draw the semicircle $\gamma(t) = e^{\mathrm{i}t}$ $(t \in [-\pi/2, \pi/2])$ in $D = \mathbb{C} \setminus \{0\}$. Define two explicit polygonal paths λ_1, λ_2 from $-\mathrm{i}$ to i in D such that
$$\int_{\lambda} f = \int_{\gamma} f$$
is true for all f differentiable in D when $\lambda = \lambda_1$, but false when $\lambda = \lambda_2$.

5. Let $\gamma : [0, 1] \to \mathbb{C}$ be given by $\gamma(0) = \gamma(1) = 0$ and
$$\gamma(t) = \begin{cases} t + \mathrm{i}t\sin(\pi/t) & \text{for } t \in [0, \frac{1}{2}] \\ (1-t) + \mathrm{i}(1-t)\sin(\pi/(1-t)) & \text{for } t \in [\frac{1}{2}, 1] \end{cases}$$
Show that γ is a closed path but not a contour, and draw a sketch.

6. Integrate the following functions around the path γ in Exercise 5:
   (i) $\cos^3(z^2)$
   (ii) $\sum_{n=1}^{\infty} z^n/n$
   (iii) $1/(z - \frac{1}{3}\sqrt{2})$

7. Let f be differentiable in D. For a closed path γ in D, beginning and ending at z_0, the *integral value* I_γ is the complex number

$$I_\gamma = \int_\gamma f$$

Show that the set of integral values forms a commutative group $I(f, D)$ under the operation

$$I_\gamma + I_\delta = I_{\gamma+\delta}$$

Determine the group of integral values in the following cases:

(i) $f(z) = 1/z, D = \mathbb{C} \setminus \{0\}$
(ii) $f(z) = \cos z, D = \mathbb{C} \setminus \{0\}$
(iii) $f(z) = 2/(z-1) + 3/(z+1), D = \mathbb{C} \setminus \{\pm 1\}$
(iv) $f(z) = 1/(z-1) + 2/(z+1), D = \mathbb{C} \setminus \{\pm 1\}$

8. *The Fundamental Group*. Let D be a domain and $z_0 \in D$. For closed paths γ, δ in D that begin and end at z_0, define $\gamma \simeq \delta$ to mean that γ, δ are fixed end point homotopic in D. Show that $\simeq$ is an equivalence relation. Let $[\gamma]$ denote the equivalence class containing γ, and let $\pi(D, z_0)$ be the set of equivalence classes.

Define the operation $*$ on $\pi(D, z_0)$ by

$$[\gamma] * [\delta] = [\gamma + \delta]$$

Check that $*$ is well-defined, and show that $\pi(D, z_0)$ is a group under $*$, specifying the identity element and the inverse of $[\gamma]$.

For any other point $z_1 \in D$ and any path σ in D from z_0 to z_1, define $g : \pi(D, z_0) \to \pi(D, z_1)$ by

$$g([\gamma]) = [-\sigma + \gamma + \sigma]$$

Show that g is an isomorphism of groups and deduce that $\pi(D, z_0)$ is independent of the choice of $z_0 \in D$. (This result relies on D being path-connected.) For this reason, the group $\pi(D, z_0)$ is usually denoted by $\pi(D)$ and called the *fundamental group* of D.

9. Describe (without formal proof, since we have not developed suitable techniques) the fundamental groups of the following domains:

(i) $\mathbb{C}$
(ii) $\{z \in \mathbb{C} : |z| < 1\}$
(iii) $\{z \in \mathbb{C} : |1 < z| < 2\}$
(iv) $\mathbb{C} \setminus \{0\}$
(v) $\mathbb{C} \setminus \{\pm 1\}$
(vi) $\mathbb{C} \setminus \mathbb{Z}$

10. Let f be differentiable in the domain D, and let γ, δ be closed contours in D beginning and ending at the same point z_0. Show that if $\gamma \simeq \delta$ in the sense of Exercise 8, then the integral values I_γ and I_δ are equal. Prove that the map $h : \pi(D) \to I(f, D)$ for which $h([\gamma]) = I_\gamma$ is a well-defined group homomorphism. Describe the homomorphism h for each f in Exercise 7.

11. *Integrals Along Arbitrary Paths.* For fixed $z_1, z_2 \in D$, suppose that γ_0 is a fixed path in D, and γ is a variable path in D, both from z_1 to z_2. Show that

$$\int_{\gamma} f = \int_{\gamma_0} f + I_{\sigma}$$

for some $\sigma \in I(f, D)$. Prove that if γ is deformed continuously in a homotopy via closed paths in D, then I_{σ} remains constant.

For $z_1 = -\mathrm{i}, z_2 = \mathrm{i}$, determine all possible values of $\int_{\gamma} f$ for each f, D in Exercise 7.

10 Taylor Series

We now reach a stage in the theory were we can take a great leap forward and show, as promised repeatedly, that any differentiable complex function has a local power series representation. On the slim assumption that the derivative of f exists throughout a domain D, we find that near any point $z_0 \in D$ there is a power series expansion

$$f(z_0 + h) = \sum_{n=0}^{\infty} a_n h^n \quad \text{for } z_0 + h \in N_r(z_0)$$

which is valid – that is, converges – on any disc $N_r(z_0) \subseteq D$, see Figure 10.1.

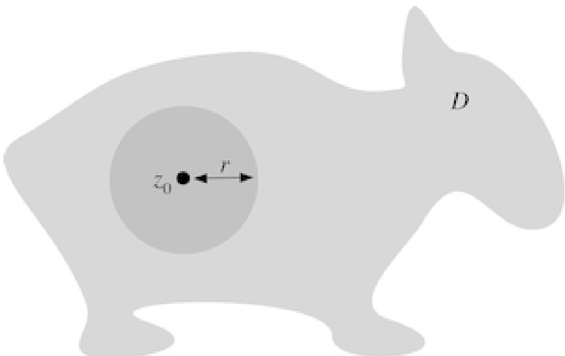

Figure 10.1 The power series round z_0 converges on any disc in D.

This theorem releases a tidal wave of theorems, because we can apply the results of Chapter 4 on power series to any differentiable function f. For instance, Corollary 4.21 states that a power series can be differentiated term by terms as many times as we like. Taylor's Theorem, Corollary 4.22, gives an explicit formula:

$$f(z_0 + h) = \sum_{n=0}^{\infty} \frac{f^{(n)}(z_0)}{n!} h^n \quad \text{for } z_0 + h \in N_r(z_0) \subseteq D$$

Thus if we insist only that the first derivative f' exists in D, it follows that all higher derivatives exist, and the function is equal to its Taylor series in $N_r(z_0)$. We derive subtler consequences in this and subsequent chapters, and Cauchy's Theorem plays a prominent role. It is this sequence of results that gives complex analysis its own special flavour.

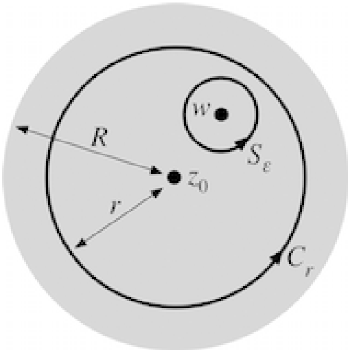

Figure 10.2 Contours in the proof of the Cauchy Integral Formula.

10.1 Cauchy Integral Formula

The proof that any differentiable function can be expressed as a power series depends on a result of Cauchy, itself of intrinsic interest:

LEMMA 10.1 (Cauchy Integral Formula for a Circle). *Let f be differentiable in the disc $N_R(z_0) = \{z \in \mathbb{C} : |z - z_0| < R\}$. For $0 < r < R$, let C_r be the path $C_r(t) = z_0 + re^{it}$ ($t \in [0, 2\pi]$). Then for $|w - z_0| < r$ we have*

$$f(w) = \frac{1}{2\pi i}\int_{C_r} \frac{f(z)}{z - w}dz$$

Proof. Fix w such that $|w - z_0| < r$. The function $F(z) = (f(z) - f(w))/(z - w)$ is differentiable in the domain

$$D = \{z \in \mathbb{C} : |z - z_0| < R, z \neq w\}$$

Let $0 < \varepsilon < r - |w - z_0|$. Then the circle S_ε, centre w, radius ε, is

$$S_\varepsilon(t) = z_0 + \varepsilon e^{it} \quad (t \in [0, 2\pi])$$

and lies in D, as do all points inside C_r and outside S_ε, see Figure 10.2.

By the Generalised Cauchy Theorem, Theorem 8.9,

$$\int_{C_r} F(z)dz = \int_{S_\varepsilon} F(z)dz \tag{10.1}$$

Now $\lim_{z\to w} F(z) = f'(w)$, so for some $\delta > 0, M \geq 0$, we have

$$0 < |z - w| < \delta \text{ implies } |F(z)| \leq M$$

The Estimation Lemma, Lemma 6.41, implies that for $\varepsilon < \delta$,

$$\left|\int_{S_\varepsilon} F(z)dz\right| \leq M \cdot 2\pi\varepsilon$$

From (10.1),

$$\left|\int_{C_r} F(z)\mathrm{d}z\right| \le 2M\pi\varepsilon$$

Since ε is arbitrary, this implies that

$$\int_{C_r} F(z)\mathrm{d}z = 0$$

Therefore

$$\begin{aligned}\int_{C_r} \frac{f(z)}{z-w}\mathrm{d}z &= \int_{C_r} \frac{f(w)}{z-w}\mathrm{d}z \\ &= f(w)\int_{C_r} \frac{1}{z-w}\mathrm{d}z \\ &= f(w)\cdot 2\pi\mathrm{i}\end{aligned}$$

In other words,

$$f(w) = \frac{1}{2\pi\mathrm{i}}\int_{C_r} \frac{f(z)}{z-w}\mathrm{d}z \qquad \square$$

10.2 Taylor Series

Using the Cauchy Integral Formula we can now expand $f(z_0+h)$ as a power series in h, with coefficients expressed as integrals.

LEMMA 10.2. *Let f be differentiable in $N_R(z_0)$. Then there exist $a_n \in \mathbb{C}$ such that*

$$f(z_0+h) = \sum_{n=0}^{\infty} a_n h^n$$

where the series converges absolutely for $|h| < R$. Further, if

$$C_r(t) = z_0 + r\mathrm{e}^{\mathrm{i}t} \ (t \in [0, 2\pi])$$

then

$$a_n = \frac{1}{2\pi\mathrm{i}}\int_{C_r} \frac{f(z)}{(z-z_0)^{n+1}}\mathrm{d}z$$

Proof. Fix h with $0 < |h| < R$ and initially suppose that r satisfies $|h| < r < R$, Figure 10.3.

The Cauchy Integral Formula gives

$$\begin{aligned}f(z_0+h) &= \frac{1}{2\pi\mathrm{i}}\int_{C_r} \frac{f(z)}{z-(z_0+h)}\mathrm{d}z \\ &= \frac{1}{2\pi\mathrm{i}}\int_{C_r} f(z)\left\{\frac{1}{z-z_0} + \frac{h}{(z-z_0)^2} + \cdots + \frac{h^m}{(z-z_0)^{m+1}}\right. \\ &\qquad\qquad \left. + \frac{h^{m+1}}{(z-z_0)^{m+1}(z-z_0-h)}\right\}\mathrm{d}z \\ &= \sum_{n=0}^{m} a_n h^m + A_m\end{aligned}$$

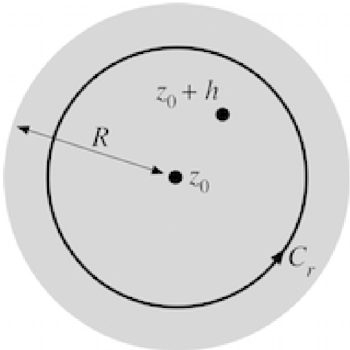

Figure 10.3 Contour for proof of Lemma 10.2.

where

$$a_n = \frac{1}{2\pi \mathrm{i}} \int_{C_r} \frac{f(z)}{(z - z_0)^{n+1}} \mathrm{d}z$$

$$A_m = \frac{1}{2\pi \mathrm{i}} \int_{C_r} \frac{f(z)h^{m+1}}{(z - z_0)^{m+1}(z - z_0 - h)} \mathrm{d}z$$

We demonstrate that $\lim_{m\to\infty} A_m = 0$.

First, since f is differentiable it is continuous, so $\phi(t) = |f(C_r(t))|$ is a continuous real function on $[0, 2\pi]$. From real analysis, ϕ is bounded, so

$$|\phi(t)| \leq M \quad \text{for } z \text{ on } C_r$$

Now $|h| < r, |z - z_0| = r$, and

$$|z - z_0 - h| \geq ||z - z_0| - |h|| = r - |h|$$

Therefore the Estimation Lemma gives

$$|A_m| \leq \frac{1}{2\pi} \frac{M|h|^{m+1}}{r^{m+1}(r - |h|)} 2\pi r = \frac{M|h|}{r - |h|} \left(\frac{|h|}{r}\right)^m$$

Since we chose $|h| < r$, this tends to zero as m tends to infinity. Thus

$$\lim_{m\to\infty} \left(f(z_0 + h) - \sum_{n=0}^{m} a_n h^n \right) = 0$$

which means that

$$f(z_0 + h) = \sum_{n=0}^{\infty} a_n h^n$$

This expansion is valid for $|h| < R$, and

$$a_n = \frac{1}{2\pi \mathrm{i}} \int_{C_r} \frac{f(z)}{(z - z_0)^{n+1}} \mathrm{d}z$$

for $|h| < r$. The latter restriction can now be seen to be unnecessary, for the integral is differentiable for $0 < |z - z_0| < R$, so the integral is unchanged if r is varied in the range $0 < r < R$. □

Once we know that a power series expansion exists, we can use our knowledge of power series to deduce:

THEOREM 10.3 (Taylor Series). *If f is differentiable in a domain D, then all higher derivatives of f exist throughout D. In any disc $N_R(z_0) \subseteq D$ the Taylor series expansion*

$$f(z_0 + h) = \sum_{n=0}^{\infty} \frac{f^{(n)}(z_0)}{n!} h^n \tag{10.2}$$

is valid. Further, if $0 < r < R$ and $C_r(t) = z_0 + re^{it}$ $(t \in [0, 2\pi])$, then

$$f^{(n)}(z_0) = \frac{n!}{2\pi i} \int_{C_r} \frac{f(z)}{(z - z_0)^{n+1}} dz$$

Proof. From Lemma 10.2,

$$f(z_0 + h) = \sum_{n=0}^{\infty} a_n h^n \quad \text{for } |h| < R$$

In other words, putting $z = z_0 + h$,

$$f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n \quad \text{for } |z - z_0| < R$$

Now, a power series may be differentiated term by term as often as we please, and by Corollary 4.22

$$f^{(n)}(z_0) = n! a_n = \frac{n!}{2\pi i} \int_{C_r} \frac{f(z)}{(z - z_0)^{n+1}} dz$$

This gives the desired integral expression for $f^{(n)}(z_0)$. Substituting $a_n = (f^{(n)}(z_0))/n!$ in the power series gives the Taylor expansion. □

Theorem 10.3 was first proved by Cauchy in 1831 by the method above. The series is named after Brooke Taylor, who in 1715 was the first to publish the idea that a function can be expanded as a power series of the form

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

Taylor's theory was restricted to real functions, and it will come as no surprise that the idea was previously known to others; specifically to James Gregory, who was aware of it some 45 years earlier, and to Newton in 1691. During the eighteenth century there were various attempts to base the theory of real analysis on power series, the most famous being that of Joseph-Louis Lagrange in 1797. Cauchy used power series extensively in complex analysis. It is a curious quirk of fate that in 1829 he quoted the counterexample $f(x) = e^{-1/x^2}$ to show that not every infinitely differentiable real function is equal to its Taylor series. Just two years later he went on to show that *all* differentiable complex functions have valid power series expansions.

We now introduce some standard terminology, which until now we have avoided.

DEFINITION 10.4. A real function $f : D \to \mathbb{R}$, where $D \subseteq \mathbb{R}$, or a complex function $f : D \to \mathbb{C}$, where $D \subseteq \mathbb{C}$, is *analytic* if for each $\alpha \in D$ it has a power series expansion

$$f(\alpha + h) = \sum_{N=0}^{\infty} a_n h^n$$

valid in some neighbourhood of α. □

Cauchy demonstrated that in the real case there are functions that are infinitely differentiable but not analytic. But in the complex case he proved that any function that is differentiable *once* in a domain must be analytic. At a stroke he showed that in this respect complex analysis is simpler than real analysis, and reduced the general study of differentiable complex functions to computations with power series, giving the sequence of theorems that unfolds in the next few sections.

REMARK 10.5. A complex function on a domain is differentiable *if and only if* it is analytic. The two words just emphasise different points of view – existence of a derivative, existence of a power series expansion – and may be used interchangeably. □

10.3 Morera's Theorem

First we have a partial converse to Cauchy's Theorem, due to Giacinto Morera in 1889:

THEOREM 10.6 (Morera's Theorem). *If $f : D \to \mathbb{C}$ is continuous in a domain D and $\int_\gamma f = 0$ for all closed contours γ in D, then f is differentiable in D.*

Proof. By Theorem 6.44, if $\int_\gamma f = 0$ for all closed contours γ in D, then there exists a differentiable function F in D whose derivative is f. But now we know that F is twice (indeed infinitely) differentiable, so $f = F'$ is differentiable. □

This theorem explains our warning in Chapter 6 that it is futile to try to find an antiderivative F for a non-differentiable complex function f. It cannot have one. So there is a class of functions, including $f(z) = |z|$, that are continuous but not differentiable. Such functions are integrable, via the formula

$$\int_\gamma f = \int_a^b f(\gamma(t))\gamma'(t)\mathrm{d}t$$

but the Fundamental Theorem of Contour Integration is no use whatsoever in this context, because f has no antiderivative – in stark contrast to the real case, where every continuous function has an antiderivative. (It can be argued that here real analysis is simpler than complex.)

To summarise what we know about the existence of derivatives and antiderivatives for complex functions:

If f is differentiable in a domain D, then all higher derivatives of f exist. The function f has an antiderivative only when f is differentiable, and even then, only *local* antiderivatives are guaranteed. Specifically, if $D_1 \subseteq D$ is simply connected, then f has an antiderivative in D_1, Theorem 8.12. In particular, we can guarantee the existence of an antiderivative of a differentiable function in any disc contained in its domain.

10.4 Cauchy's Estimate

Theorem 10.3 includes a generalisation of Cauchy's Integral Formula to the higher derivatives of a differentiable function:

$$f^{(n)}(z_0) = \frac{n!}{2\pi \mathrm{i}} \int_{C_r} \frac{f(z)}{(z-z_0)^{n+1}} \mathrm{d}z$$

where $N_R(z_0) \subseteq D$ and $0 < r < R$. With the standard conventions that $f^{(0)}(z) = f(z)$ and $0! = 1$, this formula is true for all integers $n \geq 0$. Using it, we can give an upper bound for $|f^{(n)}(z_0)|$:

LEMMA 10.7 (Cauchy's Estimate). *If $f : D \to \mathbb{C}$ is differentiable for $z - z_0 < R$, $0 < r < R$, and $|f(z)| \leq M$ for $|z - z_0| < r$, then*

$$|f^{(n)}(z_0)| \leq \frac{Mn!}{r^n}$$

Proof.

$$\begin{aligned} |f^{(n)}(z_0)| &= \left| \frac{n!}{2\pi \mathrm{i}} \int_{C_r} \frac{f(z)}{(z-z_0)^{n+1}} \mathrm{d}z \right| \\ &\leq \frac{n!}{2\pi} \frac{M}{r^{n+1}} \cdot 2\pi r \\ &= \frac{Mn!}{r^n} \end{aligned}$$

□

Cauchy's Estimate yields an important theorem of Joseph Liouville, which has an unexpected application to a purely algebraic problem:

THEOREM 10.8 (Liouville's Theorem). *If f is differentiable and bounded in the whole complex plane, then f is constant.*

Proof. Suppose that $|f(z)| \leq M$ for all $z \in \mathbb{C}$. Cauchy's Estimate applied to the derivative gives

$$|f'(z)| \leq \frac{M}{r}$$

Since f is differentiable on $\mathbb{C}$ we may let $r \to \infty$, making M/r as small as we please. Since $|f'(z)|$ is independent of r, we have

$$f'(z) = 0$$

Thus $f'(z) = 0$ throughout $\mathbb{C}$, and (integrating) f is constant. □

The application is:

THEOREM 10.9 (Fundamental Theorem of Algebra). *Let $P(z) = z^n + a_1 z^{n-1} + \cdots + a_n$ be a polynomial, where $n \geq 1$ and $a_1, \ldots, a_n \in \mathbb{C}$. Then there exists $w \in \mathbb{C}$ with $P(w) = 0$.*

Proof. For a contradiction, suppose that $P(z) \neq 0$ for all $z \in \mathbb{C}$. Then $1/P(z)$ is differentiable throughout $\mathbb{C}$.

Suppose $z \neq 0$. Then

$$\frac{P(z)}{z^n} = 1 + \frac{a_1}{z} + \cdots + \frac{a_n}{z^n} \to 1 \text{ as } |z| \to \infty$$

So there exists $k > 0$ such that

$$\left|\frac{P(z)}{z^n}\right| \geq \tfrac{1}{2} \quad \text{for } |z| > k$$

Thus

$$\left|\frac{1}{P(z)}\right| \leq \frac{2}{|z^n|} \leq \frac{2}{k^n} \quad \text{for } |z| > k$$

The same bound works for $1/P(z)$ throughout $\mathbb{C}$. For if $z \leq k$ we take a circle C_R centre z_0, radius R, which is so large that $|z| > k$ for all z on C_R. Then

$$\left|\frac{1}{P(z)}\right| \leq \frac{2}{k^n} \quad \text{for all } z \text{ on } C_R$$

and Cauchy's Estimate gives

$$\left|\frac{1}{P(z)}\right| \leq \frac{2}{k^n}$$

By Liouville's Theorem, $1/P(z)$ is constant, so $P(z)$ is constant. But this contradicts $n \geq 1$. Therefore $P(w) = 0$ for some $w \in \mathbb{C}$. □

It follows in the usual way that any polynomial $P(z)$ of degree n, with complex coefficients, can be expressed as a produce of terms of degree 1:

$$P(z) = (z - \alpha_1)(z - \alpha_2) \cdots (z - \alpha_n)$$

where the $\alpha_j \in \mathbb{C}$.

10.5 Zeros

We now broaden our perspective from polynomials, and look at the zeros of arbitrary differentiable functions.

DEFINITION 10.10. A *zero* of a differentiable function $f : D \to \mathbb{C}$ is a point $z_0 \in D$ for which $f(z_0) = 0$. □

Expanding f in a Taylor series about the zero z_0, we have

$$f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n \quad \text{for } |z - z_0| < R$$

where $N_R(z_0) \subseteq D$. Then $a_0 = f(z_0) = 0$, and two distinctly different things can occur: either all of the other a_m are zero, in which case $f(z) = 0$ for all $z \in N_R(z_0)$, or there exists $m \geq 1$ such that

$$a_0 = a_1 = \cdots = a_{m-1} = 0, \text{ but } a_m \neq 0 \tag{10.3}$$

DEFINITION 10.11. If (10.3) holds, then z_0 is a *zero of order m*, or a *zero of finite order* if m is unspecified. □

The notion of a zero of order m relates directly to that of an infinitesimal of order m in Chapter 15.

Formula (10.2) for the Taylor coefficients implies that a zero of order m can be characterised by the condition

$$f(z_0) = f'(z_0) = \cdots = f^{(m-1)}(z_0) = 0, \text{ but } f^{(m)}(z_0) \neq 0$$

Another useful expression for such a zero is to write

$$f(z_0) = (z - z_0)^m g(z) \quad (|z - z_0| < R)$$

where

$$g(z) = \sum_{n=0}^{\infty} a_{m+n}(z - z_0)^n$$

is differentiable for $|z - z_0| < R$ and $g(z_0) = a_m \neq 0$.

This leads to a fundamentally important idea:

DEFINITION 10.12. A zero z_0 of a differentiable function $f : D \to \mathbb{C}$ is *isolated* if some disc centred upon it contains no other zeros of f; that is, there exists $\delta > 0$ such that

$$0 < |z - z_0| < \delta \text{ implies } f(z) \neq 0$$ □

LEMMA 10.13. *A zero of finite order is isolated.*

Proof. Write $f(z_0) = (z - z_0)^m g(z)$ for $|z - z_0| < R$, where g is differentiable and $g(z_0) \neq 0$. Then g is continuous at z_0, so, taking $\varepsilon = \frac{1}{2}|g(z_0)|$ there exists $\delta > 0$ such that

$$|z - z_0| < \delta \text{ implies } |g(z) - g(z_0)| < \varepsilon$$

Therefore, when $|z - z_0| < \delta$ we have

$$|g(z)| \geq ||g(z_0)| - |g(z_0) - g(z)|| > 2\varepsilon - \varepsilon = \varepsilon$$

In particular, $g(z) \neq 0$. But if $0 < |z - z_0| < \delta$ then $|z - z_0|^m \neq 0$, so $f(z) = (z - z_0)^m g(z) \neq 0$. □

COROLLARY 10.14. *Let S be a set of zeros of a differentiable function f in D, having a limit point $z_0 \in D$. Then f is identically zero in any disc $N_R(z_0) \subseteq D$.*

Proof. Because z_0 is a limit point of S, there is a sequence $\{z_n\}_{n\geq 1}$ in S that tends to z_0. Then $f(z_0) = \lim_{n\to\infty} f(z_n) = 0$. So z_0 is a zero of f that is not isolated. Hence it does not have finite order, so

$$f(z_0 + h) = \sum_{n=0}^{\infty} a_n h_n$$

in any disc $N_R(z_0) \subseteq D$, where all a_n are zero. □

From this we deduce:

PROPOSITION 10.15. *If f is differentiable in a domain D and S is a set of zeros of f with a limit point $z_0 \in D$, then f is identically zero on D.*

Proof. Corollary 10.14 gives $f(z) = 0$ for z in any disc $N_R(z_0) \subseteq D$.

For any other $z \in D$, choose a path $\gamma : [a, b] \to D$ from z_0 to z, Figure 10.4. We show that $f(\gamma(t)) = 0$ for all $t \in [a, b]$. By continuity, we can find $\delta > 0$ such that

$$a \leq t < a + \delta \text{ implies } \gamma(t) \in N_R(z_0)$$

so $f(\gamma(t)) = 0$ for $a \leq t < a + \delta$. Let s be the least upper bound of those $x \in [a, b]$ such that $f(\gamma(x)) = 0$ for $a \leq t < x$. Then $a + \delta \leq s \leq b$.

By continuity, $f(\gamma(s)) = 0$. If $s < b$ then $\gamma(s)$ is a non-isolated zero, so f is identically zero in a neighbourhood of $\gamma(s)$. Then there exists an interval $[s, s + \kappa]$ with $\kappa > s$, on which $f(\gamma(t)) = 0$, contradicting the definition of s. Hence $s = b$, so $f(z) = f(\gamma(b)) = 0$. □

An important consequence is:

THEOREM 10.16 (Identity Theorem). *If f and g are differentiable in a domain D and $f(z) = g(z)$ for all $z \in S \subseteq D$ where S has a limit point in D, then $f = g$ throughout D.*

Proof. Apply Proposition 10.15 to $f - g$. □

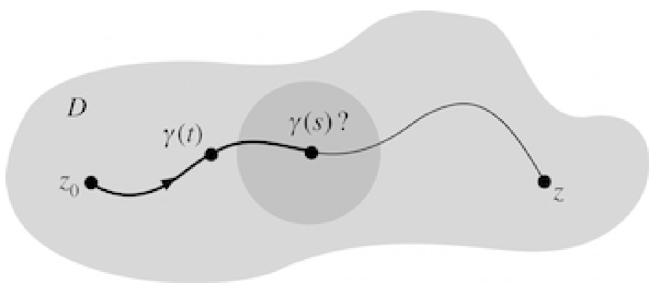

Figure 10.4 Proof of Proposition 10.15.

Example 10.17. It is essential for the relevant limit point of S to be in D: if not, the theorem can be false. Let

$$f(z) = \sin(1/z)$$
$$g(z) = 0$$

in $D = \mathbb{C} \setminus \{0\}$. Then $f(z) = g(z)$ for $z = \pm 1/(n\pi)$ and $z_0 = 0$ is a limit point of $S = \{\pm 1/(n\pi)\}$, but $f \neq g$ in D. □

10.6 Extension Functions

Combining the Identity Theorem with power series leads to a method that can often extend the domain of definition of a complex differentiable function.

DEFINITION 10.18. A function $f : D \to \mathbb{C}$ is an *extension function* of $h : S \to \mathbb{C}$ if $S \subseteq D$ and $f(z) = h(z)$ for all $z \in S$. (We also say that f *extends* h.) □

Example 10.19. $f(z) = 1/(1 - z)$ on $D = \mathbb{C} \setminus \{1\}$ is an extension function of $h(z) = \sum_{n=0}^{\infty} z^n$ on $S = \{z \in \mathbb{C} : |z| < 1\}$. □

Suppose that D is a domain and $S \subseteq D$ has a limit point in D. The Identity Theorem shows that if a function $f : S \to \mathbb{C}$ has a differentiable extension function $g : D \subseteq \mathbb{C}$, then the extension is unique.

As an application, suppose that $f : D \to \mathbb{C}$ is differentiable and consider the Taylor expansion $f(z) = \sum_{n=0}^{\infty} a_n(z - z_0)^n$ of f in a disc $N_r(z_0) \subseteq D$. Then f is the only possible extension of the Taylor expansion to the whole of D. This means that although the Taylor expansion may not converge on the whole of D, it contains all the information needed to determine f uniquely throughout D. We use this fact to great advantage in Chapter 14.

Example 10.20. Consider the power series

$$z - \frac{z^2}{2} + \cdots + (-1)^n \frac{z^n}{n} + \cdots \tag{10.4}$$

on the small disc $S = \{z \in \mathbb{C} : |z| < 1/1\,000\,000\}$. We can certainly extend this function to a larger disc, because the radius of convergence of the power series is 1. However, we can extend beyond this. For instance, if $K = \{t \in \mathbb{R} : t \leq -1\}$ and $D = \mathbb{C} \setminus K$, then $f(z) = \log(1 + z)$ is the (unique) differentiable extension of the power series to D, Figure 10.5. □

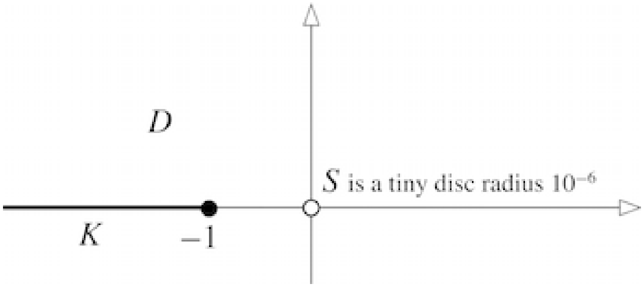

Figure 10.5 Extending the power series (10.4) to $\mathbb{C} \setminus K$.

The set S need not be a domain. In particular, it may be a subset of $\mathbb{R}$:

Example 10.21. $f(z) = \sin z$ for $z \in \mathbb{C}$ is the unique differentiable extension function of $f(x) = \sin x$ for $x \in \mathbb{R}$. This is a consequence of the usual convergent Taylor expansion in the real case. □

Of course, if D is a domain then it is open, so $D \cap \mathbb{R}$ is open in $\mathbb{R}$. A differentiable function $f : D \to \mathbb{C}$ has a power series expansion in a neighbourhood of any $x_0 \in D \cap \mathbb{R}$. So if a real function $h : S \to \mathbb{R}$ extends to a differentiable complex function on a domain, it must already have a power series expansion about any point in $S \subseteq D \cap \mathbb{R}$. Thus the only real functions that extend to differentiable complex functions are real *analytic* functions.

We may take S to be even more restricted, provided that it has at least one limit point in D:

Example 10.22. If $f(1/n) = 1/n^2$ for all positive integers n, then $f(z) = z^2$ is the unique analytic extension of f to the whole complex plane, because

$$S = \{1/n : n \text{ is a positive integer}\}$$

has the limit point $0 \in \mathbb{C}$. □

The Taylor expansion is valid on discs (Theorem 10.3), and may therefore be used to define extension functions. It is therefore easy to see that the radius of convergence of the Taylor expansion of a function f about a point z_0 is equal to the distance from z_0 to the *nearest* point z_1 at which no differentiable extension function of f may be defined. Such a point is called a *singularity* of f, see Chapter 11. Singularities are obstacles that determine the radius of convergence of a Taylor series.

(There is no relation between this use of ‘singularity’ and the term ‘singular’ in Definition 6.20. ‘Singular’ is just one of those words that get overused in mathematics.)

10.7 Local Maxima and Minima

The complex numbers are not ordered, so we cannot speak of maxima and minima of a complex function f. We can, however, consider maximum and minimum values of the modulus $|f|$, since this is real.

DEFINITION 10.23. If $f : D \to \mathbb{C}$, then

(i) The modulus $|f|$ has a *local maximum* at $z_0 \in D$ if there exists $\varepsilon > 0$ such that $N_\varepsilon(z_0) \subseteq D$ and $|f(z)| \le |f(z_0)|$ for all $z \in N_\varepsilon(z_0)$. The local maximum is *strict* if $|f(z)| < |f(z_0)|$ for all $z \in N_\varepsilon(z_0) \setminus \{z_0\}$.
(ii) The modulus $|f|$ has a *local minimum* at $z_0 \in D$ if there exists $\varepsilon > 0$ such that $N_\varepsilon(z_0) \subseteq D$ and $|f(z)| \ge |f(z_0)|$ for all $z \in N_\varepsilon(z_0)$. The local minimum is *strict* if $|f(z)| > |f(z_0)|$ for all $z \in N_\varepsilon(z_0) \setminus \{z_0\}$. □

The problem of finding local maxima of $|f|$ in a domain turns out to be easy (or impossible, depending on how you look at it, since there are none.)

PROPOSITION 10.24. *A differentiable function has no strict local maximum of its modulus in its domain. If it has a local maximum in its domain, it is constant.*

Proof. Suppose that f is differentiable in D and let z_0 be a local maximum in D. Then $|f(z)| \le |f(z_0)|$ for all $z \in N_\varepsilon(z_0)$. For $0 < r < \varepsilon$, the circle $C_r(t) = z_0 + re^{it}$ $(t \in [0, 2\pi])$ lies inside $N_\varepsilon(z_0)$, so $|f(z_0 + re^{it})| \le |f(z_0)|$ for all $t \in [0, 2\pi]$. The Cauchy Integral formula gives

$$\begin{aligned} f(z_0) &= \frac{1}{2\pi i}\int_{C_r} \frac{f(z)}{z - z_0} dz \\ &= \frac{1}{2\pi i}\int_0^{2\pi} \frac{f(z_0 + re^{it})}{re^{it}} ire^{it} dt \\ &= \frac{1}{2\pi i}\int_0^{2\pi} f(z_0 + re^{it}) dt \end{aligned}$$

so that

$$|f(z_0)| \le \frac{1}{2\pi}\int_0^{2\pi} |f(z_0 + re^{it})| dt \le \frac{1}{2\pi}\int_0^{2\pi} |f(z_0)| dt \le |f(z_0)|$$

Hence

$$|f(z_0)| = \frac{1}{2\pi}\int_0^{2\pi} |f(z_0 + re^{it})| dt \tag{10.5}$$

If the *strict* inequality $|f(z_0 + re^{it})| < |f(z_0)|$ were to hold for any $t \in [0, 2\pi]$, then by continuity it would also hold in a small interval, giving the strict inequality

$$\frac{1}{2\pi}\int_0^{2\pi} |f(z_0 + re^{it})| dt < |f(z_0)|$$

contradicting (10.5).

Hence $|f(z_0 + re^{it})| = |f(z_0)|$ for all $t \in [0, 2\pi]$. This holds for any $r < \varepsilon$, so $|f|$ is constant in $N_\varepsilon(z_0)$. By Proposition 4.15, f is constant in $N_\varepsilon(z_0)$. The Identity Theorem now shows that f is constant throughout D. □

For the complementary notion of a local minimum of $|f|$, it is clear that if f is non-constant and has a zero at z_0 then z_0 is isolated and $|f|$ has a strict local minimum there. If f is non-zero on D, observe that $|f|$ has a strict local minimum if and only if $|1/f| = 1/|f|$ has a strict local maximum. We can apply Proposition 10.24 to $1/f$ to get:

PROPOSITION 10.25. *A differentiable function that does not equal zero anywhere in its domain has no strict local minimum of its modulus in its domain. If it has a local minimum in its domain, it is constant.* □

10.8 The Maximum Modulus Theorem

The question of maxima or minima of $|f|$ on an arbitrary subset of the domain of f is somewhat different.

DEFINITION 10.26. Let $f : D \to \mathbb{C}$ be differentiable in a domain D, and let $S \subseteq D$. Then $|f|$ has a *local maximum on S* at $z_0 \in S$ if

(i) z_0 is a limit point of S.
(ii) For some $\varepsilon > 0$, $|f(z)| < |f(z_0)|$ whenever $z \in N_\varepsilon(z_0) \cap S$. □

Condition (i) is essential, for otherwise some neighbourhood $N_\varepsilon(z_0)$ contains no points of S other than z_0, and then condition (ii) is vacuously true.

Example 10.27. If $f(z) = e^z$ and $S = \{z \in \mathbb{C} : |z| \leq 1\}$, then $|f(x + iy)| = e^x$ and $|f|$ has a local maximum on S at the point $z_0 = 1$. □

Using Proposition 10.24, we see that if $N_\varepsilon(z_0) \subseteq S$, then $|f|$ cannot have a strict local maximum at z_0. To explore this further, we need:

DEFINITION 10.28. A point z_0 is a *boundary point* of S if every neighbourhood of z_0 contains a point in S and a point not in S, other than z_0 itself. In other words, a boundary point is a limit point both of S and its complement $\mathbb{C} \setminus S$.

The *boundary* ∂S of S is its set of boundary points. □

Example 10.29. The boundary of $S = \{z \in \mathbb{C} : |z| \leq 1\}$ and of of its complement $\{z \in \mathbb{C} : |z| > 1\}$ is the circle $\{z \in \mathbb{C} : |z| = 1\}$. □

We can now rephrase Proposition 10.24 to give:

THEOREM 10.30 (Maximum Modulus Theorem). *If a differentiable function is not constant, then any local maximum value of its modulus on an arbitrary subset of its domain occurs on the boundary of that set.* □

We also have from Proposition 10.25:

THEOREM 10.31 (Minimum Modulus Theorem). *If a differentiable function is not constant, then any local minimum value of its modulus on an arbitrary subset of its domain occurs either at a zero of the function, or on the boundary of that set.* □

Example 10.32. If $f(z) = z^2$ on the set $S = \{z \in \mathbb{C} : |z| \leq 1\}$, then the maximum values of $|f(x + iy)| = x^2 + y^2$ occur all round the boundary of S, while the minimum value occurs at the origin. □

10.9 Exercises

1. Find the Taylor series at 0 of $f(z) = \mathrm{Log}(1 + z)$, where Log is the principal value. What is the disc of convergence? Answer the same questions for $g(z) = \exp(\alpha \,\mathrm{Log}(1 + z))$, where $\alpha \in \mathbb{C}$.

2. Find the first three terms and radius of convergence for the Taylor series at 0 of $F(z) = [1 + \mathrm{Log}\,(1 - z)]^{-1}$.

3. Taylor expand the following functions around 0, and find the radius of convergence.

(i) $\sin^2 z$

(ii) $z^2(z + 2)^{-2}$

(iii) $(az + b)^{-1}$ $(a, b \in \mathbb{C}, b \neq 0)$

(iv) $\int_0^z \exp(w^2)\mathrm{d}w$

(v) $\begin{cases} (\sin z)/z & (z \neq 0) \\ 1 & (z = 0) \end{cases}$

(vi) $\int_0^z (\sin w)/w \,\mathrm{d}w$

4. Define the numbers c_n by the Taylor series

$$\sec z = \sum_{n=0}^{\infty} (-1)^n \frac{c_{2n}}{(2n)!} z^{2n}$$

Prove that

$$c_0 = 1$$
$$0 = c_0 + c_2 \binom{2n}{2} + c_4 \binom{2n}{4} + \cdots + c_{2n} \binom{2n}{2n}$$

Show that c_{2n} is always an integer and calculate it for $n \leq 5$.

5. Let

$$(1 - z - z^2)^{-1} = \sum F_n z^n$$

Prove that

$$F_0 = F_1 = 1 \qquad F_n = F_{n-1} + F_{n-2} \quad (n \geq 2)$$

This is the recursive definition of the *Fibonacci numbers* $1, 1, 2, 3, 5, 8, 13, \ldots$.

By expanding $(1 - z - z^2)^{-1}$ in partial fractions, prove that

$$F_n = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^{n+1} - \left(\frac{1-\sqrt{5}}{2}\right)^{n+1}\right]$$

6. Investigate analogous results to Exercise 5 on the expansion of $(1 - az - bz^2)^{-1}$, for $a, b \in \mathbb{C}$.

7. If

$$\exp\left(\frac{z}{1-z}\right) = \sum a_n z^n$$

prove that

$$a_0 = 1 \qquad a_n = \sum_{s=1}^{n} \frac{1}{s!}\binom{n-1}{s-1}$$

8. Let $f(z)$ have Taylor series $\sum a_n z^n$ for $|z| < R$. Let $\omega = e^{2\pi i/3}$ and define

$$g(z) = \tfrac{1}{3}(f(z) + f(\omega z) + f(\omega^2 z))$$

Show that

$$g(z) = \sum a_{3n} z^{3n}$$

for $|z| < R$. Find similar expressions for $\sum a_{3n+1} z^{3n+1}$ and $\sum a_{3n+2} z^{3n+2}$. (Hint: $1 + \omega + \omega^2 = 0$.)

9. Define three functions by

$$\alpha(z) = \sum \frac{z^{3n}}{(3n)!} \qquad \beta(z) = \sum \frac{z^{3n+2}}{(3n+2)!} \qquad \gamma(z) = \sum \frac{z^{3n+1}}{(3n+1)!}$$

Prove the series converge for all z. Using Exercise 8, prove the following:

(i) $\alpha'(z) = \beta(z) \quad \beta'(z) = \gamma(z) \quad \gamma'(x) = \alpha(z)$

(ii) $\alpha(z) = \frac{1}{3}[e^z + 2e^{-z/2}\cos(z\sqrt{3}/2)]$; find similar expressions for $\beta(z)$, $\gamma(z)$.

(iii) $\alpha(z + w) = \alpha(z)\alpha(w) + \beta(z)\gamma(w) + \gamma(z)\beta(w)$

(iv) $\alpha^3(z) + \beta^3(z) + \gamma^3(z) = 1$

(v) $\alpha(z) = \beta(z) \iff z = (3n - 1) \cdot \frac{2\pi}{3\sqrt{3}} \ (n \in \mathbb{Z})$

10. Generalise Exercise 8 to find an expression for $\sum a_{pn} z^{pn}$ $(p = 2, 3, \ldots)$, and (harder) for $\sum a_{pn+k} z^{pn+k}$ $(p = 2, 3, \ldots; k = 0, 1, \ldots, p - 1)$.

11. Show that the Cauchy estimate is an equality if and only if $f(z) = Kz^n$ for some $K \in \mathbb{C}$, $n = 1, 2, 3, \ldots$.

12. Let D be a disc centre z_0, and let f be differentiable in a domain containing D. Prove that the 'mean value' of $f(z)$ as z runs over ∂D (defined by a suitable integral) is equal to $f(z_0)$.

13. Let f be differentiable throughout $\mathbb{C}$, and suppose that $|f(z)| \leq K|z|^c$ for a real constant K and positive integer c. Prove that f is a polynomial function of degree $\leq c$.

14. Let f and g be differentiable on the strip $D = \{z \in \mathbb{C} : -2 < \operatorname{im} z < 2\}$. Suppose that $f(z) = g(z)$ for all z such that $|z| < 0.01$. By considering Taylor expansions first about 0, then 1, and so on by induction, prove that $f(z) = g(z)$ on D.

15. Suppose that $f(z) = \sum a_n(z - z_0)^n$ in a disc D centre z_0, radius R. If $0 \leq r < R$, show that

$$\frac{1}{2\pi}\int_0^2 |f(z_0 + re^{i\theta})|^2 d\theta = \sum |a_n|^2 r^{2n}$$

(*Parseval's Inequality*). Hence show that

$$\sum |a_n|^2 r^{2n} \leq M(r) = \sup_\theta |f(z_0 + re^{i\theta})|$$

Use this to give an alternative proof of the Maximum Modulus Theorem.

16. If $p(z)$ is a polynomial of degree n, show that for each $R > 0$ the 'level curve' of $|p(z)|$, defined to be $\{z \in \mathbb{C} : |p(z)| = R\}$, has at most n connected components.

17. Suppose that $x^2 + y^2 \leq a^2$. Prove that $(1 + x)^2 + y^2$ attains its maximum value when $x = a, y = 0$. (Hint: apply the Maximum Modulus Theorem to $1 + z$ on the disc $|z| = a$.)

18. Suppose that $x^2 + y^2 \leq 1$. Prove that $(x^2 - y^2 - 1)^2 + 4x^2y^2$ attains its maximum value when $x = 0, y = \pm 1$.

19. If f is differentiable in a domain D, prove that the zeros of f are either all of finite order and isolated, or f is identically zero on D.

20. Let

$$f(x) = \int_x^\infty t^{-1} e^{x-t} dt$$

where x is real and positive. By repeated integration by parts, show that if

$$h_n(x) = (-1)^n n! x^{-(n+1)}$$

then

$$f(x) = h_0(x) + h_1(x) + \cdots + h_{n-1}(x) + (-1)^n \int_x^\infty e^x t^{-(n+1)} dt$$

Show that the series

$$\sum_{n=0}^\infty h_n(x) \tag{10.6}$$

diverges for all x. Show also that

$$\left| f(x) - \sum_{n=o}^N h_n(x) \right| < N! x^{-(N+1)}$$

so that *for large enough x* the series (10.6) provides an arbitrarily good approximation to $f(x)$, even though it diverges.

(The catch is that, the better the approximation required, the larger x has to be. No *particular* choice of x gives an arbitrarily good approximation if we take N large enough. A series with this property is called an *asymptotic* series.)

11 Laurent Series

The Taylor series expansion is too limited for many applications. A useful generalisation was given by Laurent in 1843. Weierstrass may have discovered it two years earlier, but his paper was not published until after his death. Laurent considered 'power series' involving negative powers as well as positive. The benefits that accrue are hinted at by the following example. The function $f(z) = e^{-1/z^2}$ is very badly behaved as regards Taylor series expansion. We have seen that, restricted to the real line, its Taylor series about the origin is $0 + 0x + 0x^2 + \cdots$, which does not converge to $f(x)$. On the complex plane it is, if such a statement makes sense, even less capable of being represented by a Taylor series. The natural series representation is obtained by starting with the Taylor series for e^z and replacing z by $-1/z^2$, giving

$$f(z) = 1 - z^{-2} + \frac{1}{2!}z^{-4} + \frac{1}{3!}z^{-6} + \cdots$$

and this is a series of the 'negative powers' type. It converges for all z for which $-1/z^2$ is defined, namely $z \neq 0$.

11.1 Series Involving Negative Powers

The general series of this type can be written in the form

$$\sum_{n=-\infty}^{\infty} a_n(z - z_0)^n$$

which is to be thought of as a compact notation for

$$\left(\sum_{n=0}^{\infty} a_n(z - z_0)^n\right) + \left(\sum_{n=1}^{\infty} a_{-n}(z - z_0)^{-n}\right)$$

and hence converges if and only if the two series in parentheses converge. We know that power series converge inside a disc. Consequently power series with negative powers alone should converge *outside* a disc. For instance, that for e^{-1/z^2} converges outside the disc $|z| = 0$. Therefore those with both positive and negative powers should converge in a region *between* two concentric circles. Recall that such a region is called an annulus. More precisely, if $R_1, R_2 \in \mathbb{R}$ with $0 \leq R_1 < R_2 \leq \infty$, and $z_0 \in \mathbb{C}$, then

$$\{z \in \mathbb{C} : R_1 \leq |z - z_0| \leq R_2\}$$

is an annulus.

We begin with an existence theorem for series expansions of the above kind.

THEOREM 11.1 (Laurent's Theorem). *If f is differentiable in the annulus $R_1 \le |z - z_0| \le R_2$ where $0 \le R_1 < R_2 \le \infty$, then*

$$f(z_0 + h) = \sum_{n=0}^{\infty} a_n h^n + \sum_{n=1}^{\infty} b_n h^{-n}$$

where $\sum a_n h^n$ converges for $|h| < R_2$, $\sum b_n h^{-n}$ converges for $|h| > R_1$, and in particular both sides converge in the interior of the annulus.

Further, if $C_r(t) = z_0 + re^{it}$, $(t \in [0, 2\pi])$ then

$$a_n = \frac{1}{2\pi i} \int_{C_r} \frac{f(z)}{(z - z_0)^{n+1}} dz$$

$$b_n = \frac{1}{2\pi i} \int_{C_r} f(z)(z - z_0)^{n-1} dz$$

Before giving the proof:

REMARK 11.2. In the more compact notation we set $c_n = a_n$ $(n \ge 0)$ and $c_{-n} = b_n$ $(n \ge 1)$. Then

$$f(z_0 + h) = \sum_{n=-\infty}^{\infty} c_n h^n$$

which converges in the interior of the annulus, and

$$c_n = \frac{1}{2\pi i} \int_{C_r} \frac{f(z)}{(z - z_0)^{n+1}} dz$$

for all $n \in \mathbb{Z}$. □

Proof of Theorem 11.1. If $R_1 < |h| < R_2$, choose r_1, r_2 such that

$$R_1 < r_1 < |h| < r_2 < R_2$$

and let

$$C_{r_1}(t) = z_0 + r_1 e^{it} \quad (t \in [0, 2\pi])$$

$$C_{r_2}(t) = z_0 + r_2 e^{it} \quad (t \in [0, 2\pi])$$

as in Figure 11.1.

We show first that

$$f(z_0 + h) = \frac{1}{2\pi i} \int_{C_{r_1}} \frac{f(z)}{(z - (z_0 + h))} dz - \frac{1}{2\pi i} \int_{C_{r_2}} \frac{f(z)}{(z - (z_0 + h))} dz \qquad (11.1)$$

To do so, enclose $z_0 + h$ in a small circle

$$S_\varepsilon(t) = z_0 + \varepsilon e^{it} \quad (t \in [0, 2\pi])$$

Then the function

$$F(z) = \frac{f(z)}{z - (z_0 + h)}$$

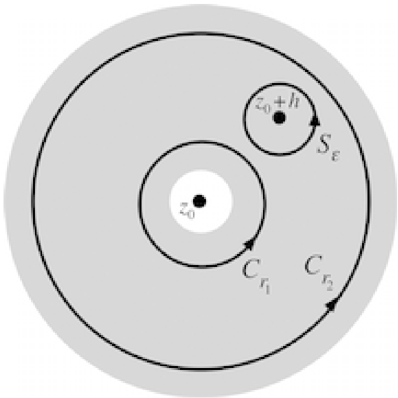

Figure 11.1 An annulus, showing paths used to prove Theorem 11.1.

is differentiable in

$$S = \{z \in \mathbb{C} : R_1 < |z - z_0| < R_2, z \neq z_0 + h\}$$

and the contours $-C_{r_2}, C_{r_1}, S_\varepsilon$ satisfy the hypotheses of Theorem 8.9. Therefore

$$\int_{-C_{r_2}} F(z)\mathrm{d}z + \int_{C_{r_1}} F(z)\mathrm{d}z + \int_{S_\varepsilon} F(z)\mathrm{d}z = 0$$

Hence

$$\int_{S_\varepsilon} F(z)\mathrm{d}z = \int_{C_{r_2}} F(z)\mathrm{d}z - \int_{C_{r_1}} F(z)\mathrm{d}z$$

But by Cauchy's Integral Formula,

$$\int_{S_\varepsilon} F(z)\mathrm{d}z = 2\pi\mathrm{i} \cdot f(z_0 + h)$$

giving (11.1).

(Alternatively we can make cuts and integrate round the two halves, as shown in Figure 11.2. The parts of the contours along the cuts cancel in pairs when integrated.)

All we need do now is work out the two integrals in (11.1) as power series, and calculate the coefficients. Unfortunately this is the longer part of the proof, although the calculations are routine.

First, we choose ρ_1, ρ_2 with

$$r_1 < \rho_1 < |h| < \rho_2 < r_2$$

which enforces the conditions:

(i) $|z - (z_0 + h)| > r_2 - \rho_2$ for z on C_{r_2}; and
(ii) $|z - (z_0 + h)| > \rho_1 - r_1$ for z on C_{r_1}.

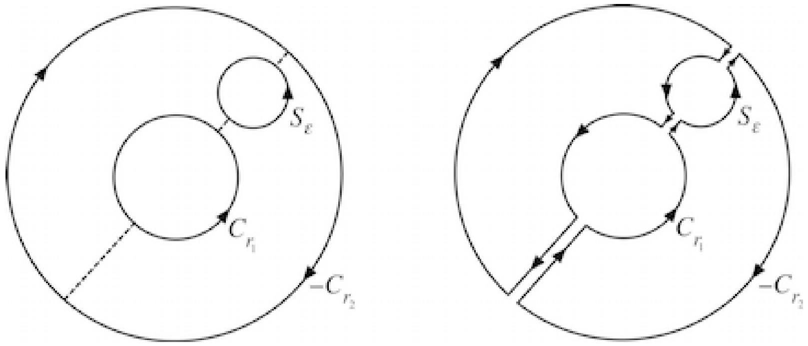

Figure 11.2 Alternative proof using cuts.

As in the proof of the Taylor series expansion, Lemma 10.2, we get

$$\frac{1}{2\pi \mathrm{i}}\int_{C_{r_1}} \frac{f(z)}{z-(z_0+h)}\mathrm{d}z = \sum_{n=0}^{\infty} a_n h^n$$

for $|h| < \rho_2$, where

$$a_n = \frac{1}{2\pi \mathrm{i}}\int_{C_{r_1}} \frac{f(z)}{(z-z_0)^{n+1}}\mathrm{d}z$$

The treatment of the second integral is similar, but will be given in full. Since

$$\frac{1}{h} + \frac{z-z_0}{h^2} + \cdots + \frac{(z-z_0)^{n-1}}{h^n} - \frac{(z-z_0)^n}{h^n(z-z_0-h)} = \frac{-1}{z-z_0-h}$$

(summing a geometric series) we have

$$\begin{aligned} -\frac{1}{2\pi \mathrm{i}}\int_{C_{r_2}} \frac{f(z)}{(z-z_0-h)}\mathrm{d}z &= \frac{1}{2\pi \mathrm{i}}\int_{C_{r_2}} f(z)\left[\frac{1}{h} + \cdots + \frac{(z-z_0)^{n-1}}{h^n} - \frac{(z-z_0)^n}{h^n(z-z_0-h)}\right]\mathrm{d}z \\ &= \sum_{m=1}^{n} b_m h^{-m} - B_n \end{aligned}$$

where

$$\begin{aligned} b_m &= \frac{1}{2\pi \mathrm{i}}\int_{C_{r_2}} f(z)(z-z_0)^{m-1}\mathrm{d}z \\ B_n &= \frac{1}{2\pi \mathrm{i}}\int_{C_{r_2}} \frac{f(z)(z-z_0)^n}{h^n(z-z_0-h)} \end{aligned}$$

Finally we estimate the size of B_n. There exists $M > 0$ such that $|f(z)| \le M$ on C_{r_1} since this circle is closed and bounded. By (i) above, $|z - z_0 - h| > \rho_1 - r_1$. Also $|h| > \rho_1, |z - z_0| = r_1$. Hence

$$|B_n| \le \frac{1}{2\pi}\frac{Mr_1^n}{\rho_1^n(\rho_1 - r_1)} \cdot 2\pi r_1 = \frac{Mr_1}{\rho_1 - r_1}\left(\frac{r_1}{\rho_1}\right)^n$$

which tends to 0 as $b = n \to \infty$, since $r_1/\rho_1 < 1$. It follows that

$$-\frac{1}{2\pi \mathrm{i}} \int_{C_{r_2}} \frac{f(z)}{(z - z_0 - h)} \mathrm{d}z = \sum_{m=1}^{\infty} b_m h^{-m}$$

To finish the proof we must replace C_{r_1} and C_{r_2} by C_r in the expressions for a_n and b_n. Since all three paths are homotopic inside the annulus, or by using cuts, this is immediate. □

Note that we can no longer assert that $a_n = f^{(n)}(z_0)/n!$ since $f(z)$ need not be differentiable for $|z - z_0| < R_1$ under the stated hypotheses.

DEFINITION 11.3. The *Laurent series* or *Laurent expansion of $f(z)$ about z_0* is the series

$$\sum_{n=-\infty}^{\infty} c_n h^n$$

where $h = z - z_0$ and c_n is defined above. □

THEOREM 11.4. *The Laurent expansion of f about z_0 is unique.*

Proof. Suppose that

$$f(z) = \sum_{n=-\infty}^{\infty} d_n (z - z_0)^n$$

Then

$$(z - z_0)^{-m-1} f(z) = \sum_{n=-\infty}^{\infty} d_n (z - z_0)^{n-m-1} = f_1(z) + \frac{d_m}{z - z_0} + f_2(z)$$

where

$$f_1(z) = \sum_{n=-\infty}^{m-1} d_n (z - z_0)^{n-m-1}$$

$$f_2(z) = \sum_{n=m+1}^{\infty} d_n (z - z_0)^{n-m-1}$$

But each of f_1, f_2 has an antiderivative in $R_1 < |z - z_0| < R_2$, because we can integrate term by term. That is, set

$$F_1(z) = \sum_{n=-\infty}^{m-1} \frac{1}{m-n} d_n (z - z_0)^{n-m}$$

$$F_2(z) = \sum_{n=m+1}^{\infty} \frac{1}{m-n} d_n (z - z_0)^{n-m}$$

It is easy to check that the series converge absolutely for $R_1 < |z - z_0| < R_2$ and that $F_1'(z) = f_1(z)$, $F_2'(z) = f_2(z)$. Hence

$$\int_{C_r} f(z)(z-z_0)^{-m-1}\mathrm{d}z = \int_{C_r} \frac{d_m}{z-z_0}\mathrm{d}z = 2\pi \mathrm{i} d_m$$

so $d_m = c_m$ as defined above. □

Example 11.5. Let $f(z) = \mathrm{e}^z + \mathrm{e}^{1/z}$. We know that

$$\mathrm{e}^z = \sum_{n=0}^{\infty} \frac{1}{n!} z^n \quad \text{for all } z$$

$$\mathrm{e}^{1/z} = \sum_{n=0}^{\infty} \frac{1}{n!} z^{-n} \quad \text{for all } z \neq 0$$

so

$$f(z) = \sum_{m=\infty}^{\infty} c_m z^m$$

where

$$\begin{aligned} c_m &= 1/m! && (m \geq 1) \\ c_0 &= 2 \\ c_m &= 1/(-m)! && (m \leq -1) \end{aligned}$$

and the expansion is valid for $z \neq 0$. □

Example 11.6. $f(z) = \dfrac{1}{z} + \dfrac{1}{1-z}$. In a similar way, $f(z) = \sum c_m z^m$ where

$$\begin{aligned} c_m &= 0 && (m < -1) \\ c_m &= 1 && (m \geq -1) \end{aligned}$$

□

Example 11.7. $f(z) = \dfrac{1}{z-1} - \dfrac{1}{z-2}$. Writing this as

$$\frac{1}{z(1-1/z)} + \frac{1}{2(1-z/2)}$$

we obtain an expansion $f(z) = \sum c_m z^m$ where

$$\begin{aligned} c_m &= 1 && (m < -1) \\ c_m &= 2^{-(m+1)} && (m \geq 0) \end{aligned}$$

valid in the annulus $1 < |z| < 2$. □

11.2 Isolated Singularities

If f is differentiable in a *punctured disc*

$$0 < |z - z_0| < R \quad \text{where } R > 0$$

we say that z_0 is an *isolated singularity* of f. We can use the Laurent expansion to study such singularities. There is a Laurent series

$$f(z) = \sum_{n=0}^{\infty} a_n(z - z_0)^n + \sum_{n=1}^{\infty} b_n(z - z_0)^{-n}$$

valid for $0 < |z - z_0| < R$. This series can behave in three radically different ways:

(1) All $b_n = 0$. If we define $f(z_0) = a_0$ we obtain a function that is differentiable on the whole disc $|z - z_0| < R$, with Taylor series

$$\sum_{n=0}^{\infty} a_n(z - z_0)^n$$

In this case z_0 is said to be a *removable singularity*. It arises from our choice of a domain of definition of f, rather than from any intrinsic feature of f.

For example, consider

$$f(z) = \frac{\sin z}{z} \quad (z \neq 0)$$

Around $z_0 = 0$ we have

$$f(z) = 1 - \frac{z^2}{3!} + \frac{z^4}{5!} - \cdots$$

so by defining $f(0) = 1$ we get a function differentiable for all $z \in D$.

(2) Only finitely many $b_n \neq 0$. Then

$$f(z) = \frac{b_m}{(z - z_0)^m} + \cdots + \frac{b_1}{z - z_0} + \sum_{n=0}^{\infty} a_n(z - z_0)^n$$

where $b_m \neq 0$. In this case we say that f has a *pole of order m* at z_0.

For example

$$\begin{aligned} f(z) &= z^{-4} \sin z \quad (z \neq 0) \\ &= \frac{1}{z^3} - \frac{1}{3!z} + \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n+1}}{(2n+5)!} \end{aligned}$$

has a pole of order 3 at $z_0 = 0$.

(2) Infinitely many $b_n \neq 0$. Then we say that z_0 is an *isolated essential singularity*.

For example

$$\begin{aligned} f(z) &= \sin(1/z) \quad (z \neq 0) \\ &= \frac{1}{z} - \frac{1}{3!z^3} + \frac{1}{5!z^5} - \cdots \end{aligned}$$

has an isolated essential singularity at $z_0 = 0$.

We investigate the behaviour of a differentiable function near an isolated singularity, according to these three possibilities.

11.3 Behaviour Near an Isolated Singularity

Removable singularities are trivial and uninteresting – which make it all the more important to recognise them. The next lemma is usually sufficient for this purpose.

LEMMA 11.8. *The following are equivalent for a function f that is differentiable for $0 < |z - z_0| < R$:*

(i) *z_0 is a removable singularity of f.*
(ii) *$\lim_{z \to z_0} f(z)$ exists and is finite.*
(iii) *There exist $M > 0, \delta > 0$ such that $|f(z)| < M$ for $0 < |z - z_0| < \delta$.*

Proof. Trivially (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii), so we need prove only that (iii) $\Rightarrow$ (i).
Suppose (iii) holds, and take a Laurent expansion

$$f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n + \sum_{n=1}^{\infty} b_n (z - z_0)^{-n}$$

Now

$$b_n = \frac{1}{2\pi \mathrm{i}} \int_{C_r} f(z)(z - z_0)^{n-1} \, \mathrm{d}z$$

where $C_r(t) = z_0 + r\mathrm{e}^{\mathrm{i}t}$ $(t \in [0, 2\pi])$ and $0 < r < R$. So

$$|b_n| \leq \frac{1}{2\pi} M r^{n-1} \cdot 2\pi r = M r^n$$

if we let $r \to 0$ it follows that $|b_n| = 0$ for all $n \geq 1$. Thus z_0 is a removable singularity and (i) holds. □

We immediately deduce the useful:

COROLLARY 11.9. *If any coefficient $b_n \neq 0$ $(n \geq 1)$ then f is unbounded on every open disc with centre z_0.* □

For example, if

$$f(z) = \frac{z^2}{(\mathrm{e}^z - 1) \sin z} \quad (z \neq 0)$$

then

$$\begin{aligned}
\lim_{z \to 0} f(z) &= \lim_{z \to 0} \left(\frac{z}{\mathrm{e}^z - 1} \right) \left(\frac{z}{\sin z} \right) \\
&= \lim_{z \to 0} \left(1 + \frac{z}{2!} + \cdots \right)^{-1} \left(1 - \frac{z^2}{3!} + \cdots \right)^{-1} \\
&= 1
\end{aligned}$$

so $z_0 = 0$ is a removable singularity.
There is a similar, slightly more complicated, criterion for poles.

PROPOSITION 11.10. *If f is differentiable for $0 < |z - z_0| < R$, then f has a pole of order m at z_0 if and only if*

$$\lim_{z \to 0} (z - z_0)^m f(z) = l \neq 0$$

Proof. If f has a pole of order m then

$$(z - z_0)^m f(z) = b_m + \cdots + b_1 (z - z_0)^{m-1} + \sum_{n=0}^{\infty} a_n (z - z_0)^n$$

so that

$$\lim_{z \to 0} (z - z_0)^m f(z) = b_m \neq 0$$

Conversely, if the limit is $l \neq 0$, then $g(z) = (z - z_0)^m f(z)$ has a removable singularity at z_0 by Lemma 11.8, so there is a series

$$g(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n$$

valid for $|z - z_0| < R$, and $a_0 = l \neq 0$. But now

$$f(z) = \frac{a_0}{(z - z_0)^m} + \cdots + \frac{a_{m-1}}{z - z_0} + \sum_{n=0}^{\infty} a_{n+m} (z - z_0)^n$$

where $a_0 \neq 0$, so f has a pole of order m at z_0. □

For example, consider

$$f(z) = \frac{5z + 3}{(1 - z)^3 \sin^2 z} \quad (0 < |z| < 1)$$

Then $\lim_{z \to 0} z^2 f(z) = 3 \neq 0$, so there is a double pole (pole of order 2) at the origin. Also, $\lim_{z \to 1} (z - 1)^3 f(z) = -8/\sin^2(1) \neq 0$, so there is a triple pole (pole of order 3) at $z_0 = 1$.

COROLLARY 11.11. *The function $f(z)$ has a pole of order m at z_0 if and only if $1/f(z)$ has a removable singularity at z_0 which, if removed, gives rise to a zero of order m at z_0. This in particular occurs when $1/f(z)$ has a zero of order m at z_0.* □

Proof. If $f(z)$ has a pole of order m at z_0 then $g(z) = (z - z_0)^m f(z)$ has a removable singularity at z_0. Further, these exists $\delta > 0$ such that $g(z) \neq 0$ for $|z - z_0| < \delta$. So

$$1/f(z) = (z - z_0)^m / g(z)$$

and $1/g(z)$ is differentiable for $|z - z_0| < \delta$. Therefore, with the singularity removed, $1/f(z)$ has a zero of order m at z_0.

The converse is proved by an almost identical argument. □

COROLLARY 11.12. *If $f(z)$ has a pole of order m at z_0, then*

$$\lim_{z \to z_0} |f(z)| = +\infty$$ □

Proof.

$$\begin{aligned}\lim_{z\to z_0} |f(z)| &= \lim_{z\to z_0} |g(z)/(z-z_0)^m| \\ &= |l| \cdot \lim_{z\to z_0} |z-z_0|^{-m} \\ &= +\infty\end{aligned}$$

□

Thus near a pole the behaviour of f is really quite good. However, near an isolated essential singularity the behaviour is much wilder. The following result is classical:

THEOREM 11.13 (Weierstrass–Casorati Theorem). *In every neighbourhood of an isolated essential singularity z_0 a differentiable function f takes values arbitrarily close to any assigned complex number. Specifically, given $r > 0, \varepsilon > 0$, and $w \in \mathbb{C}$ there exists z_1 such that $|z_1 - z_0| < R$ and $|f(z_1) - w| < \varepsilon$.*

Proof. We have

$$f(z) = \sum_{n=0}^{\infty} a_n(z-z_0)^n + \sum_{n=1}^{\infty} b_n(z-z_0)^{-n}$$

with infinitely many $b_n \neq 0$. If we define $\phi(z) = f(z) - w$ then the Laurent series of $\phi(z)$ differs from that of $f(z)$ only in the coefficient a_0, which becomes $a_0 - w$. Hence $\phi(z)$ also has an isolated essential singularity at z_0, so we need prove only that $\phi(z)$ can be made arbitrarily small in any neighbourhood of z_0. That is, there exists z_1 with

$$0 < |z_1 - z_0| < r \qquad |\phi(z_1)| < \varepsilon$$

If z_0 is a limit of zeros of ϕ this is trivial. Failing this, there exists $\rho > 0$ such that $\phi(z) \neq 0$ for $0 < |z_1 - z_0| < \rho$. Either there exists z_1 with $0 < |z_1 - z_0| < r$ and $1/\phi(z_1) > 1/\varepsilon$, or else $1/\phi(z)$ is bounded for $0 < |z_1 - z_0| < r$. If the former, then the theorem is proved. If the latter, then $1/\phi(z)$ has a removable singularity at z_0 by Lemma 11.8, so by Corollary 11.11 $\phi(z)$ has at worst a pole at z_0, which contradicts z_0 being essential. Thus the latter case cannot occur. □

In point of fact, a stronger and more satisfying result is true. Émile Picard (1856–1941) proved that in every neighbourhood of an isolated essential singularity, a differentiable function f takes every value, *with at most one exception*. For instance, $\sin(1/z)$ takes every value in $0 < |z| < r$ for any $r > 0$, as can easily be verified. The exception can occur: $e^{1/z}$ misses out the value 0 but attains all others in $0 < |z| < r$ for any $r > 0$. But the proof of Picard's Theorem requires machinery (elliptic modular functions) considerably beyond the reach of this text.

11.4 The Extended Complex Plane, or Riemann Sphere

In real analysis it is standard to extend the real line $\mathbb{R}$ by adjoining two points at infinity, $+\infty$ and $-\infty$. We then have a simple way to discuss and visualise limits such as $\lim_{x\to+\infty} 1/x$ and $\lim_{x\to-\infty} e^x$.

Sometimes it is more helpful to consider these two points to be identical. For example, the graph of $y = 1/x$ is a hyperbola, which heads off to $\pm\infty$ along both axes. If we identify $+\infty$ with $-\infty$ (known as the one-point-compactification of $\mathbb{R}$) the extended hyperbola closes up to form a closed curve, like an ellipse. Both of these curves are conic sections, and in projective geometry they are projectively equivalent. So the projective line extends $\mathbb{R}$ by adjoining a single point at infinity.

There is a similar extension of the complex plane, which lets us describe the behaviour of a complex function 'at infinity' by adjoining to $\mathbb{C}$ a single extra point '∞'. We adjoin just one point in the complex case because the distinction between the two ends of the real line becomes blurred when we deal with the whole plane. For instance, if we rotate $\mathbb{C}$ continuously through π, $+\infty$ and $-\infty$ in $\mathbb{R}$ swap places. The method for adjoining this extra point was introduced by Bernhard Riemann, and has an elegant geometric realisation.

Think of $\mathbb{C}$ as being embedded in the (x, y)-plane in $\mathbb{R}^3$, so that a point $x + iy \in \mathbb{C}$ is identified with $(x, y, 0) \in \mathbb{R}^3$. Let

$$S^2 = \{(\xi, \eta, \zeta) \in \mathbb{R}^3 : \xi^2 + \eta^2 + \zeta^2 = 1\}$$

be the *unit sphere*. A line joining the North pole $(0, 0, 1)$ to $(x, y, 0)$ cuts S^2 in a unique point (ξ, η, ζ), giving a one-to-one correspondence between $\mathbb{C}$ and all points on S^2 except the North pole, Figure 11.3.

It is easy to verify that

$$(\xi, \eta, \zeta) \quad \text{corresponds to} \quad \left(\frac{\xi}{1-\zeta}\right) + i\left(\frac{\eta}{1-\zeta}\right)$$

As (ξ, η, ζ) gets near to $(0, 0, 1)$ it follows (as is obvious geometrically) that $|x + iy|$ becomes very large. Thus it is reasonable to introduce the symbol

$$\infty$$

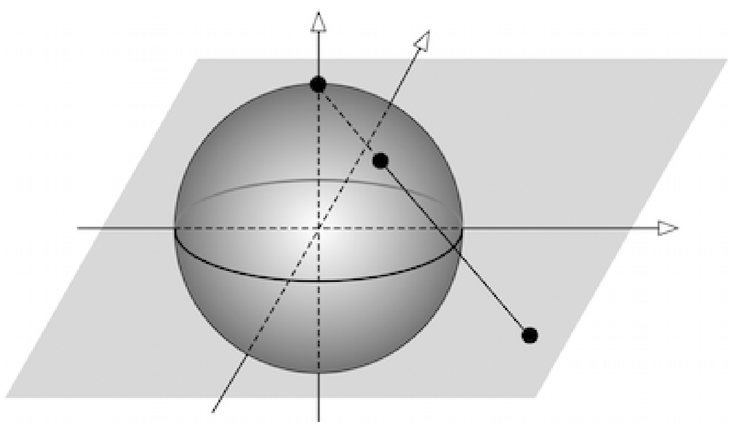

Figure 11.3 The Riemann sphere.

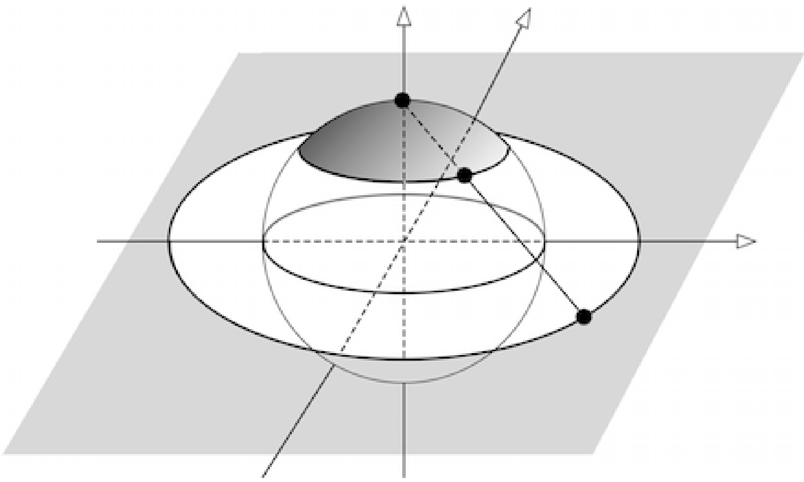

Figure 11.4 Neighbourhood of ∞ on the Riemann sphere.

to correspond to $(0, 0, 1) \in S^2$. Thus we have a one-to-one correspondence between S^2 and $\mathbb{C} \cup \{\infty\}$. We call the latter the *extended complex plane*. It may be identified with S^2, which is then known as the *Riemann sphere*.

We can now think of $\mathbb{C} \cup \{\infty\}$ either as a plane plus an extra point, or as a sphere. Correspondingly, we think of $\mathbb{C}$ as a plane, or as a sphere without a North pole. Both viewpoints are valuable, depending on the problem at hand.

Since $\{x + iy : |x + iy| > R\}$ corresponds to a 'spherical cap' between a line of latitude and the North pole (Figure 11.4) it makes sense to think of $\{z \in \mathbb{C} : |z| > R\}$ as a 'neighbourhood of ∞'. Such neighbourhoods get smaller as R gets bigger. So doing this leads to a concept of continuity on S^2, agreeing with geometric intuition (and the standard distance function on $\mathbb{R}^3$). Readers familiar with topology can render this in more precise terms: we obtain a topology on the Riemann sphere that is identical to its usual topology as a subset of $\mathbb{R}^3$.

11.5 Behaviour of a Differentiable Function at Infinity

Suppose that $f(z)$ is differentiable in $\{z \in \mathbb{C} : |z| > R\}$. Then we can define

$$g(z) = f(1/z) \quad (0 < |z| < 1/R)$$

Since $g'(z) = -z^{-2}f'(1/z)$ it follows that $g(z)$ is differentiable for $0 < |z| < 1/R$, and therefore has an isolated singularity at 0.

DEFINITION 11.14. The function f *has a removable singularity, pole of order m*, or *isolated essential singularity at* ∞ if and only if $g(z) = f(1/z)$ has the corresponding type of singularity at 0. □

Example 11.15. $f(z) = 1/z$ ($|z| > 0$). Then $g(z) = z$ ($|z| > 0$), so f has a removable singularity at ∞. □

Example 11.16. $f(z) = z$. Then $g(z) = 1/z$ ($|z| > 0$), so f has a simple pole at ∞. □

Example 11.17. $f(z) = e^z$. Then $g(z) = e^{1/z}$ ($|z| > 0$) and f has an isolated essential singularity at ∞. □

Example 11.18. $f(z) = 1/(\sin z)$ ($z \neq n\pi, n \in \mathbb{Z}$). Then $g(z) = 1/\sin(1/z)$. We cannot say that f has an isolated essential singularity at ∞, because f is not differentiable in $\{z \in \mathbb{C} : |z| > R\}$ for any $R > 0$. However, f certainly has *some* sort of singularity at ∞. In general, if $f(z)$ has a sequence of isolated singularities z_n such that $z_n \to \infty$, we say that f has an essential (but not isolated) singularity at ∞. □

If f has a removable singularity at ∞ then g has a removable singularity at 0, so may be rendered differentiable in $\{z \in \mathbb{C} : |z| < 1/R\}$. Thus we may say that f is *differentiable* (or *analytic) at* ∞, and define

$$f(\infty) = g(0) = \lim_{z \to 0} g(z)$$

and now

$$f(\infty) = \lim_{z \to \infty} f(z)$$

For instance, if $f(z) = 1/z$ then $g(0) = 0$ so $f(\infty) = 0$.

Further, we say that f has a *zero of order m at* ∞ if g has a zero of order m at 0. Thus if $f(z) = 1/z^m$ ($|z| > 0, m \in \mathbb{Z}, m > 0$) then $g(z) = z^m$ and f has a zero of order m at ∞.

In general, any statement about the behaviour of f at ∞ can be translated into one about g at 0 – and this is how to define or prove such a statement.

11.6 Meromorphic Functions

Poles are not, as singularities go, particularly nasty, and it is of interest to consider a class of functions more general than differentiable ones.

DEFINITION 11.19. If f is differentiable everywhere in a domain D except for points at which f has a pole, then f is *meromorphic* in D. □

For instance, $f(z) = 1/z$ is differentiable in $\mathbb{C} \setminus \{0\}$, and has a pole at 0, so it is meromorphic in $\mathbb{C}$. Similarly, $f(z) = 1/(\sin z)$ is differentiable in $\mathbb{C} \setminus \{n\pi : n \in \mathbb{Z}\}$, and has poles at $n\pi$, so it is meromorphic in $\mathbb{C}$.

We might go further, and consider functions meromorphic in the extended complex plane – by which we mean that the only singularities of f, including ∞, if necessary are poles. It is not necessary to introduce a special term for such functions, because they turn out to have a simple description. Recall that a *rational* function is one of the form

$$f(z) = \phi(z)/\psi(z)$$

where ϕ, ψ are polynomial functions. Then we have:

PROPOSITION 11.20. *A function is meromorphic in the extended complex plane if and only if it is rational.*

Proof. Clearly a rational function is meromorphic in $\mathbb{C} \cup \{\infty\}$.

Suppose conversely that f is meromorphic in $\mathbb{C} \cup \{\infty\}$. Since f is differentiable at ∞ or has a pole at ∞, there exists $R > 0$ such that f is differentiable in

$$\{z \in \mathbb{C} : |z| > R\}$$

So the poles, other than ∞, occur inside the closed disc

$$\{z \in \mathbb{C} : |z| \leq R\}$$

Since poles are isolated singularities, each pole has a neighbourhood that contains no other poles. Indeed, if there were an infinite set S of poles inside $|z| \leq R$, there would be a limit point of S. This is not possible because the poles are isolated. This closed disc contains only finitely many poles, say $z_1, \ldots, z_k$. Let the orders of these poles be $n_1, \ldots, n_k$. Define

$$g(z) = (z - z_1)^{n_1} \cdots (z - z_k)^{n_k} f(z)$$

This is differentiable throughout $\mathbb{C}$, hence analytic, so it has a Taylor series

$$g(z) = \sum_{n=0}^{\infty} a_n z^n \quad (z \in \mathbb{C})$$

Now the polynomial

$$\psi(z) = (z - z_1)^{n_1} \cdots (z - z_k)^{n_k}$$

has a pole of order $n_1 + \cdots + n_k$ at ∞, since

$$\begin{aligned} \psi\left(\frac{1}{z}\right) &= \left(\frac{1}{z} - z_1\right)^{n_1} \cdots \left(\frac{1}{z} - z_k\right)^{n_k} \\ &= \frac{1}{z^{n_1+\cdots+n_k}} + \cdots \end{aligned}$$

Since f has at worst a pole of order N at ∞ for some N it follows that $g(z)$ has a pole of order

$$M = n_1 + \cdots + n_k + N$$

at ∞. Now

$$g(1/z) = \sum_{n=0}^{\infty} a_n z^{-n} \quad (|z| > 0)$$

and since this has a pole of order M at 0 we must have $a_n = 0$ for $n > M$. Therefore

$$g(z) = a_0 + \cdots + a_M z^M$$

which is a polynomial, so $f(z) = g(z)/\psi(z)$, which is rational. □

A function meromorphic only in $\mathbb{C}$ need not be rational, as the example $1/(\sin z)$ shows. This is not rational since it has infinitely many poles, but we have already seen that it is meromorphic in $\mathbb{C}$. So the behaviour at infinity is important.

11.7 Exercises

1. Find the Laurent expansions of the following functions of z around $z = 0$:
(i) $(z-3)^{-1}$
(ii) $(z-a)^{-k}$ $(a \in \mathbb{C}, k = 1, 2, 3, \ldots)$
(iii) $1/(z(1-z))$
(iv) $1/((z-a)(z-b))$ $(a, b, \in \mathbb{C})$
(v) $z^3 e^{1/z}$
(vi) $\exp(z + 1/z)$
(vii) $\cos(1/z)$
(viii) $\exp(z^{-5})$
In each case, specify the largest annulus in which the expansion is valid.

2. Find the Laurent expansions of the following functions on the stated annulus:
(i) $(z-1)^{-2}(z-2)^{-1}$ on $0 < |z| < 1$
(ii) $(z-1)^{-2}(z-2)^{-1}$ on $1 < |z| < 2$
(iii) $(z-1)^{-2}(z-2)^{-1}$ on $2 < |z| < 3$
(iv) $\exp(-z^{-2})$ on $|z| > 0$
(v) $(1-z-z^2)^{-1}$ in powers of $z-1$ on $0 < |z-1| < 1$
(vi) $e^z/(1+z^2)$ on $|z| > 1$

3. Which of the following functions have a Laurent expansion around the given point z_0 (that is, in powers of $z - z_0$)? (For multivalued functions, choose one particular definition on a domain that makes it single-valued.)
(i) $\sqrt{z}$ about $z_0 = 1$
(ii) $\sqrt{z}$ about $z_0 = 0$
(iii) $\operatorname{Log} z$ about $z_0 = 0$
(iv) $\operatorname{Log} z$ about $z_0 = 3$
(v) $\sqrt{1+\sqrt{z}}$ about $z_0 = 0$
(vi) $\tan^{-1}(1+z)$ about $z_0 = 0$
(vii) $\sin^{-1}(z)$ about $z_0 = 0$
(viii) $\sqrt{(\pi/2) - \sin^{-1}(z)}$ about $z_0 = 1$
(ix) $z^2 \operatorname{cosec}(1/z)$ about $z_0 = 0$
(x) $\sqrt{(\pi/4) - \sin^{-1}(z)}$ about $z_0 = 1/\sqrt{2}$

4. Prove the validity of the Laurent expansion

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

on a suitable annulus, and state which annulus.
5. Find Laurent series for $(z^2-1)^{-1}$ and $(z^2+1)^{-1}$ in powers of $z+\mathrm{i}$ and $z-\mathrm{i}$, and say in which annuli these are valid.
6. Let $a, b \in \mathbb{C}$. Show that

$$\exp(az + bz^{-1}) = \sum_{-\infty}^{\infty} a_n z^n$$

where

$$a_n = \frac{1}{2\pi} \int_0^{2\pi} \mathrm{e}^{(a+b)\cos\theta} \cos[(a-b)\sin\theta - n\theta]\mathrm{d}\theta$$

7. Let $a \in \mathbb{C}$. Show that

$$\sin(a(z+z^{-1})) = \sum_{-\infty}^{\infty} a_n z^n$$

where

$$a_n = \frac{1}{2\pi} \int_0^{2\pi} \sin(2a\cos\theta)\cos n\theta \, \mathrm{d}\theta$$

8. Find the poles of the functions:
   (i) $1/(z^2+1)$
   (ii) $1/(z^4+16)$
   (iii) $1/(z^4+2z^2+1)$
   (iv) $1/(z^2+z-1)$
9. Describe the type of singularity at 0 of each of the following functions:
   (i) $\sin(1/z)$ $(z \neq 0)$
   (ii) $z^{-3}\sin^2 z$
   (iii) $z\cot z$ $(z \neq n\pi, n \in \mathbb{Z})$
   (iv) $\operatorname{cosec}^2 z - z^{-2}$ $(z \neq n\pi, n \in \mathbb{Z})$
   (v) $(\cos z - 1)/z^2$
   (vi) $(\sin z - z + z^3/6)/z^7$
10. Let D be a disc centre z_0, let f be differentiable on D except at z_0, and suppose that $|f(z)|$ is bounded on $D \setminus \{z_0\}$. Show that z_0 is a removable singularity of f. (Hint: negative Laurent terms are 0 – why?)
11. Find all singularities of the following functions, and say which are poles:
    (i) $(z+z^{-1})^{-1}$
    (ii) $\dfrac{\cos \pi z}{1-4z^2}$
    (iii) $\exp(z+z^{-1})$

12. Construct a function defined on $\mathbb{C} \cup \{\infty\}$ having only the following singularities:
 (i) A pole of order 2 at ∞.
 (ii) A simple pole at each of the points $e^{2\pi i k/p}$, $k = 0, \ldots, p-1$, p an odd integer ≥ 2.
 (iii) A simple pole at $z = 2$ and a pole of order 5 at $z = \sqrt{2}$.

13. Let $p(z), q(z)$ be polynomials of degrees m, n respectively. Describe the behaviour at infinity of:
 (i) $p(z) + q(z)$
 (ii) $p(z)q(z)$
 (iii) $p(z)/q(z)$

14. Find all singularities, and the behaviour at infinity, of the functions:
 (i) $(z - z^3)^{-1}$
 (ii) $z^5/(2 - z^2)^2$
 (iii) $(e^z - 1)^{-1} - 1/z$
 (iv) $\cot 1/z$
 (v) $(\cos z)z^{-2}$
 (vi) $((\cos z) - 1)z^{-2}$
 (vii) $\cot(1/z) - 1/z$
 (viii) $\sin(1/\cos(1/z))$

15. Let Γ be a circle in the complex plane, or a straight line. Is its image on the Riemann sphere also a circle?

16. Find the poles and zeros of $\tan z$. Show that $\tan z$ is meromorphic in $\mathbb{C}$, but is not a rational function.

17. Show that $(z + 1 + z^{-1})^{-1}$ has a removable singularity at $z = 0$. Find its Taylor expansion, and the radius of convergence of this.

18. Verify Picard's Theorem directly for the functions:
 (i) e^z
 (ii) $\tan^2 z$
 (iii) z^2
 (iv) $\sin z$
 (v) $e^{1/z}$
 (vi) $\cos z$
 (vii) $\tan z$
 (viii) A function of your own choice.

19. Let $f(z)$ have a pole of order n at $z = a$. Define the *principal part* $\phi(z)$ of $f(z)$ to be the sum of the negative-power terms in the Laurent expansion of $f(z)$ in powers of $z - a$. Prove that $f(z) - \phi(z)$ is differentiable at $z = a$.

20. Show that in a neighbourhood of a pole, a complex function is the sum of a rational function and a differentiable one.

21. Suppose that f is differentiable on $\mathbb{C}$ except at poles, and that ∞ is either a pole or a removable singularity of f. Show that:

(i) f has only finitely many poles.
(ii) $f - \sum_r p_r$ is constant, where the poles of f are at points a_r and the p_r are the corresponding principal parts of f (defined in Exercise 19).
(iii) f is a rational function, and $\sum_r p_r$ is its 'partial fraction' decomposition.

22. Let $f : \mathbb{C} \to \mathbb{C}$ be differentiable, with $f(z) \neq 0$ for all $z \in \mathbb{C}$. Suppose that $\lim_{z\to\infty} f(z)$ exists and is non-zero. Prove that f is constant.

12 Residues

Among the many applications of complex analysis are the explicit computation of definite integrals, the summation of series, and counting how many zeros a function has in a given region of $\mathbb{C}$. Although such questions are not as important a part of pure mathematics as they once were, they are still very relevant to practical applications. Further, the power of the method and its wide applicability demonstrate the advantage of general principles and deep theorems over manipulative ingenuity.

The basic idea in all of these applications is to use Cauchy's Theorem to exploit the exceptional nature of the term $b_1/(z - z_0)$ in the Laurent expansion of a differentiable function.

12.1 Cauchy's Residue Theorem

The residue of f at z_0 is determined by the special coefficient b_1 just described:

DEFINITION 12.1. If f has an isolated singularity at z_0 and a Laurent expansion

$$f(z_0 + h) = \sum_{n=0}^{\infty} a_n h^n + \sum_{n=1}^{\infty} b_n h^{-n} \quad (0 < |h| < R)$$

then the *residue* of f at z_0 is

$$\operatorname{res}(f, z_0) = b_1 \qquad \square$$

From Theorem 11.1 we immediately deduce that

$$\operatorname{res}(f, z_0) = \frac{1}{2\pi \mathrm{i}} \int_{C_r} f(z)\mathrm{d}z \tag{12.1}$$

where $C_r(t) = z_0 + \mathrm{e}^{\mathrm{i}t}$ $(t \in [0, 2\pi])$ and $0 < r < R$. This shows the relevance of residues to integration.

DEFINITION 12.2. A closed path γ is a *simple loop* if every point z not on γ has $w(\gamma, z) = 0$ or $w(\gamma, z) = 1$.

As usual, the set of points $z \in \mathbb{C}$ satisfying $w(\gamma, z) \neq 0$ (which in this case means $w(\gamma, z) = 1$) are said to be *inside* γ. $\square$

In applications, simple loops will all be made up of straight line segments and parts of circles, as in Figure 12.1. All the simple loops encountered will be Jordan contours (that

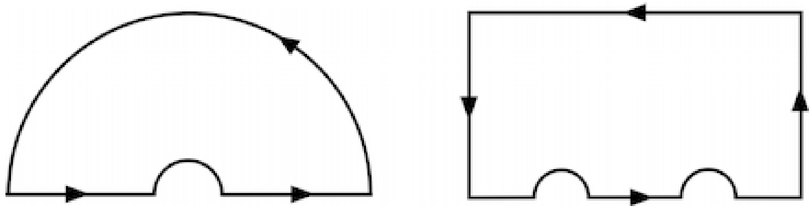

Figure 12.1 Simple loops in $\mathbb{C}$.

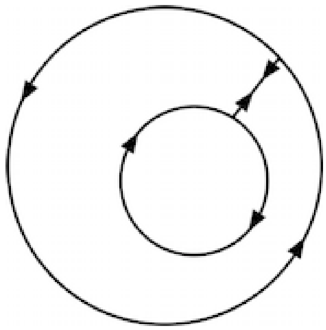

Figure 12.2 A simple loop in $\mathbb{C}$ that is not a Jordan contour.

is, they do not self-intersect), but that is not essential for the theory. (Figure 12.2 shows a simple loop according to our definition that is not a Jordan contour.) What matters is that the points inside γ all have $w(\gamma, z) = 1$. (In particular, γ winds once *anticlockwise* around such points.)

In such cases, we have:

THEOREM 12.3 (Cauchy's Residue Theorem). *Let D be a domain containing a simple loop γ and the points inside γ. If f is differentiable in D except for finitely many isolated singularities at $z_1, \ldots, z_n$ inside γ, then*

$$\int_\gamma f(z)\mathrm{d}z = 2\pi \mathrm{i} \sum_{r=1}^{n} \mathrm{res}\,(f, z_r)$$

Proof. Since D is open we can find circles $S_r(t) = z_r + \varepsilon_r \mathrm{e}^{\mathrm{i}t}$ $(t \in [0, 2\pi])$ round the z_r such that S_r and points inside it lie inside D, and such that S_r contains no singularity of f other than z_r, Figure 12.3. Then the collection of paths

$$-\gamma, S_1, \ldots, S_n$$

satisfies the hypotheses of Theorem 8.9, because

$$\begin{aligned} w(-\gamma, z) &= w(S_r, z) = 0 && (z \notin D) \\ w(-\gamma, z_r) &= -w(\gamma, z_r) = -1 && \text{since } z_r \text{ is inside } \gamma \\ w(S_m, z_r) &= 0 \text{ if } m \neq r,\ 1 && \text{if } m = r \end{aligned}$$

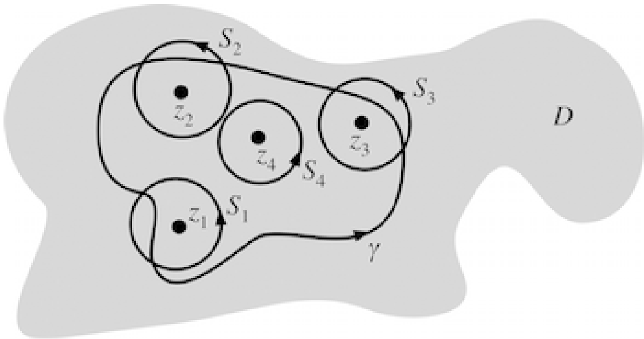

Figure 12.3 Small circular contours around each isolated singularity.

Therefore

$$w(-\gamma, z) + w(S_1, z) + \cdots + w(S_n, z) = 0$$

for all $z \notin D$.

By the Generalised Cauchy Theorem,

$$\int_{-\gamma} f + \int_{S_1} f + \cdots + \int_{S_n} f = 0$$

so that

$$\begin{aligned}\int_{\gamma} f &= \int_{S_1} f + \cdots + \int_{S_n} f \\ &= 2\pi \mathrm{i} \cdot \mathrm{res}\,(f, z_1) + \cdots + 2\pi \mathrm{i} \cdot \mathrm{res}\,(f, z_n)\end{aligned}$$

by (12.1). □

Alternative Proof. Instead of the Generalised Cauchy Theorem we can use Laurent series: the proof is instructive but less elegant. Around each z_r there is a Laurent expansion of f, which we split into three parts:

$$f(z) = Q_r(z) + \frac{\mathrm{res}\,(f, z_r)}{z - z_r} + P_r(z) \tag{12.2}$$

where

$$Q_r(z) = \sum_{n=2}^{\infty} b_n (z - z_r)^{-n}$$

$$P_r(z) = \sum_{n=0}^{\infty} a_n (z - z_r)^{n}$$

Now $P_r(z)$ is differentiable in a neighbourhood of z_r, and $Q_r(z) + \mathrm{res}\,(f, z_r)/(z - z_r)$ is differentiable for $z \neq z_r$ since all but a finite number of the b_n are zero. Therefore

$$h(z) = f(z) - \sum_{r=1}^{n} Q_r(z) - \sum_{j=1}^{n} \frac{\mathrm{res}\,(f, z_r)}{z - z_r} \tag{12.3}$$

is differentiable in D, except perhaps at $z_1, \ldots, z_n$. But from (12.2) it is also differentiable in a neighbourhood of each z_r. Hence $h(z)$ is differentiable in D, so by Cauchy's Theorem

$$\int_\gamma h(z)\mathrm{d}z = 0 \tag{12.4}$$

But $Q_r(z) = T_r'(z)$ where

$$T_r(z) = \sum_{n=2}^{\infty} \frac{b_n}{-n+1}(z-z_r)^{-n+1}$$

so that

$$\int_\gamma Q_r(z)\mathrm{d}z = 0 \tag{12.5}$$

as well. Integrating round γ and using (12.4) and (12.5), we get

$$\begin{aligned} 0 &= \int_\gamma f(z)\mathrm{d}z - 0 - \sum_{r=1}^{n} \operatorname{res}(f, z_r) \int_\gamma (z-z_r)^{-1}\mathrm{d}z \\ &= \int_\gamma f(z)\mathrm{d}z - 2\pi \mathrm{i} \sum_{r=1}^{n} \operatorname{res}(f, z_r) \end{aligned}$$

□

12.2 Calculating Residues

The Residue Theorem can be used to calculate integrals – and not just round simple loops. For it to be of much use, we must find ways to calculate residues. The following two lemmas are very useful in this respect.

LEMMA 12.4. (i) *If z_0 is a simple pole of f then*

$$\operatorname{res}(f, z_0) = \lim_{z\to z_0} (z-z_0)f(z)$$

(ii) *If $f(z) = p(z)/q(z)$ where $p(z_0) \neq 0, q(z_0) = 0, q'(z_0) \neq 0$, then*

$$\operatorname{res}(f, z_0) = p(z_0)/q'(z_0)$$

Proof. We have already done part (i) in the previous chapter as Proposition 11.10, but to recap: we have

$$f(z) = \frac{b_1}{z-z_0} + \sum_{n=0}^{\infty} a_n(z-z_0)^n$$

and so

$$(z-z_0)f(z) = b_1 + \sum_{n=0}^{\infty} a_n(z-z_0)^{n+1}$$

which tends to b_1 as $z \to z_0$.

For (ii), note that

$$\lim_{z\to z_0} \frac{(z-z_0)p(z)}{q(z)} = \lim_{z\to z_0} p(z) \Big/ \left(\frac{q(z)-q(z_0)}{z-z_0}\right)$$

since $q(z_0) = 0$, and this is equal to $p(z_0)/q'(z_0)$. □

For example, if

$$f(z) = \frac{\cos \pi z}{z^{976}}$$

then

$$\text{res}\,(f, 1) = \frac{\cos \pi}{976 \cdot 1^{975}} = \frac{1}{976}$$

LEMMA 12.5. *If z_0 is a pole of f of order m then*

$$\text{res}\,(f, z_0) = \lim_{z \to z_0} \left[\frac{1}{(m-1)!} \frac{\mathrm{d}^{m-1}}{\mathrm{d}z^{m-1}} ((z - z_0)^m f(z)) \right]$$

Proof. We have

$$f(z) = \frac{b_m}{(z - z_0)^m} + \cdots + \frac{b_1}{z - z_0} + \sum_{n=0}^{\infty} a_n (z - z_0)^n$$

so that

$$(z - z_0)^m f(z) = b_m + \cdots + b_1 (z - z_0)^{m-1} + \sum_{n=0}^{\infty} a_n (z - z_0)^{m+n}$$

Therefore

$$\frac{\mathrm{d}^{m-1}}{\mathrm{d}z^{m-1}} ((z - z_0)^m f(z)) = (m-1)! b_1 + \sum_{n=0}^{\infty} \frac{(m+n)!}{(n+1)!} a_n (z - z_0)^{n+1}$$

Now take limits. □

For example, consider

$$f(z) = \left(\frac{z+1}{z-1} \right)^3$$

which has a triple pole at $z_0 = 1$. Then

$$(z-1)^3 f(z) = (z+1)^3$$

and so

$$\frac{1}{2!} \frac{\mathrm{d}^2}{\mathrm{d}z^2} ((z-1)^3 f(z)) = \frac{6}{2!} (z+1)$$

which tends to $3 \cdot 2 = 6$ as $z \to 1$. So $\text{res}\,(f, 1) = 6$.

On occasion another technique may be brought into play: working out the appropriate part of the Laurent series. (It is a waste to work out the whole thing, because the whole point about residues is that we do not need the whole thing, but only b_1.) For instance:

$$\begin{aligned} f(z) &= \frac{1}{z^2 \sin z} \\ &= 1 \Big/ \left(z^2 \left(z - \frac{z^3}{6} + \cdots \right) \right) \\ &= \frac{1}{z^3} \left(1 - \frac{z^6}{6} + \cdots \right)^{-1} \end{aligned}$$

$$= \frac{1}{z^3}\left(1+\frac{z^6}{6}+\cdots\right)$$
$$= \frac{1}{z^3}+\frac{1}{6z}+\cdots$$

so that $\operatorname{res}(f,0) = 1/6$.

12.3 Evaluation of Definite Integrals

We now consider a number of techniques for calculating various kinds of definite integral.

I: $\displaystyle\int_0^{2\pi} Q(\cos t,\ \sin t)\mathrm{d}t$

Let $C(t) = \mathrm{e}^{\mathrm{i}t}$ $(t \in [0, 2\pi])$ be the unit circle. If

$$z = C(t) = \mathrm{e}^{\mathrm{i}t}$$

then

$$\cos t = \frac{1}{2}\left(z+\frac{1}{z}\right)$$
$$\sin t = \frac{1}{2\mathrm{i}}\left(z-\frac{1}{z}\right)$$

from which we get

$$\int_0^{2\pi} Q(\cos t, \sin t)\mathrm{d}t = \int_C Q\left(\frac{1}{2}\left(z+\frac{1}{z}\right), \frac{1}{2\mathrm{i}}\left(z-\frac{1}{z}\right)\right)\frac{\mathrm{d}z}{\mathrm{i}z}$$
$$= 2\pi\mathrm{i}\Sigma$$

where Σ is the sum of the residues of

$$\frac{1}{\mathrm{i}z}Q\left(\frac{1}{2}\left(z+\frac{1}{z}\right), \frac{1}{2\mathrm{i}}\left(z-\frac{1}{z}\right)\right) \tag{12.6}$$

inside C.

For example, consider

$$\int_0^{2\pi}(\cos^3 t+\sin^2 t)\mathrm{d}t$$

Then (12.6) becomes

$$\frac{1}{\mathrm{i}z}\left(\frac{1}{8}\left(z+\frac{1}{z}\right)^3-\frac{1}{4}\left(z-\frac{1}{z}\right)^2\right)=$$
$$\frac{1}{8\mathrm{i}}z^2-\frac{1}{4\mathrm{i}}z+\frac{3}{8\mathrm{i}}+\frac{1}{2\mathrm{i}z}+\frac{3}{8\mathrm{i}z^2}-\frac{1}{4\mathrm{i}z^3}+\frac{1}{8\mathrm{i}z^4}$$

which has a pole inside C with residue $1/2\mathrm{i}$. So the integral is $2\pi\mathrm{i}/2\mathrm{i} = \pi$.

If Q is at all complicated, the computations can become very tedious. Sometimes integrals of this kind can be found from the real and imaginary parts of an integral

$$\int_C g(z)\mathrm{d}z$$

with a suitable choice of g. For instance,

$$\int_C \frac{\mathrm{e}^z}{z}\mathrm{d}z = 2\pi\mathrm{i}$$

since e^z/z has residue 1 at $z = 0$. Therefore

$$\int_0^{2\pi} \frac{\mathrm{e}^{\cos t + \mathrm{i}\sin t}}{\mathrm{e}^{\mathrm{i}t}}\mathrm{d}t = 2\pi\,\mathrm{i}$$

so

$$\int_0^{2\pi} \mathrm{e}^{\cos t + \mathrm{i}\sin t}\mathrm{d}t = 2\pi$$

and

$$\int_0^{2\pi} \mathrm{e}^{\cos t}[\cos(\sin t)) + \mathrm{i}\sin(\sin t)]\mathrm{d}t = 2\pi$$

Equating real and imaginary parts,

$$\int_0^{2\pi} \mathrm{e}^{\cos t}\cos(\sin t) = 2\pi$$
$$\int_0^{2\pi} \mathrm{e}^{\cos t}\sin(\sin t) = 0$$

II: $\displaystyle\int_{-\infty}^{\infty} f(x)\mathbf{d}x$

The real integral

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x \tag{12.7}$$

is defined to be

$$\lim_{x_1\to-\infty,\, x_2\to\infty} \int_{x_1}^{x_2} f(x)\mathrm{d}x \tag{12.8}$$

provided the limit exists. The techniques we are about to discuss allow the calculation of

$$\lim_{R\to\infty} \int_{-R}^{R} f(x)\mathrm{d}x \tag{12.9}$$

which is known as the *Cauchy principal value* of the integral, denoted by

$$\mathrm{P}\int_{-\infty}^{\infty} f(x)\mathrm{d}x$$

If (12.8) exists then so does (12.9) and the two are equal. But the Cauchy principal value may exist when (12.8) does not. For example

$$\int_{-R}^{R} x\,\mathrm{d}x = [x^2/2]_{-R}^{R} = 0$$

so that

$$\mathrm{P}\int_{-R}^{R} x\,\mathrm{d}x = 0$$

But clearly (12.8) does not exist when $f(x) = x$.

It follows that when we use this technique below, we must take into account the convergence of (12.8). This, in part, leads to condition (ii) of the following:

PROPOSITION 12.6. *Suppose that:*

(i) *A function* f *is differentiable in the upper half-plane* $\{z \in \mathbb{C} : \operatorname{im} z \geq 0\}$ *except for a finite number of poles, none of which lies on the real axis.*
(ii) *If* $S_R(t) = R\mathrm{e}^{\mathrm{i}t}$ $(t \in [0, 2\pi])$ *then there is a constant* A *such that for large* R

$$|f(z)| \leq A/R^2$$

when z *lies on* S_R.

Then

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 2\pi\mathrm{i}\,\Sigma$$

where Σ *is the sum of the residues of the poles of* f *in the upper half-plane.*

Proof. Choose R large enough for (ii) to hold, and so that all the poles lie inside $S_R + [-R, R]$, Figure 12.4.

By Cauchy's Theorem,

$$\int_{-R}^{R} f(x)\,\mathrm{d}x + \int_{S_R} f(z)\mathrm{d}z = 2\pi\mathrm{i}\,\Sigma$$

with Σ as stated. Now let $R \to \infty$. Then

$$\left|\int_{S_R} f(z)\mathrm{d}z\right| \leq \frac{A}{R^2}\pi R = \frac{\pi A}{R}$$

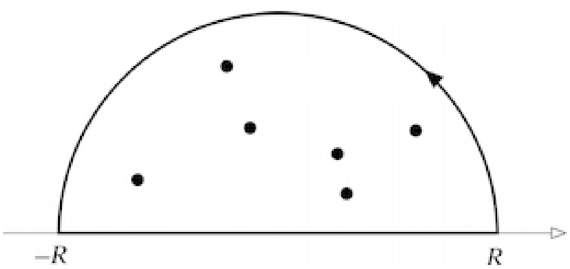

Figure 12.4 Choice of R in proof of Proposition 12.6.

which tends to 0 as $R \to \infty$. Hence

$$\lim_{R\to\infty} \int_{-R}^{R} f(x)\,\mathrm{d}x = 2\pi\mathrm{i}\,\Sigma$$

That is,

$$\mathrm{P}\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 2\pi\mathrm{i}\,\Sigma$$

Now, (ii) tells us that $|f(x)| \le A/x^2$ for large $|x|$, so it follows from real analysis that

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x$$

exists. It is therefore equal to its Cauchy principal value, hence to $2\pi\mathrm{i}\,\Sigma$. □

REMARK 12.7. Condition (ii) is certainly satisfied if $f(z) = p(z)/q(z)$ where p and q are polynomials such that q has no real zeros and the degree of q is greater than or equal to the degree of p plus 2. □

For example, consider

$$\int_{-\infty}^{\infty} \frac{\mathrm{d}x}{(x^2+a^2)(x^2+b^2)}$$

where $a, b > 0$ are real, $a \ne b$. By Remark 12.7 this satisfies (ii), and (i) is obviously true. The only poles of $1/((z^2+a^2)(z^2+b^2))$ in the upper half-plane are simple poles at $\mathrm{i}a, \mathrm{i}b$. The residue at $\mathrm{i}a$ is

$$\lim_{z\to\mathrm{i}a} \frac{z-\mathrm{i}a}{(z^2+a^2)(z^2+b^2)} = \frac{1}{2\mathrm{i}a(b^2-a^2)}$$

and similarly that at $\mathrm{i}b$ is

$$\frac{1}{2\mathrm{i}b(a^2-b^2)}$$

Therefore the value of the integral is

$$2\pi\mathrm{i}\left(\frac{1}{2\mathrm{i}a(b^2-a^2)} + \frac{1}{2\mathrm{i}b(a^2-b^2)}\right) = \frac{\pi}{ab(a+b)}$$

REMARK 12.8. In the proof of Proposition 12.6 we do not require $f(z)$ to be real for z on the real axis. □

For instance, let

$$f(z) = \frac{\mathrm{e}^{\mathrm{i}z}}{(z^2+a^2)(z^2+b^2)}$$

Then on S_R we have $|\mathrm{e}^{\mathrm{i}z}| = |\mathrm{e}^{-y+\mathrm{i}x}| = \mathrm{e}^{-y} \le 1$ for $y \ge 0$, so (ii) holds. As before, there are simple poles at $\mathrm{i}a, \mathrm{i}b$, but now the residues are

$$\frac{\mathrm{e}^{-a}}{2\mathrm{i}a(b^2-a^2)} \quad \text{and} \quad \frac{\mathrm{e}^{-b}}{2\mathrm{i}b(a^2-b^2)}$$

Hence

$$\int_{-\infty}^{\infty} \frac{\mathrm{e}^{\mathrm{i}x}}{(x^2+a^2)(x^2+b^2)} \mathrm{d}x = \pi \left(\frac{e^{-a}}{a(b^2-a^2)} + \frac{e^{-b}}{b(a^2-b^2)} \right)$$

so, equating real and imaginary parts,

$$\int_{-\infty}^{\infty} \frac{\cos x}{(x^2+a^2)(x^2+b^2)} \mathrm{d}x = \frac{\pi}{b^2-a^2} \left(\frac{e^{-a}}{a} - \frac{e^{-b}}{b} \right)$$
$$\int_{-\infty}^{\infty} \frac{\sin x}{(x^2+a^2)(x^2+b^2)} \mathrm{d}x = 0$$

Of these, the second equality is actually obvious (why?), but the first is not.

We can try to generalise this method in (at least) two ways: by getting a better estimate for $\int_{S_R} f$ or by allowing f to have poles on the real axis. The first is dealt with in (III), the second in (IV).

III: $\int_{-\infty}^{\infty} f(x)\mathrm{e}^{\mathrm{i}x}\mathrm{d}x$

PROPOSITION 12.9. *Suppose that:*

(i) *A function f is differentiable in a domain containing the upper half-plane* $\{z \in \mathbb{C} : \operatorname{im} z \geq 0\}$ *except for a finite number of poles, none on the real axis.*
(ii) *There is a constant A such that for large R*

$$|f(z)| \leq A/R \quad \text{for } |z| = R$$

when z lies on S_R. Then

$$\int_{-\infty}^{\infty} f(x)\mathrm{e}^{\mathrm{i}x}\, \mathrm{d}x = 2\pi \mathrm{i}\, \Sigma$$

where Σ is the sum of the residues of the poles of $f(z)\mathrm{e}^{\mathrm{i}z}$ in the upper half-plane.

Proof. It is possible to use the same contour as in (II) and prove that

$$\lim_{R\to\infty} \int_{S_R} f(z)\mathrm{e}^{\mathrm{i}z}\, \mathrm{d}z = 0$$

but this calculates only the Cauchy principal value. Moreover, the convergence problem in this case is much harder because all we know is that f behaves like $1/|x|$ for large $|x|$, which on its own does not imply the existence of

$$\int_{-\infty}^{\infty} f(x)\mathrm{e}^{\mathrm{i}x}\mathrm{d}x$$

More delicate arguments can overcome this obstacle, but it is easier to sidestep the whole question by using a different contour, as in Figure 12.5.

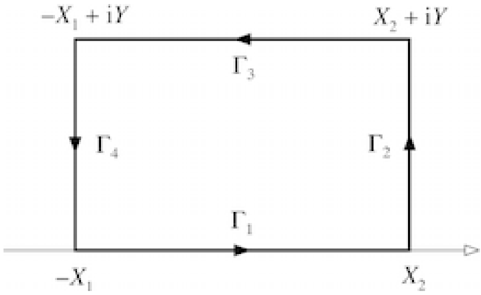

Figure 12.5 Contour for the proof of Proposition 12.9.

We prove that as $X_1, X_2, Y \to \infty$, each of

$$\int_{\Gamma_r} f(z)e^{iz}\,dz$$

for $r = 2, 3, 4$ tends to zero. It then follows as before that

$$\lim_{X_1, X_2 \to \infty} \int_{-X_1}^{X_2} f(x)e^{ix}dx = 2\pi i\,\Sigma$$

as required.

Now we carry out the necessary estimates,

$$\begin{aligned}\left|\int_{\Gamma_2} f(z)e^{iz}\,dz\right| &= \left|\int_0^Y f(X_2 + it)e^{iX_2 - t}\,idt\right| \\ &\le \int_0^Y \frac{A}{X_2}e^{-t}dt \\ &\le \frac{A}{X_2}\end{aligned}$$

for X_2 large. Similarly

$$\left|\int_{\Gamma_4} f(z)e^{iz}\,dz\right| \le \frac{A}{X_1}$$

Also

$$\begin{aligned}\left|\int_{\Gamma_3} f(z)e^{iz}\,dz\right| &= \left|-\int_{-X_1}^{X_2} f(t + iY)e^{it - Y}\,dt\right| \\ &\le \left|\int_{-X_1}^{X_2} \frac{A}{Y}e^{-Y}dt\right| \\ &= AY^{-1}e^{-Y}(X_1 + X_2)\end{aligned}$$

For fixed X_1, X_2 this tends to 0 as $Y \to \infty$. Now let $X_1, X_2 \to \infty$. □

As an example, take

$$\int_{-\infty}^{\infty} \frac{x^3e^{ix}}{(x^2 + a^2)(x^2 + b^2)}dx \quad (a, b > 0, a \ne b)$$

which does not satisfy the conditions for (II), but does for (III). The integrand has simple poles in the upper half-plane at ia and ib. Calculating the residues in the usual way, applying Proposition 12.9, and equating imaginary parts yields

$$\int_{-\infty}^{\infty} \frac{x^3 \sin x}{(x^2+a^2)(x^2+b^2)} \mathrm{d}x = \frac{\pi}{b^2-a^2}(b^2 \mathrm{e}^{-b} - a^2 \mathrm{e}^{-a})$$

(Equating real parts proves that the corresponding cosine integral is zero, but again this is obvious on other grounds.)

IV: Poles on the real axis.

If $f(z)$ has poles on the real axis we make 'indentations' in the contour by drawing small semicircles as in Figure 12.6. Suppose these semicircles have radii $\varepsilon_1, \varepsilon_2, \ldots$. Then we proceed as above, but letting $\varepsilon_1, \varepsilon_2, \ldots$ tend to 0 at the same time as $R, X_1, X_2, Y \to 0$.

There is a problem here similar to that in case (II): all we calculate is a Cauchy principal value

$$\mathrm{P}\int_a^b f(x)\mathrm{d}x = \lim_{\varepsilon \to 0}\left(\int_a^{x_0-\varepsilon} f(x)\mathrm{d}x + \int_{x_0+\varepsilon}^b f(x)\mathrm{d}x\right)$$

for a pole at x_0. This again may exist even if

$$\int_a^b f(x)\mathrm{d}x = \lim_{\varepsilon_1 \to 0}\int_a^{x_0-\varepsilon} f(x)\mathrm{d}x + \lim_{\varepsilon_2 \to 0}\int_{x_0+\varepsilon}^b f(x)\mathrm{d}x$$

does not. Thus

$$\int_{-1}^{1} \frac{\mathrm{d}x}{x}$$

does not exist, but

$$\mathrm{P}\int_{-1}^{1} \frac{\mathrm{d}x}{x} = 0$$

Consequently there is a convergence problem, once the Cauchy principal value has been obtained. Apart from this consideration, the hypotheses and conclusions of Propositions 12.6 for case (II) and 12.9 for case (III) remain valid even when poles occur on the real axis, provided that Σ is summed only over the non-real poles in the upper half-plane.

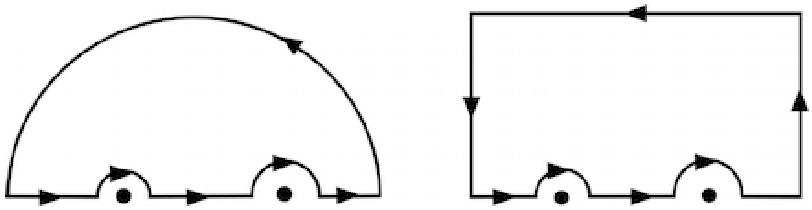

Figure 12.6 Types of contour used when there are poles on the real axis.

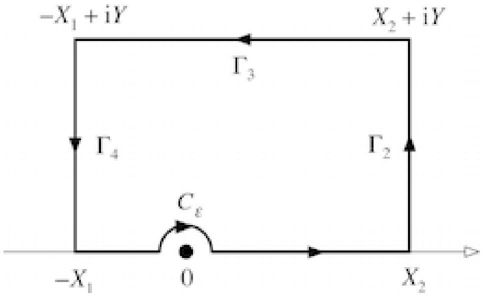

Figure 12.7 Contour used when there are poles on the real axis.

Rather than stating a cumbersome general theorem, we content ourselves with a typical example:

$$\int_{-\infty}^{\infty} \frac{e^{ix}}{x} dx$$

There is a real pole at 0, so we take a contour as in Figure 12.7.

As before, the integrals along Γ_2, Γ_3, and Γ_4 tend to zero. Since there are no poles inside the contour,

$$\lim_{X_1, X_2 \to \infty} \left[\int_{-X_1}^{-\varepsilon} \frac{e^{ix}}{x} dx + \int_{\varepsilon}^{X_2} \frac{e^{ix}}{x} dx + \int_{C_\varepsilon} \frac{e^{iz}}{z} dz \right] = 0$$

But

$$\frac{e^{iz}}{z} = \frac{1}{z} + \sum_{n=1}^{\infty} \frac{i^n z^{n-1}}{n!} = \frac{1}{z} + \phi(z)$$

where ϕ is differentiable. Hence $|\phi(z)| \leq M$ in a neighbourhood of 0, so

$$\begin{aligned} \lim_{\varepsilon \to 0} \int_{C_\varepsilon} \frac{e^{iz}}{z} dz &= \lim_{\varepsilon \to 0} \int_{C_\varepsilon} \left(\frac{1}{z} + \phi(z) \right) dz \\ &= \lim_{\varepsilon \to 0} \left(- \int_0^{\pi} \frac{1}{\varepsilon e^{it}} i\varepsilon e^{it} dt \right) + \lim_{\varepsilon \to 0} \int_{C_\varepsilon} \phi(z) dz \\ &= i\pi \end{aligned}$$

using the Estimation Lemma. Therefore

$$P \int_{-\infty}^{\infty} \frac{e^{ix}}{x} dx = i\pi$$

Equating real and imaginary parts,

$$P \int_{-\infty}^{\infty} \frac{\cos x}{x} dx = 0$$

$$P \int_{-\infty}^{\infty} \frac{\sin x}{x} dx = \pi$$

The first integral exists only as a principal value, since $\cos(x)/x$ behaves like $1/x$ for small $|x|$. But for the second we have

$$\begin{aligned} P\int_{-\infty}^{\infty} \frac{\sin x}{x}\mathrm{d}x &= \lim_{\varepsilon\to 0}\left(\int_{-\infty}^{-\varepsilon} \frac{\sin x}{x}\mathrm{d}x + \int_{\varepsilon}^{\infty} \frac{\sin x}{x}\mathrm{d}x\right) \\ &= 2\lim_{\varepsilon\to 0}\int_{\varepsilon}^{\infty} \frac{\sin x}{x}\mathrm{d}x \end{aligned}$$

This limit exists, hence by definition it equals

$$2\int_{0}^{\infty} \frac{\sin x}{x}\mathrm{d}x$$

because $\sin(x)/x \to 1$ as $x \to 0$. Hence we can remove the P from the second expression above. Further, we have also proved that

$$\int_{0}^{\infty} \frac{\sin x}{x}\mathrm{d}x = \frac{\pi}{2}$$

V: $\displaystyle\int_{-\infty}^{\infty} \frac{\mathrm{e}^{ax}}{\phi(\mathrm{e}^{x})}\mathrm{d}x$

For integrals of this type we integrate around a contour like that in Figure 12.8.

In this case, on Γ_3, if we substitute $z = -t + 2\pi\mathrm{i}$, then $-X_2 \le t \le X_1$, so that

$$\begin{aligned} \int_{\Gamma_3} \frac{\mathrm{e}^{az}}{\phi(\mathrm{e}^{z})}\mathrm{d}z &= \int_{-X_2}^{X_1} \frac{\mathrm{e}^{-at+2\pi\mathrm{i}a}}{\phi(\mathrm{e}^{-t+2\pi\mathrm{i}})}(-1)\mathrm{d}t \\ &= \mathrm{e}^{2\pi\mathrm{i}a}\int_{X_1}^{-X_2} \frac{\mathrm{e}^{-at}}{\phi(\mathrm{e}^{-t})}\mathrm{d}t \end{aligned}$$

Putting $t = -x$, this equals

$$-\mathrm{e}^{2\pi\mathrm{i}a}\int_{-X_1}^{X_2} \frac{\mathrm{e}^{ax}}{\phi(\mathrm{e}^{x})}\mathrm{d}x$$

If ϕ is such that

$$\int_{\Gamma_r} \frac{\mathrm{e}^{az}}{\phi(\mathrm{e}^{z})}\mathrm{d}z \to 0 \quad \text{as} \quad X_1, X_2 \to \infty \ \ (r = 2, 4)$$

then we obtain

$$(1 - \mathrm{e}^{2\pi a i}\int_{-\infty}^{\infty} \frac{\mathrm{e}^{ax}}{\phi(\mathrm{e}^{x})}\mathrm{d}x = 2\pi\Sigma$$

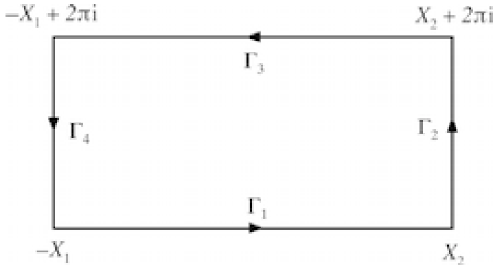

Figure 12.8 Contour for integrals of type (V).

where Σ is the sum of the residues of $e^{az/\phi(z)}$ at poles between the lines $\mathrm{im}\, z = 0$, $\mathrm{im}\, z = 2\pi$.

For example, consider

$$\int_{-\infty}^{\infty} \frac{e^{ax}}{e^{2x}+1} dx \quad (0 < a < 1)$$

The relevant singularities are at $i\pi/2, 3i\pi/2$, with corresponding residues

$$-\tfrac{1}{2}e^{i\pi a/2} \qquad -\tfrac{1}{2}e^{3i\pi a/2}$$

Hence the value of the integral is

$$\frac{2\pi i}{1-e^{2\pi i a}}(-\tfrac{1}{2}e^{i\pi a/2} - \tfrac{1}{2}e^{3i\pi a/2})$$

Putting $k = e^{i\pi a/2}$ this is easily seen to be equal to

$$\frac{\pi}{2\sin(\pi a/2)}$$

VI: Short cuts.

You should always be on the lookout for quick methods, other than those given above. We give two examples, each of independent interest.

Example 12.10. We use contour integration to prove that the integral of the 'bell curve' or 'gaussian', which is of importance in probability theory, is

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

This result is usually obtained by squaring the integral, rewriting it as

$$\left(\int_{-\infty}^{\infty} e^{-x^2} dx\right)\left(\int_{-\infty}^{\infty} e^{-y^2} dy\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)} dx dy$$

and then changing to polar coordinates. For a long time it was thought that no evaluation by contour integration existed, but several such proofs are now known. We give one of them, from Remmert [15]. Consider

$$f(z) = \frac{e^{-z^2}}{1+e^{-2az}}$$

with

$$a = (1+i)\sqrt{\frac{\pi}{2}}$$

Then $a^2 = i\pi$, so

$$f(z) - f(z+a) = e^{-z^2} \tag{12.10}$$

The poles of f are simple and located at the points $-\frac{1}{2}a + na$ with $n \in \mathbb{Z}$.

Now integrate f over the rectangle with corners $-r, s, s + i\,\mathrm{im}(a), -r + i\,\mathrm{im}(a)$ with $r, s > 0$. The only pole of f inside this rectangle is at $\frac{a}{2}$, and its residue is

$$\operatorname{res}(f, a/2) = \frac{e^{-\frac{1}{4}a^2}}{-2ae^{-a^2}} = \frac{-i}{2\sqrt{\pi}}$$

This follows from Lemma 12.4, which states that

$$\operatorname{res}\left(\frac{g}{h}, x\right) = \frac{g(x)}{h'(x)}$$

if h has a simple zero at x and $g(x) \neq 0$.

Using the residue theorem and (12.10), it is easy to see that

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \lim_{r,s \to \infty} \int_{-r}^{s} e^{-x^2} dx = 2\pi i \operatorname{res}(f, a/2) = \sqrt{\pi}$$

since the integrals along the vertical sides of the rectangles converge to zero as $r, s \to \infty$. □

Example 12.11. An integral of great importance in applied mathematics is

$$\int_{-\infty}^{\infty} e^{i\lambda x} e^{-x^2} dx \quad (\lambda \in \mathbb{R})$$

The function e^{-z^2} is differentiable throughout $\mathbb{C}$. If γ is the rectangle with vertices at $-R, R, R - i\lambda/2, -R - i\lambda/2$, then

$$\int_{\gamma} e^{-z^2} dz = 0$$

As $R \to \infty$, the integrals over the vertical edges of the rectangle tend to zero. The other two converge. Hence

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \int_{-\infty}^{\infty} e^{-(x - i\lambda/2)^2} dx$$
$$= \int_{-\infty}^{\infty} e^{-x^2} e^{i\lambda x} e^{\lambda^2/4} dx$$

Therefore

$$\int_{-\infty}^{\infty} e^{i\lambda x} e^{-x^2} dx = e^{\lambda^2/4} \int_{-\infty}^{\infty} e^{-x^2} dx$$

The latter integral is a constant; it has just been evaluated in Example 12.10 and equals $\sqrt{\pi}$. So

$$\int_{-\infty}^{\infty} e^{i\lambda x} e^{-x^2} dx = e^{\lambda^2/4}\sqrt{\pi}$$

(Although we can derive Example 12.10 from this by setting $\lambda = 0$, we need Example 12.10 to get the result.) □

12.4 Summation of Series

Residues can also be used to find explicit expressions for sums of suitable infinite series, based on the idea that the functions $\cot \pi z$ and $\operatorname{cosec} \pi z$ have simple poles at

$z = n \in \mathbb{Z}$. If f is a function that is differentiable at all $z = n \in \mathbb{Z}$ then it is easy to check that

$$\operatorname{res}(f(z)\cot \pi z, n) = \frac{f(n)}{\pi}$$
$$\operatorname{res}(f(z)\operatorname{cosec} \pi z, n) = \frac{(-1)^n f(n)}{\pi}$$

This suggests a method for summing certain series, as follows. Let C_N be the square whose vertices are

$$(N + \tfrac{1}{2})(\pm 1 \pm \mathrm{i})$$

parametrised, as usual, in the anticlockwise direction, Figure 12.9. We claim that $\cot \pi z$ and $\operatorname{cosec} \pi z$ are bounded on C_N, where the bound is *independent* of N.

To prove this, first note that on the two horizontal sides $z = x + \mathrm{i}y$, where $|y| \geq \frac{1}{2}$. Now

$$\begin{aligned}
|\cot \pi z| &= \left| \frac{\mathrm{e}^{\mathrm{i}\pi z} + \mathrm{e}^{-\mathrm{i}\pi z}}{\mathrm{e}^{\mathrm{i}\pi z} - \mathrm{e}^{-\mathrm{i}\pi z}} \right| \\
&\leq \left| \frac{|\mathrm{e}^{\mathrm{i}\pi z}| + |\mathrm{e}^{-\mathrm{i}\pi z}|}{|\mathrm{e}^{\mathrm{i}\pi z}| - |\mathrm{e}^{-\mathrm{i}\pi z}|} \right| \\
&= \left| \frac{\mathrm{e}^{-\pi y} + \mathrm{e}^{\pi y}}{\mathrm{e}^{-\pi y} - \mathrm{e}^{\pi y}} \right| \\
&= \coth |\pi y| \\
&\leq \coth \pi/2
\end{aligned}$$

Also

$$\begin{aligned}
|\operatorname{cosec} \pi z| &= (\tfrac{1}{2}|\mathrm{e}^{\mathrm{i}\pi z} - \mathrm{e}^{-\mathrm{i}\pi z}|)^{-1} \\
&\leq (\tfrac{1}{2}|\mathrm{e}^{-\pi y} - \mathrm{e}^{\pi y}|)^{-1} \\
&= (\sinh |\pi y|)^{-1} \\
&\leq (\sinh \pi/2)^{-1}
\end{aligned}$$

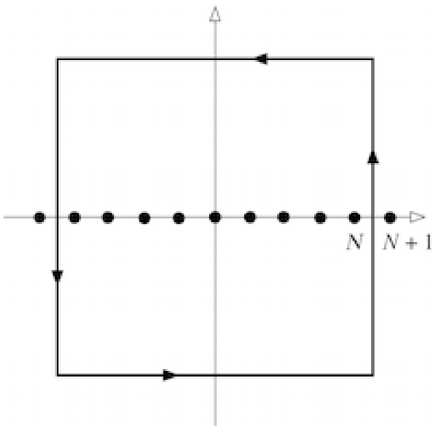

Figure 12.9 Square contour used to sum series.

On the other two sides of the square, $z = \pm(N + \frac{1}{2}) + \mathrm{i}t$, so

$$\begin{aligned}|\cot \pi z| &= |\tanh \pi t| \\ &= \left|\frac{1 - \mathrm{e}^{-2\pi t}}{1 + \mathrm{e}^{-2\pi t}}\right| \\ &\leq 1\end{aligned}$$

and

$$\begin{aligned}|\operatorname{cosec} \pi z| &= |\sin \pi z|^{-1} \\ &= |\cos(\mathrm{i}\pi t)|^{-1} \\ &= (\cosh |\pi t|)^{-1} \\ &\leq 1\end{aligned}$$

Hence there is a constant M such that $|\cot \pi z| \leq M$ and $|\operatorname{cosec} \pi z| \leq M$ for any z on C_N.

Suppose now that for large enough $|z|$ we have

$$|f(z)| \leq \frac{A}{|z|^2}$$

Then we claim that

$$\Sigma = 0$$

where Σ is the sum of the residues of $f(z) \cot \pi z$. By Cauchy's Residue Theorem,

$$\int_{C_N} f(z) \cot \pi z \,\mathrm{d}z = 2\pi \mathrm{i}\, \Sigma_N$$

where Σ_N is the sum of the residues of $f(z) \cot \pi z$ inside C_N. As $N \to 0$, clearly $\Sigma_N \to \Sigma$, so it is sufficient to prove that the integral tends to 0. But

$$\left|\int_{C_N} f(z) \cot \pi z \,\mathrm{d}z\right| \leq \frac{A}{N^2} M(8N + 4)$$

for large enough N by the Estimation Lemma. As $N \to 0$, this tends to 0 as claimed.

Usually, Σ is an infinite series, and since it is zero we can sum certain related series. This is best illustrated by an example, and the obvious one to try is $f(z) = 1/z^2$.

At an integer $n \neq 0$ the function $z^{-2} \cot \pi z$ has a simple pole with residue $1/(n^2\pi)$, whereas at the origin it has a triple pole with residue $-\pi/3$. Hence

$$\begin{aligned}0 = \Sigma &= -\frac{\pi}{3} + \sum_{n=-\infty}^{-1} \frac{1}{n^2\pi} + \sum_{n=1}^{\infty} \frac{1}{n^2\pi} \\ &= -\frac{\pi}{3} + \frac{2}{\pi} \sum_{n=1}^{\infty} \frac{1}{n^2}\end{aligned}$$

Hence

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$$

which is a theorem originally proved by Euler, using a different (non-rigorous) method.

If we use cosec πz instead of cot πz a similar result applies, allowing us to sum series of the form $\sum(-1)^n f(n)$. For instance, using $f(z) = 1/z^2$ and arguing much as above, we can prove that

$$\sum_{n=1}^{\infty}(-1)^{n+1}\frac{1}{n^2} = \frac{\pi^2}{12}$$

12.5 Counting Zeros

A rather different use for residues is to calculate the number of zeros of a function that is differentiable inside a simple loop. We begin with:

THEOREM 12.12. *Suppose that f is differentiable, apart from a finite set of poles, in a domain D containing a simple loop γ and all points inside γ. If f has no zeros or poles on γ then*

$$\frac{1}{2\pi \mathrm{i}}\int_{\gamma}\frac{f'(z)}{f(z)}\mathrm{d}z = N - P$$

where N is the number of zeros of f inside γ and P is the number of poles of f inside γ, each counted according to multiplicity.

Proof. By Cauchy's Residue Theorem, the integral is the sum of the residues of $f'(z)/f(z)$ at its poles inside γ. If z_0 is neither a zero nor a pole of f, then f'/f is differentiable at z_0. We show that:

(i) If f has a zero of order k at z_1 then f'/f has a pole with residue k.
(ii) If f has a pole of order m at z_2 then f'/f has a pole with residue $-m$.

To prove (i) note that in that case

$$f(z) = (z - z_1)^k\phi(z)$$

where $\phi(z_1) \neq 0$ and ϕ is differentiable in a neighbourhood of z_1. Therefore

$$f'(z) = k(z - z_1)^{k-1}\phi(z) + (z - z_1)^k\phi'(z)$$

so

$$\frac{f'(z)}{f(z)} = \frac{k}{z - z_1} + \frac{\phi'(z)}{\phi(z)}$$

This has a simple pole at z_1 with residue k, because ϕ'/ϕ is differentiable at z_1. This proves (i).

Similarly in case (ii)

$$f(z) = \frac{\psi(z)}{(z - z_2)^m}$$

where $\psi(z_2) \neq 0$ and ϕ is differentiable in a neighbourhood of z_2. Therefore

$$f'(z) = \frac{-m\psi(z)}{(z - z_2)^{m+1}} + \frac{\psi'(z)}{(z - z_2)^m}$$

so

$$\frac{f'(z)}{f(z)} = \frac{-m}{z - z_2} + \frac{\psi'(z)}{\psi(z)}$$

which has a simple pole at z_2 with residue $-m$.

Now sum over all poles of f'/f. □

COROLLARY 12.13. *Let γ be a simple loop in a domain D such that all points inside γ are in D. If f is differentiable in D and has no zeros on γ then the number of zeros of f inside γ is*

$$\frac{1}{2\pi \mathrm{i}} \int_\gamma \frac{f'(z)}{f(z)} \mathrm{d}z$$ □

From this we deduce another important theorem:

THEOREM 12.14 (Rouché's Theorem). *Suppose that f and g are differentiable in a domain D that contains a simple loop γ and all points inside γ. If*

$$|f(z) - g(z)| < |f(z)| \tag{12.11}$$

for all $z = \gamma(t)$ $(t \in [a, b])$ then f and g have the same number of zeros inside γ.

Proof. Let $F(z) = g(z)/f(z)$. By (12.11),

$$|1 - F(\gamma(t))| < 1 \quad (t \in [a, b]) \tag{12.12}$$

Inside D,

$$F \text{ has a zero} \iff g \text{ has a zero}$$
$$F \text{ has a pole} \iff f \text{ has a zero}$$

Thus by Theorem 12.12 it is enough to prove that

$$\int_\gamma \frac{F'(z)}{F(z)} \mathrm{d}z = 0$$

To do so, let $\Gamma(t) = F(\gamma(t))$. By (12.12),

$$|1 - \Gamma(t)| < 1 \quad (t \in [a, b])$$

so Γ lies inside the circle centre 1 radius 1, Figure 12.10. Now

$$\begin{aligned}
\int_\gamma \frac{F'(z)}{F(z)} \mathrm{d}z &= \int_a^b \frac{F'(\gamma(t))}{F(\gamma(t))} \mathrm{d}t \\
&= \int_a^b \frac{\Gamma(t)}{\Gamma(t))} \mathrm{d}t \\
&= \int_\Gamma \frac{\mathrm{d}z}{z} \\
&= 2\pi \mathrm{i} w(\Gamma, 0) \\
&= 0
\end{aligned}$$

from Figure 12.10. □

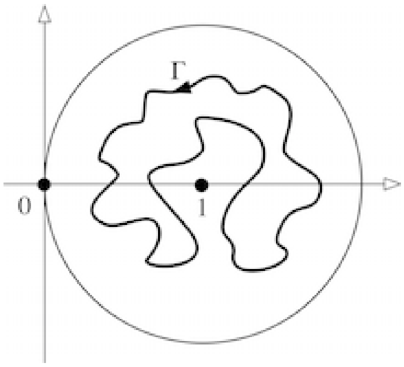

Figure 12.10 Path Γ in proof of Rouché's Theorem.

As an example of Rouché's Theorem in action, we give a second proof of the Fundamental Theorem of Algebra, Theorem 10.9. Let

$$\begin{aligned} g(z) &= z^m + a_1 z^{m-1} + \cdots + a_m \\ f(z) &= z^m \end{aligned}$$

Then

$$|f(z) - g(z)| = |a_1 z^{m-1} + \cdots + a_m|$$

and for $z \neq 0$

$$\left|\frac{1}{z^m}\right| |f(z) - g(z)| = \left|\frac{a_1}{z} + \cdots + \frac{a_m}{z^m}\right|$$

The right-hand side tends to 0 as $|z| \to \infty$, so there exists $R > 0$ such that if $|z| > R$ then

$$\left|\frac{a_1}{z} + \cdots + \frac{a_m}{z^m}\right| < 1$$

and then

$$|f(z) - g(z)| < |z^m| = f(z)$$

Therefore by Rouché's Theorem f and g have the same number of zeros inside $\{z \in \mathbb{C} : |z| < R\}$. Since f has m zeros (counting multiplicities) so does g. But g cannot have more than m zeros, so every polynomial of degree m over $\mathbb{C}$ has exactly m zeros, counting multiplicities. This is the Fundamental Theorem of Algebra.

12.6 Exercises

1. Find the residue of f at z_0 in the following cases:

(i) $f(z) = z^{-3} \sin z \ (z \neq 0),\ z_0 = 0$
(ii) $f(z) = e^z z^{-n-1} \ (z \neq 0),\ z_0 = 0$
(iii) $f(z) = \exp(1/z) \ (z \neq 0),\ z_0 = 0$

(iv) $f(z) = z^2(z^2 + a^2)^{-3}$ $(z \neq \pm \mathrm{i}a)$, $z_0 = \mathrm{i}a, -\mathrm{i}a$, where $a \in \mathbb{R}$
(v) $f(z) = (1 + z^2 + z^4)^{-1}$ $(z \neq \exp(r\pi\mathrm{i}/3),\ r = 1, 2, 4, 5)$, $z_0 = \exp(\pi\mathrm{i}/3)$

2. Find the residue of the given function at each of its isolated singular points, including infinity (provided this is also isolated – that is, not the limit of a sequence of finite singularities).

(i) $1/(z^3 - z^5)$
(ii) $\mathrm{e}^z/(z^2(z^2 + 5))$
(iii) $\cot^3 z$
(iv) $(\sin z^{-1})^{-1}$
(v) $(z \cos z^{-2})^{-1}$

3. If f has a pole of order 2 at z_0, show that the residue of f at z_0 is $h'(z_0)$, where $h(z) = (z - z_0)^2 f(z)$.

4. Let $\gamma(t) = \mathrm{e}^{\mathrm{i}t}$, $(t \in [0, 2\pi])$. Find, by residues, the value of

$$\int_\gamma \frac{\mathrm{d}z}{z^2 - 2az + 1} \quad (a > 1)$$

Hence calculate

$$\int_0^{2\pi} \frac{\mathrm{d}t}{a - \cos t}$$

What happens if $a < -1$? If $-1 \leq a \leq 1$?

5. Verify the following:

(i) $\displaystyle\int_0^\pi \frac{\mathrm{d}t}{1 + b\cos^2 t} = \frac{\pi}{\sqrt{b+1}} \quad (b > -1)$
(ii) $\displaystyle\int_0^\infty \frac{\mathrm{d}x}{1 + x^4} = \frac{\pi}{2\sqrt{2}} \quad (b > -1)$
(iii) $\displaystyle\int_0^\pi \frac{\log x}{1 + x^2}\mathrm{d}x = 0$
(iv) $\displaystyle\int_{-\infty}^\infty \frac{x \sin \pi x}{1 - x^2}\mathrm{d}x = \pi$
(v) $\displaystyle\int_{-\infty}^\infty \frac{(\log x)^2}{1 + x^2}\mathrm{d}x = \frac{\pi^3}{8}$

6. Show that

$$\int_{-\infty}^\infty \frac{(10x)^2}{(x^2 + 4)^2(x^2 + 9)^2}\mathrm{d}x = \pi$$

7. Show that

$$\int_0^\infty \frac{\cos 5x}{x^4 + a^4}\mathrm{d}x = \frac{\pi}{2a^3}\mathrm{e}^{-5a/\sqrt{2}} \sin\left(\frac{5a}{\sqrt{2}} + \frac{\pi}{4}\right) \quad (a > 0)$$

8. Evaluate:

(i) $\int_0^{2\pi} \cos^4 t + \sin^4 t \, dt$

(ii) $\int_0^{2\pi} \sin^3 t \cdot \cos t + \cos^3 t \cdot \sin t \, dt$

(iii) $\int_0^{2\pi} 2\cos^3 t + 3\cos^2 t \, dt$

9. If $n \in \mathbb{Z}, n > 0$, prove:

(i) $\int_0^{2\pi} \exp(\cos t)\cos(nt - \sin t)\, dt = 2\pi/n!$

(ii) $\int_0^{2\pi} \exp(\cos t)\sin(nt - \sin t)\, dt = 0$

10. Evaluate, by integrating suitable functions round a semicircle:

(i) $\int_0^{\infty} \frac{dx}{1 + x^2 + x^4}$

(ii) $\int_0^{\infty} \frac{\cos mx}{x^2 + a^2} dx \; (a, m > 0)$

11. Prove that

$$\int_0^{\infty} \frac{x^2}{(x^2 + a^2)^3} dx = \frac{\pi}{16a^3} \; (a > 0)$$

12. By integrating round a rectangle whose vertices lie at $R, R + i, -R + i, -R$, and letting $R \to \infty$, show that

$$\int_{-\infty}^{\infty} \frac{\cosh(cx)}{\cosh(\pi x)} dx = \sec(c/2) \quad (c \in [-\pi, \pi])$$

13. Prove that

$$\int_0^{\infty} t^{a-1}(t + 1)^{-1}\, dt = \frac{\pi}{\sin \pi a} \quad (a \in [0, 1])$$

by making the substitution $t = e^x$ and integrating $e^{az}(e^z + 1)^{-1}$ around the rectangle with vertices $\pm R, \pm R + 2\pi i$.

14. *Inversion Formula for the Laplace Transform*. Suppose that F is differentiable in $\mathbb{C}$ except for a finite number of poles, of which $z_1, \dots, z_n$ satisfy $\operatorname{re} z < a$ and none lies on the line $\operatorname{re} z = a$. If there exist $M > 0, b > 0, c > 0$ such that $|F(z)| < M/|z|^c$ for $|z| < b$, show that

$$f(t) = \lim_{R\to\infty} \int_{a-iR}^{a+iR} e^{zt}F(z)\, dz = 2\pi i \sum_{r=1}^{n} \operatorname{res}(e^{zt}F(z), z_r)$$

If $F(z) = \alpha(z^2 + \alpha^2)^{-1}$ $(\alpha > 0)$ show that $f(t) = \sin \alpha t$.

15. Use real analysis to prove *Jordan's inequality*:

$$\frac{\sin t}{t} \geq \frac{2}{\pi} \quad (t \in [0, \pi/2])$$

Hence show that if $S_R(t) = Re^{it}$ $(t \in [0, \pi])$ then

$$\lim_{R\to\infty} \int_{S_R} \frac{e^{imz}}{z}\, dz = 0 \quad (m > 0)$$

By integrating e^{imz}/z along the contours $\Gamma_1, C_\varepsilon, \Gamma_2, S_R$ defined by:

$$\begin{aligned} \Gamma_1(t) &= t \quad (t \in [-R, -\varepsilon]) \\ C_\varepsilon(t) &= e^{i(\pi - t)} \quad (t \in [0, \pi]) \\ \Gamma_2(t) &= t \quad (t \in [\varepsilon, R]) \end{aligned}$$

prove *Dirichlet's Discontinuous Factor*:

$$\int_0^\infty \frac{\sin mx}{x}\, dx = \begin{cases} 0 & \text{if } m = 0 \\ \pi/2 & \text{if } m > 0 \\ -\pi/2 & \text{if } m < 0 \end{cases}$$

16. Show that:

(i) $\dfrac{z}{e^z - 1} = 1 - \dfrac{z}{2} + \displaystyle\sum_{n=1}^{\infty} \frac{2z^2}{z^2 + 4n^2\pi^2}$

(ii) $\operatorname{cosec} z = \dfrac{1}{z} + \displaystyle\sum_{n=1}^{\infty} \frac{(-1)^n 2z}{z^2 - n^2\pi^2}$

(iii) $\operatorname{cosec}^2 z = \displaystyle\sum_{-\infty}^{\infty} \frac{1}{(z - n\pi)^2}$

17. Sum the following series, where $a \notin \mathbb{Z}$:

(i) $\displaystyle\sum_{-\infty}^{\infty} (n + a)^{-2}$

(ii) $\displaystyle\sum_{0}^{\infty} (n^2 + a^2)^{-1}$

(iii) $\displaystyle\sum_{0}^{\infty} (2n + 1)^{-2}$

(iv) $\displaystyle\sum_{0}^{\infty} (-1)^n (2n + 1)^{-3}$

18. By integrating

$$\frac{ze^{ibz}}{(a^2 - z^2)\sin \pi z}$$

round a suitable contour, show that

$$\sum_{n=1}^{\infty} (-1)^n \frac{n \sin bn}{a^2 - n^2} = \frac{\pi}{2} \frac{\sin ba}{\sin \pi a} \quad (|b| < \pi)$$

19. By considering $f(z) = 1/(z-\xi) + 1/z$, show that when $\xi \notin \mathbb{Z}$,

$$\pi \cot \pi\xi = \frac{1}{\xi} + \sum_{n=1}^{\infty} \frac{2\xi}{\xi^2 - n^2}$$

20. Using the result of Exercise 19, integrate $\pi \cot \pi x$ along a suitable contour to show that

$$\log \sin \pi z = \log \pi z + \sum_{n=1}^{\infty} \log(1 - z^2/n^2)$$

where the log is chosen to make $\log 1 = 0$ in each term. Taking exponentials, obtain the infinite product expansion of the sine function (defined as the limit of suitable finite partial products, by analogy with infinite sums):

$$\sin \pi z = \pi z \prod_{n=1}^{\infty} (1 - z^2/n^2)$$

21. If $|a| > \mathrm{e}$, use Rouché's Theorem to prove that the equation

$$\mathrm{e}^z = az^n$$

has n roots with $|z| < 1$.

22. Find the number of zeros of the following polynomials that lie inside the unit circle:

(i) $z^9 - 2z^6 + z^2 - 8z - 2$
(ii) $2z^5 - z^3 + 3z^2 - z + 8$
(iii) $z^4 - 5z + 1$

23. How many zeros of $z^4 + 4z^3 + 6z^2 - 4z + 3$ lie inside the disc $|z - 1| < 1$?

24. Prove that however small $\varepsilon > 0$ is chosen, for all large enough n the function

$$1 + z^{-1} + (2!z^2)^{-1} + (3!z^3)^{-1} + \cdots + (n!z^n)^{-1}$$

has all its zeros inside the disc $|z| < \varepsilon$.

25. Let $p(z)$ be a polynomial of degree n, and suppose that $p(z_1) = p(z_2) = 0$ with $z_1 \neq z_2$. Show that there exists a zero of $p'(z)$ within the circle centre $\frac{1}{2}(z_1 + z_2)$ and radius $\frac{1}{2}|z_1 - z_2| \cot(\pi/n)$.

26. *Residues in Reverse*. Show that the residue of $\tan^{p-1} \pi z$ at $z = \frac{1}{2}$ is $(-1)^{p/2}\pi^{-1}$, for integer $p > 0$. (Hint: integrate it round the rectangle with vertices at $-\mathrm{i}R, 1 - \mathrm{i}R, 1 + \mathrm{i}R, \mathrm{i}R$, then let $R \to \infty$ and estimate sizes.)

27. Show that

(i) $\displaystyle\int_0^{\infty} \frac{\log x}{1 + x^2}\,\mathrm{d}x = 0$

(ii) $\displaystyle\int_0^{\infty} \frac{\log x}{(1 + x^2)^2}\,\mathrm{d}x = -\frac{\pi}{4}$

13 Conformal Transformations

In mathematics we often encounter functions that preserve some structure that is of interest. For example, in Euclidean geometry rigid motions preserve lengths, angles, and areas, while changes of scale preserve the shape (but not the size) of geometric figures. Homomorphisms of groups preserve group multiplication. Arithmetic modulo n preserves addition and multiplication. Topological transformations preserve connectedness. Conversely, given an interesting class of functions, we can ask what structure they preserve. This chapter deals with a property preserved by all differentiable (equivalently analytic) complex functions, namely: angles between curves. Functions with this property are called ‘conformal’.

The conformal property can be used in two directions. By studying differentiable functions, we can prove theorems about curves; by studying curves, we can prove theorems about differentiable functions. The second technique is of great importance in the advanced ‘geometric’ theory of differentiable functions, but only the first falls within our present scope. The method has interesting applications to potential theory and fluid dynamics, and we outline the beginning of these. We also consider in moderate detail several special conformal functions; in particular ‘Möbius maps’, often called ‘Möbius transformations’, which have the remarkable property of mapping circles to circles.

13.1 Measurement of Angles

Since we are studying preservation of angles, we must discuss these first, in a manner that is appropriate for this chapter. The use of a real number to measure an angle is not always entirely satisfactory, because the same angle corresponds to many different real numbers. However, if real numbers θ and ϕ represent the same angle, then $\theta - \phi$ is an integer multiple of 2π, and conversely, so the ambiguity is not too great. For many purposes it can be avoided by making some artificial convention, such as the requirement that $-\pi < \arg(z) \leq \pi$. For other purposes it is more convenient to measure angles in a natural and unambiguous way, though one that is less familiar than the real numbers.

13.1.1 Real Numbers Modulo 2π

If $x, y \in \mathbb{R}$ we say that x and y are *congruent modulo* 2π, and write

$$x \equiv y \pmod{2\pi}$$

if there is an integer n such that $x - y = 2n\pi$. Congruence modulo 2π is an equivalence relation, so it partitions $\mathbb{R}$ into mutually disjoint equivalence classes. We denote the set of all such equivalence classes by

$$\mathbb{R}/2\pi$$

For each $x \in \mathbb{R}$ let $p(x)$ be the equivalence class (or *congruence class*) to which x belongs. This defines a function

$$p : \mathbb{R} \to \mathbb{R}/2\pi$$

and

$$p(x) = \{x + 2n\pi \ : n \in \mathbb{Z}\}$$

Given an angle measured by a real number θ, the same angle is measured by all $\theta + 2n\pi$ $(n \in \mathbb{Z})$, and by these real numbers only. Instead of picking one of them, we can represent the angle by the whole collection, namely $p(\theta)$. In other words, the natural measure of an angle is not a real number, but an element of $\mathbb{R}/2\pi$.

For instance, the real numbers

$$\ldots, -11\pi/3, -5\pi/3, \pi/3, 7\pi/3, 13\pi/3, \ldots$$

all represent the same angle, and the set

$$\{\ldots, -11\pi/3, -5\pi/3, \pi/3, 7\pi/3, 13\pi/3, \ldots\}$$

is the unique element of $\mathbb{R}/2\pi$ corresponding to this angle.

13.1.2 Geometry of $\mathbb{R}/2\pi$

Geometrically, $\mathbb{R}/2\pi$ is a circle. To see this, define

$$q : \mathbb{R} \to \mathbb{C} \qquad q(x) = \mathrm{e}^{\mathrm{i}x} \ (x \in \mathbb{R})$$

The image of q is the unit circle in $\mathbb{C}$, which we denote by

$$\mathbb{S} = \{z \in \mathbb{C} : z = \mathrm{e}^{\mathrm{i}x} \ (x \in \mathbb{R})\} = \{z \in \mathbb{C} : |z| = 1\}$$

Since $\mathrm{e}^{2\pi\mathrm{i}} = 1$, Proposition 5.3 shows that $q(x) = q(y)$ if and only if $x \equiv y \pmod{2\pi}$, so

$$q(x) = q(y) \quad \Longleftrightarrow \quad p(x) = p(y) \tag{13.1}$$

Therefore we can define

$$j : \mathbb{R}/2\pi \to \mathbb{S}$$

as follows: if $r \in \mathbb{R}/2\pi$ then $r = p(x)$ for some $x \in \mathbb{R}$, and we set

$$j(r) = q(x)$$

By (13.1) this is independent of the choice of $x \in \mathbb{R}$, and j is a bijection. Thus the elements of $\mathbb{R}/2\pi$ are in natural one-to-one correspondence with the points of a circle.

This bijection preserves continuity in a natural manner. If $X \subseteq \mathbb{C}$ we say that a function $f : X \to \mathbb{R}/2\pi$ is *continuous* if and only if the composite function $j \circ f : X \to \mathbb{S}$ is continuous in the usual sense, considering $\mathbb{S}$ as a subset of $\mathbb{C}$. This accords with the intuitive idea of geometric continuity if we think of $\mathbb{R}/2\pi$ as a circle. (In topological terms, we use j to transfer the usual topology of $\mathbb{S}$ to $\mathbb{R}/2\pi$, defining $U \subseteq \mathbb{R}/2\pi$ to be open if and only if $j(U)$ is open in $\mathbb{S}$. Continuity of maps $f : X \to \mathbb{R}/2\pi$ in this topology is equivalent to what we have just defined.)

13.1.3 Operations on Angles

We can define addition and subtraction of elements of $\mathbb{R}/2\pi$, corresponding to geometric addition and subtraction of angles. Let $r, s \in \mathbb{R}/2\pi$. Pick $x, y \in \mathbb{R}$ such that $p(x) = r, p(y) = s$, and define

$$r + s = p(x + y)$$
$$r - s = p(x - y)$$

As usual we can verify that these definitions do not depend on the choice of x and y, using (12.1).

We can restate the above in group-theoretic terms. The set $G = \{2n\pi : n \in \mathbb{Z}\}$ is a subgroup of the additive group of real numbers, hence a normal subgroup since the latter is abelian. The quotient group $\mathbb{R}/G$ is what we have called $\mathbb{R}/2\pi$. As an analogy, consider the definition of the integers $\mathbb{Z}_n$ modulo n as the quotient group $\mathbb{Z}/H$ where $H = \{kn : k \in \mathbb{Z}\}$.

In $\mathbb{Z}_n$ we can also define multiplication. You should verify that this does *not* work for $\mathbb{R}/2\pi$. It fails because 2π is not an integer. However, we can sensibly multiply an angle in $\mathbb{R}/2\pi$ by an integer, since this reduces to repeated additions or subtractions.

13.1.4 The Argument Modulo 2π

For $z \in \mathbb{C} \setminus \{0\}$ we can now define a 'mod 2π' version of $\arg z$, namely

$$\operatorname{arc} z = p(\arg z) \in \mathbb{R}/2\pi$$

The notation is deliberately chosen to resemble 'arg', and also reminds us that angles are involved. The advantages of 'arc' over 'arg' are twofold. First, there is no ambiguity. Second, and more important, the function

$$\operatorname{arc} : \mathbb{C} \setminus \{0\} \to \mathbb{R}/2\pi$$

is continuous. This is false for arg, for as z moves across the negative real axis, $\arg z$ jumps from near π to near $-\pi$, because of the convention that $-\pi < \arg z \le \pi$. In contrast, since $p(-\pi) = p(\pi)$, this defect is not shared by arc. In fact, if $z = re^{i\theta}$, $r > 0$, then $j(\operatorname{arc}(z)) = e^{i\theta}$. Using this, or directly, it is not hard to verify that

$$\operatorname{arc}(z_1 z_2) = \operatorname{arc} z_1 + \operatorname{arc} z_2 \qquad (13.2)$$

for all $z_1, z_2 \in \mathbb{C} \setminus \{0\}$. Again if we use arg this is true only up to integer multiples of 2π.

Thus we have two distinct ways to represent an angle: as a uniquely defined element of $\mathbb{R}/2\pi$ (an equivalence class of reals modulo 2π), or as a real number chosen from this class, unique only up to integer multiples of 2π. Both are useful, and there are advantages in passing between them at will. We do this in the proof of our next result.

LEMMA 13.1. *If $\gamma : [a,b] \to \mathbb{C}$ is a path and $\gamma'(t_0)$ exists and is non-zero for some $t_0 \in [a,b]$, then γ has a tangent at $z_0 = \gamma(t_0)$ making an angle* $\operatorname{arc}\gamma'(t_0)$ *with the real axis.*

Proof. Let $\gamma(t) = x(t)+\mathrm{i}y(t)$. Then the required angle θ belongs to the congruence class modulo 2π of

$$\tan^{-1}(y'(t_0)/x'(t_0)) = \arg(x'(t_0) + \mathrm{i}y'(t_0)) = \arg\gamma'(t_0)$$

Taking congruence classes, we get

$$\theta = \operatorname{arc}\gamma'(t_0)$$ □

In future we will be less explicit about passing between $\mathbb{R}$ and $\mathbb{R}/2\pi$.

DEFINITION 13.2. If γ_1 and γ_2 are two paths meeting at $z_0 = \gamma_1(t_1) = \gamma_2(t_2)$ having derivatives $\gamma_1'(t_1), \gamma_2'(t_2) \neq 0$, then the *angle between* γ_1 and γ_2 *at* z_0 is

$$\theta = \operatorname{arc}\gamma_1'(t_1) - \operatorname{arc}\gamma_2'(t_2) \in \mathbb{R}/2\pi$$

as in Figure 13.1. □

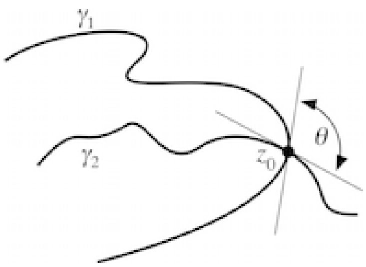

Figure 13.1 Angle between two intersecting curves.

13.2 Conformal Transformations

In this section we consider functions $f : D \to \mathbb{C}$, where D is a domain. It is convenient to distinguish the two copies of $\mathbb{C}$ by using (x, y) as coordinates in D and (u, v) as coordinates in the image $\mathbb{C}$. As usual we let $z = x + \mathrm{i}y$, and we set $w = u + \mathrm{i}v$. (The classical notation for complex functions was $w = f(z)$ before set theory came along.) If f is differentiable on D we have

$$f(x + \mathrm{i}y) = u(x, y) + \mathrm{i}v(x, y)$$

where u and v are real-valued functions of two real variables x, y. Hence f defines a function from the subset D of the (x, y)-plane to the (u, v)-plane. A smooth path γ in D, with

$$\gamma(t) = x(t) + \mathrm{i}y(t) \quad (t \in [a, b])$$

is mapped by f to a path

$$f\gamma(t) = f(\gamma(t)) = u(x(t), y(t)) + \mathrm{i}v(x(t), y(t)) \quad (t \in [a, b])$$

in the (u, v)-plane.

REMARK 13.3. Until now we have written composition as $f \circ \gamma$, but from now on we omit the circle in formulas when there is no confusion with the product. This avoids undue proliferation of circles. □

Suppose that $z_0 = \gamma(t_0)$ and $\gamma'(t_0) \neq 0$ for some $t_0 \in [a, b]$. By the chain rule,

$$(f\gamma)'(t_0) = f'(\gamma(t_0))\gamma'(t_0) = f'(z_0)\gamma'(t_0)$$

Therefore

$$\begin{aligned} \operatorname{arc}((f\gamma)'(t_0)) &= \operatorname{arc}(f'(z_0)\gamma'(t_0)) \\ &= \operatorname{arc} f'(z_0) + \operatorname{arc} \gamma'(t_0) \end{aligned} \tag{13.3}$$

by (13.2).

Suppose that γ_1 and γ_2 are two paths through z_0, say $z_0 = \gamma_1(t_1) = \gamma_2(t_2)$. Then (13.3) implies that $f\gamma_1$ and $f\gamma_2$ meet at the same angle as γ_1 and γ_2. Geometrically this is clear since both tangents are turned through the same angle $\operatorname{arc} f'(z_0)$. Alternatively, compute:

$$\begin{aligned} \operatorname{arc}((f\gamma_1)'(t_1)) - \operatorname{arc}((f\gamma_2)'(t_2)) &= \operatorname{arc} f'(z_0) + \operatorname{arc} \gamma_1'(t_1) - \operatorname{arc} f'(z_0) - \operatorname{arc} \gamma_2'(t_2) \\ &= \operatorname{arc} \gamma_1'(t_1) - \operatorname{arc} \gamma_2'(t_2) \end{aligned}$$

DEFINITION 13.4. A function $f : D \to \mathbb{C}$ that preserves angles between paths at a point z_0 is *conformal at* z_0.

If f is conformal at all $z_0 \in D$ we say it is *conformal.* □

The terms *conformal function, conformal map, conformal mapping*, and *conformal transformation* all mean the same thing. The fourth is traditional, the third is going out of fashion, the second is convenient, and the first agrees with much current terminology. One advantage of the fourth is that it focuses attention on how paths and other geometric figures transform under f.

Observe that we have proved:

THEOREM 13.5. *Let $f : D \to \mathbb{C}$ be differentiable. The f is conformal at all $z_0 \in D$ such that $f'(z_0) \neq 0$.* □

If $f'(z_0) = 0$ then f may not be conformal at z_0. For example, if $f(z) = z^2$ then the positive half of the real axis and the 'positive' half of the imaginary axis ($\{\mathrm{i}y : y > 0\}$) meet at right angles. They transform into the positive half of the real axis and the negative half of the real axis, with an angle of π.

In fact, if z_0 is a zero of f' of order m then the angle between paths meeting at z_0 is multiplied by $m+1$ on transforming by f.

We can find out a little about how f affects lengths. If $z_0, z \in \mathbb{C}$ and f is differentiable at z_0, then the ratio of the distances between $f(z)$ and $f(z_0)$ and between z and z_0 is

$$\frac{|f(z)-f(z_0)|}{|z-z_0|} = \left|\frac{f(z)-f(z_0)}{z-z_0}\right|$$

which tends to $f'(z_0)$ as $z \to z_0$. So near z_0 distances are approximately multiplied by $|f'(z_0)|$.

Some special cases illustrate the previous analysis.

Example 13.6. $f(z) = z^3$.

Here

$$\begin{aligned} u(x,y) + \mathrm{i}v(x,y) &= (x+\mathrm{i}y)^3 \\ &= (x^3 - 3xy^2) + \mathrm{i}(3x^2y - y^3) \end{aligned}$$

so that

$$\begin{aligned} u(x,y) &= x^3 - 3xy^2 \\ v(x,y) &= 3x^2y - y^3 \end{aligned}$$

Consider the paths

$$\begin{aligned} \gamma_1(t) &= 1 + \mathrm{i}t \\ \gamma_2(t) &= t + \mathrm{i} \end{aligned}$$

These are respectively the lines $x = 1, y = 1$, which meet at right angles. The paths $\Gamma_1 = f\gamma_1$ and $\Gamma_2 = f\gamma_2$ are given by

$$\begin{aligned} \Gamma_1(t) &= (1-3t^2) + \mathrm{i}(3t - t^3) \\ \Gamma_2(t) &= (t^3 - 3t) + \mathrm{i}(3t^2 - 1) \end{aligned}$$

These curves are sketched in Figure 13.2. As expected, Γ_1 and Γ_2 meet at right angles at $(-2, 2)$. □

Example 13.7. $f(z) = 1/z$.

This time

$$\begin{aligned} u(x,y) &= \frac{x}{x^2+y^2} \\ v(x,y) &= \frac{-y}{x^2+y^2} \end{aligned}$$

If $c > 0$, the circle

$$\gamma_c(t) = c\mathrm{e}^{\mathrm{i}t} \quad (t \in [0, 2\pi]) \tag{13.4}$$

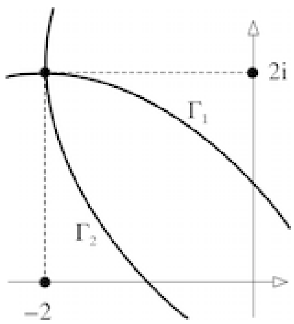

Figure 13.2 Transformed paths for Example 13.6 also meet at right angles.

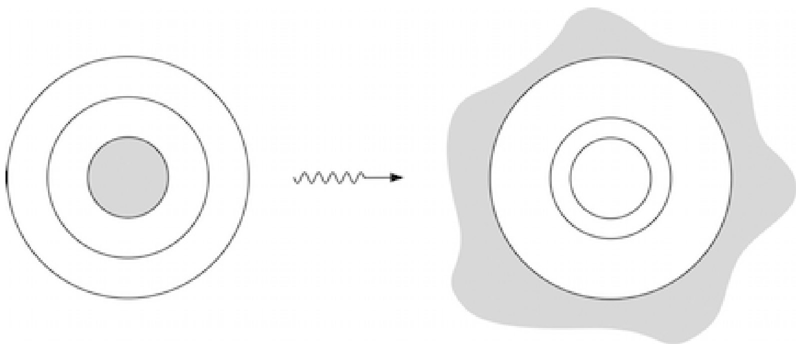

Figure 13.3 Images of circles centred at the origin for Example 13.7.

transforms into

$$\Gamma_c(t) = c^{-1}e^{-it} \quad (t \in [0, 2\pi]) \tag{13.5}$$

Hence the system of concentric circles (13.4), as c varies, maps into the system of concentric circles (13.5), Figure 13.3. Further, the lines $y = kx$ ($k \in \mathbb{R}$) through the origin are given by

$$\delta_k(t) = t + kit$$

and transform to

$$\Delta_k(t) = (t + kit)^{-1} = \frac{1}{1+k^2}\frac{1}{t} - \frac{ik}{1+k^2}\frac{1}{t}$$

which also represents lines through the origin. Now γ_c and δ_k all meet at right angles, and so do their transforms Γ_c and Δ_k, Figure 13.4. □

Example 13.8. $f(z) = \sin z$.

Here

$$u(x, y) = \sin x \cosh y$$
$$v(x, y) = \cos x \sinh y$$

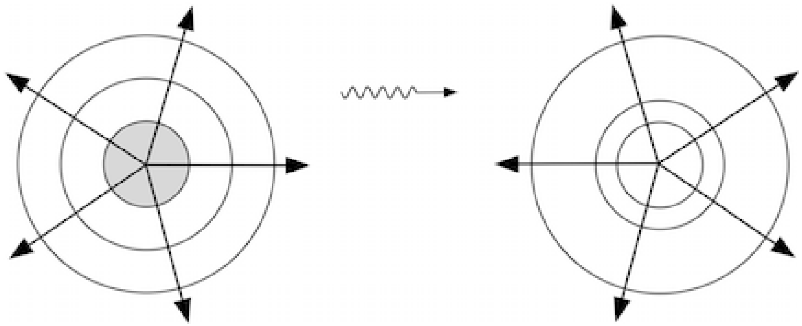

Figure 13.4 Paths for Example 13.7 and their transforms all meet at right angles.

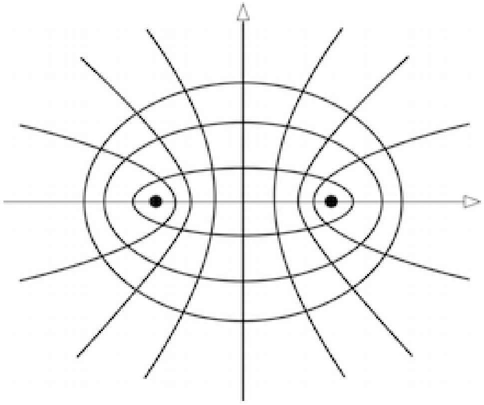

Figure 13.5 Example 13.8: Cartesian coordinate grid transforms into a system of confocal hyperbolas and ellipses. Again all curves meet at right angles.

The lines $x = c$ ($c \in \mathbb{R}$) transform into confocal hyperbolas

$$\frac{u^2}{\sin^2 c} - \frac{v^2}{\cos^2 c} = 1$$

and the lines $y = d$ ($d \in \mathbb{R}$) transform into confocal ellipses

$$\frac{u^2}{\cosh^2 d} + \frac{v^2}{\sinh^2 d} = 1$$

as in Figure 13.5. The two systems of straight lines form a grid in Cartesian coordinates, crossing at right angles. The hyperbolas and ellipses also meet at right angles, except at points where $f'(z) = 0$, namely $f(z) = \pm 1$. These are the common foci of the ellipses and hyperbolas. □

13.3 Critical Points

Theorem 13.5 proves that a differentiable complex function is conformal at any point z_0 such that $f'(z_0) \neq 0$. We now consider what happens when $f'(z_0) = 0$.

DEFINITION 13.9. Let f be a differentiable complex function. A point z_0 is a *regular point* if $f'(z_0) \neq 0$. It is a *critical point* or *singular point* if $f'(z_0) = 0$. □

Even though a complex function f need not be conformal near a critical point, the local geometry still has considerable structure. To unravel what happens at a critical point, consider the local Taylor expansion:

$$f(z_0 + h) = f(z_0) + hf'(z_0) + \cdots + h^n f^{(n)}(z_0)/n! + \cdots$$

This consists of two parts. The constant part is $w_0 = f(z_0)$, which calculates where the point z_0 in the z-plane is mapped in the w-plane. The variable part

$$f(z_0 + h) - f(z_0) = hf'(z_0) + \cdots + h^n f^{(n)}(z_0)/n! + \cdots$$

specifies how the function behaves near $w_0 = f(z_0)$. In particular, the behaviour is controlled by the first non-zero derivative $f^{(n}(z_0)$. This motivates:

DEFINITION 13.10. A differentiable function $f : D \to \mathbb{C}$ is of *order n* at $z_0 \in D$, (where $n \geq 1$) if $f^{(n)}(z_0) \neq 0$, and $f^{(k)}(z_0) = 0$ for $1 \leq k < n$. □

A function of order 1 satisfies $f'(z_0) \neq 0$, and in general, the Taylor series at a fixed point z_0 for a function of order n has the form

$$f(z_0 + h) = f(z_0) + h^n f^{(n)}(z_0)/n! + \cdots \tag{13.6}$$

Examples 13.11.

(i) $f(z) = z^n$ for integer $n \geq 1$ is of order n.
(ii) $\sin z = z - z^3/3! + \cdots$ is of order 1.
(iii) $\cos z - 1 = z^2/2! - z^4/4! + \cdots$ is of order 2. □

Equation (13.6) indicates that the behaviour of a function near a critical point is given by the order of the corresponding zero of the derivative. Indeed, a function of order $n \geq 1$ can be written as:

$$f(z_0 + h) = f(z_0) + kh^n + \text{higher order terms in } h, \text{ where } k = f^{(n)}(z_0)/n! \neq 0$$

This gives an approximation of practical value in applications:

$$f(z_0 + h) - f(z_0) \approx kh^n \text{ for suitably small } h \tag{13.7}$$

The size of h required to give a decent approximation depends on the context. As z moves from $z_0 - h$ to $z_0 + h$, the difference $f(z) - f(z_0)$ in (13.7) changes approximately from $k(-h)^n$ to kh^n. The behaviour depends on whether n is even or odd. For n odd, it moves from $-kh^n$ to kh^n, passing through zero, and for n even it travels from kh^n to zero, then back again to kh^n.

A more precise sense of the behaviour can be obtained by writing h in polar coordinates, $h = re^{i\theta}$. For suitably small h, (13.7) gives

$$f(z_0 + re^{i\theta}) - f(z_0) = kr^n e^{in\theta} + \cdots \quad (k = |f^{(n)}(z_0)|/n!)$$

so

$$|f(z_0 + re^{i\theta}) - f(z_0)| \approx kr^n$$

and

$$\text{arc}\,(f(z_0 + re^{i\theta}) - f(z_0)) \approx n\theta + \alpha \quad \text{where } \alpha = \text{arc}\,(f^{(n)}(z_0))$$

For small values of h, f transforms a tiny portion of the z-plane near z_0 to the w-plane near $w_0 = f(z_0)$, scaling r to kr^n and rotating the original angle θ to $n\theta + \alpha$, Figure 13.6.

Rotating the line from z_0 to $z_0 + h$ in the z-plane through a further angle ϕ rotates the corresponding line in the w-plane to $n(\theta + \phi) + \alpha$, so the angle between the transformed lines in the w-plane is $n\phi$. The sector of the circle radius r, angle ϕ in the z-plane is scaled to a sector in the w-plane, radius kr^n, with angle expanded to $n\phi$, Figure 13.7.

For $n = 1$ the point is regular, and f is conformal. For $n > 1$, the function f transforms z_0 to $w_0 = f(z_0)$, and a point distance r from z_0 is scaled approximately to kr^n from w_0 in the w-plane, while the angle between two paths through z_0 changes from ϕ in the z-plane to $n\phi$ in the w-plane.

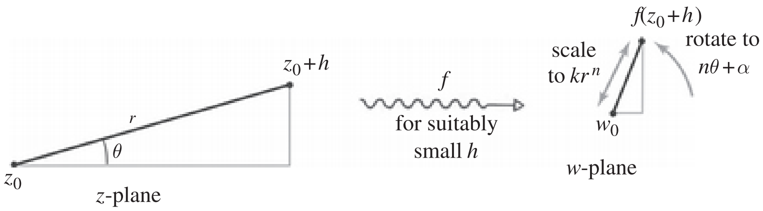

Figure 13.6 The tangent map translating, rotating and scaling from z_0 to w_0.

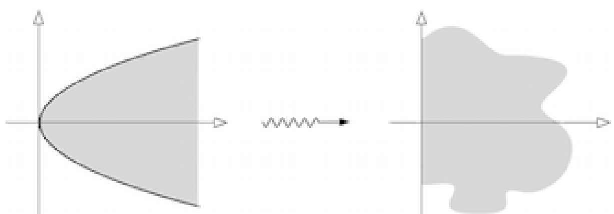

Figure 13.7 Transforming a sector of a circle centre z_0, radius r.

13.4 Möbius Maps

For fixed $a, b, c, d \in \mathbb{C}$, the function

$$f(z) = \frac{az + b}{cz + d} \quad (ad - bc \neq 0)$$

is called a *Möbius map* or *bilinear map*. (Again, 'map' can be replaced by 'function', 'mapping', or 'transformation'.) These maps have a number of important properties, and we discuss some of them.

First, note that f is differentiable for $z \neq -d/c$, and

$$f'(z) = \frac{ad - bc}{(cz + d)^2} \neq 0$$

since $ad - bc \neq 0$, again provided $z \neq -d/c$. Hence f is conformal throughout the domain $\mathbb{C} \setminus \{-d/c\}$ on which it is defined.

Now suppose that

$$f(z) = \frac{Az + B}{Cz + D} \quad (AD - BC \neq 0)$$

is a second Möbius map. Then the composite

$$gf(z) = \frac{(Aa + Bc)z + (Ab + Bd)}{(Ca + Dc)z + (Cb + Dd)}$$

is another Möbius map, because

$$(Aa + Bc)(Cb + Dd) - (Ab + Bd)(Ca + Dc) = (AD - BC)(ad - bc) \neq 0$$

So composing two Möbius maps always gives a Möbius map.

13.4.1 Möbius Maps Preserve Circles

A remarkable, and useful, property of such maps is that they transform circles (or straight lines) into circles (or straight lines). To save breath, let us agree that in this section a 'circle' may be either a circle or a straight line. (We can think of a straight line as a circle of infinite radius, remembering that ∞ is a fairly respectable concept in complex analysis.)

Suppose $p, q \in \mathbb{C}$ with $p \neq q$, and let $k > 0$. The equation

$$\frac{|z - p|}{|z - q|} = k \tag{13.8}$$

is satisfied by those points whose distances from p and q are in the ratio k. It is well known in Euclidean geometry, and can easily be checked using coordinates, that if $k \neq 1$ such points lie on a circle, and if $k = 1$ they lie on a straight line.

Here is a quick proof. By scaling and a rigid motion, we can choose coordinates in the plane so that $p = (0, 0)$ and $q = (1, 0)$. If (x, y) is distance kd from p, and distance d from q, then

$$\sqrt{x^2 + y^2} = k\sqrt{(x - 1)^2 + y^2}$$

Hence

$$x^2 + y^2 = k^2(x^2 + y^2 - 2x + 1)$$

which implies

$$\left(x - \frac{k^2}{k^2-1}\right)^2 + y^2 + \frac{k^2}{k^2-1} - \left(\frac{k^2}{k^2-1}\right)^2 = 0$$

which is the equation of a circle.

If we put

$$w = f(z) = \frac{az+b}{cz+d} \quad (ad - bc \neq 0)$$

it is easy to verify that

$$z = \frac{-dw+b}{cw-a}$$

Now (13.8) takes the form

$$\left|\frac{\frac{az+b}{cz+d} - p}{\frac{az+b}{cz+d} - q}\right| = k$$

which simplifies to

$$\frac{|w-P|}{|w-Q|} = K \tag{13.9}$$

where

$$\begin{aligned} P &= (b+pa)/(d+pc) \\ Q &= (b+qa)/(d+qc) \\ K &= k|d+qc|/|d+pc| \end{aligned}$$

Hence (13.9) also represents a circle. This proves that the Möbius map f transforms circles into circles.

13.4.2 Classification of Möbius Maps

The above calculation, while it verifies the assertion, is not as instructive as we might hope, and the following approach has its advantages. We consider several special types of Möbius map that have especially simple forms, which can easily be seen to transform circles into circles. Then we show that a general Möbius map can be obtained by composing these types.

It is perhaps not so surprising that this approach works, in view of the composition property of Möbius maps. The decomposition can be seen as analogous to the well-known fact that any rigid motion of the Euclidean plane can be obtained by composing a translation, a rotation, and a reflection.

We begin with the special types.

Translation: $w = z + k$ ($k \in \mathbb{C}$). This corresponds geometrically to moving points $\operatorname{re} k$ to the right and $\operatorname{im} k$ upwards, and clearly preserves the shape of geometric figures, in particular circles.

Rotation: $w = e^{i\theta} z$ ($\theta \in \mathbb{R}$). All points rotate round the origin through angle θ. Again this preserves the shape of geometric figures.

Magnification: $w = hz$ ($h > 0$). This produces a change of scale. If $h < 1$ it shrinks rather than magnifies, but such pedantry is irrelevant and we still use the term 'magnification'. It therefore maps geometric figures to similar geometric figures.

Inversion: $w = 1/z$. Devotees of 'inversive geometry', now largely out of fashion, will recognise the corresponding 'geometric inversion', which is well known to preserve circles. However, circles can map to straight lines, or straight lines to circles. The rest of us can check this by a coordinate calculation similar to the one above.

We can now state:

THEOREM 13.12. *Every Möbius map can be obtained by composing a translation, an inversion, a magnification, a rotation, and another translation.*

Proof. Suppose that $a, b, c, d \in \mathbb{C}$ and $ad - bc \neq 0$, where $c \neq 0$. Define

$$\begin{aligned} t_1(z) &= z + d/c \quad \text{(translation)} \\ j(z) &= 1/z \quad \text{(inversion)} \\ m(z) &= \left| \frac{ad - bc}{c^2} \right| z \quad \text{(magnification)} \\ r(z) &= \frac{(bc - ad)|c|^2}{|ad - bc|c^2} z \quad \text{(rotation)} \\ t_2(z) &= z + a/c \quad \text{(translation)} \end{aligned}$$

It is routine to verify that

$$(t_2 r m j t_1)(z) = \frac{az + b}{cz + d}$$

If $c = 0$ define

$$\begin{aligned} t_1(z) &= z + b/a \quad \text{(translation} \\ m(z) &= \left| \frac{a}{d} \right| z \quad \text{(magnification)} \\ r(z) &= \frac{a}{d} \left| \frac{d}{a} \right| z \quad \text{(rotation)} \end{aligned}$$

noting that $ad - bc \neq 0$ implies $ad \neq 0$ since $c = 0$, hence $a \neq 0, d \neq 0$. Now

$$(r m t_1)(z) = \frac{az + b}{cz + d}$$ □

COROLLARY 13.13. *Every Möbius map transforms circles into circles.* □

There are other functions $f : \mathbb{C} \to \mathbb{C}$ that preserve circularity, the most obvious being complex conjugation (which is not differentiable). A theorem of Constantin Carathéodory asserts that every such function is either a Möbius map, or a Möbius map composed with conjugation. No differentiability assumptions are needed for this theorem. We do not prove it here.

13.4.3 Extension of Möbius Maps to the Riemann Sphere

There is a natural way to extend Möbius maps to the Riemann sphere by examining their behaviour as $z \to \infty$.

PROPOSITION 13.14. *Any Möbius map can be extended uniquely to the Riemann sphere so that it is differentiable at every point.*

Proof. By continuity we can find this extension by letting $z \to \infty$, but we must then verify differentiability. If

$$f(z) = \frac{az+b}{cz+d} \quad (ad - bc \neq 0)$$

then letting $z \to \infty$ we must define

$$f(\infty) = \begin{cases} \frac{a}{c} & \text{if } c \neq 0 \\ \infty & \text{if } c = 0 \end{cases}$$

Differentiability at ∞ is an easy exercise. □

When extended in this manner, f is a *bijection* from $\mathbb{C} \cup \{\infty\}$ to itself. If we identify $\mathbb{C} \cup \{\infty\}$ with the Riemann sphere $\mathbb{S}^2$, then circles in $\mathbb{C}$ remain circles on $\mathbb{S}^2$, and straight lines in $\mathbb{C}$ *also* become circles on $\mathbb{S}^2$.

It is instructive to work out the effect of the basic types of Möbius map on $\mathbb{S}^2$. We omit the details.

13.5 Potential Theory

We now turn – briefly – to classical applications of complex analysis. These were instrumental in convincing mathematicians that the subject was worth taking seriously, before the basic concepts had been made rigorous. We discuss potential theory, a general area of mathematical physics with applications to gravitation electrostatics, magnetism, and fluid flow.

13.5.1 Laplace's Equation

The two-dimensional *Laplace equation*

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} = 0$$

for a function $\phi(x, y)$ is important in potential theory, with applications in particular to fluid dynamics. It is closely connected with complex function theory, as we now demonstrate. Let $f : D \to \mathbb{C}$ be differentiable, with $z = x + \mathrm{i}y$, and write

$$f(z) = u(x, y) + \mathrm{i}v(x, y)$$

as usual. Then

$$f'(z) = \frac{\partial u}{\partial x} + \mathrm{i}\frac{\partial v}{\partial x} = \frac{\partial v}{\partial y} - \mathrm{i}\frac{\partial u}{\partial y}$$

as in Section 4.2. By Theorem 10.3 f'' exists throughout D. If we let $f'(z) = U + iV$ then the Cauchy–Riemann Equations of Theorem 4.12 state that

$$\frac{\partial U}{\partial x} = \frac{\partial V}{\partial y} \qquad \frac{\partial V}{\partial x} = -\frac{\partial U}{\partial y}$$

Therefore

$$U = \frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \qquad V = \frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}$$

so we get

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\right) = \frac{\partial U}{\partial x} = \frac{\partial V}{\partial y} = -\frac{\partial}{\partial y}\left(\frac{\partial u}{\partial y}\right) = -\frac{\partial^2 \phi}{\partial y^2}$$

Therefore $u(x, y)$ satisfies the Laplace equation. Similarly, so does $v(x, y)$.

For instance, consider $f(z) = z\mathrm{e}^z$. Then

$$u(x, y) = x\mathrm{e}^x \cos y - y\mathrm{e}^x \sin y$$
$$v(x, y) = y\mathrm{e}^x \cos y + x\mathrm{e}^x \sin y$$

and it may be verified directly that these functions satisfy Laplace's Equation.

Solutions of Laplace's equation are called *harmonic* or *potential functions*. Pairs of functions u, v obtained from a differentiable function by the above method are called *harmonic conjugates*.

The lines $u =$ constant, $v =$ constant, are orthogonal (mutually perpendicular) in the (u, v)-plane, so by conformality the curves

$$u(x, y) = \text{constant}$$
$$v(x, y) = \text{constant}$$

are orthogonal in the (x, y)-plane. In potential theory, if u is harmonic, the lines $u(x, y) =$ constant are called *equipotential lines*, and the set of orthogonal curves $v(x, y) =$ constant are called *streamlines*. In the case when Laplace's equation describes fluid flow, streamlines are the paths along which the fluid flows. If we are given u in a domain D and wish to find the streamlines given by v, we can often use complex integration. For a fixed point $z_0 \in D$, and any $z_1 \in D$, we have

$$f(z_1) = \int_{z_0}^{z_1} f'(z)\mathrm{d}z = \int_{z_0}^{z_1} \left(\frac{\partial u}{\partial x} - \mathrm{i}\frac{\partial u}{\partial y}\right)\mathrm{d}z$$

For example, if $u(x, y) = x^2 - y^2$, which is harmonic, then picking $z_0 = 0$ for convenience we have

$$f(z_1) = \int_0^{z_1} (2x + 2iy)dz = \int_0^{z_1} 2z dz = z_1^2$$

Hence $f(x + iy) = (x + iy)^2 = (x^2 - y^2) + i(2xy)$. So the streamlines have equations $2xy =$ constant, or equivalently

$$xy = \text{constant}$$

Often, as in this case, we can *guess* what f, hence $v(x, y)$, ought to be – a process dignified by the term 'inspection'.

13.5.2 Design of Aerofoils

Conformal maps and complex function theory played a part in the design of aircraft in the early days of aviation (and, in a more sophisticated way, still do, although most modelling is now done using the numerical methods of computational fluid dynamics). In particular, the transformation from the z-plane to the w-plane given by

$$\frac{w-2}{w+2} = \left(\frac{z-1}{z+1}\right)^2$$

maps a circle in the z-plane, passing through -1 and containing $+1$ in its interior, into a 'bent teardrop' shape as in Figure 13.8 that resembles a cross-section of an aircraft's wing. This is known as a *Joukowski aerofoil*, defined by a *Joukowski transformation*, after Nikolai Zhukovsky, who discovered the transformation.

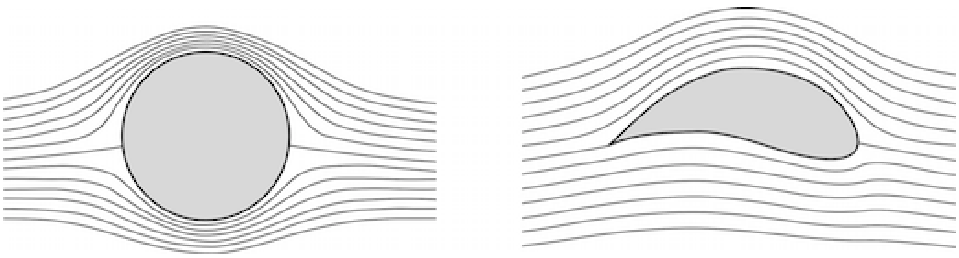

Figure 13.8 *Left*: Flow past circle (with appropriate circulation added). *Right*: Circle transformed to Joukowski aerofoil, with transformed flow.

It is used as follows. It is quite easy to solve the Laplace equation and find streamlines for the flow of a fluid round a circular disc. Now apply the Joukowski transformation: the disc maps to an aerofoil, and the streamlines round the disc map to the streamlines round the aerofoil. In order to make the flow behave in a sensible physical manner at the sharp edge of the aerofoil, the flow past the circle is modified by adding a 'circulation' term that flows round the circle. From this we can calculate properties of the flow; in particular, the amount of 'lift' imparted to the aircraft. More subtle transformations, such as the Karmann–Trefftz transform

$$\frac{w-n}{w+n}=\left(\frac{z-1}{z+1}\right)^n$$

give more accurate information.

13.6 Exercises

1. Sketch the image of the set $D = \{z \in \mathbb{C} : \operatorname{re} z > 0, \operatorname{im} z > 0, |z| < 1\}$ under the conformal map $f(z) = 1/z$. Draw the images of the lines in D parallel to the coordinate axes.
2. Show that the map

$$f(z)=\frac{\sqrt{1+k^2}}{b}\exp\left(-\mathrm{i}\left(\frac{\pi}{2}+\tan^{-1}k\right)z\right)$$

transforms the strip between the lines $y = kx, y = kx + b$ into that between $x = 0$ and $x = 1$.
3. Find a conformal map f of the annulus $2 < |z| < 5$ on to the annulus $4 < |z| < 10$, such that $f(-5) = 10$. (These points lie on the boundary so f must also be defined there.) Find another with $f(5) = 4$.
4. With a suitable choice of the square root, show that

$$f(z)=\sqrt{z-p}-\mathrm{i}\sqrt{p}$$

maps the exterior of the parabola $y^2 = 4px$ ($p > 0$) onto the right-hand half-plane $x > 0$, Figure 13.9.
5. Find a conformal map that sends the interior of the right-hand branch of the hyperbola

$$x^2-y^2=\lambda^2$$

to the upper half-plane $y > 0$, Figure 13.10. (Hint: what is $\operatorname{re} z^2$?)
6. Show that the semicircle $|z| < 1, \operatorname{re} z > 0$ is mapped by

$$f(z)=z^2+z$$

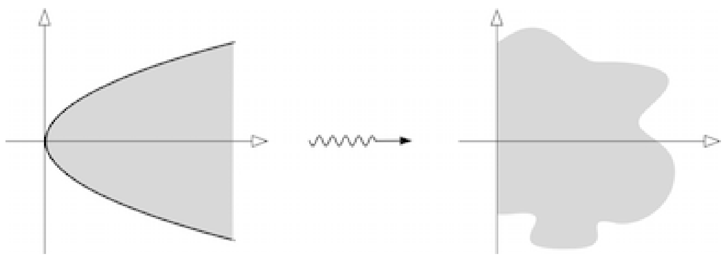

Figure 13.9 Conformal map of interior of parabola to half-plane.

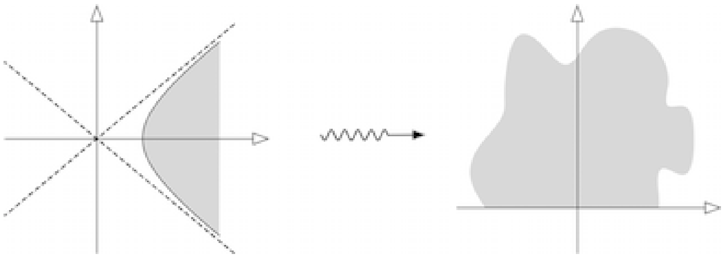

Figure 13.10 Conformal map of interior of hyperbola to half-plane.

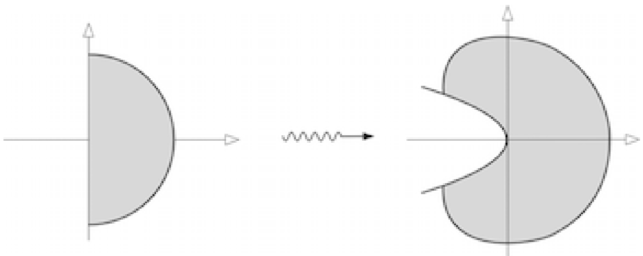

Figure 13.11 Conformal map for Exercise 6.

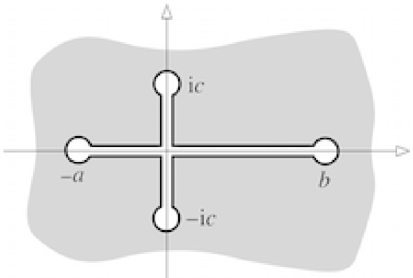

Figure 13.12 Region for Exercise 7.

to the region bounded by the parabola $x = -y^2$ and the curve whose equation in polar coordinates is

$$r = 2\cos(\theta/3)\ (|\theta| \leq 3\pi/4)$$

See Figure 13.11.

7. Show that for suitably defined square roots, when $a, b, c > 0$,

$$f(z) = \sqrt{\frac{\sqrt{z^2+c^2}+\sqrt{a^2+c^2}}{\sqrt{bz^2+c^2}-\sqrt{z^2+c^2}}}$$

maps the plane with the segments between $-a$, and b, and $-$ic and ic removed, to the upper half-plane. See Figure 13.12.

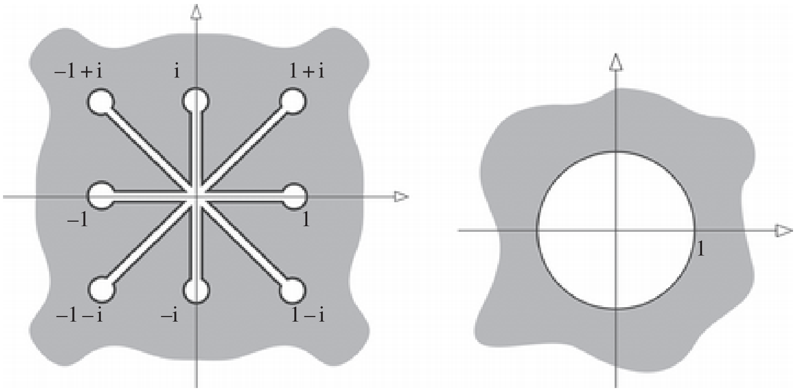

Figure 13.13 *Left*: Region for Exercise 8. *Right*: Exterior of unit circle.

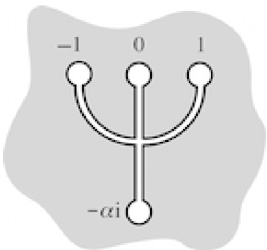

Figure 13.14 Region for Exercise 9.

8. Show that for suitably defined square roots,

$$f(z) = \frac{1}{\sqrt{2+\sqrt{5}}}\left(\sqrt{\sqrt{z^4+4}+2} + \sqrt{\sqrt{z^4+4}-\sqrt{5}}\right)$$

maps Figure 13.13 (left) to the exterior of the unit circle, Figure 13.13 (right).

9. Find a conformal map from Figure 13.14 to the upper half-plane. (Hint: compare Exercise 7.)

10. Show that

$$f(z) = \sqrt{\frac{\cos \pi z - \cosh \pi h}{1 + \cos \pi z}}$$

maps Figure 13.15 to the upper half-plane.

11. Find a conformal map from Figure 13.16 to the upper half-plane.

12. Find the Möbius map that sends $-1, \infty, \mathrm{i}$ respectively to:
(i) $\mathrm{i}, 1, 1+\mathrm{i}$
(ii) $\infty, \mathrm{i}, 1$
(iii) $0, \infty, 1$

13. Find the general form of a Möbius map that:
(i) Transforms the upper half-plane to itself.
(ii) Transforms the upper half-plane to the lower half-plane.

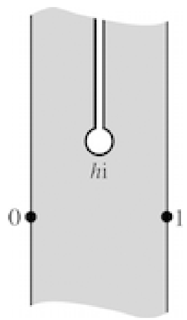

Figure 13.15 Region for Exercise 10.

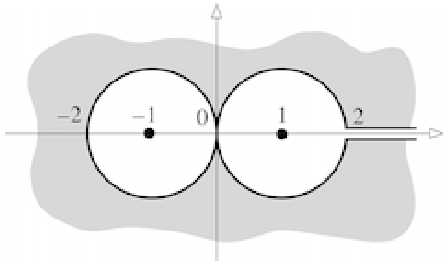

Figure 13.16 Region for Exercise 11.

(iii) Transforms the upper half-plane to the right half-plane.
(iv) Preserves the unit circle.
(v) Preserves the coordinate axes.
(vi) Transforms the upper half-plane to the interior of the unit circle.

14. *Invariance of Cross Ratio*. The *cross ratio* of four complex numbers z_1, z_2, z_3, z_4 is

$$\frac{(z_1 - z_2)(z_3 - z_4)}{(z_1 - z_4)(z_3 - z_2)}$$

Let $\mu(z) = \dfrac{az+b}{cz+d}$ be a Möbius map with $ad - bc \neq 0$, and let $w_j = \mu(z_j)$ for $j = 1, 2, 3, 4$. Prove that

$$\frac{(w_1 - w_2)(w_3 - w_4)}{(w_1 - w_4)(w_3 - w_2)} = \frac{(z_1 - z_2)(z_3 - z_4)}{(z_1 - z_4)(z_3 - z_2)}$$

(Hint: first work out how $w_1 - w_2$ relates to $z_1 - z_2$.)

15. Let f be a Möbius map. A *fixed point* of f is a point z such that $f(z) = z$.

Prove that f has at most two distinct fixed points, including ∞.

If f has a unique fixed point (including ∞) it is said to be *parabolic*. Prove it can then be written in the form

$$\frac{1}{f(z) - z_0} = \frac{1}{z - z_0} + h \quad (z_0 \neq \infty)$$

or

$$f(z) = z + h \quad \text{(a translation)}$$

If f has two distinct fixed points z_1, z_2, show that it can be written in the form

$$\frac{f(z) - z_1}{f(z) - z_2} = k\frac{z - z_1}{z - z_2} \quad (z_1, z_2 \neq \infty)$$

or

$$f(z) - z_1 = k(z - z_1) \quad (z_2 = \infty)$$

Such a map is said to be *hyperbolic* if $k > 0$, *elliptic* if $k = e^{i\alpha}$ ($\alpha \neq 0$), and *loxodromic* if $k = ae^{i\alpha}$, where $a \neq 1$ is real and $\alpha \neq 0$.

16. With the definitions of Exercise 14, prove the following:
  (i) Any Möbius map $(az + b)/(cz + d)$ is equal to one for which $ad - bc = 1$.
  (ii) Having ensured this, if $a + d$ is real then the map is elliptic if $|a + d| < 2$, hyperbolic if $|a + d| > 2$, and loxodromic if $|a + d| = 2$.
  (iii) If $a + d$ is not real, the transformation is loxodromic.

17. Show that

$$u(x, y) = x^3 - 3xy^2$$

is harmonic. Find a harmonic function $v : \mathbb{R}^2 \to \mathbb{R}$ such that

$$f(x + iy) = u(x, y) + iv(x, y)$$

is a differentiable complex function. Prove that v is unique up to the addition of a real constant.

18. For $f(z) = 1/z$, write $f(z) = u(x, y) + iv(x, y)$. Sketch the level curves $u(x, y) =$ constant, $v(x, y) =$ constant. If a level curve of u meets a level curve of v, what is the angle between them? Is there an easy way to see this, and if so, what is it?

19. Find the most general cubic polynomial

$$u(x, y) = ax^3 + bx^2y + cxy^2 + dy^3 \quad (a, b, c, d \in \mathbb{R})$$

that is harmonic. Find a differentiable function with u as its real part.

20. Verify that the Joukowski transformation does, as claims above, give rise to an aerofoil shape. Look up pages 131–134 of Kyrala [12] and see how to compute streamlines round it. A more modern book is Anderson [1]. The NASA applet at www.grc.nasa.gov/WWW/K-12/airplane/map.html is also worth investigating.

14 Analytic Continuation

When Weierstrass started his programme to rigorise complex analysis he based it on power series. Because these have well-behaved convergence properties, and can be differentiated or integrated term by term, they provide a tool of great technical value. However, this tool has limitations, which we illustrate in the final section of this chapter: many important functions cannot be represented by a *single* power series – mainly because power series converge on *discs*.

This limitation can be overcome by the method of 'analytic continuation', which lets us extend the domain of a complex function, subject to suitable conditions. It turns out that such an extension is not always unique, and the problem of describing the different possibilities and how they are related leads to a remarkable geometric concept, known as a 'Riemann surface' after its inventor. In this chapter we discuss these topics and related ones. Here we use the term 'analytic' rather than 'differentiable', to emphasise the power series viewpoint – but you should remember that in the complex case these terms are synonymous.

14.1 The Limitations of Power Series

We illustrate the problem for the function

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

It would be possible to use a simpler example, such as $1/z$, but it is more appropriate to work with something a little less special, where the general problem is more sharply defined

This function is analytic in $\mathbb{C} \setminus \{-1, 1\}$, and has simple poles at -1 and 1. Its power series expansion about $z_0 = 0$ is

$$1 + z^2 + z^4 + z^6 + \cdots \tag{14.1}$$

which converges for $|z| < 1$ but diverges for $|z| > 1$. Thus if we require functions to be defined on domains, the series (14.1) at best represents f on an open disc

$$D_1 = \{z \in \mathbb{C} : |z| < 1\}$$

Thus (14.1) tells us about a small part of f (directly, at least, although it has useful indirect implications, as we shall see).

We can begin to get round this problem by choosing a different z_0, for example $z_0 = \mathrm{i}$. To find the Taylor series for $f(z)$ around $z_0 = \mathrm{i}$ it helps to rewrite f in the form

$$f(z) = \frac{1}{2}\left(\frac{1}{1+z} + \frac{1}{1-z}\right)$$

Let $w = z - \mathrm{i}$, so that $z = w + \mathrm{i}$. Then

$$\begin{aligned}
f(z) &= \frac{1}{2}\left(\frac{1}{1+w+\mathrm{i}} + \frac{1}{1-w-\mathrm{i}}\right) \\
&= \frac{1}{2}\left[\frac{1}{1+\mathrm{i}}\left(1+\frac{w}{1+\mathrm{i}}\right)^{-1} + \frac{1}{1-\mathrm{i}}\left(1+\frac{w}{1-\mathrm{i}}\right)^{-1}\right] \\
&= \frac{1}{2(1+\mathrm{i})}\sum_{n=0}^{\infty}(-1)^n\left(\frac{w}{1+\mathrm{i}}\right)^n + \frac{1}{2(1-\mathrm{i})}\sum_{n=0}^{\infty}\left(\frac{w}{1-\mathrm{i}}\right)^n \\
&= \sum_{n=0}^{\infty}\frac{1}{2}\left[\frac{1}{1+\mathrm{i}}\left(\frac{-1}{1+\mathrm{i}}\right)^n + \frac{1}{1-\mathrm{i}}\left(\frac{1}{1-\mathrm{i}}\right)^n\right](z-\mathrm{i})^n \qquad (14.2)
\end{aligned}$$

The series (14.2) has radius of convergence $\sqrt{2}$, because that is the distance from $z_0 = \mathrm{i}$ to the nearest pole of f, so it converges on

$$D_2 = \{z \in \mathbb{C} : |z - \mathrm{i}| < \sqrt{2}\}$$

Figure 14.1 shows that (14.2) converges for some values of z for which (14.1) converges, and conversely. Thus not only do (14.1) and (14.2) represent small parts of f (namely its restrictions to D_1, D_2) but they represent *different* parts.

Similarly if we take $z_0 = 2$ we obtain a third series

$$f(z) = \sum_{n=0}^{\infty}\left[-\tfrac{1}{2}(-1)^n + \tfrac{1}{6}(-\tfrac{1}{3})^n\right](z-2)^n \qquad (14.3)$$

convergent on the disc

$$D_3 = \{z \in \mathbb{C} : |z - 2| < 1\}$$

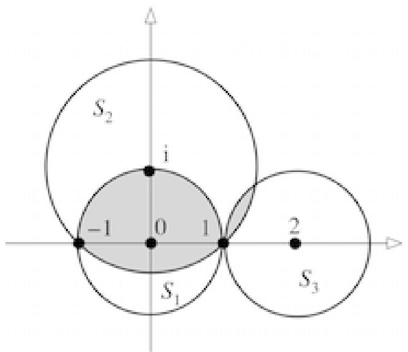

Figure 14.1 Discs of convergence of various local power series for f. Shading shows intersections of discs.

which represent yet another part of f. And something of this kind happens for any choice of z_0: the Taylor series

$$f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n \tag{14.4}$$

converges for

$$|z - z_0| < K = \min(|z_0 - 1|, |z_0 + 1|)$$

by Theorem 10.3. Thus no single choice of z_0 gives a power series expansion of $f(z)$ valid for all $z \in \mathbb{C} \setminus \{-1, 1\}$, even though f is analytic on this domain. (You should check that Laurent series do not improve the situation.)

This problem is more a limitation of a tool – power series – than a defect of analytic functions. It is no fault of $f(z)$ that our clumsy attempt to represent it using power series has apparently failed. The fact that a power series converges on a disc, so often a help, becomes a hindrance when we look at functions defined on domains that are not discs.

We could give up at this stage and ignore power series, but that would hardly be enterprising: it is a cardinal principle in mathematics not to throw away a good idea just because it does not work. If a single power series is no good, why not try a whole collection of power series? This solves part of the problem. We can certainly find, for each $z_0 \in \mathbb{C} \setminus \{-1, 1\}$, a power series expansion (14.4) around z_0 that converges to $f(z)$ for all z near z_0. Thus using several power series gives information about the whole of f.

However, it also creates another, more serious, problem. In the discussion so far we started with f and worked out the power series. Weierstrass needed to proceed in the opposite direction: use power series to *define* $f(z)$. In the example above we know that the series (14.1), (14.2), and (14.3) represent the *same* function f, because that is how they are constructed. *If we are given two power series expansions around different points, how can we tell whether they represent the same analytic function, without knowing a priori what that function is?* This is the central problem with Weierstrass's approach.

14.2 Comparing Power Series

We wish to compare two power series

$$p(z) = \sum_{n=0}^{\infty} p_n (z - z_0)^n$$

$$q(z) = \sum_{n=0}^{\infty} q_n (z - z_1)^n$$

around z_0, z_1 respectively, convergent on open discs P, Q. Everything is easy if P and Q *overlap* (that is, $P \cap Q \neq \emptyset$), as is the case for (14.1) and (14.2), or (14.2) and (14.3), see Figure 14.2. In this case we can prove:

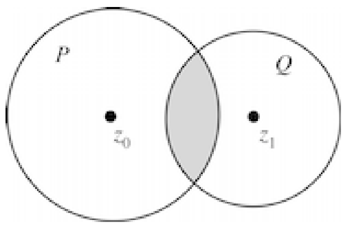

Figure 14.2 Overlapping discs of convergence.

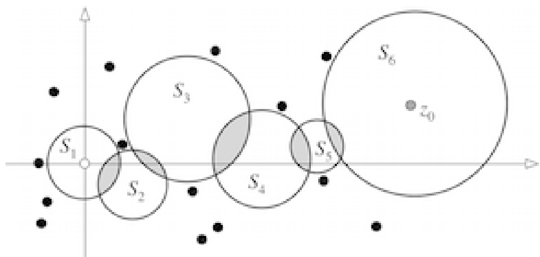

Figure 14.3 With more singularities, chains of discs may have to be longer.

LEMMA 14.1. *Let f, g be analytic on a domain D. Suppose that P and Q are open sets and $\emptyset \neq P \cap Q \subseteq D$. Suppose that $p(z) = q(z)$ for all $z \in P \cap Q$, and f, g are analytic functions defined on D such that $f(z) = p(z)$ for $z \in P$ and $g(z) = q(z)$ for $z \in Q$. Then $f(z) = g(z)$ for all $z \in D$.*

Proof. For z in the non-empty set $P \cap Q$,

$$f(z) = p(z) = q(z) = g(z)$$

Now Theorem 10.16 implies that $f(z) = g(z)$ for all $z \in D$. □

In this sense, p and q represent the same analytic function. Obviously the converse also applies: if p and q represent the same analytic function on D then they must agree on the overlap.

The existence of a non-empty overlap allows *direct* comparison of the two power series: all we have to do is look at their values. But what if the discs of convergence do not overlap, as for (14.1) and (14.3)? The answer is to construct a chain of overlapping discs from one to the other, on each of which the power series converges, and such that the power series agree on all overlaps. Thus D_1 and D_2 overlap, and (14.1) agrees with (14.2) on $D_1 \cap D_2$; and D_2 and D_3 overlap, and (14.2) agrees with (14.3) on $D_2 \cap D_3$.

It is fairly clear that for $f(z) = 1/(1-z^2)$ we can get from D_1 to any point $z_0 \neq \pm 1$ by a chain of at most three discs (two unless z_0 is real and $|z_0 > 1|$). For a more complicated function with more poles (or other singularities) we may need more discs in a chain, because the discs have to 'push between' the singularities, see Figure 14.3. In a sense this is the reason for the limitation of power series: discs are too blunt an instrument to penetrate beyond the singularities into more distant territory.

When we formalise these ideas, as in the next section, it becomes clear that the restriction to power series is inessential. This is often the way in mathematics: the solution to a special problem turns out to apply in a much more general setting. The special problem plays the useful role of a psychological springboard, hurling our thoughts into higher realms.

14.3 Analytic Continuation

The simplest type of analytic continuation is direct analytic continuation, extending a function from one domain to an overlapping domain. Repeating this process we get indirect analytic continuation. We consider these in turn.

14.3.1 Direct Analytic Continuation

DEFINITION 14.2. If f_1 is analytic on a domain D_1 and f_2 is analytic on a domain D_2, where $D_1 \cap D_2 \neq \emptyset$ and $f_1(z) = f_2(z)$ for all $z \in D_1 \cap D_2$, then f_2 is a *direct analytic continuation* of f_1 to the domain D_2, see Figure 14.4. □

Such an f_2 must be unique by Lemma 14.1.

As a simpler example than that in the previous section, take

$$f_1(z) = \sum_{n=0}^{\infty} z^n \quad (|z| < 1)$$
$$f_2(z) = 1/(1-z) \quad (z \in \mathbb{C} \setminus \{1\})$$

Then f_2 is a direct analytic continuation of f_1. Whereas f_1 is defined only in the interior of the unit disc, f_2 is defined on the whole of $\mathbb{C}$ except at 1.

Before going on to the general case, a series of overlapping domains, it is worth dealing with another phenomenon. Sometimes D_1 may be such that f_1 has no direct continuation to any D_2 not contained in D_1. In this case the boundary of D_1 is called a *natural boundary* for f_1. Now the limitation does not stem from our choice of tools, but from intrinsic properties of f_1: it so happens that D_1 is the end of the line as far as analytic continuation of f_1 is concerned.

The standard example of this phenomenon is

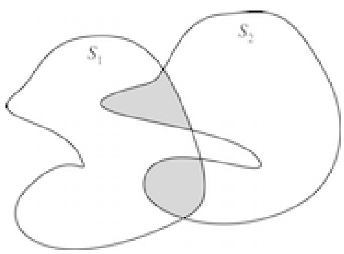

Figure 14.4 Direct analytic continuation.

$$f(z) = \sum_{n=0}^{\infty} z^{n!} \tag{14.5}$$

which converges (so is analytic) for $|z| < 1$. We prove:

PROPOSITION 14.3. *The unit circle* $\mathbb{S} = \{z \in \mathbb{C} : |z| = 1\}$ *is a natural boundary for the function* f *defined by* (14.5).

Proof. Let $z_0 = e^{ip/q}$ for $p, q, \in \mathbb{Z}, q \geq 1$. Then $z_0^{n!} = 1$ for all $n \geq q$. We show first that as $z \to z_0$ then $f(z) \to \infty$. Let $z = rz_0$ where $0 < r < 1$. Then

$$\begin{aligned} f(z) &= \sum_{n=0}^{\infty} (rz_0)^{n!} \\ &= (1 + rz_0 + \cdots + r^{(q-1)!} z_0^{(q-1)!}) + \sum_{n=q}^{\infty} r^{n!} \\ &= g(r) + h(r), \text{ say.} \end{aligned}$$

Note that $h(r)$ has the above form because $z_0^{n!} = 1$ for all $n \geq q$. For any fixed integer $N \geq 0$,

$$\sum_{n=q}^{q+N} r^{n!} \to N + 1 \quad \text{as } r \to 1$$

So for some $\varepsilon > 0$, if $1 - \varepsilon < r < 1$, then

$$\sum_{n=q}^{q+N} r^{n!} \geq \tfrac{1}{2} N$$

Then

$$h(r) = \sum_{n=q}^{\infty} r^{n!} \geq \sum_{n=q}^{q+N} r^{n!} \geq \tfrac{1}{2} N$$

so that $h(r) \to \infty$ as $r \to 1$, and $h(r) \to \infty$ as $z \to z_0$. On the other hand,

$$g(z) = 1 + z_0 + \cdots + + r^{(q-1)!} z_0^{(q-1)!} \quad \text{as } z \to z_0$$

Therefore

$$\lim_{z \to z_0} f(z) = \infty \tag{14.6}$$

Now, suppose that F is a direct analytic continuation of f to a domain D_2 not contained in D_1. Then $\partial D_1 \cap D_2$ is open in $\partial D_1 = \mathbb{S}$, so it contains some point $z_0 = e^{ip/q}$ for $p, q, \in \mathbb{Z}, q \geq 1$ since such points are clearly dense on the unit circle. There is a small disc D_ε round z_0 such that $D_\varepsilon \subseteq D_2$. Now for $1 > r > 1 - \varepsilon$ we have $rz_0 \in D_\varepsilon$, so that

$$F(rz_0) = f(rz_0)$$

As $r \to 1$,

$$F(rz_0) \to \infty$$

by (14.6). But F is analytic in D_2, so

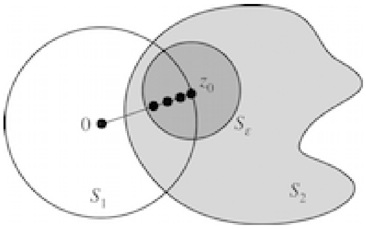

Figure 14.5 Sequence of points tending to z_0.

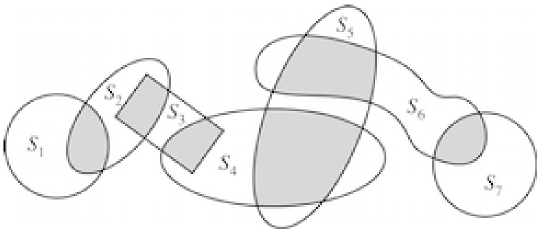

Figure 14.6 Sequence of overlapping domains.

$$F(rz_0) \to F(z_0)$$

as $r \to 1$, by continuity. This is a contradiction; therefore no direct analytic continuation of f to D_2, not contained in D_1, is possible. □

Figure 14.5 illustrates the geometry of this proof. Informally, the point is that the singularities $e^{ip/q}$ are so close together that there is no room to squeeze a disc between them, onto which f might be analytically continued.

14.3.2 Indirect Analytic Continuation

We now return to the question of performing a sequence of direct analytic continuations, one after the other.

DEFINITION 14.4. Let $D_1, \ldots, D_n$ be domains such that

$$D_r \cap D_{r+1} \neq \emptyset \quad (r = 1, \ldots, n-1)$$

as in Figure 14.6. If (f_r) is a sequence of analytic functions defined on D_r such that f_{r+1} is a direct analytic continuation of f_r $(r = 1, \ldots, n-1)$, then f_n is an *analytic continuation* of f_1 from D_1 to D_n.

An analytic continuation that is not direct is an *indirect analytic continuation.* □

Unlike the direct case, we may obtain different results by using different sequences of domains, Figure 14.7. Indeed, we can even have a sequence of overlapping domains for which $D_n = D_1$, but $f_n \neq f_1$, so that on returning to our starting point we end up with a different function from the original one. We illustrate this possibility by an example in the next section. But first, we apply the above ideas to define a broader concept of an analytic function.

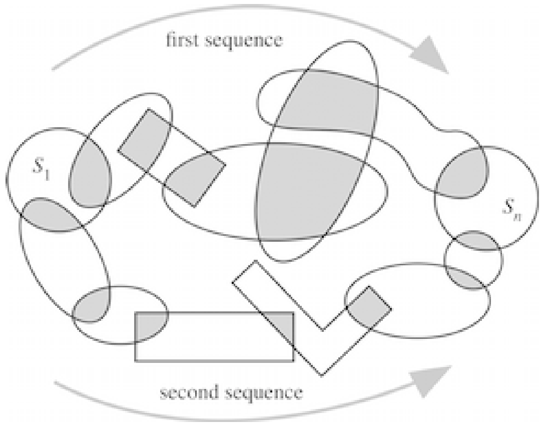

Figure 14.7 Sequences of overlapping domains that may lead to different analytic continuations.

14.3.3 Complete Analytic Functions

If f is analytic in a domain D we call the pair (f, D) a *function element*. We define a relation $\sim$ on the set of all function elements by

$$(f_1, D_1) \sim (f_2, D_2)$$

if f_2 is an analytic continuation of f_1 from D_1 to D_2. Since indirect continuations are allowed, it follows easily that $\sim$ is an equivalence relation. A *complete analytic function* is an equivalence class under $\sim$ of function elements.

In other words, a complete analytic function in the new sense is an analytic function in the old sense, together with all of its analytic continuations. It is clearly more convenient to have the analytic continuations 'built in' than to have to fit them together as the occasion warrants.

If there exist function elements $(f_1, D_1$ and (f_2, D_2) of the complete analytic function F such that for some $z \in D_1 \cap D_2$ we have $f_1(z) \neq f_2(z)$, we say that F is *multiform*. If not, then F is *uniform*. All the examples considered so far in this chapter are uniform, but we give examples of multiform functions in the next section. A multiform F is a formal version of the classical idea of a 'multivalued function', with the advantage that it is broken down into pieces on each of which it is a genuine single-valued function. The multiformity arises because of how these pieces fit together. A geometric approach to this leads to the concept of a Riemann surface, discussed informally in Section 14.5.

14.4 Multiform Functions

A simple instance of a multiform function is

$$f(z) = \sqrt{z}$$

If $z = re^{i\theta}$ then we can choose for $\sqrt{z}$ either

$$\sqrt{r}e^{i\theta/2} \quad \text{or} \quad \sqrt{r}e^{i\theta/2+\pi}$$

where $\sqrt{r}$ is real and positive. With the old concept of an analytic function, we must choose one of these arbitrarily, and then $f(z)$ is analytic only if we make a cut in the complex plane. From the present viewpoint, we can do better. We spell the method out in considerable detail – just this once.

We introduce four domains:

$$\begin{aligned} H_1 &= \{z \in \mathbb{C} : \operatorname{re} z > 0\} \\ H_2 &= \{z \in \mathbb{C} : \operatorname{im} z > 0\} \\ H_3 &= \{z \in \mathbb{C} : \operatorname{re} z < 0\} \\ H_4 &= \{z \in \mathbb{C} : \operatorname{im} z < 0\} \end{aligned}$$

which are the open half-planes to the right, top, left, and bottom of the complex plane. Let $z = re^{i\theta}$ where $r > 0, \theta \in [-\pi, \pi]$. Define

$$\begin{array}{lcll} f_1(z) & = & \sqrt{r}e^{i\theta/2} & \text{for} \quad z \in D_1 = H_1 \\ f_2(z) & = & \sqrt{r}e^{i\theta/2} & \text{for} \quad z \in D_2 = H_2 \\ f_3(z) & = & \begin{cases} \sqrt{r}e^{i\theta/2} \\ \sqrt{r}e^{i(\theta/2+\pi)} \end{cases} & \begin{array}{l} \text{for} \quad z \in D_3 = H_3, \operatorname{im} z \geq 0 \\ \text{for} \quad z \in D_3 = H_3, \operatorname{im} z < 0 \end{array} \\ f_4(z) & = & \sqrt{r}e^{i(\theta/2+\pi)} & \text{for} \quad z \in D_4 = H_4 \\ f_5(z) & = & \sqrt{r}e^{i(\theta/2+\pi)} & \text{for} \quad z \in D_5 = H_1 \\ f_6(z) & = & \sqrt{r}e^{i(\theta/2+\pi)} & \text{for} \quad z \in D_6 = H_2 \\ f_7(z) & = & \begin{cases} \sqrt{r}e^{i(\theta/2+\pi)} \\ \sqrt{r}e^{i\theta/2} \end{cases} & \begin{array}{l} \text{for} \quad z \in D_7 = H_3, \operatorname{im} z \geq 0 \\ \text{for} \quad z \in D_7 = H_3, \operatorname{im} z < 0 \end{array} \\ f_8(z) & = & \sqrt{r}e^{i\theta/2} & \text{for} \quad z \in D_8 = H_4 \end{array}$$

Each of these eight functions is analytic in its domain of definition; this is why for f_3 and f_7 we have to change our choice of $\sqrt{z}$ as we cross the imaginary axis, because otherwise our choice $-\pi < \theta \leq \pi$ would introduce a discontinuity. Further, f_{r+1} is a direct analytic continuation of f_r ($r = 1, \ldots, 7$) and f_1 is a direct analytic continuation of f_8. For each $z \in \mathbb{C} \setminus \{0\}$ the values $f_r(z)$, where defined, are one or other of the two possible square roots. Further, for each r, the function $f_r(z)$ takes one of the two values and f_{r+4} takes the other value, interpreting $r + 4$ modulo 8 (and replacing 0 by 8).

Thus we have defined a multiform function, taking two values at each $z \neq 0$. Going round the origin once, from D_1 to D_2 to D_3 to D_4 to $D_5 = D_1$, we reach a different value of $f(z)$ from the original one. However, in this case, going round a second time returns us to the original value. This last effect is a special feature of $\sqrt{z}$, as the next example shows.

14.4.1 The Logarithm as a Multiform Function

Example 14.5. An immensely important multiform function is

$$f(z) = \log z$$

Its multiformity was Euler's great discovery, arising from the Bernoulli–Leibniz controversy mentioned in Chapter 0. In terms of the present discussion, we introduce domains

$$D_{4k+r} = H_r \quad (k \in \mathbb{Z}, r = 1, 2, 3, 4)$$

with the H_r as above. To avoid the kind of two-piece definition that occurred for f_3 and f_7 above, we proceed as follows. For $z \in D_n$ and $z = re^{i\theta}$ where

$$\frac{n-2}{2}\pi < \theta \leq \frac{n}{2}\pi$$

we define

$$f_n(z) = \log r + i\theta$$

Then f_n is analytic on D_n. On D_1,

$$f_1(z) = \operatorname{Log} z$$

the principal value of the logarithm. On D_5,

$$f_5(z) = \operatorname{Log} z + 2\pi i$$

On D_{4k+1},

$$f_{4k+1}(z) = \operatorname{Log} z + 2k\pi i$$

It is not hard to check that f_{s+1} is a direct analytic continuation of f_s from D_{s+1} to D_s for all $s \in \mathbb{Z}$. For each $r = 1, 2, 3, 4$ and $k \in \mathbb{Z}$ the values of $f_{4k+r}(z)$ $(z \in H_r)$ give all the infinitely many possible values

$$\log |z| + (2k\pi + \arg z)i$$

of the logarithm. □

14.4.2 Singularities

We can now give a general definition of a singularity. If no analytic continuation of f can be defined at a point z_0 then we say that z_0 is a *singularity* of the corresponding complete analytic function. We have met several kinds of singularity before: poles, isolated essential singularities, natural boundaries. Removable singularities are not singularities at all according to the new definition, because 'filling in' the missing point is a form of analytic continuation.

For $\sqrt{z}$ and $\log z$ we encounter a new kind of singularity, called a *branch point*. Analytic continuation round such a point leads to a change in value. Notice that for $f(z) = \sqrt{z}$ there is even a natural definition of $f(z)$ at the branch point itself, namely 0, but f is not analytic there. This is not the case for the logarithm.

Multiform functions make an appearance in contour integration whenever different choices of path from z_0 to z_1 leads to different values of the integral

$$F(z_1) = \int_{z_0}^{z_1} f(z)\, dz$$

(as may occur, for instance, if f has a pole in the region between the two paths). Hence quite nice singularities of f, such as poles, give rise to much nastier singularities of F, namely branch points. This can occur even when f is uniform: for instance, $1/z$ is uniform but its integral $\log z$ is multiform, and the pole of f at 0 has become a branch point of F.

14.5 Riemann Surfaces

Riemann was a great geometer, and he invented a geometric way to envisage multiform functions, which is much more intuitive than an equivalence class of function elements. His idea is to replace $\mathbb{C}$ by a more complicated 'Riemann surface'. Roughly speaking, the idea is to 'glue together' the domains of function elements at overlaps where the functions agree. We sketch this idea in general later; our immediate aim is to describe a few key examples.

14.5.1 Riemann Surface for the Logarithm

In the case of the logarithm we can describe the construction informally in the following terms, which should not be subjected to too deep scrutiny of a logic-chopping kind. We are not attempting a rigorous definition at this point; the informal description, though it may sound far-fetched, is in fact capable of being given a rigorous rendering, described at the end of this section.

Consider a collection of copies $\mathbb{C}_k$ of $\mathbb{C}$, one for each $k \in \mathbb{Z}$. Slit each $\mathbb{C}_k$ along the negative real axis from 0 to $-\infty$. For each k join the top left-hand quadrant of $\mathbb{C}_k$ to the bottom left-hand quadrant of $\mathbb{C}_{k+1}$ along the slit. The resulting surface, Figure 14.8, resembles a spiral staircase or a concertina, with the planes stacked on top of each other in order relative to k, and with a continuous spiral path from any $\mathbb{C}_k$ to any other $\mathbb{C}_l$ going up the 'steps' where the slits were joined. These planes $\mathbb{C}_k$ are the *sheets* of the Riemann surface, and it is convenient to imagine the whole collection to be stacked on top of $\mathbb{C}$ as shown in Figure 14.8.

We define the logarithm for points on the Riemann surface by

$$\log|\hat{z}| + (2k\pi + \arg z)\mathrm{i}$$

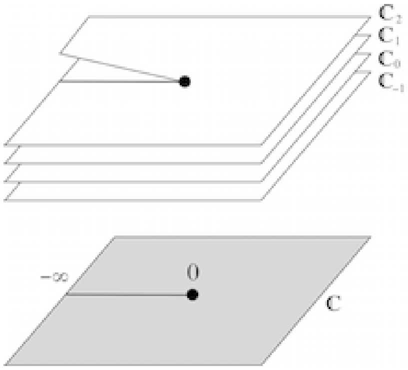

Figure 14.8 Riemann surface for the logarithm.

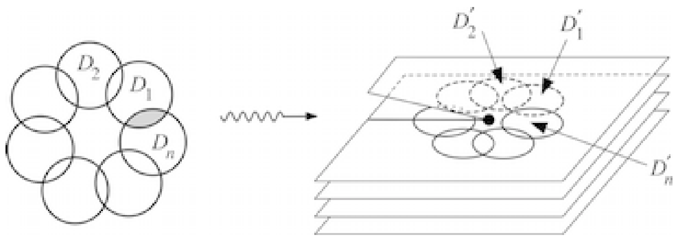

Figure 14.9 A chain of overlapping domains that closes up in $\mathbb{C}$ no longer closes up when it is transferred to the Riemann surface.

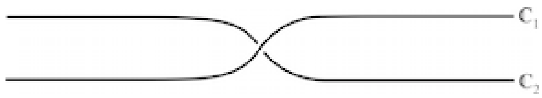

Figure 14.10 Riemann surface for the square root – schematic of how the two sheets join.

where $\hat{z}$ is the point of $\mathbb{C}_k$ lying directly above the point $z \in \mathbb{C}$. This is a *single*-valued function on the Riemann surface, and it is continuous in the sense that the values join up correctly across the cuts. (In a similar sense we can even say it is differentiable: the derivatives also join up correctly across the cuts.)

Now we describe how to relate analytic continuation in $\mathbb{C}$ to the corresponding operation on the Riemann surface. Corresponding to a closed chain of domains $D_1, \ldots, D_n = D_1 \subseteq \mathbb{C}$, we have $D_1', \ldots, D_n'$ on the Riemann surface. In $\mathbb{C}$ the values of the functions $f_1, \ldots, f_n$ agree on the relevant overlaps $D_1 \cap D_2, \ldots, D_{n-1} \cap D_n$, but because of the multiformity they do not agree on D_1 and D_n.

In contrast, on the Riemann surface we find that if we make D_1' and D_2' overlap, then D_2' and D_3', and so on, then by the time we get to D_n' it no longer coincides with D_1' because we have moved one level up the Riemann surface, see Figure 14.9. Hence D_1' and D_n' no longer overlap, and the difference between f_1 and f_n is only to be expected. The multiform nature of analytic continuation is automatically taken care of by the multiplicity of the sheets of the Riemann surface and how they join together. Even the nature of the singularity at 0 is apparent from the geometry of the surface, because here all the levels meet.

14.5.2 Riemann Surface for the Square Root

Similarly for $\sqrt{z}$ we obtain the Riemann surface by taking two copies $\mathbb{C}_1$ and $\mathbb{C}_2$ of $\mathbb{C}$, slit from 0 to $-\infty$. Then join the top left-hand quadrant of $\mathbb{C}_1$ to the bottom left-hand quadrant of $\mathbb{C}_2$ along the slit; and similarly join the top left-hand quadrant of $\mathbb{C}_2$ to the bottom left-hand quadrant of $\mathbb{C}_1$. (To do this in three-dimensional space requires allowing the surface to intersect itself, but there is no conceptual problem here, see Figure 14.10.)

Again the phenomena noted above are clear from the geometry of the Riemann surface: the two values of $f(z)$, the fact that going round the the branch point at the origin once changes the value but going round twice does not.

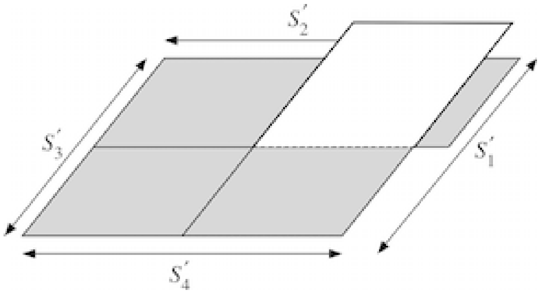

Figure 14.11 Gluing successive domains to form the Riemann surface of the logarithm.

14.5.3 Constructing a General Riemann Surface by Gluing

Obviously we cannot rely on such *ad hoc* methods to construct the Riemann surface in general. We can approach the general method through the same example of the logarithm, but building up the surface in a different way. Consider the sequence of half-planes D_s and functions f_s used for analytic continuation above. Suppose we take half-planes D_s' corresponding to the D_s, but make them pairwise disjoint. (The D_s are not disjoint; for instance $D_1 = D_5$ and so on.) Then we 'glue' the D_s' together in the following way: D_s and D_{s+1} overlap on a quadrant in which $f_s = f_{s+1}$, so we glue together the corresponding quadrants D_s' and D_{s+1}' where they overlap. Then D_2' glues on top of D_1' at the top right-hand quadrant; D_3' glues on to the top left of D_2'; D_4' glues on to the bottom left of D_3'; and D_5' glues on to the bottom right of D_4'. at this point D_5' is lying directly over D_1', but we do not glue them together since f_1 and f_5 are different, Figure 14.11. Continuing with $D_6', D_7', \ldots$ (and in the downward direction $D_0', D_{-1}', D_{-2}', \ldots$) we build up the spiral staircase Riemann surface again.

The general method for constructing a Riemann surface of a complete analytic function F follows the same lines. Recall that F is an equivalence class of pairs (f, D) where f is analytic on a domain D and $(f_1, D_1) \sim (f_2, D_2)$ if f_1, f_2 are equal on the non-empty set $D_1 \cap D_2$. We take *disjoint* copies D_λ' of all the domains D_λ occurring in the pairs (f_λ, D_λ) that belong to F. If $D_\lambda \cap D_\mu \neq \emptyset$ and $f_\lambda = f_\mu$ on it, we glue D_λ' to D_μ' at the points corresponding to the overlap $D_\lambda \cap D_\mu$.

We make this description more rigorous. First, the question of 'disjoint copies' D_λ' of D_λ. One way to achieve this is to define

$$D_\lambda' = D_\lambda \times \{\lambda\}$$

Then if $\lambda \neq \mu$ it is clear that $D_\lambda' \cap D_\mu' = \emptyset$, and there is a natural bijection

$$j_\lambda : D_\lambda \to D_\lambda'$$

defined by

$$j_\lambda(s) = (s, \lambda) \quad (s \in D_\lambda)$$

Now for that 'gluing'. This is accomplished by a simple trick, using yet another equivalence relation, but it requires a certain amount of sophistication to appreciate that it does what is required.

Let

$$(z_\lambda, \lambda) \approx (z_\mu, \mu)$$

for $z_\lambda \in D_\lambda, z_\mu \in D_\mu$, if:

(i) $z_\lambda = z_\mu$
(ii) $f_\lambda(z_\lambda) = f_\mu(z_\mu)$

Then the set of equivalence classes under $\approx$ of points in the union of all the D'_λ is defined to be the Riemann surface of F. The equivalence relation $\approx$ acts as the glue. This definition is quite elegant, but might not be immediately appealing.

The advantage of the Riemann surface construction is that we can now think of a multiform complex analytic function as a genuine function, defined not on $\mathbb{C}$ but on the appropriate Riemann surface, rather than as an equivalence class of function elements. The Riemann surface gives a 'global' view of the function, instead of chopping it up into 'local' pieces. The last few sections of this chapter revisit previously discussed material to explore some of the insights that can be gained in this way.

14.6 Complex Powers

So far we have considered powers z^a of a complex number z only for rational a, where $z^{p/q}$ is a qth root of z^p. We now define it for arbitrary $a \in \mathbb{C}$. We would like to do this so that the laws $z^{a+b} = z^a z^b$ and $(z^a)^b = z^{ab}$, $(zw)^a = z^a w^a$ remain valid. It turns out that we can 'almost' do this: however, z^a is in general multiform, and the formulas hold only if we choose appropriate values. Even when a is rational this problem already arises.

A natural approach is to write

$$z = re^{i\theta} = e^{\log r + i\theta} \tag{14.7}$$

where $\log r \in \mathbb{R}$, and let $a = \alpha + i\beta$. On the assumption that the above laws hold, we obtain:

$$\begin{aligned} z^a &= (re^{i\theta})^a \\ &= e^{a(\log r + i\theta)} \\ &= e^{(\alpha+i\beta)(\log r + i\theta)} \\ &= e^{\alpha \log r - \beta\theta} e^{i(\beta \log r + \alpha\theta)} \end{aligned} \tag{14.8}$$

We therefore define z^a by (14.8). This makes sense for all $z \neq 0$, and for all $a \in \mathbb{C}$. It amounts to requiring that

$$z^a = e^{a \log z} \tag{14.9}$$

which is an equally natural way to define z^a. It follows that in any domain for which we may define a unique branch of log (such as the cut plane $\mathbb{C}_\rho$ for any ρ) we may also define z^a as a single-valued differentiable function.

In general, however, since log is multiform, so is z^a. To see what the possibilities are, choose a particular value θ_0 for θ (the obvious choice is the principal value of $\arg z$). Then the possible θ that make (14.7) hold are of the form

$$\theta = \theta_0 + 2n\pi \quad (n \in \mathbb{Z}) \tag{14.10}$$

Write

$$(z^a)_n = \mathrm{e}^{a(\log r + \mathrm{i}(\theta_0 + 2n\pi))}$$

which is the 'nth branch' of z^a, obtained by substituting θ from (14.10) into (14.8). The question is now: how does $(z^a)_n$ depend on n?

From (14.8) we have

$$(z^a)_n = (\mathrm{e}^{-2n\pi\beta}\mathrm{e}^{2n\pi\mathrm{i}\alpha})(z^a)_0 \tag{14.11}$$

where $(z^a)_0$ is the 0th branch. Using (14.11) we can answer the above question. There are three cases:

(a) If $\beta \neq 0$ then z^a has infinitely many values, one for each n, because

$$|(z^a)_n/(z^a)_0| = (\mathrm{e}^{-2n\pi\beta})^n$$

which takes distinct values for distinct n. The Riemann surface for z^a in this case has the same form as that for the logarithm: an infinite 'spiral staircase' on whose nth layer the function is given by the branch $(z^a)_n$.

(b) If $\beta \neq 0$, so $a = \alpha \in \mathbb{R}$, and if further α is irrational, then $z^a = z^\alpha$ is again infinitely many-valued, with distinct branch values for distinct n. For if $(z^\alpha)_m = (z^\alpha)_n$ then $\mathrm{e}^{2m\pi\mathrm{i}\alpha} = \mathrm{e}^{2n\pi\mathrm{i}\alpha}$, so $\mathrm{e}^{2\pi\mathrm{i}(m-n)\alpha} = 1$. This implies that $(m-n)\alpha \in \mathbb{Z}$ by Section 5.7, which implies that $m = n$ since α is irrational. The Riemann surface for z^α is the same as for case (a), but unlike case (a), the *moduli* of distinct branches are equal. Only the arguments vary.

(c) If $\beta \neq 0$, so $a = \alpha \in \mathbb{R}$, and if further α is rational, then we can write $\alpha = p/q$ in lowest terms, where $p, q \in Z$. Now $z^\alpha = \sqrt[q]{z^p}$.

Following the discussion in (b), two branches give the same values if and only if

$$(m-n)\frac{p}{q} \in \mathbb{Z}$$

which happens if and only if q divides $m - n$, because p and q have no common factor greater than 1.

It follows that the value of $(z^\alpha)_n$ depends only on $n \bmod (q)$. The branches

$$(z^\alpha)_0, (z^\alpha)_1, \ldots, (z^\alpha)_{q-1}$$

are distinct, but the qth branch repeats:

$$(z^\alpha)_q = (z^\alpha)_0$$

and repetitions continue thereafter.

The Riemann surface, of course, is a q-sheeted spiral, with its top sheet glued to its bottom one, as in our discussion of $z^{1/2}$ but with q sheets instead of 2. Figure 14.12 illustrates the case $q = 5$.

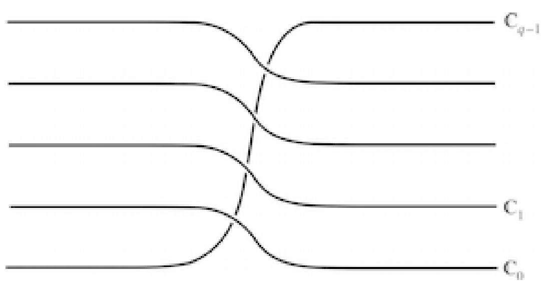

Figure 14.12 Section through the Riemann surface for $z^{p/q}$ when $q = 5$, showing how the layers glue together.

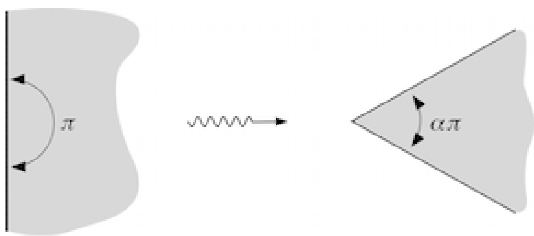

Figure 14.13 Conformal map of a half-plane to a wedge.

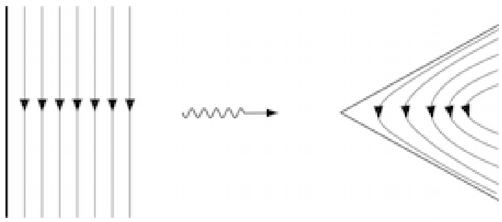

Figure 14.14 How the conformal map of a half-plane to a wedge transforms streamlines.

14.7 Conformal Maps Using Multiform Functions

Conformal mapping with multiform functions requires careful attention to how values are specified; again, the simplest way to keep track is to put everything on a Riemann surface. Here we consider only one case, of practical importance: the map $z \to z^\alpha$ for real $\alpha \geq 0$. We assume the domain is the cut plane $\mathbb{C}_\pi$, on which the map may be defined to be single-valued.

The great virtue of this map is that it transforms a half-plane into a wedge, as in Figure 14.13. The vertex angle of the wedge is $\alpha\pi$, so we have to assume $\alpha < 2$ to prevent the image overlapping itself. The map is conformal except at the origin.

For example, uniform fluid flow, with streamlines $\mathrm{re}\, z = \text{constant}$, transforms to the flow round a corner of angle $\alpha\pi$, as in Figure 14.14.

If we take $\alpha = 1/2$, the corner is right-angled. Then $z = x + \mathrm{i}y$ and $z^{1/2} = w = u + \mathrm{i}v$. The flow-lines are $x = \text{constant}$. Now $z = w^2$ so $x = u^2 - v^2$, $y = 2uv$. The streamlines in the w-plane are given by $u^2 - v^2 = \text{constant}$; they are branches of rectangular hyperbolas.

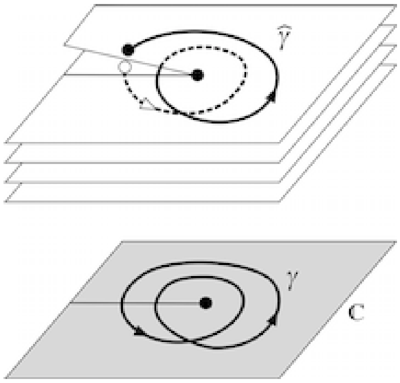

Figure 14.15 Transferring a contour to a Riemann surface.

In combination with other 'standard' conformal maps, z^α provides a useful weapon for the applied mathematician.

14.8 Contour Integration of Multiform Functions

It is conventional to interpret a contour integral $\int_\gamma f$ of a multiform f by choosing the values of $f(z)$ so that they vary continuously along the contour. (This still leaves an arbitrary initial choice to be made, which must also be specified.) A more civilised approach is to break the contour γ into a sum $\gamma = \gamma_1 + \cdots + \gamma_n$ such that:

(i) Each γ_j lies inside a domain D_j on which f may be defined as a single-valued function.
(ii) On each D_j we choose a branch of f that makes the values agree where γ_j and γ_{j+1} join.

In terms of Riemann surfaces, we can interpret this process as the definition of a contour integral *when the contour lies on the Riemann surface*. The way the contour wanders up and down the staircases on the surface automatically takes care of the choice of values of the multiform function $f(z)$, see Figure 14.15. In practice this is all more straightforward than it may sound, and the computations are no harder than the uniform case – provided you use your head.

Example 14.6. Let γ be the contour

$$\gamma(t) = (1+t)e^{it} \quad (t \in [0, 6\pi])$$

Find

$$\int_\gamma z^{1/5} dz$$

Method 1 (stupid): Defining $z^{1/5}$ on the contour by $(\gamma(t))^{1/5} = (1+t)^{1/5}e^{it/5}$ makes it vary continuously with t. By Theorem 6.6

Table 14.1 Choice of θ.

domain	θ chosen in interval
D_1	$[-\pi/4, 5\pi/4]$
D_2	$[3\pi/4, 9\pi/4]$
D_3	$[7\pi/4, 13\pi/4]$
D_4	$[11\pi/4, 17\pi/4]$
D_5	$[15\pi/4, 21\pi/4]$
D_6	$[19\pi/4, 25\pi/4]$

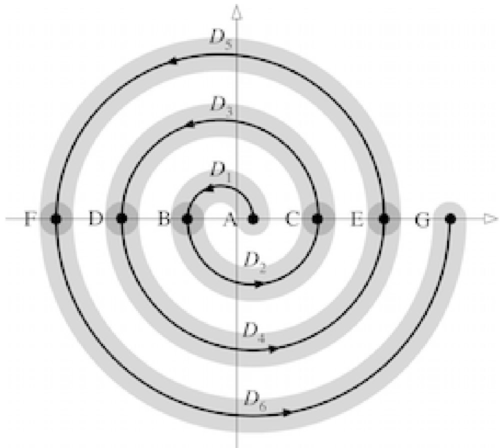

Figure 14.16 Six overlapping domains.

$$\int_\gamma f(z)\mathrm{d}z = \int_0^{6\pi} (f(\gamma(t))\gamma'(t)\,\mathrm{d}t$$
$$= \int_0^{6\pi} (1+t)^{1/5}\mathrm{e}^{\mathrm{i}t/5}(\mathrm{i}\mathrm{e}^{\mathrm{i}t} + \mathrm{i}t\mathrm{e}^{\mathrm{i}t} + \mathrm{e}^{\mathrm{i}t})\,\mathrm{d}t$$

which leads the hopeful calculator along some fascinating byways inhabited by integrals such as

$$\int_0^{6\pi} t(1+t)^{1/5}\cos(6t/5)\mathrm{d}t$$

A wise person knows when to cut their losses.

Method 2 (sound but pedestrian): Cover the contour γ by domains in which $z^{1/5}$ may be given a unique value for each z, which is differentiable. For example, we can find six domains $D_1, \ldots, D_6$, one for each of the intervals AB, BC, CD, DE, EF, FG in Figure 14.16.

To make the values match on overlaps we choose the definition of $z^{1/5}$ as follows. Write $z = r\mathrm{e}^{\mathrm{i}\theta}$ where the choice of θ is as in Table 14.1.

On each domain choose the branch of $z^{1/5}$ given by $r^{1/5}\mathrm{e}^{\mathrm{i}\theta/5}$ with θ in the range stated in the table. These choices make θ agree on overlaps.

Break up the integral as

$$\int_\gamma = \int_{\mathrm{AB}} + \int_{\mathrm{BC}} + \int_{\mathrm{CD}} + \int_{\mathrm{DE}} + \int_{\mathrm{EF}} + \int_{\mathrm{FG}}$$

In each domain D_j the function $z^{1/5}$ has an antiderivative $\frac{5}{6}z^{6/5}$ where $z^{6/5}$ is defined using the corresponding branch $z^{6/5} = r^{6/5}\mathrm{e}^{6\mathrm{i}\theta/5}$. By Theorem 6.33, if PQ is any of the arcs AB, BC, and so on, then

$$\int_{\mathrm{PQ}} z^{1/5}\mathrm{d}z = \tfrac{5}{6}\mathrm{Q}^{6/5} - \tfrac{5}{6}\mathrm{P}^{6/5}$$

When we add the six expressions of this type, all terms but two cancel, so the integral is

$$\tfrac{5}{6}\mathrm{G}^{6/5} - \tfrac{5}{6}\mathrm{A}^{6/5}$$

Here $\mathrm{A} = \gamma(0) = 1, \mathrm{G} = \gamma(6\pi) = (1 + 6\pi)\mathrm{e}^{6\mathrm{i}\pi}$. Evaluate the $6/5$th powers according to the above prescription:

$$\begin{aligned} \mathrm{A}^{6/5} &= (1 \cdot \mathrm{e}^{\mathrm{i}0})^{6/5} = 1 \\ \mathrm{G}^{6/5} &= (1 + 6\pi)\mathrm{e}^{6\pi\mathrm{i}})^{6/5} = (1 + 6\pi)^{6/5}\mathrm{e}^{36\mathrm{i}\pi/5} = -(1 + 6\pi)^{6/5}\mathrm{e}^{\mathrm{i}\pi/5} \end{aligned}$$

Therefore the integral is equal to

$$-\tfrac{5}{6}[(1 + 6\pi)^{6/5}\mathrm{e}^{\mathrm{i}\pi/5} + 1]$$

Method 3 (slicker): In terms of the parameter t we can define the required branch of $z^{1/5}$ to vary continuously *along the contour* γ if we put

$$z^{1/5} = (\gamma(t))^{1/5} = (1 + t)^{1/5}\mathrm{e}^{\mathrm{i}t/5}$$

There is a local antiderivative $\frac{5}{6}z^{6/5}$, whose value varies continuously along the path γ if we choose branches so that

$$(\gamma(t))^{6/5} = (1 + t)^{6/5}\mathrm{e}^{6\mathrm{i}t/5}$$

Now pave γ by a finite number of domains, on each of which $z^{1/5}$ may be rendered single-valued, choosing the branches to agree with those already chosen on the path. As above, the integral may be expanded as a sum of terms of the form

$$\tfrac{5}{6}[(1 + t_{j+1})^{6/5}\mathrm{e}^{6\mathrm{i}t_{j+1}/5} - (1 + t_j)^{6/5}\mathrm{e}^{6\mathrm{i}t_j/5}]$$

where the t_j subdivide the interval over which t runs. All but two terms cancel, leaving the same answer as before.

The advantage of this method is that by choosing branches in an obvious way for $\gamma(t)$ we may leave the prescription of the domains, and the branches on those domains, implicit.

Method 4 (smart): Let the Riemann surface do the work. The above analysis easily generalises to give a version of Theorem 6.33 for multiform functions, integrated along a contour *in the Riemann surface*. Given a global antiderivative F for f on the Riemann surface, we have

$$\int_\gamma f = F(z_1) - F(z_0)$$

where z_0 is the initial point, z_1 the final point, and the branches are chosen with reference to the end points of γ *on the surface*, as in Figure 14.15. This gives the result almost immediately on noting that γ winds three times anticlockwise, so the $6/5$th power winds $3 \cdot 6/5 = 18/5$ times anticlockwise, giving for $F(z_1)$ the argument $2\pi \cdot 18/5 - 36\pi/5$, if we choose argument 0 for z_0. □

We leave it as an exercise to formulate and prove generalisations of Cauchy's Theorem for contour integrals on the Riemann surface of a multiform function. For those who prefer not to use the insight that Riemann surfaces provide, however, we recommend Method 3 above as a relatively direct and simple technique.

Example 14.7. Example 14.6 is of course a bit artificial, and its main interest is pedagogical. A more typical example of integrals likely to be encountered in practice is

$$\int_0^\infty \frac{x^a}{1+x^2}\mathrm{d}x \quad (a \in \mathbb{R}, 0 < a < 1)$$

Note that the restrictions on a make this integral converge, so the question is sensible. We begin with the complex function

$$f(z) = z^a(1+z^2)^{-1}$$

which is differentiable for all $z \neq 0, \mathrm{i}, -\mathrm{i}$, and is multiform for all a in the given range.

Make a cut along the positive real axis and work in the cut plane $\mathbb{C}_0$ (in the notation of Section 2.2). On this domain z^a may be rendered single-valued: set $(r\mathrm{e}^{\mathrm{i}\theta})^a = r^a\mathrm{e}^{a\mathrm{i}\theta}$ where $\theta \in [0, 2\pi]$. Now integrate $f(z)$ along the contour γ in Figure 14.17. This runs from the real point ρ to the real point R, then once anticlockwise round the circle of radius R, then back to ρ, then once clockwise round the circle radius ρ.

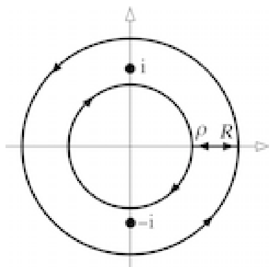

Figure 14.17 Contour for Example 14.7.

By Cauchy's Residue Theorem, Theorem 12.3,

$$\int_\gamma f(z)\mathrm{d}z = 2\pi\mathrm{i}\,(\Sigma \text{ residues inside } \gamma)$$

We intend to let $\rho \to 0$ and $R \to \infty$, so we may as well assume that $\rho < 1$ and $R > 1$. Then the singularities inside γ are $z = \pm\mathrm{i}$. We calculate the residues at these points as follows:

At $z = \mathrm{i}$ the residue is

$$\alpha = \lim_{z\to\mathrm{i}} z^a/(z+\mathrm{i}) = \frac{1}{2\pi\mathrm{i}}\mathrm{e}^{\mathrm{i}a\pi/2}$$

At $z = -\mathrm{i}$ the residue is

$$\beta = \lim_{z\to-\mathrm{i}} z^a/(z-\mathrm{i}) = \frac{1}{2\pi\mathrm{i}}\mathrm{e}^{\mathrm{i}a3\pi/2}$$

Let the circle of radius R be γ_1, and that of radius ρ be γ_2. Then Cauchy's Residue Theorem implies that

$$\int_\rho^R \frac{x^a}{1+x^2}\mathrm{d}x + \int_{\gamma_1} \frac{z^a}{1+z^2}\mathrm{d}z + \int_R^\rho \frac{(x\mathrm{e}^{2\pi\mathrm{i}})^a}{1+x^2}\mathrm{d}x + \int_{-\gamma_2} \frac{z^a}{1+z^2}\mathrm{d}z = 2\pi\mathrm{i}(\alpha+\beta) \quad (14.12)$$

Note that in the third integral we must choose the branch of z^a corresponding to argument 2π, to retain continuity. Now let $\rho \to 0$ and $R \to \infty$. Easy estimates show that the second and fourth integrals tend to zero. The first tends to

$$\int_0^\infty \frac{x^a}{1+x^2}\mathrm{d}x$$

and the third to

$$-\mathrm{e}^{2\pi\mathrm{i}a}\int_0^\infty \frac{x^a}{1+x^2}\mathrm{d}x$$

Therefore

$$(1-\mathrm{e}^{2\pi\mathrm{i}a})\int_0^\infty \frac{x^a}{1+x^2}\mathrm{d}x = 2\pi\mathrm{i}\left(\frac{1}{2\pi\mathrm{i}}\mathrm{e}^{\mathrm{i}a\pi/2} - \frac{1}{2\pi\mathrm{i}}\mathrm{e}^{\mathrm{i}a3\pi/2}\right)$$

After some manipulation, this leads to

$$\int_0^\infty \frac{x^a}{1+x^2}\mathrm{d}x = \frac{\pi}{2\cos(a\pi/2)} \qquad \square$$

Yes, but ... did you spot the nifty footwork? The chosen contour does not actually lie inside the domain $\mathbb{C}_0$. Nonetheless, the result is correct. There are several ways to justify the answer rigorously. They include:

(1) Replace γ by a contour that starts just above ρ, say at $\rho + \mathrm{i}\varepsilon$, runs horizontally to just above R at $R + \mathrm{i}\varepsilon$, circles γ_1 to just below R at $R - \mathrm{i}\varepsilon$, runs back to below

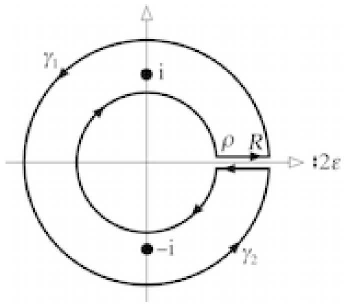

Figure 14.18 Modified contour remains inside $\mathbb{C}_0$.

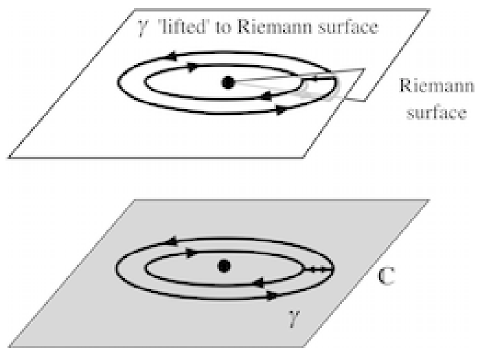

Figure 14.19 Lifting γ to the Riemann surface.

ρ at $\rho - i\varepsilon$, and the goes back round γ_2 as in Figure 14.18. Now, as $\rho \to 0$ and $R \to \infty$, make the width 2ε of the channel tend to zero and use continuity arguments.

(2) The choice of $\mathbb{C}_0$ causes problems, so reject it. Work in two overlapping domains, say $\mathbb{C}_{-\pi/4}$ and $\mathbb{C}_{\pi/4}$, and switch between them when defining the second, third, and fourth integrals in (14.12).

(3) Put it all on a Riemann surface. The contour γ then winds round the staircase and up one step, then winds back down one step to get back to the start.

Everything generalises to such contours, because they are obviously boundaries of rectangles in the Riemann surface, Figure 14.19. This argument is an example of the ‘homology’ version of Cauchy’s Theorem, when generalised to Riemann surfaces. We discuss homology in Chapter 16, but do not extend the ideas to Riemann surfaces since this would take us too far afield.

These three approaches justify the computational method rigorously. Of course, when performing the calculations, it is not necessary to trot out such a justification every time – provided you understand how it would go. Everything works fine as long as you follow the golden rule: make the multiform function vary *continuously* as you walk round the contour.

14.9 Exercises

1. Define three power series by

$$a(z) = 1 + z + z^2 + z^3 + \cdots = \sum_{n=0}^{\infty} z^n$$

$$b(z) = \mathrm{i} - (z - \mathrm{i} - 1) - \mathrm{i}(z - \mathrm{i} - 1)^2 + (z - \mathrm{i} - 1)^3 + \cdots = \sum_{n=0}^{\infty} \mathrm{i}^{n+1}(z - \mathrm{i} - 1)^n$$

$$c(z) = -1 + (z - 2) - (z - 2)^2 + \cdots = \sum_{n=0}^{\infty} (-1)^{n+1}(z - 2)^n$$

Find their discs of convergence and sketch them. Prove that $a(z) = b(z)$ on the overlap of their discs of convergence, and similarly that $b(z) = c(z)$ on the overlap of their discs of convergence. Do the discs of convergence of $a(z)$ and $c(z)$ intersect?

2. Let $f(z) = \sum_{n=0}^{\infty} a_n(z - z_1)^n$ and $g(z) = \sum_{n=0}^{\infty} b_n(z - z_2)^n$ be two power series, and assume that their discs of convergence have non-empty intersection. Prove that g is a direct analytic continuation of f if and only if there exists z_0 in both discs of convergence such that for $m = 0, 1, 2, 3, \ldots$ the equations

$$\sum_{n=0}^{m} n(n-1)\cdots(n-m)[a_n(z_0 - z_1)^{n-m} - b_n(z_0 - z_2)^{n-m}] = 0$$

hold.

3. A certain function has singularities at precisely the points z such that $\mathrm{re}\, z$ and $\mathrm{im}\, z$ are integers. You are given its Taylor expansion around the point $\frac{1}{2} + \frac{1}{2}\mathrm{i}$. What is the smallest number of stages needed to find an indirect analytic continuation, by power series, to a domain that contains $\frac{7}{2} + \frac{5}{2}\mathrm{i}$?

4. Show that the functions defined by
(i) $f(z) = 1 + z + z^2 + z^4 + z^8 + \cdots + z^{2^n} + \cdots$
(ii) $g(z) = 1 + z + z^3 + z^9 + z^{27} + \cdots + z^{3^n} + \cdots$
have natural boundaries at $|z| = 1$.

5. Suppose that $f(z) = \sum a_n z^n$ has radius of convergence 1. Set

$$z = \frac{w}{1+w} = w - w^2 + w^3 - w^4 + \cdots$$

and transform $f(z)$ to a power series in w, say $F(w) = \sum b_n w^n$. Prove that this latter power series has radius of convergence $\geq \frac{1}{2}$, and that if -1 is a singular point of f, the radius of convergence is exactly $\frac{1}{2}$.

6. Show that the series

$$\sum_{n=1}^{\infty} [(1 - z^{n+1})^{-1} - (1 - z^n)^{-1}]$$

converges when $|z| > 1$ or $|z| < 1$, but that the two functions so represented are *not* analytic continuations of each other.

7. Suppose that f and g are defined, and have no singularities, on the whole of $\mathbb{C}$. Define

$$\Phi(z) = \sum_{n=1}^{\infty} \left(\frac{1 - z^n}{1 + z^n} - \frac{1 - z^{n-1}}{1 + z^{n-1}} \right)$$

Show that

$$\tfrac{1}{2}(f(z) + g(z)) + \tfrac{1}{2}\Phi(z)(f(z) - g(z))$$

is equal to $f(z)$ when $|z| < 1$, but to $g(z)$ when $|z| > 1$.

8. Let $f(z)$ be the (multiform) function $z\sqrt{z}$. Show that at $z = 0$ there exists a first derivative that is the same for all branches, but a finite second derivative does not exist. What about $z^2 \log z$?

9. Describe the Riemann surfaces of the multiform functions:

(i) $\sqrt[7]{z + 43}$
(ii) $\sqrt{1 - z^3}$
(iii) $\cos^{-1} z$
(iv) $\tan^{-1} z$

10. Describe the Riemann surface of

$$[(z-1)(z-2)^{-2}]^{1/3} + (z-3)^{1/2}$$

11. Let $\omega \in \mathbb{C}$. Show that the values of 1^{ω} form a subgroup U_{ω} of the multiplicative group of non-zero complex numbers. Show that U_{ω} is cyclic of order q if ω is a rational number p/q in lowest terms; otherwise, U_{ω} is infinite cyclic.

Show that U_{ω} is contained in the unit circle if and only if ω is real; lies on the positive real axis if and only if $\mathrm{re}\,\omega$ is an integer; and otherwise lies on a logarithmic spiral parametrised by $t \in \mathbb{R}$ in the form $\mathrm{e}^{\alpha t}$ for some fixed complex number α.

12. Show that the function

$$f(z) = \mathrm{e}^{-\mathrm{i}\pi/8}\sqrt{z - \mathrm{i}}$$

defines a conformal map from the domain in Figure 14.20 (left) to the upper half-plane.

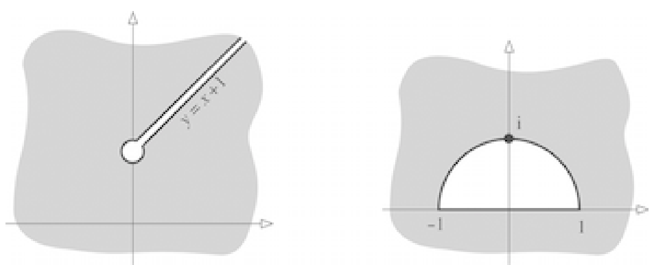

Figure 14.20 *Left*: Domain for Exercise 12. *Right*: Domain for Exercise 13.

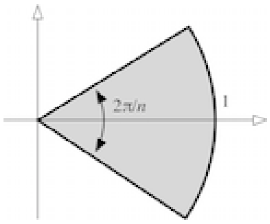

Figure 14.21 Sector occurring in Exercise 14.

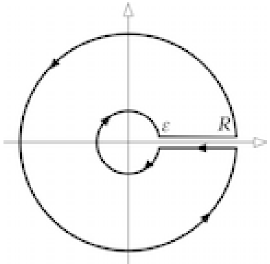

Figure 14.22 Contour for Exercise 16.

13. Show that the function

$$f(z) = e^{-i\pi/3}\left(\frac{z+1}{z-1}\right)^{2/3}$$

defines a conformal map from the domain in Figure 14.20 (right) to the upper half-plane.

14. Find the image of the sector $-\pi/n < \arg z < \pi/n$, $|z| < 1$ (Figure 14.21) under the conformal transformation

$$f(z) = z(1+z^n)^{2/n}$$

where n is a positive integer.

15. Let $\gamma(t) = e^{it}$, $(t \in [0, 4\pi])$. Calculate $\int_\gamma \sqrt{z}\,dz$ where at $t = 0$ we take $\sqrt{1} = 1$.

16. Using the contour in Figure 14.22 show that

$$\int_0^\infty \frac{x^{-k}}{1+x}dx = \pi \operatorname{cosec} k\pi \ (0 < k < 1)$$

17. Show by contour integration that

$$\int_0^1 \frac{dx}{(1+ax^2)\sqrt{1-x^2}} = \frac{\pi}{2\sqrt{1+a}} \quad (a > 0)$$

18. Using the contour of Exercise 16, show that

$$\int_0^\infty \frac{x^a}{(1+x^2)^2}dx = \frac{\pi(1-a)}{4\cos\frac{1}{2}a} \quad (-1 < a < 3)$$

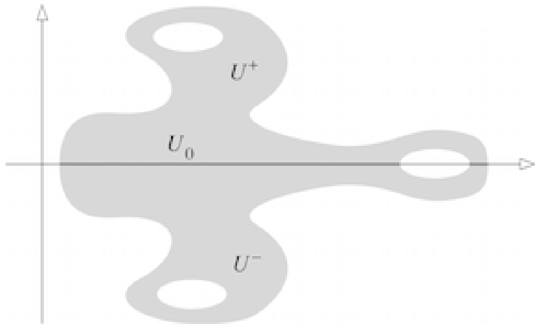

Figure 14.23 Schwartz Reflection Principle.

19. Let $\gamma(t) = (1 + \frac{1}{2}\cos t + \mathrm{i}\sin t)^5$, $(t \in [0, 2\pi])$. Find

$$\int_\gamma z^{99}\log z - \sqrt{z} + (z - \tfrac{1}{4})^{5/17}\mathrm{d}z$$

20. Describe the Riemann surface of the function

$$f(z) = \sqrt{z + z^2 + z^4 + z^8 + \cdots + z^{2^n} + \cdots}$$

21. *Schwartz Reflection Principle*. Let U be a domain in $\mathbb{C}$ that is symmetric about the real axis (that is, if $z \in U$ then $\bar{z} \in U$). Let

$$U^+ = \{z \in U : \mathrm{im}\, z > 0\}$$
$$U^- = \{z \in U : \mathrm{im}\, z < 0\}$$
$$U^0 = \{z \in U : \mathrm{im}\, z = 0\}$$

as in Figure 14.23. Suppose that

$$f : U^+ \cup U^0 \to \mathbb{C}$$

is continuous, analytic on U^+, and takes real values for $z \in U^0$. Then there is an analytic function $F : U \to \mathbb{C}$ such that $F(z) = f(z)$ for all $z \in U^+ \cup U^0$.

(Hint: define $F(z) = \overline{f(\bar{z})}$ for $z \in U^-$, and use Morera's Theorem.)

15 Infinitesimals in Real and Complex Analysis

Historically, calculus began with the study of variable quantities represented as curves in the plane. These variables could become 'arbitrarily small' or 'infinitesimal'. They were used to compute derivatives and integrals, and to formulate differential equations modelling systems that change dynamically. However, the precise meaning of an infinitesimal caused controversy. At least two different concepts existed: a 'number' (of some kind) that is smaller than any non-zero number, or a variable number that can become smaller than any positive number. The two were often confused. Moreover, it is not easy to formalise the idea of a 'variable number'. (If it varies, how can it be a number?)

Because of these difficulties, infinitesimals, in either sense, were eventually displaced by the logical epsilon-delta formulation of Weierstrass in the late nineteenth century. The real numbers form a 'complete ordered field', and therefore do not contain infinitesimals, so infinitesimals were banished from formal mathematical analysis.

Meanwhile, applied mathematicians and theoretical physicists continued to model the real world using 'arbitrarily small' quantities. The familiar formula $T = 2\pi\sqrt{L/g}$ for the period of a pendulum, derived using the approximation of simple harmonic motion, is a familiar example. Usually described as 'valid for small amplitude swings', this formula is actually incorrect for any non-zero amplitude. But is is also a very good approximation for sufficiently small amplitudes, and more accurate formulas cannot be derived or expressed using elementary functions. This procedure is typically justified by letting the approximations concerned become arbitrarily small – the main traditional view of an infinitesimal. Thus two parallel but apparently conflicting approaches to calculus arose: one with, and one without, infinitesimals. Moreover, the tacit assumptions about infinitesimals applied to two different concepts.

In this chapter we reconcile these views, offering a formal approach to infinitesimals that lets them be manipulated algebraically. Real infinitesimals can be visualised on an extended number line, and their complex analogues can be visualised in an extended complex plane. Infinitesimals are made visible to the human eye by a process of magnification, which is defined algebraically and has a natural visual interpretation. The formal representation includes the idea of an infinitesimal quantity as a process that tends to zero, and also as an element in an ordered field extension K of the reals, and in the corresponding extension of $\mathbb{C}$. We prove a simple structure theorem for such an extended system that any finite element x in the extended system (meaning $|x| < r$ for

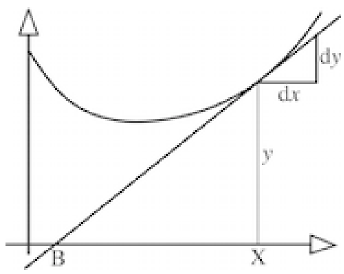

Figure 15.1 Leibniz's view of dx, dy as components of the tangent.

some real number r) is uniquely of the form $x+h$, where x is a real or complex number and h is infinitesimal or zero.

This approach provides a formal theory that encompasses both the standard epsilon-delta approach to analysis in pure mathematics, and the 'applied' approach in which infinitesimals are thought of as variables that tend to zero.

15.1 Infinitesimals

Leibniz, Newton, and their contemporaries used various imaginative devices to deal with differentiation and integration. Leibniz's first publication on calculus specified the derivative $\mathrm{d}y/\mathrm{d}x$ as the quotient of two finite numbers dy and dx. Figure 15.1 illustrates his procedure. Suppose that the tangent to a point P on the curve mets the horizontal axis at B. For any finite horizontal displacement dx, let dy be the corresponding vertical change of the tangent. Then the ratio $\mathrm{d}y/\mathrm{d}x$ is the same as y/BX.

In this interpretation, dx and dy do not represent infinitesimals. They represent finite components of the tangent vector. When the slope of the tangent is m, we can take any finite number dx to be the horizontal component, and the vertical component is then

$$\mathrm{d}y = m\,\mathrm{d}x$$

The slope of the tangent to the curve at P is the quotient $\mathrm{d}y/\mathrm{d}x$.

The devil lies in the detail. The definition makes sense only when the slope of the tangent is known, and to do that, it must be calculated. This is where infinitesimals came in. Leibniz performed the calculation by imagining the graph to be a polygon with an infinite number of infinitesimally small sides. The tangent at a point was the prolongation of one of these sides, as in Figure 15.2, where the lengths of the sides are imagined to be arbitrarily small.

He also conceived of different degrees of infinitely small or infinitely large quantities, by comparing the size of a ball to that of the Earth, or to the distance to the fixed stars. This led him to imagine infinitesimals of different orders of size. An infinitesimal of second order is infinitely smaller than an infinitesimal of first order, and so on. The reciprocal of an infinitesimal of a given order is an infinite element of the same order. For instance, if v is a first order infinitesimal, then v^2 is a second order infinitesimal and $1/v^2$ is an infinite element of second order. Leibniz's ideas were natural, in the practical

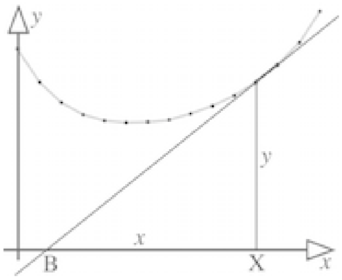

Figure 15.2 The tangent as the extension of an infinitesimal side.

sense that he imagined arbitrarily small or large quantities, and in the theoretical sense that they can be manipulated symbolically.

Their logical status was less clear, and disputed. The philosopher George Berkeley, ridiculed infinitesimals, claiming that it was beyond common sense to imagine quantities smaller than any sensible quantity, and even more nonsensical to imagine quantities that are smaller still. He referred to infinitesimals sarcastically as 'ghosts of departed quantities', neatly confusing the two different interpretations. For generations, mathematicians disputed the precise meaning of this elusive concept, seeking to circumvent the logical difficulties.

Cauchy succeeded, but his ideas were widely misunderstood, even today. He formalised the 'variable quantity' approach, defining an infinitesimal as a sequence $\alpha = (a_n)$ that tends to zero. He could then calculate $f(x+\alpha)$ for infinitesimal α as the sequence $(f(x+a_n))$. He also formulated the notion of continuity for a function in such terms: $f(x)$ is continuous if, for any infinitesimal α, the difference $f(x+\alpha)-f(x) = (f(x+a_n)-f(x))$ is also infinitesimal. As he grew confident using infinitesimals, Cauchy manipulated them in the same way as he manipulated variable quantities. He imagined points and lines after the style of the ancient Greeks, where a line is an entity that has 'length but no breadth' and a point is an entity that has 'position but no size'. A line is an entity in its own right. A point can be marked on a line and a line can be drawn through a point. Furthermore, points can be constant, remaining in a fixed position, or variable, including infinitesimals as variables that tend to zero.

What is less obvious is the nature of the number line itself. Not only is it possible to mark rational points on it, but there are also points such as $\sqrt{2}$ or π that arise in geometric constructions but are not rational. At the end of the nineteenth century, a solution was found by formulating the completeness property of the number line, so that it includes not only the rational numbers, but also the limits of convergent rational sequences. This notion of the reals cannot include infinitesimals if they are defined as 'smaller than any positive real number', because this concept is logically contradictory.

A significant change in interpretation then occurred. Instead of a line being something drawn physically by the stroke of a pen, or imagined theoretically as a Platonic entity with length but no breadth, it was re-conceptualised as the set of all real numbers – defined, either as limits of convergent (that is, Cauchy) sequences of rational numbers, or as Dedekind cuts on the rationals [21]. Now every real number occupies a fixed place on the number line. It cannot move around, or become 'arbitrarily small', so the formal real number line cannot include infinitesimals in Cauchy's sense. Indeed, the notion of a 'variable' was reformulated as a fixed but arbitrary member of some set. Movement was effectively banished from the fundamentals of mathematics. In consequence, the notion of an infinitesimal was excluded from formal set-theoretic analysis.

Yet the idea that quantities can vary, and become arbitrarily small, remains in the mental imagery of practising mathematicians, especially in applied mathematics. The recent arrival of dynamic computer graphics offers an entirely new practical way to visualise mathematical ideas. Greek geometers drew a curve as a mark in the sand; later generations drew it in ink on paper, or printed it in a book. Today we can represent a curve on a high-resolution visual display, and zoom in on part of the curve to magnify the image. Unlike the magnification of a picture in a textbook, where the thickness of the graph increases under a magnifying glass, we can now move our fingers apart on a tablet computer to magnify the picture while the graph is redrawn at the same visual thickness.

As we zoom in on a standard smooth graph such as $y = x^2, y = e^x$, or $y = \sin x$, the magnified picture looks less and less curved. At successively higher magnifications, a small portion of the graph looks more and more like a straight line, and the graph and the tangent become visually indistinguishable, see Figure 15.3. Now, under high magnification, a picture of a smooth graph magnifies locally to look like a straight line.

This suggests a new way of imagining the tangent to a differentiable function: under high magnification, the graph looks less and less curved, until at suitably high magnification a small portion of the graph looks straight. The derivative is now visually the slope of the graph, and as we cast an eye along the graph, the derivative is the changing slope of the graph itself. This property can be visualised for combinations of standard polynomial, trigonometric, logarithmic, and exponential functions, but it needs to be formally defined to reach modern standards of rigour, and suitably interpreted to apply to complex analysis and higher-dimensional vector calculus.

15.2 The Relationship Between Real and Complex Analysis

The generalisation from the real to the complex case follows the same symbolic formulation, but the visual representations are different. In the real case, we can draw

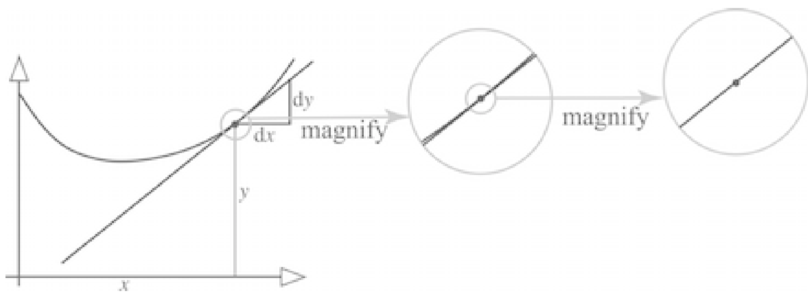

Figure 15.3 Successive magnifications of a smooth curve.

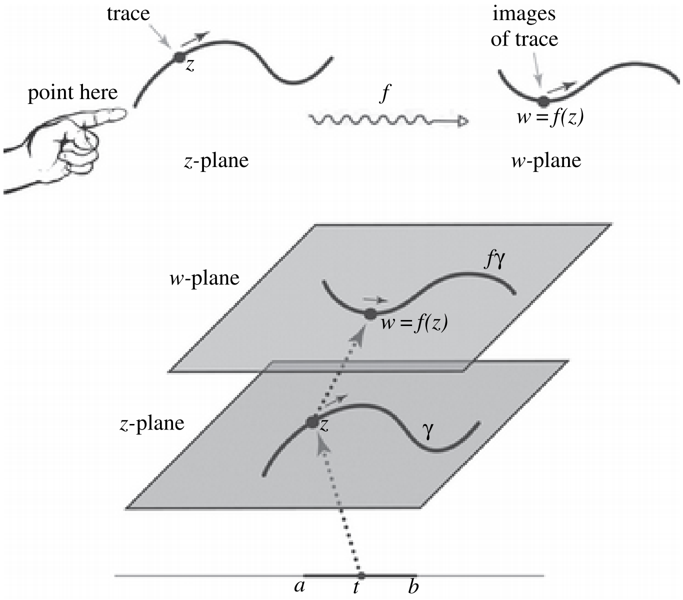

Figure 15.4 *Top*: Tracing a point in the z-plane to see its image move in the w-plane. *Bottom*: Visualising this in three dimensions by stacking planes.

the graph of a function $f : D \to \mathbb{R}$ as a subset $\{(x, f(x)) \in \mathbb{R}^2 : x \in D\}$, but this cannot be directly generalised to the complex case because the graph is a subset of $\mathbb{C} \times \mathbb{C}$, which requires four real dimensions. However, we can use alternative representations in two or three dimensions. For example, Figure 15.4 (top) represents the complex function $w = f(z)$ by drawing parts of the z-plane and the w-plane side by side. If a finger starting at a point z traces a path in the z-plane, then the corresponding point $w = f(z)$ moves around in the w-plane.

This function can be visualised in three-dimensional space by placing the w-plane above the z-plane, with an arrow connecting z to $f(z)$. This representation is similar to that used in Figure 2.9, but with an extra dimension in the z-plane.

Then, as z moves in the z-plane along a path γ, the arrow moves along the image path $f\gamma$, tracing out the motion of $w = f(z)$, Figure 15.4 (bottom).

The movement of the finger travelling smoothly along a curve in the z-plane without stopping can be represented formally by a path $\gamma : [a, b] \to \mathbb{C}$, where γ is differentiable and $\gamma'(t) \neq 0$ for $t \in [a, b]$. Recall from Definition 6.21 that we call such a path a *regular curve*. In this case, the tangent to γ at $z_0 = \gamma(t_0)$ makes an angle $\text{arc}\,(\gamma'(t_0))$ with the real axis. In Section 13.3, for an analytic function f, we distinguished between regular points z_0 where $f'(z_0) \neq 0$ and critical points where $f'(z_0) = 0$. The behaviour of the transformed path $f\gamma$ near $w_0 = f(z_0)$ depends on this distinction. The transformed image path $w = f(\gamma(t))$ satisfies $(f\gamma)'(t_0) = f'(z_0)\gamma'(t_0)$, and if z_0 is a regular point of f, that is $f'(z_0) \neq 0$, then the image path has a tangent at $w = f(z_0)$ as in Figure 15.5. The tangent makes an angle

$$\text{arc}\,(f'(z_0)\gamma'(t_0)) = \text{arc}\,(f'(z_0)) + \text{arc}\,(\gamma'(t_0))$$

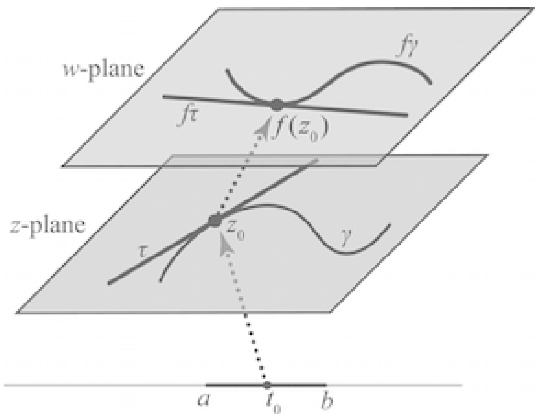

Figure 15.5 Mapping a path γ in D by an analytic function f where $f'(z_0) \neq 0$.

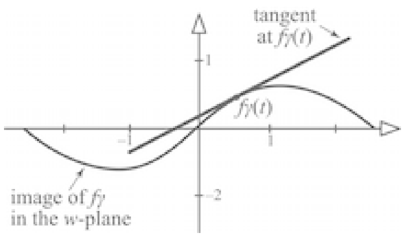

Figure 15.6 Tangent to curve in the w-plane.

with the real axis. The transformation f turns the tangent at a regular point through an angle arc$(\gamma'(t_0))$.

15.2.1 Critical Points

In Section 13.3 we compared the geometry of a differentiable complex function f near a regular point and near a critical point. For a regular point z_0, the tangent to a regular curve γ through z_0 has a corresponding tangent through $f(z_0)$.

Example 15.1. Consider the function $f(z) = \sin z$ at the origin where $f'(0) = \cos 0 = 1$. The path $\gamma(t) = t + \mathrm{i}t$ for $t \in [-1, 1]$ in the z-plane is transformed into

$$\begin{aligned} f(\gamma(t)) \; = \; & \sin(t + \mathrm{i}t) = \sin(t)\cos(\mathrm{i}t) + \cos(t)\sin(\mathrm{i}t) \\ & = \sin(t)\cosh(t) + \cos(t)\sinh(t) \end{aligned}$$

in the w-plane, see Figure 15.6. □

At a critical point, $f'(z_0) = 0$, so τ collapses to a single point: $\tau(z_0 + h) = w_0$. In Chapter 13 we noted that f need not be conformal at a critical point. In this case, a graph

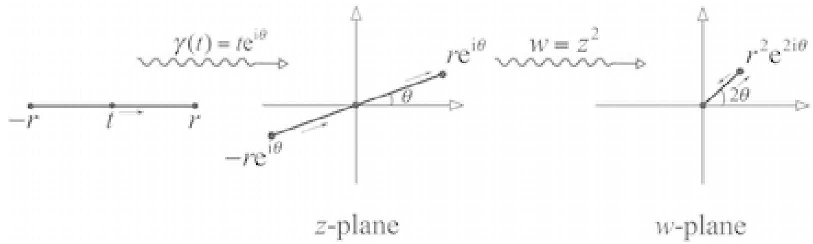

Figure 15.7 Transforming the path $\gamma(t) = te^{i\theta}$ by the function $w = z^2$.

γ in the z-plane passing through z_0 is transformed by an analytic function f in a subtler manner.

Example 15.2. Consider the function $f(z) = z^2$ at the origin where $f'(0) = 0$. The path $\gamma(t) = te^{i\theta}$ for $t \in [-r, r]$ in the z-plane is transformed by f into the composition

$$f\gamma(t) = t^2\cos(2\theta) + it^2\sin(2\theta)$$

in the w-plane. (Recall we now write $f\gamma$ instead of $f \circ \gamma$, for brevity.) The original path γ in the z-plane starts at $-re^{i\theta}$ and moves in a straight line through the origin to the point $re^{i\theta}$ on the other side; the path $f\gamma$ in the w-plane lies on the straight line at an angle 2θ to the real axis, starting at $(-r)^2e^{2i\theta} = r^2e^{2i\theta}$ moving through the origin and turning back in the opposite direction to $r^2e^{2i\theta}$, see Figure 15.7.

This apparently sudden turn in direction turns out to be a direct generalisation of the real case. If the real function $f(x) = x^2$ is drawn with separate x- and y-axes, as x increases from $-r$ to $+r$ through the origin, the y-value x^2 starts at r^2, moves down to the origin, and then turns back up again to the starting point.

An alternative way to grasp what is happening is to represent the path $\gamma : [a, b] \to \mathbb{C}$ by writing $f\gamma(t) = u(t) + iv(t)$ and looking at the graph of $(t, u(t), v(t))$ in three-dimensional (t, u, v)-space. In Figure 15.8 the w-plane is drawn horizontally and the t-axis vertically upwards. As t moves up the vertical axis from $-r$ to r, $(t, u(t), v(t))$ moves along a path in three dimensions, with the tangent at the origin being vertical where its projection onto the w-plane is a single point. Looking down on the w-plane, the point $(u(t), v(t))$ moves from $r^2e^{2i\theta}$ to the origin and turns back again.

Imagining t as time, the velocity of the moving point on the path f is $(f\gamma)'(t) = 2te^{2i\theta}$, and as t passes through the origin the velocity smoothly slows down to zero, and then reverses without any sudden change in speed. □

More generally, if a point moves along any smooth path γ through a critical point in the z-plane, then the path $f\gamma$ in the w-plane may turn back in the opposite direction. In three-dimensional (t, u, v)-space, however, it travels smoothly through the critical point. For example, for $w = z^2$, the path $\gamma(t) = 1 - e^{i\pi t}$ ($t \in [-1, 1]$) is a circle radius 1, centre 1, which travels through the critical point at the origin in the z-plane. Meanwhile the path

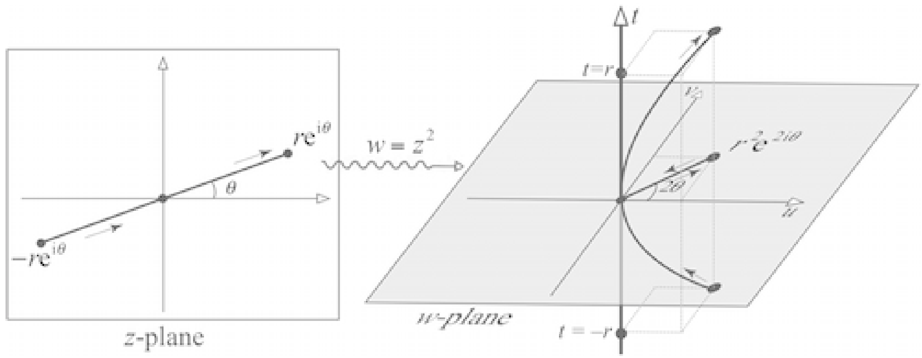

Figure 15.8 The path $f\gamma$ in (t, u, v)-space.

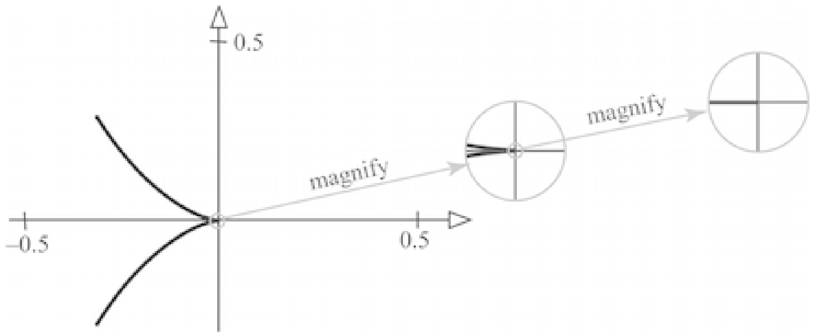

Figure 15.9 Magnifying a path in the w-plane through a critical point of f.

in the w-plane passes through the origin as in Figure 15.9. Magnifying the picture in the w-plane near the origin reveals the path moving in and out in what looks like a straight line.

15.3 Interpreting Power Series Tending to Zero as Infinitesimals

Using the idea that an infinitesimal quantity is a variable that tends to zero, the variable part of a power series can be interpreted as an infinitesimal. Moreover, it has a special property that relates to Leibniz's idea that infinitesimals have an order of size. Recall from Definition 13.10 that an analytic function $f : D \to \mathbb{C}$ is of *order* n at $z_0 \in D$ (where $n \geq 1$) if $f^{(n)}(z_0) \neq 0$, and $f^{(k)}(z_0) = 0$ for $1 \leq k < n$. By (13.6) the Taylor series about z_0 then has the form

$$f(z_0 + h) = f(z_0) + h^n f^{(n)}(z_0)/n! + \cdots$$

which is a constant $f(z_0)$ plus an infinitesimal of order n. So the behaviour of $f(z)$ near a critical point z_0 is given by the order of the variable part of the analytic function at the point. The definition applies to regular points where $n = 1$, and to critical points where $n > 1$, to give a single theory in all cases.

Using polar coordinates, we observed that if $h = re^{i\theta}$, where h is small, then f transforms a small region of the z-plane near z_0 to the w-plane near $w_0 = f(z_0)$. The radius r scales to kr^n, and the original angle θ is multiplied by n and rotates to $n\theta + \alpha$, as we saw in Chapter 13, Figure 13.6. Moreover, we noted that a further rotation of the line from z_0 to $z_0 + h$ in the z-plane through ϕ rotates the corresponding line in the w-plane to $n(\theta + \phi) + \alpha$, as in Figure 13.7.

The case of a straight line passing through z_0 can now be considered as two half-lines where $\phi = \pi$. The image in the w-plane (for suitably small h) is then made up of two half-lines, one at an angle θ to the real axis, the other at an angle $\theta + \pi$. Under the transformation f, the angle between them is increased to $n\pi$. For n odd, this gives two half-lines in opposite directions and, for n even, it gives two half-lines in the same direction.

This is a natural generalisation of the real case where, for n odd, as x increases through zero, so does y, while for n even, the graph of $y = x^n$ has a minimum at $x = 0$, and as x passes through the origin, y travels down to zero and turns back up again.

The language used in the preceding discussion has been expressed in terms of a variable h that is 'sufficiently small', linking to the Leibniz notion of infinitesimals of various orders of size and the dynamic idea of 'arbitrarily small' quantities in applied mathematics. We now introduce a formal set-theoretic definition of infinitesimals and prove a structure theorem that reveals how they can be visualised on a number line or in a complex plane.

15.4 Real Infinitesimals as Variable Points on a Number Line

The structure of the human visual system makes it natural for us to imagine points moving along a line. Given a vertical line in the real plane of the form $x = v$, consider where various polynomials intersect this vertical line as it is moved to the left or right, Figure 15.10. A constant function $y = k$ always meets the line at the same height k, but the line $y = x$ meets the line at a height v, and as the line moves, v varies. As v gets smaller, the variable point v moves below any fixed positive constant k. In this sense, as v tends to zero, the variable point becomes less than k for any real number k – one

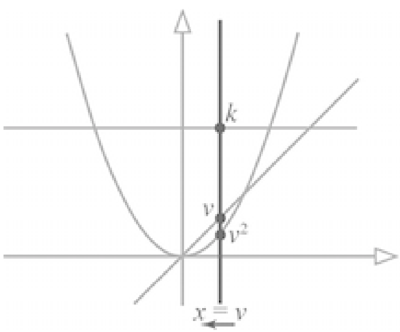

Figure 15.10 Infinitesimals as variable points on a line.

interpretation of ‘infinitesimal’. In the same sense, the variable point v^2 is smaller still – infinitesimal compared to v.

The vertical line $x = v$ can now be seen to have two kinds of points. Some are familiar constant points that remain in the same position, while others are variable, and can become arbitrarily small. Others can become even smaller.

15.5 Infinitesimals as Elements of an Ordered Field

The plan now is to extend the real numbers $\mathbb{R}$ to a larger ordered field K that contains infinitesimals, and to extend real functions f so that we can calculate $f(x)$ for $x \in K$. Taking a cue from the previous section, we might begin with the field $\mathbb{R}(v)$ of rational functions in v,

$$\frac{a_0 + a_1 v + \cdots + a_n v^n}{b_0 + b_1 v + \cdots + b_m v^m}$$

where $a_0, a_1, \ldots, a_n, b_0, b_1, \ldots, b_m \in \mathbb{R}$ and $b_m \neq 0$, and give it the structure of an ordered field by defining v to satisfy $0 < v < k$ for all positive $k \in \mathbb{R}$. This lets us evaluate rational functions that include infinitesimals.

For instance, if $f(x) = x^2/(x-1)$ we can extend f to include calculations with the infinitesimal v by substitution:

$$f(x+v) = (x+v)^2/(x+v-1) \in \mathbb{R}(v)$$

To extend real functions $f : D \to \mathbb{R}$ so that we can calculate $f(x+h)$ when $x \in D$ and h is infinitesimal, we work in a suitable larger field K. The field $K = \mathbb{R}(v)$ is fine for rational functions, but not for more general functions. Leibniz’s version of calculus deals with polynomials, trigonometric functions, exponential functions, and, more generally, any function that can be expressed as a power series. In this case we need a larger field that includes power series in an infinitesimal. If we include a single infinitesimal ε, then we also need all the elements that are produced by performing the usual operations of addition, subtraction, multiplication, and division with power series so that the system is an ordered field. In particular, the field should include all power series in ε and inverses of infinitesimals such as $1/\varepsilon$.

Foundational Example 15.3. (The Superreal Field Generated by a Single Infinitesimal). The superreal field $\mathbb{R}_\varepsilon$ is defined to consist of all series of the form

$$\sum_{r=n}^{\infty} a_r \varepsilon^r$$

where n is an integer (which may be negative). To obtain a field, the series are added, subtracted and multiplied term by term. Division in the form

$$\left(\sum_{r=m}^{\infty} a_r \varepsilon^r\right) \Big/ \left(\sum_{r=n}^{\infty} b_r \varepsilon^r\right) = \sum_{r=m+n}^{\infty} c_r \varepsilon^r$$

is achieved by solving the equation

$$\left(\sum_{r=m+n}^{\infty} c_r\varepsilon^r\right)\left(\sum_{r=n}^{\infty} b_r\varepsilon^r\right) = \left(\sum_{r=m}^{\infty} a_r\varepsilon^r\right)$$

recursively for successive coefficients $c_{m+n}, c_{m+n+1}, \ldots$.

Here we do not need to worry about convergence, because all the operations are purely formal. It is possible to give an alternative set-theoretic definition in which the superreals are defined to be functions $s : \mathbb{Z} \to \mathbb{R}$, where each s has an integer n such that $s(r) = 0$ for $r < n$. The function $s : \mathbb{Z} \to \mathbb{R}$ then corresponds to the infinite series $\sum_{r=n}^{\infty} a_r\varepsilon^r$. Knowing that the superreal numbers can be given a formal set-theoretic definition, we can safely think of them as infinite series in ε. □

DEFINITION 15.4. An element $\sum_{r=n}^{\infty} a_r\varepsilon^r$ is an *infinitesimal of order n* if $a_n \neq 0$ and $n \geq 1$. It is *finite* if $n \geq 0$ and *infinite of order k* if $n < 0, a_n \neq 0$ and $k = -n$. □

Examples 15.5. (i) $\varepsilon^3 + 17\varepsilon^{25}$ is a third order infinitesimal.
(ii) $0, 25, \pi + \varepsilon, \varepsilon^4$ are finite.
(iii) $1/\varepsilon$ and $\varepsilon^{-11} - 15\varepsilon$ are infinite.
(iv) $\sin\varepsilon = \varepsilon - \varepsilon^3/3! + \cdots$ is a first order infinitesimal. □

The term ‘order’ is used here in a special sense referring to the ‘order of size’ of an infinitesimal as distinct from the order $a < b$ in the field. There may be two elements $a < b$ where the order of a as an infinitesimal is greater than the order of b. An example is $a = -\varepsilon^2$ and $b = \varepsilon$. If there is a danger of ambiguity, we refer to the order of a single element a in the sense of Definition 15.4 as the *order of infinitesimality* of a, and the order $a < b$ between two elements as the *linear order* between a, b.

The order of infinitesimality of $a = \sum_{r=n}^{\infty} a_r\varepsilon^r$ is the subscript n of the dominant term a_n, which is the first non-zero term in the sum.

The linear order $a < b$ also depends on the dominant terms of a, b. If $a = \sum_{r=n}^{\infty} a_r\varepsilon^r, b = \sum_{r=m}^{\infty} b_r\varepsilon^r$, then without loss of generality we can assume $m \leq n$. If $m = n$ then the linear order of a, b is the same as the linear order of a_m, b_n. If $m < n$, then a_m dominates and the linear order is given by the sign of a_m. (In the latter case, $b_m = 0$, so the same rule applies: in this case the linear order of a, b is the same as the linear order of $a_m, 0$.)

Examples 15.6. (i) $\varepsilon^2 < 3\varepsilon$.
(ii) $-1/\varepsilon < 0$ are finite.
(iii) $\varepsilon < r$ for every positive $r \in \mathbb{R}$. □

The superreal field generated by just one infinitesimal has many of the properties that Leibniz wanted. Every non-zero element is either finite or has a specific order as an infinitesimal, or as an infinite element. Any real function expressible as a power series $f(x) = \sum_{r=n}^{\infty} a_r x^r$ can be extended to define $f(x+\delta)$ for an infinitesimal $\delta = \sum_{s=m}^{\infty} b_s \varepsilon^s$ by substituting

$$f(x+\delta) = \sum_{r=n}^{\infty} a_r \left[x + \sum_{s=m}^{\infty} b_s \varepsilon^s \right]^r$$

and then rearranging the result as $\sum_{r=n}^{\infty} c_r x^r$ where $c_r \in \mathbb{R}_\varepsilon$. This process looks messy, but it can easily be done term by term. (See the exercises at the end of the chapter.)

Our choice of $\mathbb{R}_\varepsilon$ is made initially for pedagogical purposes; in particular, to investigate intuitive models for limits and derivatives in terms of genuine infinitesimals. This field has an explicit description, but it does not have all the properties we would like. For example, the exponential function is not defined at $1/\varepsilon$. Indeed, $\exp(1/\varepsilon)$ has a series involving *all* negative powers of ε, not just a finite number. This feature causes no problems in the sequel, but it is a limitation of the current approach to infinitesimals, and it can be avoided altogether by working with the hyperreal numbers, Section 15.10.

The same technique works in the complex case, where $\mathbb{C}_\varepsilon$ is the supercomplex field of elements $\sum_{r=n}^{\infty} c_r \varepsilon^r$ where the coefficients c_r are complex numbers. Complex functions $f(z)$ given by power series can be extended to apply to $f(z+\delta)$, where δ is infinitesimal. However, complex analysis is much simpler than real analysis because every differentiable complex function can be represented locally by a power series. Introducing infinitesimals into real analysis therefore requires a more powerful extension of the real numbers.

We begin with a simple definition that proves amazingly effective.

DEFINITION 15.7. An ordered field K that contains the real numbers as a proper ordered subfield is a *super ordered field*. □

Examples of super ordered fields include the field of rational functions, the superreal numbers $\mathbb{R}_\varepsilon$, and many others, such as the hyperreal numbers that are studied in nonstandard analysis, Section 15.9.

A super ordered field has all the familiar properties of an ordered field. It also contains infinite and infinitesimal elements defined as follows:

DEFINITION 15.8. In any super ordered field K, an element $x \in K$ is:

positive infinite if $x > r$ for all $r \in \mathbb{R}$
negative infinite if $x < r$ for all $r \in \mathbb{R}$
finite if $a < x < b$ for some $a, b \in \mathbb{R}$
positive infinitesimal if $0 < x < r$ for all positive $r \in \mathbb{R}$
negative infinitesimal if $0 < -x < r$ for all positive $r \in \mathbb{R}$. □

Using the standard properties of an ordered field it is straightforward to prove that an element x is infinitesimal if and only if $1/x$ is (positive or negative) infinite. (These and other properties of arithmetic and order are left as exercises at the end of the chapter.)

15.6 Structure Theorem for any Ordered Extension Field of $\mathbb{R}$

The key to visualising the infinitesimal structure in the field $\mathbb{R}_\varepsilon$ is a simple but powerful structure theorem:

THEOREM 15.9. *Every element x in a super ordered field K is either infinite, or uniquely of the form $x = c + h$ where $c \in \mathbb{R}$ and h is zero or infinitesimal.*

Proof. We need consider only the case where x is finite, so that $a < x < b$, where $a, b \in \mathbb{R}$. Let $S = \{r \in \mathbb{R} : r < x\}$. Then S is non-empty because $a \in S$, and bounded above (by b). By completeness of $\mathbb{R}$, the set S has a unique least upper bound c. Let $h = x - c$, so $x = c + h$. Now we use completeness to show that if $h \neq 0$, then h is infinitesimal. There are two cases:

For $h > 0$, suppose that there exists $k \in \mathbb{R}$ such that $0 < k < h$. Then $c + k < c + h = x$, so by definition of S, $c + k \in S$. But this implies that $c + k$ is greater than the least upper bound c. By contradiction, h is infinitesimal.

For $h < 0$, suppose there exists $k \in \mathbb{R}$ such that $0 < k < -h$. Then $x = c + h < c - k < c$, so by definition $c - k$ is an upper bound of S. But $c - k$ is less than the least upper bound c, again giving a contradiction. So h is infinitesimal.

To prove uniqueness, suppose that $c + h = d + k$, where $c, d \in \mathbb{R}$ and h, k are infinitesimal. Then $c - d = k - h$ is infinitesimal or zero, and real. Hence it is zero. We have now proved that any finite $x \in K$ is uniquely of the form $x = c + h$, where $c \in \mathbb{R}$ and h is infinitesimal or zero. □

We extract a key concept:

DEFINITION 15.10. The real number c is called the *standard part* of x and is written $c = \operatorname{st}(x)$. □

This structure theorem is the key to imagining infinitesimal quantities on a number line. Over the centuries, we have expanded our interpretation of a number line, from the Greek idea of a line where the distance between two points is the same regardless of direction, to a number line with positive and negative numbers, rational numbers, irrational numbers, and real numbers. There is no logical reason to stop there. A super ordered field, has finite, infinite, and infinitesimal elements. The structure theorem tells us that any finite element is (uniquely) a real number plus an infinitesimal. An infinitesimal is, of course, too tiny to see to a normal scale, but this is not as problematic as it seems to be at first sight (pun intended). For the real numbers, on a normal scale we cannot see the difference between the irrational number π and the rational number given by π to a thousand decimal places. However, if we magnified a small part of the picture by a factor 10^{1000}, the difference would become apparent. This simple idea offers the way ahead: we can magnify a super ordered field to distinguish infinitesimally close points visually. First, we establish some simple ideas relating finite elements and infinitesimals.

PROPOSITION 15.11. *Let F be the subset of finite elements in K and I the subset of elements that are zero or infinitesimal. Then $a, b \in F$, $c, d \in I$ imply $a + b, a - b, ab \in F$, $c + d, c - d, ac \in I$.*

Proof. These follow from the definitions. □

Proposition 15.11 states that the sum, difference, and product of finite numbers are also finite, but the quotient a/b need not be finite (if b is infinitesimal). The sum and difference of infinitesimal elements are infinitesimal, and the product of a non-zero finite element by an infinitesimal is again infinitesimal. (Anyone familiar with ring theory will recognise that F is a ring and I is an ideal of F; however, this is not essential to follow the ideas in this chapter.)

PROPOSITION 15.12. *For $x, y \in F$, the standard part map* $\mathrm{st} : F \to \mathbb{R}$ *satisfies*

$$\begin{aligned} \mathrm{st}(x+y) &= \mathrm{st}(x) + \mathrm{st}(y) \\ \mathrm{st}(x-y) &= \mathrm{st}(x) - \mathrm{st}(y) \\ \mathrm{st}(xy) &= \mathrm{st}(x)\,\mathrm{st}(y) \\ \mathrm{st}(x/y) &= \mathrm{st}(x)/\mathrm{st}(y) \quad (\text{for } \mathrm{st}(y) \neq 0) \end{aligned}$$

Proof. Use algebraic manipulation of the definition, taking care with division. □

15.7 Visualising Infinitesimals as Points on a Number Line

The structure theorem offers new ways of visualising infinitesimal quantities as points on a number line. A standard picture cannot distinguish points that differ by an infinitesimal, because the difference is too small to see, or represent infinite quantities, because they are too far away. But a linear map $m(x) = ax + b$ for appropriate values of $a, b \in K$ reveals infinitesimal or infinite detail:

DEFINITION 15.13. The map $m : K \to K$ where $m(x) = (x - c)/d$ for $c, d \neq 0$ is *the d-lens pointed at c*. If d is infinitesimal, then m is a *microscope pointed at d*. If c is infinite then m is a *telescope pointed at infinity*.

The set

$$V = \{x \in K : m(x) \text{ is finite}\}$$

is the *field of view* of m. □

The field of view V of a d-lens contains those elements $x \in K$ whose image $m(x)$ is finite, and these images can be marked on a finite line in a picture. For instance, if c is a real number and δ is infinitesimal, the field of view V of a δ-microscope pointed at c contains $c, c+\delta, c+2\delta$. These differ by infinitesimal quantities, but they map to $m(c) = 0, m(c+\delta) = 1, m(c+2\delta) = 2$, which are now visibly different. At the same time, the points $c+\delta$ and $c+\delta+3\delta^2$ map to $m(c+\delta) = 1, m(c+\delta+3\delta^2) = 1+3\delta$, which differ by an infinitesimal and cannot be distinguished at this scale. See Figure 15.11.

When we make a map in a real-world situation, say a map of England including London, we mark the location of London using the same name. This technique can be used to advantage when magnifying infinitesimal detail using a δ-lens. Renaming the image points with their original names, as in Figure 15.12, the picture may be interpreted

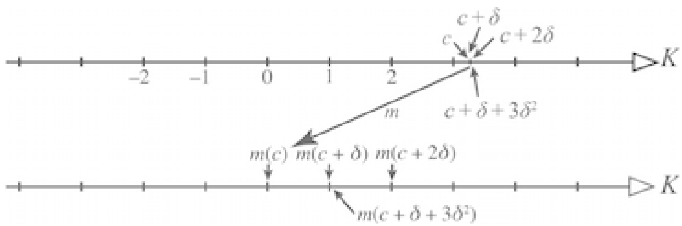

Figure 15.11 Scaling up K using a δ-lens pointed at c.

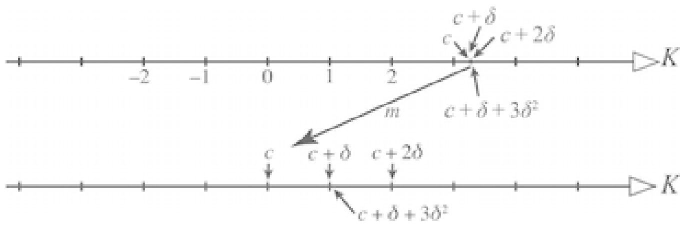

Figure 15.12 Scaling up infinitesimal detail.

as scaling up part of the number line K to distinguish points that differ by an infinitesimal amount.

Using a specific map, it is possible to see only a certain level of detail, because some differences in the image are too small to distinguish, and some are too far away. Leibniz imagined infinitesimals of order $1, 2, 3, \ldots$ and in the above pictures we have included only examples of polynomials in an infinitesimal that follow Leibniz's intuition. However, we must be prepared for more general situations that occur in practice. For example, if h is an infinitesimal of order 1 and h has a square root $r \in K$, then $r^2 = h$ and r must have order $\frac{1}{2}$. To deal with the general case, instead of referring to the specific order of infinitesimality as a whole number, we compare the relative order of size of two elements a, b as follows:

DEFINITION 15.14. If $b \in K$ and $b \neq 0$, an element $a \in K$ is:

of *higher order* than b if a/b is infinitesimal,
of the *same order* as b if a/b is finite,
of *lower order* than b if a/b is infinite. □

When visualising infinitesimal detail using a δ-lens pointed at c, the field of view V consists of elements that differ from c by an element of the same or higher order. Those differing by the same order map onto distinct points: those that differ by a higher order quantity are indistinguishable to the naked eye. We can model this visually using:

DEFINITION 15.15. The *optical δ-lens* or *optical microscope* μ pointed at c is $\mu : V \to \mathbb{R}$ where $\mu(x) = \mathrm{st}((x - c)/\delta)$. □

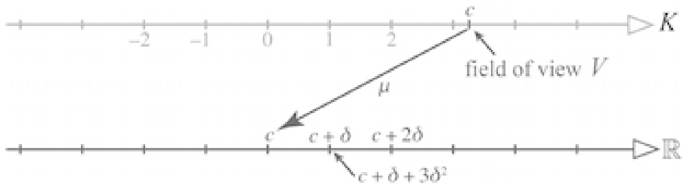

Figure 15.13 Scaling points in the field of view onto the whole real line.

The image is now a drawing of the real line. Points in V that differ by an element of the same order as δ map on to different real numbers and points that differ by a higher order than δ map onto the same real number. Meanwhile, the δ-lens $m(x) = (x - c)/\delta$ maps points outside the field of view V onto infinite elements of K, which are beyond the range of a finite drawing. In Figure 15.13 the line representing the field K is greyed out to show that only the point c and the field of view V are magnified to give a visible point on the real line.

In general, the optical microscope $\mu : V \to \mathbb{R}$ always maps the field of view V onto the whole of $\mathbb{R}$ because $\mu(c + x\delta) = x$ for any $x \in \mathbb{R}$.

A special case occurs when we take $c = 0$ and $\delta = 1$. Now the field of view is the set of finite numbers F, and the optical δ-lens is the standard part function st. The image is the whole real line and the points mapping onto a particular real number $x \in \mathbb{R}$ consist precisely of x and those points differing from x by an infinitesimal. This suggests:

DEFINITION 15.16. The *monad* M_x around $x \in K$ is

$$M_x = \{x + h \in K : h \in I\}$$

where I is the subset of K consisting of zero and infinitesimals. □

Some sources use the alternative name 'halo' in place of 'monad'.

In this definition, x may be any element of K, so any infinite value of x is also surrounded by its own monad. If x is finite, then by the structure theorem of Theorem 15.9, there is only one real number $c \in M_x$ and the standard part map st : $F \to \mathbb{R}$ collapses the monad M_c down onto the real number c.

DEFINITION 15.17. The relationship $x \approx y$ is defined on K by $x \approx y$ if and only if $x - y \in I$. □

This is an equivalence relation and the monads are the equivalence classes. It lets us imagine the picture of the super ordered field K as points on a line, where the finite points F consist of the real numbers and each real number c is surrounded by the monad M_c of elements $c + h$ where h is infinitesimal.

This vision lets us see that any super ordered field K is not complete. We need:

THEOREM 15.18. *For $a \in \mathbb{R}$, the monad M_a is bounded above by any real number $b > a$, but has no least upper bound.*

Proof. If $l \in K$ is a least upper bound of M_a, then either $l \in M_a$ or $l \in M_b$ for $b > a$. In the first case, for any positive infinitesimal h, $l + h \in M_a$ is greater than the supposed upper bound l of M_a. In the second, $l - h \in M_b$ is an upper bound for M_a that is smaller than the least upper bound l. By contradiction, M_a has no least upper bound. □

COROLLARY 15.19. *A super ordered field is not complete.* □

This should not be a surprise. A complete ordered field is unique up to isomorphism and does not contain infinitesimals. So a super ordered field that contains infinitesimals cannot be complete. The structure theorem, Theorem 15.9, tells us why.

Nevertheless, we now have a fundamentally new way of visualising 'the number line'. Not only does it include the familiar real numbers; it can include infinitesimals and infinite quantities that can be seen by human eyes only through appropriate d-lenses.

15.8 Complex Infinitesimals

The generalisation of infinitesimals in any super real field K can be extended to the complex case by using quantities of the form $x + \mathrm{i}y$ where $x, y \in K$ and $\mathrm{i}^2 = -1$.

DEFINITION 15.20. Given a super ordered field K, the corresponding *super complex field* is the field $K[\mathrm{i}]$ of polynomials in i where $\mathrm{i}^2 = -1$. It is an extension field of the field of complex numbers $\mathbb{C} = \mathbb{R}[\mathrm{i}]$. It is a field because the polynomial $t^2 + 1$ has no zeros in any ordered field, so is irreducible. Alternatively, a direct proof follows the usual one for $\mathbb{C}$ by writing $(x + \mathrm{i}y)^{-1}$ in the form $u + \mathrm{i}v$.

An element $z = x + \mathrm{i}y$ where $x, y \in K$ is *infinitesimal* if both x and y are infinitesimal, *finite* if both x and y are finite, and *infinite* if at least one of x, y is infinite. The *standard part* of a finite element $z = x + \mathrm{i}y$ is

$$\mathrm{st}\,(z) = \mathrm{st}\,(x) + \mathrm{i}\,\mathrm{st}\,(y)$$

An element w is of *higher order* than a non-zero element z if w/z is infinitesimal; of the *same order* if w/z is finite and non-zero; and of *lower order* if w/z is infinite.

Let F_{i} be the set of finite elements in $K[\mathrm{i}]$ and I_{i} the set of elements that are infinitesimal or zero. (Again F_{i} is a ring and I_{i} is an ideal in F_{i} with the properties specified in Proposition 15.11.) □

THEOREM 15.21 (Structure Theorem for a Super Complex Field). *Every element $z = x + \mathrm{i}y$ in a super complex field is either infinite or uniquely of the form $z = w + h$ where $w \in \mathbb{C}$ is given by $w = \mathrm{st}\,(z)$ and h is zero or infinitesimal.*

For finite u, v, the standard part map satisfies

$$\begin{aligned}
\mathrm{st}\,(u + v) &= \mathrm{st}\,(u) + \mathrm{st}\,(v) \\
\mathrm{st}\,(u - v) &= \mathrm{st}\,(u) - \mathrm{st}\,(v) \\
\mathrm{st}\,(uv) &= \mathrm{st}\,(u)\,\mathrm{st}\,(v) \\
\mathrm{st}\,(u/v) &= \mathrm{st}\,(u)/\mathrm{st}\,(v) \quad (\text{for } \mathrm{st}\,(v) \neq 0) \\
\mathrm{st}\,(x + \mathrm{i}y) &= \mathrm{st}\,(x) + \mathrm{i}\,\mathrm{st}\,(y)
\end{aligned}$$

Proof. If z is finite, then both x and y are finite elements in K, and the structure theorem for super ordered fields shows that they are uniquely of the form $x = c + p, y = d + q$

where $c, d \in \mathbb{R}$ and $p, q \in I$, so $z = w + h$ where $w = c + \mathrm{i}d \in \mathbb{C}$, and $h = p + \mathrm{i}q \in I_{\mathrm{i}}$. The properties of the map st follow by direct calculation. □

Many definitions and properties generalise from the real case, such as:

DEFINITION 15.22. The monad M_z around $z \in K[\mathrm{i}]$ is

$$\{z + h \in K[\mathrm{i}] : h \in I_{\mathrm{i}}\}$$ □

The monads partition $K[\mathrm{i}]$ into non-overlapping subsets consisting of points that differ by an infinitesimal. This includes the case where z may be infinite, so that even an infinite element is surrounded by a monad of points differing from it by an infinitesimal. Finite elements in $K[\mathrm{i}]$ consist of the underlying field of complex numbers together with the monad M_z surrounding each complex number z.

DEFINITION 15.23. If $d \neq 0$, the *d-lens pointed at w* in a super complex field is

$$m(z) = \frac{z - w}{d}$$

The *field of view* of m is the set of points z in the super complex field such that $(z-w)/d$ is finite.

The *optical d-lens pointed at w* is

$$\mu(z) = \mathrm{st}\,(m(z) = \mathrm{st}\left(\frac{z - w}{d}\right)$$

If d is infinitesimal, the d-lens is called a *microscope*, and if w is infinite, it is called a *telescope*. □

Multiplying by a complex number in the form $re^{\mathrm{i}\theta}$ scales by a factor r and turns through an angle θ, so it is sensible to take $\theta = 0$ and to select d to be of the form $d = \delta + \mathrm{i}0$ where $\delta > 0$. (Such an element will be called a positive infinitesimal.) Then the δ-lens simply scales by a factor δ and does not rotate the image. For this reason, we will always assume that the scaling factor δ is a positive element of the subfield K.

Example 15.24. Figure 15.14 shows the point $z_0 = re^{\mathrm{i}\theta}$, where $r, \theta \in \mathbb{R}$, lying on the circle $|z| = r$ in $K[\mathrm{i}]$, $z_0 + h = re^{\mathrm{i}\theta+\varepsilon}$ lies on the circle where ε is an infinitesimal in K. Let μ be an optical microscope $\mu(z) = \mathrm{st}\,(z - z_0)/\delta$, where δ is the same order as ε.

Then

$$\begin{aligned} h &= re^{\mathrm{i}\theta+\varepsilon} - re^{\mathrm{i}\theta} \\ &= re^{\mathrm{i}\theta}(e^{\mathrm{i}\varepsilon} - 1) \\ &= z_0(\mathrm{i}\varepsilon - \varepsilon^2/2 + \cdots) \\ &= z_0\mathrm{i}\varepsilon + \text{ higher order terms} \end{aligned}$$

So $\mu(z_0) + h = \mu(z_0 + z_0\mathrm{i}\varepsilon)$, and the point $z_0 + h$ on the circle is mapped optically to the same point as $z_0 + z_0\mathrm{i}\varepsilon$.

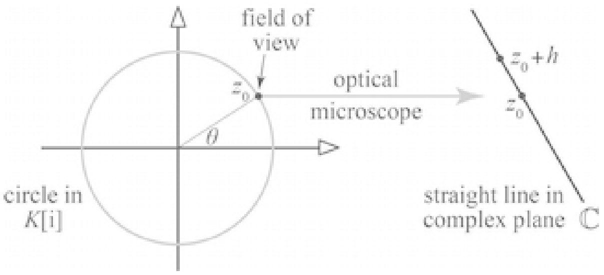

Figure 15.14 An infinitesimal part of a circle under optical magnification.

Multiplying by $\mathrm{i}\varepsilon$ rotates the plane through a right angle (multiplying by i) and scales by the real factor ε, so $z_0 + z_0\mathrm{i}\varepsilon$ is on the tangent at right angles to the radius from 0 to z_0. Furthermore, for any $\lambda \in \mathbb{R}$, taking $\varepsilon = \lambda\delta$ maps $z_0 + z_0\mathrm{i}\varepsilon$ onto a general point $z_0 + z_0\mathrm{i}\lambda$ on the tangent, so that the optical magnification transforms the part of the unit circle in the field of view of μ onto the whole tangent line.

In this very precise sense, an infinitesimal portion of the graph of a circle is 'locally straight' and is indistinguishable from an infinitesimal part of the tangent line when viewed through an optical microscope. □

15.9 Non-standard Analysis and Hyperreals

In the early days of calculus, functions were given by formulas. When the formula is a polynomial or a power series – which includes functions such as exponentials and trigonometric functions – the value of $f(x + h)$ for infinitesimal h can be calculated in the superreal numbers $\mathbb{R}_\varepsilon$ in the real case, or the supercomplex numbers $\mathbb{C}_\varepsilon$ in the complex case. The derivative may then be calculated as

$$f'(x) = \mathrm{st}\left(\frac{f(x+h) - f(x)}{h}\right) \quad \text{for infinitesimal } h$$

The fields $\mathbb{R}_\varepsilon$ and $\mathbb{C}_\varepsilon$ are sufficient for differentiation. By the Fundamental Theorem of Calculus, they are also sufficient for integration, using an antiderivative. However, there is no mechanism in these fields to deal with more general functions such as sequences $s : \mathbb{N} \to \mathbb{R}$, and they do not permit a suitable definition of the Riemann integral. When the real numbers were formulated set-theoretically at the end of the nineteenth century, infinitesimals had no place in formal epsilon-delta analysis.

In 1966, Abraham Robinson [17] made the formal use of infinitesimals rigorous using mathematical logic. He introduced an extension field $^*\mathbb{R}$ of $\mathbb{R}$, the *hyperreal numbers*, which includes infinitesimals and provides natural extensions of all of the basic functions of analysis. It is based on a distinction between two different types of logical statement, as follows.

The axioms for arithmetic and order in $\mathbb{R}$ quantify properties of *elements* of $\mathbb{R}$. These include

$$\text{for all } x, y \in \mathbb{R} : x + y = y + x$$

or

$$\text{there exists } 0 \in \mathbb{R} \text{ such that: for all } x \in \mathbb{R}, x + 0 = x$$

or

$$\text{for all } x, y, z \in \mathbb{R} : x > y, y > z \text{ implies } x > z$$

Just one axiom – the axiom of completeness – quantifies properties of *sets*:

$$\text{for all sets } S \subseteq \mathbb{R}, \text{ if } S \text{ is non-empty and bounded above,}$$
$$\text{then it has a least upper bound.}$$

Corollary 15.19 proves that super ordered fields are not complete, because monads are bounded above but fail to have least upper bounds. This is a subtle clue. When Robinson introduced the hyperreal numbers, he observed that properties that quantify elements (first order logic) generalise to ${}^*\mathbb{R}$, but properties quantifying sets (second order logic) may not. Writing properties using the universal quantifier $\forall$ (meaning 'for all') and $\exists$ ('there exists'), most properties of arithmetic quantify elements, such as

$$\forall x, y, \in \mathbb{R} : x + y = y + x$$

or

$$\exists 0 \in \mathbb{R} : \forall x \in \mathbb{R} : x + 0 = x$$

These use first order logic, but the completeness axiom – which quantifies sets – does not.

The solution is to find a system of hyperreal numbers ${}^*\mathbb{R}$ for which *any* first order logical statement that is true for the real numbers $\mathbb{R}$ remains true in the extended system ${}^*\mathbb{R}$. For example

$$\forall x, y, \in {}^*\mathbb{R} : x + y = y + x$$

or

$$\exists 0 \in {}^*\mathbb{R} : \forall x \in {}^*R : x + 0 = x.$$

remain true.

However, second order logical statements, such as the completeness axiom, need not extend to ${}^*\mathbb{R}$.

DEFINITION 15.25. The system of *hyperreal numbers* ${}^*\mathbb{R}$ is a proper ordered field extension of $\mathbb{R}$ where every subset $D \subseteq \mathbb{R}$ has an extension $D \subseteq {}^*D \subseteq {}^*\mathbb{R}$ and every function $f : D \to \mathbb{R}$ has an extension $f : {}^*D \to \mathbb{R}$ that satisfies:

Transfer Princple: Every true statement about the real numbers $\mathbb{R}$ expressed in first order logic is true in the extended system ${}^*\mathbb{R}$. □

(It is convenient use the same symbol for the extended function as for f to avoid proliferation of *s. This should not cause confusion.) We talk of 'the' hyperreals, but the construction does not lead to a unique result; however, any of them does the same job, so we assume that a particular one has been selected.

There are two possible approaches to the hyperreal numbers. The first is to assume that such a system exists and then build up a theory based on the definition. The second is to provide a formal existence proof, or better, a construction, of a hyperreal number system.

The first approach, though it does not prove the actual existence of hyperreal numbers, has great potential. For example, assuming its existence, it is possible to prove that the extension $^*\mathbb{N}$ of the natural numbers $\mathbb{N}$ must have infinite elements. The proof is simple. The first order statement

$$\forall x \in \mathbb{R}\ \exists n \in \mathbb{N} : n > x$$

states that for every real number x, there is always a bigger natural number n. Its extension says that

$$\forall x \in {}^*\mathbb{R}\ \exists n \in {}^*\mathbb{N} : n > x$$

and, since we know that $^*\mathbb{R}$ contains infinite elements, so does $^*\mathbb{N}$.

This immediately gives a new way of thinking about the limit of a sequence as a function $s : \mathbb{N} \to \mathbb{R}$ that extends to $s : {}^*\mathbb{N} \to {}^*\mathbb{R}$. Consider the elements $s(N)$ for infinite N and take the standard part $\mathrm{st}\,(s(N))$. If this is the same for all infinite N, then this common value is the limit of the sequence.

For instance, if

$$s_n = s(n) = \frac{6n^2 + n}{3n^2 + 1}$$

then the limit is

$$\mathrm{st}\, s(N) = \mathrm{st}\left(\frac{6N^2 + N}{3N^2 + 1}\right) = \mathrm{st}\left(\frac{6 + 1/N}{3 + 1/N^2}\right) = \frac{6 + 0}{3 + 0} = 2$$

Limits still have to be calculated, but the definitions become simpler. For instance, the usual definition of the limit s of a sequence s_n of real numbers is

$$\forall \varepsilon > 0\ \exists N \in \mathbb{N} : n > N \Rightarrow |s_n - s| < \varepsilon$$

Using the hyperreals, this becomes, as just explained:

$$\text{For all infinite } n \in {}^*\mathbb{N} : s = \mathrm{st}\,(s_n)$$

In general, definitions using the hyperreals (called non-standard analysis) involve fewer quantifiers than the standard definition. For instance, pointwise continuity of $f : D \to \mathbb{R}$ has a standard definition:

$$\forall x \in D\ \forall \varepsilon \in \mathbb{R}\ (\varepsilon > 0)\ \exists \delta \in \mathbb{R}\ (\delta > 0)\ \forall x' \in D : |x - x'| < \delta \Rightarrow |f(x) - f(x')| < \varepsilon$$

and uniform continuity of $f : D \to \mathbb{R}$ has a standard definition:

$$\forall x \in D\ \forall x' \in D\ \forall \varepsilon \in \mathbb{R}\ (\varepsilon > 0)\ \exists \delta \in \mathbb{R}\ (\delta > 0) : |x - x'| < \delta \Rightarrow |f(X) - f(x')| < \varepsilon$$

Each has four quantifiers that are highly complicated for the human mind to hold and manipulate all at once. The corresponding non-standard definitions are, for pointwise continuity:

$$\forall x \in D\ \forall x' \in {}^*D : x - x' \text{ infinitesimal } \Rightarrow f(x) - f(x') \text{ infinitesimal}$$

and for uniform continuity:

$$\forall x, x' \in {}^*D : x - x' \text{ infinitesimal } \Rightarrow f(x) - f(x') \text{ infinitesimal}$$

which have only two quantifiers and are much simpler.

15.10 Outline of the Construction of Hyperreal Numbers

It is possible to use the idea of hyperreals based on the transfer principle, without proving that hyperreals exist. But to put the idea on a sound logical basis, it is important to show that such a system can be constructed from known mathematics.

Broadly speaking, the construction of $^*\mathbb{R}$ from $\mathbb{R}$ follows a similar overall strategy to the one that Cantor used to construct the real numbers $\mathbb{R}$ from the rational numbers $\mathbb{Q}$. His construction begins with the set of all Cauchy sequences (a_n) of rational numbers; where a sequence is Cauchy if for any $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that $m, n > N \Rightarrow |a_m - a_n| < \varepsilon$. Then he sets up an equivalence relation:

$$(a_n) \sim (b_n) \iff (a_n - b_n) \to 0$$

Then $\mathbb{R}$ is defined to be the set of equivalence classes of Cauchy sequences, and the operations of arithmetic and order are derived from operations on Cauchy sequences. The construction of the hyperreal numbers $^*\mathbb{R}$ from $\mathbb{R}$ follows a similar overall pattern. It begins with all sequences (r_n) of real numbers, sets up an appropriate (and cunningly defined) equivalence relation $\sim$, and defines $^*\mathbb{R}$ to be the set of equivalence classes. The field $^*\mathbb{R}$ is considered as an extension of $\mathbb{R}$ by identifying the real number r with the equivalence class $[r_n]$, where $r_n = r$ for all n. The detailed construction requires the notion of an 'ultrafilter', akin to the axiom of choice. For a more detailed discussion, see Robinson [17] or Stewart and Tall [21].

Once the equivalence relation is set up, extending real sets and functions to hyperreal sets and functions is exquisitely simple. The extension of any subset $S \subseteq \mathbb{R}$ to $^*S \subseteq {}^*\mathbb{R}$ is defined by:

$$^*S = \{[r_n] \in {}^*\mathbb{R} : r_n \in S \text{ for all } n \in \mathbb{N}\}$$

The extension set $^*S$ is precisely the set of equivalence classes of sequences whose terms all lie in S. The extension of any function $f : S \to \mathbb{R}$ to $f : {}^*S \to {}^*\mathbb{R}$ is equally straightforward. Just define

$$f([r_n]) = [f(r_n)] \text{ for all } [r_n] \in {}^*S$$

If subsets of the real numbers or real functions are formulated set-theoretically using statements that quantify only elements, then those sets and functions can be extended to the hyperreal numbers using the same statements. For instance, if $S = \{x \in \mathbb{R} : x \geq 0\}$ and $f(x) = \sqrt{x}$, then $^*S = \{x \in {}^*\mathbb{R} : x \geq 0\}$. Because every $x \in S$ satisfies $x \geq 0$ and $f(x)^2 = x$, it follows that for every $x \in {}^*S$, the same equation holds, so $f(x)$ is the (positive) square root of x. For a positive infinitesimal x, this means that $\sqrt{x}$ is also a positive infinitesimal whose square is x.

The theory also extends to more general real functions $f : S \to \mathbb{R}^n$, where $S \subseteq \mathbb{R}^m$, to give an extension $f : {}^*S \to {}^*\mathbb{R}^n$ that satisfies the transfer theorem for logical statements quantifying several elements in $\mathbb{R}$. In particular, since every complex function $w = f(z)$ can be expressed as a real function of two variables taking (x, y) to (u, v), where $z = x + \mathrm{i}y$ and $w = u + \mathrm{i}v$, the hyperreal theory generalises naturally to the complex case.

15.11 Hypercomplex Numbers

To treat complex analysis in a similar manner we need the complex analogue of ${}^*\mathbb{R}$. The *hypercomplex numbers* ${}^*\mathbb{C}$ are of the form $a + \mathrm{i}b$, where $a, b \in {}^*\mathbb{R}$ and $\mathrm{i}^2 = -1$. (As for hyperreals, the word 'the' assumes a particular choice for ${}^*\mathbb{R}$, hence for ${}^*\mathbb{C}$.) They generalise the real case so that any function $f : S \to \mathbb{C}$ extends to a function $f : {}^*S \to {}^*\mathbb{C}$ that satisfies the transfer principle: any logical statement quantifying elements in $\mathbb{C}$ remains true in ${}^*\mathbb{C}$. For example, suppose that a complex analytic function $f : D \to \mathbb{C}$ on a domain D is expressed as a power series $f(z_0 + h) = \sum_{n=0}^{\infty} a_n h^n$ for $h \in \mathbb{C}, |h| < R$. Then the extended function $f : {}^*D \to {}^*\mathbb{C}$ is given by the same formula for $h \in {}^*\mathbb{C}, |h| < R$. This makes the infinitesimal extension of complex analysis simpler than the real case.

In particular, for any infinitesimal $\varepsilon \in \mathbb{C}$, the values of functions can be calculated using power series in ε, so we need only work in the field $\mathbb{C}_\varepsilon$. This, of course, is a subfield of the hypercomplex numbers ${}^*\mathbb{C}$, defined in the same way as $\mathbb{R}_\varepsilon$ but using complex coefficients in the power series in ε.

For any infinitesimal δ, the optical δ-lens μ pointing at $z_0 \in \mathbb{C}$ is defined by

$$\mu(z) = \operatorname{st} \frac{z - z_0}{\delta}$$

Its field of view is the set of elements $z_0 + h$ where h/δ is finite, so $h = \delta k$ where k is finite. This lies within the monad M_{z_0} of points lying an infinitesimal distance from z_0.

The field of view need not be the whole of the monad. For instance, if we take any infinitesimal ε and choose $\delta = \varepsilon^2$ then $z_0 + \varepsilon$ does not lie in the field of view of the optical δ-lens μ because $\varepsilon/\delta = 1/\varepsilon$ is infinite. Essentially the optical microscope lets us see detail differing from z_0 by the same order as δ, but lower order detail is too small and higher order detail is too far away to be seen in an optical image.

Writing $1/\delta$ in polar coordinates as $1/\delta = r\mathrm{e}^{\mathrm{i}\theta}$, the lens magnifies distances from z_0 by the (infinite) factor r and rotates through the angle θ. Choosing δ to be any positive infinitesimal in ${}^*\mathbb{R}$ gives $\theta = 0$ so that the optical microscope μ magnifies the picture by a factor r without any rotation. For this reason, when using an optical microscope we take the magnification factor δ to be a positive infinitesimal in ${}^*\mathbb{R}$.

DEFINITION 15.26. A subset $S \subseteq {}^*\mathbb{C}$ is called an *infinitesimal δ-shape near $z_0 \in \mathbb{C}$* if it is of the form $S = \{z_0 + h : z_0 \in \mathbb{C}, h \text{ is a } \delta\text{-infinitesimal}\}$. □

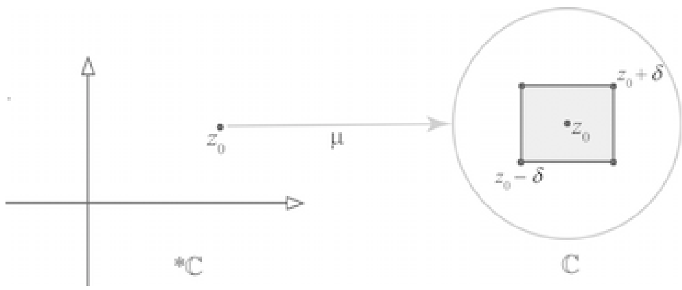

Figure 15.15 Magnifying an infinitesimal δ-shape using an optical δ-microscope.

Example 15.27. For an infinitesimal $\delta \in {}^*\mathbb{R}$, the subset

$$\{z_0 + h \in {}^*\mathbb{C} : z_0 \in \mathbb{C}, -\delta \leq h \leq \delta\}$$

is an infinitesimal δ-shaped square in ${}^*\mathbb{C}$ of side 2δ, centred on z_0. Using the standard convention that the image $\mu(z)$ is also denoted by the symbol z, this lets us see the infinitesimal shape in the complex plane $\mathbb{C}$ using the optical microscope μ, see Figure 15.15. □

An analytic function $f : D \to \mathbb{C}$ then extends naturally to $f : {}^*D \to {}^*\mathbb{C}$, and an infinitesimal shape near z_0 in ${}^*D$ transforms to an infinitesimal shape in ${}^*\mathbb{C}$ near $w_0 = f(z_0)$.

In the z-plane, let $z_0 + h$ be in the field of view of the optical microscope $\mu(z) = \operatorname{st}((z - z_0)/\delta)$. Then

$$\mu(z_0 + h) = \operatorname{st}(h/\delta)$$

Using an optical microscope $\nu(w) = \operatorname{st}((w - w_0)/\delta)$ in the w-plane with the same magnifying factor δ,

$$\begin{aligned} \nu(f(z_0 + h)) &= \operatorname{st}((f(z_0 + h) - f(z_0))/\delta) \\ &= \operatorname{st}(((f(z_0 + h) - f(z_0))/h)\operatorname{st}(h/\delta) \\ &= f'(z_0)\mu(z_0 + h) \end{aligned}$$

The complex number $\mu(z_0 + h)$ is transformed to its optical image $\nu(f(z_0 + h))$ multiplied by $f'(z_0)$. Writing $f'(z_0)$ in polar coordinates, $f'(z_0) = re^{i\theta}$, this scales the picture by a factor $r = |f'(z_0)|$ and turns it through an angle $\theta = \operatorname{arc}(f'(z_0))$, see Figure 15.16.

If $f'(z_0) \neq 0$, the transformation of the optical image is conformal, turning the infinitesimal shape precisely through the angle $\arg f'(z_0)$ and scaling lengths by the factor $|f'(z_0)|$.

If $f'(z_0) = 0$, the shape collapses to a single point. In this case, the detail in the w-plane involves infinitesimals of higher order. In Section 15.2 the Taylor series of f near z_0 is written in the form

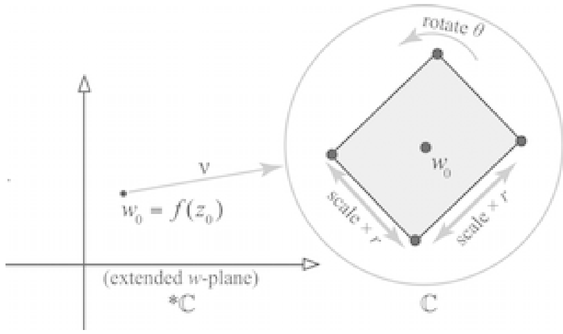

Figure 15.16 Magnifying a transformed infinitesimal shape.

$$f(z_0 + h) = f(z_0) + h^n f^{(n)}(z_0)/n! + \cdots$$

where $n \geq 1$ and $f^{(n)}(z_0) \neq 0$. This is interpreted as the sum of the constant $f(z_0)$ plus the infinitesimal $h^n f^{(n)}(z_0)/n! + \cdots$ of order n in the variable h.

Using an optical microscope of order δ^n gives

$$\nu(w) = \text{st}\,((w - w_0)/\delta^n)$$

so

$$\begin{aligned}\nu(f(z_0 + h)) &= \text{st}\,((f(z_0 + h) - f(z_0)/\delta^n)\\ &= \text{st}\,(h^n f^{(n)}(z_0))/n!/\delta^n)\\ &= \text{st}\,(h^n/\delta^n) f^{(n)}(z_0))/n!\\ &= \text{st}\,(h/\delta)^n f^{(n)}(z_0)/n!\\ &= \mu(z_0 + h)^n f^{(n)}(z_0)/n!\end{aligned}$$

Write $\mu(z_0 + h) = r\mathrm{e}^{\mathrm{i}\theta}$ and $f^{(n)}(z_0)/n! = k\mathrm{e}^{\mathrm{i}\alpha}$ in polar coordinates, to get

$$\nu(f(z_0 + r\mathrm{e}^{\mathrm{i}\theta})) = r^n \mathrm{e}^{\mathrm{i}n\theta} k\mathrm{e}^{\mathrm{i}\alpha}$$

so

$$|\nu(f(z_0 + r\mathrm{e}^{\mathrm{i}\theta}))| = kr^n$$

and

$$\text{arc}\,\nu(f(z_0 + h)) = n\theta + \alpha$$

This effect can be seen by starting with an infinitesimal δ-shape in the z-plane in the form of a sector of a circle centre z_0, with two infinitesimal sides of radius δ, at (finite) angles θ and $\theta + \phi$ to the real axis. The shape is then transformed in the w-plane as an infinitesimal sector that can be seen through an optical δ^n-lens as a sector of a circle, centre $w_0 = f(z_0)$, with the radius to $f(z_0 + h)$ turned through an angle $n\theta + \alpha$, the angle between the two radii stretched to $n\phi$, and the lengths of each radius scaled from r to kr^n, see Figure 15.17.

In the case of an infinitesimal δ-shape in the form of a circular path $f(\gamma(t))$ ($t \in [0, 2\pi]$) then, as the circle is traced once round the critical point, the image path $f\gamma$

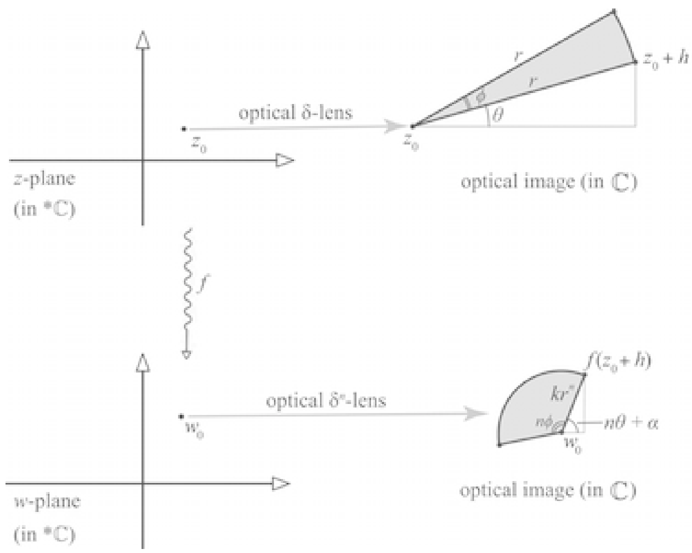

Figure 15.17 Transformation of an infinitesimal shape at a critical point of order n.

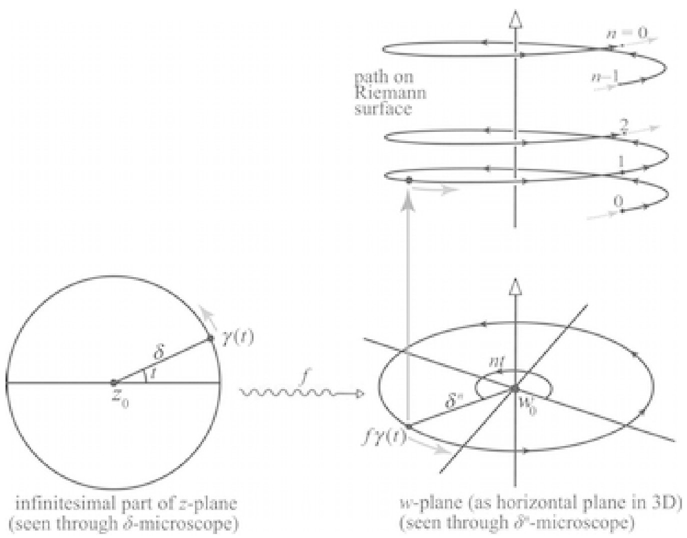

Figure 15.18 Optical image of a point moving round an infinitesimal circular path.

traces round the corresponding infinitesimal circle in the w-plane n times. This relates to the theory of Riemann surfaces, where a point z in the z-plane moves once round a critical point as its image $w = f\gamma(z)$ traces round the corresponding path in the w-plane n times. It is visualised in Figure 15.18 with part of the z-plane magnified by an optical δ-microscope pointed at z_0 and the corresponding part of the w-plane magnified by an optical δ^n-microscope pointed at w_0. In the figure the image in the w-plane is represented by a horizontal plane in three dimensions and the path on the Riemann surface directly above has n levels. As $\gamma(t)$ moves in the z-plane from 0 to $2\pi/n$ the corresponding point on the Riemann surface moves once round the initial level and then moves up on successive levels as t moves from $2r\pi/n$ to $2(r+1)\pi/n$, returning to the initial level when $t = 2\pi$.

This phenomenon applies to every point $z_0 \in S$. Furthermore, each infinitesimal δ-circle lies within the monad M_{z_0}, and each monad contains just one complex number z_0. If $z \neq z_0$ then an infinitesimal δ-circle around z lies in the distinct monad M_z. The behaviour of the function near any point $z \in \mathbb{C}$ therefore depends only on the order of infinitesimality of the function f at the point z.

This completes the picture of the behaviour of regular analytic functions from an infinitesimal viewpoint. An analytic function $f : D \to \mathbb{C}$ has a natural extension $f : {}^*D \to {}^*\mathbb{C}$ which generalises to a function from ${}^*D$ to the Riemann surface of f with multiple levels involving points in ${}^*\mathbb{C}$. This extension then has a unique analytic inverse from the Riemann surface back to ${}^*D$.

15.12 The Evolution of Meaning in Real and Complex Analysis

At this point we are in a position to look back on the historical evolution of real and complex analysis to gain a broader view of the current state of the theory as part of an ongoing development that will continue to evolve in the future. The historical path taken depends on the prevailing ideas at the time and may later be seen in an entirely different light. We review relevant parts of Chapter 0 and then move on to the modern era.

15.12.1 A Brief History

Initially the calculus developed by Leibniz and Newton focused on the relationship between variables represented as curves in the plane. On the continent, the infinitesimal methods of Leibniz prevailed, while in England, Newton began by thinking of a variable x changing in time, which he called a 'fluxion', and its rate of change x', which he called a 'fluent'. He also formulated the Binomial Theorem in the form

$$(x+o)^n = x^n + nx^{n-1}o + \tfrac{1}{2}n(n-1)x^{n-2}o^2 + \cdots$$

leading to the expansion

$$\frac{(x+o)^n - x^n}{o} = nx^{n-1} + o \cdot \tfrac{1}{2}n(n-1)x^{n-2} + \cdots$$

This holds good not only for a whole number n, but also for fractional and negative powers, which expresses $(x+o)^n$ as a power series in o, and as o becomes small, leads to the term in o being small enough to neglect, giving the derivative of x^n as nx^{n-1}.

In the eighteenth century, work on the calculus by Euler and others focused on real and complex power series manipulated as symbols to give significant results whose logical status would be questioned by later generations. It was not until the turn of the nineteenth century that Wessel, Argand, Gauss, and others began interpreting the complex numbers as points in the plane. Derivatives were calculated using the same formula as the real case, while complex integrals were taken along contours in the plane from z_0 to z_1. As in the real case, if f has an antiderivative F in its domain D then $\int_{z_0}^{z_1} f(z)\mathrm{d}z = F(z_1) - F(z_0)$, so the Fundamental Theorem of Calculus generalises directly. But then the theory of real and complex functions began to diverge spectacularly. In real analysis, functions with unexpected properties were found: functions differentiable once but not twice; functions differentiable n times but not $n+1$; and infinitely differentiable functions (such as $f(x) = e^{-1/x^2}$ for $x \neq 0$ and $f(0) = 0$) that do not equal their Taylor series expansion, even though it converges. There are even functions that are continuous everywhere and differentiable nowhere. Mathematicians realised that precise definitions and proofs were needed to cope with all these possibilities.

Meanwhile, it was discovered that a complex function f differentiable in a domain D can be represented as a power series in a small disc around any point in D. At the same time, the arithmetisation of analysis led to the set-theoretic definition of the real numbers as a complete ordered field, which cannot include infinitesimals.

This resulted in two parallel developments in the twentieth century with very different philosophies. Pure mathematics excluded infinitesimals because they did not fit into the real number system, while applied mathematicians used them informally but productively to model dynamic situations. Robinson's theory of non-standard analysis redeemed infinitesimals using mathematical logic. However, it was ignored by most applied mathematicians, who had no interest in that level of rigour, thought in terms of approximations, and were happy to discard any sufficiently small term if that proved fruitful, without asking further questions. It was actively resisted by many pure mathematicians, who had invested considerable effort in epsilon-delta analysis and preferred to remain with a familiar theory, or who preferred to avoid the subtle logical machinery of ultrafilters, with its reliance on the axiom of choice.

In principle, any theorem in standard real analysis that can be proved using non-standard analysis has a standard proof, so non-standard analysis may seem to offer no advantages. However, proofs using non-standard analysis are often simpler than epsilon-delta, once the initial investment of mastering the set-up has been made. Some important theorems currently have only non-standard proofs, because an epsilon-delta proof gets bogged down in too many ε's. So non-standard analysis has become an accepted technical tool in mainstream research. However, it is still a very specialised area.

15.12.2 Non-standard Analysis in Mathematics Education

The intuitive appeal of infinitesimals, as opposed to the difficult machinery of epsilon-delta methods, offers some educational advantages for students meeting analysis for the first time. Some programmes have used this approach successfully, but they have not been widely adopted.

The developments in this chapter reveal these apparently opposing views to be two sides of the same underlying coin. Applied mathematics focuses on variables that can 'become small' while pure mathematics focuses on fixed mathematical objects defined in epsilon-delta terms. Both can now be seen in a broader set-theoretical context in which real and complex numbers lie in larger structures that contain infinitesimals. The elements of such extension fields of $\mathbb{R}$ and $\mathbb{C}$ may be termed 'quantities'. A quantity

x is defined to be infinite if $|x| > r$ for all $r \in \mathbb{R}$. It is finite if it is not infinite, and infinitesimal if $1/x$ is infinite. Elements in $\mathbb{R}$ are called 'real constants' and those in $\mathbb{C}$ are 'complex constants'. This gives a formal language to describe an extension of the number line or complex plane that resonates with the use of the term 'infinitesimal' in historical development.

The structure theorems prove that every finite quantity x is either a constant or uniquely of the form $x = a + \varepsilon$, where a is either a real or complex constant and ε is a (real or complex) infinitesimal. The unique constant a is called the standard part of x. Optical microscopes let us see infinitesimal detail as a genuine picture on a line or in the plane. This returns us to an intuitive human way of visualising infinitesimal detail that resonates with earlier conceptions of the calculus that were rejected in mathematical analysis at the beginning of the twentieth century.

In 1972 Stroyan [22] introduced the idea of infinitesimal microscopes and telescopes. These were used in Keisler's undergraduate text *Foundations of Infinitesimal Calculus* based on Robinson's non-standard analysis, Keisler [10]. In 1980 Tall [23] modified Stroyan's definition to include optical lenses and applied the theory to the simpler system of superreal numbers.

In his book *Visual Complex Analysis*, Needham [14] sees a complex analytic function $w = f(z)$ transforming the z-plane into the w-plane. Figure 15.19 shows $f(z) = z^2$ transforming a shape in the z-plane to its image in the w-plane.

The black T-shape in the z-plane is transformed visually into a T-shape in the w-plane that is scaled in size and rotated through an angle. The scaling and rotation become more precise as the shape becomes smaller. In particular, near a point z where $f'(z) \neq 0$, a small shape is scaled approximately by a factor $|f'(z)|$ and turned through the angle $\arg f'(z)$. Needham introduced the term 'amplitwist' to indicate that the transformation 'amplifies' the scale by a factor $|f'(z)|$ and 'twists' the shape through the angle $\arg f'(z)$. He used the term 'infinitesimal' in a technical way that 'does not refer to some mystical, infinitely small quantity'. Instead (in Needham [14], pages 20–21) he suggested 'two intuitive ways of speaking', formulated as follows:

> ... if two quantities X and Y depend on a third quantity δ, then
> $\lim_{\delta\to 0} \frac{X}{Y} = 1 \iff$ '$X = Y$ for infinitesimal δ'
> $\iff$ X and Y are ultimately equal as δ tends to zero.

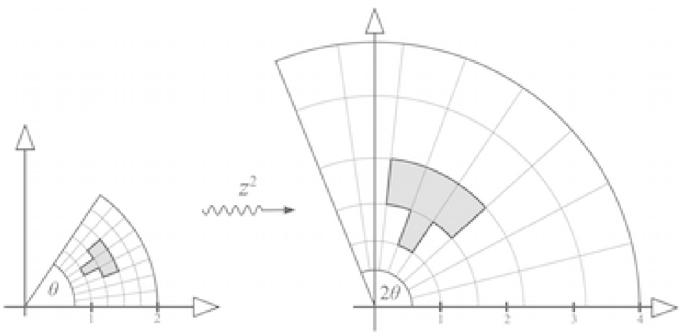

Figure 15.19 A visual transformation from the z-plane to the w-plane.

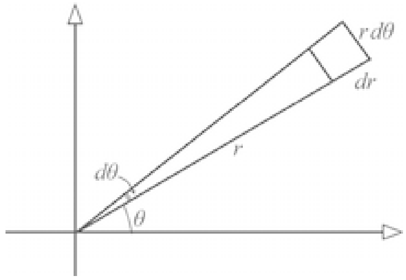

Figure 15.20 Polar coordinates r, θ, with increments dr, dθ.

This does not explicitly say what an 'infinitesimal' is. The approach in the current chapter defines an infinitesimal formally as an element in an ordered field extension K of the real numbers or in the corresponding complex extension field $K[\mathrm{i}]$.

15.12.3 Human Visual Senses

Thinking of an infinitesimal as a variable that is 'becoming small' is a natural human way to visualise infinitesimal ideas. Our eyes and brains have been wired by evolution to detect moving objects. This way of thinking also has practical advantages over thinking of an infinitesimal only as a tiny physical quantity. For example, using polar coordinates, Figure 15.20 shows (r, θ) being given a further increment dr in the radius and dθ in the angle.

If $r, \theta, \mathrm{d}r, \mathrm{d}\theta$ are all finite quantities, we can draw this as a finite picture. But if r and θ are finite and $\mathrm{d}r, \mathrm{d}\theta$ are infinitesimal, we cannot draw the whole picture to the same scale. Using an optical δ-lens to magnify the infinitesimal shape sides dr by $r\mathrm{d}\theta$ to give a finite picture requires $\mathrm{d}r, \mathrm{d}\theta$, and δ all to have the same order. If $r \neq 0$, then r has higher order than δ, so if a δ-lens is used magnify the shape dr to reveal a finite picture, the origin, distance r away, lies outside the field of view. On the other hand, if r is drawn to a finite scale, the optical picture of the infinitesimal shape reduces to a single point.

This means that while we may imagine infinitesimals as arbitrarily small fixed elements in an extension field of the real or complex numbers, when making a physical drawing, it is more natural to think of them as variable quantities that become very small.

Recent research into how the human eye works shows that imagining points moving smoothly is a natural human facility. When we read text, our eye has a sharp focus which only reads a few consecutive letters (around four or five with standard size type) and our eye jumps along the line taking in information, attempting to build up the meaning of successive stretches of text. If you look at a paragraph on this page and focus on how you read it, you can sense these jumps (called saccades) as the eye moves along the text. (Try it now.) However, if you look at an object moving smoothly, there is an initial jump to lock on to the object in the sharp central vision, and the eye then follows the object in a smooth movement. This means that it is natural to imagine a point moving smoothly along a line.

15.12.4 Computer Graphics

New technologies are evolving which let users dynamically control movement on a high-resolution visual display, smoothly zooming in on graphs or moving objects around. Software such as *Mathematica* and *GeoGebra* let a human operator code mathematical ideas symbolically, and manipulate pictures dynamically to explore mathematical relationships.

Placing the w-plane above the z-plane lets a three-dimensional view be drawn on a two-dimensional dynamic display, where $z = x + \mathrm{i}y$ is represented on the z-plane as $(x, y, 0)$ and $w = u + \mathrm{i}v$ is represented on the w-plane as $(u, v, 1)$. As $z = x + \mathrm{i}y$ moves around in the z-plane, the corresponding value $w = f(z) = u + \mathrm{i}v$ moves around in the w-plane. The transformation from z to w can be signified by an arrow linking z to $f(z)$. As z varies, this gives a mapping diagram from the z-plane to the w-plane as shown earlier in the lower part of Figure 15.4.

Figure 15.21 shows a display created in *GeoGebra* (see Flashman [6] to explore the dynamic software on the internet). It is in three framed areas. The central area shows the mapping diagram for the function $f(z) = z^3 + pz + q$ from the z-plane to the w-plane for $p = -1.9, q = 3.3$. On the left is an area showing the z-plane and a circle radius $\delta = 1.0$ with $m = 24$ points equally spaced around it. The area on the right shows the graph of the real function $y = x^3 + px + q$. On the screen are boxes to input the formula for the function f and the variables p and q, as well as sliders with movable points to vary p, q, δ, and m. The central area can also be selected to change the viewpoint or the scale.

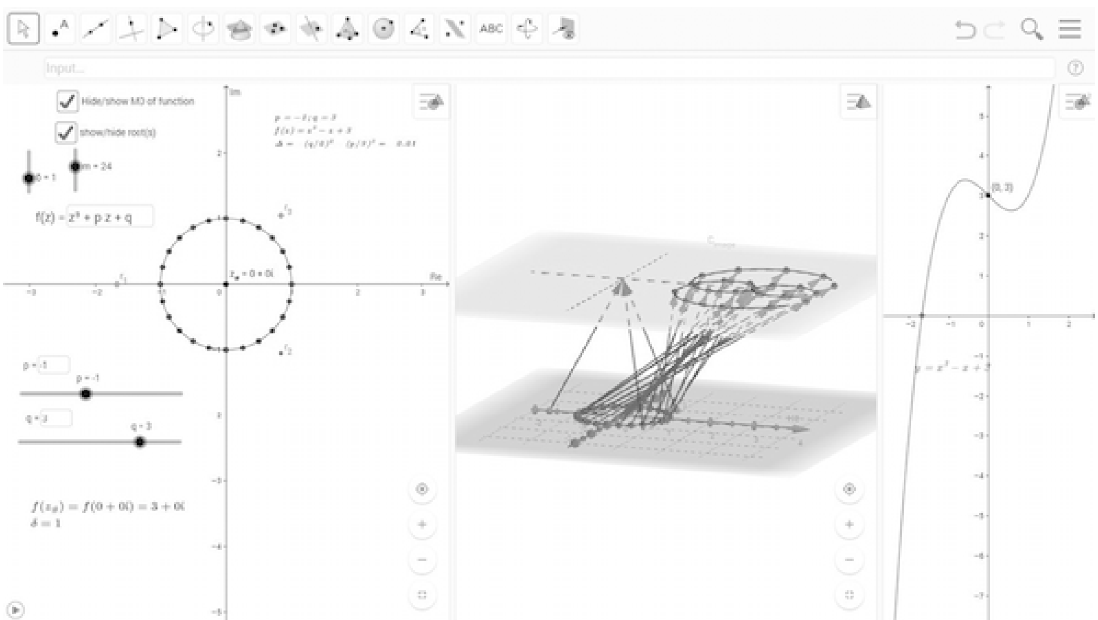

Figure 15.21 Using *Geogebra* to explore the cubic function $f(z) = z^3 + pz + q$. You can interact with this worksheet at www.cambridge.org/gb/S&T-geogebra

New possibilities arise that cannot be adequately described using a static picture in a textbook. The design provides two boxes to tick: one (MD) hides or shows the Mapping Diagram, which consists of the arrows from the points around the circle radius δ in the z-plane to the w-plane; the other shows or hides the roots of the equation $f(z) = 0$.

The three roots marked r_1, r_2, r_3 are shown in the panel on the left, where r_3 is on the real axis, and r_1, r_2 are complex conjugates. Moving q up or down moves the real graph in the right frame dynamically up or down by the same amount. The graph as pictured has one real root near -2, and two complex conjugates near $1 \pm \mathrm{i}$. As q decreases, the real cubic curve moves down by the same amount until the function has three real roots. As this happens, the roots r_1, r_2 in the left frame move towards each other until they coalesce as a repeated real root, then separate into two distinct real roots, so that the cubic has three real roots.

The user can manipulate the picture dynamically to focus on aspects of interest. For example, the left panel has an icon ▶ which, when clicked, causes the chosen parameter (here p) to vary dynamically over its given range (here -5 to 5) and, as it does so, it displays the changing situation of the roots of the real cubic as the equation shifts between having one real and two complex conjugate roots to having three real roots.

The *GeoGebra* environment allows the user to create and explore dynamic figures in complex analysis, Flashman [5].

15.12.5 Summary

Interpretations of complex analysis continue to evolve into the future. New approaches in modern times are part of a more comprehensive overall picture in which intuition may be translated into rigour and conversely, proving structure theorems returns formal rigour to more sophisticated forms of symbolic manipulation and visual intuition.

This broader vision reveals the fundamental distinction between real and complex analysis. Real analysis considers functions that exhibit many different properties and require highly sophisticated construction of the hyperreal numbers to cope with the logical introduction of infinitesimals. Meanwhile, complex differentiable functions are given locally by power series around any point in a connected open domain and these can be manipulated in the much simpler extension field $\mathbb{C}_\varepsilon$ generated by a single infinitesimal ε. This simpler extension encompasses the more powerful theory of analytic continuation and Riemann surfaces. From this higher vantage point we can see, at a single glance, the distinction between the technically complicated theory of real analysis and the subtle sophistication of complex analysis.

15.13 Exercises

1. Figure 15.22 shows the path $\gamma : [-\pi, \pi] \to \mathbb{C}$ given by $\gamma(t) = t + \mathrm{i}t$ and the analytic function $w = f(z) = \cos z$. Calculate $f\gamma, \gamma'(t)$, and $(f\gamma)'(t)$ as functions of t. Interpreting t as time, consider the points $\gamma(t)$ and $(f\gamma)(t)$ moving along the paths $\operatorname{im}\gamma, \operatorname{im} f\gamma$. By comparing the motion in the w-plane in the intervals $-\pi \le t \le 0$

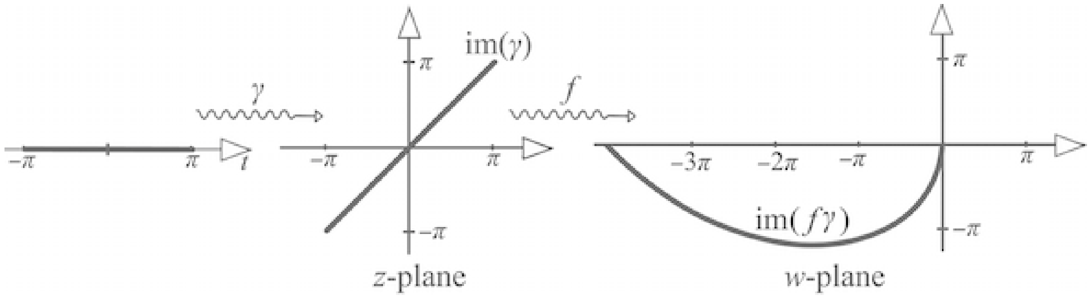

Figure 15.22 Geometry for Exercise 4.

and $0 \leq t \leq \pi$, or otherwise, explain what happens in the w-plane as t passes through 0.

2. Let $\gamma : [-1, 1] \to \mathbb{C}$ be given by $\gamma(t) = t^3 + \mathrm{i}t^2$. Use a graph plotter to draw the parametric graph of (t^3, t^2) and magnify the graph at various points. Conjecture what the highly magnified portions of the graph will look like for $t = 0$ and $t \neq 0$. Calculate $\gamma(t)$ numerically for $t = -0.001, t = 0, t = 0.001$ and draw the graph for $-0.001 \leq t \leq 0.001$ on a standard piece of paper. What happens if the graph is scaled from $t = -10^{-6}$ to $+10^{-6}$ on a standard piece of paper?

3. For $f(z) = z^2$, plot the paths $\gamma_1(t) = 1 + \mathrm{i}t\ (-1 \leq t \leq 1)$, $\gamma_2(t) = -1 + \mathrm{i}t\ (-1 \leq t \leq 1)$, $\gamma_3(t) = \mathrm{i}t\ (-1 \leq t \leq 1)$ in the z-plane and their transformed paths $f\gamma_1, f\gamma_2, f\gamma_3$ in the w-plane where $w = f(z)$. Imagine t increasing smoothly in time from $t = -1$ to $t = 1$ and trace the movement of the corresponding point along each graph with your finger. One has a different behaviour from the other two. Which one? Explain why, by considering the critical point of f.

4. Determine the Taylor expansion $f(z_0 + h)$ of the following functions about a point z_0 and deduce the order of the variable part $f(z_0 + h) - f(z_0)$ at that point:

(i) $\sin z - z$ at $z_0 = 0$
(ii) $\cos z$ at $z_0 = 0$ and at $z_0 = \mathrm{i}$
(iii) z^5 at $z_0 = 0$ and at $z_0 \neq 0$
(iv) The principal value $\mathrm{Log} z$ at $z = 1$

5. Let K be a proper ordered field extension of the real numbers $\mathbb{R}$. For $x \in K$, write down the definition for x to be finite, positive infinite, negative infinite, infinite, positive infinitesimal, negative infinitesimal, infinitesimal. In each of the following cases, decide whether the given statement is true or false for all $x, y \in K$. If it is true, give a proof; if it is false, provide a counterexample (for instance, using the field $K = \mathbb{R}(v)$ of rational functions in a variable v).

(i) x negative infinite implies $1/x$ is negative infinitesimal
(ii) x infinite implies $1/x$ is infinitesimal
(iii) x infinitesimal implies $1/x$ is infinite
(iv) x finite, y infinitesimal implies xy is finite
(v) x finite and non-zero, y infinitesimal implies xy is infinitesimal

(vi) x infinitesimal, y infinitesimal implies x/y is infinitesimal
(vii) x finite, y infinite implies xy is infinite

6. For $x \in K[\mathrm{i}]$ where $\mathrm{i}^2 = -1$, write down the definition for x to be finite, infinite, infinitesimal. In each of the following cases, decide whether the given statement is true or false for all $x, y \in K[\mathrm{i}]$. If it is true, give a proof; if it is false, provide a counterexample.
(i) x infinite implies $1/x$ is infinitesimal
(ii)) x infinitesimal implies $1/x$ is infinite
(iii) x finite, y infinitesimal implies xy is finite
(iv) x finite, y infinitesimal implies xy is infinitesimal
(v) x infinitesimal, y infinitesimal implies x/y is infinitesimal
(vi) x finite, y infinite implies xy is infinite.

7. In the field $K = \mathbb{R}(v)$ of rational functions in a variable v, place the following elements in linear order and in order of infinitesimality:

$$v, 1 - v, 1/(1 - v), -1/v, v/(v - 1)$$

8. Prove Proposition 15.11 that if F is the subset of finite elements in a super ordered field K and I is the subset of elements that are zero or infinitesimal, then

$$a, b \in F \text{ and } c, d \in I \text{ implies } a + b, a - b, ab \in F, c + d, c - d, cd \in I$$

9. Prove Proposition 15.12 that if $x, y \in F$, where F is the subset of finite elements in a super ordered field, then the standard part map st satisfies

$$\begin{aligned}
\mathrm{st}\,(x + y) &= \mathrm{st}\,(x) + \mathrm{st}\,(y),\\
\mathrm{st}\,(x - y) &= \mathrm{st}\,(x) - \mathrm{st}\,(y),\\
\mathrm{st}\,(xy) &= \mathrm{st}\,(x)\,\mathrm{st}\,(y)\\
\mathrm{st}\,(x/y) &= \mathrm{st}\,(x)/\mathrm{st}\,(y) \text{ (for st}\,(y) \neq 0).
\end{aligned}$$

10. Calculate the coefficients $c_{m+n}, c_{m+n+1}, c_{m+n+2}, c_{m+n+3}$ which arise when division

$$\left(\sum_{r=m}^{\infty} a_r \varepsilon^r\right) / \left(\sum_{r=n}^{\infty} b_r \varepsilon^r\right) = \left(\sum_{r=m+n}^{\infty} c_r \varepsilon^r\right)$$

in the superreal field $\mathbb{R}_\varepsilon$ and the supercomplex field $\mathbb{C}_\varepsilon$, is determined by solving the equation

$$(c_{m+n}\varepsilon^{m+n} + \cdots + (c_r\varepsilon^r + \cdots)(b_n\varepsilon^n + \cdots + (b_r\varepsilon^r + \cdots) = (a_m\varepsilon^m + \cdots + (a_r\varepsilon^r + \cdots)$$

for successive coefficients $c_{m+n}, c_{m+n+1}, \ldots$.

11. Show that there are some positive infinitesimals $\delta \in \mathbb{R}_\varepsilon$ that fail to have a square root in $\mathbb{R}_\varepsilon$. Nevertheless, show that, for any $z = x + \mathrm{i}y \in \mathbb{C}_\varepsilon$ where $x, y \in \mathbb{R}_\varepsilon$, the modulus $|z| = \sqrt{x^2 + y^2}$ can be calculated as an element in $\mathbb{R}_\varepsilon$. (Hint: each of x, y is either zero, or of the form $a\varepsilon^n(1 + \delta)$, where δ is infinitesimal.)

12. For an infinitesimal δ, calculate which of the following are finite and which are infinite (using power series for trigonometric or exponential functions where necessary):

(i) $3\delta + 1/\delta$
(ii) $\sin(\mathrm{i}\pi + \delta)$
(iii) $\delta^2 \sin(1/\delta)$
(iv) $\sin(\mathrm{i}\pi + \delta)$
(v) $\mathrm{e}^{\mathrm{i}\pi+\delta}$
(vi) $e^{\mathrm{i}\pi+k\delta}$ for $k \in \mathbb{C}$

In those cases where the element is finite, calculate its standard part.

13. For $f : D \to \mathbb{C}$ (where D is a domain in $\mathbb{C}$) given by the following formulas, calculate the standard part of $(f(z+h) - f(z))/h$ in a super complex field, where h is an infinitesimal, and hence find the derivative $f'(z)$ for $z \in D$.

(i) $z^3 - 2z^2$
(ii) $1/(z-2)$ for $z \neq 2$
(iii) $(1+z^3)^n$ for $n \in \mathbb{N}$
(iv) $\sin z$ (in a super complex field including power series expansions)
(v) $z^2 \sin z$

14. Write down the definition for hyperreal numbers $^*\mathbb{R}$, hypercomplex numbers $^*\mathbb{C}$ and the transfer principle for hyperreal and hypercomplex numbers.

(i) Use the first order statement $\forall x \in \mathbb{R}\ \exists n \in \mathbb{N} : n > x$ to deduce that there exist infinite hypernatural numbers $N \in {}^*\mathbb{N}$, where $N \notin \mathbb{N}$.
(ii) By considering the property

$$\forall m, n \in \mathbb{N} : m > n \Rightarrow m - n \in \mathbb{N}$$

use the transfer principle to show that if m is an infinite hypernatural number and n is a natural number, then $m - n$ is an infinite hypernatural number.

(iii) Let $E = \{2n \in \mathbb{N} : n \in \mathbb{N}\}$, $O = \{2n-1 \in \mathbb{N} : n \in \mathbb{N}\}$ be the set of even and odd natural numbers. Use the transfer principle to show that every infinite hypernatural number N is either even (in $^*E$) or odd (in $^*O$) and that N is even if and only if $N+1$ is odd.

15. Using the hypercomplex numbers, by considering the standard part of z_N for infinite N, determine whether the sequence (z_n) converges. If it does, calculate the limit, if not, explain why.

(i) $(n^3 + \sin n)/(n^3 - 2n^2)$
(ii) $1 + k + k^2 + \cdots + k^{n-1}$ for $k \in \mathbb{C}, k < 1$
(iii) $1 + \frac{1}{2} + (\frac{1}{2})^2 + \cdots + (\frac{1}{2})^{n-1}$ for n odd, $\dfrac{2n^3 + 3n - 17}{(n^2 + 3n)(n+1)}$ for n even
(iv) $\dfrac{n^3 + \sin n}{n^3 - 2n^2}$ for n odd, $2 + \dfrac{1}{n}$ for n even

16. Visit the *Geogebra* website www.geogebra.org/ and use it to investigate some of the curves in $\mathbb{C}$ and $\mathbb{R}^3$ discussed in this book.

16 Homology Version of Cauchy's Theorem

In Chapter 8 we proved the classical version of Cauchy's Theorem, using step paths to simplify the geometry and topology involved. In Chapter 9 we developed the concept of homotopy, and reformulated Cauchy's Theorem in that framework. Homotopy is one of the basic ideas in algebraic topology. A related topological concept, homology, emerged a little earlier from work on the classification of surfaces. This notion and its extensions are of great importance in algebraic topology. When applied to domains in $\mathbb{C}$, it provides fresh insight into Cauchy's Theorem and its consequences.

In complex analysis we view homology as a topological property of a domain $D \subseteq \mathbb{C}$, but we define it in terms of properties of paths in D. There are then deep connections between the homology of D and the behaviour of complex integration along paths in D. In this respect homology is like homotopy, which is also a topological property defined in terms of paths. The two concepts are related, but different. For homotopy, two paths in a domain are considered to be equivalent if one can be continuously deformed into the other. In the context of homology, two paths in a domain are considered to be equivalent if they differ by a path that forms the boundary ∂R of a rectangle R whose image lies in D, as described in Section 9.3 – and, by extension, if they differ by a finite set of such boundary paths.

In this chapter, we develop a homology version of Cauchy's Theorem, and examine a few of its consequences, including a homology version of Cauchy's Residue Theorem. The material here is a natural continuation from Chapter 9. We have postponed it until now to avoid delaying the more standard material in Chapters 10–14. The main ideas are simple, elegant, and geometrically intuitive, although the formal setting takes a little getting used to. Also, proofs have a habit of getting tangled up in combinatorial arguments about step paths. Here, algebra often comes to our rescue. The topology is fairly subtle, and results that seem obvious sometimes require a little care if we want a rigorous proof. We work throughout with step paths that lie in a domain $D \subseteq \mathbb{C}$. Recall that a domain, in this context, is *open* and *connected*; moreover, being connected is equivalent to being *path-connected*. Cauchy's Theorem concerns the integral of a complex function on D over a closed step path in D.

REMARK 16.1. For the figures in the chapter, we often do not draw paths explicitly as step paths, because lots of little steps everywhere just make the figures more complicated without adding any understanding. We draw the steps only when they matter. □

16.0.1 Outline of Chapter

We sketch the main ideas of this chapter now. Formal definitions and proofs come later. Throughout this chapter, we abbreviate 'step path' to 'path', except when we want to emphasise that we are working with step paths. Our terminology and approach to homology are tailored to the material in this text, and adapted to the simpler geometry of step paths. Topologists generally set up homology in a more general, more abstract, and more elegant manner, see for example Hatcher [8] or Hocking and Young [9].

Homology provides a natural context for several features of complex integrals that we have previously noticed:

- Integrals are additive: the integral of a given function along a sum of paths is the sum of its integrals along the separate paths.
- Moreover, the integral of the function along the reverse of a path is minus the integral along the path, allowing subtraction as well as addition.
- Cauchy's theorem for a boundary, Theorem 9.6, states that if R is a rectangle and $\phi : R \to D$ is continuous, and $f : D \to \mathbb{C}$ is differentiable, then the integral of f along the boundary of ϕ vanishes.
- On the other hand, the integral of f along an arbitrary closed path need not vanish. Its value depends on how the path winds around points not in D.

All of these facts are closely related, suggesting that they can all be put together in a natural context. The question is: what?

Our definition of a sum of paths in Section 2.8 involves joining them end to end, and we have to impose special conditions for that to be possible. This corresponds to the classical view that a function should be integrated 'along' a path. However, we can extend integration to any finite set of paths, even if they do not join up to create a single path, just by adding the separate integrals together. A path γ can be repeated any number of times to give a path $n\gamma$ for $n \in \mathbb{N} \setminus \{0\}$. For negative n we can define $n\gamma$ to be $-n$ times the reverse path $-\gamma$. Finally, the empty set of paths plays the role of 0. So we can define an arbitrary integer combination

$$m_1\gamma_1 + \cdots + m_n\gamma_n$$

of paths γ_j, for $m_j \in \mathbb{Z}$.

At first glance, these combinations seem to form a group (indeed, an abelian group) under addition, but there is a technical problem: if γ is a path, $\gamma + (-\gamma)$ ought to be the zero element of this group, which is empty. Actually, as described above, it is the set $\{\gamma, -\gamma\}$, which is not empty.

On the other hand, the integral of f along $\gamma + (-\gamma)$ is always zero, so this discrepancy has no effect on integrals. All of which suggests that we must somehow *redefine* $\gamma + (-\gamma)$ to be zero. Mathematicians have a standard trick to resolve this kind of difficulty: introduce an equivalence relation in which $\gamma + (-\gamma)$ is equivalent to 0, and work with the equivalence classes. The equivalence relation we need is homology, and the clue for how to define it is Cauchy's Theorem for a boundary. Essentially, we want any boundary path to be equivalent to zero. Everything else follows from that.

Since we are trying to construct a group, a natural way to achieve all this is to build in the group property right at the start. Then Cauchy's Theorem for a boundary tells us how to define the equivalence relation of homology using group theory, where it takes a very simple, though more abstract, form. To accomplish this, we define integration over 'chains'. A chain is a formal integer combination

$$\Gamma = m_1\gamma_1 + \cdots + m_n\gamma_n$$

where the γ_j are step paths (not necessarily closed) and the $m_j \in \mathbb{Z}$. We explain what this means, how to formalise it, and how to visualise it, in Section 16.1. When all $m_j = 0$ the chain is empty, and this possibility is allowed in the definition.

Be warned: addition here is not quite the same as addition of paths as defined in Section 2.4, and $-\gamma$ is not quite the same as the reverse path to γ. As we have just seen, the classical operations $+$ and $-$ run into technical problems, and we have to get round those. However, the classical notions and the abstract group operations are closely related. The precise relationship is given in Lemma 16.15: we can identify them 'modulo homology'.

The group of chains has two important subgroups. The first is the subgroup of 'cycles', which correspond to formal sums of *closed* step paths, which we call 'loops'. Inside the subgroup of cycles we define a further subgroup of 'boundaries'. A boundary is now a formal integer combination

$$\beta = m_1\beta_1 + \cdots + m_r\beta_r$$

where each $\beta_j = \partial R_j$ is the boundary of a rectangle in D, and the $m_j \in \mathbb{Z}$. Every boundary of a rectangle is a loop, so in particular boundaries are cycles.

By 'rectangle in D' we mean a continuous map $\phi : R \to D$ such that the image $\phi(R) \subseteq D$, just as in Cauchy's Theorem for a boundary. Chains, in contrast, are defined by the image $\phi(\partial R)$, and in that case the map is not required to be defined on the interior of R, let alone to have its image a subset of D. Indeed, the distinction between these two cases is central to homology.

To complete the set-up, two cycles Ω_1, Ω_2 are said to be 'homologous' if their difference $\Omega_1 - \Omega_2$ is a boundary. Another term for being a boundary is 'homologous to zero'.

The machinery of chains, cycles, boundaries, and homology is beautifully adapted to natural properties of complex integrals along paths. Let $f : D \to \mathbb{C}$ be a continuous function. Integration over paths can be generalised to integration over chains, hence also over cycles or boundaries, by defining

$$\int_\Gamma f = m_1 \int_{\gamma_1} f + \cdots + m_n \int_{\gamma_n} f$$

Boundaries come into play because we proved in Theorem 9.6 that the integral round the boundary of a rectangle whose image lies inside a domain is zero:

$$\int_{\partial R} f = 0$$

So the integral of f over a cycle Ω is unchanged if we add or subtract a series of boundaries of rectangles – that is, if we add a boundary to Ω. In other words, for any fixed f, the integral depends only on the homology class of the cycle.

We also need to define winding numbers for chains, which we do by defining

$$w(\Gamma, z_0) = m_1 w(\gamma_1, z_0) + \cdots + m_n w(\gamma_n, z_0)$$

We already know that the winding number for a path can be specified by an integral as in (7.6). This relationship also holds for chains.

Once these concepts have been set up, we come to the main aim of the chapter, which is to prove:

THEOREM 16.2 (Homology Version of Cauchy's Theorem). *Let $f : D \to \mathbb{C}$ be a continuous function, and let Ω be a cycle in D. Then the following are equivalent:*

(i) *Ω is homologous to zero (that is, it is a boundary) in D.*
(ii) *All integrals over Ω vanish:*

$$\int_\Omega f = 0 \quad \text{for all differentiable } f.$$

(iii) *All winding numbers of Ω around points outside D vanish:*

$$w(\Omega, z_0) = 0 \quad \text{for all } z_0 \notin D.$$

The proof of this theorem occupies most of this chapter. When the proof is complete, we derive some useful consequences.

16.0.2 Group-theoretic Interpretation

Homology can be interpreted group-theoretically, as Emmy Noether noticed when attending lectures on topology by Heinz Hopf in 1926–7. Leopold Vietoris and Walther Mayer independently discovered the same interpretation at much the same time. Chains $\mathcal{C}$, being formal integer combinations of specific objects (here paths), form an abelian group under formal addition. Cycles $\mathcal{Z}$ form a subgroup of $\mathcal{C}$, and boundaries form a subgroup $\mathcal{B}$ of $\mathcal{Z}$. Because the group of chains is abelian, these are normal subgroups. The equivalence relation of homology on $\mathcal{Z}$ determines a quotient group

$$\mathcal{H} = \mathcal{Z}/\mathcal{B} = \text{cycles/boundaries}$$

Specifically, the equivalence classes for homology are precisely the cosets of $\mathcal{B}$ in $\mathcal{Z}$.

All four groups $\mathcal{C}, \mathcal{Z}, \mathcal{B}$, and $\mathcal{H}$ depend on D. The group $\mathcal{H}$ is the *first homology group* of D, usually written $H_1(D)$. (In the topology of higher-dimensional spaces in $\mathbb{R}^n$ and $\mathbb{C}^n$ there are also 'higher' homology groups $H_k(D)$ for $k = 0, 1, 2, 3, \ldots$, but only $H_1(D)$ concerns us here.) The elements of $\mathcal{H}$ can be thought of as the 'essentially different' topological types of loops in D, up to homology.

With hindsight, our previous results related to Cauchy's Theorem offer strong hints at the existence of this algebraic structure. As already noted, integrals vanish on boundaries. Paths can be added to give new paths, and integrals are additive. So are winding numbers. Our aim is to develop a formal algebraic context that unifies these hints. Algebraic topology originated from these and related considerations.

To make the discussion as self-contained as possible, our approach will be classical (a polite way to say 'old-fashioned'). The modern approach to homology is more general and more abstract.

16.1 Chains

Let $D \subseteq \mathbb{C}$ be a domain. Recall that a (step) path in D is a map

$$\gamma : [a,b] \to D$$

with the step property (sum of finitely many line segments parallel to the real or imaginary axis). To keep the terminology simple, we define:

DEFINITION 16.3. A *loop* in D is a *closed* step path $\gamma : [a,b] \to D$, so $\gamma(a) = \gamma(b)$. □

Let

$$\mathcal{P} = \{\text{all paths in } D\}$$

We turn $\mathcal{P}$ into an abelian group using an algebraic trick (which can turn *any* set into a subset of an abelian group):

DEFINITION 16.4. A *chain* Γ in D is a formal integer combination

$$\Gamma = m_1\gamma_1 + \cdots + m_n\gamma_n \tag{16.1}$$

where the $\gamma_j \in \mathcal{P}$ are *distinct* paths in D and the $m_j \in \mathbb{Z}$.

The set of all chains is denoted by $\mathcal{C}$. □

A more respectable, but less convenient, definition views the m_j as the values of a function $m : \mathcal{P} \to \mathbb{Z}$, which maps γ_j to m_j. Thus we define

$$\mathcal{Z}(D) = \{m : \mathcal{P} \to \mathbb{Z} : m(\gamma) = 0 \text{ for all except a finite number of } \gamma \in \mathcal{P}\}$$

We can pass between functions m and formal sums (16.1) as follows:

(i) Given (16.1), define m by:

$$m(\gamma) = \begin{cases} m_j & \text{if} \quad \gamma = \gamma_j \\ 0 & \text{if} \quad \gamma \neq \gamma_j \text{ for any } j \end{cases}$$

(ii) Given m, define

$$\Gamma = \{\gamma \in \mathcal{L} : m(\gamma) \neq 0\}$$
$$m_j = m(\gamma_j) \quad \text{for } \gamma_j \in \Gamma$$

We can add two such functions:

$$(m_1 + m_2)(\gamma) = m_1(\gamma) + m_2(\gamma)$$

and we then have:

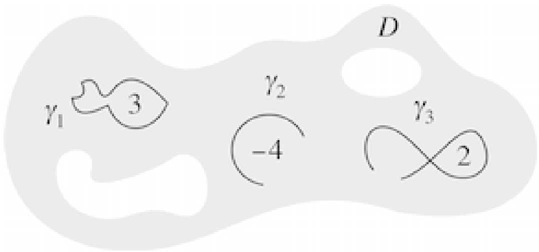

Figure 16.1 Visualising a chain. This is $\Gamma = 3\gamma_1 - 4\gamma_2 + 2\gamma_3$. It so happens that γ_1 is a loop.

PROPOSITION 16.5. *The set of chains $\mathcal{C}(D)$ is an abelian group under the operation $+$.*

Proof. The zero element of $\mathcal{C}(D)$ is the function m for which $m(\gamma) = 0$ for all $\gamma \in \mathcal{L}$.

The additive inverse of a function m is the function $-m$ defined by

$$(-m)(\gamma) = -(m(\gamma))$$

It is easy to see that $+$ is associative and commutative. □

Technically, $\mathcal{C}$ is the 'free abelian group' generated by the set of paths $\mathcal{P}$.

At first sight, it is not obvious how to visualise a *formal* sum of paths, especially one with integer coefficients. One way is to draw the images of the different paths γ_j in the sum, and label each by the corresponding 'multiplicity' m_j, as in Figure 16.1.

The *image* of a cycle $\Gamma = \sum m_j\gamma_j$ is defined to be the union of the images of the γ_j for those j with $m_j \neq 0$.

We extend the definition of the integral and the winding number from loops to chains:

DEFINITION 16.6. Let $\Gamma = \sum m_j\gamma_j$ be a chain, where the γ_j are paths and $z_0 \in \mathbb{C}$. Then define

(i)

$$\int_\Gamma f = \sum_j m_j \int_{\gamma_j} f$$

(ii)

$$w(\Gamma, z_0) = \sum_j m_j w(\gamma_j, z_0)$$ □

The definition trivially implies:

PROPOSITION 16.7. *The integral is additive:*

$$\int_{\Gamma_1} f + \int_{\Gamma_2} f = \int_{\Gamma_1+\Gamma_2} f$$

$$\int_{-\Gamma} f = -\int_\Gamma f$$ □

Using Theorem 7.8 is it easy to prove:

PROPOSITION 16.8. *The winding number is additive:*

$$\begin{aligned} w(0, z_0) &= 0 \\ w(\Gamma_1 + \Gamma_2, z_0) &= w(\Gamma_1, z_0) + w(\Gamma_2, z_0) \\ w(-\Gamma, z_0) &= -w(\Gamma, z_0) \end{aligned}$$

□

We can now generalise Section 7.5:

THEOREM 16.9. *Let* $\Gamma = \sum m_j\gamma_j$ *be a chain, and let* z_0 *be any point in* $\mathbb{C}$ *that does not lie on the image of* Γ*. Then*

$$w(\Gamma, z_0) = \frac{1}{2\pi \mathrm{i}} \int_\Gamma \frac{\mathrm{d}z}{z - z_0}$$

Proof.

$$\begin{aligned} w(\Gamma, z_0) &= \sum_j m_j w(\gamma_j, z_0) \\ &= \sum_j \frac{1}{2\pi \mathrm{i}} m_j \int_{\gamma_j} \frac{\mathrm{d}z}{z - z_0} \\ &= \frac{1}{2\pi \mathrm{i}} \sum_j m_j \int_{\gamma_j} \frac{\mathrm{d}z}{z - z_0} \\ &= \frac{1}{2\pi \mathrm{i}} \int_\Gamma \frac{\mathrm{d}z}{z - z_0} \end{aligned}$$

□

We sometimes need to make an explicit distinction between a path $\gamma : [a, b] \to D \subseteq \mathbb{C}$ and its image. To do so we write the image as

$$\hat{\gamma} = \{\gamma(t) : t \in [a, b]\}$$

We include the hat only when it might be confusing to omit it.

16.2 Cycles

Next we need a special type of chain, which we call a *cycle*. A cycle is a 'closed chain'. That is, the group of cycles is generated by chains whose images decompose into a series of closed loops. That is, if the usual sum $\gamma_1 + \cdots + \gamma_n$ is a loop, then the corresponding formal sum in $\mathcal{C}$ is a cycle. Then all formal sums of cycles are again cycles. By definition the cycles form a subgroup of $\mathcal{C}$, which we denote by $\mathcal{Z}$.

Next we define a subgroup $\mathcal{B}$ of $\mathcal{Z}$, whose elements we call 'boundaries', which consists of cycles that can be ignored when calculating integrals (and therefore winding numbers). That is, cycles Ω such that

$$\int_\Omega f = 0 \quad \text{for all differentiable } f : D \to \mathbb{C}$$

(As a convention, we use Γ, γ when discussing chains and Ω, ω when discussing cycles.) If we let

$$f(z) = \frac{1}{2\pi \mathrm{i}} \frac{1}{z - z_0} \quad (z_0 \notin D)$$

then Theorem 16.9 implies that all boundary cycles automatically also satisfy

$$w(\Omega, z_0) = 0 \quad (z_0 \notin D)$$

Which cycles should we use? By Theorem 9.4 the group $\mathcal{B}$ should include all boundaries ∂R of rectangles R, as in Section 9.3. Several special loops of this type are particularly useful.

16.2.1 Sums and Formal Sums of Paths

We have already defined $\gamma + \delta$ and $-\gamma$ for paths: respectively, adjoin the two paths and combine them into a single one, and reverse the path. These operations are *not* those of $\mathcal{C}$. However, we want them to be the same 'up to homology', see Section 16.4. That is, they should be equal modulo $\mathcal{B}$. We can then think of these geometric operations as being the corresponding operations in $\mathcal{C}$ (or the subgroup $\mathcal{Z}$, which is more important here) whenever homologous chains can be considered equivalent.

Recall from Section 6.8 that if $\gamma : [a, b] \to D$ and $\delta : [b, c] \to D$ (not necessarily closed paths) we defined $-\gamma : [a, b] \to D$ and $\gamma + \delta : [a, c] \to D$ by

$$-\gamma(t) = \gamma(b + a - t)$$

and

$$(\gamma + \delta)(t) = \begin{cases} \gamma(t) & \text{if} \quad t \in [a, b] \\ \delta(t) & \text{if} \quad t \in [b, c] \end{cases}$$

(The original definition was slightly more general, allowing the interval for δ to be $[c, d]$ and then shifting it to make $c = b$. That option is not needed here.) See Figure 16.2.

Although we are using the same symbols $+$ and $-$, these operations are not (quite!) those of the group $\mathcal{C}$, and we need to sort out the precise relationship. To avoid confusion, we temporarily write the operations in $\mathcal{C}$ as $\oplus$ and $\ominus$, and use $+$ and $-$ for the operations on paths that we defined in Chapter 6.

First: they really are different. The path-sum $\gamma + \delta$, considered as an element of $\mathcal{C}$, corresponds to the function $m : \mathcal{P} \to \mathbb{Z}$ such that

$$m(\gamma + \delta) = 1$$

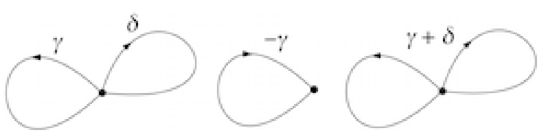

Figure 16.2 Previous definitions of $-\gamma$ and $\gamma + \delta$. Here both paths are loops, so they can also be thought of as cycles.

$$m(\sigma) = 0 \quad (\sigma \neq \gamma + \delta)$$

In contrast, $\gamma \oplus \delta \in \mathcal{C}$ corresponds to the function $m' : \mathcal{P} \to \mathbb{Z}$ such that

$$m'(\gamma) = 1$$
$$m'(\delta) = 1$$
$$m'(\sigma) = 0 \quad (\sigma \neq \gamma, \delta)$$

These are different functions because, as elements of $\mathcal{P}$, we have $\gamma \oplus \delta \neq \gamma$, $\gamma \oplus \delta \neq \delta$, so

$$m'(\gamma \oplus \delta) = 0 \qquad m(\gamma + \delta) = 1$$

Similarly, $\ominus\gamma$ (cycle) is not the same as $-\gamma$ (path).

16.3 Boundaries

The role of homology is, in effect, to *force* these technically different forms of addition and reversal to be the same. It exploits a standard trick encountered throughout mathematics: define an equivalence relation for which objects that differ in inessential ways are equivalent. Then their equivalence classes are *the same*. For instance, in arithmetic modulo n, a difference between integers that is divisible by n is inessential. We would like to set $n = 0$, but that is technically false. So we define r, s to be congruent modulo n if $r - s$ is divisible by n. Congruence is an equivalence relation, and the set of equivalence classes is denoted $\mathbb{Z}_n$. Now, although r and s are different in $\mathbb{Z}$, they represent the same congruence class.

In this case $\mathbb{Z}_n$ is the quotient group of the additive group of integers $\mathbb{Z}$ by the subgroup of $n\mathbb{Z}$ of multiples of n; that is, $\mathbb{Z}_n = \mathbb{Z}/(n\mathbb{Z})$. (The multiplicative structure can also be included, because $n\mathbb{Z}$ is an ideal of the ring $\mathbb{Z}$.) We can play the same game with cycles in place of $\mathbb{Z}$, and whatever differences we wish to ignore playing the role of $n\mathbb{Z}$. In this case, the differences that we want to forget about are boundaries, in the following sense:

DEFINITION 16.10. The group $\mathcal{B}$ of *boundaries* in D is the subgroup of $\mathcal{Z}$ generated by all boundaries ∂R of rectangles in D. □

To make the argument clear, we continue to use $\oplus, \ominus$ in place of $+, -$ in $\mathcal{C}$ to distinguish the group operations from the standard operations on paths. To get round one further technicality we go the whole hog and throw in all reparametrisations of γ that preserve orientation. Consider an orientation-preserving homeomorphism $\rho : [a,b] \to [a,b]$. That is, ρ is continuous and has a continuous inverse ρ^{-1}, and $\rho(a) = a, \rho(b) = b$. Equivalently, ρ is monotonic strictly increasing on $[a, b]$ and satisfies $\rho(a) = a, \rho(b) = b$. Now $\gamma\rho(t) = \gamma(\rho(t))$ defines a loop $\gamma\rho$. By Proposition 9.15, this change has no effect on integrals, hence on winding numbers.

If we allow ρ to reverse orientation, the integral changes sign, which is why we require orientation to be preserved. An alternative is to allow orientation to be reversed as well. Then $-\gamma$ is just a special reparametrisation, and any orientation-reversing

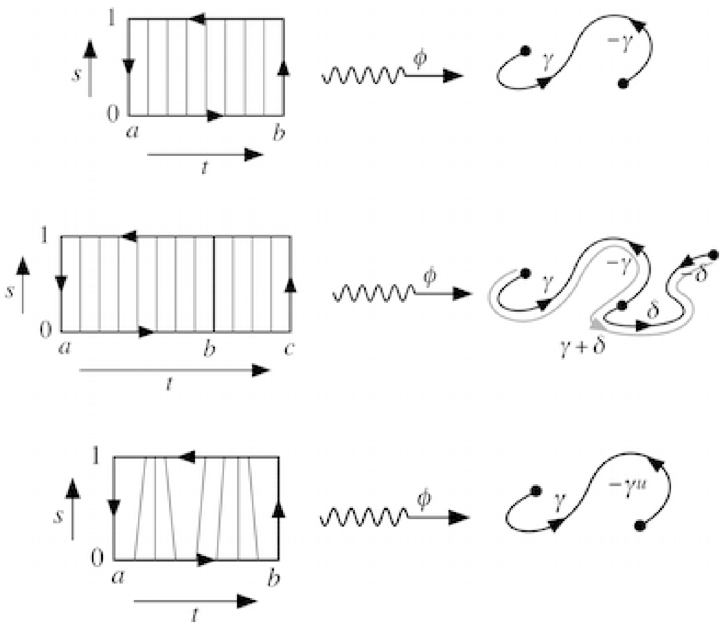

Figure 16.3 Maps of rectangles used in the proof of Proposition 16.11. *Top*: Reversal. *Middle*: Sum. *Bottom*: Reparametrisation.

reparametrisation is the map $\gamma \mapsto -\gamma$ composed with an orientation-preserving reparametrisation.

PROPOSITION 16.11. *Let γ, δ be paths in D and let $\gamma\rho$ be an orientation-preserving reparametrisation of γ. Then*

$$\gamma \oplus (-\gamma) \in \mathcal{B} \tag{16.2}$$

$$(\gamma \oplus \delta) - (\gamma + \delta) \in \mathcal{B} \tag{16.3}$$

$$\gamma \ominus \gamma\rho \in \mathcal{B} \tag{16.4}$$

Proof. In each case we specify a map $\phi : R \to D$ with a suitable boundary. Figure 16.3 illustrates the underlying geometry. Assume that $\gamma : [a,b] \to D$ and $\delta : [b,c] \to D$; also that $\rho : [a,b] \to [a,b]$ is the reparametrisation.

For (16.2), let $(t,s) \in [a,b] \times [0,1] = R$ and set

$$\phi(t,s) = \gamma(t)$$

(so ϕ is independent of s). Then $\phi(R) = \gamma([a,b]) \subseteq D$, and $\partial\phi = \gamma + (-\gamma)$.

For (16.3), let $(t,s) \in [a,c] \times [0,1] = R$ and set

$$\phi(t,s) = \begin{cases} \gamma(t) & \text{if} \quad t \in [a,b] \\ \delta(t) & \text{if} \quad t \in [b,c] \end{cases}$$

Now $\phi(R) = \gamma([a,b]) \cup \delta([c,d]) \subseteq D$, and $\partial\phi$ can be decomposed into $(\gamma \oplus \delta) \ominus (\gamma + \delta)$.

For (16.4), let $(t,s) \in [a,b] \times [0,1] = R$ and set

$$\phi(t,s) = \gamma((1-s)t + s\rho(t))$$

Then $\phi(R) = \gamma([a, b]) \subseteq D$, and $\partial\phi = \gamma - \gamma\rho$. □

PROPOSITION 16.12. *If* $\Omega \in \mathcal{B}$ *then*

$$\int_\Omega f = 0 \quad \text{for all differentiable } f : D \to \mathbb{C}$$

and

$$w(\Omega, z_0) = 0 \quad (z_0 \notin D)$$

Proof. It is enough to check that these conditions hold for boundaries of rectangles in D, because both the integral and winding number are additive by Propositions 16.7 and 16.8. □

16.4 Homology

The rather pedantic distinction between $\oplus$ and $+$, and the other items discussed above, does not cause any great problems, because – as we now prove – it disappears when we pass to homology classes. But first, we must define homology:

DEFINITION 16.13. Two cycles Ω_1, Ω_2 in D are *homologous*, written $\Omega_1 \sim \Omega_2$, if

$$\Omega_1 \ominus \Omega_2 \in \mathcal{B}$$

The *homology class* of Ω is its $\sim$-equivalence class:

$$[\Omega] = \{\Omega' : \Omega' \sim \Omega\}$$ □

It is easy to see that $[\Omega]$ is the coset of $\mathcal{B}$ in $\mathcal{Z}$ that contains Ω. We therefore define:

DEFINITION 16.14. The *first homology group* of D is the quotient group

$$H_1(D) = \mathcal{H} = \mathcal{Z}(D)/\mathcal{B}(D)$$

It is abelian since $\mathcal{Z}(D)$ is abelian. □

We denote the identity element of $\mathcal{H}$ by 0.

LEMMA 16.15. *For all loops* γ, δ *and orientation-preserving reparametrisations* ρ,

$$\gamma \oplus \delta \sim \gamma + \delta$$
$$\ominus\gamma \sim -\gamma$$
$$\gamma\rho \sim \gamma$$

Proof. In each case the left- and right-hand sides differ by an element of $\mathcal{B}$, by Proposition 16.11. Therefore they define the same element of $\mathcal{Z}/\mathcal{B} = \mathcal{H}$. □

This lemma shows that the distinction between $\oplus$ and $+$, and the other similar distinctions, disappear when we work modulo $\mathcal{B}$. *We therefore use the ordinary symbols* $+, -$ *for cycles as well as loops, and stop using* $\oplus, \ominus$. They have served their purpose.

LEMMA 16.16. *If $\gamma \sim \gamma'$ and $\delta \sim \delta'$ then $\gamma + \delta \sim \gamma' + \delta'$.*

Proof. Both $\gamma - \gamma'$ and $\delta - \delta'$ lie in $\mathcal{B}$. Now add. □

REMARK 16.17. Homology 'really' works with the image $\hat{\gamma}$ of a path rather than the path $\gamma : [a,b] \to D$. More precisely, what matters is the image *and its orientation*. Orientation is relevant because γ and $-\gamma$ have the same image, but the integral along one is minus that along the other. We want $\gamma - \gamma \in \mathcal{B}$ but not $\gamma + \gamma \in \mathcal{B}$ to make integrals vanish on cycles in $\mathcal{B}$. □

Example 16.18. Let $D = \mathbb{C} \setminus \{0\}$. Define $\gamma : [0,1] \to D$ by

$$\gamma(t) = \mathrm{e}^{2\pi \mathrm{i} t}$$

and $\delta : [0,1] \to D$ by

$$\delta(t) = 2\mathrm{e}^{2\pi \mathrm{i} t}$$

as in Figure 16.4 (left).

We claim that $\gamma \sim \delta$.

To prove this, introduce the path $\sigma = [1,2]$ and its reverse $-\sigma = [2,1]$, each parametrised by $t \in [0,1]$.

Now $\gamma - \delta$ is homologous to $\gamma + \sigma - \delta - \sigma$ because modulo $\mathcal{B}$ we have

$$\gamma + \sigma - \delta - \sigma\gamma - \delta + \sigma - \sigma = \gamma + \delta + 0 = \gamma + \delta$$

Then we claim that $\gamma + \sigma - \delta - \sigma = 0$ modulo $\mathcal{B}$. This follows because $\gamma + \sigma - \delta - \sigma$ is the boundary of a rectangle. Specifically, let $R = [0,1] \times [0,1]$ and define a map

$$\phi : [0,1] \times [0,1] \to D$$

by

$$\phi(r,s) = (1+s)\mathrm{e}^{2\pi \mathrm{i} r}$$

Then it is straightforward to check that $\partial\phi = \gamma + \sigma - \delta - \sigma$, see Figure 16.4 (right).

But now, modulo $\mathcal{B}$, we have $\gamma - \delta = 0$, so $\gamma \sim \delta$. □

REMARK 16.18. It is possible to define the relation of homology using winding numbers. (Obtaining the homology *group* is less straightforward.) In this approach a boundary is defined to be any cycle Ω such that $w(\Omega, z_0) = 0$ when $z_0 \notin D$. The proof of Cauchy's Theorem still requires most of the same ideas, and we feel that this definition of homology loses contact with the original geometric ideas that inspired it. Eventually we will show that our definition is equivalent to the one based on winding numbers, but we have to prove Theorem 16.2 to do this.

The following result is crucial:

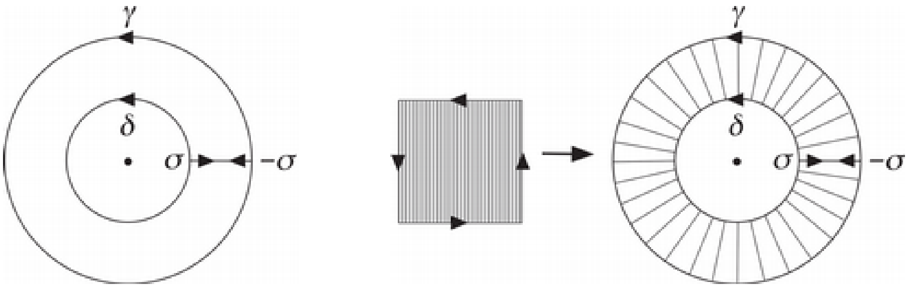

Figure 16.4 *Left*: The paths $\gamma, \delta, \sigma, -\sigma$. *Right*: Mapping a rectangle so that its boundary is $\gamma + \sigma - \delta - \sigma$.

THEOREM 16.19. *Let $f : D \to \mathbb{C}$ be differentiable on a domain D. Let Ω_1, Ω_2 be cycles in D. If $\Omega_1 \sim \Omega_2$ then*

$$\int_{\Omega_1} f = \int_{\Omega_2} f$$

Proof. Since $\Omega_1 \sim \Omega_2$ is equivalent to $\Omega_1 - \Omega_2 \in \mathcal{B}$, and the integral is additive, the statement is equivalent to

$$\text{If } \Omega \in \mathcal{B} \text{ then } \int_{\Omega} f = 0$$

Again using additivity, this follows if the integral over every generator of $\mathcal{B}$ is zero. But $\mathcal{B}$ is generated by boundaries of rectangles whose images lie in D, and the integral over the boundary of a rectangle is zero by Cauchy's Theorem for a Boundary, Theorem 9.6. □

The winding number is a homology invariant:

COROLLARY 16.20. *If $\Omega_1 \sim \Omega_2$ then*

$$w(\Omega_1, z_0) = w(\Omega_2, z_0) \quad \text{for all } z_0 \notin D$$ □

16.5 Proof of Cauchy's Theorem, Homology Version

We now give the proof of Theorem 16.2, the homology version of Cauchy's Theorem. Cauchy's Theorem for a Boundary, Theorem 9.6, gives the implication (i) ⇒ (ii). That (ii) ⇒ (iii) follows from Theorem 16.9. So it remains only to prove that (iii) ⇒ (i). Recall that condition (i) states that Ω is homologous to zero in D, or equivalently, $\Omega \in \mathcal{B}$. Condition (iii) states:

$$w(\Omega, z_0) = 0 \quad \text{for all } z_0 \notin D$$

which we call the *non-winding condition*.

Most of the proof that (iii) ⇒ (i) is an exercise in making careful distinctions between paths and their images, and using the relatively simple geometry of step paths to control the topology. This does involve some rather pedantic distinctions, needed only for this proof.

Let

$$\Omega = \sum n_j \omega_j \tag{16.5}$$

for loops ω_j in D, and $n_j \in \mathbb{Z}$. We may assume that the ω_j are distinct and that all $n_j \neq 0$. (If not, $\Omega = 0 \in \mathcal{B}$ anyway.)

We assume the non-winding condition and deduce that Ω is homologous to zero.

The idea of the proof is to simplify Ω by repeatedly changing it to some cycle Ω' that is homologous to Ω, and to use induction on some suitable measure of the complexity of the cycle. The definition of Ω' is chosen so that the complexity decreases at each stage. A simple result along these lines is Corollary 16.20, which implies that if $\Omega \sim \Omega'$ and $z_0 \notin D$, then

$$w(\Omega, z_0) = w(\Omega', z_0)$$

Thus, if Ω satisfies (iii) then so does Ω'. Trivially, if Ω' satisfies (i) then so does Ω. Therefore, if (iii) implies (i) for Ω', the implication also holds for Ω. In other words:

When proving (iii) implies (i) for Ω we can replace it by any homologous Ω'.

16.5.1 Grid of Rectangles

To construct suitable Ω' we proceed as follows.

As in the proof of Lemma 8.6, we form a grid of rectangles. Extend each edge of $\hat{\Omega}$ (horizontal or vertical) to a line in $\mathbb{C}$. These lines divide $\mathbb{C}$ into finitely many *rectangles*, and the lines cross at *grid points*, see Figure 16.5. Some of the 'rectangles' are infinite; these play no role as we shortly see.

Let P be the set of all grid points that lie on $\hat{\Omega}$, along with any corners or 'ends' where an edge of $\hat{\Omega}$ terminates without continuing. (The corresponding path 'doubles back' at such a point. In the generality employed here, this can happen.)

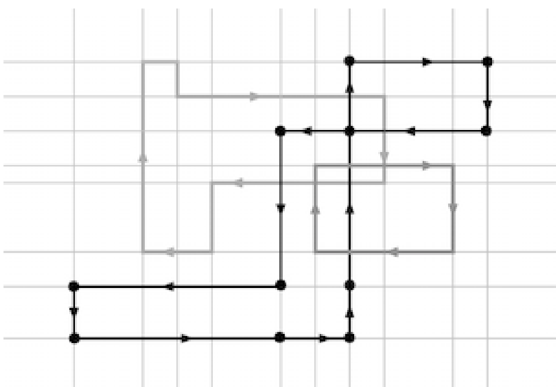

Figure 16.5 Rectangles and grid points. Here Ω is a sum of three loops.

We use P to split the ω_j into sums of paths, such that the image of each path is an edge of some rectangle. Let the parameter values of ω_j be $t \in [a, b]$. Since ω_j is a loop we can change parameters if necessary so that $\omega_j(a) \in P$. (Start at a corner or some other grid point, and note that a change of parameter is a homology.) Then the set of t for which $\omega_j(t) \in P$ (where the path goes through a point in P) is a finite set, and can be arranged in order as

$$a = t_0 < t_1 < \cdots < t_{n-1} < t_n = b$$

Let ω_j^k be the restriction of ω_j to $[t_{k-1}, t_k + 1]$ for $k = 1, \ldots, n$. Then

$$\omega_j = \omega_j^1 + \cdots + \omega_j^n$$

We call each ω_j^k a *segment* of Ω. Its image $\hat{\omega}_j^k$ is an edge of some rectangle, and by definition $\hat{\omega}_j^k \subseteq \hat{\Omega}$. Moreover, ω_j^k has an *orientation* in the direction of increasing t. Thus we can think of ω_j^k as an oriented edge of a rectangle.

Now Ω is (homologous to) a sum of segments, over all ω_j^k, and $\hat{\Omega}$ is the union of their images.

Example 16.22. Consider Figure 16.6. Here Ω is a single loop, but this overlaps itself several times. Its path is indicated by the grey polygon, and the loop visits P in the order

ABEDGHIFCBEHGDEBA

Therefore

$$\Omega = [\mathrm{A}, \mathrm{B}] + [\mathrm{B}, \mathrm{E}] + [\mathrm{E}, \mathrm{D}] + \cdots + [\mathrm{B}, \mathrm{A}]$$

Several segments occur more than once, sometimes in reverse orientation. For example AB and BA occur, and BE occurs twice, along with its reverse EB. In such cases, the image of Ω overlaps itself. □

We have now changed Ω by a series of homologies into a sum of integer multiples of segments. (To avoid proliferation of dashes, we continue to call the changed cycle Ω.)

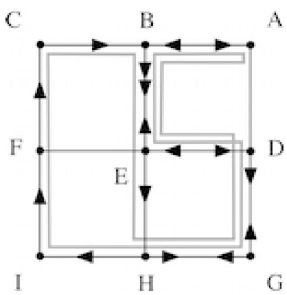

Figure 16.6 A loop with several overlaps.

Collect together all segments with the same image. Note that if σ is a segment then

$$\sigma + (-\sigma) \sim \sigma - \sigma \sim 0$$

since it is a boundary; indeed a generator of type (ii). So we can cancel these two segments without changing the homology class of Ω. Since $\sigma + (-\sigma)$ is a loop, the rest of Ω remains a cycle. That is:

Without loss of generality, Ω contains no pair of opposite segments.

Therefore we can write

$$\Omega = \sum_j m_j \sigma_j$$

where the σ_j are distinct segments, σ_j and $-\sigma_j$ do not both occur, and the m_j are non-zero integers. (There is one exception: all m_j may be zero. But then $\Omega \in \mathcal{B}$ and we are finished.)

For induction purposes we require:

DEFINITION 16.23. The *complexity* of Ω is

$$c(\Omega) = \text{number of segments + number of relevant rectangles}$$

Here, as in the proof of Lemma 8.6, a rectangle R_r is *relevant* if $w(\Omega, z_r) = 0$ for z_r at the centre of R_r. All infinite rectangles are irrelevant, and some finite ones may also be irrelevant. □

Observe that cancelling opposite segments decreases the complexity, so this 'without loss of generality' step in the proof does not affect induction based on the complexity.

16.5.2 Proof of Theorem 16.2.

We prove that (iii) implies (i) by induction on the complexity $c(\Omega)$.

The induction starts with $c(\Omega) = 0$. Then Ω is the empty cycle. This corresponds to the zero element $0 \in \mathcal{C}$, and since $\mathcal{B}$ is a subgroup of $\mathcal{C}$, we have $0 \in \mathcal{B}$. So the empty cycle is homologous to zero. This starts the induction.

For the induction step, take Ω with $c(\Omega) > 0$. Define A to be the *top right* point of P. That is,

(i) $\operatorname{im} \mathrm{A} \geq \operatorname{im} \mathrm{B}$ for all $\mathrm{B} \in P$
(ii) $\operatorname{re} \mathrm{A} \geq \operatorname{re} \mathrm{B}$ for all $\mathrm{B} \in P$ such that $\operatorname{im} \mathrm{B} = \operatorname{im} \mathrm{A}$

Since P is finite its top right point exists and is unique.

We consider the geometry of Ω 'near A'. There are three possibilities:

(i) A lies on a vertical edge of $\hat{\Omega}$ but not on any horizontal edge.
(ii) A lies on a horizontal edge of $\hat{\Omega}$ but not on any vertical edge.
(iii) A lies on a horizontal edge of $\hat{\Omega}$ and on a vertical edge.

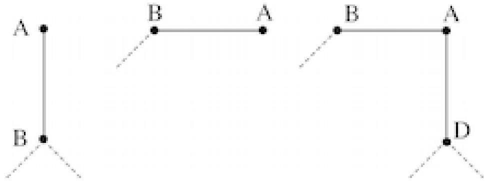

Figure 16.7 Three possible cases for the geometry near A. Dotted lines indicate possible connections to the rest of the cycle.

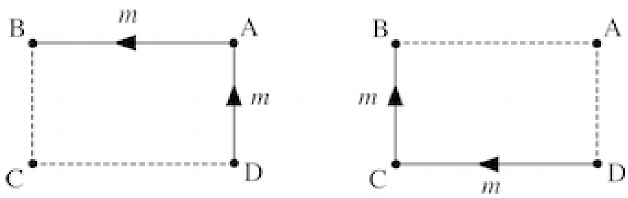

Figure 16.8 Rerouting segments to get rid of R.

Figure 16.7 shows these three possibilities, where A, B, D $\in P$.

In cases (i) and (ii), there must be a segment BA in that orientation, and another with the reverse orientation AB, because Ω is a cycle. These two segments cancel, contrary to our previous assumption, so case (i) cannot occur. Similarly, case (ii) cannot occur. So only (iii) is possible.

Orient AB from A to B and DA from D to A. Then all segments in Ω with image AB occur m times, where $0 \neq m \in Z$. If $m > 0$ they have orientation AB; if $m < 0$ they have orientation BA. Since Ω is a cycle, the segments of Ω with image DA must also occur m times, where $0 \neq m \in Z$. If $m > 0$ they have orientation DA, of $m < 0$ they have orientation AD. (Any loop that arrives at A must leave A.)

Let C be the fourth corner of the rectangle $R =$ ABCD. Then A, B, C, D are grid points and A, B, D $\in P$. The geometry on R is shown in Figure 16.8 (left), with m indicating $|m|$ copies of a segment in the direction of the arrow for $m > 0$, and in the opposite direction for $m < 0$.

16.5.3 Rerouting Segments

We now claim that by using a homology we can re-route the segments of Ω that lie in DAB to corresponding segments in DCB, as in Figure 16.8 (right). To set this up, we pause from the proof of Theorem 16.2 to prove:

LEMMA 16.24. *With the above notation, if z lies in the interior of R then*

$$w(\Omega, z) = m \tag{16.6}$$

Proof. Draw a vertical ray λ from z upwards, to infinity. This cuts AB at one (interior) point X, as in Figure 16.9 .

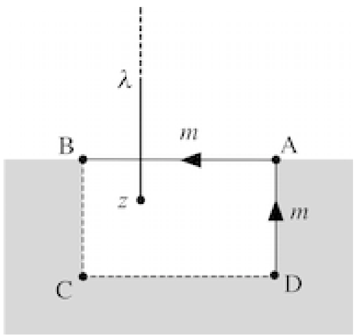

Figure 16.9 Rectangle R and ray λ. Shaded area indicates half-plane containing Ω_2.

Let Ω_1 be Ω with all segments in DA and AB removed (this is not a cycle). Let Ω_2 be Ω_1 plus the rerouted segments in DC and CB (all of multiplicity m). Then Ω_2 is a cycle. We claim that

$$\Omega_2 \sim \Omega$$

Segments of Ω lying in AB cut λ m times (with AB oriented to the left). The cycle Ω_2 does not intersect λ because A is the top right point of Ω, so Ω_2 lies in the half-plane below AB.

By continuous choice of argument, as in Section 7.4,

$$w(\Omega_2, z) = 0 \tag{16.7}$$

Let

$$\delta = m\,\partial R \in \mathcal{Z}$$

Clearly

$$\Omega = \Omega_2 + \delta$$

because segments of the right-hand side on DC or CB cancel, and the other segments of δ restore the segments on DA and AB.

We now face a minor technical problem: we do not (yet) know that DC and CB lie in the domain D. We therefore need to prove this, as follows.

If z is in the interior of R, then

$$w(\delta, z) = m$$

by continuous choice of argument. AB crosses λ only at X, where it crosses m times in the anticlockwise direction.

Therefore if z is in the interior of R,

$$w(\Omega, z) = w(\Omega_2 + \delta, z) = w(\Omega_2, z) + w(\delta, z) = 0 + m = m \neq 0$$

using (16.7) and additivity of w. □

16.5.4 Resumption of Proof of Theorem 16.2.

By (iii), which we are assuming, we have $z \in D$. Therefore the interior of R lies in D.

If BC $\in \hat{\Omega}$ then BC $\subseteq D$. Otherwise, continuity of the winding number at points not in $\hat{\Omega}$ implies that $w(\Omega, z) = 0$ if z lies on the interior of BC. The same goes for CD.

Finally, suppose that z = C. If C lies in $\hat{\Omega}$ then $z \in D$. If not, we again appeal to continuity of w to deduce that $w(\Omega, z) = m \neq 0$. As above, $z \in D$. So the whole of R, interior plus boundary, lies in D. But now

$$\delta = m\,\partial R \sim 0$$

since ∂R is a boundary, so lies in $\mathcal{B}$. We finally conclude that

$$\Omega_2 \sim \Omega - \delta \sim \Omega$$

The complexity $c(\Omega_2) < c(\Omega)$. The number of segments does not change, but we have removed the rectangle R. This is relevant since $w(\Omega, z_R) = m \neq 0$, where z_R is the centre of R. On the other hand, it is easy to see that no other rectangle can be irrelevant for Ω_2 but relevant for Ω. The reason is that if z_k is the centre of rectangle R_k, the continuous argument definition of w implies that

$$w(R_k, z_l) = \begin{cases} 1 & \text{if} \quad k = l \\ 0 & \text{if} \quad k \neq l \end{cases}$$

Additivity then proves that for all rectangles R_k except R we have

$$w(\Omega_2, z_k) = (\Omega, z_k)$$

Thus we have removed one relevant rectangle but not introduced any new ones, and $c(\Omega_2) < c(\Omega)$ as claimed.

Inductively, $\Omega_2 \sim 0$. But $\Omega \sim \Omega_2$ so $\Omega = 0$. This completes the induction step, proving Theorem 16.2. □

16.6 Cauchy's Residue Theorem, Homology Version

We can now use Theorem 16.2 to prove a homology version of Cauchy's Residue Theorem, Theorem 12.3.

First:

LEMMA 16.25. *Let Ω_1, Ω_2 be cycles in D such that*

$$w(\Omega_1, z) = w(\Omega_2, z) \textit{ for all } z \notin D$$

Then

$$\Omega_1 \sim \Omega_2$$

Proof. For all $z \notin D$,

$$w(\Omega_1 - \Omega_2, z) = w(\Omega_1, z) - w(\Omega_2, z) = 0$$

By Theorem 16.2, $\Omega_1 - \Omega_2 \sim 0$. So $\Omega_1 \sim \Omega_2$. □

Suppose $z_1, \ldots, z_k$ are distinct points in $\mathbb{C}$, and define the *k-times punctured plane* to be

$$D = \mathbb{C} \setminus \{z_1, \ldots, z_k\}$$

It is easy to see that D is a domain. (Recall: this means open and connected.)

THEOREM 16.26. *If D is the k-times punctured plane, then*

$$H_1(D) \cong \mathbb{Z}^k = \mathbb{Z} \oplus \cdots \oplus \mathbb{Z} \quad (k \text{ times})$$

Proof. We do more: we find independent generators for $H_1(D)$; that is, elements such that any element is a unique integer combination of them.

Let Ω be a cycle in D. Then for $j = 1, \ldots, k$, no z_j lies on $\hat{\Omega}$.

Since $\hat{\Omega}$ is a closed set, its complement is open, so there exists $\varepsilon > 0$ such that for all $j = 1, \ldots, k$

$$N_\varepsilon(z_j) \cap \hat{\Omega} = \emptyset$$

Let R_j be a square (with horizontal and vertical sides) of side $\varepsilon/2$ centred at z_j. Then $R_j \subseteq N_\varepsilon(z_j)$ so $R_j \cap \hat{\Omega} = \emptyset$. (We use a square to make ∂R_j a step path.)

Define cycles

$$\rho_j = \partial R_j$$

oriented anticlockwise. Then clearly

$$w(\rho_j, z_l) = \begin{cases} 1 & \text{if} \quad j = l \\ 0 & \text{if} \quad j \neq l \end{cases} \tag{16.8}$$

using a suitable cut plane and a continuous choice of argument. Let

$$m_j = w(\Omega, z_j)$$

We claim that

$$\Omega \sim \sum m_j \rho_j$$

This follows because

$$w(\Omega, z_l) = m_l = w\left(\sum m_j \rho_j, z_l\right)$$

by additivity and (16.8). Now apply Lemma 16.25.

We also claim that the m_j are unique: no other $\mathbb{Z}$-linear combination of the ρ_j is homologous to Ω. Equivalently, we must show that if

$$\rho = \sum m_j \rho_j \sim 0 \tag{16.9}$$

then all $m_j = 0$.

Suppose (16.9) holds. Then $w(\rho, z) = 0$ for all $z \notin D$ by Theorem 16.2. That is, $w(\rho, z_l) = 0$ for all $l = 1, \ldots, k$. But by Proposition 16.8,

$$w(\rho, z_l) = w\left(\sum m_j \rho_j, z_l\right) = m_l$$

Thus $m_l = 0$ for all $l = 1, \ldots, k$.

Clearly every ρ is a cycle, so the map

$$\Omega = (m_1, \ldots, m_k)$$

induces an isomorphism

$$H_1(D) \to \mathbb{Z}^k$$ □

We say that the *homology class* of Ω is the corresponding $(m_1, \ldots, m_k) \in \mathbb{Z}^k$.

THEOREM 16.27 (Cauchy's Residue Theorem, Homology Version). *Let D be a k-times punctured plane and suppose that $f : D \to \mathbb{C}$ is differentiable. Let Ω by a cycle in D of homology class $(m_1, \ldots, m_k) \in \mathbb{Z}^k$. Then*

$$\int_\Omega f = 2\pi \mathrm{i} \sum_{j=1}^{k} m_j \operatorname{res}(f, z_j)$$

Proof. $\Omega \sim \sum_{j=1}^{k} m_j \rho_j$, so

$$\int_\Omega f = \sum_{j=1}^{k} \int_{\rho_j} f$$

Since f is differentiable, we can apply Theorem 12.3, Cauchy's Residue Theorem for a loop, to obtain

$$\int_{\rho_j} f = 2\pi \mathrm{i} \operatorname{res}(f, z_j)$$ □

We could go on to state a more general version of Cauchy's Residue Theorem for other domains, such as $\mathbb{C}$ with a discrete infinite set of points removed, or a disc with finitely many points or discs removed. We could also consider analogues for Riemann surfaces. But it seems sensible to stop here, having opened up the possibilities of homology and used it to prove two powerful theorems in complex analysis.

16.7 Exercises

1. Let $D = \{z \in \mathbb{C} : 1 < |z| < 2\}$. Prove that $H_1(D) \cong \mathbb{Z}$.
2. Let D be an open disc with k distinct points removed. What is $H_1(D)$? Prove your answer correct.
3. Let D be $\mathbb{C}$ with k disjoint closed discs removed. What is $H_1(D)$? Prove your answer correct.
4. Suppose that $K = \{z_k : k \in \mathbb{N}\}$ is an infinite set of distinct points, and that K is *discrete*; that is, for each k there exists $\varepsilon > 0$ such that $N_\varepsilon(z_k) \cap K = \{z_k\}$. Prove a version of Cauchy's Residue Theorem for the domain $D = \mathbb{C} \setminus K$.
5. Construct a closed step path in $\mathbb{C} \setminus \{0\}$ that is homologous to zero, whose complement has four distinct connected components.

6. The 0*th homology group* $H_0(D)$ can be defined in a manner analogous to that for $H_1(D)$, as follows. A 0-cycle is a finite formal sum of pairs of (not necessarily distinct) points in D. A 0-boundary is a finite formal sum of pairs of points in D that can be joined by a step path in D.

Show that 0-cycles form a group $\mathcal{Z}_0$ and 0-boundaries from a group $\mathcal{B}_0$. Define $H_0(D) = \mathcal{Z}_0/\mathcal{B}_0$.

Show that the pair (z_1, z_2) is a 0-boundary if and only z_1 and z_2 lie in the same connected component of D. Deduce that $H_0(D)$ is generated by all cosets of $\mathcal{B}$ of the form $(z_K, z_K) + \mathcal{B}$, where one point $z_K \in C$ is chosen for each connected component K. (That is, $H_0(D)$ is isomorphic to the free abelian group on the connected components.) In particular, if D has a finite number k of connected components, prove that $H_0(D) \cong \mathbb{Z}^k$.

7. Prove the *Dog Walking Theorem*: If σ_1, σ_2 are loops, both parametrised by $t \in [a, b]$, and there is a point z_0 such that

$$|\sigma_1(t) - \sigma_2(t)| < |\sigma_1(t) - z_0| \quad \text{for all } t \in [a, b]$$

then

$$w(\sigma_1, z_0) = w(\sigma_2, z_0)$$

and $\sigma_1 \sim \sigma_2$ in $\mathbb{C} \setminus \{z_0\}$.

Explain the name of this theorem. (Hint: σ_1 is the man, σ_2 the dog, and z_0 is a tree.)

Use it to give another proof of Rouché's Theorem 12.14.

8. *Jordan Contour Theorem for Loops*. Recall that a loop is a closed *step* path. Say that a loop $\sigma : [a, b] \to \mathbb{C}$ is *simple* if it does not cross itself; that is, $\sigma(t_1) = \sigma(t_2)$ with $t_1 \neq t_2 \in [a, b]$ only when the t_j equal a or b.

Use the trick in the proof of Theorem 16.2, of rerouting the segments of a cycle (here one loop) round the rectangle R adjacent to the top right point of $\hat{\sigma}$ to prove that:

(i) $\mathbb{C} \setminus \hat{\sigma}$ has exactly two connected components: one bounded, the other unbounded.

(ii) Let $I(\sigma)$ be the bounded component in (i) and $O(\sigma)$ be the unbounded component. Prove that $w(\sigma, z) = \pm 1$ if $z \in I(\sigma)$ and $w(\sigma, z) = 0$ if $z \in O(\sigma)$.

(iii) $\partial I(\sigma) = \partial O(\sigma) = \hat{\sigma}$.

9. If σ_1, σ_2 are simple loops and $\hat{\sigma}_1 \subseteq I(\sigma_2)$, prove that $I(\sigma_1) \subseteq I(\sigma_2)$ and $O(\sigma_1) \supseteq O(\sigma_2)$.

10. Suppose that a simple loop σ lies in a domain D. Prove that there exists $\varepsilon > 0$ such that $|z - \sigma(t)| < \varepsilon$ for $z \in \mathbb{C}$ and some $t \in [a, b]$ implies that $z \in D$. (Hint: use the Paving Lemma.)

11. Using the result of Exercises 9 and 10, prove that if σ is a simple loop in D there exists $\varepsilon > 0$ such that there are simple loops σ^I, σ^O such that:

(i) Every point of $\hat{\sigma}^O \cup \hat{\sigma}^I$ lies within distance ε of some point of $\hat{\sigma}$.
(ii) σ^O lies in $O(\sigma)$ and σ^I lies in $I(\sigma)$.
(iii) The regions $I(\sigma^O) \setminus (\hat{\sigma} \cup I(\sigma))$ and $I(\sigma) \setminus (\hat{\sigma}^I \cup I(\sigma^I))$ are connected.

12. If D is a domain (hence connected) and σ is a simple loop in D, prove that both $D \cap O(\sigma)$ and $D \cap I(\sigma)$ are connected.

13. Define the *edges* of a loop as follows. Define a corner to be a point at which the image of the loop switches between horizontal and vertical. Assume that $\sigma(a)$ is a corner, and let Q be the set of $t \in [a, b]$ such that $\sigma(t)$ is a corner. Arrange the elements of Q in order:

$$a = t_0 < t_1 < \cdots < t_n = b$$

Then an edge is the image of σ restricted to some interval $[t_{j-1}, t_j]$ for $j = 1, \dots, n$.

Define the edges of a cycle to be those of its component loops.

Say that a cycle σ is in *general position* if no two edges are collinear (that is, both lie in the same horizontal or vertical line in $\mathbb{C}$).

Prove that every cycle in a domain D is homologous to a cycle in general position.

(Hint: inductively displace successive edges parallel to themselves through a small distance, so that the displaced edge differs from the original edge by the boundary of a rectangle lying inside D. Fix up the geometry at corners, where as stated there may be a gap or a small overlap between the displaced edges.)

14. Show that any non-zero cycle is homologous to a sum of disjoint simple loops σ_j, each occurring in the sum with multiplicity ± 1. (Hint: use Exercise 13 to put it in

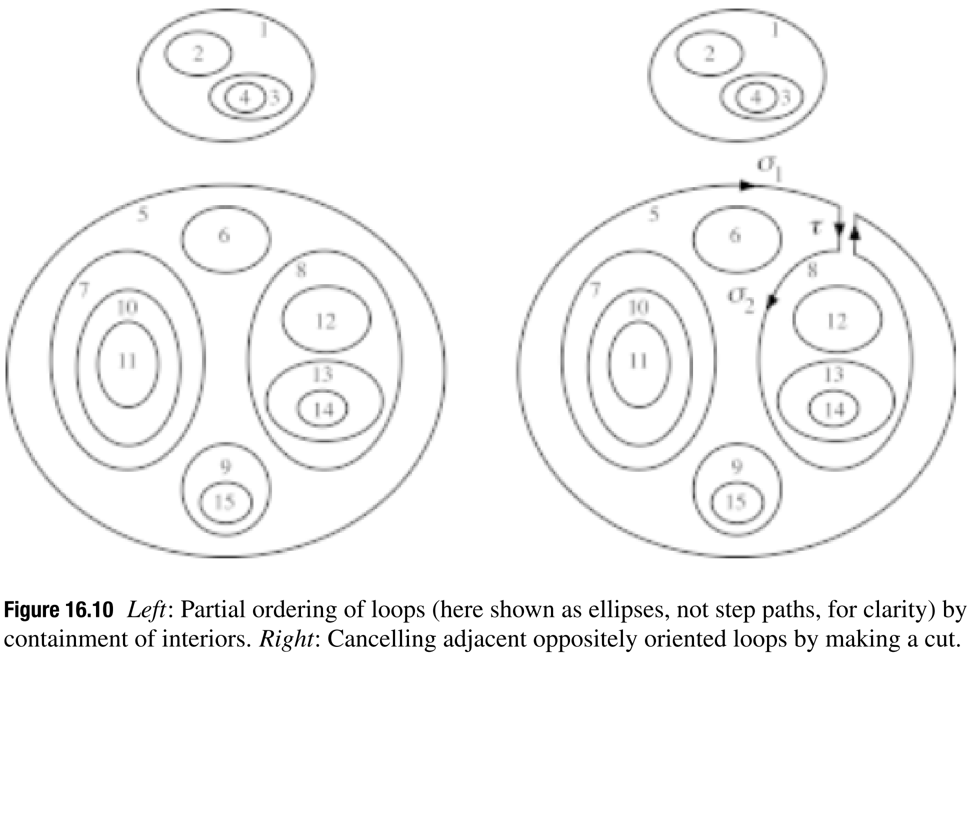

Figure 16.10 *Left*: Partial ordering of loops (here shown as ellipses, not step paths, for clarity) by containment of interiors. *Right*: Cancelling adjacent oppositely oriented loops by making a cut.

general position, then reroute edges near crossing points to eliminate the crossings, using small rectangles inside D.)

15. Using induction on the number of simple loops in such a decomposition, obtain a different proof of Theorem 16.2, which is modelled along the classical lines of simplifying loops by 'making cuts' that cancel each other out. See for instance Figure 11.2.

Warning: This exercise is hard, but the following sketch may help. Think of it as a mini-research project.

First, show that disjoint simple loops are partially ordered by the relation $\sigma_1 < \sigma_2 \iff I(\sigma_1) \subsetneq I(\sigma_2)$. This ordering can be complicated, see Figure 16.10.

Show that if (iii) holds then there must be two simple loops $\sigma_1 < \sigma_2$ with opposite orientations, such that there is no simple loop σ_3 with $\sigma_1 < \sigma_3 < \sigma_2$. Using some of the exercises above, show that there is a step path τ joining a point on σ_1 to a point on σ_2 that lies in D and intersects no other σ_j appearing in the decomposition. Use τ to write

$$\sigma_1 + \sigma_2 = \rho + \tau'$$

where ρ is a simple loop in D and τ' is a simple loop in D that is homologous to zero, see Figure 16.9. This reduces the number of simple loops by 1, and induction reduces to the case of one simple loop σ. Show that $I(\sigma)$ lies in D and complete the proof using property (iii).

17 The Road Goes Ever On . . .

The end of this book is fast upon us, but there is as yet no discernible end to complex analysis itself. It remains a vigorous and growing part of mainstream mathematics. We sample this rich area by giving brief descriptions of five areas of current research: the Riemann Hypothesis, modular functions, several complex variables, complex manifolds, and complex dynamics.

17.1 The Riemann Hypothesis

One of the more unexpected applications of complex analysis occurred in the second half of the nineteenth century, leading to major advances in number theory. At first sight these two areas have little connection: complex analysis is about continuous quantities and number theory is discrete. Even more surprisingly, the connection that emerged did so in the context of statistical properties of prime numbers. The starting point was a new kind of complex function, based not on power series, but on *Dirichlet series* of the form

$$\phi(s) = \sum_{n=1}^{\infty} \frac{a_n}{n^s}$$

where the a_n and s are complex numbers and the domain is restricted to make the series converge. For example, if the a_n are bounded, we require $\operatorname{re} s > 1$. More strongly, if the partial sums $a_1 + \cdots + a_n$ are bounded, we require only $\operatorname{re} s > 0$. Classical concepts such as the gamma function (and, later, more sophisticated new ideas) were then used to extend the definition of ϕ to the entire complex plane, except for isolated poles.

Prime numbers are of vital importance in mathematics, yet they notoriously lack any clear pattern. Given a list of all primes up to some particular number, there seems to be no simple way to predict the next one. However, hints of regularity emerge when we think about entire ranges of primes. For instance: how many primes are there up to some specified limit x? It is difficult to answer this question exactly, but good approximations are another matter. In 1797–8 Adrien-Marie Legendre counted how many primes occur up to various limits, using tables of primes provided by Jurij Vega and Anton Felkel. Denoting the number of primes less than x by $\pi(x)$, as we now do, he observed empirically that $\pi(x)$ seems to be close to $x/(\log x - 1.083\,66)$. In a letter of 1849, Gauss wrote that when he was about 15 he noticed that $\pi(x)$ is approximately $x/\log x$ for large x. In 1838 Dirichlet wrote to Gauss to tell him he had found a similar but more accurate approximation to $\pi(x)$, the *logarithmic integral*

$$\mathrm{Li}(x) = \int_0^x \frac{\mathrm{d}t}{\log t}$$

For example, when $x = 10^9$,

$$\pi(x) = 50\,847\,534$$
$$\mathrm{Li}(x) = 50\,849\,234.9$$
$$x/\log x = 48\,254\,942.4$$

The approximations here are asymptotic: that is, the ratio of the approximate formula to the true value $\pi(x)$ was conjectured to tend to 1 as $x \to \infty$. The error in an asymptotic formula can be quite big; it just has to be significantly smaller than the exact quantity.

All of these observations were purely empirical, and a rigorous proof that these formulas are asymptotic to $\pi(x)$ became a major open problem in number theory, called the Prime Number Theorem (at a time when 'theorem' was also used to refer to a conjecture). The eventual proof of the Prime Number Theorem relied on deep applications of complex analysis.

The relation between these apparently disconnected areas of mathematics goes back to Euler, who realised in 1737 that uniqueness of prime factorisation has an implication for real analysis, namely the formula

$$\sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_p \frac{1}{1 - p^{-s}}$$

where p runs through the primes. To prove this, observe that summing a geometric series gives

$$\frac{1}{1 - p^{-s}} = \frac{1}{1^s} + \frac{1}{p^s} + \frac{1}{p^{2s}} + \cdots$$

Expand the product on the right-hand side, and observe that each term $1/n^s$ appears exactly once as 1 divided by the sth power of a product of prime powers.

Euler considered only positive integer s, but in 1848 and 1850 the Russian mathematician Pafnuty Chebyshev attempted to prove the Prime Number Theorem by applying real analysis to Euler's series, assuming s to be real and greater than 1, so that the series converges. He managed to prove that for large enough x the ratio of $\pi(x)$ to $x/\log x$ lies between two constants: one slightly bigger than 1, the other slightly smaller.

Riemann realised that real analysis was too limited to complete the proof, but complex analysis was more powerful and might do the trick. He noticed that Euler's series converges for any complex s with $\mathrm{re}\, s > 1$. Its sum, called the *zeta function* $\zeta(s)$, is complex analytic. He wrote up his ideas in 1859. They included an explicit formula expressing $\pi(x)$ *exactly* in terms of the zeta function. To state it, Riemann defined a related function

$$\Pi(x) = \pi(x) + \frac{1}{2}\pi(x^{1/2}) + \frac{1}{3}\pi(x^{1/3}) + \frac{1}{4}\pi(x^{1/4}) + \cdots$$

which counts the prime powers up to x. From this it is easy to deduce $\pi(x)$. Then he proved that

$$\Pi(x) = \mathrm{Li}(x) - \sum_{\rho} \mathrm{Li}(x^{\rho}) + \int_x^{\infty} \frac{\mathrm{d}t}{t(t^2-1)\log t}$$

where the sum is over all zeros of the zeta function, excluding negative even integers. (We explain where those come from in a moment – their role is not apparent in the formula for the infinite series.) The zeros and poles of a complex analytic function are important because they characterise the function completely if we also know their orders. It turns out that the zeros of the zeta function include all negative even integers, along with infinitely many others.

In 1896 Jacques Hadamard and Charles Jean de la Vallée-Poussin independently proved the Prime Number Theorem using Riemann's zeta function, a triumph for what became known as analytic number theory – the application of complex analysis to properties of whole numbers. This approach has now blossomed into a huge area of number theory.

Riemann took the crucial step of using analytic continuation to extend the definition of $\zeta(s)$ so that s can be any complex number except 1. This step depends on the gamma function, introduced in Section 6.11, which is a generalisation of the factorial function $n!$ of a whole number $n \in \mathbb{N}$.

To extend the domain of definition of the zeta function, we first note that the series for $\zeta(s)$ converges only when $\mathrm{re}\, z > 1$. However, it is easy to verify that

$$\left(1 - \frac{2}{2^s}\right)\zeta(s) = \frac{1}{1^s} - \frac{1}{2^s} + \frac{1}{3^s} - \cdots$$

The series on the right converges for $\mathrm{re}\, z > 0$, so the formula lets us extend the definition of the zeta function to such z, and it remains an analytic function of z. Then, starting from a formula of Adolf Hurwitz, Riemann deduced the *functional equation for the zeta function*: if $0 < \mathrm{re}\, s < 1$, then

$$\zeta(s) = 2^s \pi^{s-1} \sin\left(\frac{s\pi}{2}\right)\Gamma(1-s)\zeta(1-s)$$

where Γ denotes the gamma function, Section 6.11. This formula can then be used to define $\zeta(s)$ for all complex $z \neq 1$, by assuming it holds whenever $0 \leq \mathrm{re}\, s$, giving an analytic continuation of the zeta function to $\mathbb{C} \setminus \{1\}$. The same formula shows that

$$\zeta(-2n) = 0 \quad n = 1, 2, 3, \ldots$$

because the factor $\sin\left(\frac{s\pi}{2}\right)$ vanishes at these numbers. The negative even integers are the *trivial zeros* of the zeta function.

In his paper, Riemann [16] observed that there also exist non-trivial zeros of the zeta function – those that are not negative even integers. In particular he found zeros at

$$\tfrac{1}{2} \pm 14.135\mathrm{i} \qquad \tfrac{1}{2} \pm 21.022\mathrm{i} \qquad \tfrac{1}{2} \pm 25.011\mathrm{i}$$

but did not publish these results. His calculations suggested that *all* non-trivial zeros seem to have real part $\frac{1}{2}$. This statement is now known as the *Riemann Hypothesis*. (His actual statement was about a closely related function and is equivalent to this.) He wrote:

One would, however, wish for a strict proof of this; I have, though, after some fleeting futile attempts, provisionally put aside the search for such, as it appears unnecessary for the next objective of my investigation.

As time passed, it became clear that the Riemann Hypothesis is of central importance in mathematics. It implies strong bounds on the error in the Prime Number Theorem – the order of magnitude of the difference between the approximate formula and $\pi(x)$. For example, in 1901 Niels Helge von Koch showed that if the Riemann Hypothesis is correct then the error $|\mathrm{Li}(x) - \pi(x)|$ grows no faster than a constant times $\sqrt{x}\log x$, and in 1976 Lowell Schoenfeld gave the explicit bound

$$|\mathrm{Li}(x) - \pi(x)| < \frac{1}{8\pi}\sqrt{x}\log x \qquad (x \geq 2657)$$

The Riemann Hypothesis also tells us about the size of gaps between consecutive primes. Better still, there are far-reaching generalisations with major implications throughout number theory, and it has applications to testing numbers to see whether they are prime, which is important for encrypted Internet communications. There is a lot of circumstantial evidence in favour of the Riemann Hypothesis, and computer calculations have verified it for the first ten trillion zeros.

Despite a huge amount of effort, however, the Riemann Hypothesis has neither been proved nor disproved. Hilbert included it in his famous list of unsolved problems in 1900. The Clay Mathematics Institute is currently offering a prize of one million dollars for solutions to six major unsolved problems in mathematics, and the Riemann Hypothesis is one of them. A seventh, the Poincaré Conjecture, was solved in 2002–3 by Grigori Perelman, who declined the prize.

17.2 Modular Functions

Extending the idea of periodic functions (such as exp) leads to *doubly periodic* functions, which satisfy $f(z) = f(z+p) = f(z+q)$ for two complex numbers p, q that are linearly independent over the reals (not real multiples of each other). These are also known as *elliptic functions* because some of them can be used to give a formula for the arc-length of an ellipse.

Further generalisation leads to modular functions, which transform in a nice way under a discrete group of Möbius maps. Here 'discrete' means that each transformation is separate from the others rather than being part of some continuous family. For example, $\mathbb{Z}^2$ is a discrete subgroup of $\mathbb{R}^2$.

These functions preoccupied many mathematicians at the end of the nineteenth century. These united in one package group theory, differential equations, algebraic function theory, topology, and complex analysis. This is still an important field of research. For example, modular functions are heavily involved in Andrew Wiles's celebrated proof of Fermat's Last Theorem.

The starting point is the *modular group*, which consists of all Möbius maps

$$\mu(z) = \frac{az+b}{cz+d}$$

for which $a, b, c, d \in \mathbb{Z}$ and $ad - bc \neq 0$. This group is also called the *special linear group* over $\mathbb{Z}$, denoted $\mathbf{SL}_2(\mathbb{Z})$. If $\mathbb{H}$ denotes the upper half-plane $\{z \in \mathbb{C} : \operatorname{im} z \geq 0\}$, then (subject to some technical conditions omitted here) a *modular function* is a function $f : \mathbb{H} \to \mathbb{C}$ that is analytic on $\mathbb{H}$ and satisfies the condition

$$f(\mu(z)) = (cz + d)^k f(z) \quad \forall z \in \mathbb{H}$$

for any $\mu \in \mathbf{SL}_2(\mathbb{Z})$, where the *weight* k is an odd positive integer.

Generalisations of these functions, known as *modular forms*, have many applications to number theory. Wiles deduced Fermat's Last Theorem from his proof of a long-conjectured relationship between modular forms and certain number-theoretic equations called elliptic curves because of the connection with elliptic functions. An ellipse is *not* an elliptic curve.

17.3 Several Complex Variables

Functions of *several* complex variables may be studied. They turn out to be far trickier than we might expect, with startling new phenomena.

A function of several complex variables is a map

$$f : D \to \mathbb{C}$$

where the domain D is an open subset of $\mathbb{C}^n$ for some $n \in \mathbb{N}, n \geq 2$. We require f to be analytic, in the sense that for each $a = (a_1, \ldots, a_n) \in U$ there is a convergent power series expansion

$$f(z_1, \ldots, z_n) = \sum b_{k_1,\ldots,k_n}(z_1 - a_1)^{k_1} \cdots (z_n - a_n)^{k_n}$$

There is an equivalent condition analogous to the Cauchy–Riemann Equations, which can be proved using a form of Cauchy's Integral Formula. A more common term in this context is that such an f is *holomorphic* on U.

Some of the basic theorems and concepts of complex analysis generalise to several variables, but many others do not. Analytic continuation has a natural definition, and is again unique but possibly multivalued. One of the most influential new discoveries is the Hartogs' phenomenon, named after Friedrich Hartogs. In one-variable complex analysis, there are analytic functions with a natural boundary – a closed curve beyond which they cannot be analytically continued. Equation (14.5) gives an example. Hartogs' extension theorem of 1906 states, roughly speaking, that no such thing occurs for two or more variables. The domain on which a function is holomorphic is always unbounded.

Hartogs proved more. In one-variable complex analysis, for any domain D there exists an analytic function $f : D \to \mathbb{C}$ for which the boundary ∂D is a natural boundary. Hartogs proved that if

$$\Delta^2 = \{(z_1, z_2) \in \mathbb{C}^2 : |z_1| < 1, |z_2| < 1\}$$
$$H_\varepsilon = \{(z_1, z_2) \in \Delta^2 : |z_1| < \varepsilon \text{ or } 1 - \varepsilon < |z_2|\}$$

for $0 < \varepsilon < 1$, then any function f holomorphic on H_ε has a holomorphic extension F to Δ^2. Indeed, F can be defined using Cauchy's Integral Formula.

The theory was developed by Hartogs and Kiyoshi Oka in the 1930s. Hartogs proved that every isolated singularity is removable when $n \geq 2$, another crucial difference from the one-variable case. The subject really began to take off in 1945 with work of Henri Cartan, Hans Grauert, and Reinhold Remmert. It soon evolved into the more general context of complex manifolds, which we consider next.

17.4 Complex Manifolds

The Riemann surface alone has opened up broad vistas, by capturing the essential structure of a complex function in a single geometric object. All kinds of information, such as the presence of branch points and singularities, can be read off from it – and the 'rigidity' of analytic functions means that they are essentially determined by the position and nature of their singularities. Many questions that seem baffling without Riemann surfaces become transparent when this extra geometry is invoked.

The notion of a *complex manifold* arises when we generalise Riemann surfaces to several complex variables. The basic definition is modelled on real manifolds: these are multidimensional analogues of surfaces, obtained by patching together open subsets of $\mathbb{R}^n$ whose coordinates are smoothly related when the patches overlap. In the complex case we use open subsets of $\mathbb{C}^n$ whose coordinates are analytically related when the patches overlap.

In differentiable topology a smooth real manifold may have several distinct differentiable structures, but in all dimensions except 4 the number of such structures is finite. However, a fixed manifold of even dimension can possess infinitely many different complex structures. These can be classified by associating each complex structure with a point in a so-called moduli space which is itself a complex algebraic variety – roughly, the set of zeros of a system of polynomials in several complex variables.

The whole area is deeply related to topology and to algebraic geometry, and we will not attempt to summarise it. Because all of these areas are highly abstract, until recently much of this work was viewed by most non-mathematicians (among those aware of its existence) as mere generalisation for its own sake – pretty, intellectually clever, but far too wild and abstract to have sensible applications outside pure mathematics. Such judgements are usually superficial and premature when they refer to the mathematical mainstream: most ideas that are important in pure mathematics eventually acquire significant uses. And so it has proved here. Complex manifolds and automorphic functions are now central to the physics of Quantum Field Theory and the study of gauge fields in particle physics, for instance.

17.5 Complex Dynamics

Let $f : Z \to Z$ be a map from a set Z to itself. Then we can apply f repeatedly to any $z_0 \in X$ to get the sequence

$$z_0,\ f(z_0),\ f(f(z_0)),\ f(f(f(z_0))),\ldots$$

This process is classically known as *iterating* f. It can be defined inductively by

$$f^0(z_0) = z_0 \qquad f^{n+1}(z_0) = z_{n+1} = f(z_n)$$

The modern term is *discrete dynamical system.* If we think of the subscript n as 'time', ticking in whole number steps, the sequence (z_n) specifies how an initial state x_0 changes over time. Usually the set Z is a smooth manifold – a multidimensional analogue of a smooth surface. A central aim of dynamical systems theory is to understand sequences of this kind, especially their long-term behaviour.

Complex dynamics studies discrete dynamical systems when $Z = \mathbb{C}$ (or some subset of $\mathbb{C}$) and f is differentiable. The first serious work on this topic was carried out by Pierre Fatou in 1917 and Gaston Julia in 1918. Their methods were analytic. With the advance of computer graphics it became possible to draw accurate pictures, which revealed surprisingly intricate shapes of great beauty. The associated mathematics is also very beautiful, from a logical point of view. However, many interesting and important problems remain open at this stage.

For this brief discussion, we restrict attention to polynomial f. Analogous concepts can be studied for general differentiable f, but some definitions are more complicated. See Devaney [3] for the details. A key concept is the *Julia set* of f. This can be characterised in several equivalent ways, by proving appropriate theorems. Perhaps the simplest is to distinguish two types of behaviour for the sequence (z_n). Depending on z_0, either the set of all z_n remains bounded for all $n \in \mathbb{N}$, or it diverges to infinity. The Julia set $J(f)$ is the boundary of the set of z_0 for which it diverges to infinity. For polynomial f this is also the boundary of the set of z_0 for which it remains bounded. The Julia set is *invariant* under f, that is, $f^{-1}(J(f)) = f(J(f)) = J(f)$.

For example, if $f(z) = z^2$ then $J(f)$ is the unit circle. On $J(f)$, the map f sends $e^{i\theta}$ to $e^{2i\theta}$, doubling the angle. If $f(z) = z^2 - 2$ it can be proved that $J(f)$ is the line segment $[-2, 2]$, and the dynamics of f on this interval is more complicated.

The geometry of Julia sets is fascinating, even for very simple maps f. It is best understood (and even then not completely) for quadratic maps

$$f(z) = f_c(z) = z^2 + c \qquad c \in \mathbb{C}$$

Sometimes $J(f_c)$ is quite simple. For instance, if $c < \frac{1}{4}$ then $J(f_c)$ is a simple closed curve. For other values of c it is a highly intricate fractal, as in Figure 17.1. For some c the Julia set is a disconnected dust cloud; for others it is connected.

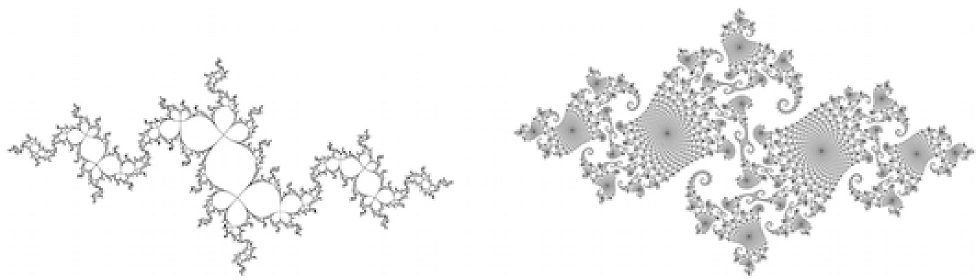

Figure 17.1 Two examples of Julia sets.

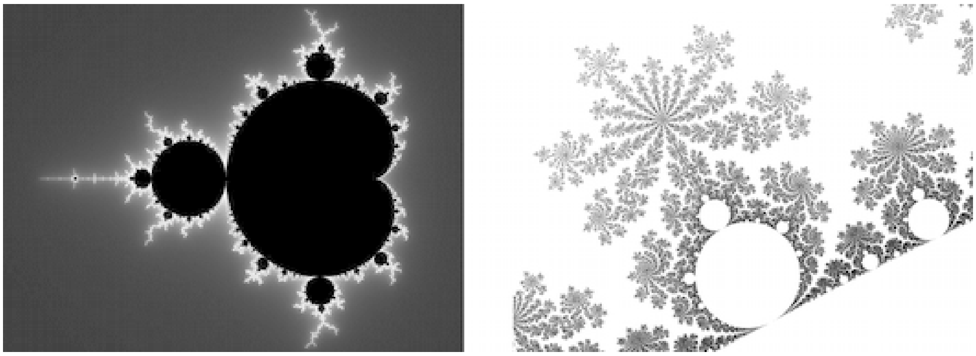

Figure 17.2 *Left*: The Mandelbrot set. [From Wikimedia Commons under the GNU Free Documentation License.] *Right*: Close-up of the boundary of part of the Mandelbrot set.

To bring order to the Julia sets, we consider how their topology depends on the constant $c \in \mathbb{C}$. The *Mandelbrot set* $M \subseteq \mathbb{C}$ is defined to be the set of c for which $J(f_c)$ is connected. Figure 17.2 (left) shows the Mandelbrot set: it is the black cactus-like region. The boundary of the Mandelbrot set is infinitely complicated – a small part is shown in Figure 17.2 (right). Numerous images can be found on the Internet, including movies that zoom into the set at ever greater magnification.

Near any point c the Mandelbrot set looks like the Julia set $J(f_c)$. It has been proved that the Mandelbrot set is connected. It is conjectured to be locally connected (look up the definition!) but this conjecture remains open as we write. Analogues for other types of differentiable complex function have been studied, and many problems remain open.

17.6 Epilogue

In a sense, the wheel has come full circle. The contour of history has wound round the singularity of mathematical discovery, and closed. More precisely, it has returned to its starting point but climbed one level up the Riemann surface of scientific advances. In its early days, complex analysis was (almost) a branch of physics; the connections with potential theory and fluid mechanics were widely exploited. Towards the end of the nineteenth century Felix Klein offered a 'proof' of a theorem along the following lines: *think of the Riemann surface as being made of thin metal, and an electric current flowing through it* . . . . Today this would not be considered a logically convincing argument, but the physical intuition behind it led to some important mathematical discoveries. Now we are witnessing the converse process, with mathematical intuition leading to important ideas in physics. There is a two-way trade between mathematics and its applications. And whatever the attractions of beauty for its own sake, this trade is vital for the health of both mathematics and science.

References

[1] J. Anderson. *Fundamentals of Aerodynamics* (2nd edn), McGraw-Hill, Toronto 1991.

[2] M. Bader. *Space-Filling Curves*, Texts in Computational Science and Engineering **9**, Springer, Heidelberg 2013.

[3] R. L. Devaney. *Chaotic Dynamical Systems* (2nd edn), Addison-Wesley, Redwood City 1989.

[4] K. Falconer. *Fractal Geometry*, Wiley, Chichester 1990.

[5] M. Flashman. *Mapping Diagrams to Visualize Complex Analysis*, 2015; www.geogebra.org/book/title/id/Ni69jyKs

[6] M. Flashman. 2017; Retrieved from www.geogebra.org/o/bGUrTgZq.

[7] J. Gray. *Plato's Ghost*, Princeton University Press, Princeton 2008.

[8] A. Hatcher. *Algebraic Topology*, Cambridge University Press, Cambridge 2009.

[9] J. G. Hocking and G. S. Young. *Topology*, Addison-Wesley, Reading 1961.

[10] H. J. Keisler. *Foundations of Infinitesimal Calculus*, Prindle, Weber & Schmidt, Boston 1976.

[11] M. Kline. *Mathematical Thought from Ancient to Modern Times*, Oxford University Press, Oxford 1972.

[12] A. Kyrala. *Applied Functions of a Complex Variable*, Wiley, New York 1972.

[13] B. Mandelbrot. *The Fractal Geometry of Nature* (2nd edn), Freeman, San Fancisco 1982.

[14] T. Needham. *Visual Complex Analysis*, Oxford University Press, Oxford 1997.

[15] R. Remmert. *Theory of Complex Functions*, Springer, New York 1998.

[16] B. Riemann. 'Ueber die Anzahl der Primzahlen unter einer gegebenen Grösse', *Monatsberichte der Berliner Akademie* (1859): 136–144.

[17] A. Robinson. *Non-standard Analysis*, North Holland, Amsterdam 1966.

[18] H. Sagan. *Space-Filling Curves*, Springer, New York 1994. Reissued 2013 in Universitext series.

[19] I. Stewart. *Why Beauty is Truth*, Basic Books, New York 2007.

[20] I. Stewart. *Significant Figures*, Profile, London 2017.

[21] I. Stewart and D. O. Tall. *Foundations of Mathematics* (2nd edn), Oxford University Press, Oxford 2015.

[22] K. D. Stroyan. 'Uniform continuity and rates of growth of meromorphic functions', in *Contributions to Non-Standard Analysis* (eds. W. J. Luxemburg and A. Robinson), North-Holland, Amsterdam 1972, 47–64.

[23] D. O. Tall. 'Looking at graphs through infinitesimal microscopes, windows and telescopes', *Mathematical Gazette* **64** (1980): 22–49.

[24] D. O. Tall and M. Katz. 'A cognitive analysis of Cauchy's conceptions of function, continuity, limit, and infinitesimal, with implications for teaching the calculus', *Educational Studies in Mathematics* **86** (2014): 97–124.

[25] J. Wallis. *A Treatise of Algebra*, John Playford, London 1685.

Index

Praise for the first edition:

'The authors strike a pleasant balance between informality and rigour and the reader is always kept well informed of the general picture … it has a large number of worked examples and exercises and I recommend it as a text for students.'
A. F. Beardon, ***Bulletin of the London Mathematical Society***

'Ian Stewart and David Tall are excellent mathematical communicators, and their expertise in both mathematics and mathematics education shines out of this book. It provides not only clear explanations of the technicalities but also a wealth of visual representations from which students can develop intuition for the meanings of the key definitions and for the deep connections between the results. Complex analysis is a beautiful subject, and there is no-one better to learn it from.'
Lara Alcock,
Loughborough University

This new edition of a classic textbook develops complex analysis from the established theory of real analysis by emphasising the differences that arise as a result of the richer geometry of the complex plane. Key features of the authors' approach are to use simple topological ideas to translate visual intuition into rigorous proof, and, in this edition, to address the conceptual conflicts between pure and applied approaches head-on.

Beyond the material of the clarified and corrected original edition, there are three new chapters: Chapter 15 on infinitesimals in real and complex analysis; Chapter 16 on homology versions of Cauchy's Theorem and Cauchy's Residue Theorem, linking back to geometric intuition; and Chapter 17 outlines some more advanced directions in which complex analysis has developed, and continues to evolve into the future.

With numerous worked examples and exercises, clear and direct proofs, and a view to the future of the subject, this is an invaluable companion for any modern complex analysis course.

IAN STEWART, FRS is Emeritus Professor of Mathematics at the University of Warwick. He is author or coauthor of over 190 research papers and is the bestselling author of over 120 books, from research monographs and textbooks to popular science and science fiction. His awards include the Royal Society's Faraday Medal, the IMA Gold Medal, the AAAS Public Understanding of Science Award, the LMS/IMA Zeeman Medal, the Lewis Thomas Prize, and the Euler Book Prize. He is an honorary wizard of the Discworld's Unseen University.

DAVID TALL is Emeritus Professor of Mathematical Thinking at the University of Warwick and is known internationally for his contributions to mathematics education. He is author or coauthor of over 200 papers and 40 books and educational computer software, covering all levels from early childhood to research mathematics.

Cover illustration: © Getty Images.

Online Resources
www.cambridge.org/Stewart&Tall2ed

List of the changes from the first to the second edition

Cover designed by Rob Lock

CAMBRIDGE
UNIVERSITY PRESS
www.cambridge.org

ISBN 978-1-108-43679-3

9 781108 436793