Graduate Texts in Mathematics

GTM

Richard Beals
Roderick S. C. Wong

More Explorations in Complex Functions

Springer

Graduate Texts in Mathematics 298

Graduate Texts in Mathematics

Graduate Texts in Mathematics bridge the gap between passive study and creative understanding, offering graduate-level introductions to advanced topics in mathematics. The volumes are carefully written as teaching aids and highlight characteristic features of the theory. Although these books are frequently used as textbooks in graduate courses, they are also suitable for individual study.

Richard Beals • Roderick S. C. Wong

More Explorations in Complex Functions

Springer

Richard Beals
Department of Mathematics
Yale University
New Haven, CT, USA

Roderick S. C. Wong
City University of Hong Kong
Kowloon Tong, Hong Kong

ISSN 0072-5285 ISSN 2197-5612 (electronic)
Graduate Texts in Mathematics
ISBN 978-3-031-28287-4 ISBN 978-3-031-28288-1 (eBook)
https://doi.org/10.1007/978-3-031-28288-1

Mathematics Subject Classification: 30-01, 33-01, 30D35, 33E05, 11M06

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

In the preface to *Explorations in Complex Functions*, the authors noted that “A first course in complex analysis introduces keys that open many doors. … The doors open on many subjects of interest. Too many subjects, in fact, to cover in a single follow-up course. … Our purpose is to provide brief, but self-contained introductions to many of the subjects alluded to above.” We felt that such a book could be useful – for independent reading, as a source of material for presentation in a seminar, or as a text for a second course in the subject. The first author used the material of *Explorations* in this way in a one-semester course for two successive years. The courses covered two different (though overlapping) subsets of roughly half of the chapters past the “basics.”

Anyone familiar with complex analysis could see that *Explorations* did not exhaust the topics that such a book might cover. Eventually the authors decided that the book did not completely exhaust themselves, either. We envision the same kind of uses for the present book – independent reading, seminar topics, or for a semester or year-long second course in complex analysis that gives a broad overview of some important parts of the subject.

The present book is independent of, and has minimal overlap with, *Explorations*, and is essentially self-contained. We begin with two chapters that are meant to be used as a resource, rather than as regular reading – sections of these chapters can be drawn on as needed as background for later chapters. Both the introductory chapters contain proofs, or sketches of proofs, of all the material that they cover. The first chapter is (almost) the same as in *Explorations*, reviewing material that is standard in an introductory course. The second chapter covers, quickly, some topics from that book and some additional topics that will be used in more than one chapter in this book. These additions include Carathéodory’s theorem that a conformal map between Jordan domains extends to the boundaries, and an introduction to weak solutions and Weyl’s lemma.

The order of the remaining chapters is somewhat arbitrary. Chapter 3 and Chapter 4 are stand-alone introductions to complex dynamics and to univalent function theory, respectively. Chapter 3 treats iteration of a rational function. It covers basic facts about the Fatou and Julia sets and the roles played by different types of fixed points.

Chapter 4 begins with a capsule history of the Bieberbach conjecture, introduces the basic results of Koebe and Bieberbach, and continues through Carathéodory convergence and Loewner's equation. After covering the Robertson and Milin conjectures, the chapter ends with Weinstein's short proof of de Branges's theorem: the verification of the Bieberbach conjecture.

The next three chapters can be treated as a unit leading to the uniformization theorem: the characterization of simply connected Riemann surfaces. Chapter 5 follows Perron's approach to the Dirichlet problem via subharmonic functions. General Riemann surfaces, universal covers, cover transformations, and some consequences of the uniformization theorem are covered in Chapter 6. Chapter 7 contains the proof of the uniformization theorem itself.

Chapter 8 and Chapter 9 carry the theory of Riemann surfaces further. Chapter 8 is a stand-alone introduction to quasiconformal mapping through modules, extremal ring domains, the Beurling–Ahlfors extension, Hölder continuity, and the Beltrami equation. This chapter paves the way for the use of the uniformization theorem and quasiconformal equivalence to attack the problem of moduli of Riemann surfaces in Chapter 9 on Teichmüller theory.

The remaining five chapters are (largely) stand-alone introductions to topics of both theoretical and applied interest. Chapter 10 treats the Bergman kernel and the Bergman metric, with applications to conformal mapping for simply connected and multiply connected plane domains and to the Dirichlet problem. Chapter 11 introduces theta functions, particularly for hyperelliptic curves, and the approaches of Riemann and Weierstrass to the Jacobi inversion problem.

The final three chapters have applications to approximation theory and to asymptotics. Chapter 12 deals with Padé approximants and the connections with continued fractions, orthogonal polynomials, and the Stieltjes transform. Chapter 13 treats the original Riemann–Hilbert problem and some of its generalizations and applications, such as integral transforms and integral equations. Chapter 14 covers Darboux's method for computing asymptotics of Maclaurin expansions, and some recent generalizations.

Altogether, there is more material here than one could expect to cover in a yearlong course in complex analysis. How much *can* be covered in one or two semesters will depend on the degree of preparation of the class. The authors hesitate, therefore, to make specific suggestions – especially since the choice of topics will depend very much on the interests of the instructor and/or the students. It has been pointed out to the authors that material selected from Chapters 2, 4, 5, 7, and 8 contains the function theory background for some stochastic equations of current interest, such as SLE.

For an overview of dependence relations of chapters, see Fig. 0.1.

The authors gratefully acknowledge the patience and support of their wives, Nancy Beals and Edwina Wong, and the encouragement and helpfulness of their Springer editors, Loretta Bartolini and Elizabeth Loew. The authors are also grateful to Professor Wei-Yuan Qiu for many helpful comments on Chapter 3.

Dependence relations among chapters:

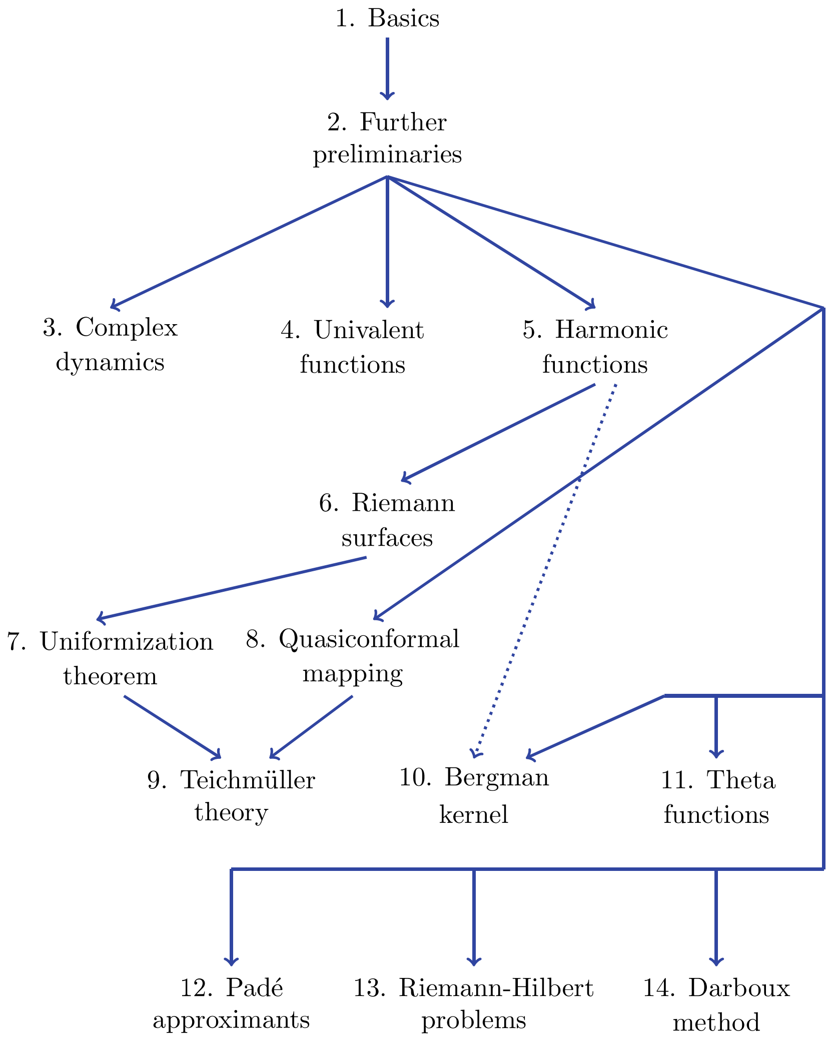

Fig. 0.1 Chart of dependence relations among chapters

New Haven, Connecticut, USA
Kowloon Tong, Hong Kong

Richard Beals
Roderick S. C. Wong

Contents

Chapter 1
Basics

This chapter begins with a brief summary of facts from a standard introductory complex variables course: Cauchy's formula and consequences, isolated singularities, residues, and the complex logarithm. Also included are four topics that are not as standard for an elementary course, but are used in the following chapters: reflection properties, infinite products, analytic continuation, and harmonic functions. For all this material we give brief discussions and sketches of proofs.

1.1 Introduction; notation

Throughout, a *domain* Ω is a connected non-empty open subset of the complex plane $\mathbb{C}$. A function $f : \Omega \to \mathbb{C}$ is said to be C^n, or a C^n *function* if all partial derivatives of f of order $\leq n$ exist and are continuous. The space of such functions is denoted $C^n(\Omega)$. The space $C^\infty(\Omega)$ of C^∞ functions is defined similarly. Partial derivatives are often denoted by subscripts: f_x, f_y, f_{xx},f_{xy}, etc.

A (parametrized) *curve* in Ω is a continuous function γ defined on a real interval $I = [a, b]$ and having values in Ω. We commit the usual abuse of terminology by using the term "curve" interchangeably for the continuous function $\gamma : I \to \Omega$ and for the image $\gamma(I)$ in $\mathbb{C}$. The image carries an orientation from the parametrization. The curve γ is said to be *smooth* if γ is a C^1 function on the closed interval. The curve γ is said to be *piecewise smooth* if it is smooth on each of finitely many subintervals whose union is $[a, b]$.In some contexts a curve, or a part of a curve, may be referred to as an *arc* or a *path*..

Similarly, a curve γ is said to be *analytic* if it is real-analytic, i.e. for each $t_0 \in [a, b]$, γ is given for nearby values by a convergent power series

$$\gamma(t) = \sum_{n=0}^{\infty} a_n (t - t_0)^n.$$

R. Beals and R. S. C. Wong, *More Explorations in Complex Functions*, Graduate Texts in Mathematics 298, https://doi.org/10.1007/978-3-031-28288-1_1

Note that this means that γ can be extended so as to be defined on an open neighborhood of its (original) domain, defined by these same power series for complex values of t.

A curve $\Gamma : [a, b] \to \mathbb{C}$ is said to be *closed* if the endpoints coincide: $\gamma(a) = \gamma(b)$. A curve is said to be *simple* if its image has no self-intersections.

In this chapter it is assumed that any domain Ω that occurs is bounded and that the boundary $\partial\Omega$ is the union of finitely many pairwise disjoint simple smooth closed curves, oriented so that Ω lies to the left of each boundary curve.

If $\gamma : [a.b] \to \Omega$ is a curve and f is a continuous function defined on the image of γ, then the integral

$$\int_\gamma f = \int_\gamma f(z)\,dz$$

is defined to be the limit as $\max\{|z_{j+1} - z_j|\}$ tends to zero of the Riemann sums over partitions $a = x_0 < x_1 < \cdots < x_{n+1} = b$,

$$\sum_{j=0}^{n} f(x_j)[\gamma(x_{j+1}) - \gamma(x_j)].$$

In the case of double integrals it will often be convenient to write $dm(z)$ for $dx\,dy$:

$$\iint_\Omega f(x+iy)\,dx\,dy = \iint_\Omega f(z)\,dm(z).$$

(As usual, it is understood that x, y are real, and that a function of $z = x + iy$ can be considered as a function of x and y, and conversely.)

If $z = x + iy$, the *complex conjugate* $\bar{z}$ is $x - iy$. Thus the *real part* $\operatorname{Re} z$ and *imaginary part* $\operatorname{Im} z$ of $z = x + iy$ are

$$x = \operatorname{Re} z = \frac{1}{2}(z + \bar{z}); \qquad y = \operatorname{Im} z = \frac{1}{2i}(z - \bar{z}).$$

The *polar decomposition* of $z = x + iy$ is essentially the representation of z in polar coordinates

$$z = r\,e^{i\theta} = r\cos\theta + i\sin\theta, \tag{1.1.1}$$

so

$$r = -\sqrt{x^2 + y^2}, \qquad \theta = \tan^{-1}\frac{y}{x}.$$

1.2 The Cauchy–Riemann equations and Cauchy's integral theorem

Consider a function

$$f(x + iy) = u(x, y) + i\,v(x, y), \quad x + iy \in \Omega,$$

where u and v are real-valued C^1 functions. The complex-valued function f is said to be *holomorphic* (differentiable in the complex sense), if and only if u and v satisfy the *Cauchy–Riemann equations*:

$$u_x = v_y, \qquad u_y = -v_x, \tag{1.2.1}$$

where the subscripts denote partial differentiation.

Green's theorem (or an argument due to Goursat that uses only pointwise differentiability) yields the basic theorem of the subject.

Theorem 1.2.1. (Cauchy integral theorem) *If f is holomorphic in a domain Ω, and continuous on the closure of Ω, then*

$$\int_{\partial\Omega} f(\zeta)\, d\zeta = 0. \tag{1.2.2}$$

Let us pause to look at the Cauchy–Riemann equations and Cauchy's theorem from the point of view of differential forms and Green's theorem. The pairs of 1-forms dz, $d\bar{z}$, and dx, dy are related by

$$dz = dx + idy, \qquad d\bar{z} = dx - idy;$$

$$dx = \frac{dz + d\bar{z}}{2}, \qquad dy = \frac{dz - d\bar{z}}{2i}.$$

Thus

$$df = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy = \frac{1}{2}\left[\frac{\partial f}{\partial x} - i\frac{\partial f}{\partial y}\right] dz + \frac{1}{2}\left[\frac{\partial f}{\partial x} + i\frac{\partial f}{\partial y}\right] d\bar{z}.$$

It is natural to express this as

$$df = \frac{\partial f}{\partial z} dz + \frac{\partial f}{\partial \bar{z}} d\bar{z} = \partial f\, dz + \bar{\partial} f\, d\bar{z},$$

where

$$\partial = \frac{\partial}{\partial z} = \frac{1}{2}\left[\frac{\partial}{\partial x} - i\frac{\partial}{\partial y}\right]; \qquad \bar{\partial} = \frac{\partial}{\partial \bar{z}} = \frac{1}{2}\left[\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right]. \tag{1.2.3}$$

With $f = u + iv$ we find that

$$\partial f = \frac{1}{2}\left[(u_x + v_y) + i(v_x - u_y)\right], \qquad \bar{\partial} f = \frac{1}{2}\left[(u_x - v_y) - i(v_x + u_y)\right]. \tag{1.2.4}$$

Thus the Cauchy–Riemann equations (1.2.1) are equivalent to the single equation $\bar{\partial} f = 0$. Moreover they imply that for holomorphic $f = u + iv$,

$$\partial f = u_x + iv_x = f', \qquad \bar{\partial} \bar{f} = u_x - iv_x = \overline{f'}. \tag{1.2.5}$$

A standard form of Green's theorem is that if Ω is a domain, then

$$\int_{\partial\Omega}[P\,dx + Q\,dy] \;=\; \iint_{\Omega}[Q_x - P_y]\,dx\,dy. \tag{1.2.6}$$

It is an exercise, using the identities above, to show that (1.2.6) is equivalent to the equation

$$\int_{\partial\Omega}[f\,dz + g\,d\bar{z}] \;=\; 2i\iint_{\Omega}[\bar{\partial} f - \partial g]\,dx\,dy. \tag{1.2.7}$$

In particular, taking $g = 0$ we obtain a result known as the *Cauchy–Green formula*

$$\int_{\partial\Omega} f\,dz \;=\; 2i\iint_{\Omega}\bar{\partial} f\,dx\,dy. \tag{1.2.8}$$

The case $\bar{\partial} f = 0$ is Cauchy's formula (1.2.2).

Another application of these identities is to the calculation of the *area* of the image of a domain Ω under an injective holomorphic function f whose first and second partial derivatives are continuous up to the boundary. If $f = u + iv$, the area is, taking account of the Cauchy–Riemann equations:

$$\begin{aligned}\iint_{\Omega}\begin{vmatrix} u_x & v_x \\ u_y & v_y \end{vmatrix}\,dx\,dy \\ &= \iint_{\Omega}\begin{vmatrix} u_x & v_x \\ -v_x & u_x \end{vmatrix}\,dx\,dy \\ &= \iint_{\Omega}[u_x^2 + v_x|^2\,dx\,dy.\end{aligned}$$

Taking into account (1.2.5) we have two area formulas for injective holomorphic f:

$$\text{Area}\{f(\Omega)\} \;=\; \iint_{\Omega}|f'|^2\,dx\,dy \;=\; \iint_{\Omega}\partial f\,\overline{\partial f}\,dx\,dy. \tag{1.2.9}$$

Since $\bar{\partial}\partial f = 0$,

$$\partial f\,\overline{\partial f} \;=\; \bar{\partial}(f'\bar{f}).$$

It follows from (1.2.8) that

$$\text{Area}\,\{f(\Omega)\} \;=\; \frac{1}{2i}\int_{\partial\Omega} f'(z)\overline{f(z)}\,dz. \tag{1.2.10}$$

1.3 The Cauchy integral formula and applications

Much of basic complex function theory consists of exploring (fairly immediate) consequences of the Cauchy integral theorem, Theorem 1.2.1. One such consequence is the *Cauchy integral formula*. If f is holomorphic in a general domain Ω, and continuous on the closure, we can apply (1.2.2) to the function

$$g(w) \;=\; \frac{1}{2\pi i}\cdot\frac{f(w)}{w-z}, \qquad w \in \Omega,$$

on the domain Ω_ε formed by removing from Ω a small disk centered at z,

$$D_\varepsilon(z) \;=\; \{w : w = z + re^{i\theta},\ 0 \le r < \varepsilon, 0 \le \theta \le 2\pi\}.$$

The integral over the boundary of D_ε, oriented in the positive (counter-clockwise) direction, approaches $2\pi f(z)$ as $\varepsilon \to 0$; see the calculation (1.3.4). Taking the limit yields the formula (1.3.1). This formula can be differentiated arbitrarily often.

Theorem 1.3.1. (Cauchy integral formula) *If f is holomorphic in a domain Ω, and continuous on the closure of Ω, then for each $z \in \Omega$,*

$$f(z) \;=\; \frac{1}{2\pi i}\int_{\partial\Omega}\frac{f(\zeta)}{\zeta - z}\,d\zeta. \tag{1.3.1}$$

More generally, each derivative can be written as an integral:

$$f^{(n)}(z) \;=\; \frac{n\,!}{2\pi i}\int_{\partial\Omega}\frac{f(\zeta)\,d\zeta}{(\zeta - z)^{n+1}}. \tag{1.3.2}$$

Thus a holomorphic function is infinitely differentiable. Moreover, if

$$|z - z_0| \;<\; r \;=\; \inf_{\zeta\in\partial\Omega}|\zeta - z_0|,$$

then the expansion

$$\frac{1}{\zeta - z} \;=\; \frac{1}{(\zeta - z_0)\cdot\left[1 - \dfrac{z - z_0}{\zeta - z_0}\right]} \;=\; \sum_{n=0}^{\infty}\frac{(z - z_0)^n}{(\zeta - z_0)^{n+1}}$$

converges uniformly for $\zeta \in \partial\Omega$. This gives:

Theorem 1.3.2. (Taylor expansion) *If f is holomorphic in a disk $D_r(z_0)$, then f has a convergent Taylor expansion*

$$f(z) \;=\; \sum_{n=0}^{\infty}a_n(z - z_0)^n, \quad |z - z_0| < r; \qquad a_n \;=\; \frac{f^{(n)}(z_0)}{n\,!}. \tag{1.3.3}$$

In other words, a holomorphic function is an *analytic* function of z.

Remark. If f is holomorphic in a neighborhood of 0, the Taylor expansion centered at $z = 0$,

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

is often referred to as the *Maclaurin expansion*.

Other easy consequences of the Cauchy integral formula are various *mean value* and *maximum* principles. For example, if f is holomorphic in a domain that includes the closure of a disk $D_r(z)$, then a change of variables

$$\zeta = z + re^{i\theta}$$

gives

$$f(z) = \frac{1}{2\pi i}\int_{|\zeta - z| = r} \frac{f(\zeta)}{\zeta - z}\, d\zeta = \frac{1}{2\pi}\int_0^{2\pi} f(z + re^{i\theta})\, d\theta. \tag{1.3.4}$$

One can also take the real or imaginary part of this formula.

Theorem 1.3.3. (Mean value property) *If f is holomorphic in a domain Ω, then the value of f at each point $z_0 \in \Omega$ is the mean of the values on any circle $\{z : |z - z_0| = r\}$ that is small enough so that $D_r(z_0)$ is contained in Ω. The real and imaginary parts of f have the same property.*

It is an easy consequence of Theorem 1.3.3 that the maximum value of the modulus $|f(z)|$, or of the real or imaginary part of f, occurs at the boundary of Ω. A closer examination of (1.3.4), taking into account the Taylor expansion, shows that no such maximum value can occur at an interior point of Ω, unless f is constant near the point.

Theorem 1.3.4. (Maximum modulus principle) *If f is holomorphic in Ω and continuous on the closure of Ω, then the maximum value of the modulus $|f(z)|$ is attained on the boundary. The same is true for the real and imaginary parts of f.*

By assumption a domain Ω is connected, so it is easily seen that if f is constant near a point, it is constant throughout Ω. Therefore Theorem 1.3.4 has a more precise form.

Theorem 1.3.5. (Strong maximum modulus principle) *If f is holomorphic in a domain Ω, and the maximum modulus is attained at a point of Ω itself, then f is constant. The same is true for the real and imaginary parts of f.*

We note here another frequently used consequence of the Cauchy integral formula.

Proposition 1.3.6. *Suppose that $\{f_n\}_{n=1}^{\infty}$ is a sequence of functions holomorphic in a domain Ω, and suppose that the sequence converges to a function f, uniformly on each compact subset of Ω. Then f is holomorphic in Ω.*

In fact if $z \in \Omega$, the convergence is uniform on a small circle Γ that contains z. Therefore in the disk bounded by Γ the limit function f is given by the Cauchy integral formula, from which it follows that f is holomorphic in that disk.

An *entire function* is a function f that is holomorphic in the entire plane $\mathbb{C}$. For each $R > 0$ and each $z \in \mathbb{C}$, (1.3.1) and (1.3.2) give

$$f(z) = \frac{1}{2\pi i} \int_{|\zeta - z| = R} \frac{f(\zeta)}{\zeta - z} \, d\zeta$$

and, more generally,

$$f^{(n)}(z) = \frac{n!}{2\pi i} \int_{|\zeta - z| = R} \frac{f(\zeta)}{(\zeta - z)^{n+1}} \, d\zeta.$$

Since the circle of integration has length $2\pi R$ and the modulus of the denominator is R^{n+1}, it is easy to see that constraints on the growth of f can imply vanishing of high order derivatives.

Theorem 1.3.7. (Liouville's theorem) *If f is entire and bounded, then f is constant.*

Theorem 1.3.8. (Extended Liouville theorem) *If f is entire and*

$$|f(z)| \leq C(|z|^n + 1)$$

for some integer $n \geq 0$, then f is a polynomial of degree $\leq n$.

1.4 Change of contour, isolated singularities, residues

The Cauchy integral theorem is often used to justify a *change of contour* in an integration. This is particularly useful in the rest of this section. Rather than formulate a general theorem, we illustrate with an example. Suppose that the domain Ω is bounded by one large circle Γ and two smaller, disjoint circles, Γ_1, Γ_2, that are enclosed by Γ, as in Figure 1.1 on the left. Suppose that f is holomorphic in Ω and continuous on the closure. Then

$$\int_{\Gamma} f(z)\, dz = \int_{\Gamma_1} f(z)\, dz + \int_{\Gamma_2} f(z)\, dz, \tag{1.4.1}$$

where each circle is oriented in the positive (counter-clockwise) direction.

In fact, Theorem 1.2.1 implies that the integral of f over the contour on the right in Figure 1.1 is zero. In the limit, as the gap is closed, the integrals over the flat parts of the contour cancel, and we are left with (1.4.1) in the form

$$\int_{\Gamma} f(z)\, dz - \int_{\Gamma_1} f(z)\, dz - \int_{\Gamma_2} f(z)\, dz = 0.$$

An *isolated singularity* for a holomorphic function is a point z_0 such that f is holomorphic in a punctured disk $\Omega = \{z : 0 < |z - z_0| < r\}$.

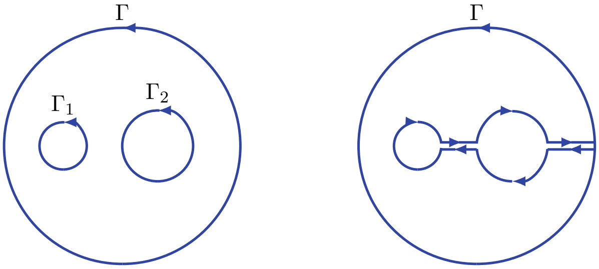

Fig. 1.1 Change of contour in integration.

An isolated singularity z_0 is said to be a *removable singularity* if a value $f(z_0)$ can be assigned to f at z_0 in such a way that the extended function is holomorphic in some disk $\{z : |z - z_0| < r\}$.

An isolated singularity z_0 is said to be a *pole* if there is some integer $n > 0$ such that

$$f(z) = \frac{a_{-n}}{(z-z_0)^n} + \frac{a_{1-n}}{(z-z_0)^{n-1}} + \cdots + a_0 + a_1(z-z_0) + \ldots \qquad (1.4.2)$$

in some punctured disk $\{0 < |z - z_0| < r\}$, with $a_{-n} \neq 0$. The expansion (1.4.2) is called the *Laurent expansion* of f at z_0. The *order* of the pole is n. A *simple pole* is a pole of order 1.

Suppose that the function f is bounded and holomorphic in the punctured disk $\{z : 0 < |z - z_0| < R\}$. Choosing a smaller radius, we may assume that f is continuous up to the circle $\{z : |z - z_0| = r\}$. Let $g(z) = (z - z_0)f(z)$ and $g(z_0) = 0$, so $g(z)$ is continuous at z_0. Using the Cauchy integral formula for $\{z : \varepsilon < |z - z_0| < r\}$ and letting $\varepsilon \to 0$, we find that g is given by the Cauchy integral formula and is therefore holomorphic near 0. If follows that the same is true for $f = g/(z - z_0)$. Thus

Proposition 1.4.1. *Suppose that z_0 is an isolated singularity of f and suppose that $f(z)$ is bounded for* $0 < |z - z_0| < r$. *Then z_0 is a* removable singularity: *$f(z)$ has a limit at $z = z_0$ and extends to be holomorphic in $D_r(z_0)$.*

Corollary 1.4.2. *Suppose that z_0 is an isolated singularity of f. Suppose that for some integer n, $g(z) = (z - z_0)^n f(z)$ is bounded as $z \to z_0$, and suppose that n is the least such integer. If n is negative, it follows that z_0 is a removable singularity, at which f has a zero of order $-n$. If n is positive, then f has a pole of order n at z_0,*

An isolated singularity that is neither removable nor a pole is called an *essential singularity*. An example is the function $g(z) = \exp(1/z)$ on $\Omega = \mathbb{C} \setminus \{0\}$. In this case the behavior near 0 is quite different. It is an exercise to show that g takes any given non-zero value a infinitely often in each neighborhood of 0. A weaker version of this is easily proved for essential singularities in general. (For a stronger version, see Theorem 2.5.2.)

Theorem 1.4.3. (Casorati–Weierstrass theorem) *Suppose that f is holomorphic in a domain Ω and has an essential singularity at $z_0 \in \Omega$. In each punctured neighborhood $D_\varepsilon = \{z : 0 < |z - z_0| < \varepsilon\}$, f comes arbitrarily close to any given complex number a.*

Proof: Suppose, to the contrary, that $|f(z) - a| \geq \delta > 0$ in D_ε. Then $g(z) = 1/[f(z) - a]$ has an isolated singularity at z_0. Moreover, g is bounded as $z \to z_0$, so the singularity is removable. If $g(z_0) \neq 0$, then f has a removable singularity at z_0. If g has a zero of degree $n > 0$ at z_0, then f has a pole of order n at z_0. □

Let us return to the Laurent expansion (1.4.2). Suppose that f is holomorphic in $\{z : 0 < |z - z_0| < R\}$. Then $(z - z_0)^{-1-n} f(z)$ can be integrated term-by-term over the boundary of the domain $\{z : \varepsilon < |z - z_0| < r < R\}$. Taking $\varepsilon \to 0$, we find that

$$a_n = \frac{1}{2\pi i} \int_{|z-z_0|=r} \frac{f(z)\,dz}{(z - z_0)^{n+1}}. \tag{1.4.3}$$

In particular, the coefficient a_{-1} is defined to be the *residue* $\mathrm{res}(f, z_0)$ of f at z_0:

$$\mathrm{res}(f, z_0) = \frac{1}{2\pi i} \int_{|z-z_0|=r} f(z)\,dz. \tag{1.4.4}$$

A function f is said to be *meromorphic* in a domain Ω if f is holomorphic except at isolated points that are poles of f. An application of Cauchy's theorem to the domain minus sufficiently small disks centered at the poles gives the following.

Theorem 1.4.4. (Residue theorem) *If f has finitely many poles in Ω and is continuous on the closure, then*

$$\frac{1}{2\pi i} \int_{\partial\Omega} f(\zeta)\,d\zeta = \sum_{z \in \Omega} \mathrm{res}(f, z). \tag{1.4.5}$$

The residue theorem can be used to *count* poles and zeros (taking into account multiplicities). In fact, suppose that near $z = z_0$, $f(z) = (z - z_0)^n g(z)$, where n is an integer, g is holomorphic and $g(z_0) \neq 0$. then

$$\frac{f'(z)}{f(z)} = \frac{n}{z - z_0} + \frac{g'(z)}{g(z)}$$

has residue n at z_0. As a consequence:

Theorem 1.4.5. (Counting zeros and poles) *If f is meromorphic in Ω, and continuous and nowhere zero at the boundary, then*

$$\frac{1}{2\pi i} \int_{\partial\Omega} \frac{f'(\zeta)}{f(\zeta)}\,d\zeta$$
$$= \textit{number of zeros minus number of poles of } f \textit{ in } \Omega, \tag{1.4.6}$$

where the zeros and poles are counted according to multiplicity.

Corollary 1.4.6. *If f is meromorphic in Ω and continuous on the boundary, then it takes each value in the complement of $f(\partial\Omega)$ the same number of times (counting multiplicity) in each connected component of this complement..*

Proof: If f does not take the value a on the boundary, then the integral

$$N(a) = \frac{1}{2\pi i}\int_{\partial\Omega}\frac{f'(\zeta)}{f(\zeta)-a}\,d\zeta$$

counts the number of times f takes the value a minus the number of poles. The number of poles is constant, and $N(a)$, being integer-valued and continuous with respect to a, is also constant on the connected component of the complement that contains a. □

Here are two more applications of these ideas.

Theorem 1.4.7. (Rouché's theorem) *Suppose that f and g are holomorphic in Ω and continuous on the closure. If $|f(z)-g(z)| < |f(z)|$ on the boundary $\partial\Omega$, then f and g have the same number of zeros in Ω.*

In fact the function $f_s(z) = (1-s)f(z) + sg(x) = f(z) - s[f(z)-g(z)]$, $0 \le s \le 1$, has no zeros on $\partial\Omega$, so the number of zeros in Ω is

$$\frac{1}{2\pi i}\int_{\partial\Omega}\frac{f_s'(\zeta)}{f_s(\zeta)}\,d\zeta.$$

This is an integer-valued continuous function of s, so it has the same value at $s = 0$ and at $s = 1$. But $f_0 = f$, $f_1 = g$.

Theorem 1.4.8. (Inverse function theorem) *Suppose that f is holomorphic near z_0 and $f'(z_0) \ne 0$. Then f has an inverse that is holomorphic near $f(z_0)$.*

In fact it follows from the series expansion at z_0 that for small $r > 0$, $f(z) \ne f(z_0)$ if z is inside or on the curve $\Gamma = \{z : |z-z_0| = r\}$. Therefore if a is close enough to $f(z_0)$, the integral

$$\frac{1}{2\pi i}\int_{\Gamma}\frac{\zeta\, f'(\zeta)}{f(\zeta)-a}\,d\zeta$$

is the unique value of z inside the curve such that $f(z) = a$. This expression is a holomorphic function of a.

1.5 The logarithm and powers

In view of (1.1.1), the complex *logarithm* $\log z$, $z \ne 0$, is defined by

$$\log z = \log(|z|e^{i\arg z}) = \log|z| + i\arg z. \tag{1.5.1}$$

Here $\log|z|$ denotes the usual choice for positive argument; thus $\log|z|$ is real. Of course $\arg z$ is defined only up to addition of an integer multiple of 2π. By a *branch* of the logarithm in a domain Ω, we mean a choice that is holomorphic throughout Ω. (Such a choice may not be possible, e.g. in a deleted neighborhood of the origin $\{z : 0 < |z| < r\}$.) A branch is called the *principal branch* if $\Omega \cap \mathbb{R}$ is not empty and $\log z$ is real on this intersection.

An important concept here is that of a *simply connected* domain,usually defined to be one that is connected and in which each closed curve can be continuously shrunk to a point. An equivalent definition is that Ω is connected and, given two curves γ_0 and γ_1 in Ω that join points z and w, there is a family of curves $\gamma_t : [0, 1] \to \Omega, 0 < t < 1$, such that $\gamma_t(0) = z$, $\gamma_t(1) = w$, and the map $(s, t) \to \gamma_t(s)$ is continuous, $0 \le s, t \le 1$. (Showing that the two definitions are equivalent is an interesting exercise.)

Suppose that Ω is a simply connected domain. Suppose also that 0 is not in Ω. Then a branch of the logarithm may be obtained by choosing $z_0 \in \Omega$, choosing $\log z_0$, and setting

$$\log z \;=\; \log z_0 + \int_{z_0}^{z} \frac{d\zeta}{\zeta}. \tag{1.5.2}$$

Because of the assumption that Ω is simply connected, the integral is independent of the path of integration from z_0 to z: see Section 1.8 for details.

Corresponding to a branch of the logarithm, and to each $\alpha \in \mathbb{C}$, there is a branch of the power z^α:

$$z^\alpha \;=\; e^{\alpha \log z}. \tag{1.5.3}$$

This is independent of the branch of the logarithm if and only if α is an integer.

The next result is a generalization of Theorem 1.4.4 to the case in which the function f is defined only in a neighborhood of the curve of integration.

Theorem 1.5.1. *(***Argument principle***) Suppose that f is holomorphic in a neighborhood of a closed curve Γ, and suppose that z_0 is not in the image $f(\Gamma)$. Then the integral*

$$n(z_0) \;=\; \frac{1}{2\pi i}\int_\Gamma \frac{f'(\zeta)\,d\zeta}{f(\zeta) - z_0}$$

is an integer: the number of times that the curve $f(\Gamma)$ wraps around z_0 in the positive direction.

Proof: Let $g(\zeta) = f(\zeta) - z_0$. By assumption, $g \neq 0$ for $\zeta \in \Gamma$, so we may choose a branch of the logarithm at a point $\zeta_0 \in \Gamma$ and follow the logarithm continuously along the curve. When we return to the starting point ζ_0, the logarithm will have the same real part as initially, but the imaginary part will differ by $2\pi n$, where the integer n can be interpreted as the number of times that $g(\Gamma)$ wraps around the origin in the positive direction. Equivalently, n is the number of times that $f(\Gamma)$ wraps around z_0 in the positive direction. Thus

$$\frac{1}{2\pi i}\int_\Gamma \frac{f'(\zeta)\,d\zeta}{f(\zeta) - z_0} = \frac{1}{2\pi i}\int_\Gamma \frac{g'(\zeta)\,d\zeta}{g(\zeta)} = \frac{2n\pi i}{2\pi i} = n. \qquad \square$$

1.6 Infinite products

Infinite products are often written in the form

$$\prod_{n=1}^{\infty}(1-a_n), \tag{1.6.1}$$

where the a_n are complex numbers. The key tool to be used is the following estimate.

Lemma 1.6.1. *Suppose* $|z| \le 1/2$. *Then the principal branch of* $\log(1-z)$ *satisfies*

$$|\log(1-z)+z| \;\le |z|^2 \;\le\; \frac{|z|}{2}. \tag{1.6.2}$$

Proof: Integrating along the line segment from 1 to $1+z$,

$$\begin{aligned}\log(1-z) = \int_1^{1-z}\frac{ds}{s} &= -\int_0^z \frac{dt}{1-t}\\ = -\int_0^z (1+t+\dots)\,dt &= -z-\frac{z^2}{2}-\frac{z^3}{3}-\dots,\end{aligned}$$

so

$$|\log(1-z)+z| \;\le\; \frac{|z|^2}{2}(1+|z|+|z|^2+\dots) \;\le\; \frac{|z|^2}{2}\cdot\frac{1}{1-|z|} \;\le |z|^2. \qquad \square$$

The (formal) product (1.6.1) is said to converge if

$$\lim_{M,N\to\infty}\prod_{n=M}^{N}(1-a_n) \;=\; 1. \tag{1.6.3}$$

This implies that the partial products $\prod_M^{\infty}(1-a_n)$ have a non-zero limit, as soon as M is large enough that $n \ge M$ implies $1-a_n \ne 0$. In particular, a *necessary* condition for convergence is that $1-a_n \to 1$, i.e. $a_n \to 0$. Suppose that $|a_n| \le 1/2$ for $n \ge M$. Then, taking the principal branch of the logarithm

$$\log\left|\prod_M^N(1-a_n)\right| \;=\; \sum_{n=M}^{N}|\log(1-a_n)|.$$

The product is said to be *absolutely convergent* if

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

converges. Absolute convergence implies convergence. It follows from (1.6.2) that for large enough n,

$$\frac{|a_n|}{2} \le |\log(1 + |a_n|) \le \frac{3|a_n|}{2},$$

Therefore the product converges absolutely if and only if $\sum_{n=1}^{\infty} |a_n| < \infty$.

1.7 Reflection principles

Theorem 1.7.1. *Suppose that Ω is a domain that is symmetric under reflection about the real axis: $\overline{\Omega} = \Omega$. Suppose also that f is holomorphic on the intersection of Ω with the upper half-plane $\mathbb{H} = \{z : \operatorname{Im} z > 0\}$, continuous up to $I = \Omega \cap \mathbb{R}$, and real on I. Then f has a holomorphic extension to the remainder of Ω, with*

$$f(\bar{z}) = \overline{f(z)}, \quad z \in \Omega_+. \tag{1.7.1}$$

Proof: The prescription (1.7.1) defines f so as to be holomorphic in $\Omega \cap \mathbb{C}_-$, and continuous in all of Ω. As a domain, Ω is connected, so I is not empty. We need to show that f is holomorphic near I. Consider a complex neighborhood $D_r(x_0)$ of a point $x_0 \in I$, whose closure is contained in Ω. Let

$$\Delta_\pm = D_r(x_0) \cap \{z : \pm \operatorname{Im} z > 0\}, \tag{1.7.2}$$

and

$$g(z) = \frac{1}{2\pi i} \int_{|\zeta - x_0| = r} \frac{f(\zeta)\, d\zeta}{\zeta - z}, \qquad |z - x_0| < r.$$

This function is holomorphic in $D_r(x_0)$. For $z \in \Delta_+$ the lower semicircle of the contour can be moved to the x-axis, showing that $g = f$ on Δ_+. Similarly, $g = f$ on Ω_- if we use (1.7.1) to define f on Δ_-. It follows that (1.7.1) extends f holomorphically across I. □

Theorem 1.7.2. *Suppose that Ω and I are as in the previous theorem. Suppose that f is holomorphic in $\Omega \cap \mathbb{H}$, nowhere zero, and continuous up to I. Suppose also that $|f(x)| = 1$ for $x \in I$. Then f has a holomorphic extension to the remainder of Ω, with*

$$f(\bar{z}) = 1/\overline{f(z)}. \tag{1.7.3}$$

Proof: As in the previous proof, it is sufficient to work in a small disk $D_r(x_0)$. For small r a branch g of $\log f$ can be chosen in Δ_+. By the assumption on $|f|$, the limit of ig is real on $D_r(x_0) \cap \mathbb{R}$. Therefore ig can be continued to all of $D_r(x_0)$. The continuation of g, given by (1.7.1) for ig, exponentiates to the continuation of f given by (1.7.3). □

1.8 Analytic continuation

There are two situations that give rise to the consideration of analytic continuation. An example of one such situation is the function f defined by the series

$$f(z) = 1 + z + z^2 + z^3 + \cdots + z^n + \ldots . \tag{1.8.1}$$

The series converges if and only if $|z| < 1$. On the other hand, the sum is $1/(1-z)$, which is holomorphic in the complement of the point $z = 1$. It is natural to consider $1/(1-z)$ as a *continuation* of f: the extension of f to a function holomorphic on a larger domain. A natural question: is such an extension unique?

An example of a second such situation is the logarithm. Starting with the usual choice in a neighborhood of $z = 2$, and following along a curve that circles the origin in the positive direction, one comes back not to $\log 2$ but to $\log 2 + 2\pi i$ – but it is natural to think of this "branch" as an analytic continuation of the original. For a visualization, see Figure 1.2.

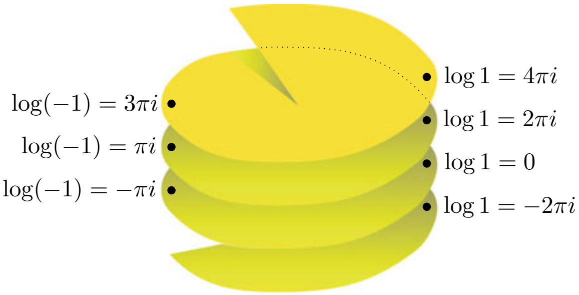

Fig. 1.2 Analytic continuation of the logarithm.

In general, suppose that f_0 is holomorphic in an open disk D_0 centered at z_0, suppose that $\gamma : [0, 1] \to \mathbb{C}$ is a curve with $\gamma(0) = z_0$, and suppose that D_0 does not contain γ (we are systematically conflating γ as a mapping and γ as a set of points, i.e. the image of the mapping). It may still be the case that we can find successive points $z_j = \gamma(t_j)$ along the curve and functions f_j holomorphic in disks D_j centered at z_j such that $D_j \cap D_{j+1} \neq \emptyset$, $f_j = f_{j+1}$ on $D_j \cap D_{j+1}$, and the union of the D_j covers γ. The result is a function f, holomorphic in a neighborhood of the curve γ, that agrees with f_0 near z_0. The function f is said to be a *continuation of f_0 along the curve γ*.

Proposition 1.8.1. (Uniqueness of analytic continuation) *If two functions that are holomorphic in a connected domain Ω agree on a non-empty open subset of Ω, then they agree on all of Ω.*

Proof: It suffices to prove that if f is holomorphic in Ω and vanishes near a point $z_0 \in \Omega$, then f is identically zero. Let z be another point of Ω and let $\gamma : [0, 1] \to \Omega$

be a smooth curve with $\gamma(0) = z_0$ and $\gamma(1) = z$. If $f(\gamma(s)) = 0$ for $0 \le s \le t$, then it follows that each derivative of f vanishes at $z = \gamma(t)$. Thus the Taylor expansion of f vanishes at $\gamma(t)$, so f vanishes in a neighborhood of $\gamma(t)$. It follows from this argument that f vanishes along the entire curve, so $f(z) = 0$. □

Recall from Section 1.5: a domain $\Omega \subset \mathbb{C}$ is said to be simply connected if each closed curve in Ω can be deformed continuously to a point (a constant curve). For example, the plane $\mathbb{C}$ is simply connected, but the complement of any non-empty bounded subset A is not. As noted above, an equivalent definition is that any two curves from a point z_0 to a point z_1 can be deformed continuously from one to the other.

Theorem 1.8.2. (Monodromy theorem) *Suppose that the domain Ω is simply connected. Suppose that f_0 is holomorphic in a domain $\Omega_0 \subset \Omega$, and suppose that f_0 can be continued along each curve in Ω. Then f_0 has a unique holomorphic extension to all of Ω.*

Proof: Take $z_0 \in \Omega_0$. It is enough to show that the continuation of f_0 along a curve $\gamma : [0, 1] \to \Omega$ that starts at z_0 leads to a value $f(\gamma(1))$ that depends only on $z_1 = \gamma(1)$, not on the particular curve γ. Suppose that γ_0 and γ_1 are two such curves from z_0 to z_1. Then there is a family of curves γ_t from z_0 to z_1, $0 < t < 1$, that interpolates continuously from γ_0 to γ_1.

Suppose that f_0 is continued along each curve γ_t. Let T be the supremum of those t such that $\gamma_t(z_1) = \gamma_0(z_1)$ It follows from Proposition 1.8.1 that T is positive. It follows from continuity that $\gamma_T(z_1) = \gamma_0(z_1)$. Then $T = 1$, since, otherwise, Proposition 1.8.1 implies that equality at z_1 can be extended past $t = T$. □

1.9 Harmonic functions

A function f that maps a domain $\Omega \to \mathbb{C}$ is said to be *harmonic* if is belongs to $C^2(\Omega)$ and satisfies *Laplace's equation*, the differential equation

$$\Delta f \;\equiv\; f_{xx} + f_{yy} \;=\; 0. \tag{1.9.1}$$

As an example, suppose that $f : U \to \mathbb{C}$ is holomorphic. Writing $f(x + iy) = u(x, y) + iv(x, y)$, where u and v are real-valued, we note that the Cauchy–Riemann equations imply

$$u_{xx} + u_{yy} \;=\; (v_y)_x - (v_x)_y \;=\; 0; \qquad v_{xx} + v_{yy} \;=\; -(u_x)_y + (u_y)_x \;=\; 0.$$

Thus the real and imaginary parts of a holomorphic function are holomorphic. This proves the first half of the following proposition.

Proposition 1.9.1. *Suppose that U is a simply connected domain in $\mathbb{C}$. If $f : U \to \mathbb{C}$ is holomorphic, then its real part u is harmonic. Conversely, if U is simply connected* *and a C^2 function $u : U \to \mathbb{R}$ is harmonic, then u is the real part of a holomorphic function $f : U \to \mathbb{C}$.*

Proof: It is enough to prove the second statement for a disk D whose closure lies in U; then the result follows by analytic continuation. We may translate and take $D = D_r(0)$. If u is the real part of f, the Cauchy–Riemann equation tell us that the gradient of the imaginary part v is given by $v_x = -u_y$, $v_y = u_x$. Therefore, for $z = x + iy \in D$,

$$\begin{aligned} v(x, y) &= v(0) + \int_0^1 \frac{d}{ds} v(sx, sy)\, ds \\ &= v(0) + \int_0^1 \left\{ -xu_y(sx, sy) + yu(sx, sy) \right\} ds. \end{aligned} \tag{1.9.2}$$

Conversely, choose $v(0)$ arbitrarily and define v by (1.9.2). The assumption that u is harmonic implies that u and v, so defined, satisfy the Cauchy–Riemann equations. Therefore $f = u + iv$ is holomorphic. □

The function v is called a *harmonic conjugate* of u, and is often denoted u^*. It is unique up to an additive constant.

Corollary 1.9.2. *A harmonic function u is infinitely differentiable, and its Taylor series sums to u in any disk that is contained in the domain of u.*

Corollary 1.9.3. *If u is harmonic and g is holomorphic, then $u \circ g$ is harmonic where it is defined. In particular, dilations $u_{(r)}(x, y) = u(rx, ry)$ are harmonic.*

Proof: Locally u is the real part of f, so $u \circ g$ is the real part of $f \circ g$. □

Remarks and further reading

Most undergraduate texbooks on complex analysis cover the basic complex analysis in this chapter. Three classic complex analysis texts—Ahlfors [6], Hille [107], and Titchmarsh [206]—cover, in addition, several of the topics in later chapters: For a discussion of the development of the subject through the work of Cauchy, Riemann, Weierstrass, and others, see Neuenshwander [153].

The special issue of the Journal *Primus*, vol. 27, issue 8-9 (2017) on "Revitalizing Complex Analysis" contains a number of papers that explore topics in this chapter and their applications.

Chapter 2
Further preliminaries

This chapter covers additional material that is used in more than one subsequent chapter. The various sections here are meant to be read or consulted as needed for later chapters, so that those chapters or sequence of chapters can be read independently.

Four fundamental domains in complex function theory are: the complex plane $\mathbb{C}$ itself; the *unit disk* $\mathbb{D}$,

$$\mathbb{D} \;=\; \{z \in \mathbb{C} \,:\, |z| = 1\};$$

the *upper half-plane* $\mathbb{H}$,

$$\mathbb{H} \;=\; \{z \in \mathbb{C} \,:\, \operatorname{Im} z > 0\};$$

and the *Riemann sphere* $\mathbb{S}$:

$$\mathbb{S} \;=\; \mathbb{C} \cup \{\infty\}.$$

The Riemann sphere is given a complex structure by taking a base of neighborhoods of ∞ to be the sets

$$\{z \in C \,:\, |z| > R \geq 0\}.$$

A function f is said to be holomorphic at ∞ if it is holomorphic in a neighborhood of ∞ and $g(z) = f(1/z)$ has a removable singularity at 0; equivalently, f is holomorphic and bounded in some neighborhood of ∞. Poles and essential singularities at ∞ are defined similarly.

It is also useful to have a special notation for the *unit circle* $\mathbb{T}$, the boundary of the unit disk:

$$\mathbb{T} \;=\; \partial\mathbb{D} \;=\; \{\zeta \in \mathbb{C} \,:\, |z| = 1\}.$$

Sections 2.1 and 2.2 cover the automorphisms (bijective holomorphic self-maps) of $\mathbb{C}$, $\mathbb{D}$, $\mathbb{H}$, and $\mathbb{S}$, and the geometries associated to these domains.

Section 2.3 introduces normal families and theorems of Ascoli–Arzelà and Montel. In Section 2.4 this material is applied to the proof of the Riemann mapping theorem.

The triply-punctured sphere and theorems of Picard and of Montel are introduced in Section 2.5. Carathédory's extension theorem is proved in Section 2.6.

R. Beals and R. S. C. Wong, *More Explorations in Complex Functions*, Graduate Texts
in Mathematics 298, https://doi.org/10.1007/978-3-031-28288-1_2

Sections 2.7 and 2.8 cover basic facts about Hilbert spaces and L^p spaces and measure. Convolution, approximation, and weak solutions are covered in Section 2.9. The gamma function and some of its properties are introduced in Section 2.10.

2.1 Linear fractional transformations

A *linear fractional transformation*, or *Möbius transformation*, is a function of the form

$$f(z) = \frac{az+b}{cz+d}, \qquad ad - bc \neq 0. \tag{2.1.1}$$

Computing the composition of two such functions shows that the group of such transformations is a homomorphic image of the group $GL(2, \mathbb{C})$ of invertible 2×2 complex matrices:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \to f_A, \qquad f_A(z) = \frac{az+b}{cz+d}.$$

Note that if α is a non-zero constant, then αA and A induce the same mapping. In particular α can be chosen so that αA has determinant 1.

Proposition 2.1.1. *(a) Each linear fractional transformation is a bijective map from the Riemann sphere $\mathbb{S}$ to itself.*
(b) Each linear fractional transformation f has either one or two fixed points, i.e. solutions of $f(z) = z$, $z \in \mathbb{S}$.
(c) Given any two ordered triples of distinct points (z_1, z_2, z_3) and (w_1, w_2, w_3), there is a unique linear fractional transformation f such that $f(z_j) = w_j$, $j = 1, 2, 3$.
(d) Each linear fractional transformation is conformal: if two smooth curves in $\mathbb{S}$ meet at an angle, then the images under f meet at the same angle.
(e) Each linear fractional transformation f has the property that if Γ is either a straight line or a circle in $\mathbb{C}$, then $f(\Gamma) \cap \mathbb{C}$ is either a straight line or a circle.

Proof: Parts (a) and (b) are easily checked. For (c), it is enough to show that given (z_1, z_2, z_3), there is a unique linear fractional transformation f such that $f(z_1) = 0$, $f(z_2) = 1$, and $f(z_3) = \infty$. (See below).

Part (d) is clear geometrically. Given z_0 in the domain of f, let us define $g(w) = f(z_0 + w) - f(z_0)$, so that $g(0) = 0$. In the limit as $w \to 0$, g is multiplication by $f'(z_0) \neq 0$. So letting $f'(z_0) = re^{i\theta}$, in the limit g dilates by a factor r and rotates by θ, both of which actions preserve angles. (If z_0 or $f(z_0)$ equals ∞, this argument can be modified accordingly.)

In part (e), note that the statement is *not* that lines are taken to lines and circles are taken to circles. One way to verify the result is to note that any linear fractional transformation is a product of linear fractional transformations of the form $f(z) = az + b$ and, if necessary, the inversion $R(z) = 1/z$. The first type maps lines to lines and circles to circles, so the problem reduces to the study of R. Taking into account rotations, it is enough to consider $R(\Gamma)$ when Γ is a vertical line or a circle with center on the real axis. In each case, consideration of where $R(\Gamma)$ intersects $\mathbb{R} \cup \{\infty\}$ will identify the nature of the image, and aid in verifying that it is indeed a straight line or circle. Details are left to the reader. □

Three distinct points in $\mathbb{S}$ determine a unique line or circle in the plane; if one of the points is the point at ∞, then they determine a line. Suppose that the three points are $(z_1, z_2, z_3) \in \mathbb{C}$. The unique linear fractional transformation that takes these points, in order, to $(0, 1, \infty)$ is

$$f(z) \;=\; \frac{z - z_1}{z - z_3} \cdot \frac{z_2 - z_3}{z_2 - z_1}.$$

The expression on the right is called the *cross ratio* of the quadruple (z, z_1, z_2, z_3). It is commonly denoted $[z, z_1, z_2, z_3]$. The following is a consequence of Proposition 2.1.1.

Corollary 2.1.2. *(a) The cross ratio is invariant under linear fractional transformations: given four distinct points z_0,z_1,z_2,z_3 in $\mathbb{C}$ and a linear fractional transformation g,*

$$[g(z_0), g(z_1), g(z_2), g(z_3)] \;=\; [z_0, z_1, z_2, z_3].$$

(b) A point $z \in \mathbb{C}$ lies on the line or circle determined by three distinct points z_1, z_2, z_3 in $\mathbb{S}$ if and only if the cross ratio $[z, z_1, z_2, z_3]$ is real.

By an *automorphism* of a domain Ω with a complex structure, we mean a bijective holomorphic map of the domain to itself. The set of such mappings is a group $\mathrm{Aut}(\Omega)$ under composition.

Proposition 2.1.3. *The automorphism group of $\mathbb{C}$ is the set of linear fractional transformations of the form $f(z) = az + b$, $a \neq 0$.*

Proof: Obviously any such linear fractional transformation is an automorphism of $\mathbb{C}$. Conversely, suppose that f is an automorphism of $\mathbb{C}$. Then f has an isolated singularity at ∞. Since f is single-valued, this singularity is not a multiple pole nor (by Casorati–Weierstrass) an essential singularity. Moreover f is not bounded, so it is not bounded in a neighborhood of ∞. Therefore the singularity is a simple pole with residue $a \neq 0$. Then $f(z) - az$ is entire and bounded near ∞, hence constant. □

Proposition 2.1.4. *The automorphism group of $\mathbb{S}$ is the set of all linear fractional transformations.*

Proof: Any linear fractional transformation f is an automorphism of $\mathbb{S}$. Bijectivity is easy to check, and holomorphy needs only to be checked at $z = \infty$ and at $f^{-1}(\infty)$. Conversely, if g is an automorphism of $\mathbb{S}$, compose with a linear fractional transformation f such that $f(g(\infty)) = \infty$. Then $h = f \circ g$ restricted to $\mathbb{C}$ is an automorphism of $\mathbb{C}$, hence a linear fractional transformation, so $g = f^{-1} \circ h$ is also a linear fractional transformation. □

Lemma 2.1.5. (Schwarz's lemma) *If f is an automorphism of $\mathbb{D}$ such that $f(0) = 0$, then f is a rotation: $f(z) = \omega z$, $|\omega| = 1$.*

Proof: Let $g(z) = f(z)/z$. Then $g : \mathbb{D} \to \mathbb{D}$ is holomorphic, so the maximum principle implies $|f(z)/z| \leq 1/r$ for $|z| \leq r < 1$. Taking the limit as $r \to 1$ and noting that the same argument applies to f^{-1}, we find that $|f(z)| = |z|$. By the strong maximum principle, $f(z)/z$ is constant. □

Proposition 2.1.6. *The automorphism group of the unit disk $\mathbb{D}$ consists of linear fractional transformations of the form*

$$f(z) = \omega \frac{z-a}{1-\bar{a}z}, \qquad a \in \mathbb{D}, \ |\omega| = 1. \tag{2.1.2}$$

Proof: If f has the form (2.1.2), then $|z| = 1$ implies $|f(z)| = 1$, so f maps each component of the complement of the unit circle onto such a component. Since $f(a) = 0$, it follows that $f(\mathbb{D}) = \mathbb{D}$.

Conversely, suppose that g is an automorphism of $\mathbb{D}$. Let f be given by (2.1.2) with $a = g(0)$ and $\omega = 1$. Then $h = f \circ g$ is an automorphism with $h(0) = 0$. By Lemma 2.1.5, h is constant, so $g = f^{-1} \circ h$ is a linear fractional transformation. □

The linear fractional transformations

$$C(z) = \frac{z-i}{z+i}, \qquad C^{-1}(w) = i\,\frac{1+w}{1-w} \tag{2.1.3}$$

are the *Cayley transform* and its inverse. It is easily seen that C maps the real line to the unit circle. Since $C(i) = 0$, C maps the upper half-plane to the unit disk and the lower half-plane to $\{z : |z| > 1\}$.

Proposition 2.1.7. *The automorphism group of the upper half-plane $\mathbb{H}$ consists of the linear fractional transformations that have real coefficients and positive determinants.*

Proof: If f is an automorphism of $\mathbb{H}$, then $g = C \circ f \circ C^{-1}$ is an automorphism of $\mathbb{D}$. Therefore g is a linear fractional transformation and so is $f = C^{-1} \circ g \circ C$. If f has the form (2.1.1), we may assume that $ad - bc$ is real, and then computing $\operatorname{Im} f(i)$ shows that $ad - bc$ must be positive. Now f maps $\mathbb{R} \cup \{\infty\}$ to itself. Checking $f(\infty)$, $f(0)$, and $f(1)$ shows that each coefficient is real. □

2.2 Geometries

Each of the domains $\mathbb{C}$, $\mathbb{S}$, $\mathbb{D}$, and $\mathbb{H}$ carries one or more natural metrics and geometries.

A. Geometries on $\mathbb{C}$

There are two commonly used geometries and three commonly used metrics on $\mathbb{C}$. The first is the euclidean geometry of $\mathbb{C}$ as identified with $\mathbb{R}^2$, with metric $|z-w|$. The second geometry and a related metric come from the identification of $\mathbb{C}$ as a subset of the Riemann sphere via *stereographic projection*. This standard pictorial representation is obtained by considering $\mathbb{C}$ as the x, y plane of the three-dimensional space

$$\mathbb{R}^3 \;=\; \mathbb{C}\times\mathbb{R} \;=\; \{(w,t) : w\in\mathbb{C},\ t\in\mathbb{R}\},$$

and relating it to the 2-sphere of radius 1 centered at the origin:

$$S \;=\; \{(w,t) : |w|^2+t^2 \;=\; 1\}.$$

A point $\omega=(w,t)$ on S is mapped to a point $z=\pi(\omega)$ in $\mathbb{C}$ by following the line from the north pole $N=(0,1)\in S$ through (w,t) to its intersection $(z,0)$ with $\mathbb{C}\times\{0\}$; see Figure 2.1.

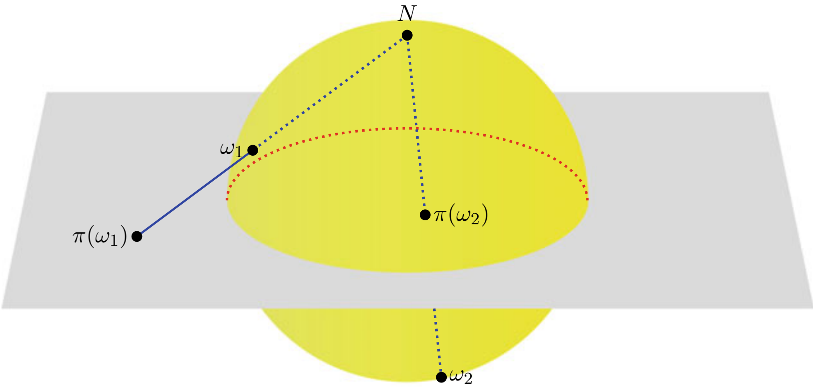

Fig. 2.1 Stereographic projection.

The line determined by $(0,1)$ and $\omega=(w,t)$ is the set of points

$$(1-\lambda)(0,1)+\lambda(w,t), \quad \lambda\in\mathbb{R}.$$

Thus for $t\neq 1$,

$$\pi(w,t) \;=\; \frac{w}{1-t}. \tag{2.2.1}$$

It follows that

$$|\pi(w,t)|^2 \;=\; \frac{|w|^2}{(1-t)^2} \;=\; \frac{1-t^2}{(1-t)^2} \;=\; \frac{1+t}{1-t}.$$

Therefore

$$t \;=\; \frac{|z|^2-1}{|z|^2+1}; \qquad 1-t \;=\; \frac{2}{|z|^2+1},$$

and

$$\pi^{-1}(z) \;=\; \left(\frac{2z}{|z|^2+1}, \frac{|z|^2-1}{|z|^2+1}\right). \tag{2.2.2}$$

As $\omega \in S$ approaches the north pole, $\pi(\omega)$ approaches ∞, so we let $\pi(0,1) = \infty$.

We define a second distance function in the plane by using the euclidean distance of the pull-back to $\mathbb{C}$,

$$d(z_1, z_2) \;=\; \frac{1}{2}\left|\pi^{-1}(z_1) - \pi^{-1}(z_2)\right|, \tag{2.2.3}$$

The computation is made simpler by noting that for (w_j, t_j) in the sphere,

$$|(w_1, t_1) - (w_2, t_2)|^2 \;=\; 2 - \mathrm{Re}\,(\bar{w}_1 w_2 + t_1 t_2). \tag{2.2.4}$$

Using this and (2.2.2) we find that

$$d(z_1, z_2) \;=\; \frac{|z_1 - z_2|}{\sqrt{|z_1|^2+1}\sqrt{|z_2|^2+1}}. \tag{2.2.5}$$

Taking the limit as $z_2 \to \infty$ gives

$$d(z, \infty) \;=\; \frac{1}{\sqrt{1+|z|^2}}. \tag{2.2.6}$$

B. Hyperbolic geometry in the disk

The fundamental idea for geometry in $\mathbb{D}$ is that the metric $\rho_{\mathbb{D}}$ should be invariant under $\mathrm{Aut}(\mathbb{D})$ and that the diameters should be geodesics. The geometry is then uniquely determined by setting a scale factor. Different sources use different scale factors. In [22] we followed [107] and chose the scaling so that in the limit at the origin, the metric is euclidean:

$$\lim_{\varepsilon \to 0} \frac{\rho_{\mathbb{D}}(\varepsilon, 0)}{|\varepsilon|} \;=\; 1. \tag{2.2.7}$$

This choice also fits nicely with the Teichmüller metric in Section 9.4.

Since the diameter $(-1, 1)$ is to be a geodesic, we want additivity:

$$\rho_{\mathbb{D}}(r,t) \;=\; \rho_{\mathbb{D}}(r,s) + \rho_{\mathbb{D}}(s,t) \qquad \text{if } -1 < r < s < t < 1. \tag{2.2.8}$$

By a slight abuse of notation, write $\rho_{\mathbb{D}}(z, 0) = \rho_{\mathbb{D}}(z)$. Note that invariance implies that $\rho_{\mathbb{D}}(z) = \rho_{\mathbb{D}}(|z|)$. The automorphism $f(z) = (z - r)/1 - rz)$ and the invariance assumption reduce (2.2.8) to the case for the triple $0 < r < r + \delta$, thus to $(f(0), 0, f(r+\delta))$. Letting $\delta \to 0$ and taking into account (2.2.7), we find that

$$\rho'_{\mathbb{D}}(r) \;=\; \frac{1}{1-r^2} \;=\; \frac{1}{2}\left(\frac{1}{1+r} + \frac{1}{1-r}\right),$$

so

$$\rho_{\mathbb{D}}(r, 0) \;=\; \rho_{\mathbb{D}}(r) \;=\; \frac{1}{2}\log\frac{1+r}{1-r}, \qquad -1 < r < 1. \tag{2.2.9}$$

In general, suppose that z_1, z_2 are distinct points of $\mathbb{D}$. There is a (unique) automorphism that takes z_1 to 0 and z_2 to an element of $(0, 1)$. Invariance then gives the general formula for the *hyperbolic metric*, or *Poincaré metric*

$$\rho_{\mathbb{D}}(z_1, z_2) \;=\; \frac{1}{2}\log\frac{|1-\bar{z}_1 z_2| + |z_1 - z_2|}{|1-\bar{z}_1 z_2| - |z_1 - z_2|} \;=\; \tanh^{-1}\frac{|z_1 - z_2|}{|1-\bar{z}_1 z_2|}. \tag{2.2.10}$$

The set of geodesics is, by construction, invariant under $\mathrm{Aut}(\mathbb{D})$. We know that the image of a diameter must be an arc of a circle, and since the automorphisms are conformal, such an arc must meet the boundary $\mathbb{T}$ in two right angles. Conversely, given such a circular arc, there is an element of $\mathrm{Aut}(\mathbb{D})$ that moves the endpoints to $-1, 1$. Therefore

Proposition 2.2.1. *The geodesics for the hyperbolic metric $\rho_{\mathbb{D}}$ in $\mathbb{D}$ are the diameters of $\mathbb{D}$ and the circular arcs that meet the boundary in two right angles.*

Two infinitesimal versions are the line element

$$ds \;=\; (ds)_{\mathbb{D}} = \frac{|dz|}{1-|z|^2} \tag{2.2.11}$$

and the Riemann metric

$$ds^2 \;=\; \frac{dx^2 + dy^2}{(1-r^2)^2}. \tag{2.2.12}$$

The *Poincaré density* $\eta_{\mathbb{D}}(z)$ is, by definition, the limiting ratio between the Poincaré distance and the euclidean distance at $z\mathbb{D}$:

$$\eta_{\mathbb{D}}(z) \;=\; \frac{|dz|}{ds} \;=\; \frac{1}{1-r^2}. \tag{2.2.13}$$

C. Hyperbolic geometry in the upper half-plane

The fundamental idea is the same as in the case of the disk. The metric $\rho_{\mathbb{H}}$, known as the *Poincaré metric* is invariant under $\mathrm{Aut}(\mathbb{H})$. This can be constructed by the same method used for $\rho_{\mathbb{D}}$, starting with the positive imaginary axis and deriving a differential equation that determines $\rho_{\mathbb{H}}$ on that half-line, up to a scale factor. Instead, we shall take advantage of the Cayley transform $C : \mathbb{H} \to \mathbb{D}$ to transplant $\rho_{\mathbb{D}}$:

$$\rho_{\mathbb{H}}(z_1, z_2) = \rho_{\mathbb{D}}(C(z_1), C(z_2)) \;=\; \rho_{\mathbb{D}}\left(\frac{z_1 - i}{z_1 + i}, \frac{z_2 - i}{z_2 + i}\right)$$
$$= \frac{1}{2} \log \frac{|\bar{z}_1 - z_2| + |z_1 - z_2|}{|\bar{z}_1 - z_2| - |z_1 - z_2|}. \tag{2.2.14}$$

In particular

$$\rho_{\mathbb{H}}(it, is) \;=\; \frac{1}{2} \log \frac{t}{s} \qquad \text{if } 0 < s < t.$$

The infinitesimal version is

$$(ds)_{\mathbb{H}} \;=\; \frac{|dz|}{2 \operatorname{Im} z}; \qquad ds^2 \;=\; \frac{dx^2 + dy^2}{4y^2}. \tag{2.2.15}$$

Thus the Poincaré density $\eta_{\mathbb{H}}$ for $\rho_{\mathbb{H}}$ is

$$\eta_{\mathbb{H}}(z) \;=\; \frac{(ds)_{\mathbb{H}}}{|dz|} \;=\; \frac{1}{2 \operatorname{Im} z}. \tag{2.2.16}$$

It follows from this construction that the geodesics, as images under C^{-1} of geodesics in $\mathbb{D}$, are lines and half-circles that meet the real axis at right angles.

Remark. There seem to be two common normalizations for $\rho_{\mathbb{H}}$. The other is twice this one. We have chosen to use the normalization (2.2) since this is what is used in our sources for Chapter 9.

This construction can be carried over to any conformal image of $\mathbb{D}$. If $f : \Omega \to \mathbb{D}$ is a conformal map, then we set

$$\rho_{\Omega}(z_1, z_2) \;=\; \rho_{\mathbb{D}}\left(f(z_1), f(z_2)\right).$$

Thus the infinitesimal distance ds_{Ω} is

$$ds_{\Omega}(z) \;=\; \lim_{\varepsilon \to 0} \frac{\rho_{\mathbb{D}}((f(z + \varepsilon), f(z))}{\varepsilon} \;=\; \frac{|f'(z)\, dz|}{1 - |f(z)|^2},$$

and the associated *metric density* is

$$\eta_{\Omega}(z) \;=\; \frac{|f'(z)|}{1 - |f(z)|^2}. \tag{2.2.17}$$

In Section 9.4 we will need the following two results.

Proposition 2.2.2. *Metric density decreases with respect to set inclusion: if Ω_1 and Ω_2 are simply connected domains with $\Omega_1 \subset \Omega_2 \subset \mathbb{C}$ and $\Omega_1 \neq \Omega_2 \neq \mathbb{C}$, then*

$$\eta_{\Omega_2}(z) \;<\; \eta_{\Omega_1}(z), \qquad z \in \Omega_1. \tag{2.2.18}$$

Proof. Given $z \in \Omega_1$, let f_j be a conformal map of Ω_j onto $\mathbb{D}$ that takes z to 0, such that $f_j'(0) > 0$. Then $f = f_2 \circ f_1^{-1}$ is a conformal map of $\mathbb{D}$ to a proper subset of $\mathbb{D}$, and $f(0) = 0$, $f'(0) > 0$. By Schwarz's lemma, $f'(0) < 1$. Near z, $f_2 = f \circ f_1$, so $|f_2'(z)| = |f'(0)f_1(0)| < |f_1'(0)|$. □

Proposition 2.2.3. *The metric density ρ_Ω satisfies*

$$\frac{1}{4\,d(z,\partial\Omega)} \le \eta_\Omega(z) \le \frac{1}{d(z,\partial\Omega)}, \tag{2.2.19}$$

where $d(z,\partial\Omega)$ is the (euclidean) distance to the boundary:

$$d(z,\partial\Omega) = \inf_{\zeta\in\partial\Omega} |z-\zeta|.$$

Proof. Given $z \in \Omega$, let ϕ be the conformal map of $\mathbb{D}$ onto Ω such that $\phi(0) = z$, $\phi'(0) > 0$. Let Ω_1 be the disk of radius $r = d(z,\partial\Omega)$ centered at z. Then $f(w) = \phi^{-1}(rw + z)$ maps $\mathbb{D}$ into $\mathbb{D}$, so

$$1 \ge |f'(0)| = r[\phi^{-1}]'(z) = \frac{d(z,\partial\Omega)}{\phi'(0)} = d(z,\partial\Omega)\cdot\eta_\Omega,$$

which proves the second inequality in (2.2.19).

Now $\psi(w) = [\phi(w) - z]/\phi'(0)]$ is a normalized map of $\mathbb{D}$ into $\mathbb{C}$. The Koebe one-quarter theorem, Theorem 4.1.4, says that $\psi(\mathbb{D})$ contains $D_{1/4}(0)$. The largest disk about 0 has radius $r/\phi'(0)$, so

$$\frac{1}{4} \le \frac{r}{\phi'(0)} = \frac{d(z,\partial\Omega)}{\eta_\Omega},$$

which proves the first inequality in (2.2.19). □

2.3 Normal families

A collection $\mathscr{F}$ of real or complex-valued functions defined on a domain Ω in $\mathbb{C}$ is said to be a *normal family* if each sequence $\{f_n\}$ in $\mathscr{F}$ contains a subsequence that converges uniformly on each compact subset of Ω. A normal family $\mathscr{F}$ is said to be *complete* if the limit of any such convergent sequence belongs to F.

The standard criterion for a normal family is the following, specialized to domains in $\mathbb{C}$. A family $\mathscr{F}$ of real or complex-valued functions defined on a domain $\Omega \subset \mathbb{C}$ is said to be *bounded on compact subsets* if for each compact subset $C \subset \Omega$, the supremum of $|f(z)|$, $f \in \mathscr{F}$, $z \in C$, is finite. The family $\mathscr{F}$ is said to be *equicontinuous on compact subsets* if for each compact subset $C \subset \Omega$ and each $\varepsilon > 0$, there is a $\delta > 0$ such that for each $f \in \mathscr{F}$ and points $z, w \in \mathbb{C}$, if $|z-w| < \delta$ then $|f(z) - f(w)| < \varepsilon$.

Theorem 2.3.1. *(Ascoli–Arzelà) Suppose that $\mathscr{F}$ is a family of real or complex-valued functions defined on a domain $\Omega \subset \mathbb{C}$, that is bounded and equicontinuous on compact subsets. Then $\mathscr{F}$ is a normal family.*

Proof: Let $\{z_m\}$ be a countable dense subset of Ω. The boundedness assumption implies, in particular, that for any sequence $\{f_n\}$ in $\mathscr{F}$, there is a subsequence $\{f_{1,n}\}$ that converges at z_1. Some subsequence of $\{f_{1,n}\}$ converges at z_2; denote this subsequence by $\{f_{2,n}\}$, and find a subsequence of $\{f_{2,n}\}$ that converges at z_3. Having found nested subsequences $\{f_{m+1,n}\} \subset \{f_{m,n}\}$ for each m, we note that the sequence $\{f_{n,n}\}$ converges at each z_n.

Suppose now that C is a compact subset of $\mathbb{C}$ that is the closure of an open set. Then the z_n that lie in C are dense in C. It is an exercise to show, using the assumptions of equicontinuity and boundedness, that $\{f_{n,n}\}$ converges uniformly on C to a bounded continuous function f. □

Corollary 2.3.2. *(**Montel's theorem**) Suppose that $\mathscr{F}$ is a family of holomorphic functions on a domain $\Omega \subset \mathbb{C}$, that is bounded on compact subsets. Then $\mathscr{F}$ is a normal family.*

Proof: It is enough to prove equicontinuity on compact subsets, and for that it is enough to prove that the family of derivatives

$$\mathscr{F}' = \{f' : f \in \mathscr{F}\}$$

is bounded on compact sets. To see that this is true, given $z_0 \in \Omega$, choose $r > 0$ such that $\{z : |z - z_0| \le r\}$ is contained in Ω and use the estimate (1.3.2) to show that the derivative is bounded on each disk $D_s(z_0)$, $0 < s < r$. □

Anticipating a bit, we note here that analogous results are true for harmonic functions.

Corollary 2.3.3. *Suppose that $\mathscr{F}$ is a family of harmonic functions on a domain $\Omega \subset \mathbb{C}$, and bounded on compact subsets. Then $\mathscr{F}$ is a normal family.*

Proof: The proof is essentially the same as for Corollary 2.3.2, with the (scaled) version of the Poisson integral formula (5.1.6) in place of the Cauchy integral formula. □.

It is important to know something about the limit functions in the holomorphic or harmonic case.

Proposition 2.3.4. *If $\{f_m\}$ is a uniformly convergent sequence of holomorphic (resp. harmonic) functions on a domain $\Omega \subset \mathbb{C}$, then the limit f is holomorphic (resp. harmonic).*

Proof: Given $z_0 \in \Omega$, choose r as in the proof of Corollary 2.3.3. The f_n are eventually defined on $D_r(z_0)$. In the holomorphic case, the Cauchy integral formula for $f_n(z)$, $z \in D_r(z_0)$, gives a formula for $f(z)$. The same argument, using the Poisson formula, applies to the harmonic case. □

Corollary 2.3.5. *(**Vitali's theorem**) If the sequence $\{f_n\}$ of holomorphic functions on a domain $\Omega \subset \mathbb{C}$ is uniformly bounded on compact subsets and converges at each point of a set with an accumulation point in $\mathbb{S}$, then it converges uniformly on compact sets in Ω.*

Proof. The sequence is a normal family. The limit of any convergent subsequence is determined uniquely by its values $\{a_n\}$ on a set with an accumulation point, since all derivatives at that point are uniquely determined by the $\{a_n\}$. Therefore the sequence itself converges. □

Proposition 2.3.6. *If $\{f_m\}$ is a uniformly convergent sequence of injective holomorphic functions on a domain $\Omega \subset \mathbb{C}$, then the limit f is either constant or injective.*

Proof: Suppose that f is not constant. Given $z_0 \in \Omega$, choose r as before, but in such a way that $|f(z) - f(z_0)| \geq \delta > 0$ for $|z - z_0| = r$. By Rouché's theorem, 1.4.7, eventually f_n has the same number of values $f(z_0)$ in $D_r(z_0)$ as f does. By assumption each f_n is injective, so f must also be injective. □

The special properties of holomorphic and harmonic functions allow a useful generalization of these criteria. The proofs make new use of the same ideas.

Proposition 2.3.7. *Suppose that $\mathscr{F}$ is a family of holomorphic or harmonic functions on a domain $\Omega \subset \mathbb{C}$, and suppose that for any compact $C \subset \Omega$,*

$$\sup_{f \in \mathscr{F}} \iint_C |f|\, dx\, dy \; < \; \infty. \tag{2.3.1}$$

Then $\mathscr{F}$ is a normal family.

Proof: Consider the holomorphic case. Let z_0 be any point of Ω and again let $r > 0$ be such that the closure of $D_r(z_0)$ is contained in Ω. Given any f holomorphic on $D_r(z_0)$ and any z in $D_{2r/3}(z_0)$, we can write a version of the Cauchy integral formally by smearing the original formula over the circles of radius $2r/3$ to r:

$$\begin{aligned} f(z) &= \frac{3}{r}\int_{2r/3}^{r}\left(\frac{1}{2\pi i}\int_{|\zeta - z_0| = s}\frac{f(\zeta)\,d\zeta}{\zeta - z}\right) ds \\ &= \frac{1}{2\pi i}\iint_{2r/3 < |z - z_0| < r}\frac{f(\zeta)\,dx\,dy}{\zeta - z}. \end{aligned}$$

Differentiating with respect to z, and using the assumption (2.3.1) gives us uniform estimates for f' on $D_{r/3}(z_0)$, $f \in \mathscr{F}$. The same idea, using the Poisson formula, applies in the harmonic case. □

Remark. This section exemplifies different mathematical traditions. Texts on complex analysis speak of normal families and Montel, and give the above criterion and proof, without mentioning Ascoli or Arzelà. Functional analysis and real analysis texts often contain the Ascoli–Arzelà theorem, by name, but do not mention Montel or give a name to the "normal family" concept.

For a pictorial proof of the Ascoli–Arzelà theorem, see [22], Section 5.3.

2.4 Conformal equivalence and the Riemann mapping theorem

A *conformal mapping* of one complex domain to another is a bijective C^1 function that preserves, at each point, the size and the orientation of the angle between any two C^1 curves passing through that point. Some calculation shows that the necessary and sufficient condition for this, pointwise, is that the Cauchy–Riemann equations hold, and that the derivative with respect to z be non-zero. Thus a conformal mapping is a bijective holomorphic function.

Given two domains in $\mathbb{C}$ or $\mathbb{S}$, a natural question is whether they are *conformally equivalent*: does there exist a bijective holomorphic map from one onto the other? Since linear fractional transformations are conformal maps, we know that a half-plane and a disk, are conformally equivalent. On the other hand, by Liouville's theorem, a holomorphic map from $\mathbb{C}$ to a disk in $\mathbb{C}$ must be constant, so $\mathbb{C}$ and $\mathbb{D}$ are not conformally equivalent. Moreover it is easily seen that a bijective holomorphic image of a simply connected domain is simply connected. Therefore the plane minus a single point is also not conformally equivalent to the unit disk.

The original version of the following theorem was formulated by Riemann for domains with some assumptions about the boundary. The definitive result is the following, due to Koebe [122].

Theorem 2.4.1. (Riemann mapping theorem) *If Ω is an open, simply connected, proper subset of the plane, then there is a bijective holomorphic map f that maps Ω onto the unit disk $\mathbb{D}$.*

Given a point $z_0 \in \Omega$, we may specify $f(z_0) = 0$, $f'(z_0) > 0$. These conditions determine f uniquely.

Remark. Given that $\Omega \subset \mathbb{C}$ is a simply connected domain, the assumption that it is a proper subset is equivalent to the assumption that the boundary $\partial\Omega$ contains at least two points. Therefore the theorem is often stated with this as the extra condition on Ω.

The proof involves a number of steps. The first step is a reduction to the case of bounded Ω. Choose a not in Ω. Then $z - a$ is never zero on the simply connected domain Ω, so we may choose a branch of $\sqrt{z-a}$ that is holomorphic on Ω; see Section 1.5. This branch maps Ω bijectively onto a domain Ω_1. Choose a point $b \in \Omega_1$. For some $\varepsilon > 0$, the disk $\{z : |z - b| < \varepsilon\}$ is contained in Ω_1. If z is in Ω_1

then $-z$ is not, so Ω_1 lies outside the disk $\{z : |z+b| < \varepsilon\}$. The map $z \to 1/(z+b)$ takes Ω_1 bijectively onto a bounded domain Ω_2. Thus we may replace Ω_1 by Ω_2, and assume that Ω itself is bounded.

For the next step, choose a point $z_0 \in \Omega$ and let $\mathscr{F}$ be the family of bijective holomorphic maps f from Ω into the unit disk $\mathbb{D}$ such that $f(z_0) = 0$ and $f'(z_0) > 0$. Note that this family is not empty: $f(z) = \varepsilon(z - z_0)$ will belong to $\mathscr{F}$ if $\varepsilon > 0$ is small enough. Note also that

$$\sup_{f\in\mathscr{F}} f'(z_0) \le \frac{1}{r}$$

if r is such that the closure of $D_r(z_0)$ is contained in Ω. It follows from Corollary 2.3.3 that $\mathscr{F}$ is a normal family. Therefore $\mathscr{F}$ contains an element f such that

$$f'(z_0) \;=\; \sup_{g\in\mathscr{F}} g'(z_0).$$

By Proposition 2.3.6, f is injective. We need to show that f is surjective. Suppose f omits a point $a \in \mathbb{D}$, and suppose first, for simplicity, that $a > 0$. A branch of the square root can be chosen so that

$$g(z) \;=\; \sqrt{\frac{z-a}{az-1}}$$

is holomorphic on $f(\Omega) \subset \mathbb{D}$. The linear fractional transformation under the radical sign maps $\mathbb{D}$ to $\mathbb{D}$, so the composition $g \circ f$ maps Ω into $\mathbb{D}$. Note that $g(0) = \sqrt{a}$. Let

$$h(z) \;=\; \frac{z-\sqrt{a}}{\sqrt{a}z-1}, \tag{2.4.1}$$

and let $f_1 = h \circ g \circ f$. Then f_1 is bijective from Ω into $\mathbb{D}$, and

$$f_1'(z_0) \;=\; h'(\sqrt{a})\, g'(0)\, f'(z_0). \tag{2.4.2}$$

But

$$g'(z) \;=\; \frac{1}{2g(z)}\,\frac{(az-1)-a(z-a)}{(az-1)^2} \;=\; \frac{1}{2g(z)}\,\frac{a^2-1}{(az-1)^2},$$

and

$$h'(z) \;=\; \frac{(\sqrt{a}z-1)-\sqrt{a}(z-\sqrt{a})}{(\sqrt{a}z-1)^2} \;=\; \frac{a-1}{(\sqrt{a}z-1)^2},$$

so

$$g'(0) \;=\; \frac{a^2-1}{2\sqrt{a}}, \qquad h'(\sqrt{a}) \;=\; \frac{1}{a-1}$$

and

$$f_1'(z_0) \;=\; \frac{a+1}{2\sqrt{a}}\, f'(z_0).$$

But since $0 < a < 1$ we have $a + 1 - 2\sqrt{a} = (1 - \sqrt{a})^2 > 0$, so $f_1'(z_0) > f'(z_0)$, contradicting the assumption that $f'(z_0)$ is maximal.

The preceding argument assumed that $f(\Omega)$ omitted a point $a \in \mathbb{D}$ and that $a > 0$. Otherwise, we may assume that the omitted point has the form ωa, where $|\omega| = 1$ and $a > 0$, and take

$$f_1(z) = \omega h(g(\bar{\omega} f(z)))$$

with g and h defined as before. Again we find that $f_1'(z_0) > f'(z_0)$, a contradiction. This contradiction shows that our function f of (2.4.2) maps Ω onto $\mathbb{D}$.

Finally, uniqueness follows easily from Lemma 2.1.5. □

2.5 The triply-punctured sphere, Montel, and Picard

By the *triply-punctured sphere* we mean the Riemann sphere with the points $0, 1, \infty$ removed. We will denote it by $\mathbb{S} \setminus 3$:

$$\mathbb{S} \setminus 3 = \mathbb{S} \setminus \{0, 1, \infty\} = \mathbb{C} \setminus \{0, 1\}. \tag{2.5.1}$$

The study of $\mathbb{S} \setminus 3$ is closely connected to the *elliptic modular function* λ. Theorem 2.4.1 makes possible a quick conceptual construction of λ. Let Ω be the domain

$$\Omega = \{z : 0 < \operatorname{Re} z < 1,\ |z - \tfrac{1}{2}| > \tfrac{1}{2}\}. \tag{2.5.2}$$

See Figure 2.2.

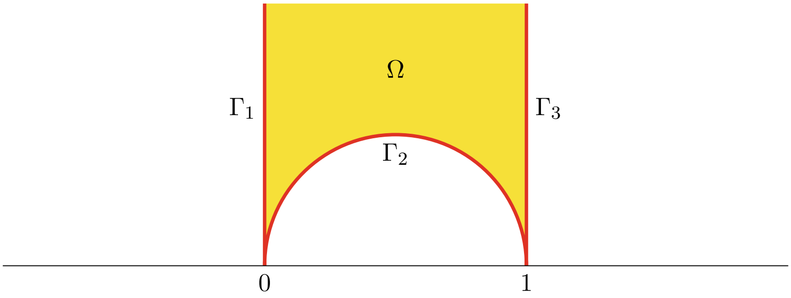

Fig. 2.2 Fundamental domain of λ.

Let Φ be a conformal map of $\mathbb{D}$ onto Ω. Since the boundary of Ω consists of analytic arcs, Theorem 1.7.2, together with some consideration of the images under Φ of circles $\{z : |z| = r\}$ as $r \uparrow 1$, shows that Φ has a continuous extension to the boundary. (Or see Section 2.6.) By composing with an automorphism of $\mathbb{D}$, we may suppose that $\Phi(1) = \infty$, $\Phi(-i) = 0$, and $\phi(1) = \infty$. Let $C : \mathbb{H} \to \mathbb{D}$ be the Cayley transform. Then $\Psi = \Phi \circ C$ is a conformal map of $\mathbb{H}$ onto Ω that extends continuously to map $\mathbb{R}$ onto the boundary of Ω, fixing 0, 1, and ∞:

$$\Psi : \mathbb{H} \to \Omega; \qquad \lim_{z \to w} \Psi(w) \ = \ w, \quad w = 0, 1, \infty. \tag{2.5.3}$$

Let λ be the map

$$\lambda \ = \ \Psi^{-1} : \Omega \to \mathbb{H}. \tag{2.5.4}$$

As we shall see, λ extends to a *covering map* of $\mathbb{H}$ onto $\mathbb{S} \setminus 3$. This means that any point $z \in \mathbb{S} \setminus 3$ has a connected neighborhood U with the property that λ maps each connected component of $\lambda^{-1}(U)$ conformally onto U.

Theorem 2.5.1. *The function λ is the unique conformal map of the domain (2.5.2) onto $\mathbb{H}$ whose extension to the boundary fixes* 0, 1, *and* ∞. *Moreover,* λ *extends to* $\mathbb{H}$ *as a covering map of* $\mathbb{H}$ *onto* $\mathbb{S} \setminus 3$.

Proof. Suppose that $\mu : \Omega \to \mathbb{H}$ is another such map. Then $\mu^{-1} \circ \lambda$ is an automorphism of $\mathbb{H}$ whose extension fixes three distinct points. Therefore $\mu^{-1} \circ \lambda = \mathbf{1}$, the identity map, so $\mu = \lambda$.

Since λ extends to map the semicircle that is the lower boundary Γ_2 onto the real interval $[0, 1]$, it extends across by reflection to the reflection $r(\Omega)$ through the semicircle Γ_2. By following the boundary and using conformality, it is easy to see that the lower boundary of $r(\Omega)$ consists of the semicircles in $\mathbb{H}$ centered at $1/4$ and at $3/4$ and having radius $1/4$. This gives us an extension that is a conformal map

$$\lambda : \Omega \cup \Gamma_2 \cup r(\Omega) \to \mathbb{C} \setminus \{0, 1\} \ = \ \mathbb{S}\backslash 3.$$

This can be extended again across each of its lower bounday arcs, and the process continued; see Figure 2.3.

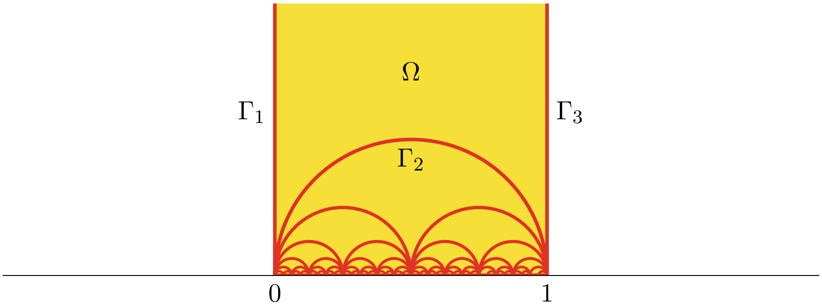

Fig. 2.3 Extension to a half strip.

Each of the bounded, three-sided subdomains in this figure, when reflected across one of its lower boundary arcs, gives an extension that it defined on a similar domain having half the diameter and maps to the opposite half-plane of $\mathbb{C}$. Continuing indefinitely results in a countably-many-to-one locally conformal map of the strip,

$$\lambda : \{z \ : \ \operatorname{Im} z > 0, \ 0 < \operatorname{Re} \ < 1\} \to \mathbb{S}\backslash 3.$$

Finally, continued reflection through the vertical boundaries results in an extension

$$\lambda : \mathbb{H} \to \mathbb{C} \setminus \{0, 1\} \;=\; \mathbb{S}\backslash 3.$$

This is a covering map—tracing through the construction shows that each point of $S\backslash 3$ has a neighborhood U whose inverse image is the disjoint union $\bigcup U_\alpha$ such that $\lambda : U_\alpha \to U$ is conformal. □

For our purpose here it is convenient to replace λ by $\lambda^* = \lambda \circ \mathbb{C}^{-1}$, C the Cayley transform, so that λ^* is a covering map from $\mathbb{D}$ onto $\mathbb{S} \setminus 3$.

Remark. The classical elliptic modular function has (2.5.2) as fundamental domain, but maps $0 \to 1$, $1 \to \infty$, $\infty \to 0$. Therefore it is $h \circ \lambda$, where $h \in \mathrm{Aut}(\mathbb{H})$ maps $0 \to 1$, etc.

Let us pass to two important consequences of the existence of such maps λ or λ^*. The first is Picard's theorem.

Theorem 2.5.2. *If an entire function $f : \mathbb{C} \to \mathbb{C}$ omits two points of $\mathbb{C}$, then f is constant.*

Proof. If f omits two points of $\mathbb{C}$, choose the fractional transformation h such that $f_* = h \circ f$ omits $0, 1, \infty$. and consider f_* as a holomorphic function from $\mathbb{C}$ to $\mathbb{S} \setminus 3$. We know that there is a covering map from $\mathbb{D}$ to $\mathbb{S} \setminus 3$, and the identity map $\mathbb{C} \to \mathbb{C}$ is also a covering map. Both $\mathbb{D}$ and $\mathbb{C}$ are simply connected, so the monodromy theorem, Theorem 1.8.2, implies that the map f_* can be lifted to a map $\widetilde{f_*}$ from one covering space to another, in such a way that the diagram

$$\begin{tikzcd}
\mathbb{C} \arrow[r, "\widetilde{f_*}"] \arrow[d, "\mathbf{1}"'] & \mathbb{D} \arrow[d, "\lambda^*"'] \\
\mathbb{C} \arrow[r, "f_*"] & \mathbb{S} \setminus 3.
\end{tikzcd} \tag{2.5.5}$$

is commutative. In particular, $\widetilde{f_*}$ is a bounded holomorphic function on $\mathbb{C}$, so it is constant. Therefore $f_* = \lambda \circ \widetilde{f_*}$ is constant. □

Exactly the same procedure reduces the following theorem of Montel to Montel's theorem, Corollary 2.3.2.

Theorem 2.5.3. *Suppose that $\mathscr{F}$ is a family of rational functions on some domain, and suppose that there are at least three points in $\mathbb{S}$ that are omitted by each $f \in \mathscr{F}$. Then $\mathscr{F}$ is a normal family.*

Proof. Choose a linear fractional transformation h such that each $f_* = h \circ f$ omits $0, 1, \infty$, and restrict to a simply connected subdomain Ω. Passing to the diagram (2.5.5) with Ω in place of $\mathbb{C}$, and any $f \in \mathscr{F}$, we find that the family of lifts $\{\widetilde{f_*}\}$ is a normal family, by Corollary 2.3.2. It follows easily that $\mathscr{F}$ is a normal family. □

2.6 Jordan domains and Carathéodory's extension theorem

A *Jordan curve* in $\mathbb{C}$ is a closed curve γ that is *simple*, i.e. does not intersect itself; see Figure 2.4 According to the *Jordan curve theorem*, the complement of γ consists of two simply connected components, one that is bounded and one that is unbounded. We follow here the usual practice of remarking that this is intuitively obvious, and referring elsewhere for a proof, e.g. Newman [155].

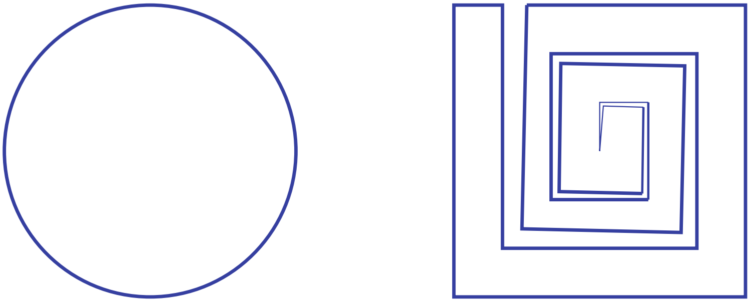

Fig. 2.4 Two Jordan curves.

The bounded component of the complement of a Jordan curve γ is called a *Jordan domain*. In other words, a Jordan domain in $\mathbb{C}$ is a bounded domain whose boundary is a Jordan curve. In view of Theorem 2.4.1, any two Jordan domains are conformally equivalent. What one would wish is that a conformal map from one Jordan domain to another extends continuously to the boundary curves, yielding a homeomorphism between the boundaries. In fact, the wish has been granted: [39].

Theorem 2.6.1. (Carathéodory) *A conformal mapping f from one Jordan domain onto another has a continuous extension to the closures. The restriction to the boundary is a homeomorphism.*

Proof: We may assume that one of the domains is the unit disk. Suppose that f is a conformal map of $\mathbb{D}$ onto a Jordan domain Ω. Note that the images of the disks $D(r, 0)$, $0 > r < 1$, fill out Ω and, therefore, eventually cover any given compact subset of Ω. Therefore if a sequence $\{z_n\}$ in $\mathbb{D}$ converges to a point $\zeta \in \partial\mathbb{D}$, some subsequence of $\{f(z_n)\}$ converges to a point of $\partial\Omega$.

The *diameter* of a bounded set $S \subset \mathbb{C}$ is defined to be the supremum of the distance between two points of S. By assumption the boundary Γ is homeomorphic to the unit circle, from which it follows that, given $\varepsilon > 0$, if two points of Γ are sufficiently close together, then exactly one of the two subarcs of Γ determined by them has diameter $\leq \varepsilon$.

Let us look closely at f near a point ζ of the boundary of $\mathbb{D}$. For $0 < r < 1$, let $\gamma_r = \gamma_r(\zeta)$ and $B_r = B_r(\zeta)$ be the arc and domain that are the parts of the circle $\{z : |z - \zeta| = r\}$ and of the disk $D_d(\zeta)$ that are contained in $\mathbb{D}$:

$$\gamma_r \;=\; \{z \,:\, |z-\zeta| = r\} \cap \mathbb{D}, \qquad B_r \;=\; \{z \,:\, |z-\zeta| < \varrho\} \cap \mathbb{D};$$

see Figure 2.5.

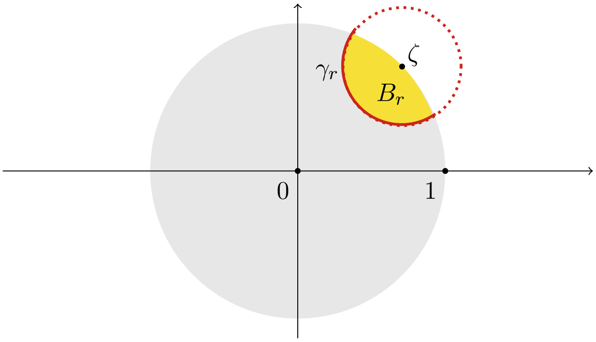

Fig. 2.5 Domain $B_r(\zeta)$ and arc $\gamma_r(\zeta)$, $r = .4$.

The image $f(\gamma_r)$ of γ_r has length $L(r)$ that can be estimated using the Cauchy–Schwarz inequality. Since $\int_{\gamma_r} |dz| < \pi r$, we obtain

$$L(r)^2 = \left[\int_{\gamma_r} |f'(z)|\,|dz|\right]^2 < \pi r \int_{\gamma_r} |f'(z)|^2\,|dz|$$
$$= \pi r \int_{\gamma_r} |f'(\zeta + re^{i\theta})|^2 r\,d\theta.$$

Therefore

$$\int_0^\delta L(r)^2 \frac{dr}{r} \;<\; \pi \int_0^\delta \int_{\gamma_r} |f'(\zeta + re^{i\theta})|^2 r\,d\theta\,dr \;=\; \text{Area}(\text{B}_\delta) \;<\; \infty.$$

The fact that the integral on the left is finite implies that there is a sequence $r_n \to 0$ such that $L(r_n) \to 0$.

Now the closure of γ_{r_n} meets $\partial\mathbb{D}$ in two points α_n and β_n. The finiteness of $L(r_n)$ implies that as one approaches α_n and β_n on γ_{r_n}, the image points converge to limits a_n and b_n on $\partial\Omega$. Clearly $|a_n - b_n| \le L(r_n) \to 0$. Passing to a subsequence, if necessary, we may assume that a_n and b_n converge to a point $w \in \partial\Omega$.

By the remarks above, this means that, given $\varepsilon > 0$, for large n exactly one of the two subarcs of $\partial\Omega$ determined by a_n, b_n has diameter $< \varepsilon$. Denote this subarc by η_n. Together $f(\gamma_{r_n})$ and η_n form a Jordan curve that encloses a domain $\Omega_n \subset \Omega$. We claim that $\Omega_n = f(B_{r_n})$ In fact γ_{r_n} divides $\mathbb{D}$ into two disjoint simply connected subdomains and the images under f are two connected subdomains of Ω whose complement in Ω is $f(\gamma_{r_n})$. Note also that since the diameter of the two parts of the bounding curve approach 0, the diameter of $\Omega_n \to 0$. In fact every point of the curve $f(\gamma_n) \cup \eta_n$ is within the maximum M of diam $f(\gamma_n)$ and diam η_n of a_n, so the complement of $D_M(a_m)$. Therefore Ω_n has diameter $\leq 2M$.

To complete the proof, it is enough to show that f is uniformly continuous on $\mathbb{D}$. In fact this will imply that it extends to be continuous on the closure. A conformal map allows us to interchange the roles of the two domains Ω and $\mathbb{D}$ and conclude that f^{-1} also extends continuously to the boundary, from which it follows that f is a homeomorphism from one boundary onto the other.

If f were not uniformly continuous, there would be two sequences $\{s_m\}$, $\{t_m\}$ in $\mathbb{D}$ and a constant $\varepsilon > 0$ such that $|s_m - t_m| \to 0$ but $|f(s_m) - f(t_m)| \geq \varepsilon$. Passing to subsequences, we may assume that both the original sequences converge to a point ζ which is necessarily on the boundary $\partial\mathbb{D}$. Then, given n, both sequences will eventually belong to B_{r_n}, so their images will belong to Ω_n. Since the diameter of Ω_n is eventually $< \varepsilon$, this is a contradiction. □

2.7 Hilbert spaces

For some applications we need the basics of Hilbert space theory. The starting point is an *inner product space*. This is a complex vector space H, equipped with an *inner product* (u, w), defined for each pair u, w in H and having the properties

$$(a_1u_1 + a_2u_2, w) = a_1(u_1, w) + a_2(u_2, w), \quad a_j \in \mathbb{C}, \quad u_j, w \in H; \tag{2.7.1}$$

$$(u, w) = \overline{(w, u)}, \qquad u, w \in H; \tag{2.7.2}$$

$$(u, u) > 0 \quad \text{if} \quad u \in H \quad \text{and} \quad u \neq 0. \tag{2.7.3}$$

Let

$$||u|| = \sqrt{(u, u)}.$$

A basic property is the *Cauchy–Schwarz inequality*

$$|(u, w)| \leq ||u||\,||w||. \tag{2.7.4}$$

The proof can be reduced to the case $||u|| = ||w|| = 1$. Then for each $a \in \mathbb{C}$ with $|a| = 1$,

$$\begin{aligned} 0 \leq ||u - aw||^2 &= (u - aw, u - aw) \\ &= ||u||^2 - (u, aw) - (aw, u) + ||aw||^2 \\ &= 2 - 2\,\mathrm{Re}\,\{\bar{a}(u, w)\}. \end{aligned}$$

We may choose a with $|a| = 1$ in such a way that $\mathrm{Re}\,\{\bar{a}(u, w)\} = |(u, w)|$.

The Cauchy–Schwarz inequality implies the triangle inequality

$$||u + w|| \leq ||u|| + ||w||.$$

This and the positivity property (2.7.3) imply that $d(u, w) = ||u - w||$ is a metric. The space H is said to be a *Hilbert space* if H is complete with respect to this metric.

First example: the space $l^2(\mathbb{Z})$ of two-sided complex sequences $\mathbf{x} = (x_n)_{-\infty}^{\infty}$ such that

$$\sum_{n=-\infty}^{\infty} |x_n|^2 \;<\; \infty,$$

with inner product

$$(\mathbf{x}, \mathbf{y}) \;=\; \sum_{n=-\infty}^{\infty} x_n \bar{y}_n.$$

The Cauchy–Schwarz inequality, applied to partial sums, implies that the inner product is well defined. This space is easily shown to be complete.

Second example: the space of continuous functions $u : \mathbb{R} \to \mathbb{C}$ that are periodic, with period 2π:

$$u(x + 2\pi) \;=\; u(x), \qquad (u, w) \;=\; \frac{1}{2\pi} \int_{-\pi}^{\pi} u(x)\overline{w(x)}\, dx.$$

The completion of this inner product space with respect to the associated metric is $L^2_{\text{per}}(\mathbb{R})$; it can be identified with the corresponding space for the interval $[0, 2\pi]$,

$$L^2([0, 2\pi]).$$

Two elements u, w of an inner product space are said to be *orthogonal*, written $u \perp w$, if $(u, w) = 0$. Note that

$$u \perp w \;\Rightarrow\; ||u + w||^2 \;=\; ||u||^2 + ||w||^2.$$

An *orthonormal set* in an inner product space H is a subset consisting of elements $\{\varphi_j\}$ such that

$$(\varphi_j, \varphi_k) \;=\; \begin{cases} 1 & \text{if } \; j = k; \\ 0 & \text{if } \; j \neq k. \end{cases}$$

For our purposes the index set $\{j\}$ here is finite or countable. Let us suppose that it is the integers $\mathbb{Z}$. If $\{\varphi_n\}$ is an orthonormal set in H, let

$$u_n \;=\; \sum_{|j| \le n} (u, \varphi_j)\, \varphi_j.$$

An easy calculation shows that u_n and $u - u_n$ are orthogonal, so

$$\sum_{|j| \le n} |(u, \varphi_j)|^2 \;=\; ||u_n||^2 \;=\; ||u||^2 - ||u - u_n||^2.$$

This implies *Bessel's inequality*:

$$\sum_{|j|\leq n} |(u,\varphi_j)|^2 \;\leq\; ||u||^2, \tag{2.7.5}$$

and also *Bessel's equality*:

$$||u-u_n|| \;\to 0 \;\Leftrightarrow\; \sum_{j=-\infty}^{\infty} |(u,\varphi_j)|^2 \;=\; ||u||^2. \tag{2.7.6}$$

The orthonormal set $\{\varphi_j\}$ is said to be *complete*, or *an orthonormal basis* if u_n converges to u for every $u \in H$. Note that u_n is the element closest to u in the subspace H_n spanned by $\{\varphi_j\}_{-n}^{n}$. In fact if w belongs to H_n, then

$$||u-(u_n+w)||^2 \;=\; ||u-u_n||^2 + ||w||^2$$

is minimal when $w = 0$.

An element w of H induces a linear transformation

$$f_w(u) \;=\; (u,w)$$

from H to $\mathbb{C}$. This transformation is *bounded*:

$$|f_w(u)| \;\leq C\,||u||$$

where the constant C can be taken to be $||w||$. It is an important fact about Hilbert spaces that the converse is true. A helpful identity here is the *parallelogram identity*: the sum of the squares of the diagonals of a parallelogram equals the sum of the squares of the sides.

$$||u+w||^2 + ||u-w||^2 \;=\; 2||u||^2 + 2||w||^2. \tag{2.7.7}$$

Proposition 2.7.1. *If $f : H \to \mathbb{C}$ is a bounded linear transformation, then there is a unique element w of H such that*

$$f(u) \;=\; (u,w), \quad \textit{all } u \in H.$$

Proof: We may assume that f is not identically zero. Let H_1 be the null space: $H_1 = \{u \in H : f(u) = 0\}$. Since f is bounded, H_1 is closed. If w exists, it must be orthogonal to H_1. Take any w_0 that is not in H_1 and look for an element u_0 of H_1 that is closest to w_0. Then, as argued above, $w_1 = w_0 - u_0$ is orthogonal to H_1, so w should be a multiple of w_1. To put this argument into effect, we need to show that there is indeed a $u_0 \in H_1$ closest to w_0. We choose a sequence $\{u_n\}$ in H_1 such that

$$\lim_{n\to\infty} ||u_n - w_0|| \;=\; \inf_{u\in H_1} ||u - w_0||.$$

Then applying (2.7.7) with $u_n - w_0$ and $u_m - w_0$ in place of u and w,

$$\begin{aligned}||u_n - u_m||^2 &= 2||u_n - w_0||^2 + 2||u_m - w_0||^2 - ||(u_n + u_m) - 2w_0||^2 \\ &= 2||u_n - w_0||^2 + 2||u_m - w_0||^2 - 4\left|\left|\tfrac{1}{2}(u_n + u_m) - w_0\right|\right|^2.\end{aligned}$$

Considering how the u_n were chosen, and that $\frac{1}{2}(u_n + u_m)$ belongs to H_1, we see that the sequence $\{u_n\}$ converges to $u_0 \in H$. The argument shows that u_0 is unique, so we see that the orthogonal complement of H_1 is one-dimensional. For any $\lambda \in \mathbb{C}$, $w = \lambda(w_0 - u_0)$ is orthogonal to H_1. Therefore the function

$$f_w(u) \;=\; (u, w)$$

agrees with f on H_1. If we choose λ so that $\lambda||w||^2 = f(w)$, then f_w and f agree on the orthogonal complement as well. □

2.8 L^p spaces and measure

Suppose that Ω is either a domain in $\mathbb{C}$ or an open interval in $\mathbb{R}$. For convenience we denote the variable in Ω by z in either case. We denote by $C(\Omega)$ the space of continuous functions $f : \Omega \to \mathbb{C}$. The *support* of a function in any of these spaces is the smallest subset S of Ω such that $f = 0$ on the complement $\Omega \setminus S$. The subspace of $C(\Omega)$ that consists of functions with compact support is denoted $C_c(\Omega)$.

Given an index p, $1 \le p \le \infty$, the L^p norm of a function $f \in C_c(\Omega)$ is defined to be

$$\begin{cases} ||f||_p \;=\; \left[\iint_\Omega |f(z)|^p \, dm(z)\right]^{1/p}, \quad 1 \le p < \infty; \\ ||f||_\infty \;=\; \sup_z |f(z)| \end{cases} \tag{2.8.1}$$

Here $m(z)$ is a convenient shorthand for $dx\, dy$ in the two-dimensional case $z = x + iy$, and $m(x) = dx$ in the one-dimensional case. Eventually m will be interpreted as a measure.

The defining properties of a norm in a vector space of functions are

(i) $||f|| \;>\; 0$ unless $f = 0$;

(ii) $||af|| \;=\; |a| \, ||f||$ if $a \in \mathbb{C}$;

(iii) $||f + g|| \;\le\; ||f|| + ||g||$.

It is clear that (2.8.1) satisfies the first two of these properties, but (iii) is not obvious except for $p = 1$ and $p = \infty$. To prove (iii) for $1 < p < \infty$ we introduce the concept of a *dual index*, and prove an important inequality. The dual index for p, $1 \le p \le \infty$, is the index q such that

$$\frac{1}{p} + \frac{1}{q} \;=\; 1, \tag{2.8.2}$$

Thus 1 and ∞ are dual. Otherwise $q = p/(p-1)$. The key inequality (when generalized to the completions L^p, L^q) is *Hölder's inequality*.

Proposition 2.8.1. *If f and g belong to $C_c(\Omega)$ and p, q are dual indices, then*

$$\left|\int_\Omega f(z)\,g(z)\,dm(z)\right| \;\leq\; ||f||_p ||g||_q \tag{2.8.3}$$

Proof: We may normalize and assume that $||f||_p = ||g||_q \;=\; 1$. We use the elementary inequality

$$ab \;\leq\; \frac{a^p}{p} + \frac{b^q}{q} \quad \text{if } \; a, b > 0, \;\; p, q > 0, \;\; \text{and} \;\; \frac{1}{p} + \frac{1}{q} = 1.$$

Integrating this inequality, with $a = |f(z)|$, $b = |g(z)|$, gives (2.8.3). □

Proposition 2.8.2. *If f belongs to $C_c(\Omega)$ and p, q are dual indices, then*

$$||f||_p \;=\; \sup\left\{\left|\int f(z)\,g(z)\,dm(z)\right|,\;\; g \in C^0_c(\Omega),\;\; ||g||_q = 1\right\}. \tag{2.8.4}$$

Proof: Assume $1 < p < \infty$ and normalize again with $||f||_p = 1$. It follows from (2.8.3) that $||f||_p$ is at most equal to the right side of (2.8.4). To prove equality, define $g(z) = 0$ if $f(z) = 0$, and otherwise let

$$g(z) \;=\; |f(z)|^{p-1}\,\frac{\overline{f(z)}}{|f(z)|}.$$

Then $||g||_q = ||f||_p = 1$ and $f(z)g(z) = |f(z)|^p$, so equality holds in (2.8.4). □

Corollary 2.8.3. *For $1 \leq p \leq \infty$,*

$$||f + g||_p \;\leq ||f||_p + ||g||_p.$$

For $1 \leq p < \infty$, the space $L^p(\Omega)$ is defined to be the completion of $C_c(\Omega)$ with respect to the norm (2.8.1). The elements of $L^p(\Omega)$ can therefore be considered as equivalence classes of Cauchy sequences of continuous functions. They can also be considered equivalence classes of measurable functions. Let us briefly discuss Lebesgue measure in $\mathbb{C}$; the transfer to $\mathbb{R}$ will be clear.

The *outer measure* $m^*(S)$ of a set $S \subset \mathbb{C}$ is the infimum of the sum of the areas $A(R_n)$, where $\{R_n\}$ is a sequence of open rectangles with sides parallel to the coordinate axes that covers S: $S \subset \bigcup R_n$. It is an exercise to show that the outer measure of a countable set is zero, and the outer measure of a disk or a rectangle is its area. The *inner measure* $m_*(S)$ of a bounded set S is defined to be $m^*(R) - m^*(R \setminus S)$ for any rectangle R that contains S. This is independent of the choice of R. A bounded set S is said to be *measurable* if $m^*(S) = m_*(S)$, and the common value is the measure $m(S)$. An unbounded set S is said to be measurable if its intersection with each rectangle R is measurable; $m(S)$ is the supremum of $m(S \cap R)$. Two things are especially worth noting here. One is that not every set is measurable. The other is that non-measurable sets do not occur easily; it takes some ingenuity to prove that they exist. Some facts: The complement of a measurable set is measurable. Open sets and closed sets are measurable. The union or intersection of a countable collection of measurable sets is measurable, If they are pairwise disjoint, the measure of the union is the sum of the measures (countable additivity). The measure of a decreasing sequence of measurable sets is the limit of their measures.

A statement about $\mathbb{C}$ is said to hold *almost everywhere* (abbreviated a.e.) if it is true of every $z \in \mathbb{C}$ except (possibly) for a set having measure zero.

By definition, a function $f : \mathbb{C} \to \mathbb{C}$ is *measurable* if $S \subset \mathbb{C}$ open implies $f^{-1}(S)$ is measurable. The elements of each L^p space, $1 \le p < \infty$ are measurable functions; two such functions represent the same element of L^p if and only if they are equal a.e. If $\{f_n\}$ is a Cauchy sequence in L^p, then it may not converge a.e. but some subsequence converges a.e.

We defined the L^∞ norm of a function belonging to $C_c(\Omega)$. The completion of this space with respect to this norm is the space C_0 of continuous functions with limit 0 at $\partial\Omega$. The space $L^\infty(\Omega)$ is defined to be the space of measurable functions f with the property that for some M, $|f(z)| \le M$ a.e.

2.9 Convolution, approximation, and weak solutions

Limits of nice functions (holomorphic, harmonic, ...) may give rise to functions that satisfy the same equations ($f_{\bar z} = 0$, $u_{xx} + u_{yy} = 0$, ...) in a "weak" sense, or in "the sense of distributions." In this case and some other important cases, weak solutions are necessarily "strong solutions": solutions in the ordinary sense. A tool for proving this—approximation by convolution—occurs in many other contexts as well.

The functions considered in this section map a domain $\Omega \subset \mathbb{C}$ to $\mathbb{C}$. Given an integer $n > 0$, let $C_c^n(\Omega)$ denote the space of such functions whose partial derivatives of order up to n exist and belong to $C_c(\Omega)$. Similarly, $C_c^\infty(\Omega)$ is the space of functions that belong to $C_c^n(\Omega)$ for every n. It is not obvious that $C_0^\infty = C_c^\infty(\mathbb{C})$ contains any non-zero functions. However, as we shall see, it is dense in each of the spaces $L^p = L^p(\mathbb{C})$, $1 \le p < \infty$. A starting point is the function

$$G(z) \;=\; c\,\exp(1 - |z|^2) \cdot \max\{1 - |z|, 0\}, \tag{2.9.1}$$

It is an exercise to show that, for any $c > 0$, this function belongs to C_c^∞. It is positive for $|z| < 1$ and otherwise vanishes. We choose c so that

$$\iint_{\mathbb{C}} G(z)\,dm(z) \;=\; 1. \tag{2.9.2}$$

Given $0 < \varepsilon \le 1$, let $G_\varepsilon(z) \;=\; \varepsilon^{-2} G(z/\varepsilon)$ Then the support of G_ε is $\{z : |z| \le \varepsilon\}$ and

$$\iint_{\mathbb{C}} G_\varepsilon(z)\,dm(z) \;=\; 1.$$

Thus G_ε has L^1 norm 1, but is more and more concentrated near 0. As we shall see, convolution with G_ε is a systematic way of taking smooth averages of translates of a function, in such a way that the averages get close to the function itself.

The *convolution* of two functions $f, g \in C_c$ is the function $f * g$ defined by

$$f * g(z) = \iint_{\mathbb{C}} f(z-w)g(w)\, dm(w).$$

A change of variables shows that

$$f * g = g * f.$$

Consider a Riemann sum approximation

$$f * g(z) = g * f(z) \sim \sum_j g(z - x_j) f(x_j)(x_{j+1} - x_j). \qquad (2.9.3)$$

Each translate $g_j(z) = g(z - x_j)$ has the same L^p norm as g, so the L^p norm of the approximation (2.9.3) is at most

$$\left[\sum_j |f(x_j)|(x_{j+1} - x_j) \right] ||g||_p.$$

Taking a limit gives

$$||f * g||_p \le ||f||_1 ||g||_p, \qquad 1 \le p < \infty. \qquad (2.9.4)$$

Since C_c is dense in each L^p, $1 \le p < \infty$, it follows that convolution can be extended to pairs $f \in L^1$, $g \in L^p$, and (2.9.4) remains valid.

If f belongs to C_c^1, then the partial derivatives satisfy

$$(f * g)_x = f_x * g, \qquad (f * g)_y = f_y * g. \qquad (2.9.5)$$

Theorem 2.9.1. *Let $\{G_\varepsilon\} \subset C_c^\infty$ be the family of functions defined above. Then for each $f \in L^p$, $1 \le p < \infty$, the functions $f_\varepsilon = G_\varepsilon * f$ belong to C^∞ and*

$$\lim_{\varepsilon \to 0} ||f_\varepsilon - f||_p = 0. \qquad (2.9.6)$$

Proof: For $p > 1$, each G_ε belongs to the dual space L^q, $1/p + 1/q = 1$, so Hölder's inequality (2.8.3) gives a bound for $G_\varepsilon * g$, and translations of $G_\varepsilon * g$ involve translations of G_ε. For $p = 1$, the same argument works using boundedness of G_ε. Combining this argument with (2.9.5) and iterations, we see that each G_ε is infinitely differentiable.

It is enough to prove (2.9.6) on a dense subset. For $f \in C_c$, a change of variables implies that

$$f_\varepsilon(z) \;=\; G_\varepsilon * f(z) \;=\; \iint_{\mathbb{C}} G(w)[f(z-\varepsilon w) - f(z)]\, dm(w). \tag{2.9.7}$$

This identity and (2.9.2) imply that g_ε converges uniformly to g. moreover, for $0 < \varepsilon \leq 1$, g_ε vanishes outside a fixed bounded set (the set of points at distance ≤ 1 from the support of f). Therefore (2.9.6) holds. □

Corollary 2.9.2. *The space C_c^∞ is dense in each L^p, $1 \leq p < \infty$.*

Now let us turn to weak solutions. The idea is that if, say, f is a C^1 function and $\phi \in C_c^1$, then

$$\iint_{\mathbb{C}} g\,\phi \;=\; -\iint_{\mathbb{C}} f\,\phi_x, \qquad g = f_x. \tag{2.9.8}$$

If we assume that f and g are functions that belong to the space of locally integrable functions L^1_{loc}, meaning that $\int_B |g| < \infty$ and $\int_B |f| < \infty$ for each bounded set B, then both sides of (2.9.8) are well defined for each $\phi \in C_0^1$. If equality holds for each such "test function" ϕ, then f is said to satisfy $f_x = g$ in the weak sense, or $f_x = g$ weakly. Thus f can in principle be a weak solution without having a continuous derivative. However if the partial derivative f_x exists and is continuous, then an integration by parts in (2.9.8) leads to

$$\iint_{\mathbb{C}} (f_x - g)\phi \;=\; 0$$

for every test function ϕ, from which it follows that $f - g = 0$ a.e., so essentially f is an ordinary ("strong") solution of $f_x = g$. Equations of higher order are treated in the same way.

Without loss of generality, and to accommodate equations of any order, we may take the set of test functions in any case to be C_c^∞.

The three cases to be considered here, for later use, are the weak solution of $f_{\bar z} = 0$ in a domain Ω, characterized by

$$\iint_{\Omega} f\,\phi_{\bar z} \;=\; 0, \qquad \text{all } \phi \in C_c^\infty(\Omega), \tag{2.9.9}$$

the weak solution of $\Delta f \equiv f_{xx} + f_{xx} = 0$ in a domain Ω, characterized by

$$\iint_{\Omega} f\,\Delta\phi \;=\; 0, \qquad \text{all } \phi \in C_0^\infty(\Omega) \tag{2.9.10}$$

and the system $f_{\bar z} = p$, $f_z = q$, under the assumption that p and q are weak solutions of $p_{\bar z} = q_z$. The results concerning (2.9.9) and (2.9.10) are both known as *Weyl's lemma*.

Theorem 2.9.3. *Suppose $f \in L^p_{loc}(\Omega)$ is a weak solution of $f_{\bar z} = 0$ in Ω. Then f is holomorphic in Ω.*

Proof: Extend f to $\mathbb{C}$ by setting $f = 0$ on the complement of Ω. The extended f still satisfies (2.9.9) for each $\phi \in C_c(\Omega)$. Given such a function ϕ, note that for sufficiently small $\varepsilon > 0$, the support of the convolution $G_\varepsilon * \phi$ is also a compact subset of Ω. Let $f_\varepsilon = G_\varepsilon * f$. Then for small ε, using the fact that G_ε is an even function, we have

$$\begin{aligned}
\iint_\Omega (f_\varepsilon)_{\bar z}\, \phi(z)\, dm(z) &= \iint_\Omega (G_\varepsilon)_{\bar z} * f(z)\, \phi(z)\, dm(z) \\
&= \iint_\Omega (G_\varepsilon)_{\bar z}(z - w)\ f(w)\, \phi(z)\, dm(w)\, dm(z) \\
&= \iint_\Omega f(w) \left[\iint_\Omega (G_\varepsilon)_{\bar z}(z - w)\, \phi(z)\, dm(z) \right] dm(w) \\
&= - \iint_\Omega f(w)\, G_\varepsilon * \phi_{\bar w}(w)\, dm(w) \ = \ 0.
\end{aligned}$$

It follows that for any given open disk D whose closure is in Ω, f_ε is eventually holomorphic in D. The restrictions $f_\varepsilon|_D$ converge to $f|_D$ in $L^1(D)$. By Proposition 2.3.7, f is holomorphic in D. Thus f is holomorphic in Ω. □

The proof of Theorem 2.9.4 adapts readily to the harmonic case.

Theorem 2.9.4. *Suppose $f \in L^p_{loc}(\Omega)$ is a weak solution of $\Delta f_{\bar z} = 0$ in Ω. Then f is harmonic in Ω.*

Now we come to the third case mentioned above. Suppose here that p and q are functions in C^1. We can construct a function $f \in C^2$ such that $f_z = p$, $f_{\bar z} = q$ if and only if p and q satisfy the equality-of-mixed-partials condition $p_{\bar z} = q_z$. In fact the unique solution with $f(0) = 0$ must be given by the formula

$$\begin{aligned}
f(z, \bar z) &= \int_0^1 \frac{\partial f}{\partial s}(sz, s\bar z)\, ds \\
&= \int_0^1 [z f_z(sz, s\bar z) + \bar z f_{\bar z}(sz, s\bar z)]\, ds \\
&= \int_0^1 z\, p(sz, s\bar z) + \bar z q(sz, s\bar z)]\, ds. \qquad (2.9.11)
\end{aligned}$$

It is an exercise to verify that the function defined by last integral satisfies $f_z = p$, $f_{\bar z} = q$. Now we want to carry this over to the case of weak solutions of $p_{\bar z} = q_z$.

Theorem 2.9.5. *Suppose that p and q are continuous functions whose weak derivatives $p_{\bar z}$ and q_z belong to L^1_{loc} and are weak solutions of $p_{\bar z} = q_z$, i.e. for each $\phi \in C_c^\infty$,*

$$\iint_{\mathbb{C}} [p\, \phi_{\bar z} - q\, \phi_z] \ = \ 0. \qquad (2.9.12)$$

Then there is a C^1 function f such that $f_z = p$, $f_{\bar z} = q$.

Proof: With G_ε as above, let $p_\varepsilon = G_\varepsilon * p$ and $q_\varepsilon = G_\varepsilon * q$. The same type of argument that we used in the proof of Theorem 2.9.3 shows that $(p_\varepsilon)_{\bar z} = (q_\varepsilon)_z$. Therefore we may construct f_ε by using the integral in (2.9.11) with p and q replaced by p_ε and q_ε. Since p and q are assumed to be continuous, the formula (2.9.7), with f replaced by p or q shows that p_ε and q_ε converge to p and q uniformly on bounded sets. Therefore f_ε converges to f in C^1, with $f_z = p$, $f_{\bar{z}} = q$. □

2.10 The gamma function

The gamma function $\Gamma(z)$ is usually defined for $\operatorname{Re} z > 0$ by

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1}\, dt. \tag{2.10.1}$$

An integration by parts yields the *functional equation*

$$\Gamma(z+1) = z\,\Gamma(z). \tag{2.10.2}$$

In particular, $\Gamma(1) = 1$ and $\Gamma(n+1) = n\,!$, $n = 1, 2, 3, \dots$. Moreover, since the left side of (2.10.2) is defined for $\operatorname{Re} z > -1$, we may use (2.10.2) to extend the definition of $\Gamma(z)$ to the strip $-1 < \operatorname{Re} z \le 0$. Continuing this process allows us to extend Γ to the complement of the non-positive integers. The result is a function meromorphic in $\mathbb{C}$ with simple poles at the negative integers. The residue of Γ at $-n$, $n = 0, 1, 2, 3, \dots$, is $(-1)^n n\,!$

A particularly important case is $\Gamma(1/2)$. Now

$$\Gamma(1/2) = \int_0^\infty e^{-t} t^{-1/2}\, dt = 2\int_0^\infty e^{-t} d(t^{1/2}) = 2\int_0^\infty e^{-s^2}\, ds = \int_{-\infty}^\infty e^{-s^2}\, ds.$$

and

$$\begin{aligned}\left[\int_{-\infty}^\infty e^{-s^2}\, ds\right]^2 &= \iint e^{-(x^2+y^2)}\, dx\, dy = \iint e^{-r^2} r\, dr\, d\theta \\ &= 2\pi \int_0^\infty e^{-r^2} r\, dr = \pi \int_0^\infty e^{-u}\, du = \pi, \end{aligned} \tag{2.10.3}$$

so

$$\Gamma(1/2) = \sqrt{\pi}. \tag{2.10.4}$$

The asymptotics of the gamma function are given by *Stirling's approximation*:

Theorem 2.10.1.

$$\Gamma(x) = \left(\frac{x}{e}\right)^x \left[\left(\frac{2\pi}{x}\right)^{1/2} + O(x^{-3/2})\right] \tag{2.10.5}$$

as $x \to +\infty$.

Proof. By (2.10.2),

$$\Gamma(x) = \frac{\Gamma(x+1)}{x} = \frac{1}{x}\int_0^\infty e^{-t}t^x\,dt.$$

The integrand is maximal at $t = x$, so with $t = xu$ we have

$$\begin{aligned}\Gamma(x) &= \frac{1}{x}\int_0^\infty e^{-xu}(xu)^x\,x\,du\\ &= \left(\frac{x}{e}\right)^x \int_0^\infty \left(u\,e^{1-u}\right)^x\,du.\end{aligned}$$

This integrand decays exponentially away from $u = 1$. Comparing Taylor expansions at $u = 1$ shows that

$$u\,e^{1-u} = e^{-(u-1)^2/2}[1 + \tfrac{1}{3}(u-1)^3 + O((u-1)^4)].$$

Therefore, setting $(u-1) = s/\sqrt{x}$, we have

$$\int_0^\infty \left(u\,e^{1-u}\right)^x\,du = \int_{-\infty}^\infty e^{-s^2/2}\left[1 + x\{\tfrac{1}{3}s^3x^{-3/2} + O(s^4x^{-2})\}\right]\frac{ds}{\sqrt{x}}.$$

Since $e^{-s^2/2}s^3$ is an odd function, its integral vanishes and we obtain

$$\int_0^\infty \left(u\,e^{1-u}\right)^x\,du = \int_{-\infty}^\infty e^{-s^2/2}\left[1 + O(s^4x^{-1})\right]\frac{ds}{\sqrt{x}}.$$

The identity (2.10.3) is equivalent to

$$\int_{-\infty}^\infty e^{-s^2/2}\,ds = \sqrt{2\pi}.$$

Combining these estimates, we get (2.10.5). □

Remark. The estimate (2.10.5) can be shown to hold uniformly in any closed sector that omits the negative half-axis:

$$\Gamma(z) = \left(\frac{z}{e}\right)^z\left[\left(\frac{2\pi}{z}\right)^{1/2} + O(|z|^{-3/2})\right]$$

as $z \to \infty$, uniformly for $|\arg z| \le \pi - \delta$, for any $\delta > 0$. See [21] or [22].

Remarks and further reading

The material in this chapter is covered in many standard texts on complex analysis, functional analysis, or distribution theory, according to the topic. The topics in Sections 2.1 and 2.2 are treated more leisurely in [22]. Normal families are covered extensively in Schiff [184]. There are many texts on conformal mapping; two classics are Cohn [46] and Nehari [152].

Riemann's original argument for the Riemann mapping theorem came into much criticism by Schwarz and others. Subsequent arguments were given, under various assumptions on the boundary of the simply connected domain, before Koebe's definitive result; see Tazzioli [200] and Gray [92].

Boundary behavior of conformal maps is treated comprehensively in Pommerenke [171]. For much more on the material in Sections 2.7 and 2.8 see, for example, Folland [78], Rudin [182], or Stein and Shakarshi [195], [196]. For Section 2.9, see Georgiev [90] or Duistermaat and Kolk [61]. The gamma function is treated in any book on special functions, e.g. [21].

Chapter 3
Complex dynamics

Mathematically, a dynamical system generally consists of a product space $X \times T$ and a function $f : X \times T \to X$, with X thought of as "space" and T as "time." Time is usually taken to be continuous, with $T = \mathbb{R}$ or $T = [0, \infty)$, or discrete, with $T = \mathbb{Z}$ or $\{0, 1, 2, \dots\}$. The simplest situation is the autonomous discrete case with f independent of t. In other words, $f : X \to X$ and one studies the iterates f, $f \circ f$, $f \circ f \circ f, \dots$. To have a convenient notation that will not be confused with powers of f (if X has a multiplication), we set $f^{\circ 0} = \mathbf{1}$, the identity map, and

$$f^{\circ(n+1)} = f \circ f^{\circ n}, \qquad n = 1, 2, 3, \dots \ .$$

If f is invertible, this can be extended to $f^{\circ n}$, $n = -1, -2, \dots,$. (We shall always assume that f itself is not the identity map.)

One wants to understand the *orbits*

$$\{f(x), f^{\circ 2}(x), f^{\circ 3}(x), \dots\}, \qquad x \in X.$$

It is useful also to consider the *backward orbit* of a point z_0: the set

$$\{z : f^{\circ n}(z) = z_0, \quad \text{some } n \geq 1.\} \tag{3.0.1}$$

Two maps f and g from X to itself are said to be *conjugate* if there is an invertible map $\phi : X \to X$ such that

$$f \circ \phi = \phi \circ g.$$

Then

$$f^{\circ n} = \phi \circ g^{\circ n} \circ \phi^{-1},$$

so the dynamics of g can be read off from the dynamics of f and conversely.

Among the most-studied cases are those in which X is a space of one real or one complex dimension and f is a rational function—even a polynomial. These cases allow surprisingly intricate behavior, as anyone knows who has encountered the

R. Beals and R. S. C. Wong, *More Explorations in Complex Functions*, Graduate Texts
in Mathematics 298, https://doi.org/10.1007/978-3-031-28288-1_3

Mandelbrot set. This study began in earnest in the 19th century, with the examination of Newton's method for approximating zeros. It flourished in the early 20th century with work by Fatou, Julia, and others. It revived later in the 20th century, sparked in part by connections with chaos theory and fractals and was given striking visual form by the development of powerful computers and computational techniques.

In this chapter we introduce the subject, concentrating on the case $X = \mathbb{S}$, the Riemann sphere, and f a rational function: $f = P/Q$ where P and Q are polynomials with no common zero. The *degree* of f is

$$\deg f \;=\; \max\{\deg P,\ \deg Q\}.$$

The rational functions of degree 1 are the linear fractional transformations.

A residue calculus argument shows that f takes each value in $\mathbb{S}$ $\deg f$ times, counting multiplicity. This shows that degree is multiplicative:

$$\deg(f \circ g) \;=\; \deg f \cdot \deg g. \tag{3.0.2}$$

In particular,

$$\deg f^{\circ n} \;=\; (\deg f)^n. \tag{3.0.3}$$

This fact suggests (or even insists) that the study of rational dynamics is much simpler in the case of degree 1. We deal with this case immediately.

As noted above, the case $\deg f = 1$ is the case $f \in \mathrm{Aut}(\mathbb{S})$. To normalize, we note that f has exactly one or two fixed points. Consider first the case of a single fixed point, which we may take to be $z = \infty$. Then

$$f(z) \;=\; z + b, \qquad b \neq 0, \qquad f^{\circ n}(z) \;=\; z + nb,$$

and for each $z \in \mathbb{S}$, $f^{\circ n}(z) \to \infty$ as $n \to \pm\infty$.

Now suppose f has two fixed points, which we may take to be 0 and ∞. Then $f(z) = az$ for some $a \neq 0$. If $|a| < 1$, then $|f^{\circ n}(z)| = |a|^n|z|$ shrinks any bounded set to $\{0\}$ at a geometric rate. If $|a| > 1$, $f^{\circ n}$ shrinks closed sets not containing the origin to the point at ∞. Inversion $z \to 1/z$ conjugates one case to the other; thus as dynamical systems they are identical.

If $|a| = 1$, the situation is more interesting. Here $a = e^{2\pi i\theta}$ and the behavior is very different depending on whether θ is rational or not: whether some iterate $f^{\circ m}$ is the identity map (so that the orbit of each point consists of finitely many points), or not.

As we shall see, things can become much more interesting than this, even for f a quadratic. In Section 3.1 we define the Fatou set $\mathrm{F}(f)$ and the Julia set $\mathrm{J}(f)$, and examine the examples $f(z) = z^2 + c$ with $c = 0$, $c = -2$, and $c = 6$. General properties of F and J are developed in Section 3.2.

Fixed points and periodic points of f play a dominant role in the study of dynamics. Section 3.3 establishes general properties for rational f of degree ≥ 2. In Section 3.4 we look more closely at the general features of attracting and repelling fixed points. Section 3.5 introduces the basic theory of "neutral" fixed points. The special case of "parabolic" fixed points is examined in more detail in Section 3.6.

Finally, in Section 3.7 we describe the classification theorem, which characterizes the connected components of the Fatou set. We also give one more illustration of dependence on parameters by a brief discussion of the Mandelbrot set.

3.1 Fatou sets and Julia sets; some examples

We begin here a systematic consideration of discrete dynamics with f a rational function of degree ≥ 1. For practical purposes this means $\deg f > 1$, since we obtained a complete picture (up to normalization) for degree 1 in the Introduction. Even the case of quadratic polynomials illustrates many of the general phenomena. We note that a given quadratic polynomial f can be conjugated by an affine map $\phi(\zeta) = a\zeta + b$ to either of the canonical forms

$$f_c(z) \,=\, \zeta^2 + c; \qquad f^c(\zeta) \,=\, \zeta(\zeta + c); \tag{3.1.1}$$

see Exercise 2.

The *Fatou set* $\mathrm{F} = \mathrm{F}(f)$ of f is the set of points $z \in \mathbb{S}$ such that the family of iterates $\{f^{\circ n}\}_{n=1}^{\infty}$, when restricted to some small enough neighborhood U of z, is a normal family: for any subsequence $\{f^{\circ n_k}\}$ some subsequence $\{f^{\circ n_{k_j}}\}$ converges uniformly (in $\mathbb{S}$) on compact subsets of U. The complement $\mathrm{J} = \mathrm{J}(f)$ is the *Julia set* of f. Thus, by definition, F is open and J is closed.

Two maps $f : \Omega \to \Omega$ and $g : \Omega' \to \Omega'$ are said to be *conformally conjugate* if there is a conformal map $\Phi : \Omega' \to \Omega$ such that

$$g \,=\, \Phi^{-1} \circ f \circ \Phi. \tag{3.1.2}$$

Again this means that the dynamics are the same. In fact

$$g^{\circ n} \,=\, \Phi^{-1} \circ f^{\circ n} \circ \Phi. \tag{3.1.3}$$

Remark. We may always study the dynamics of f at up to three prescribed points $z_j \in \mathbb{S}$ by using a Möbius transformation Φ so that this amounts to studying the dynamics of g at prescribed points w_j. In particular, behavior near a pole or at infinity can be reduced to behavior at a regular point or at a finite point.

Example. Consider the triple

$$\begin{aligned}
&g : \Omega \to \Omega, \quad && \Omega \,=\, \{z \,:\, z \notin [-2, 2]\}, \quad && g(z) \,=\, z^2;\\
&f : \Omega' \to \Omega', \quad && \Omega' \,=\, \{z \,:\, |z| > 1\}, \quad && f(z) \,=\, z^2 - 2;\\
&\Phi : \Omega \to \Omega', \quad && \Phi(z) \,=\, z + z^{-1}. &&
\end{aligned} \tag{3.1.4}$$

Then (3.1.2) is satisfied; see Exercise 3. The same is true if replace the domain Ω' by $\Omega'' = \{z : |z| < 1\}$. Now it is easily checked that the Julia sets of g and f are

$$\mathrm{J}(g) \;=\; \{z : |z| = 1\}, \qquad \mathrm{J}(f) \;=\; [-2, 2]; \tag{3.1.5}$$

Exercise 4.

These very simple Julia sets are not representative of the general situation, even for quadratic polynomials.

A more instructive example is

$$f(z) \;=\; z^2 - 6. \tag{3.1.6}$$

with the two-valued inverse map

$$g(z) \;=\; \sqrt{6+z}.$$

Proposition 3.1.1. *If $f(z) = z^2 - 6$, then*
(a) The fixed points of f in $\mathbb{S}$ are $\{-2, 3, \infty\}$.
(b) If $|z| \geq 3$ then $|f(z)| \geq 3$, with equality only if $z = \pm 3$.
(c) If $|z| > 3$ then $|f(z)| - 3 > 2(|z| - 3)$, so $f^{\circ n}(z) \to \infty$.
(d) If $|z| \leq \sqrt{3}$ then $|f(z)| \geq 3$.
(e) If $|z| \leq 3$, then $|\mathrm{Im\,g(z)}| \leq |\mathrm{Im\,z}|/2\sqrt{3}$.

Proof. This is left as Exercise 5.

By following up properties (a)—(e), we can show that the Julia set $\mathrm{J}(f)$ is a Cantor set. Note first that it follows from (b), (c), (d) that, with the exception of ± 3 and $\pm\sqrt{3}$,

$$f^{\circ n}(z) \;\to\; \infty \quad \text{if} \;\; |z| \geq 3 \;\; \text{or} \;\; |z| \leq \sqrt{3}. \tag{3.1.7}$$

It follows from (e) that if $|z| < 3$ and $\mathrm{Im\,z} \neq 0$, then $|\mathrm{Im\,f^{\circ n}(z)}|$ increases geometrically so long as it remains in $D_3(0)$. Because of this we can sharpen (3.1.7):

$$f^{\circ n} \;\to\; \infty \quad \text{if } z \notin [-3.3]. \tag{3.1.8}$$

For real x, we know that $f^{\circ n}(x) \to \infty$ if

$$x \;\in\; A_0 \;=\; (-\sqrt{3}, \sqrt{3}).$$

But then the same is true if $x \in g(A_0)$, and so on. Let

$$A_n \;=\; g^{\circ n}(A_0) \;=\; \{x \in (-3, 3)\,;\; |f^{\circ n}(x)| < 3, |f^{\circ (n+1)}(x)| > 3\}.$$

Then the sets A_n are disjoint, and for real x,

$$f^{\circ n}(x) \to \infty \quad \text{if and only if} \quad x \in \bigcup_{n=0}^{\infty} A_n. \tag{3.1.9}$$

Now (each branch of) g is continuous and strictly monotone on $[-3, 3]$, so we simply need to track the endpoints $\pm\sqrt{3}$. The first stage is

$$A_1 = g(A_0) = \left(-(6+\sqrt{3})^{1/2}, -(6-\sqrt{3})^{1/2}\right) \cup \left((6-\sqrt{3})^{1/2}, (6+\sqrt{3})^{1/2}\right).$$

Note that these are subintervals of $[-3, -\sqrt{3}]$ and $[\sqrt{3}, 3]$, respectively.

Let $B_n = \bigcup_{j=0}^{n} A_j$.

Lemma 3.1.2. *B_n consists of $2^{n+1} - 1$ disjoint intervals. The 2^n intervals of A_n interlace the $2^n - 1$ intervals of B_{n-1}.*

Proof. This is clearly the case at $n = 1$. Suppose that assertion is true for n. Then the positive and negative parts of $g(A_n)$ each consist of 2^n disjoint intervals, so A_{n+1} consists of 2^{n+1} disjoint intervals. The positive part of $g(B_n)$ consists of 2^n disjoint intervals, interlaced by the intervals of $g(A_n)$. These last are the positive part of A_{n+1}. By symmetry the intervals of A_{n+1} interlace those of $g(B_{n-1}) = B_n \setminus A_0$. Since A_n is disjoint from A_0, and all the intervals in the positive part of $g(B_n)$ lie to the right of A_0, the interlacing property carries over to all of B_n. □

Recall that the original Cantor set C is obtained by the processing of successively removing the open middle third of a union of closed intervals, starting with the unit interval. In general, a Cantor set is any set that is homeomorphic to C.

Proposition 3.1.3. *Let $f(z) = z^2 - 6$. The Julia set $\mathrm{J}(f)$ is a Cantor set contained in the interval $[-3, 3]$. For every z in the Fatou set, $f^{\circ n}(z) \to \infty$.*

Proof. Since $B_\infty = \bigcup B_n$ is an open set in $\mathbb{R}$, it follows from (3.1.8) and (3.1.9) that the Fatou set contains

$$B_\infty \cup (\mathbb{S} \setminus [-3, 3]).$$

By Lemma 3.1.2, to show that C is a Cantor set it is enough to show that the lengths of the intervals in the complement of B_n in $[-3, 3]$ shrink to zero. This follows immediately from the fact that

$$|g'(x)| = \frac{1}{2\sqrt{6+x}} \leq \frac{1}{2\sqrt{3}}.$$

It follows that points whose orbits converge to 3 are dense in the complement of B_∞ in $[-3.3]$. Therefore

$$\mathrm{F}(f) = B_\infty \cup (\mathbb{S} \setminus [-3, 3]), \qquad \mathrm{J}(f) = C \equiv [-3, 3] \setminus B_\infty. \qquad \square$$

We close this with three images that give some idea of the possible variations in form of $J(f)$, even for quadratic polynomials. The first is a Cantor set: the appearance of some connectedness is deceiving and is violated on a small enough scale. The second is a case of a figure known as a *Douady rabbit*. The third begins to show how elaborate Julia sets of the second type can become (Figures 3.1, 3.2, and 3.3).

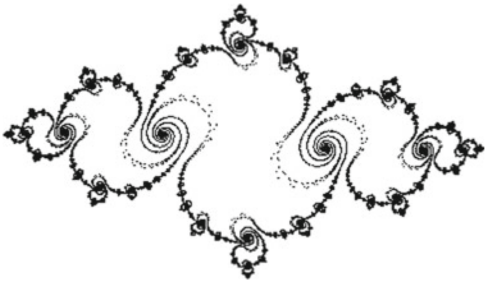

Fig. 3.1 Julia set for $f(z) = z^2 + (-.766 + .083i)$: a Cantor set.

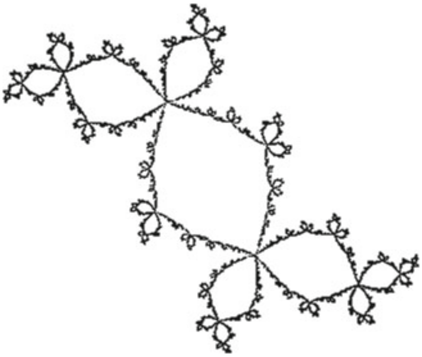

Fig. 3.2 Julia set for $f(z) = z^2 + (-.122 + .745i)$: a Douady rabbit.

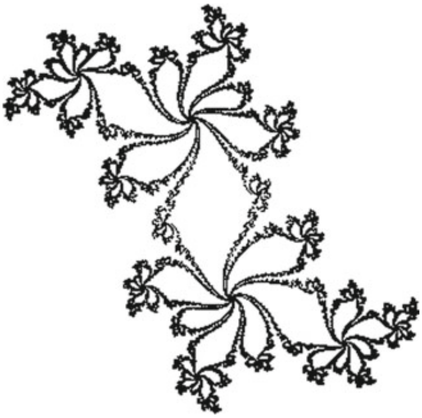

Fig. 3.3 Julia set for $f(z) = z^2 + (.125 + .604i)$: an elaborated rabbit.

These three figures were obtained from the Java Script Julia set generator. https://marksmath.org>visualization>julia_sets. The reader is invited to explore the site for other examples with $f(z) = z^2 + c$, for one's choice of c.

In subsequent sections we shall see how the general theory accounts for some of the properties of these figures, such as self-similarity.

3.2 Julia sets: invariance, density, and self-similarity

We assume throughout that $f : \mathbb{S} \to \mathbb{S}$ is a rational map of degree $d \geq 2$. Let f^{-1} denote the d-valued inverse. In this section we begin discussion of the general properties of the Julia set $\mathrm{J} = \mathrm{J}(f)$.

Proposition 3.2.1. *(a) The Julia set* $\mathrm{J} = \mathrm{J}(f)$ *is not empty.*
(b) J *is fully invariant for* f*:*

$$f(\mathrm{J}) = \mathrm{J} = f^{-1}(\mathrm{J}). \tag{3.2.1}$$

(c) For any $k \geq 1$, $\mathrm{J}(f^{\circ k}) = \mathrm{J}(f)$.

Proof. Part (a) is proved by contradiction. If $\mathrm{J} = \emptyset$ then $\{f^{\circ n}\}$ is a normal family on $\mathbb{S}$, so some subsequence $\{f^{\circ n_j}\}$ converges uniformly on $\mathbb{S}$ to a function $h : \mathbb{S} \to \mathbb{S}$ that is everywhere meromorphic, hence rational. The limit h cannot be constant, since the $f^{\circ n_j}$ are surjective. Since the convergence is uniform, eventually the number of zeros of h will be the number of zeros of $f^{\circ n_j}$. But $f^{\circ n_j}$ has degree d^{n_j} and we have assumed $d \geq 2$.

It is enough to prove statements (b) and (c) for the Fatou set. Now $z_0 \in \mathrm{F}(f)$ if and only if $\{f^{\circ n}\}$ is a normal family on some neighborhood of z_0. This is true if and only if $\{f^{\circ(n\pm1)}\}$ has this property on some neighborhood of $f^{\pm}(z_0)$. This proves (b).

Any subfamily of a normal family on a given domain is normal, so $z_0 \in \mathrm{F}(f)$ implies $z_0 \in \mathrm{F}(f^{\circ k})$. Conversely, suppose $z_0 \in \mathrm{F}(f^{\circ k})$. Then

$$\{f^{\circ n}\} = \bigcup_{j=0}^{k-1} \{f^{\circ j}(f^{\circ nk})\}.$$

Suppose $z_0 \in \mathrm{F}(f^{\circ k})$. Given a sequence in $\{f^{\circ n}\}$, some subsequence belongs to one of the families on the right, each of which is normal in a neighborhood of z_0. Therefore $\{f^{\circ n}\}$ itself is normal in that neighborhood, showing that $z_0 \in \mathrm{F}(f)$. □

Let us look more closely at points $z \in \mathrm{J}(f)$. Let U be a neighborhood of such a point, and consider the family of functions $\mathcal{G} = \{g_n\}$, $g_n = f^{\circ n}|_U$. By assumption $\mathcal{G}$ is not a normal family. By Montel's theorem, Theorem 2.5.3, the exceptional set $E_z = \mathbb{S} \setminus \bigcup g_n(U)$ of values omitted by every g_n contains at most two points. Suppose that E_z consists of a single point a. Replacing f by $h^{-1} \circ f \circ h$, where h is a linear fractional transformation that takes ∞ to a, we may assume that $E_z = \{\infty\}$, and take $U = \mathbb{C}$. By definition, $f^{-1}(E_z) \subset E_z$, so $f^{-1}(\infty) = \infty$ and there are no other poles, so f is a polynomial. If E_z consists of two points, we may take them to be 0 and ∞ and take $U = \mathbb{C} \setminus \{0\}$. Then either $f(0) = 0$ and $f(\infty) = \infty$, so $f(z) = Cz^n$ or $f(0) = \infty$ and $f(\infty) = 0$, so $f(z) = Cz^{-n}$. In either case we have proved the first part of the following.

Proposition 3.2.2. *(a) The exceptional set E_z, $z \in \mathrm{J}(f)$, is independent of z, and is contained in the Fatou set.*
(b) If z_0 belongs to $\mathrm{J}(f)$, then the backward orbit (3.0.1) of z_0 is dense in $\mathrm{J}(f)$.
(c) Any completely invariant non-empty subset of $\mathrm{J}(f)$ is dense in $\mathrm{J}(f)$.
(d) If U is a union of connected components of $\mathrm{F}(f)$ that is completely invariant, then $\mathrm{J}(f) = \partial U$.

Proof. As noted, the preceding discussion proves (a).

Given $z \in \mathrm{J}(f)$ and a neighborhood U of z, $\bigcup f^{\circ b}(U)$ is the complement of E, so by (a) it contains $\mathrm{J}(f)$. Therefore for any $z_0 \in \mathrm{J}$, some point of U is in the backward orbit of z_0. This proves (b), and (c) is a consequence of (b).

Complete invariance of U implies complete invariance of the boundary, and the assumption that U is made up of connected components of $\mathrm{F}(f)$ implies that $\partial U \subset \mathrm{J}$, so (d) follows from (c). □.

Corollary 3.2.3. *If* J *has an interior point, then* $\mathrm{J} = \mathbb{S}$.

Proof. Suppose that the open set U is contained in J. By invariance, the same is true of $\bigcup f^{\circ n}(U)$. But this set is $\mathbb{S} \setminus E$. Since J is closed, it must be all of $\mathbb{S}$. □

We supplement part (b) of Proposition 3.2.2 by considering forward orbits. Consider $\mathrm{J}(f)$ as a metric space relative to the spherical metric of $\mathbb{S}$. It is closed in $\mathbb{S}$, and therefore complete. It is a general property of complete metric spaces that the intersection of a countable family of dense open subsets is itself dense; see Exercise 9. We refer to such an intersection as a *generic* subset.

Proposition 3.2.4. *There is a generic subset $V \subset \mathrm{J}$ such that for each z in V, the forward orbit $\{f^{\circ n}(z)\}$ is dense in* J.

Proof. Given $j = 1, 2, \ldots$, let U_{j1},U_{j2},..., U_{jm_j} be an open cover of J by disks of radius $1/j$. Let V_{jk} consist of all pre-images of points of U_{jk}. Then each V_{jk} is open and dense in J, so the intersection V is generic. If z belongs to V, then each U_{jk} contains some point of the forward orbit of z. Therefore the orbit of z is dense in J. □

In the next section we begin a general discussion of fixed points of f. It is useful here to discuss one case. A finite fixed point z_0 of f is said to be *super-attracting* if $f'(z_0) = 0$. If ∞ is fixed, it is super-attracting for f if 0 is super-attracting for $g(z) = 1/f(1/z)$, i.e. $f(z)/z \to \infty)$ as $z \to \infty$.

For example, each of the examples in Section 3.1 has a super-attracting fixed point at ∞. In fact at ∞ we consider the expansion

$$\frac{1}{f(1/z)} = \frac{1}{\dfrac{1}{z^2} + O(1)} = \frac{z^2}{1 + O(z^2)} = z^2 + O(z^4).$$

Lemma 3.2.5. *If z_0 is a super-attracting fixed point of f, then z_0 belongs to the Fatou set* $\mathrm{F}(f)$.

Proof. For small enough δ, if $z \in U = D_\delta(z_0)$, then $|f(z)| \le C|z - z_0|^2$. If δ is chosen so that $C\delta^2 < \delta$, then U is invariant for f and $\{f^{\circ n}\}$ is a normal family. □

Theorem 3.2.6. *The Julia set* $\mathrm{J}(f)$ *contains no isolated points.*

Proof. Suppose $z_0 \in \mathrm{J}$. We may assume that z_0 is finite. If no backward iterate $(f^{-1})^{\circ n}(z_0)$ is equal to z_0, then, since these pre-images are dense in J, it follows that z_0 is not isolated. Otherwise z_0 is a periodic point for f: $f^{\circ m}(z_0) = z_0$, for some $m > 0$ that we may assume to be minimal. There are $(\deg f)^m$ solutions to $f^{\circ m}(z) = z_0$, so if z_0 were the only solution it would have positive multiplicity. (Recall that we are assuming that $\deg f > 1$.) But then the orbit would be super-attracting and belong to F. This leaves us with the possibility that $f^{\circ m}(z) = z_0$ has distinct solutions z_0 and z_1. Then $f^{\circ j}(z_0) \neq z_1$ for all j, since otherwise, because of the m-periodicity of z_0, it would be true for some $0 < j < m$. But then

$$f^{\circ j}(z_0) = f^{\circ(m+j)}(z_0) = f^{\circ m}(z_1) = z_0,$$

contradicting the minimality of m. Then again no preimage of z_1 is equal to z_0, but they are dense in J, so z_0 is not isolated. □

Corollary 3.2.7. *If $S \subset \mathrm{J}$ is countable, then the complement* $\mathrm{J} \setminus S$ *is dense in* J.

Proof. Let $\{z_n\}_{n=1}^{\infty}$ be an enumeration of S. Given $U \subset \mathbb{S}$ open, with $U \cap \mathrm{J}$ not empty, we may find a sequence of open disks D_n such that $D_1 \supset D_2 \supset \dots$, the closure of D_1 is contained in U, and the closure $\overline{D}_n$ contains a point of J but does not contain z_n. Then $\bigcap \overline{D}_n \subset \mathrm{J}$ is non-empty and does not contain any of the z_n. □

We are now in a position to address *self-similarity*, the "fractal" nature of Julia sets. We say that two points z_1, z_2 of J have *conformally equivalent locations* in J if there are neighborhoods U_j of x_j, and a conformal map Φ of U_2 onto U_1, such that $\Phi(z_2) = z_1$ and $\Phi(U_2 \cap \mathrm{J}) = U_1 \cap \mathrm{J}$.

The *critical points* of a rational map $f : \mathbb{S} \to \mathbb{S}$ are those points where f is not locally injective. At finite points of $\mathbb{C}$ where f is finite, these are the zeros of f'. At poles, they are the points where the pole has order ≥ 2. The number of critical points, counting multiplicity, for f of degree d, is $2d - 2$; see Exercise 8.

Let us say that a point z in $\mathrm{J}(f)$ is *atypical* if every backward orbit contains a critical point; otherwise we say that z is typical. Since there are finitely many critical points, and each orbit is countable, there are only countably many atypical points. It follows from Corollary 3.2.7 that typical points are dense in J.

Theorem 3.2.8. *Suppose that z is a typical point of* $\mathrm{J}(f)$. *Then the set of points z' that have a conformally equivalent location is dense in* $\mathrm{J}(f)$.

Proof. We know from Proposition 3.2.2 (b) that the backward orbit of z is dense in J. Suppose $f^{\circ n}(z') = z$ for some $n \geq 1$. Because there are no critical points in the backward orbit, the derivative $[f^{\circ n}]'(z') \neq 0$. Therefore $f^{\circ n}$ restricted to some neighborhood U' of z' serves as the desired conformal map. □

3.3 Fixed points and periodic points

It is clear from the examples in Section 3.1 that fixed points of f play a key role in the dynamics. A first classification of a fixed point $f(z_0) = z_0$ is based on the *multiplier*. For a finite fixed point z_0,

$$\lambda = f'(z_0).$$

We leave the determination of λ in the case $z_0 = \infty$ to the reader.

A fixed point z_0 is said to be

$$\begin{array}{ll} \textit{attracting} & \text{if } |\lambda| < 1; \\ \textit{super-attracting} & \text{if } \lambda = 0; \\ \textit{repelling} & \text{if } |\lambda| > 1; \\ \textit{neutral} & \text{if } |\lambda| = 1. \end{array}$$

Obviously super-attracting is a special case of attracting. A fixed point that is attracting but not super-attracting is termed *geometrically attractive*.

(It might seem more natural to use a different definition of "attracting fixed point" that does not reference $f'(z_0)$, but see Exercise 10.) The neutral case needs further refinement; see Section 3.5.

In the example (3.1.6), -2 and 3 are repelling fixed points. As noted above, ∞ is a super-attracting fixed point.

Proposition 3.3.1. *(a) If z_0 is an attracting fixed point of of f, then z_0 belongs to the Fatou set* F(f).
(b) If z_0 is a repelling fixed point of f, then z_0 belongs to the Julia set J(f).

Proof. (a) By assumption $f(z_0) = z_0$ and $|f'(z_0)| < 1$. Then there is a bounded neighborhood U of z_0 such that $f : U \to U$. Therefore each iterate maps U into U, so $\{f^{\circ n}\}$ is a normal family on U.

(b) If z_0 belonged to F(f), then a subsequence of $\{f^{\circ n}\}$ would converge uniformly near z_0. In particular, by the Cauchy estimate for derivatives, the derivatives would converge at z_0. But

$$[f^{\circ n}]'(z_0) = f'(z_0)^n \to \infty.$$ □

Corollary 3.3.2. *(a) If $f^{\circ n}(z) = z_0$ is an attracting fixed point of f, then z belongs to the Fatou set.*

(b) If $f^{\circ n}(z) = z_0$ is a repelling fixed point of f, then z belongs to the Julia set.

Periodic points play a similar role. A point $z_1 \in \mathbb{S}$ is said to be *periodic* with *period* $k > 1$ if the partial orbit

$$\{z_0,\ z_1, z_2 \ldots, z_k = z_0\}, \qquad z_j \ = \ f^{\circ j}(z_0), \tag{3.3.1}$$

contains k distinct points. Clearly each z_j is also periodic with period k. Equivalently, z_0 is periodic with period k if z_0 is a fixed point for $f^{\circ k}$, but not for any $f^{\circ j}$, $0 < j < k$. The set (3.3.1) is called a *periodic orbit*, or a *cycle*. In the case (3.3.1), the derivative

$$[f^{\circ k}]'(z_0) \ = \ f'(z_0)f'(z_1)\cdots f'(z_{k-1})$$

is independent of the choice of the point z_j in the periodic orbit. We have the same classification for periodic orbits as for fixed points: if z is periodic with period k and multiplier $\lambda = [f^{\circ k}]'(z)$, then, as in the case of a fixed point, the orbit is said to be

attracting	if $\lvert\lambda\rvert < 1$;
super-attracting	if $\lambda = 0$;
repelling	if $\lvert\lambda\rvert > 1$;
neutral	if $\lvert\lambda\rvert = 1$.

The analogue of Proposition 3.3.1 and its corollary are valid for periodic orbits.

Proposition 3.3.3. *(a) Any attracting or super-attracting periodic orbit of f belongs to* $\mathrm{F}(f)$.
(b) Any repelling periodic orbit of f belongs to $\mathrm{J}(f)$.

Proof. Since elements of periodic orbits are fixed points for some iterate $f^{\circ k}$ of f, this follows from Proposition 3.3.1 and the fact that $\mathrm{F}(f^{\circ k}) = \mathrm{F}(f)$. □

Corollary 3.3.4. *(a) If $f^{\circ n}(z)$ belongs to an attracting or super-attracting orbit of f, then z belongs to* $\mathrm{F}(f)$.
(b) If $f^{\circ n}(z)$ belongs to a repelling orbit of f, then z belongs to $\mathrm{J}(f)$.

3.4 Attracting, super-attracting, and repelling fixed points

We continue to assume that f is rational, $\deg f = d > 1$. The next result goes back to Kœnigs [123] in 1884.

Theorem 3.4.1. *Suppose that z_0 is a geometrically attracting fixed point. Then there is a conformal map ϕ defined on a neighborhood U of z_0 to a disk $D_\rho(0)$ such that $\phi(z_0) = 0$ and*

$$\phi(f(z)) \ = \ \lambda\phi(z), \qquad z \in U. \tag{3.4.1}$$

The map ϕ is unique up to multiplication by a constant.

Proof. For convenience, we conjugate so that $z_0 = 0$. Choose $\delta > 0$ such that f is single-valued on a neighborhood $V \subset \mathbb{D}$ and

$$|f(z) - \lambda z| \leq C|z|^2, \qquad z \in V. \tag{3.4.2}$$

Then $|f(z)| \leq |\lambda z| + C|z|^2 = (|\lambda| + C\delta)|z|$ for $z \in V$, so by choosing V small enough we also have $f(V) \subset V$. Choose $\delta > 0$ such that $U = \phi^{-1}(V) \subset D_\delta(0)$. Let

$$\phi_n(z) = \lambda^{-n} f^{\circ n}(z), \qquad z \in U.$$

Then

$$\phi_n \circ f = \lambda^{-n} f^{\circ(n+1)} = \lambda\phi_{n+1}.$$

If we show that the ϕ_n converge to ϕ on U, it follows that

$$\phi(f(z)) = \lambda\phi(z). \tag{3.4.3}$$

Iterating (3.4.3) gives

$$|f^{\circ n}(z)| \leq \rho^n |z|, \qquad z \in U. \tag{3.4.4}$$

Taking (3.4.2) into account, we get

$$|\phi_{n+1}(z) - \phi_n(z)| = \left| \frac{f(f^{\circ n}(z)) - \lambda f^{\circ n}(z)}{|\lambda|^{n+1}} \right| \leq \frac{C|f^{\circ n}(z)|^2}{|\lambda|^{n+1}} \leq \frac{C\rho^{2n}|z|^2}{|\lambda|^{n+1}}.$$

Thus, choosing $\rho < \sqrt{|\lambda|}$, we get uniform convergence of ϕ_n on U. Since f is injective in $D_\delta(0)$, the same is true of $f^{\circ n}$ and, therefore of ϕ_n. It follows from Proposition 2.3.6 that ϕ is either conformal or constant. But

$$\phi_n'(0) = \lambda^{-n}[f^{\circ n}]'(0) = \lambda^{-n} f'(0)^n = 1,$$

so $\phi'(0) = 1$ and ϕ is conformal.

Finally, suppose that ψ was a second such map. Then

$$\phi \circ f = \lambda\phi, \qquad \psi \circ f = \lambda\psi,$$

so, writing $w = \psi(z)$, we have

$$\lambda\phi \circ \psi^{-1}(w) = \phi \circ \psi^{-1}(\lambda w).$$

Expanding $\phi \circ \psi^{-1}(w)$ gives

$$\lambda(a_1 w + a_2 w^2 + \dots) = a_1 \lambda w + a_2 (\lambda w)^2 + \dots$$

so $a_j = 0$ for $j \geq 1$, and $\phi \circ \psi^{-1}(w) = a_1 w$, □

The *basin of attraction* of an attracting or super-attracting fixed point z_0 is the set $A = A(z_0)$ of points z such that the $f^{\circ n}(z)$ converge to z_0. Because f is continuous on $\mathbb{S}$, this is an open set. The connected component $A_0 = A_0(z_0)$ of A that contains z_0 is the *immediate basin of attraction* of z_0. Clearly A is invariant under f and f^{-1}.

Theorem 3.4.1 can be extended.

Proposition 3.4.2. *The map ϕ of Theorem 3.4.1 can be extended to a holomorphic map on the basin of attraction $A(z_0)$ that intertwines f and multiplication by λ:*

$$\phi(f(z)) = \lambda\phi(z), \qquad z \in A(z_0). \tag{3.4.5}$$

Proof. Formally, (3.4.5) extends to give

$$\phi(z) = \lambda^{-1}\phi(f(z)) = \lambda^{-1}[\lambda^{-1}\phi(f(f(z))) = \ldots = \lambda^{-n}\phi \circ f^{\circ n}(z). \tag{3.4.6}$$

Given $z \in A(z_0)$, $f^{\circ n}(z)$ will eventually belong to the original domain of ϕ. Thus (3.4.6) serves to define ϕ throughout $A(z)$ and to extend the intertwining identity (3.4.5) one step at a time. □

Theorem 3.4.3. *The immediate basin of attraction A_0 of an attracting fixed point z_0 of f contains a critical point of f.*

Proof. If z_0 is super-attracting, then z_0 itself is a critical point. Suppose that z_0 is geometrically attracting and let $\phi : A_0 \to \mathbb{C}$ be the map in Proposition 3.4.2. Then ϕ has a local inverse ψ_0 defined in a disk $D_\rho = D_\rho(0)$. There is some maximal disk D_r such that ψ has a holomorphic extension ψ_r to D_r. Since $\phi^{-1} \circ \psi$ and $\psi^{-1} \circ \phi$ are the identity map near 0 and z_0 respectively, they continue to be the identity up to D_r and $\phi(D_r)$, respectively. It follows from (3.4.5) that

$$f(z) = \psi(\lambda\psi^{-1}(z)), \quad z \in D_r. \tag{3.4.7}$$

Therefore f is injective on D_r. Our standing assumption is that f has degree ≥ 2, hence is not injective on $\mathbb{C}$, so $D_r \neq \mathbb{C}$.

Writing (3.4.7) as

$$f(\psi(w)) = \psi(\lambda w), \quad w \in D_r \tag{3.4.8}$$

we note that the right side is holomorphic in a neighborhood of the closure $\overline{D_r}$. Near any point w_0 of ∂D_r that is not a critical point of f, f has a local inverse g and we get a holomorphic extension $\psi(w) = g(\lambda\psi(w))$ to a neighborhood of w_0. It follows that there is a critical point of f on ∂D_r. Otherwise ψ can be extended across ∂D_r, contradicting the assumption of maximality. (We assumed here that the critical point was not a pole of f. The case of a pole can be dealt with by passing from f to a conformally equivalent function. The same remark applies to the rest of the proof.)

Now (3.4.8) extends by continuity to $w \in \overline{D_r}$, so

$$f^{\circ n}(\psi(w)) = f^{\circ(n-1)}(\psi(\lambda w)) = \cdots = \psi(\lambda^n w) \to \psi(0) = 0, \quad w \in \overline{D_r}. \tag{3.4.9}$$

In particular, the critical points on ∂D_r belong to A_0. □

Remark. The basin of attraction of a periodic orbit of f is defined to be the set of points z such that $\{f^{\circ n}(z)\}$ tends to the orbit. Recall that each periodic orbit corresponds to a fixed point of some iterate of f, and that the Julia set of an iterate is the same as the Julia set of f itself. It follows that each of the preceding results in this chapter has a corresponding extension to attracting periodic orbits.

Moreover, critical points of iterates of f are critical points of f. Therefore Theorem 3.4.3 puts an absolute limit on the total number of attracting points and attracting periodic orbits of f.

Proposition 3.4.4. *Let A be the basin of attraction of an attracting or super-attracting fixed point z_0 of f. Then $\mathrm{J}(f) = \partial A$.*

Proof. Boundary points of A belong to J; see Exercise 16. Thus A is a union of components of F. It is easily seen that A is completely invariant for f, so the result follows from Proposition 3.2.2 (d). □

The following result helps show why Julia sets can be so complicated.

Corollary 3.4.5. *Suppose that f has a repelling fixed or periodic point z_0 with multiplier λ that is not real. Then $\mathrm{J}(f)$ is not contained in a proper smooth submanifold of $\mathbb{S}$.*

Proof. We may assume that z_0 is a fixed point. There is a branch g of f^{-1} defined in a neighborhood of z_0, and z_0 is a fixed point of g with multiplier $1/\lambda$. By Proposition 3.4.2 there is a conformal map ϕ defined in a neighborhood U of z_0 that locally conjugates f^{-1} to g, with $g(\zeta) = \lambda^{-1}\zeta$. Choose $z_1 \in J \cap U$ and let $\zeta_1 = \phi(z_1)$. The points $\zeta_j = g^{\circ j}(\zeta_1)$ lie on a logarithmic spiral centered at $\phi(z_0)$. Thus J is not smooth at z_0.

Since J is not smooth at the center z_0 of the spiral and is not in the exceptional set $E \subset \mathrm{F}$, for any small neighborhood U_1 of z_1, $\bigcup f^{\circ n}(U_1)$ contains z_0. Therefore U_1 contains a preimage $(f^{\circ n})^{-1}$ of a neighborhood of z_0. By shrinking U_1 to z_1 itself, we see that J is not smooth at z_1. □

We end this section with a quick look at the super-attracting case. The analog of Theorem 3.4.1 was proved by Boettcher [28] in 1904.

Theorem 3.4.6. *Suppose that z_0 is a super-attracting fixed point of the rational function f. Then there is a neighborhood of z_0 and conformal map $\phi : U \to D_\rho(0)$ such that $\phi(z_0) = 0$ and*

$$\phi(f(z)) = \phi(z)^p, \qquad z \in U, \tag{3.4.10}$$

where $p \geq 2$. The map ϕ is unique up to multiplication by a $(p-1)$-th root of unity.

Proof. We may translate coordinates so that $z_0 = 0$, and change scale, $z \to cz$, so that in $D_\delta(0)$,

$$f(z) = z^p [1 + \epsilon(z)], \qquad |\epsilon(z)| \le C|z|. \tag{3.4.11}$$

Choose $0 < \delta < 1/2$ so that $|f(z)| \le (2|z|)^p$ if $z \in U = D_\delta(0)$. Then $f : U \to U$ and inductively

$$|f^{\circ n}(z) \le (2|z|)^{p^n}, \qquad z \in U.$$

Let

$$\phi_n(z) = [f^{\circ n}(z)]^{1/p^n} = [z^{p^n}(1 + \dots)]^{1/p^n} = z\,(1 + \dots)^{1/p^n}.$$

This is well defined and conformal in a neighborhood of 0. Then

$$\phi_n \circ f = [f^{\circ n} \circ f]^{1/p^n} = \phi_{n+1}^p,$$

With ϕ_0 the identity map,

$$\phi_n = \prod_{j=1}^{n} \frac{\phi_j}{\phi_{j-1}} \tag{3.4.12}$$

and

$$\begin{aligned}\frac{\phi_{n+1}}{\phi_n} &= \frac{[f \circ f^{\circ n}]^{1/p^{n+1}}}{[f^{\circ n}]^{1/p^n}} \\ &= \frac{\{[f^{\circ n}]^p[1 + \epsilon(f^{\circ n})]\}^{1/p^{n+1}}}{[f^{\circ n}]^{1/p^n}} \\ &= [1 + \epsilon(f^{\circ n})]^{1/p^{n+1}}.\end{aligned}$$

Therefore

$$1 \le \left|\frac{\phi_{n+1}(z)}{\phi_n(z)}\right| \le 1 + |\epsilon(f^{\circ n}(z))| = 1 + O(2|z|^{p^n})$$

so the product (3.4.12) converges uniformly in a neighborhood of 0; see Section 1.6. The product is either conformal or constant. If $\varepsilon(z)$ in (3.4.11) is identically zero then there is nothing more to prove. Otherwise $|\varepsilon(z)| \ge \delta|z|^k$ for z near 0, some positive δ and k, and this can be used to show that the product is not constant.

Finally, suppose that ψ is another such conformal map. Then $\phi \circ \psi^{-1}$ commutes with taking the p-th power. Expanding

$$\phi \circ \psi^{-1}(z) = a_1 z + a_2 z^2 + \dots,$$

we have

$$(a_1 z + a_2 z^2 + \dots)^p = a_1 z^p + a_2 z^{2p} + \dots,$$

so $a_1^p = a_1$ and it is easily seen by induction that $a_n = 0$ for $n > 1$. □

Remark. Since $\log|\phi(f(z))| = p \log|\phi(z)|$, we may extend the harmonic function $G(z) = \log|\phi(z)|$ to the entire basin of attraction $A(z_0)$ by

$$G(z) = p^{-n} G(f^{\circ n}(z)) \tag{3.4.13}$$

for z such that $f^{\circ n}(z)$ is in the domain of ϕ.

3.5 Neutral fixed points

Once again we concentrate on the dynamics of a rational map f of degree $d \geq 2$. Recall that a *neutral* fixed point or periodic point of f is one whose multiplier λhas modulus $|\lambda| = 1$, so $\lambda = e^{2\pi i\theta}$ for some real θ. A fundamental distinction here is whether θ is rational or irrational. A fixed point or periodic point with θ rational is said to be *parabolic*. A more fundamental distinction, from the point of view of dynamics, is whether the point belongs to the Julia set or not.

Proposition 3.5.1. *Any parabolic fixed point or periodic point z_0 of f belongs to* $\mathrm{J}(f)$.

Proof. We know that some $\lambda^j = 1$, so the iterate $g = f^{\circ j}$ has multiplier $\lambda^j = 1$. We are assuming that $d = \deg f \geq 2$. Therefore g has degree d^j, so g is not the identity map. Expanding in an affine coordinate ζ centered at z_0, we have

$$g(\zeta) = \zeta + a_k \zeta^k + a_{k+1}\zeta^{k+1} + \ldots, \qquad a_k \neq 0, \;\; \zeta = z - z_0.$$

Then

$$g^{\circ n}(\zeta) = \zeta + n a_k \zeta^k + O(\zeta^{nk+1}),$$

so $[g^{\circ n}]^{(k)}(0) = n\, k\,!\, a_k \to \infty$ as $n \to \infty$. Therefore z_0 belongs to $\mathrm{J}(g) = \mathrm{J}(f)$. □

Suppose now that z_0 is an irrationally neutral fixed point, i.e. $\lambda = e^{2\pi i\theta}$, with θ not rational. This implies that powers of λ are dense in the unit circle. (In fact a much stronger statement is true: see Exercise 17.) We may assume that $z_0 = 0$. We want to know whether there is a local conjugation by $\zeta = h(z)$ such that in some neighborhood of 0

$$f(h(z)) = h(\lambda z), \qquad h(0) = 0, \;\; h'(0) = 1. \tag{3.5.1}$$

Lemma 3.5.2. *A solution to (3.5.1) in a disk $D_r(0)$, $r > 0$, is injective in the disk.*

Proof. Suppose that $h(z_1) = h(z_2)$. Then

$$h(\lambda z_1) = f(h(z_1)) = f(h(z_2)) = h(\lambda z_2), \ldots, \;\; h(\lambda^n z_1) = h(\lambda^n z_2)$$

for all n. Since the λ^n are dense in the circle, $h(e^{i\psi} z_1) = h(e^{i\psi} z_2)$ for all ψ. Therefore the functions $g_j(w) = h(w z_j)$ agree in the interior of $\mathbb{D}$ as well: $g_1(w) = g_2(w)$ if $|w| < 1$. Then $z_1 = g_1'(0) = g_2'(0) = z_2$. □

Lemma 3.5.3. *A solution to (3.5.1) exists if and only if $\{f^{\circ n}\}$ is uniformly bounded in some neighborhood of* 0.

Proof. If there is such a solution in a neighborhood U of 0, then

$$g(z) = h^{-1} \circ f \circ h(z) = \lambda z, \qquad z \in U,$$

so the functions $g^{\circ n}$ are uniformly bounded in U, and the same is true for the iterates $f^{\circ n} = h \circ g^{\circ n} \circ h^{-1}$.

Conversely, suppose boundedness. Then let

$$\varphi_n(z) = \frac{1}{n} \sum_{j=0}^{n-1} \lambda^{-j} f^{\circ j}(z) = z, \qquad z \in U.$$

Note that

$$\varphi_n'(0) = \frac{1}{n} \sum_{j=0}^{n-1} \lambda^{-j} f'(0)^j = 1.$$

The family $\{\varphi_n\}$ is bounded and

$$\varphi_n \circ f(z) = \lambda \varphi_n + \frac{1}{n} \left(\lambda^{-(n-1)} f^{\circ n}(z) - z \right).$$

Therefore a subsequence converges, and its limit φ can be taken as h^{-1}. In fact $\varphi'(0) = 1$, so $h'(0) = 1$. □

It was shown by Pfeifer [165] in 1917 that there is a choice of θ such that no solution of (3.5.1) exists. In fact it was shown in 1938 by Cremer [49] that no solution exists if $\liminf_{n\to\infty} |1 - \lambda^n|^{1/n} = 0$. The question whether there is a choice of θ such that (3.5.1) *does* have a solution was finally settled by Siegel in 1952. (The question of exactly *which* θ permit a solution was settled by Brjuno [34] (sufficiency) in 1965 and Yoccoz [222] (necessity in 1988.)

Theorem 3.5.4. *There is a $\lambda = e^{2\pi i\theta}$ such that the equation (3.5.1) has no solution for any polynomial f.*

Proof. Let f be a polynomial with $f'(0) = \lambda$: $f(z) = z^d + \cdots + \lambda z$, and suppose that (3.5.1) has a solution h in $D_\delta(0)$. Then $f^{\circ n}(z) = z$ has d^n solutions, the zeros $\{z_j\}$ of

$$0 = f^{\circ n}(z) - z = z^{d^n} + \cdots + (\lambda^n - 1)z = h(\lambda^n(h^{-1}(z))) - z.$$

Since there is only one zero in $D_\delta(0)$, namely $z = 0$, we have

$$|1 - \lambda^n| = \prod_{z_j \neq 0} |z_j| \geq \delta^{d^n}.$$

But suppose that $\lambda = e^{2\pi i\theta}$, where

$$\theta \;=\; \sum_{k=1}^{\infty} 2^{-q_k}, \tag{3.5.2}$$

and $\{q_k\}$ is a strictly increasing sequence of positive integers. Then

$$|1-\lambda^{2^{q_k}}| \;\sim\; 2^{q_k-q_{k+1}}; \tag{3.5.3}$$

see Exercise 18. Then

$$q_{k+1} \;\le\; C(\delta)\, 2^{\delta_k}. \tag{3.5.4}$$

But by choosing q_k to grow rapidly enough, we may be sure that, for any given degree d and any given $\delta > 0$, (3.5.4) eventually fails. □

Note that θ, as constructed here, is an irrational number that can be approximated very efficiently by rationals. Siegel showed that when this condition fails badly, it is possible to solve (3.5.1). A *Diophantine* number is a number $\theta \in \mathbb{R}$ such that for some positive constants c and m,

$$\left|\theta - \frac{p}{q}\right| \;\ge\; \frac{c}{q^m}, \tag{3.5.5}$$

for all integers p, q with $q > 0$. In particular, Liouville showed that if θ is irrational but algebraic, then θ is Diophantine; see Exercise 19.

Let us express (3.5.5) in terms of $\lambda = e^{2\pi i\theta}$. Given an integer $n > 0$, the distance $|\lambda^n - 1|$ is

$$|e^{i\psi}-1| \;=\; 2\sin\left(\frac{\psi}{2}\right), \qquad |\psi| \;=\; \inf_{k\in\mathbb{Z}} |2n\pi\theta - 2k\pi|.$$

Since $|\psi| \le \pi/2$, we have $2|\sin(\psi/2)| \sim |\psi|$, so

$$|\lambda^n - 1| \;\ge c' n^{1-m}.$$

For use in the proof to follow, it will be convenient to rewrite this in the form

$$\frac{1}{|\lambda^n-1|} \;\le\; c_0\,\frac{n^\mu}{\mu\,!}, \tag{3.5.6}$$

where we take μ to be some positive integer.

Theorem 3.5.5. (Siegel) *Suppose that f is holomorphic in a neighborhood of* 0, *and* $f(0) = 0$. *Suppose that the multiplier $\lambda = f'(0)$ is $\lambda = e^{2\pi i\theta}$, where θ is Diophantine. Then there is a solution to (3.5.1) in a neighborhood of* 0.

Proof. We follow the proof as presented in [42]. Given any function h, we will write approximations in a form like $h = k + \hat{h}$, where k is some more-or-less explicit first approximation, and $\hat{h}$ is smaller than h in some suitable sense. In particular we begin with $f(z) \;=\; \lambda z + \hat{f}(z)$, and we want to take $h(z) = z + \hat{h}(z)$ and solve (3.5.1) in the form

$$\hat{h}(\lambda z) - \lambda \hat{h}(z) = \hat{f}(h(z)). \tag{3.5.7}$$

The first step is to look for a conjugation ψ, $\psi(z) = z + \hat{\psi}(z)$, that gives an approximate solution to (3.5.1) in some disk

$$\psi^{-1} \circ f \circ \psi(z) = g(z) = \lambda z + \hat{g}(z), \qquad z \in D_r(0). \tag{3.5.8}$$

The idea is to replace f by g and work on a slightly smaller disk. When $\hat{g} = 0$, we have reached our goal.

Replacing $h(z)$ by $z + \hat{h}(z)$ in the right side of (3.5.7) leads to

$$\hat{\psi}(\lambda z) - \lambda \hat{\psi}(z) = \hat{f}(z) = \sum_{n=2}^{\infty} b_n z^n, \tag{3.5.9}$$

which has the solution
$$\hat{\psi}(z) = \sum_{n=2}^{\infty} \frac{b_n}{\lambda^n - \lambda} z^n.$$

Using this, we want to compare $\hat{g}$, as defined by (3.5.8), to $\hat{f}$ in (3.5.9). For this purpose we use (3.5.6) and the assumptions

$$|\hat{f}'(z) \le \delta, \ \text{ for } z \in D_r(0); \qquad \frac{1}{|\lambda^n - 1|} \le c_0 \frac{n^\mu}{\mu\,!}.$$

This allows us to use Cauchy estimates for the coefficients of $\hat{f}$,

$$|nb_n| \le \frac{\delta}{r^{n-1}},$$

together with (3.5.6) to estimate $\hat{\psi}$ in a smaller disk of radius $(1 - \eta)r$:

$$\begin{aligned}
|\hat{\psi}'(z)| &\le \sum_{n=2}^{\infty} \frac{n|b_n|}{|\lambda^n - \lambda|} r^{n-1}(1-\eta)^{n-1} \\
&\le \frac{c_0 \delta}{\mu\,!} \sum_{n=2}^{\infty} n^\mu (1-\eta)^{n-1} \\
&< c_0 \delta \sum_{n=1}^{\infty} \binom{n+\mu}{\mu} (1-\eta)^n = \frac{c_0 \delta}{\eta^{\mu+1}}.
\end{aligned}$$

Let us assume now that δ is chosen so that $c_0 \delta \le \eta^{\mu+2}$, so

$$|\hat{\psi}'(z)| \le \eta, \qquad z \in D_{(1-\eta)r}(0).$$

This estimate implies that $\psi : D_{(1-3\eta)r}(0) \to D_{(1-4\eta)r}(0)$. The argument principle and the fact that
$$|\psi(z)| \ge (1 - 2\eta)r, \quad z \in \partial D_{(1-\eta)r}(0),$$

while $\psi(z) = 0$ in $D_{(1-\eta)r}(0)$ only for $z = 0$, implies that ψ takes every value in $D_{(1-2\eta)r}(0)$ exactly once in $D_{(1-\eta)r}(0)$.

Now consider

$$g = \psi^{-1} \circ f \circ \psi, \quad z \in D_{(1-4\eta)r}(0).$$

Each factor enlarges the radius by at most ηr. By (3.5.8) and (3.5.9),

$$\begin{aligned} \hat{g}(z) + \hat{\psi}(\lambda z + \hat{g}(z)) &= \lambda\hat{\psi}(z) + \hat{f}(z + \hat{\psi}(z)) \\ &= \hat{\psi}(\lambda z) - \hat{\psi}(\lambda z + \hat{g}(z)) + \hat{f}(z + \hat{\psi}(z)) - \hat{f}(z). \end{aligned}$$

Let $C = \sup\{|\hat{g}(z)|, z \in D_{(1-4\eta)r}(0)\}$. Then

$$\begin{aligned} C &\le \sup|\hat{\psi}'(z)|\, C + \sup|\hat{f}(z + \hat{\psi}(z)) - \hat{f}(z)| \\ &\le \eta C + \delta\, \frac{c_0\delta}{\eta^{\mu+1}}\, r, \end{aligned}$$

which gives

$$C \le \frac{c_0\delta^2 r}{\eta^{\mu+1}} \frac{1}{1-\eta}.$$

Therefore Cauchy's estimate gives

$$|\hat{g}'(z)| \le \frac{c_0\delta^2 r}{\eta^{\mu+2}} \frac{1}{1-\eta}, \qquad z \in D_{(1-5\eta)r}(0). \tag{3.5.10}$$

Recapitulating, we have passed from f with $|\hat{f}'| \le \delta$ in $D_r(0)$ to g with the estimate (3.5.10), using the assumptions: f defined on $D_{r_0}(0)$, $r \le r_0$, and

$$0 < \eta < \frac{1}{5}, \qquad c_0\delta < \eta^{\mu+2}, \qquad \delta < \eta. \tag{3.5.11}$$

If we require that $\eta \le c_1$ for small enough c_1 then $\eta < 1/5$ and the second inequality in (3.5.11) will imply the third. Starting from η_0 and δ_0 that satisfy (3.5.11), we choose inductively

$$\begin{aligned} r_{n+1} &= (1 - 5\eta_n)r_n; \\ \eta_{n+1} &= \tfrac{1}{2}\eta_n = \eta_0 2^{-n-1}; \\ \delta_{n+1} &= c_0\delta_n^2(2\eta_n)^{-(\mu+2)}. \end{aligned}$$

If $c_0\delta_n < \eta_n^{\mu+2}$, then

$$\begin{aligned} c_0\delta_{n+1} &= c_0^2\delta_n^2(2\eta_n)^{-(\mu+2)} \\ &\le (\eta_n^{2\mu+4})\eta_n^{-(\mu+2)}2^{-(\mu+2)} = \eta_{n+1}^{\mu+2}. \end{aligned}$$

In the process we choose functions ψ_n, g_n,

$$g_0 = f, \qquad g_n = \psi_n^{-1} \circ g_{n-1} \circ \psi_n,$$

i.e.

$$g_n = \psi_n^{-1} \cdots \circ \psi_1^{-1} \circ f \circ \psi_1 \circ \cdots \circ \psi_n.$$

The limiting radius of definition is

$$R = r_0 \prod_{n=0}^{\infty} (1 - 5\eta_n) = r_0 \prod_{n=0}^{\infty} \left(1 - \frac{5\eta_0}{2^n}\right) > r_0 e^{-10\eta_0}$$

and

$$|\hat{g}_n'(z)| \leq \frac{\delta_n r_n}{1 - \eta_n} \to 0$$

on $D_R(0)$. Thus $g_n(z) \to \lambda z$ uniformly on $D_R(0)$,

$$\psi_1 \circ \psi_2 \circ \cdots \circ \psi_n \to \psi$$

uniformly, and ψ *conjugates* f to multiplication by λ, i.e.

$$\psi^{-1} \circ f \circ \psi(z) = \lambda z. \qquad \square$$

The connected component of the Fatou set that contains a neutral fixed point for which (3.5.1) has a solution is called a *Siegel disk.*.

A reward for all this effort is Figure 3.4; the larger region on the lower left is a Siegel disk.

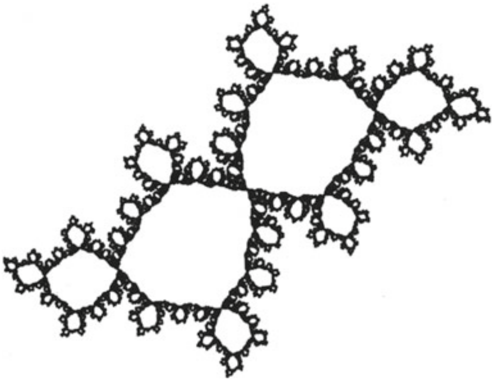

Fig. 3.4 Julia set for $f(z) = z^2 + e^{i2\pi\xi}z$, $\xi = (1/4)^{1/3}$, with a Siegel disk. (Figure reproduced from [141] with the permission of Princeton University Press.).

3.6 Parabolic fixed points

In this section we return for a closer look at how iterates of a rational function f behave near a parabolic fixed point in the Julia set. We assume again that $\deg f \geq 2$.

We begin with the case when the multiplier $\lambda = 1$, and work in a local coordinate with the fixed point at the origin. Then for some n,

$$f(z) \;=\; z[1 + az^n + O(z^{n+1})], \quad a \neq 0.$$

We assume that the multiplicity $n + 1$ of the fixed point is ≥ 2. To get some perspective, suppose that some sequence $\{z_k = f^{\circ k}(z_0)\}$ converges to 0. Then

$$\begin{aligned} z_{k+1} &= z_k \left[1 + az_k^n + \ldots\right] \;\sim\; z_k \left[1 + naz_k^n + \ldots\right]^{1/n} \\ &= z_k \left(1 + \left(\frac{z_k}{\omega}\right)^n + ...\right)^{-1/n}, \end{aligned} \tag{3.6.1}$$

for any choice of ω with $\omega^n = -1/an$. Set $z_k = (c_k)^{-1/n}$. Then (3.6.1), if we ignore the remainder term, leads to the functional equation

$$c_{k+1} \;=\; c_k \left(1 + \frac{1}{c_k}\right) \;=\; c_k + 1$$

with the solution $c_k = k$. Thus we might expect convergent sequences to look like

$$z_k \;\sim\; \frac{\omega}{k^{1/n}}, \qquad \omega^n \;=\; -\frac{1}{na}. \tag{3.6.2}$$

Similarly f^{-1} has a branch g defined near 0 with

$$g(z) \;=\; z[1 - az^n + O(z^{n+1})].$$

Therefore a sequence $\{z'_k = g^{\circ k}(z'_0)\}$ converging to 0 can be expected to have the form

$$z'_k \;=\; \frac{\omega'}{k^{1/n}}, \qquad (\omega')^n \;=\; \frac{1}{na}. \tag{3.6.3}$$

In view of these remarks we refer to the n solutions ω_j, $j = 1, \ldots, n$ of the equation $\omega^n = -1/an$ as *attraction directions* for f at the fixed point, and the n solutions ω'_n of $(\omega')^n = 1/an$ as *repulsion directions* for f at the fixed point.

The preceding construction showed that what behaves nicely under f is not the z_k, but rather

$$c_k \;=\; \left(\frac{\omega}{z}\right)^n \;=\; \frac{c}{z^n}, \qquad c \;=\; -\frac{1}{na}.$$

This suggests a change of variables that puts the fixed point at ∞:

$$w \;=\; \phi(z) \;=\; \frac{c}{z^n}; \qquad z \;=\; \phi^{-1}(w) \;=\; \left[\frac{c}{w}\right]^{1/n}, \tag{3.6.4}$$

for some choice of the branch. Under this coordinate change, the attraction and repulsion directions become

$$\phi(\omega_j) \;=\; 1, \qquad \phi(\omega'_j) \;=\; -1.$$

Under ϕ, the function f is conjugated to F, i.e.

$$\begin{aligned} F(w) &= \phi \circ f \circ \phi^{-1}(w) \\ &= \phi\left(\left[\frac{c}{w}\right]^{1/n}\left[1 + a\frac{c}{w} + O(w^{-1-1/n})\right]\right) \\ &= w\left[1 + a\frac{c}{w} + O(w^{-1-1/n})\right]^{-n} \\ &= w\left[1 - \frac{nac}{w} + O(w^{-1-1/n})\right].. \end{aligned}$$

Since $c = -1/na$,

$$F(w) \;=\; w + 1 + O(|w|^{-1/n}). \tag{3.6.5}$$

Similarly, the branch g of f^{-1} that fixes 0 conjugates to $G = \phi \circ g \circ \psi$ and

$$G(w) \;=\; w - 1 + O(|w|^{-1/n}). \tag{3.6.6}$$

It is especially simple to analyze the behavior in certain sector-like regions. First, we may choose $R > 0$ large enough so that

$$|w| \geq \frac{R}{\sqrt{2}} \quad \text{implies} \quad |F(w) - (w+1)| \leq \frac{1}{2}. \tag{3.6.7}$$

Lemma 3.6.1. *The domain*

$$\Omega_R \;=\; \{w = x + iy \;:\; x + |y| > R\} \tag{3.6.8}$$

is mapped into itself by F. For any $w \in \Omega_R$, $F^{\circ k}(w) \to \infty$..

Proof. If $w = x + iy$ then $2|w|^2 - (x + |y|)^2 \geq (x - |y|)^2$. Therefore $w \in \Omega_R$ implies $|w| \geq R/\sqrt{2}$. Setting $F(w) = x' + iy'$, we have

$$|x' - (x+1)| \;\leq\; \frac{1}{2}, \quad |y' - y| \;\leq\; \frac{1}{2},$$

so $x' + |y'| \geq (x + \frac{1}{2}) + (|y| - \frac{1}{2}) > R$. Moreover

$$\mathrm{Re\ F^{\circ k}(w)} \;\geq\; \mathrm{Re\ w} + \frac{\mathrm{k}}{2},$$

so $F^{\circ k}(w) \to \infty$. □

The various inverse maps ψ can be realized on the complement of the real interval $(-\infty, 0]$ by taking

$$\psi(w) = \omega w^{-1/n}, \qquad w \notin (-\infty, 0], \tag{3.6.9}$$

as ω runs through the roots of $\omega^n = -1/na$.

For each choice of ω with $\omega^n = -1/an$, the map $w \to \psi(w) = \omega w^{-1/n}$, taking the principal branch of the n-th root, takes the domain Ω_R to one petal of a petal-shaped region $P = P_R$ as shown in Figure 3.5. The angle of each petal at the origin is π/n. The arrows on the left indicate the direction of travel under F. It is not difficult to see that any orbit in Ω_R maps to a sequence in P that converges to 0 along a curve that is tangent to the corresponding petal at the origin. More precisely:

Lemma 3.6.2. *Suppose that* $w_0 \in \Omega_R$. *Let* $w_k = F^{\circ k}(w_0)$ *and* $z_k = \omega w_k^{1/n}$, $\omega^n = -1/an$, *Then*

$$z_k = \omega k^{-1/n}(1 + o(1)). \tag{3.6.10}$$

Proof. Note that for any fixed m, $(k+m)^{1/n} = k^{1/n} + O(m/k)$, so we may replace z_0 by any later point in the orbit and number from there. In particular, we may assume that we have reached the region where $|F(w)| \geq |w|$. Moreover, given $\varepsilon > 0$, we may assume that we have reached the region where

$$F(w) = w + 1 + \eta(w), \qquad |\eta(w)| \leq \varepsilon.$$

Then, renumbering the sequence from here,

$$w_k = w_0 + k + \eta_k, \qquad |\eta_k| \leq k\varepsilon.$$

Therefore

$$z_k = w_k^{-1/n} = k^{-1/n}\left[1 + \frac{w_0}{k} + \frac{\eta_k}{k}\right]^{-1/n} = k^{-1/n}[1 + O(\varepsilon)].$$

Since ε is arbitrary, this proves (3.6.10). □

Under various choices of the inverse map ψ given by (3.6.9), the petal P is mapped by F to rotations of a scaled copy of P. The arrows on the right in Figure 3.5 indicate the direction of travel under f. Similarly, the maps

$$w \to \omega' w^{-1/n}, \qquad (\omega')^n = \frac{1}{2an} \tag{3.6.11}$$

take P to rotations of a scaled copy of P. Reversing the arrows in Figure 3.5 shows the direction of travel under G and g.

Putting all this together yields the *Leau–Fatou flower*, a covering of a neighborhood of the parabolic fixed point by overlapping petals that alternate between attraction and repulsion directions. The case $n = 3$ is indicated in Figure 3.5. The arrows indicate the direction of travel of points under f.

It is clear from this that any z with the property that the orbit of z enters P_j approaches the fixed point 0, in the limit, from the direction ω_j of the axis of symmetry

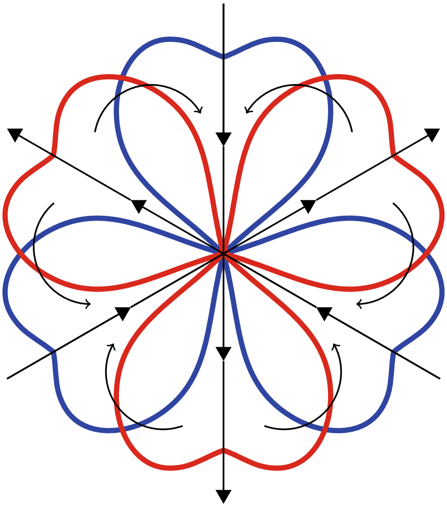

Fig. 3.5 Leau–Flatou flower, $n = 3$.

of P_f. The set of such points is denoted A_j. Clearly, each A_j belongs to the Fatou set, as does the entire basin of attraction of 0.

Proposition 3.6.3. *The boundary $\partial A_j(0)$ of each basin of attraction $A_j(0)$ belongs to the Julia set.*

Proof. Consider the orbit $z_0 \to z_1 \to \ldots$ of a point $z_0 \in \partial A_j$ If the orbit reaches 0 in finitely many steps, then since $0 \in$ J it follows that $z_0 \in$ J. Now z_0 is not in any of the A_j, so it does not converge to the fixed point 0. Therefore there is a subsequence that is bounded away from 0. But $f^{\circ k}$ converges to 0 at each point of A_j, so $\{f^{\circ k}\}$ cannot be a normal family in any neighborhood of 0. Therefore $z_0 \in$ J. □

It is easily seen, in light of this discussion, and taking account of Lemma 3.6.2, that we have confirmed (3.6.2) and (3.6.3):

Proposition 3.6.4. *(a) Suppose that a sequence $\{z_k = f^{\circ k}(z_0)\}_{k=0}^{\infty}$ converges to* 0 *and no $z_k = 0$. Then for some j,*

$$\lim_{k\to\infty} k^{1/n} z_k = \omega_j.$$

(b) Suppose that a sequence $\{z'_k = g^{\circ k}(z'_0)\}_{k=0}^{\infty}$ converges to 0*, where $g = f^{-1}$, and no $z'_k = 0$. Then for some j*

$$\lim_{k\to\infty} k^{1/n} z'_k = \omega'_j.$$

All attraction and repulsion directions occur.

Remarks. 1. The flower grew through work of Leau [130], Julia [118], and Fatou [69].

2. To this point we have been considering only the case of a parabolic fixed point z_0 with multiplier 1. Suppose that the multiplier $\lambda \neq 1$, $\lambda^m = 1$. Then z_0 is a fixed

point of $f^{\circ m}$ with multiplier 1. Similarly, any point in a parabolic periodic orbit is a parabolic fixed point of some iterate of f, and therefore a fixed point with multiplier 1 of some further iterate. Recall that the Fatou and Julia sets are unchanged under iteration.

3.7 Perspectives: classification and the Mandelbrot set

One long-time goal of the theory is to understand the connected components of the Fatou set of a rational f of degree ≥ 2. In principle there are three possibilities for such a component U: it might be *periodic*: $f^{\circ m}(U) = U$ for some minimal $m \geq 1$, or it might be *pre-periodic*: some $f^{\circ k}(U)$, $k \geq 1$ is periodic, or it might be *wandering*: the images $\{f^{\circ n}(U)\}$ are all distinct. Much of the progress was made by Fatou and Julia. The final pieces were supplied by Siegel, Sullivan, and Herman.

Sullivan [198] introduced methods of quasiconformal mapping to prove in 1985 that for rational f there are no wandering components. In 1984 Herman [105] produced a new type, now known as a *Herman ring*. By definition this is a periodic component U of the Fatou set that is doubly connected and $f^{\circ n}|_U$ is conjugate to either a rotation on an annulus or a rotation followed by an inversion. Figure 3.6 shows the Julia set of a cubic rational map that contains a Herman ring.

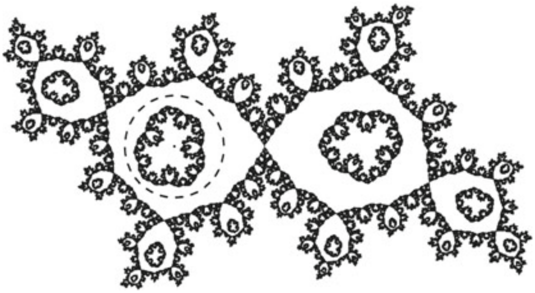

Fig. 3.6 Julia set that contains a Herman ring. Figure reproduced from [141] with the permission of Princeton University Press.

Theorem 3.7.1. (Classification Theorem) *If U is a periodic component of the Fatou set of a rational function f, then either*

(a) U contains an attracting or super-attracting fixed point or periodic point;

(b) U is the basin of attraction of a parabolic fixed point with multiplier 1;

(c) U is a Siegel disk; or

(d) U is a Herman ring.

For the proofs of Sullivan's non-wandering theorem and the existence of Herman rings, and for a thorough discussion of components of the Fatou set, see Chapter IV of Carleson and Gamelin [42].

Finally, we discuss dependence on parameters and the Mandelbrot set. We saw in Section 3.1 that even for quadratics $f(z) = z^2 + c$, the Julia set varies considerably with c. In this case the critical points are 0 and ∞, and the fixed points are the solutions of $z^2 - z + c = 0$.

Proposition 3.7.2. *If $f(z) = f_c(z) = z^2 + c$, then $\mathrm{J}(f)$ is connected if and only if $\{f^{\circ n}(0)\}$ is bounded.*

Proof. By the maximum principle, iterates of f are bounded on bounded components of the Fatou set F, so the basin of attraction A of ∞ is connected. By Theorem 3.4.6 there is a conformal map ϕ defined in a neighborhood of ∞ such that $\phi(z) = z + O(1)$ and

$$\phi(f(z)) = \phi(z)^2; \qquad \log|\phi(f(z))| = 2\log|\phi(z)|.$$

By the remark following Theorem 3.4.6, the harmonic function $G(z) = \log|\phi(z)|$ can be extended to A. We may define $U_r = \{z : G(z) > r\}$. Then $f : U_r \to U_{2r}$. For sufficiently large r, ϕ is defined on U_r. We may extend ϕ further by

$$\phi(z) = [\phi(f(z))]^{1/2}, \qquad z \in U_{r/2},$$

so long as $U_{r/2}$ does not contain the critical point 0 of f. This extension is injective on $U_{r/2}$. The process can be continued as long as we do not reach 0.

Therefore if $\{f^{\circ n}(0)\}$ is bounded, i.e. $0 \notin A(\infty)$, it follows that ϕ extends to all of A, A is simply connected, and its boundary ∂A is connected. But by Proposition 3.4.4, $\partial A = \mathrm{J}$.

Suppose that $f^{\circ n}(0) \to \infty$. We shall show that J is totally disconnected. We know that J is bounded. choose a disk $U \supset \mathrm{J}$, and choose N so that $f^{\circ n}(0)$ is not in the closure $\overline{U}$ for $n \geq N$. Given $z_0 \in \mathrm{J}$, there is an N such that for $n \geq N$ there is a branch g_n of $(f^{\circ n})^{-1}$, holomorphic on U, with $g_n(f^{\circ n}(z_0)) = z_0$. The functions $\{g_n\}$ are a normal family. Any limit point of $\{g_n(z)\}$, $z \in A(\infty) \cap U$ belongs to J, so any limit of a subsequence of the $\{g_n\}$ maps $A(\infty) \cap U$ into J. Since J contains no open sets of $\mathbb{C}$, the limit is constant. Therefore the diameter of $g_n(\overline{U})$ tends to 0. Since $g_n(\partial U)$ is disjoint from J, it follows that $\{z_0\}$ is a connected component of J. □

The set of c such that $z^2 + c$ has a connected Julia set,

$$\mathrm{M} = \{c : \sup|f_c^{\circ n}(0)| < \infty, \ f_c(z) = z^2 + c\}$$

is called the *Mandelbrot set*, Figure 3.7. It was studied also by Brooks and Matelski [33], but it was the computer images in Mandelbrot [139] that showed the complexity of this set and made it famous as a "fractal." The term "Mandelbrot set" is due to Douady and Hubbard [59]. Douady and Hubbard [58] proved that M is connected.

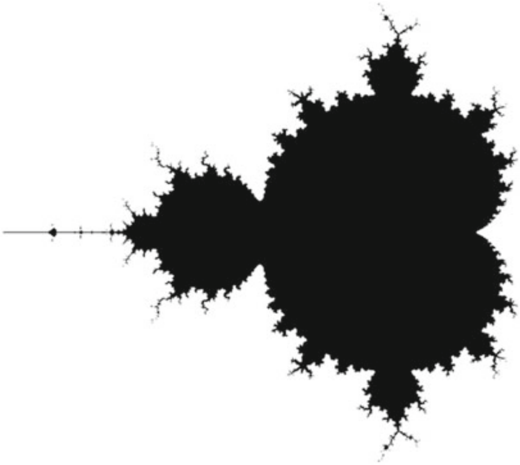

Fig. 3.7 The Mandelbrot set.

Theorem 3.7.3. *The Mandelbrot set is a closed, simply connected subset of the closed disk $\overline{D_2(0)}$, consisting of those c such that*

$$|f_c^{\circ n}(0)| \;\le\; 2, \qquad n = 1, 2, 3, \dots\ . \tag{3.7.1}$$

Moreover

$$\mathrm{M} \cap \mathbb{R} \;=\; [-2, 1/4]. \tag{3.7.2}$$

Proof. Suppose $|c| > 2$. Then $|f_c(0)| \ge |c|^2 - |c| = |c|(|c| - 1)$ and by induction

$$|f_c^{\circ n}(0)| \;\ge\; |c|(|c| - 1)^{2^{n-1}} \;\to\; \infty,$$

so $\mathrm{M} \subset \overline{D_2(0)}$.

Suppose $m \ge 1$ and $f_c^{\circ m}(0)| = 2 + \delta > 2$. If $|c| > 2$, then $c \notin \mathrm{M}$. If $|c| \le 2$, then $f_c^{\circ(m+1)}(0) \ge 2 + 4\delta$, and inductively

$$|f_c^{\circ(m+k)}(0)| \;\ge\; 2 + 4^k\delta \;\to\; \infty,$$

so again $c \notin \mathrm{M}$. This proves that $c \in \mathrm{M}$ satisfies (3.7.1), and this characterization implies that M is closed. Then $\mathbb{C} \setminus \mathrm{M}$ is open. For any $n \ge 1$, $f_c^{\circ n}(0)$ is holomorphic in c, so if (3.7.2) is true for c in some open set, it is true on the boundary. Therefore $\mathbb{C} \setminus \mathrm{M}$ has no bounded components, so M is simply connected.

Any finite limit point z of $f_c^{\circ m}(0)$ would be a fixed point of f_c. For $c > 1/4$, f_c is strictly increasing on $[0, \infty)$ and has no real fixed points, so $f_c^{\circ n}(0) \to \infty$ and $c \notin \mathrm{M}$. We know that $c \notin \mathrm{M}$ if $c < -2$, so suppose that $-2 \le c \le 1/4$. Let a be the larger of the two real roots of $f_c(z) = z$,

$$a \;=\; \frac{1}{2} + \frac{1}{2}\sqrt{1 - 4c}.$$

Note that $a^2 + c = a$ and that $|c| = |f_c(0)| \leq a$. Inductively,

$$|f_c^{\circ(n+1)}(0)| \;=\; |[f_c^{\circ n}]^2 + c| \leq |a^2 + c| = a,$$

so $c \in \mathrm{M}$. □

The next result accounts for the large smooth blob that is the dominant part of M and shows that the smaller pieces immediately next to the blob are tangent to it.

Theorem 3.7.4. M *contains the cardioid*

$$\mathrm{C} = \left\{ \frac{2\lambda - \lambda^2}{4} \;:\; |\lambda| < 1 \right\}. \tag{3.7.3}$$

Moreover $\partial C \subset \partial \mathrm{M}$.

Proof. . As noted before, a fixed point z_c for f_c is $z_c = \frac{1}{2} \pm \frac{1}{2}\sqrt{1-4c}$. The multiplier is therefore $\lambda = 2z_c = 1 \pm \sqrt{1-4c}$, so the parameter associated to a given multiplier λ is

$$c = c(\lambda) = \frac{2\lambda - \lambda^2}{4}.$$

Therefore f_c has a (finite) attracting fixed point z_c if and only if $c \in \mathrm{C}$. If so, then by Theorem 3.4.3, the immediate basin of attraction $A_0(z_c)$ contains the unique finite critical point 0. Thus $f_c^{\circ n}(0) \to z_c$, so $c \in \mathrm{M}$.

Let Ω be the component of the interior of M that contains C. Now $f_c^{\circ n}(0)$ is a polynomial $P_n(c)$, and $\{P_n\}$ is a normal family on Ω. On the interior of C the P_n converge to the fixed point z_c, so by analyticity this is true on Ω. If $\zeta \notin \overline{C}$ then $|\lambda| > 1$, so $P_n(c)$ cannot converge to z_c unless $P_n(c) = z_c$ for sufficiently large z. But for each n, $P_{n+1}(c) = P_n(c)$ has only finitely many solutions, so the set of such c is at most countable. Therefore $\Omega = C$. □

It is not difficult to account for the next most obvious feature on M, the disk-like piece to the left of the main cardioid; see Exercise 20. This suggests a way to account for further features of M as well.

For much more information about quadratic polynomial dynamics and the structure of M to see Chapter VIII of Carleson and Gamelin [42] and the references there.

Exercises

1. Determine the Julia set for each of the cases $f \in \mathrm{Aut}(\mathbb{S})$ discussed in the introduction.
2. Verify that any quadratic can be put into a form (3.1.1).

3. Show that (3.1.4) satisfies (3.1.2).
4. Verify (3.1.5).
5. Verify Proposition 3.1.1
6. Show that the points in the Cantor set associated with (3.1.6) are the points that have the form

$$x = \sqrt{6 + \eta_1\sqrt{6 + \eta_2\sqrt{6 \pm \ldots}}}, \qquad \eta_k = \pm 1,$$

for any choice of the signs η_k.

7. Verify the assertions in Proposition 3.2.2 for the examples in (3.1.4).
8. Prove that the number of critical points of a rational function of degree d is $2d - 2$.
9. Prove that if $\{U_n\}$ is a sequence of dense open sets in a complete metric space X, then $\bigcap U_n$ is dense in X.
10. Suppose that f is a rational map, $\deg f \geq 1$, and suppose that z_0 is a fixed point with the property that some subsequence of $\{f^{\circ n}\}$ converges to z_0 uniformly in some neighborhood of z_0. Prove that $|f'(z_0)| < 1$.
11. Suppose f is rational, degree ≥ 2, with a fixed point at infinity. Determine the multiplier λ.
12. Show that for any $n > 0$, there is a c such that $f(z) = z^2 + c$ has a parabolic fixed point with multiplier $\lambda = \exp(2\pi i/n)$.
13. Show that a fixed point z_0 of a rational map has the Liapunov stability property—that for any $\varepsilon > 0$, if z is sufficiently close to z_0 then $|f^{\circ n}(z) - z_0| < \varepsilon$ for all $n \geq 0$—if and only if z_0 belongs to $\mathrm{F}(f)$.
14. *Newton's method* for approximating the roots of a real polynomial P: given a point $x \in \mathbb{R}$, go to the point $(x, P(x))$, and follow the tangent line to a the point x' where it meets the real axis.

(a) Show that $x' = x - P(x)/P'(x)$.

(b) Given a general complex polynomial P, let f be the rational function

$$f(z) = z - \frac{P(z)}{P'(z)}.$$

Show that the fixed points of f are ∞, which is repelling, and the zeros of P, which are attracting.

(c) Suppose that P is a quadratic with two distinct zeros. Show that $\mathrm{J}(f)$ consists of a straight line and ∞.

(d) Determine the Julia set of a quadratic with a double zero.

15. Given $|a| < 0$, the linear fractional transformation

$$b_a(z) = \frac{z - a}{1 - \bar{a}z}$$

takes the unit disk $\mathbb{D}$ to itself. A *Blaschke product* is a product $f = \omega \prod_{j=1}^{m} b_{a_j}$, where $|\omega| = 1$.

(a) Show that $\mathrm{J}(f)$ is a subset of $\{z : |z| = 1\}$.

(b) Suppose that one of the factors is b_0. Show that 0 and ∞ are attracting fixed points and $\mathrm{J}(f)$ is the entire unit circle.

16. Suppose that z lies in the closure of the basin of attraction A of an attracting fixed point or attracting periodic orbit of a rational function f. Prove that if $z \in \mathrm{F}(f)$, then $z \in A$.
17. Suppose that $\theta \in \mathbb{R}$ is irrational. A result of Kronecker says that the powers of $\omega = e^{2\pi i\theta}$ are *equidistributed* in the unit circle $\partial\mathbb{D}$. An equivalent formulation is in terms of the fractional parts $\{2n\pi\theta\}$. Here $\{x\}$ is defined to be $x - m$, where m is the integer such that $m \le x < m + 1$. Then the theorem says that for any subinterval $[a, b) \subset [0, 1)$, as $N \to \infty$, the average number of values $\{2n\pi\theta\}$, $|n| \le N$, (counting multiplicity) that lie in $[a, b)$ approaches $b - a$. This assertion has a number of equivalent formulations. First, let f be the characteristic function of the interval $[a, b)$: $f(x) = 1$ if $x \in [a, b)$, otherwise $f(x) = 0$. Then

$$\lim_{N\to\infty} \frac{1}{2N+1} \sum_{-N}^{N} f(\{2n\pi\}) = \int_a^b f(x)\,dx. \tag{3.7.4}$$

Second, (3.7.4) is true for all such characteristic functions if and only if it is true for all real linear combinations of such functions. Show that, in turn, (3.7.4) is true for all such combinations if and only if (3.7.4) is true for all continuous functions $f : [0, 1) \to \mathbb{R}$. Use the Weierstrass polynomial approximation theorem (see the Remark after Corollary 5.1.8) to conclude that (3.7.4) is true for all such continuous functions f if and only if it is true for each power function $f_m(x) = x^m, m = 0, 1, 2, \ldots$. Finally, prove Kronecker's theorem by verifying that (3.7.4) is indeed true for each power f_m. (It is only at this last step that we use the assumption that θ is irrational.)
18. Suppose that $\{q_k\}$ is an increasing sequence of positive integers. Prove that (3.5.4) implies (3.5.2).
19. Suppose that θ is an algebraic irrational: i.e. θ is a zero of a polynomial P with integer coefficients, and the minimum degree of such a polynomial of degree $m > 2$. Prove Liouville's result that there is a constant $c > 0$

$$\left|\theta - \frac{p}{q}\right| \ge \frac{c}{q^m}$$

for every pair of integers p, q, $q > 0$. (Suppose that $P(\theta) = 0$, where P has integer coefficients and has minimal degree m. Then $P'(\theta) \neq 0$ and

$$\left|q^m P\left(\frac{p}{q}\right) - P(\theta)\right|$$

is a non-zero integer. By the Mean Value Theorem, this expression is

$$q^m |P'(\theta^*)| \left|\frac{p}{q} - \theta\right|$$

for some θ^* between θ and p/q.)

20. Let $f_c(z) = z^2 + c$. Find the attracting fixed points of $f_c \circ f_c$ that are not fixed points of f_c. Hint: the solutions of $f_c(z) = z$ are solutions of $f_c \circ f_c(z) = z$, so $f_c(z) - z$ divides $f_c \circ f_c(z) - z$:

$$f_c \circ f_c(z) - z \;=\; [f_c(z) - z][g(z)],$$

where g is a quadratic in z. Compute g, use $g(z) = 0$ to show that the multiplier $\lambda = (f_c \circ f_c)'$ satisfies $|\lambda| < 1$ if and only if $|c + 1| < 1/4$. Adapt the argument in Theorem 3.7.4 to show that the disk $|c + 1| < 1/4$ is contained in the Mandelbrot set M, and that the boundary of the disk is contained in the boundary of M.

Remarks and further reading

The second flourishing of complex dynamics in the 20th century was celebrated in a number of expositions, including books by Beardon [23] and Steinmetz [197] and a review article by Lyubich [138]. Our presentation here relied mainly on the systematic and thorough treatment by Carleson and Gamelin [42], and the discursive and profusely illustrated notes of Milnor [141]. The book by Carleson and Gamelin contains proofs of all major results, that by Milnor is particularly complete on historical detail and anecdote. Both treat topics beyond rational dynamics, such as entire functions and dynamics on Riemann surfaces, as well as a more thorough treatment of polynomial dynamics. Sullivan's work and subsequent developments have made use of quasiconformal mapping and methods of Teichmüller theory; see the exposition by Shishikura in one of the supplementary chapters in the second edition of Ahlfors's lectures [5].

Chapter 4
Univalent functions and de Branges's theorem

The Riemann mapping theorem says that any simply connected domain $\Omega \subset \mathbb{C}$ that is not all of $\mathbb{C}$ can be mapped conformally onto $\mathbb{D}$. Moreover if we fix $\phi(0) \in \Omega$ and require $\phi'(0) > 0$, then the conformal map $\phi : \mathbb{D} \to \Omega$ is unique.

The study of *univalent* (injective) functions turns this around, by looking at injective holomorphic functions $f : \mathbb{D} \to \mathbb{C}$. Here the usual normalization is $f(0) = 0$, $f'(0) = 1$. This fixes the position, orientation, and scale of the image $f(\mathbb{D})$. Then the series expansion of f at 0 has the form

$$f(z) \;=\; z + a_2 z^2 + a_3 z^3 + \cdots + \;. \tag{4.0.1}$$

In principle, the coefficients $\{a_n\}$ encode all the information about the conformal image $\Omega = f(\mathbb{D})$. The set of such normalized conformal maps of $\mathbb{D}$ is denoted S. (The S stands for the German *schlicht*, meaning "simple." The functions in S are often called "schlicht functions.")

A particularly important example comes about as follows. The linear fractional transformation

$$h(z) \;=\; \frac{1+z}{1-z}$$

maps $\mathbb{D}$ to the right half-plane $\{z : \mathrm{Re}\, z > 0\}$. Therefore h^2 maps $\mathbb{D}$ to the complement of the half-line $\{x : x \le 0\}$. We can adjust this to get a function in S by a translation and dilation. The result is the *Koebe function*

$$\begin{aligned} K(z) = \frac{1}{4}[h(z)^2 - h(0)^2] \;&=\; \frac{z}{(1-z)^2} \\ &= z + 2z^2 + 3z^3 + 4z^4 + \ldots \;. \end{aligned} \tag{4.0.2}$$

The image of $\mathbb{D}$ under K is the complement of the half-line $\{x : x \le -1/4\}$; see Exercise 1. More generally, given $\theta \in \mathbb{R}$, define the rotated Koebe function

$$K_\theta(z) \;=\; e^{-i\theta} K(e^{i\theta} z) \;=\; z + \sum_{n=2}^{\infty} a_n z^n, \quad a_n = n\, e^{i(n-1)\theta}. \tag{4.0.3}$$

R. Beals and R. S. C. Wong, *More Explorations in Complex Functions*, Graduate Texts in Mathematics 298, https://doi.org/10.1007/978-3-031-28288-1_4

Then $|a_n| = n$, and the complement of the image $K_\theta(\mathbb{D})$ is a rotation around the origin of the half-line $\{x : x \le -1/4\}$.

Koebe proved that there is a $\delta > 0$ such that for each $f \in S$, the image $f(\mathbb{D})$ necessarily contains the disk $D_\delta(0)$. The example $f = K$ shows that $\delta \le 1/4$, and Koebe conjectured that $\delta = 1/4$ is sharp, i.e. $f \in S$ implies $f(\mathbb{D}) \supset D_{1/4}(0)$. This is correct. The result is referred to as "Koebe's one-quarter theorem," although it was Bieberbach who proved it. Given that any such image must contain $D_{1/4}(0)$ and must be simply connected, we might consider the "largest" possible such domain to be one whose complement is a ray running to ∞ from a point a with $|a| = 1/4$—in other words, the image of a Koebe function. As a result, we might suspect that for any univalent function with expansion (4.0.1), the coefficients must satisfy $|a_n| \le n$, with equality only for the Koebe functions.

Bieberbach [27] proved in 1916 that this is true for $n = 2$, i.e. $|a_2| \le 2$, with equality only for the Koebe functions. He went on to pose the full *Bieberbach conjecture*: each coefficient in the expansion (4.0.1) of a normalized conformal map f of $\mathbb{D}$ into $\mathbb{C}$ satisfies

$$|a_n| \le n, \tag{4.0.4}$$

with equality only for the Koebe functions.

Proving (or disproving) the Bieberbach conjecture was the outstanding challenge of research on univalent functions for much of the 20th century. Aside from some special cases, progress was mainly made on one coefficient at a time. The full conjecture was finally proved by de Branges in 1984 [52].

In Section 4.1 we give Bieberbach's proof of the result for $n = 2$ and derive some of the important consequences for the general theory. These include the proof of the one-quarter theorem, and proofs of the growth and distortion theorems of Koebe.

The rest of this chapter is devoted to the theory that leads up to the proof of the full Bieberbach conjecture. Section 4.2 begins with an outline of progress before 1984, and its relation to the proof of the full conjecture. Section 4.3 introduces *slit mappings* and gives Carathéodory's proof that they are dense in S. To prove Bieberbach's conjecture it is enough to prove it for slit mappings. Section 4.4 treats Loewner's theory of slit mappings, which enabled him to prove the conjecture for the third coefficient.

Section 4.5 introduces the conjectures of Robertson and of Milin. It is shown that the Robertson conjecture implies the Bieberbach conjecture and that the Milin conjecture implies the Robertson conjecture.

The achievement of de Branges was to prove the Bieberbach conjecture by proving the Milin conjecture. In Section 4.7 we give an expanded version of Weinstein's later, very condensed, proof of the Milin conjecture [213].

Weinstein's proof relies on some particular facts from the Loewner theory, on the generating function representation of the Legendre polynomials, and on Legendre's addition formula. In the preparatory section, Section 4.6, we develop the needed version of the Loewner theory. We also derive some general properties of the Legendre polynomials from the generating function representation, and outline the relation to the addition theorem and to its its interpretation in connection with surface harmonics.

Some further history is outlined in the section on remarks and further reading.

4.1 Bieberbach's theorem and some consequences

As noted in the introduction to this chapter, the term *univalent* means injective, i.e. not taking the same value twice. In complex function theory the term is primarily used for holomorphic functions, also (especially in the older literature) called *schlicht functions*.

If $g : \mathbb{D} \to \mathbb{C}$ is univalent, then there is a unique translation $h(z) = az + b$ such that $f = h \circ g$ satisfies the normalization conditions

$$f(0) = 0, \qquad f'(0) = 1. \tag{4.1.1}$$

As we noted in the introduction, the set of univalent maps f that are defined on $\mathbb{D}$ and satisfy (4.1.1) is denoted S.

If f belongs to S, then the function

$$g(z) = \frac{1}{f(1/z)} = z\left[1 + a_2 z^{-1} + a_3 z^{-2} + \ldots\right]^{-1}, \quad |z| > 1, \tag{4.1.2}$$

is univalent and has a simple pole at ∞. Let us consider functions of this type. By a translation we may get rid of the constant term in the expansion and have

$$h(z) = z + b_1 z^{-1} + b_2 z^{-2} + b_3 z^{-3} + \ldots, \quad |z| > 1. \tag{4.1.3}$$

A key result due to Gronwall [93] is known as the *Gronwall area theorem*.

Theorem 4.1.1. *(*Area Theorem*) If a function h given by the formula (4.1.3) is univalent, then*

$$\sum_{n=1}^{\infty} n|b_n|^2 \le 1. \tag{4.1.4}$$

Proof: For $r > 1$, let E_r be the complement of the image $\{h(z) : |z| \ge r\}$, and let

$$\Gamma_r = \{h(z) : |z| = r\}.$$

Univalence implies that Γ_r is a simple closed curve that encloses E_r. By (1.2.10), the area of E_r is

$$A_r = \frac{1}{2i}\int_{\Gamma_r} \overline{h(z)}h'(z)\,dz$$

$$= \frac{1}{2i}\int_{\Gamma_r}\left[\bar{z}+\sum_{n=0}^{\infty}\bar{b}_n\bar{z}^{-n}\right]\left[1-\sum_{m=0}^{\infty}mb_m z^{-m-1}\right]dz.$$

On Γ_r, write $z = re^{i\theta}$ so that the integral becomes

$$\frac{1}{2}\int_0^{2\pi}\left[re^{-i\theta}+\sum_{n=0}^{\infty}\bar{b}_n r^{-n}e^{in\theta}\right]\left[re^{i\theta}-\sum_{m=1}^{\infty}mb_m r^{-m}e^{-im\theta}\right]d\theta.$$

The series converge uniformly, so we may take the product and integrate term-by-term. Since

$$\frac{1}{2}\int_0^{2\pi}e^{ip\theta}\,d\theta = \begin{cases}\pi & \text{if } p=0,\\ 0 & \text{if } p=\pm1,\pm2,\ldots\,.\end{cases} \tag{4.1.5}$$

it follows that

$$A_r = \pi\left[r^2-\sum_{n=1}^{\infty}n|b_n|^2r^{-2n}\right].$$

Letting r decrease to 1, the limit of the left side is the outer measure $m^*(E)$ of the complement E of the image of the map h. Therefore

$$0 \le m^*(E) = \pi\left[1-\sum_{n=1}^{\infty}n|b_n|^2\right]. \qquad \square$$

In particular, equality holds in (4.1.4) if and only if the complement of the image of h has measure zero. If $h = g$ has the form (4.1.2), where f belongs to S, then this is equivalent to saying that the complement of the image of f has measure zero.

Corollary 4.1.2. *If h given by (4.1.3) is univalent then for each n, $|b_n|^2 \le 1/n$.*

The remaining ingredient in Bieberbach's proof of his conjecture in the case $n = 2$ is the square root transformation. Suppose f belongs to S. Then

$$f(z^2) = z^2\left[1+\sum_{n=1}^{\infty}a_{n+1}z^{2n}\right], \quad |z|<1.$$

By assumption, f is univalent, so the term in brackets is never 0. Therefore we may choose a branch of the square root that is 1 at $z = 0$ and define

$$f_2(z) \equiv f(z^2)^{1/2} = z\left[1+\frac{a_2}{2}z^2+\ldots\right], \quad |z|<1. \tag{4.1.6}$$

The function f_2 is single-valued (Exercise 2) so it belongs to S.

Theorem 4.1.3. (Bieberbach) *If f belongs to S and has the expansion (4.1.2), then $|a_2| \le 2$. The equality is strict unless $f = K_\theta$ for some θ.*

Proof: Let

$$g(z) = \frac{1}{f_2(1/z)} = \frac{1}{f(1/z^2)^{1/2}} = z - \frac{a_2}{2} z^{-1} + \dots . \tag{4.1.7}$$

By Corollary 4.1.2, $|a_2| \le 2$. Equality implies that the remaining coefficients are zero, so

$$g(z) = z - \frac{e^{i\theta}}{z}, \tag{4.1.8}$$

for some $\theta \in \mathbb{R}$. It follows that $f = K_\theta$. □

We are also in a position to prove Koebe's conjecture.

Theorem 4.1.4. (One-quarter theorem) *If f belongs to S, then the image $f(\mathbb{D})$ contains $D_{1/4}(0)$, the disk of radius $1/4$ centered at the origin.*

Proof: Suppose that f omits the value w. The function

$$g(z) = \frac{w\, f(z)}{w - f(z)}$$

is the composition $h \circ f$ of a linear fractional transformation with f, so it is also univalent and is easily seen to belong to S. The coefficient of z^2 in its expansion is

$$\frac{g''(0)}{2} = a_2 + \frac{1}{w},$$

where a_2 is the coefficient of x^2 in the expansion of f. Therefore

$$\left|\frac{1}{w}\right| = \left|\left(a_2 + \frac{1}{w}\right) - a_2\right| \le 4, \tag{4.1.9}$$

showing that no value in $D_{1/4}(0)$ can be omitted. □

In fact equality can happen in (4.1.9) only if $|a_2| = 2$, so any function in S that is not a Koebe function has a disk $D_r(0)$ larger than $D_{1/4}(0)$ in its image.

Bieberbach's theorem has other consequences for the general theory of univalent functions. The following simple lemmas are key.

Lemma 4.1.5. *For f in S and g in* Aut$(\mathbb{D})$ *the function*

$$h = \frac{f \circ g - f(g(0))}{f'(g(0))\, g'(0)} \tag{4.1.10}$$

belongs to S.

Proof: This function is the composition of $f \circ g$ with an affine map, so it is univalent. Clearly $f(0) = 0$, and the denominator is chosen so that $h'(0) = 1$. □

Lemma 4.1.6. *For f in S and t in $\mathbb{D}$,*

$$\left|(1-|t|^2)\frac{f''(t)}{f'(t)} - 2\bar{t}\right| \le 4. \tag{4.1.11}$$

Proof: In (4.1.10) we choose $t \in \mathbb{D}$ and take

$$g(z) \;=\; \frac{z+t}{1+\bar{t}z}.$$

Then

$$g(0) \;=\; t, \qquad g'(0) \;=\; 1-|t|^2, \qquad g''(0) \;=\; -2\bar{t}(1-|t|^2)$$

so

$$\begin{aligned} h''(0) = \frac{(f\circ g)''(0)}{f'(g(0))\,g'(0)} &= \frac{f''(g(0))g'(0)^2 + f'(g(0))g''(0)}{f'(g(0))g'(0)} \\ = \frac{f''(t)\,g'(0)}{f'(t)} + \frac{g''(0)}{g'(0)} &= \frac{f''(t)}{f'(t)}(1-|t|^2) - 2\bar{t}. \end{aligned}$$

But $h''(0)$ is twice the second coefficient in the expansion of h, so Bieberbach's theorem gives (4.1.11). □

Theorem 4.1.7. *(*Koebe's distortion theorem*) For f in S,*

$$\frac{1-\rho}{(1+\rho)^3} \;\le\; |f'(z)|^2 \;\le\; \frac{1-\rho}{(1-\rho)^3}, \qquad \rho = |z|. \tag{4.1.12}$$

Proof: Replacing t by z in (4.1.11) and multiplying by $|z|/(1-\rho^2)$ gives the estimate

$$\left|\frac{zf''(z)}{f'(z)} - \frac{2\rho^2}{1-\rho^2}\right| \le \frac{4\rho}{1-\rho^2}.$$

Thus

$$\frac{2\rho^2-4\rho}{1-\rho^2} \;\le\; \mathrm{Re}\left\{\frac{zf''(z)}{f'(z)}\right\} \;\le\; \frac{2\rho^2+4\rho}{1-\rho^2}. \tag{4.1.13}$$

Since f is univalent and $f'(0) = 1$, we may take the principal branch of $\log f'$. Writing $z = \rho e^{i\theta}$, we have

$$\frac{\partial}{\partial\rho}\log f'(z) \;=\; \frac{z}{|z|}\frac{f''(z)}{f'(z)}$$

Multiplying by $\rho = |z|$ and taking the real part gives

$$\rho\frac{\partial}{\partial\rho}\left\{\mathrm{Re}\,\log f'(z)\right\} \;=\; \mathrm{Re}\left\{\frac{z\,f''(z)}{f'(z)}\right\}.$$

From this and (4.1.13) we obtain

$$\frac{2\rho-4}{1-\rho^2} \;\le\; \frac{\partial}{\partial\rho}\log|f'(\rho e^{i\theta})| \;\le\; \frac{2\rho+4}{1-\rho^2}.$$

Integrating from 0 to ρ gives

$$\log\frac{1-\rho}{(1+\rho)^3} \;\le\; \log|f'(\rho e^{i\theta})| \leqslant \log\frac{1+\rho}{(1-\rho)^3},$$

and exponentiating yields (4.1.12). □

Theorem 4.1.8. *(*Growth theorem) For f in S,

$$\frac{\rho}{(1+\rho)^2} \;\le\; |f(z)| \;\le\; \frac{\rho}{(1-\rho)^2}, \qquad \rho = |z|. \tag{4.1.14}$$

Proof: Let $z=\rho e^{i\theta}$ be fixed, $0<\rho<1$. Since $f(0)=0$,

$$f(z) \;=\; \int_0^\rho f'(\sigma e^{i\theta})e^{i\theta}d\sigma.$$

By the distortion theorem, we have

$$|f(z)| \;\le\; \int_0^\rho |f'(\sigma e^{i\theta})|d\sigma \;\le\; \int_0^\rho \frac{1+\sigma}{(1-\sigma)^3}d\sigma = \frac{\rho}{(1-\rho)^2}.$$

This gives the upper bound in (4.1.14).

To establish the lower bound, since $\rho(1+\rho)^{-2} < \frac{1}{4}$ for $0 \;\le\; \rho < 1$, we only need to consider the case $|f(z)| < \frac{1}{4}$. Then by the Koebe one-quarter theorem, the straight line segment from 0 to $w=f(z)$ lies entirely within $f(\mathbb{D})$. Let Γ be the pre-image of this segment. Then Γ is a simple arc from 0 to z, and

$$f(z) \;=\; \int_\Gamma f'(\zeta)d\zeta.$$

But $\arg f'(\zeta)d\zeta = \arg dw =$ constant along Γ. Therefore

$$|f(z)| \;=\; \left|\int_\Gamma f'(\zeta)d\zeta\right| \;=\; \int_\Gamma |f'(\zeta)||d\zeta|$$

By the distortion theorem,

$$|f(z)| \;\ge\; \int_0^\rho \frac{1-\sigma}{(1+\sigma)^3}d\sigma = \frac{\rho}{(1+\rho)^2},$$

completing the proof. □

The bounds in (4.1.12) and (4.1.14) are sharp. They are attained by the Koebe function; see Exercise 3.

4.2 The Bieberbach conjecture: history and strategy

The attack on the Bieberbach conjecture is a case study in research work on a natural and difficult problem.

One approach is to start with some subclass of S. The conjecture was proved in 1921, for f with range $f(\mathbb{D})$ that is starlike with respect to 0, by Nevanlinna [154], and in 1931-32, for f with real coefficients, independently by Rogosinski [180] and Dieudonné [55].

Another approach is to look for the sharpest uniform upper bound that one can find for all f in S. Successive results were

$\lvert a_n\rvert \leq e \cdot n$	Littlewood [134], 1925
$\lvert a_n\rvert \leq \frac{3}{4}e \cdot n$	Bazilevic [18], 1951
$\lvert a_n\rvert \leq (\frac{1}{2}e + 1.51) \cdot n$	Baernstein [14], 1974
$\lvert a_n\rvert \leq \frac{1}{2}e \cdot n$	Milin [140], 1965
$\lvert a_n\rvert \leq 1.243 \cdot n$	Fitzgerald [72], 1972
$\lvert a_n\rvert \leq \sqrt{7/6} \cdot n$	Horowitz [110], 1976

We mention also an asymptotic result of Hayman [97] in 1955:

$$\lim_{n\to\infty} \frac{|a_n|}{n} = \alpha(f) \leq 1,$$

with equality if and only if f is a Koebe function.

Finally, one can attack one coefficient at a time

$\lvert a_3\rvert \leq 3$	Loewner [136], 1923
$\lvert a_4\rvert \leq 4$	Garabedian and Schiffer [81], 1955
$\lvert a_6\rvert \leq 6$	Ozawa [161], Pederson [163], 1976
$\lvert a_5\rvert \leq 5$	Pedersen and Schiffer [164], 1980

The Koebe functions are examples of *slit mappings*: functions $f \in S$ with the property that the complement of $f(\mathbb{D})$ is a Jordan path from some finite point in $\mathbb{C}$ to the point at ∞. Carathédory [38] introduced a notion of convergence of domains that allowed him to show that slit mappings are dense in S, in the sense of uniform convergence on compact subsets of $\mathbb{D}$. Therefore attacks on the Bieberbach conjecture can focus on slit mappings. Loewner used this fact, and a construction of a well-chosen family of slit mappings, to prove his result for the third coefficient. Loewner's method came to play an important part in the proof of the full conjecture. (The original paper [136] was written by Karl Löwner. After emigrating to the U.S, the author became Charles Loewner.)

4.3 The Carathéodory convergence theorem

A function f in S is called a *slit mapping* if the complement of $f(\mathbb{D})$ in $\mathbb{C}$ is a Jordan path. An example is the Koebe function (4.0.2), where the complement of $K(\mathbb{D})$ is a half-line. Since $f(\mathbb{D})$ is simply connected, this curve must tend to infinity. The image $f(D)$ itself is termed a *slit domain*. As we shall see, slit mappings are dense in S. A preliminary result is Carathéodory's convergence theorem. This involves a particular notion of convergence of a sequence of simply connected domains $\Omega_n \subset \mathbb{C}$. If 0 is an interior point of $\bigcap \Omega_n$ then the *kernel* of $\{\Omega_n\}$ is the largest domain Ω that contains 0 and has the property that each compact subset of Ω lies in all but finitely many Ω_n. (It is important to remember that a domain is, by definition, a connected set.) Any domain that is the union of such domains has this property: see Exercise 9; therefore there is a largest such domain. If 0 is not an internal point of $\bigcap \Omega_n$, then the kernel is taken to be $\{0\}$. In either case, $\{\Omega_n\}$ is said to *converge* to its kernel in the sense of Carathéodory if every subsequence of $\{\Omega_n\}$ has the same kernel.

This clearly needs some illustration. Let Ω_n be the complement of the path consisting of a half-line and a portion of the unit circle

$$\Gamma_n = [1, \infty) \cup \{e^{i\theta} : 0 \le \theta \le 2\pi(1 - 1/n)\}.$$

It is easily checked that the kernel is $\mathbb{D}$ and that Ω_n converges to $\mathbb{D}$; see Figure 4.1. This example also hints at how to accomplish an approximation by slit mappings.

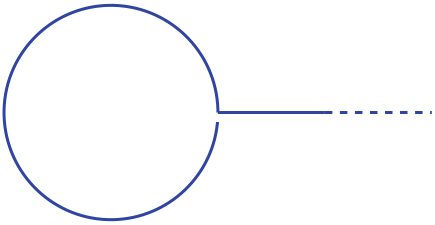

Fig. 4.1 Slit domain approximation to a disk.

Theorem 4.3.1. (Carathéodory) *Let $\{\Omega_n\}$ be a sequence of simply connected plane domains not equal to $\mathbb{C}$, such that 0 is an interior point of $\bigcap \Omega_n$. Suppose the kernel Ω of $\{\Omega_n\}$ is not all of $\mathbb{C}$. Let $f_n : \mathbb{D} \to \Omega_n$ be conformal, with $f_n(0) = 0$ and $f_n'(0) > 0$. Then $f_n \to f$ uniformly on each compact subset of $\mathbb{D}$ if and only if $\{\Omega_n\}$ converges to Ω in the sense of Carathéodory. In the case of convergence, Ω is simply connected, and the inverse maps f_n^{-1} converge to f^{-1} uniformly on each compact subset of Ω.*

Proof: Suppose that $f_n \to f$ uniformly on compact subsets of $\mathbb{D}$. By Proposition 2.3.6, f is either constant or univalent. Some disk $D = D_\rho(0)$ belongs to all Ω_n. Then the functions $g_n = f_n^{-1} : \Omega_n \to \mathbb{D}$ map 0 to 0 and $g_n(\Omega_n)$ contains D, so by

Schwarz's lemma, $|g_n'(0)| < 1/\rho$. Therefore $f_n'(0) > \rho$, so the limit f is univalent. We need to show that $f(\mathbb{D}) = \Omega$, and that $\{\Omega_n\}$ converges to Ω in the sense of Carathéodory.

We first show that $f(\mathbb{D}) \subset \Omega$. Let E be a compact subset of $f(\mathbb{D})$, and let Γ be a smooth Jordan curve that encloses E. Let $\delta > 0$ be the distance from E to Γ, and $\Gamma_1 = f^{-1}(\Gamma)$. We will show that $E \subset \Omega_n$ for all sufficiently large n. Fix $z_0 \in E$. Then $|f(z) - z_0| \geq \delta$ for $z \in \Gamma_1$. By the uniform convergence of $\{f_n\}$ on Γ_1, $|f_n(z) - f(z)| < \delta$ for all $z \in \Gamma_1$ and sufficiently large n. In view of

$$|f_n(z) - f(z)| \ < \ |f(z) - z_0|, \qquad z \in \Gamma,$$

Rouché's theorem implies that $f_n(z) - z_0 = f(z) - z_0 + [f_n(z) - f(z)]$ has the same number of zeros inside Γ_1 as $f(z) - z_0$, namely, one. This shows that z_0 is in Ω_n for all $n > n_0$, where n_0 depends on E but not on z_0. In other words, $E \subset \Omega_n$ for all $n > n_0$. By the definition of the kernel Ω, this means that $f(\mathbb{D}) \subset \Omega$.

The inverse functions $g_n = f_n^{-1}$ are defined on E for all $n \geq n_0$, and $|g_n(w)| \leq 1$. Therefore the g_n are a normal family. Renumbering a convergent subsequence, we get $\{g_n\}$ that converges uniformly on compact subsets of $f(\mathbb{D})$ to a function g holomorphic on $f(\mathbb{D})$ with $g(0) = 0$ and $g'(0) >$. Indeed, restricting n to the subsequence,

$$0 \ < \ \frac{1}{f'(0)} = \lim_{n\to\infty} \frac{1}{f_n'(0)} = \lim_{n\to\infty} g_n'(0) = g'(0).$$

Thus, g is univalent.

The next step is to show that $g = f^{-1}$. Fix $z_0 \in \mathbb{D}$ and let $w_0 = f(z_0)$. Choose $\varepsilon > 0$ so that the circle $\Gamma = \{z : |z - z_0| = \varepsilon\}$ lies in $\mathbb{D}$, and let $\Gamma_1 = f(\Gamma)$. Let δ be the distance of w_0 from Γ_1. Then $|f_n(z) - w_0| \geq \delta$ for $z \in \Gamma$, while $|f_n(z) - f(z)| < \delta$ on Γ for all large n. As above, it follows by Rouché's theorem that for large n there is precisely one z_n inside Γ such that $f_n(z_n) = w_0$. Thus $|z_n - z_0| < \varepsilon$ and $z_n = g_n(w_0)$. Therefore, if n is so large that $|g_n(w_0) - g(w_0)| < \varepsilon$, then for g_n in the convergent subsequence

$$|g(w_0) - z_0| \ \leq \ |g(w_0)) - g_n(w_0)| + |z_n - z_0| < 2\varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, $g(w_0) = z_0$. Since $z_0 \in \mathbb{D}$ is arbitrary, $g = f^{-1}$.

We know now that any convergent subsequence of $\{g_n\}$ converges, uniformly on compact subsets of $f(\mathbb{D})$, to f^{-1}. A further application of Montel's theorem shows that the whole sequence $\{g_n\}$ converges to f^{-1}.

Now let Ω be the kernel of $\{\Omega_n\}$, and let F be a compact subset of Ω. All but finitely many g_n are defined on F. Since $f(\mathbb{D}) \subset \Omega$, Theorem 2.3.5 applies, so $\{g_n\}$ converges (to f^{-1}) uniformly on F. Since $g_n(F) \subset \mathbb{D}$, we have $F \subset f(\mathbb{D})$. This is true for every such F, so we have proved that $\Omega = f(\mathbb{D})$. The preceding argument applies to any subsequence of the original $\{\Omega_n\}$, showing that its kernel is $f(\mathbb{D})$. Thus, all subsequences of $\{\Omega_n\}$ have the same kernel.

Suppose now that $\{\Omega_n\}$ converges in the sense of Caratheodory to a domain $\Omega \neq f(\mathbb{D})$. Then the sequence $\{f_n'(0)\}$ is bounded. Indeed, if $|f_n(0)| > n$ for some (renumbered) subsequence, the Koebe one-quarter theorem shows that $f_n(\mathbb{D})$ contains $D = D_{n/4}(0)$, and it follows that the subsequence has kernel $\mathbb{C}$. This contradiction shows that there exists $c \in \mathbb{R}$ such that $f_n'(0) < c$ for all n. By Theorem 4.1.8

$$|f_n(z)| \le f_n'(0)\frac{|z|}{(1-|z|)^2}, \qquad z \in \mathbb{D},$$

which shows that the sequence $\{f_n\}$ is uniformly bounded on each compact subset, hence is normal. Some subsequence converges to a holomorphic f, uniformly on compact subsets of $\mathbb{D}$. By the first part of this proof, f maps $\mathbb{D}$ onto Ω. Again it follows that the whole sequence $\{f_n\}$ converges to f, uniformly on compact subsets of $\mathbb{D}$. □

We now reach the punch line.

Theorem 4.3.2. *If f is in S, there is a sequence of slit mappings $\{f_n\}$ in S such that $f_n \to f$ uniformly on compact subsets of $\mathbb{D}$.*

Proof: Suppose first that f extends holomorphically to a larger disk $D_{1+\delta}(0)$. Then $f(\partial\mathbb{D})$ is an analytic Jordan curve. Let

$$w_0 = f(1), \qquad w_n = f(e^{2\pi i(1-1/n)})$$

and let Γ_n be the path that consists of an arc in the complement of $f(\overline{\mathbb{D}})$ running from from ∞ to w_0 and the arc

$$w = f(e^{i\theta}), \qquad 0 \le \theta \le 2\pi\left(1-\frac{1}{n}\right);$$

see Figure 4.2.

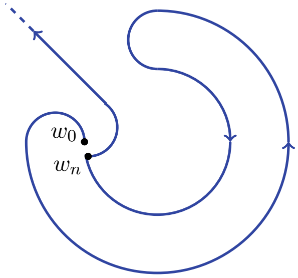

Fig. 4.2 Approximating a general univalent function by slit maps.

The complement Ω_n of Γ_n is simply connected. Let g_n be the conformal map of $\mathbb{D}$ onto Ω_n with $g_n(0) = 0$, $g_n'(0) > 1$. It is geometrically clear that $\Omega = f(\mathbb{D})$ is the kernel of the family $\{\Omega_n\}$ and that $\Omega_n \to \Omega$ in the sense of Carathéodory. Therefore Theorem 4.3.1 implies that $g_n \to f$ uniformly on compact subsets of

$\mathbb{D}$. By Cauchy's formula for the derivative, this implies that $g_n'(0) \to f'(0) = 1$. Therefore the functions

$$h_n(z) = \frac{g_n(z)}{g_n'(0)}$$

are slit mappings in S that converge to f uniformly on compact subsets of $\mathbb{D}$.

For a general f in S, let $f_\sigma(z) = f(\sigma z)/\sigma$, $0 < \sigma < 1$. Then f_σ extends to $D_{1/\delta}$, so it can be approximated by slit mappings. As $\sigma \to 1$, $f_\sigma \to f$, so f can be approximated by slit mappings. □

The importance of slit mappings for the main subject of this chapter is made clear by the following.

Corollary 4.3.3. *If Bieberbach's conjecture is true for the subset S_s of S consisting of slit mappings, then it is true for S.*

The proof is left as Exercise 10.

4.4 Slit mappings and Loewner's equation

In Section 4.3 we saw that slit mappings are dense in S. Given such a mapping $f \in S$, the associated slit Γ is $\mathbb{C} \setminus f(\mathbb{D})$, the set of values that are not attained by f.

Loewner [136] attacked the converse problem of determining such a mapping f from knowledge of the slit Γ. After a suitable parametrization of the equation of the slit, he was able to determine the associated map $f \in S$ from the limiting value of the solution of a certain differential equation. In fact we shall see that

$$f(z) = \lim_{t \to \infty} e^t g(z, t),$$

where $g(z, t) = g_t(z))$ is the solution of a first-order differential equation in t with initial condition $g(0, z) = z$ for $z \in \mathbb{D}$.

The first steps in the argument are to shrink the map f to a family of maps $\{f_t\}_{t \geq 0}$ in S by shrinking the slit toward ∞, and to choose a canonical parametrization. Then $f_0 = f$ and f_t converges to the identity map as $t \to \infty$. The decisive step is to examine $g_t = f_t^{-1} \circ f$ and show that g_t satisfies a first-order differential equation with respect to t.

Suppose that $f \in S$ is a slit mapping. Let Γ be the Jordan arc that is the complement of $\Omega = f(\mathbb{D})$, parametrized by a map

$$t \to \sigma(t), \quad 0 \leq t < b, \qquad \lim_{t \to b} \sigma(t) = \infty.$$

Let

$$\Gamma_t = \{\sigma(s) : t \leq s < b\}, \qquad \Omega_t = \mathbb{C} \setminus \Gamma_t.$$

Thus $\Omega_0 = \Omega$, the domains Ω_t increase with t, and $\bigcup \Omega_t = \mathbb{C}$. Let f_t,

$$f_t(z) \;=\; \beta(t)[z + b_2(t)z^2 + b_3(t)z^3 + \ldots],$$

be the conformal map from $\mathbb{D}$ onto Ω_t with $f_t'(0) = \beta(t) > 0$. Then $f_0 = f$. Given $t \in [0, b)$, the Carathéodory convergence theorem says that $f_s \to f_t$ uniformly on compact subsets of $\mathbb{D}$. It follows that the coefficients $b_n(t)$ are continuous functions of t.

Suppose that $s < t$. Then the function $f_t^{-1} \circ f_s$ maps $\mathbb{D}$ to a proper subset of itself and fixes $z = 0$. By Schwarz's lemma, its derivative at $z = 0$, which is positive, is < 1. Therefore $\beta(t) = f_t'(0)$ is strictly increasing with t. Since $b(0) = 1$, we may choose the parametrization $\sigma(t)$ so that $\beta(t) = e^t$. This is called the *standard parametrization* of Γ.

We claim that in the standard parametrization, $b = \infty$. In fact

$$\left|\frac{z}{f_t(z)}\right| \;\le\; \frac{|z|}{r} \;\le\; \frac{1}{r}$$

for $z \in \mathbb{D}$. In particular, $r \le |f_t'(0)| = e^t$ for t close to b. Since r is arbitrary, this shows that $e^t \to \infty$ as $t \to b$, so $b = \infty$. Thus our parametrization is

$$f_t(z) \;=\; e^t[z + b_2(t)z^2 + b_3(t)z^3 + \ldots], \qquad 0 < t < \infty.$$

The function $g_t(z) = f_t^{-1} \circ f$ maps $\mathbb{D}$ conformally onto $\mathbb{D}$ minus the pre-image of Γ_t, which is an arc that extends inward from the boundary. This function has an expansion

$$g_t(z) \;=\; e^{-t}[z + a_2(t)z^2 + a_3(t)z^3 + \ldots], \tag{4.4.1}$$

where each $a_n(t)$ is a polynomial in $b_2(t), \ldots b_n(t)$; Exercise 11. In particular, $a_0(z) = z$.

Next, following Duren [64], we prove convergence of g_t to f and establish *Loewner's differential equation*, (4.4.3).

Theorem 4.4.1. *Let f be a slit map, let $\sigma(t)$ be the standard representation of the omitted path Γ, and let the functions f_t and g_t be defined as above. Then*

$$\lim_{t\to\infty} e^t g_t(z) \;=\; f(z) \tag{4.4.2}$$

uniformly on compact subsets of $\mathbb{D}$. There is a continuous $k : \mathbb{D} \to \partial\mathbb{D}$ such that g_t satisfies Loewner's differential equation

$$\frac{\partial g_t}{\partial t}(z) \;=\; -g_t(z)\,\frac{1 + k(z)\,g_t(z)}{1 - k(z)\,g_t(z)}. \tag{4.4.3}$$

Proof: Since $g_t = f_t^{-1} \circ f$ and f maps compact subset of $\mathbb{D}$ to compact subsets of $f(\mathbb{D}) = \Omega$, to prove the first statement of the theorem it is enough to show that

$$\lim_{t\to\infty} e^t f_t^{-1}(w) \;=\; w$$

uniformly on compact sets in $\mathbb{C}$. Theorem 4.1.8 gives

$$\frac{e^t|z|}{(1+|z|)^2} \le |f_t(z)| \le \frac{e^t|z|}{(1-|z|)^2}$$

With $z = f_t^{-1}(w)$, this leads to

$$[1 - |f_t^{-1}(w)]^2 \le e^t \left|\frac{f_t^{-1}(w)}{w}\right| \le [1 + |f_t^{-1}(w)]^2. \tag{4.4.4}$$

Therefore $|f_t^{-1}(w)| \le 4|we^{-t}|$, so $f_t^{-1} \to 0$ uniformly on bounded sets. Hence (4.4.4) implies

$$e^t \left|\frac{f_t^{-1}(w)}{w}\right| \to 1. \tag{4.4.5}$$

It follows that the functions

$$h_t(w) = e^t \frac{f_t^{-1}(w)}{w}, \qquad 0 \le t < \infty$$

are a normal family. Any convergent subsequence has holomorphic limit h with $|h(w)| = 1 = h(0)$, so, by the strong maximal principle, $h \equiv 1$. Therefore the h_t converge to 1 as $\tau \to \infty$, uniformly on compact sets. This proves (4.4.3).

Now for $0 \le s < t < \infty$, let

$$h_{st}(z) = f_t^{-1}(f_s(z)) = e^{s-t}[z + c_2(s,t)z^2 + \dots].$$

This function maps $\mathbb{D}$ conformally onto $\mathbb{D}$ minus a Jordan arc J_{st}, that extends inward from a point $\lambda(t) = f_t^{-1}(\sigma(t))$ on $\partial\mathbb{D}$. Let B_{st} be the portion of $\partial\mathbb{D}$ that maps to J_{st}. By the Carathéodory extension theorem, Theorem 2.6.1, f_t^{-1} maps $\partial\mathbb{D}$ onto the (two-sided) slit $\Gamma_t \setminus \Gamma_s$, so $\lambda(s) = f_t^{-1}(\sigma(s))$ is an interior point of the arc B_{st}. As $s \uparrow t$ or $t \downarrow s$, B_{st} shrinks to $\lambda(s)$ or to $\lambda(t)$, respectively.

We claim that λ is continuous. The function h can be continued by reflection across the complement of B_{st} in $\partial\mathbb{D}$. The continuation maps the (full) complement of B_{st} onto the complement of the union of J_{st} and its reflection J_{st}^*. By Koebe's one-quarter theorem, J_{st} lies outside the disk $D_r(0)$, $r = e^{s-t}/4$. Therefore J_{st}^* lies in the disk

$$\{z : |z| < 4e^{t-s}\}.$$

Since $z/h_{st}(z) \to e^{t-s}$ as $t \to 0$, the reflection satisfies

$$\lim_{z\to\infty} \frac{h_{st}(z)}{z} = e^{t-s}.$$

By the maximum modulus theorem,

$$\left|\frac{h_{st}(z)}{z}\right| \le 4e^{t-s}, \quad z \notin B_{st}.$$

Letting t decrease to s, a normal families argument shows that $h_{st}(t)$ converges to a function that is holomorphic and bounded on the complement of $\lambda(s)$ with limit 1 at $\lambda(s)$. Thus

$$\lim_{t\downarrow s} h_{st}(z) = z,$$

uniformly on compact sets not containing $\lambda(s)$.

Now given $s \geq 0$ and $\varepsilon > 0$, choose $\delta > 0$ such that if $s < t < s + \delta$, then the circle $C = \{z : |z - \lambda(0)| < \varepsilon\}$ encloses B_{st}. The image $\widetilde{C}$ of C under $h(z, s, t)$ encloses $J_{st} \cup J^*_{st}$, so in particular it encloses $\lambda(t)$. Since $h_{st}(t) \to z$ uniformly on C as $t \to s$, it follows that for t sufficiently close to s, the diameter of $\widetilde{C}$ is $< 3\varepsilon$. Thus for any $z_0 \in C$, as $t \downarrow s$,

$$\begin{aligned} |\lambda(t) - \lambda(z)| &\leq |\lambda(t) - z_0| + |z_0 - h_{st}(z_0)| + |h_{st}(z_0) - \lambda(t)| \\ &\leq \varepsilon + \varepsilon + 3\varepsilon. \end{aligned}$$

This proves continuity from the right, and the same constructions prove continuity from the left.

Finally, note that $h_{st}(z)/z$ has no zeros, and extends to have value e^{t-s} at $z = 0$. Therefore we may choose a branch of the logarithm so that

$$\Phi(z) = \Phi(z, s, t) = \log \frac{h_{st}(z)}{z}, \qquad \Phi(0) = t - s.$$

Now Φ is holomorphic in $\mathbb{D}$ and continuous in the closure. The properties of h_{st} imply that

$$\operatorname{Re} \Phi(z) = 0, \quad \text{for } |z| = 1,\ z \notin B_{st}; \qquad \operatorname{Re} \Phi(z) < 0, \quad \text{for } z \in B_{st}. \tag{4.4.6}$$

Therefore the extended Poisson integral formula of Theorem 5.1.6 gives

$$\Phi(z) = \frac{1}{2\pi} \int_\alpha^\beta \operatorname{Re} \Phi(e^{i\theta}) \frac{e^{i\theta} + z}{e^{i\theta} - z}\, d\theta, \tag{4.4.7}$$

where $e^{i\alpha}$ and $e^{i\beta}$ are the endpoints of B_{st} with the positive orientation. Then

$$s - t = \Phi(0) = \frac{1}{2\pi} \int_\alpha^\beta \operatorname{Re} \Phi(e^{i\theta})\, d\theta. \tag{4.4.8}$$

By definition, $h_{st}(g_s(z)) = g_t(z)$. Therefore if we replace z in (4.4.7) by $g_s(z)$ we get

$$\log \frac{g_t(z)}{g_s(z)} = \frac{1}{2\pi} \int_\alpha^\beta \operatorname{Re} \Phi(e^{i\theta}) \frac{e^{i\theta} + g_s(z)}{e^{i\theta} - g_t(z)}\, d\theta, \tag{4.4.9}$$

As $t \downarrow s$ the interval shrinks. We may apply the mean value theorem to the real part and the imaginary part of (4.4.9) separately in order to replace the variable $e^{i\theta}$ by some intermediate values, divide by $t - s$, and take advantage of (4.4.8) to conclude that the derivative from the right is

$$\frac{\partial}{\partial s}\log g_s(z) \;=\; -\frac{\lambda(s)+g_s(z)}{\lambda(s)-g_s(z)}, \tag{4.4.10}$$

recalling that B_{st} contracts to $\lambda(s)$. The same argument applies to the derivative from the left, taking $s \uparrow t$. Setting $k(t) = 1/\lambda(t)$, we have obtained (4.4.3). □

Let us look more closely at the family of functions f_t.

Theorem 4.4.2. *Let f_t, $0 \le t < \infty$, be the normalized conformal map of $\mathbb{D}$ onto $\Omega_t = \mathbb{C} \setminus \Gamma_t$,*

$$f_t(z) \;=\; e^t[z + b_2(t)z^2 + b_3(t)z^3 + \ldots], \qquad 0 < t < \infty. \tag{4.4.11}$$

Then $f_0(z) = f(z)$, the normalized conformal map with $f(\mathbb{D}) = \mathbb{C} \setminus \Gamma$, and

$$\lim_{t\to\infty} \frac{f_t(z)}{e^t\, z} \;=\; 1, \qquad z \in \mathbb{C}, \tag{4.4.12}$$

uniformly on compact subsets of $\mathbb{D}$. Moreover

$$\frac{\partial f_t}{\partial t}(z) \;=\; z\, f_t'(z)\, \frac{1+k(t)\,z}{1-k(t)\,z}, \qquad k(t) = \frac{1}{\lambda(t)}, \tag{4.4.13}$$

and

$$\operatorname{Re}\left\{\frac{\partial f_t(z)/\partial t}{z f_t'(z)}\right\} \;>\; 0. \tag{4.4.14}$$

Proof: It is clear that $f_0(z,0) = f(z)$, since $\Gamma_0 = \Gamma$. The assertion (4.4.12) follows from (4.4.5) by taking $w = f(z)$.

By definition,

$$f_t(g_t(w)) \;=\; f(z).$$

Differentiating with respect to t gives

$$f_t'(g_t(w))\,\frac{\partial}{\partial t}(g_t(w)) + \frac{\partial}{\partial t} f_t(g_t(w)) \;=\; 0. \tag{4.4.15}$$

Using (4.4.3), and replacing $g_t(w)$ by z, converts (4.4.15) to (4.4.13). Then (4.4.14) follows, since $|k| = 1$, $|z| < 1$ implies

$$\operatorname{Re}\frac{1+kz}{1-kz} \;>\; 0. \qquad \square$$

Remarks. An important modification of Loewner's equation (4.4.3) is a stochastic version proposed by Schramm [187], who designated it SLE for *stochastic Loewner evolution*. It is now generally called the *Schramm–Loewner evolution*. Loewner's equation can be written in an equivalent form

$$\frac{\partial g_t}{\partial t} \;=\; -g_t\,\frac{\zeta(t)+g_t}{\zeta(t)-g_t},$$

where the “driving function” ζ is a continuous mapping to ∂D. Schramm’s version replaces the deterministic term ζ by a scaled Brownian motion on ∂D. The resulting equation, denoted SLE$_\kappa$, is

$$\frac{\partial g_t}{\partial t} = -g_t \frac{\sqrt{\kappa}\, B(t) + g}{\sqrt{\kappa}\, B(t) - g},$$

(This is *chordal* SLE; there are other versions, including *radial* SLE.)

There are a number of deep mathematical and physical applications. See Lawler and Limic [129] and Kemppainen [119].

4.5 The Robertson and Milin conjectures

Suppose that f belongs to S, with expansion

$$f(z) = z + a_2 z^2 + \cdots$$

Let h be the square root transform of f; i.e.

$$h(z) = [f(z^2)]^{1/2} = z[1 + b_2 z^2 + b_4 z^4 + \cdots].$$

Comparing coefficients of z^n in the equation

$$a_1 z + a_2 z^2 + \alpha_3 z^3 + \cdots = z(b_0 + b_2 z + b_4 z^4 + \cdots)^2, \qquad a_1 = b_0 = 1,$$

we find that

$$a_n = b_0 b_{2n} + b_3 b_{2n-2} + \cdots + b_{2n} b_0, \qquad n = 1, 2, \cdots .$$

By Schwarz’s inequality,

$$|a_n|^2 \le (|b_0|^2 + |b_2|^2 + \cdots + |b_{2n}|^2)^2, \tag{4.5.1}$$

The *Robertson conjecture* is that the Bieberbach conjecture is true *because* of this inequality: i.e. that for $n = 1, 2, 3, \ldots,$

$$|b_0|^2 + |b_2|^2 + \cdots + |b_{2n}|^2 \le n^2, \tag{4.5.2}$$

with equality for some n if and only if f is a Koebe function. Thus

Theorem 4.5.1. *If Robertson’s conjecture is true, then Bieberbach’s conjecture is true.*

Robertson’s conjecture is itself a consequence of a conjecture of Milin. For this we need first

Lemma 4.5.2. (Lebedev–Milin) *Let*

$$P \;=\; p_1x + p_2x^2 + \cdots$$

be an element of the ring $\mathfrak{P}$ *of formal power series over* $\mathbb{C}$*, and let*

$$Q \;=\; E \circ P = q_0 + q_1x + q_2x^2 + \cdots, \tag{4.5.3}$$

where E *is the exponential series* $\sum_{n=0}^{\infty} x^n/n\,!$*. Then for* $n = 0, 1, 2, \cdots$,

$$\begin{aligned} &|q_0|^2 + |q_1|^2 + \cdots + |q_n|^2 \\ &\le\; (n+1)\exp\left\{\frac{1}{n+1}\sum_{k=1}^{n}(n+1-k)\left[k|p_k|^2 - \frac{1}{k}\right]\right\}. \end{aligned} \tag{4.5.4}$$

Proof: Formal differentiation of (4.5.3) gives

$$\begin{aligned} &q_1 + 2q_2x + 3q_3x^2 + \ldots \\ &\quad= (q_0 + q_1x + q_2x^2 + \cdots)(p_1 + 2p_2x + 3p_3x^2 + \cdots). \end{aligned}$$

A comparison of the coefficients yields

$$nq_n \;=\; \sum_{k=0}^{n-1}(n-k)\,p_{n-k}\,q_k, \qquad n = 1, 2\cdots.$$

By the Schwarz inequality,

$$n^2|q_n|^2 \;\le\; \sum_{k=1}^{n} k^2|p_k|^2 \sum_{k=0}^{n-1}|q_k|^2. \tag{4.5.5}$$

For $n = 1, 2, \cdots$, let

$$\pi_n \;=\; \sum_{k=1}^{n} k^2|p_k|^2, \qquad \gamma_n = \sum_{k=0}^{n}|q_k|^2.$$

Then (4.5.5) can be written as

$$\gamma_n - \gamma_{n-1} \le \frac{1}{n^2}\pi_n\gamma_{n-1}.$$

Using $1 + x \le e^x$, we obtain

$$\begin{aligned} \gamma_n &\le \left(1 + \frac{1}{n^2}\pi_n\right)\gamma_{n-1} \;=\; \frac{n+1}{n}\left(\frac{n}{n+1} + \frac{\pi}{n(n+1)}\right)\gamma_{n-1} \\ &= \frac{n+1}{n}\left(1 + \frac{\pi_n - n}{n(n+1)}\right)\gamma_{n-1} \;\le\; \frac{n+1}{n}\exp\left\{\frac{\pi_n - n}{n(n+1)}\right\}\gamma_{n-1}. \end{aligned}$$

Repeated application of this inequality yields

$$\begin{aligned}\gamma_n &\le (n+1)\exp\left\{\sum_{k=1}^{n}\frac{\pi_k - k}{k(k+1)}\right\}\\ &= (n+1)\exp\left\{\sum_{k=1}^{n}\frac{\pi_k}{k(k+1)} + 1 - \sum_{k=1}^{n+1}\frac{1}{k}\right\}.\end{aligned}$$

Since

$$\frac{1}{k+1} = \frac{1}{k} - \frac{1}{k+1},$$

it follows from summation by parts that

$$\begin{aligned}\sum_{k=1}^{n}\frac{\pi_k}{k(k+1)} &= \sum_{k=1}^{n}\pi_k\left(\frac{1}{k} - \frac{1}{k+1}\right) = \sum_{k=1}^{n}\frac{1}{k}(\pi_k - \pi_{k-1}) - \frac{\pi_n}{n+1}\\ &= \sum_{k=1}^{n}k|p_k|^2 - \frac{1}{n+1}\sum_{k=1}^{n}k^2|p_k|^2.\end{aligned}$$

Therefore

$$\begin{aligned}\gamma_n &\le (n+1)\exp\left\{\sum_{k=1}^{n}\left(1 - \frac{k}{n+1}\right)k|p_k|^2 + 1 - \sum_{k=1}^{n+1}\frac{1}{k}\right\}\\ &= (n+1)\exp\left\{\frac{1}{n+1}\sum_{k=1}^{n}(n+1-k)\left[k|p_k|^2 - \frac{1}{k}\right]\right\},\end{aligned}$$

thus proving (4.5.4). □

Now let h be any odd function in S,

$$h(z) = z + b_2z^3 + b_4z^5 + \cdots,$$

and let $f(z) = [h(\sqrt{z})]^2$. Then f belongs to S: see Exercise 5. Using (4.5.4) we convert Robertson's inequality (4.5.3) into an inequality for the coefficients c_k in

$$\log\frac{f(z)}{z} = \sum_{k=1}^{\infty}c_kz^k. \tag{4.5.6}$$

Clearly this has the form

$$\log\frac{f(z)}{z} = \log\frac{[h(\sqrt{z})]^2}{z} = 2\log(1 + b_2z + b_4z^2 + \cdots).$$

Therefore

$$1 + b_2z + b_4z^2 + \cdots = \exp\left\{\frac{1}{2}\sum_{k=1}^{\infty}c_kz^k\right\}.$$

Using Lemma 4.5.2 with $q_k = b_{2k}$ and $p_k = \frac{1}{2}c_k$, we obtain

$$|1 + |b_2|^2 + \cdots + |b_{2n}|^2 \;\le\; (n+1)\exp\left\{\frac{1}{4(n+1)}\sum_{k=1}^{n}(n+1-k)\left[k|c_k|^2 - \frac{4}{k}\right]\right\}.$$

If the exponent is negative, then the above exponential is ≤ 1. This, together with Theorem 4.5.1 gives the following result.

Theorem 4.5.3. *If, for each f in S, the coefficients c_s defined by (4.5.6) satisfy*

$$\sum_{k=1}^{n}(n+1-k)\left[k|c_k|^2 - \frac{4}{k}\right] \;\le\; 0, \qquad n = 1, 2, 3, \ldots, \tag{4.5.7}$$

then the Bieberbach conjecture is true.

Milin conjectured in 1971 that the inequality (4.5.7) actually holds.

4.6 Preparation for the proof of de Branges's theorem

In this section we adapt the constructions in Section 4.4 to obtain the specific results that are the basis from which Weinstein's proof of de Branges's theorem proceeds. Here we start with a function that is not a slit mapping and approximate it by slit mappings. For convenience we repeat some steps of the arguments in Section 4.4.

Suppose that $f : \mathbb{D} \to \mathbb{C}$ belongs to S, and

$$f(z) \;=\; z + \sum_{n=2} a_n z^n. \tag{4.6.1}$$

Given $0 < r < 1$, the function

$$f_r(z) \;=\; \frac{f(rz)}{r} \;=\; z + \sum_{n=2}^{\infty}(r^{n-1}a_n)z^n,$$

restricted to $\mathbb{D}$, belongs to S and has a holomorphic extension to $D_{1/r}(0)$. The coefficients $a_n(r) = r^{n-1}a_n$ converge to a_n, so for our purposes we may replace f by f_r and assume that f extends smoothly to $\partial\mathbb{D}$.

Theorem 4.6.1. *Suppose that $f \in S$ extends smoothly to the boundary $\partial\mathbb{D}$. Then there is a family $\{f_t\}_{t>0} \subset S$ with the properties*

(a) $f_0(z) \;=\; f(z)$*;*

(b) $f_t(z) \;=\; e^t z + \sum_{n=2}^{\infty} a_n(t) z^n$*;*

(c) $\log\dfrac{f_t(z)}{e^t z} \;=\; \sum_{k=1}^{\infty} c_k(t) z^k, \qquad c_k(\infty) \;=\; \dfrac{2}{k}$*;*

(d) $\operatorname{Re}\left\{\dfrac{\partial f_t(z)/\partial t}{z f_t'(z)}\right\} > 0.$

Proof. Let $\Omega = f(\mathbb{D})$. The boundary curve $\partial\Omega = f(\partial\mathbb{D})$ encloses $D_{1/4}(0)$ and meets the half-line $(-\infty, 1/4]$. Let

$$s_0 = \sup\{s > 0\,;\,,\ -s \in f(\partial\mathbb{D})\}.$$

We may choose a parametrization of $\sigma(\tau)$ of $\partial\Omega$, $0 \le \tau \le s_0$, with $\sigma(0) = \sigma(s_0) = s_0$. We then parametrize the curve

$$\Gamma = \partial\Omega \cup [s_0, \infty] \tag{4.6.2}$$

by

$$\Gamma(\tau) = \begin{cases} \sigma(\tau), & 0 \le \tau \le s_0; \\ -\tau, & s_0 < \tau < \infty. \end{cases}$$

Now define a family of slit domains for $s > 0$ by

$$\Omega_s = \mathbb{C} \setminus \Gamma_s, \qquad \Gamma_s = \{\Gamma(\tau) : \tau \ge s\};$$

see Figure 4.3. Then Ω_s is simply connected, and the Ω_s converge to $\Omega_0 = \Omega$ in the sense of Carathéodory as $s \to 0$. Note that

$$\Omega_t \subset \Omega_s \qquad \text{if } t < s,$$

and

$$\Omega_s = \mathbb{C} \setminus (-\infty, -s], \qquad \text{if } s \ge s_0. \tag{4.6.3}$$

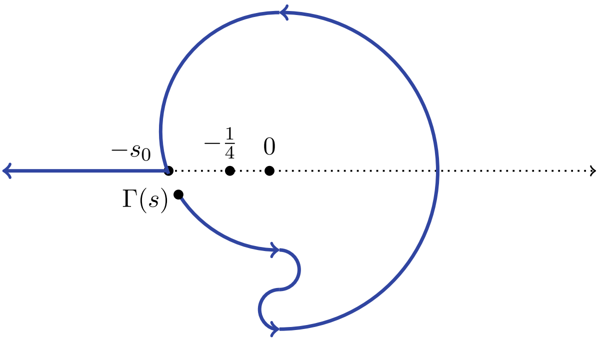

Fig. 4.3 The approximating curve Γ_s, s close to 0.

Let $\widetilde{f}_t$ be the conformal map of $\mathbb{D}$ onto Ω_t that satisfies $\widetilde{f}_t(0) = 0$, $\widetilde{f}_t'(0) > 0$. For $t < s$,

$$\widetilde{f}_t(\mathbb{D}) = \Omega_t \subset \Omega_s = \widetilde{f}_s(\mathbb{D}),$$

so $\widetilde{f}_s^{-1} \circ \widetilde{f}_t : \mathbb{D} \to \mathbb{D}$ is well defined. By Schwarz's lemma, it follows that

$$(\widetilde{f}_s^{-1})'(0))(\widetilde{f}_t)'(0) = (\widetilde{f}_s^{-1})'(\widetilde{f}_t(0))\widetilde{f}_t'(0) < 1,$$

so $\widetilde{f}_t'(0) < \widetilde{f}_s'(0)$. Thus $\widetilde{f}_t'(0)$ is strictly increasing. Moreover (4.6.3) implies that

$$\widetilde{f}_s(z) = \frac{4sz}{(1-z)^2}, \qquad \text{for } s \geq s_0; \tag{4.6.4}$$

see Exercise 15. Therefore $\widetilde{f}_s'(0) = 4s$ for $s \geq s_0$. Note that $\widetilde{f}_0 = f$, so $\widetilde{f}'(0) = 1$.

It follows from these considerations that we may reparametrize by taking

$$f_t = \widetilde{f}_s, \qquad t = \log \widetilde{f}_s'(0), \qquad 0 \leq t \leq \infty. \tag{4.6.5}$$

This produces an expansion in the form (b).

Since f_t vanishes only at $0 \in \mathbb{D}$, we may choose the branch of the logarithm so that $\log(f_t(z)/e^t z)$ is holomorphic in $\mathbb{D}$ and equals 0 at the origin. Therefore f_t has an expansion of the form in (c). Let t and s be related as in (4.6.5), and set $t = t_0$ if $s = s_0$. For $t \geq t_0$ it follows from (4.6.4) that $t = \log 4s$ and that

$$\frac{f_t(z)}{e^t z} = \frac{1}{(1-z)^2}, \qquad \text{for } t \geq t_0.$$

Therefore

$$\log \frac{f_t(z)}{e^t z} = -2\log(1-z) = \sum_{k=1}^{\infty} \frac{2}{k} z^k, \qquad t \geq t_0.$$

This proves part (c). Part (d) is contained in (4.4.14). □

Another important ingredient involves the *Legendre polynomials* $\{P_n\}_0^\infty$. These can be defined by the *generating function*

$$G(x,s) = \sum_{n=0}^{\infty} P_n(x)\, s^n = \frac{1}{(1-2xs+s^2)^{1/2}}, \qquad |x| < 1. \tag{4.6.6}$$

Expanding the right side gives

$$G(x,s) = \sum_{n=0}^{\infty} \frac{1 \cdot 3 \cdots (2n-1)}{2^n n\,!} (2xs - s^2)^n. \tag{4.6.7}$$

Collecting coefficients of s^n shows that $P_n(x)$ is a polynomial of degree n.

The representation (4.6.6) can be used to show that

$$\frac{d}{dx}\left\{(1-x^2)\frac{d}{dx}P_n(x)\right\} = -(n+1)n\,P_n(x); \tag{4.6.8}$$

see Exercise 18. In other words, the P_n are eigenfunctions of the operator

$$L = \frac{d}{dx}(1-x^2)\frac{d}{dx}$$

acting on functions in $L^2(I)$, $I = (-1, 1)$.

Beyond the representation (4.6.8), we need *Legendre's addition formula*

$$\begin{aligned} &P_n(\cos\varphi \sin\theta \sin\theta' + \cos\theta\cos\theta')) &(4.6.9)\\ &= P_n(\cos\theta)P_n(\cos\theta') + 2\sum_{k=1}^{n}\frac{(n-k)!}{(n+k)!}\cos(k\varphi)P_n^k(\cos\theta)P_n^k(\cos\theta'). \end{aligned}$$

The functions P_n^k that occur in (4.6.9) are known as the *associated Legendre functions*. They are closely related to the derivatives of the Legendre polynomials:

$$P_n^k(x) = (-1)^k(1-x^2)^{k/2}P_n^{(k)}(x).$$

Here we give a brief outline of the discussion in [21], which contains the details of the proof of the addition formula (4.6.9).

In spherical coordinates

$$(x, y, z) = (r\cos\varphi\sin\theta, r\sin\varphi\sin\theta, r\cos\theta) \qquad (4.6.10)$$

the Laplacian in $\mathbb{R}^3$ is

$$\begin{aligned} \Delta &= \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\\ &= \frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r} + \frac{1}{r^2\sin^2\theta}\frac{\partial^2}{\partial\varphi^2} + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\sin\theta\frac{\partial}{\partial\theta}. \end{aligned}$$

With $r = 1$, the second and third terms constitute the Laplacian L_S on the unit sphere in $\mathbb{R}^3$ with respect to the coordinates φ, θ. Solutions of $L_S Y = 0$ are known as *spherical harmonics*. Separation of variables, i.e. looking for solutions having the form $Y(\theta, \varphi) == \Phi(\varphi)\Theta(\theta))$, leads one to choose $\Phi(\varphi) = e^{im\varphi}$, $m \in \mathbb{Z}$. Then the equation for Θ becomes

$$\frac{1}{\sin\theta}\frac{d}{d\theta}\left\{\sin\theta\frac{d\Theta}{d\theta}\right\} + \left[(n+1)n - \frac{m^2}{\sin^2\theta}\right]\Theta = 0, \qquad (4.6.11)$$

Letting $x = \cos\theta$ converts (4.6.11) to the spherical harmonic equation

$$\frac{d}{dx}(1-x^2)\frac{du}{dx} + \left[(n+1)n - \frac{m^2}{(1-x)^2}\right]u(x) = 0.$$

For $m = 0$ the solution is a multiple of P_n, and in general the solution Θ_{nm} is a multiple of P_n^m.

With suitable normalization constants $\{c_{nm}\}$, the resulting functions

$$Y_{nm} \; = \; c_{nm} e^{im\varphi} P_n^m(\cos\theta), \qquad |m| \leq n, \;\; n = 0, 1, 2, \dots \; ,$$

are an orthonormal basis for the L^2 space of the sphere, consisting of eigenfunctions of L_S. In particular, the function on the left in (4.6.9) can be expanded with respect to the $\{Y_{nm}\}$, and the right side of (4.6.9) is the expansion.

In the preceding discussion we used the notation of [21]. The version used (implicitly) by Weinstein [213] and expounded in the next section uses the version with θ and φ interchanged, and also sets two of the variables equal. The result is the identity

$$\begin{aligned} & P_n^m(\cos^2\varphi + \sin^2\varphi\cos\theta) \\ & = \; P_n(\cos\varphi)^2 + 2\sum_{m=1}^{n} \frac{(n-m)!}{(n+m)!}\cos(m\phi)(P_m(\cos m\phi))^2 \end{aligned} \tag{4.6.12}$$

4.7 Proof of de Branges's Theorem

Let $\alpha_0 = \beta_0 = 0$ and $\alpha_k = \frac{4}{k} - k|c_k(0)|^2$, $\beta_k = k$, $k = 1, 2, \dots$. The convolution

$$\gamma_n = \sum_{k=0}^{n} \alpha_k \beta_{n-k}$$

from the product

$$\left(\sum_{k=1}^{\infty} \alpha_k z^k\right)\left(\sum_{k=1}^{\infty} \beta_k z^k\right) = \sum_{n=2}^{\infty} \gamma_n z^n$$

suggests that the finite sum in the Milin conjecture is just the coefficient of z^{n+1} in the product of the two series

$$\sum_{k=1}^{\infty}\left(\frac{4}{k} - k|c_k(0)|^2\right) z^k, \qquad \sum_{k=1}^{\infty} k\, z^k = \frac{z}{(1-z)^2}.$$

Indeed, we have

$$\begin{aligned} & \sum_{n=1}^{\infty}\left\{\sum_{k=1}^{n}(\frac{4}{k} - k|c_k(0)|^2)(n-k+1)\right\} z^{n+1} \\ & \quad = \frac{z}{(1-z)^2}\sum_{k=1}^{\infty}\left(\frac{4}{k} - k|c_k(0)|^2\right) z^k. \end{aligned} \tag{4.7.1}$$

This is the first step in Weinstein's argument. Let us denote the left-hand side of (4.7.1) by $\Phi(z)$:

$$\Phi(z) = \sum_{n=1}^{\infty} \left\{ \sum_{k=1}^{n} \left(\frac{4}{k} - k|c_k(0)|^2 \right) (n-k+1) \right\} z^{n+1}. \tag{4.7.2}$$

The next step is to show that $\Phi(z)$ can be expressed as

$$\Phi(z) = \sum_{n=1}^{\infty} \left(\int_0^{\infty} h_n(t)dt \right) z^{n+1}, \tag{4.7.3}$$

where $h_n(t) \geq 0$ for all $t \geq 0$ and $n = 1, 2, \cdots$.

In the following, we keep z fixed and define $w = w_t(z)$ by

$$\frac{z}{(1-z)^2} = \frac{e^t w}{(1-w)^2}, \qquad t \geq 0, \tag{4.7.4}$$

so that $w_0(z) = z$. Recall from Theorem 4.6.1 (c) that $|c_k(\infty)| = 2/k$. Hence

$$\begin{aligned} \int_0^{\infty} \frac{d}{dt} \left[\sum_{k=1}^{\infty} \left(\frac{4}{k} - k|c_k(t)|^2 \right) w_t^k \right] dt \;&=\; \sum_{k=1}^{\infty} \left(\frac{4}{k} - k|c_k(t)|^2 \right) w_t^k \Bigg|_0^{\infty} \\ &= \sum_{k=1}^{\infty} \left(\frac{4}{k} - k|c_k(\infty)|^2 \right) w_{\infty}^k - \sum_{k=1}^{\infty} \left(\frac{4}{k} - k|c_k(0)|^2 \right) w_0^k. \end{aligned}$$

Since $w_\infty < \infty$ and $w_0 = z$, the identities (4.7.1) and (4.7.2) imply that

$$\Phi(z) \;=\; \int_0^{\infty} -\frac{z}{(1-z)^2} \frac{d}{dt} \left[\sum_{k=1}^{\infty} \left(\frac{4}{k} - k|c_k(t)|^2 \right) w^k \right] dt. \tag{4.7.5}$$

It follows from (4.7.4) that

$$\frac{dw}{dt} \;=\; -w\frac{1-w}{1+w}. \tag{4.7.6}$$

so

$$\begin{aligned} \Phi(z) \;=\; \int_0^{\infty} \frac{e^t w}{1-w^2} \frac{1+w}{1-w} &\left[\sum_{k=1}^{\infty} k[c_k(t)\overline{c_k(t)}]' w^k \right. \\ &\left. + \sum_{k=1}^{\infty} (4 - k^2|c_k(t)|^2) w^k \frac{1-w}{1+w} \right] dt. \end{aligned} \tag{4.7.7}$$

Write $z = re^{i\theta}$. From Theorem 4.6.1 (c), we have

$$\frac{\partial f_t(z)/\partial t}{f_t(z)} \;=\; 1 + \sum_{k=1}^{\infty} c_k'(t) r^k e^{ik\theta},$$

which is a Fourier expansion. By (4.1.5) the coefficients are given by

$$r^k c_k'(t) \;=\; \frac{1}{2\pi}\int_0^{2\pi} \frac{\partial f_t(z)/\partial t}{f_t(z)} e^{-ik\theta} d\theta,$$

which in turn gives

$$r^{2k} c_k'(t) \;=\; \frac{1}{2\pi}\int_0^{2\pi} \frac{\partial f_t(z)/\partial t}{f_t(z)} \bar{z}^k \, d\theta.$$

Taking the limit $r \to 1$ gives

$$c_k'(t) \;=\; \lim_{r\to 1} \frac{1}{2\pi}\int_0^{2\pi} \frac{\partial f_t(z)/\partial t}{f_t(z)} \bar{z}^k \, d\theta$$

and

$$k c_k'(t)\overline{c_k(t)} = \lim_{r\to 1} \frac{1}{2\pi}\int_0^{2\pi} \frac{\partial f_t(z)/\partial t}{f_t(z)} k\, \overline{c_k(t) z^k} d\theta.$$

Similarly, we have

$$k c_k(t)\overline{c_k'(t)} = \lim_{r\to 1} \frac{1}{2\pi}\int_0^{2\pi} \frac{\overline{\partial f_t(z)/\partial t}}{\overline{f_t(z)}} k\, c_k(t) z^k \, d\theta.$$

With these representations, equation (4.7.7) becomes

$$\begin{aligned}\Phi(z) \;=\; \int_0^{\infty} \frac{e^t w}{1-w^2} \Bigg\{ & \frac{1+w}{1-w}\left[1 + \sum_{k=1}^{\infty}\left(\lim_{r\to 1}\frac{1}{2\pi}\int_0^{2\pi} \frac{\partial f_t(z)/\partial t}{f_t(z)} k\overline{c_k(t)z^k} d\theta\right) w^k\right] \\ & + \frac{1+w}{1-w}\left[1 + \sum_{k=1}^{\infty}\left(\lim_{r\to 1}\frac{1}{2\pi}\int_0^{2\pi} \frac{\overline{\partial f_t(z)/\partial t}}{f_t(z)} k c_k(t) z^k d\theta\right) w^k\right] \\ & -2\left(\frac{1+w}{1-w}\right) + \frac{4w}{1-w} + \sum_{k=1}^{\infty} -k^2 |c_k(t)|^2 w^k \Bigg\}\, dt \end{aligned} \tag{4.7.8}$$

Denote the term inside the curly brackets in (4.7.8) by A. Then

$$A \;=\; \frac{1+w}{1-w}\left(1 + \sum_{k=1}^{\infty} d_k w^k\right), \qquad d_k \;=\; \lim_{r\to 1}\frac{1}{2\pi}\int_0^{2\pi} \frac{\partial f_t(z)/\partial t}{f_t(z)} k\, \overline{c_k(t) z^k}\, d\theta.$$

Simple calculation yields

$$A \;=\; 1 + \sum_{k=1}^{\infty} \left[2(1 + d_1 + \cdots + d_k) - d_k\right] w^k.$$

Similarly, we put

$$B \;=\; \frac{1+w}{1-w}\left(1 + \sum_{k=1}^{\infty} e_k w^k\right), \qquad e_k \;=\; \lim_{r\to 1}\int_0^{2\pi} \frac{\overline{\partial f_t(z)/\partial t}}{\overline{f_t(z)}} k\, c_k(t) z^k \, d\theta,$$

and

$$B = 1 + \sum_{k=1}^{\infty}[2(1 + e_1 + \cdots + e_k) - e_k]w^k.$$

It follows From (4.7.8) that

$$\begin{aligned}\Phi(z) \;=\; & \int_0^{\infty} \frac{e^t w}{1-w^2}\left\{1 + \sum_{k=1}^{\infty}\left(\lim_{r\to 1}\frac{1}{2\pi}\int_0^{2\pi}\frac{\partial f_t(z)/\partial t}{f_t(z)}\right.\right.\\ & \times\left.[2(1 + \cdots + k\overline{c_k(t)z^k}) - k\overline{c_k(\tau)z^k}]d\theta\right) w^k\\ & +1 + \sum_{k=1}^{\infty}\left(\lim_{r\to 1}\frac{1}{2\pi}\int_0^{2\pi}\frac{\overline{\partial f_t(z)/\partial t}}{\overline{f_t(z)}}\right.\\ & \times\left.[2(1 + \cdots + kc_k(t)z^k) - kc_k(t)z^k]d\theta\right) w^k\\ & \left.-2 + \sum_{k=1}^{\infty} -k^2|c_k(t)|^2 w^k\right\} dt \end{aligned} \tag{4.7.9}$$

To proceed further, we differentiate the equation in Theorem 4.6.1 (c) with respect to z, and obtain

$$z\frac{\partial f_t(z)/\partial z}{f_t(z)} = 1 + \sum_{k=1}^{\infty} k\, c_k(t)z^k,$$

so that (4.7.9) can be written as

$$\begin{aligned}\Phi(z) \;=\; & \int_0^{\infty}\frac{e^t w}{1-w^2}\left\{\sum_{k=1}^{\infty}\lim_{r\to 1}\frac{1}{2\pi}\int_0^{2\pi}\left[\frac{\partial f_t(z)/\partial t}{f_t(z)}\bigg/ z\frac{\partial f_t(z)/\partial z}{f_t(z)}\right]\right.\\ & \times\left(1 + \sum_{l=1}^{\infty} lc_l(t)z^l\right)[2(1 + \cdots + k\overline{c_k(t)z^k}) - k\overline{c_k(t)z^k}]\, d\theta\, w^k\\ & + \sum_{k=1}^{\infty}\lim_{r\to 1}\frac{1}{2\pi}\int_0^{2\pi}\left[\frac{\overline{\partial f_t(z)/\partial t}}{f_t(z)}\bigg/ z\frac{\overline{\partial f_t(z)/\partial z}}{f_t(z)}\right]\\ & \times\left(1 + \sum_{l=1}^{\infty} l\overline{c_l(t)z^l}\right)[2(1 + \cdots + kc_k(t)z^k) - kc_k(t)z^k]d\theta w^k\\ & \left.+ \sum_{k=1}^{\infty} -k^2|c_k(t)|^2 w^k\right\} dt. \end{aligned} \tag{4.7.10}$$

To simplify (4.7.10), we put

$$F(z,t) = \frac{\partial f_t(z)/\partial t}{z\,\partial f_t(z)/\partial z}.$$

From Theorem 4.6.1 (b), we have

$$\frac{\partial f_t(z)}{\partial t} = e^t z + \sum_{k=2}^{\infty} a_k'(t) z^k.$$

Then

$$F(z,t) = \frac{e^t + \sum\limits_{k=2}^{\infty} a_k'(t) z^{k-1}}{\partial f_t(z)/\partial z}.$$

Since $\partial f_t(z,t)/\partial z$ is holomorphic and nonvanishing in the unit disk, the function $F(z,t)$ is holomorphic. Furthermore, since $\partial f_t(z)/\partial z = e^t + \cdots$, we have $F(0,t) = 1$. (This fact will be used later.) The first inner integral in (4.7.10) is equal to

$$\begin{aligned}
&\frac{1}{2\pi}\int_0^{2\pi} F(z,t)\left[1 + \sum_{l=1}^{\infty} l\, c_l(t) z^l\right][2(1 + \cdots + k\overline{c_k(t)z^k}) - k\overline{c_k(t)z^k}]\, d\theta \\
&= \frac{1}{2\pi}\int_0^{2\pi} F(z,t)\left[1 + \cdots + k\, c_k(t) z^k - \frac{1}{2} k\, c_k(t) z^k + \frac{1}{2} k\, c_k(t) z^k + \sum_{l=k+1}^{\infty} l\, c_l(t) z^l\right] \\
&\qquad \times [2(1 + \cdots + k\overline{c_k(t)z^k}) - k\overline{c_k(t)z^k}] d\theta \\
&= \frac{1}{2\pi}\int_0^{2\pi} F(z,t) \frac{1}{2} |2(1 + \cdots + k c_k(t) z^k) - k c_k(t) z^k|^2\, d\theta \\
&+ \frac{1}{2\pi}\int_0^{2\pi} F(z,t) \frac{1}{2} k c_k(t) z^k \left[2(1 + \cdots + k\overline{c_k(t)z^k}) - k\overline{c_k(t)z^k}\right] d\theta \\
&+ \frac{1}{2\pi}\int_0^{2\pi} F(z,t) \left(\sum_{l=k+1}^{\infty} c_l(t) z^l\right) [2(1 + \cdots + k\overline{c_k(t)z^k}) - k\overline{c_k(t)z^k}]\, d\theta.
\end{aligned} \tag{4.7.11}$$

There are now three terms on the right-hand side of equation (4.7.11). In the second term, for $m \le k-1$, since $z = re^{i\theta}$ we have

$$\frac{1}{2\pi}\int_0^{2\pi} F(z,t) z^k \overline{z^m} d\theta = \frac{r^{2m}}{2\pi i}\int_{|z|=r} F(z,t) z^{k-m-1} dz = 0$$

by Cauchy's theorem. Similarly, for $m = k$ we have

$$\frac{1}{2\pi}\int_0^{2\pi} F(z,t) z^k z^k d\theta = \frac{r^{2k}}{2\pi i}\int_{|z|=r} \frac{F(z,t)}{z} dz = r^{2k} F(0,t) = r^{2k}$$

since, as noted above, $F(0,t) = 1$. Taking the limit $r \to 1$ in (4.7.10) shows that the second term on the right-hand side of (4.7.11) is $\frac{1}{2}k^2|c_k(t)|^2$. Similarly the third term in (4.7.11) is zero.

Summarizing, the first inner integral in (4.7.10) is equal to the sum of the first term in (4.7.11) and the contribution $\frac{1}{2}k^2|c_k(t)|^2$ from the second term. The second inner integral in (4.7.10) is the complex conjugate of the first inner integral, so it contributes $\frac{1}{2}k^2|c_k(t)|^2$ to (4.7.11). Summing with respect to k cancels the last series in (4.7.10). Thus (4.7.10) becomes

$$\Phi(z) = \int_0^\infty \frac{e^t w}{1-w^2} \sum_{k=1}^\infty \lim_{r\to 1} \frac{1}{2\pi} \int_0^{2\pi} \mathrm{Re}\left\{ \frac{\partial f_t(z)}{\partial t} \Big/ z\frac{\partial f_t(z)}{\partial z} \right\} \times |2(1+\cdots+kc_k(t)z^k) - k\,c_k(t)z^k|^2\, d\theta\, w^k\, dt. \tag{4.7.12}$$

Denote the term being summed in (4.7.12) by $A_k(t)$. Then (4.7.12) becomes

$$\Phi(z) = \int_0^\infty \frac{e^t w}{1-w^2} \left(\sum_{k=1}^\infty A_k(t) w^k \right) dt. \tag{4.7.13}$$

By Theorem 4.6.1 (d), we have Re $F \geq 0$, from which it follows that

$$A_k(t) \geq 0 \quad \text{for} \quad t \geq 0, \qquad k = 1, 2, \cdots . \tag{4.7.14}$$

If we show that

$$\frac{e^t w^{k+1}}{1-w^2} = \sum_{n=0}^\infty \Lambda_k^n(t) z^{n+1} \tag{4.7.15}$$

with $\Lambda_k^n(t) \geq 0$ for $t \geq 0$, then we have proved the Milin conjecture (4.5.7). Indeed, from (4.7.13) and (4.7.15), we have

$$\Phi(z) = \sum_{n=0}^\infty \left(\int_0^\infty \sum_{k=1}^\infty A_k(t) \Lambda_k^n(t) dt \right) z^{n+1}. \tag{4.7.16}$$

The function $h_n(t)$ in (4.7.3) is explicitly given by

$$h_n(t) = \sum_{k=1}^\infty A_k(t) \Lambda_k^n(t). \tag{4.7.17}$$

If $\Lambda_k^n(t) \geq 0$, then it follows from (4.7.2) and (4.7.16) that

$$\sum_{k=1}^n \left(\frac{4}{k} - k|c_k(t)|^2 \right) (n-k+1) = \int_0^\infty h_n(t) dt \geq 0.$$

To show that $\Lambda_k^n(t) \geq 0$ for $t \geq 0$, we first establish the equation

$$\frac{z}{1 - 2z(\cos^2\phi + \sin^2\phi\cos\theta) + z^2} = \frac{e^t w}{1-w^2} + 2\sum_{k=1}^\infty \frac{e^t w^{k+1}}{1-w^2} \cos\theta, \tag{4.7.18}$$

where $\sin\phi = e^{-t/2}$ and $z/(1-z)^2 = e^t w/(1-w)^2$. Note that the right-hand side of this equation is a Fourier cosine series. Thus, we may write it as

$$\frac{z}{1 - 2z(\cos^2\phi + \sin^2\phi\cos\theta) + z^2} = \frac{a_0}{2} + \sum_{k=1}^\infty a_k \cos k\theta.$$

From (4.1.5) again, we see that the coefficient a_k has the integral representation

$$a_k = \frac{2}{\pi}\int_0^\pi \frac{z\cos k\theta}{1-2z(\cos^2\phi+\sin^2\phi\cos\theta)+z^2}\,d\theta. \tag{4.7.19}$$

We first consider the special case $t = 0$; i.e. when $\sin^2\phi = 1$, $\cos^2\phi = 0$ and $w = z$ (see (4.7.4)). In this case

$$\frac{2}{\pi}\int_0^\pi \frac{w\cos k\theta}{1-2w\cos\theta+w^2}d\theta = \frac{2w^{k+1}}{1-w^2}; \tag{4.7.20}$$

see Exercise 19.

For the general case, we just need the identities

$$\begin{aligned}\frac{z\cos k\theta}{1-2z(\cos^2\phi+\sin^2\phi\cos\theta)+z^2} &= \frac{\cos k\theta}{(1-z)^2/z+2\sin^2\phi(1-\cos\theta)}\\ = \frac{\cos k\theta}{e^{-t}(1-w)^2/w+2e^{-t}(1-\cos\theta)} &= \frac{e^t w\cos k\theta}{1-2w\cos\theta+w^2}.\end{aligned}$$

Substituting this into (4.7.19), we obtain from (4.7.20)

$$a_k = \frac{2e^t w^{k+1}}{1-w^2},$$

proving (4.7.18).

Inserting (4.7.15) into (4.7.18) gives

$$\begin{aligned}&\frac{z}{1-2z(\cos^2\phi+\sin^2\phi\cos\theta)+z^2}\\ &= \sum_{n=0}^\infty \Lambda_0^n(t)z^{n+1} + 2\sum_{k=1}^\infty\sum_{n=0}^\infty \Lambda_k^n(t)z^{n+1}\cos k\theta.\end{aligned} \tag{4.7.21}$$

Here we use the generating function for the Legendre polynomials (4.6.6):

$$\frac{z}{\sqrt{1-2z(\cos^2\phi+\sin^2\phi\cos\theta)+z^2}} = \sum_{n=0}^\infty P_n(\cos^2\phi+\sin^2\phi\cos\theta)z^n \tag{4.7.22}$$

and the addition formula (4.6.12):

$$\begin{aligned}&P_n(\cos^2\phi+\sin^2\phi\cos\theta) =\\ &= [P_n(\cos\theta)]^2 + 2\sum_{k=1}^n \frac{(n-k)!}{(n+k)!}[P_n^k(\cos\theta)]^2\cos k\phi\end{aligned} \tag{4.7.23}$$

Applying (4.7.23) to (4.7.22) gives

$$\frac{z}{\sqrt{1-2z(\cos^2\phi+\sin^2\phi\cos\theta)+z^2}} \tag{4.7.24}$$

$$= \sum_{n=0}^{\infty}[P_n(\cos^2\theta)]^2 z^n + 2\sum_{n=0}^{\infty}(\sum_{k=1}^{n}\frac{(n-k)!}{(n+k)!}[P_n^k(\cos\phi)]^2\cos k\theta)z^n.$$

Since $2\cos k\theta = e^{ik\theta} + e^{-ik\theta}$, we have

$$\sum_{k=-\infty}^{\infty} \Lambda_{|k|}^n(t)e^{ik\theta} = \Lambda_0^n(t) + 2\sum_{k=1}^{\infty}\Lambda_k^n(t)\cos k\theta.$$

Therefore, equation (4.7.21) can be written as

$$\frac{1}{1-2z\left(\cos^2\phi + \sin^2\phi\cos\theta\right)+z^2} = \sum_{n=0}^{\infty}\left(\sum_{k=-\infty}^{\infty}\Lambda_{|k|}^n(t)e^{ik\theta}\right)z^n. \qquad (4.7.25)$$

Similarly,

$$\sum_{k=-n}^{n}\frac{(n-|k|)!}{(n+|k|)!}[P_n^{|k|}(\cos\phi)]^2 e^{ik\theta}$$
$$= [P_n(\cos\phi)]^2 + 2\sum_{k=1}^{n}\frac{(n-k)!}{(n+k)!}[P_n^k(\cos\phi)]^2\cos k\theta,$$

and (4.7.24) becomes

$$\frac{1}{\sqrt{1-2z(\cos^2\phi+\sin^2\phi\cos\theta)+z^2}} = \sum_{n=0}^{\infty}\left(\sum_{k=-n}^{n}\frac{(n-|k|)!}{(n+|k|)!}[P_n^{|k|}(\cos\phi)]^2e^{ik\theta}\right)z^n.$$

Squaring both sides of the last equation gives

$$\begin{aligned}
&\frac{1}{1-2z(\cos^2\phi+\sin^2\phi\cos\theta)+z^2}\\
&= \sum_{n=0}^{\infty}\sum_{m=0}^{n}\sum_{j=-m}^{m}\sum_{l=-n+m}^{n-m}\frac{(m-|j|)!}{(m+|j|)!}\frac{(n-m-|l|)!}{(n-m+|l|)!}\\
&\quad\times[P_m^{|j|}(\cos\phi)]^2[P_{n-m}^{|l|}(\cos\phi)]^2e^{i(j+l)\theta}z^n\\
&= \sum_{n=0}^{\infty}\sum_{k=-n}^{n}\sum_{j=-n}^{n}\sum_{m=|j|}^{n-|k-j|}\frac{(m-|j|)!}{(m+|j|)!}\frac{(n-m-|k-j|)!}{(n-m+|k-j|)!}\\
&\quad\times[P_m^{|j|}(\cos\phi)]^2[P_{n-m}^{|k-j|}(\cos\phi)]^2e^{ik\theta}z^n. \qquad (4.7.26)
\end{aligned}$$

Comparing (4.7.25) with (4.7.26) gives

$$\Lambda_k^n(t) = 0 \qquad \text{for} \quad k > n,$$

since the summation on k in equation (4.7.25) ranges from $-\infty$ to $+\infty$ while the same summation in (4.7.26) ranges from $-n$ to n. For $k = 0, 1, \cdots, n$, we have

$$\Lambda_k^n(t) = \sum_{j=-n}^{n} \sum_{m=|j|}^{n-|k-j|} \frac{(m-|j|)!(n-m-|k-j|)!}{(m+|j|)!(n-m+|k-j|)!} [P_m^{|j|}(\cos\phi)]^2 [P_{n-m}^{|k-j|}(\cos\phi)]^2,$$

which is clearly non-negative.

To this point we have proved the inequality $|b_n| \le n$, $n = 2, 3, \ldots$. If equality holds for any n, then this argument shows that it holds for all n. In particular, it holds for $n = 2$, so Bieberbach's theorem implies that f is a Koebe function. This completes Weinstein's proof of Bieberbach's conjecture.

Exercises

1. Prove that $K(\partial\mathbb{D}) = (-\infty, -1/4]$, where K is the Koebe function.
2. Prove that the function f_2 of (4.1.6) is single-valued.
3. Use the Koebe function to prove that the bounds in (4.1.12) and (4.1.14) are sharp.
4. (a) Prove that if f belongs to S, then f' has no zeros in $\mathbb{D}$.
   (b) Is the converse true?
5. Suppose that $h \in S$ is an odd function of z. Show that $h(z) = \sqrt{f(z^2)}$ where f belongs to S.
6. (a) Suppose $f \in S$, and m is a positive integer. Show that, for a suitable choice of the m-th root, the function

$$g(z) = f(z^m)^{1/m} \tag{4.7.27}$$

   belongs to S, and has the symmetry property $g(e^{2\pi i/m}z) = g(z)$.
   (b) Conversely, show that if $g \in S$ has the preceding symmetry property, then g has the form (4.7.27) for some $f \in S$. Is f unique?
7. Prove the sharper version of the Koebe one-quarter theorem: if $f \in S$ omits the value w; then $|w| \ge 1/(2 + |a_2|)$.
8. Suppose $f \in S$. Prove that for any compact set $K \in \mathbb{D}$ and any $\varepsilon > 0$ there is a polynomial p such that $p\big|_{\mathbb{D}} \in S$ and $|p(z) - f(z)| < \varepsilon$ for $z \in K$. In other words, polynomials are dense in S.
9. Let $\{\Omega_n\}$ be a sequence of domains, and consider the family of domains Ω with the property that any compact subset of Ω is contained in all but finitely many Ω_n. Prove that any domain that is the union of domains with this property also has this property.
10. Prove Corollary 4.3.3.
11. Prove the assertion about the form of the coefficients of g_τ in (4.4.1).
12. Suppose that the function $k(t)$ in Loewner's equation (4.4.3) is identically -1. Show that the function f generated by (4.4.3) is the Koebe function. In fact, given $z \in \mathbb{D}$, let $g(t) = g_t(z)$. Write (4.4.3) in the form $[\log F]'(g)\, g' = 1$ for a certain choice of the function F, so that

$$F(g(t)) = ce^t$$

where the constant of integration c depends on z. The initial condition $g(0) = z$ determines $c(z)$, and the asymptotic condition $e^t g(t) \to f(z)$ determines $f(z)$.

13. (a) Suppose that the function $k(t)$ in (4.4.3) generates $f \in S$. Show that for any real θ, the function $e^{i\theta}k(t)$ generates the function

$$f_\theta(z) = e^{-i\theta} f(e^{i\theta} z).$$

(b) Show that f_θ also belong to S and is a slit map.

14. Suppose that $k(t)$ in (4.4.3) is constant. What f is generated? (Remember that $f(t) \in \partial\mathbb{D}$.)
15. Prove (4.6.4).
16. Prove that $P_n(-1) = 1$ and $P_n(1) = (-1)^n$.
17. Differentiate both sides of (4.6.7) with respect to s and equate coefficients of s^n to prove the recursion relation

$$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_n(x).$$

18. Prove (4.6.8). Hint: use (4.6.8) to derive a partial differential equation for $G(x, s)$, and show that the right-hand side of (4.6.7) satisfies that equation.)
19. Prove (4.7.20). Hint: The integrand is even in θ, so convert the left side into an integral from $-\pi$ to π, use the identity $1 - 2w\cos\theta + w^2 = (1 - we^{i\theta})(1 - we^{-i\theta})$ to write the integrand as the sum of two (two-sided) series in $e^{i\theta}$, integrate term-by-term, and sum.)

Remarks and further reading

Standard references for univalent functions, pre-de Branges, are Duren [64] and Pommerenke [170]. De Brange's original manuscript ran to 385 typed pages and made much use of ideas from the theory of operators in Hilbert space. As recounted in [52], de Brange's participation in a seminar in Leningrad that included Milin led to the simplified proof in [52]. Though much shorter, this proof still gives some indication of the operator-theoretic considerations that led to deBranges's new approach to the problem. Another short version is due to Fitzgerald and Pommerenke [73], and has even made its way into textbooks, e.g. [47] and [102].

Some of the history surrounding de Branges's proof is recounted in the symposium volume [14]. For a comprehensive current account of the theory, see Thomas, Tuneski, and Vasudevarao [205]. For a somewhat different approach, see Rosenblum and Rovnyak [181].

Weinstein's proof of de Brange's theorem, presented here, depends on a positivity result involving Legendre polynomials P_n. These are among the simplest cases of Jacobi polynomials $P_n^{(\alpha,\beta)}$. De Brange's argument, and that of Fitzgerald and Pommerenke, depended on a positivity result for certain sums of Jacobi polynomials proved by Askey and Gasper [12]. To prove their result, Askey and Gasper established positivity for a still more esoteric class of hypergeometric functions:

$$ {}_3F_2(n-r, r+n+2, n+\tfrac{1}{2} \ : \ 2+1, n+\tfrac{3}{2} \ : \ s) \ > \ 0, \qquad 0<s<1. $$

Wilf [216] has pointed out that Weinstein's argument actually gives an independent proof of the non-negativity of the Askey–Gasper polynomials.

Chapter 5
Harmonic and subharmonic functions; the Dirichlet problem

Harmonic and subharmonic functions play an important role in many developments in complex analysis. As we noted in Section 1.9, a real harmonic function is, locally, the real part of a holomorphic function and has some of the same properties.

The *Dirichlet problem* is the problem of finding a function that is harmonic on a given domain U and has prescribed values on the boundary ∂U. An important special case with $U = \mathbb{D}$ has an explicit solution given by Poisson's integral formula. This formula is derived in Section 5.1. A number of consequences, such as maximum principles and the Harnack inequalities, are worked out in Sections 5.1 and 5.2.

Subharmonic functions and Perron's principle provide a mechanism of attack for the Dirichlet problem in a general domain. These are introduced in Section 5.3 and applied to the solution of the Dirichlet problem in Section 5.4. An alternative approach to the Dirichlet problem is outlined in Section 5.5.

The results and techniques introduced in this chapter lead eventually to the uniformization theorem for Riemann surfaces, the subject of Chapters 6 and 7.

5.1 Harmonic functions and the Poisson integral formula

As noted above, the *Dirichlet problem* is the problem of finding a function that is harmonic in a given domain and that has prescribed values on the boundary of the domain. As we shall see, if the domain is a disk, the problem has a very explicit solution.

Theorem 5.1.1. *Suppose that $D \subset \mathbb{C}$ is a disk and f is a continuous real function on the boundary ∂D. Then there is a unique function u, harmonic in D and continuous on the closure of D, such that $u = f$ on ∂D.*

R. Beals and R. S. C. Wong, *More Explorations in Complex Functions*, Graduate Texts in Mathematics 298, https://doi.org/10.1007/978-3-031-28288-1_5

Proof: In view of Corollary 1.9.3, it is enough to consider $D = \mathbb{D}$, the unit disk. The strategy of the proof is to assume we have a solution, derive an explicit formula that it must satisfy, and then show that the formula does indeed provide the solution.

Suppose that u is a solution. By Proposition 1.9.1, there is a function holomorphic $g : \mathbb{D} \to \mathbb{C}$ having real part u. We may assume that $g(0) = u(0)$. The function g has an expansion

$$g(z) \;=\; \sum_{n=0}^{\infty} \alpha_n z^n \tag{5.1.1}$$

that converges uniformly on disks $D_r(0)$, $r < 1$. The real part u has the expansion

$$\begin{aligned} u(re^{i\theta}) &= \sum_{n=0}^{\infty} r^n \,\mathrm{Re}\,(\alpha_n e^{in\theta}) \\ &= \sum_{n=-\infty}^{\infty} a_n r^{|n|} e^{in\theta}, \qquad a_n \;=\; \begin{cases} \alpha_n/2, & n > 0; \\ \mathrm{Re}\,\alpha_0, & n = 0; \\ \bar{\alpha}_n/2, & n < 0. \end{cases} \end{aligned} \tag{5.1.2}$$

Given $\varepsilon > 0$, the dilated function $g_\varepsilon(z) = g(z/(1+\varepsilon))$ is holomorphic in $D_{1+\varepsilon}(0)$. By assumption, the restriction of $u_\varepsilon = \mathrm{Re}\, g_\varepsilon$ to $\partial\mathbb{D}$ converges uniformly to f as $\varepsilon \to 0$. Convergence of the sum shows that

$$u_\varepsilon(e^{i\theta}) \;=\; u\left(\frac{e^{i\theta}}{1+\varepsilon}\right) \;=\; \sum_{n=-\infty}^{\infty} a_n (1+\varepsilon)^{-|n|} e^{in\theta}.$$

By assumption, this series converges uniformly, so we can identify the coefficients $a_n(1+\varepsilon)^{-|n|}$ by integrating term by term, using the identity

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{in\theta} e^{-im\theta}\, d\theta \;=\; \begin{cases} 1 \;\; if \;\; m = n; \\ 0 \;\; if \;\; m \neq n. \end{cases} \tag{5.1.3}$$

This gives

$$\frac{a_n}{(1+\varepsilon)^{|n|}} \;=\; \frac{1}{2\pi}\int_{-\pi}^{\pi} u_\varepsilon(\varphi) e^{-in\varphi}\, d\varphi.$$

Convergence of u_ε to f on $\partial\mathbb{D}$ gives

$$a_n \;=\; \frac{1}{2\pi}\int_{-\pi}^{\pi} f(e^{i\theta}) e^{-in\varphi}\, d\varphi. \tag{5.1.4}$$

Note that

$$|a_n| \;\le\; \frac{1}{2\pi}\int_{-\pi}^{\pi} |f(e^{i\theta})|\, d\theta.$$

Therefore the series

$$\sum_{n=-\infty}^{\infty} a_n r^{|n|} e^{in\theta} \tag{5.1.5}$$

converges uniformly for $0 \le r \le R < 1$. Since $a_{-n} = \bar{a}_n$, the terms

$$a_n e^{in\theta} + a_{-n} e^{-in\theta}$$

are real and harmonic, so (5.1.4) and (5.1.5) together define a function u that is real and harmonic in $\mathbb{D}$.

We have shown that if u is a solution to the Dirichlet problem with boundary value f, then it is necessarily given on $\mathbb{D}$ by the formula

$$u(re^{i\theta}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} P_r(\theta - \varphi) f(e^{i\varphi})\, d\varphi, \qquad 0 \le r < 1, \tag{5.1.6}$$

where P_r is the *Poisson kernel*

$$\begin{aligned} P_r(\theta) = \sum_{n=-\infty}^{\infty} r^{|n|} e^{in\theta} &= 1 + \sum_{1}^{\infty} \left\{ (re^{i\theta})^n + (re^{-in\theta})^n \right\} \\ &= 1 + \left\{ \frac{re^{i\theta}}{1 - re^{i\theta}} + \frac{re^{-i\theta}}{1 - re^{-i\theta}} \right\} \\ &= \frac{1 - r^2}{1 - 2r\cos\theta + r^2}. \end{aligned} \tag{5.1.7}$$

This proves uniqueness. We have also shown that the function u defined by (5.1.6) is harmonic in $\mathbb{D}$.

To see that u is continuous up to the boundary $\partial\mathbb{D}$ and equal to f on $\partial\mathbb{D}$, note that P_r, $0 \le r < 1$ has the properties

$$\begin{aligned} &\text{(i)}\ \ P_r > 0; \\ &\text{(ii)}\ \ \frac{1}{2\pi}\int_{-\pi}^{\pi} P_r(\theta)\, d\theta = 1; \\ &\text{(iii)}\ \ \lim_{r \to 1} \frac{1}{2\pi}\int_{\delta \le |\theta| \le \pi} P_r(\theta) = 0 \end{aligned}$$

for each $\delta > 0$; Exercise 6. Using these properties and the continuity of f, it is not difficult to prove that $u(re^{i\theta})$ converges uniformly to $f(\theta)$ as $r \to 1-$. (See the proof of Theorem 2.9.1.) □

The formula (5.1.6) is known as the *Poisson integral formula.*

Corollary 5.1.2. (Weierstrass!approximation theorem) *If $f : \partial\mathbb{D} \to \mathbb{C}$ is continuous, then for any $\varepsilon > 0$ there is a trigonometric polynomial, i.e. a function of the form*

$$g(\theta) = \sum_{|k| \le m} a_k e^{ik\theta}, \tag{5.1.8}$$

such that $|g(\theta) - f(e^{i\theta})| < \varepsilon$, all θ.

Proof. In view of the previous discussion, the functions

$$u(re^{i\theta}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} P_r(\theta-\varphi) f(\varphi)\,d\varphi = \sum_{-\infty}^{\infty} a_n r^{|n|} e^{in\theta},$$

where the bounded sequence $\{a_n\}$ is defined by (5.1.4), converge to f uniformly as $r \to 1$. For any given $0 < r < 1$, the partial sums of the series on the right are trigonometric polynomials, and they converge uniformly to $u(re^{i\theta})$. □

Remark. The other well-known Weierstrass approximation theorem, that a continuous function on a bounded closed interval can be approximated uniformly by polynomials, is a consequence. In fact the interval can be rescaled to $[0, \pi]$, and the function reflected about π so that $f(2\pi) = f(0)$ and f can be considered as an element of $C(\partial\mathbb{D})$. Then f can be approximated within $\varepsilon/2$ by a trigonometric polynomial (5.1.8), and each $a_k e^{ik\theta}$ can be approximated within $\varepsilon/4m$ by a polynomial, by taking enough terms of the series expansion of $e^{ik\theta}$.

Corollary 5.1.3. (Mean value property) *If u is harmonic in a neighborhood of a point $z \in C$, then for sufficiently small $r > 0$,*

$$u(z) = \frac{1}{2\pi}\int_{-\pi}^{\pi} u(z + re^{i\theta})\,d\theta, \tag{5.1.9}$$

i.e. u(z) is the mean value of u over any sufficiently small circle centered at z.

Proof: After a translation and dilation, we may assume that $z = 0$ and $r = 1$. In this case, the result follows from Theorem 5.1.1. □

Corollary 5.1.4. (strict maximum principle) *If u is harmonic in a bounded domain $U \subset \mathbb{C}$ and U has a local maximum at a point $z \in U$, then U is constant.*

Proof: The mean value property and the assumption that u has a local maximum at z imply that u has this same value on each sufficiently small circle centered at z. Thus u is constant, hence holomorphic, near z. If w is any other point of U we may find a curve that joins z to w and a simply connected neighborhood V of the curve with $V \subset U$. By uniqueness of analytic continuation, u is constant in V, so $u(w) = u(z)$. □

Let us pass to consideration of the Dirichlet problem for a Jordan domain: a domain in $\mathbb{C}$ whose boundary is a curve with no self-intersections.

Theorem 5.1.5. *Suppose that $\Omega \subset \mathbb{C}$ is a Jordan domain with boundary Γ. For any continuous function $f : \Gamma \to \mathbb{C}$, the Dirichlet problem has a unique solution.*

Proof. By the Riemann mapping theorem, there is a conformal map Φ that maps $\mathbb{D}$ onto Ω. By Theorem 2.6.1, Φ extends to a bijective continuous map from the closure to the closure. Therefore Φ^{-1} and Φ can be used to transfer the Dirichlet problem for Ω to the Dirichlet problem for $\mathbb{D}$. The details are left as Exercise 4. □

More general domains are considered in later sections.

We know that a real function u that is harmonic in $\mathbb{D}$ is the real part of a function ϕ that is holomorphic in $\mathbb{D}$. Moreover, if we require that $\phi(0)$ be real, then ϕ is unique. We may extend the Poisson integral formula to exhibit ϕ in the case when $u = f$ on $\partial\mathbb{D}$. In fact for $z \in \mathbb{D}$,

$$\operatorname{Re} \frac{e^{i\theta} + z}{e^{i\theta} - z} = \frac{1 - |z|^2}{|e^{i\theta} - z|^2}.$$

If $z = re^{i\varphi}$, then the last expression on the right is $P_r(\theta - \varphi)$. Since the quotient on the left is holomorphic in z, $z \in \mathbb{D}$, the proof of Theorem 5.1.1 also proves the following *extended Poisson formula*:

Theorem 5.1.6. *If $f : \partial\mathbb{D} \to \mathbb{R}$ is continuous, then the function*

$$\phi(z) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(e^{i\theta}) \frac{e^{i\theta} + z}{e^{i\theta} - z} \, d\theta$$

is holomorphic in $\mathbb{D}$ and the real part is continuous and equal to f at the boundary.

5.2 Harnack's principle; removable singularities

We begin with a simple consequence of the Poisson formula.

Lemma 5.2.1. *Suppose that u is harmonic and non-negative in $\mathbb{D}$. Then for $z \in \mathbb{D}$*

$$\frac{1-r}{1+r} \cdot u(0) \le u(z) \le \frac{1+r}{1-r} \cdot u(0), \qquad r = |z|. \tag{5.2.1}$$

Proof: Suppose first that u is continuous on the closure of $\mathbb{D}$. The Poisson kernel (5.1.7) satisfies

$$\frac{1-r}{1+r} = \frac{1-r^2}{(1+r)^2} \le \frac{1-r^2}{1-2r\cos\theta + r^2} \le \frac{1-r^2}{(1-r)^2} = \frac{1+r}{1-r}.$$

Then the Poisson formula (5.1.6) gives (5.2.1). If u is not continuous on the closure, we approximate u as in the proof of Theorem 5.1.1. □

The inequalities (5.2.1) are the simplest case of the *Harnack inequalities* for solutions of elliptic equations.

Theorem 5.2.2. *(***Harnack's principle***) If $\{u_n\}$ is a sequence of harmonic functions on a domain U, with $u_1 \le u_2 \le u_3 \ldots$, then $u_\infty(p) = \lim_{n\to\infty} u_n(p)$ is either harmonic or identically infinite.*

Proof. Given $p \in U$, let us rescale coordinates, for convenience, so that $p = 0$ and $\mathbb{D} \subset U$. Let $u_\infty(0) = \lim_n u_n(0)$. If $u_\infty(0)$ is finite, then inequalities (5.2.1) show that $\lim_{n\to\infty} u_n$ is finite in $\mathbb{D}$. Similarly, if $u_\infty(0) = \infty$, these inequalities show that the limit is identically ∞ in $\mathbb{D}$. Since domains are, by assumption, connected, this implies that the set where $\lim_{n\to\infty} u_n(p) = \infty$ is either empty or all of U. □

An important example of a harmonic function is $\log|z|$ in the punctured plane $\mathbb{C} \setminus \{0\}$. In fact $\log|z| = \mathrm{Re}\, \log z$ for any determination of $\log z$, $z \neq 0$.

As in the case of holomorphic functions, there is a removable singularity theorem for harmonic functions. The singularity of $\log|z|$ at the origin is obviously not removable, but $\log|z|$ allows us to prove a type of one-sided singularity result that will be useful later.

Lemma 5.2.3. *Suppose that u is real-valued and harmonic in the punctured disk $\mathbb{D} \setminus \{0\}$ and is continuous at the boundary of $\mathbb{D}$. If u is bounded above, then $u \le v$, where v is the function harmonic on $\mathbb{D}$, continuous on the closure, and equal to u on the boundary.*

Proof: We may subtract u from v and reduce to the case that $u \equiv 0$ on the boundary of $\mathbb{D}$. Then $v \equiv 0$, and we want to show that $u \le 0$. Let $h > 0$ be an upper bound for u. For any $0 < r < 1$, let

$$u_r(z) = \frac{h \log|z|}{\log r}.$$

Then u_r is harmonic and $u \le u_r$ on the boundary of the annulus $\{z : r < |z| < 1\}$, so by the maximum principle $u \le u_r$ on the annulus. As $r \to 0$, $u_r \to 0$ pointwise on the punctured disk, so $u \le 0$ on the punctured disk. □

Corollary 5.2.4. *If u is bounded and harmonic in a punctured neighborhood of a point, then u extends to be harmonic in a full neighborhood of that point.*

Proof. Apply Lemma 5.2.3 to u and to $-u$. □

5.3 Subharmonic functions and Perron's principle

A real-valued function u defined on a domain U in $\mathbb{C}$ is said to be *subharmonic* if in each coordinate disk D with closure in U, $u \le h$, where h is the harmonic function such that $h = u$ on ∂D.

It may be helpful in the arguments that follow to think of the one-dimensional analogue. A real-valued solution of $u_{xx} = 0$ is a linear function $u(x) = \alpha x + \beta$. A convex function $v : (a, b) \to \mathbb{R}$ is characterized by the property that for any subinterval (c, d), $a < c < d < b$, if u is linear, and if $v \le u$ at the endpoints c, d, then $v \le u$ on all of (c, d); see the left part of Figure 5.1.

If u and v are subharmonic on a domain, then so is the maximum $u \vee v$,

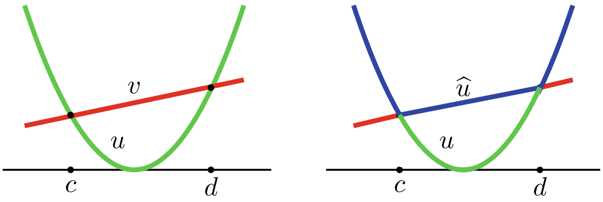

Fig. 5.1 One-dimensional analogues of subharmonicity and of harmonic regularization.

$$[u \vee v](p) = \max\{u(p), v(p)\}.$$

If u is subharmonic in U and p_0 belongs to U, a *harmonic regularization* $\hat{u}$ of u at p_0 is obtained by replacing u in a coordinate neighborhood $D = \{p : |p - p_0| < r\}$ centered at p_0 by the harmonic function that agrees with u on ∂D. The new function $\hat{u}$ is subharmonic and $u \le \hat{u}$. For the analogous construction in the one-dimensional case, see the right part of Figure 5.1.

The usefulness of subharmonic functions for attacking the Dirichlet problem and proving the uniformization theorem was established by Perron [166]. A *Perron family* is a non-empty family $\mathscr{F}$ of subharmonic functions such that:

(a) if u and v belong to $\mathscr{F}$ then so does $u \vee v$;

(b) if u belongs to $\mathscr{F}$, so does each harmonic regularization of u.

Theorem 5.3.1. (Perron's principle) *Suppose that $\mathscr{F}$ is a Perron family of functions on a domain U. Then $\bar{u} = \sup\{u : u \in \mathscr{F}\}$ is either harmonic or identically infinite.*

Proof: Let V be a coordinate neighborhood in U. Suppose that p, q are two points of V. We may choose a sequence in $\mathscr{F}$ that converges to $\bar{u}(p)$ and another that converges to $\bar{u}(q)$. Taking advantage of properties (a) and (b) of the definition, we may replace these with a single sequence that is non-decreasing at p and at q and is harmonic in V. By Theorem 5.2.2, $\bar{u}(p)$ is infinite if and only if $\bar{u}(q)$ is infinite. Since p and q were arbitrary, $\bar{u}$ is either infinite in all of V or finite in all of V. The result follows from connectedness of U. □

Suppose that U is a bounded domain with boundary ∂U, and that $g : \partial U \to \mathbb{R}$ is continuous. Perron's approach to the Dirichlet problem was to define the family $\mathscr{F}(g)$ to consist of all subharmonic functions u that are continuous on the closure $\overline{U}$ and $\le g$ on the boundary. This is clearly a Perron family. Each $u \in \mathscr{F}(g)$ is bounded by $\sup g$, so the supremum u is harmonic. We shall refer to u as the *Perron function* for the Dirichlet problem for the pair (U, g). The question is whether the Perron solution is a solution, i.e. whether $u = g$ on ∂U. This is not necessarily the case, indeed there may not be a solution: see Exercise 11.

5.4 Regular points and the solution of the Dirichlet problem

A boundary point p_0 of a domain U is said to have a *barrier* if there is a subharmonic function v such that v is continuous on $\overline{U}$, $v \leq 0$, and $v = 0$ only at p_0. A local barrier can be converted to a barrier:

Lemma 5.4.1. *If p_0 is a boundary point of U and there is a continuous, subharmonic function $u \leq 0$ in the intersection of a neighborhood of p_0 with $\overline{U}$, then p_0 has a barrier.*

Proof: Choose $r > 0$ so that u is defined on $D_r(p_0) \cap \overline{U}$. Let $c = \sup\{u(p) : |p - p_0| = r\}$, and let $v = \max\{u, c\}$ on the domain of u and $v = c$ on the remainder of $\overline{U}$. □

A boundary point p_0 of a domain is said to be *regular* if there exists a barrier at p_0. The extreme example of a point that is not regular is an isolated point of the boundary; see Exercise 12. See Figure 5.2.

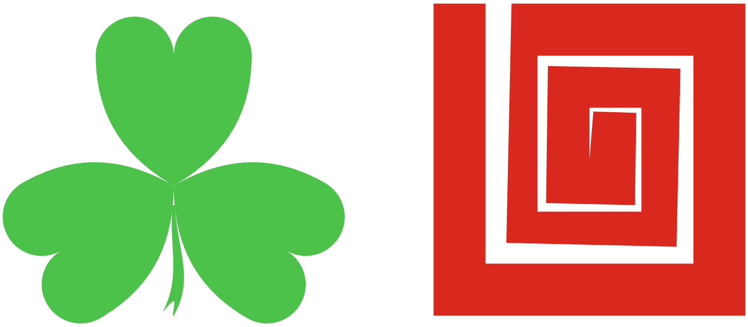

Fig. 5.2 Two domains for which every boundary point is regular.

As we shall see, if u is the Perron function for the pair (U, g), then $u \to g$ at each regular point of ∂U. This is not very useful unless we can identify at least some class of regular points. How does one tell when a barrier, or local barrier, exists? The perfect candidate for a local barrier would seem to be

$$v(p_0 + re^{i\theta}) = \operatorname{Re} \frac{1}{\log(re^{i\theta})} = \frac{\log r}{(\log r)^2 + \theta^2}. \tag{5.4.1}$$

This function is harmonic and negative for $0 < r < 1$, but θ^2 is not single-valued. However this suggests a way to remedy the situation.

Proposition 5.4.2. *Suppose that p_0 is a boundary point of a bounded domain U and suppose that for some $0 < R < 1$ and some real θ_0 the line segment*

$$L \;=\; \{p = p_0 + re^{i\theta_0} : 0 < r < R\} \qquad (5.4.2)$$

is contained in the complement of $\overline{U} \setminus \{p_0\}$. *Then* p_0 *is a regular point.*

Proof: In this case there is a single-valued branch of $\log(p - p_0)/R$, holomorphic on the complement of the segment (5.4.2) in $D_R(p_0)$, so (5.4.1) is a local barrier and Lemma 5.4.1 applies. □

Remark. It is clear that in place of a straight segment in the complement of $\overline{U} \setminus p_0$, it is enough to have any C^1 curve that has p_0 as one endpoint and otherwise lies in the complement of $\overline{U}$.

Theorem 5.4.3. (Perron) *If* U *is a bounded domain,* $g : \partial U \to \mathbb{R}$ *a continuous function, and* p_0 *a regular point of* ∂U, *then the Perron function* u *for* (U, g) *converges to* $g(p_0)$ *at* p_0.

Proof: Let $\mathscr{F}(U, g)$ be the Perron family. Given $\varepsilon > 0$, there is a $\delta > 0$ such that $|g(p) - g(p_0)| < \varepsilon$ if $g(p)$ is defined and $|p - p_0| < \delta$. Let v be a barrier at p_0. Let $||g|| = \max |g|$. For sufficiently large M,

$$Mv + 2||g|| \;<\; 0 \quad \text{on } \partial U \setminus D_\delta(p_0).$$

Let $w \;=\; Mv + g(p_0) - \varepsilon$. This function is subharmonic, continuous on ∂U, and $w(p_0) \;=\; g(p_0) - \varepsilon$. In $D_\delta(p_0) \cap \overline{U}$,

$$w \;=\; Mv + g(p_0) - \varepsilon < g(p_0) - \varepsilon \;\le\; g$$

by the choice of δ. By the choice of M, on the complement of $\overline{U} \setminus D_\delta(p_0)$ we have

$$w \;=\; Mv + 2||g|| + g(p_0) - 2||g|| - \varepsilon < g(p_0) - 2||g|| - \varepsilon < g.$$

Therefore w belongs to $\mathscr{F}(U, g)$. Note that $w(p_0) \;=\; g(p_0) - \varepsilon$. We have shown that the Perron function u satisfies

$$\liminf_{p \to p_0} u(p) \;\ge\; g(p_0) - \varepsilon. \qquad (5.4.3)$$

Let us apply this argument in the case of $-g$ to produce $w^* \in \mathscr{F}(u, -g)$ such that $w^* \ge -g - \varepsilon$. Now let v be any element of $\mathscr{F}(U, g)$. Then $v + w^*$ is continuous on $\overline{U}$, harmonic in U, and ≤ 0 on ∂U. Therefore $v \le -w^*$ in U. Since $v \in \mathscr{F}(U, g)$ was arbitrary, $u \le -w^*$. Therefore

$$\limsup_{p \to p_0} u(p) \;\le\; -\liminf_{p \to p_0} w^*(p) \;\le -(-g(p_0) - \varepsilon) \;=\; g(p_0) + \varepsilon. \qquad (5.4.4)$$

Since $\varepsilon > 0$ is arbitrary, (5.4.3) and (5.4.4) together show that u has limit $g(p_0)$ at p_0. □

5.5 The L^2 approach to the Dirichlet problem

Given a bounded domain Ω, consider the space $C^1(\overline{\Omega})$ of real functions u such that u and its first derivatives are continuous on the closure $\overline{\Omega}$. The associated *Dirichlet integral* is

$$D(u) \;=\; \iint_{\Omega} [u_x^2 + u_y^2]\, dx\, dy. \tag{5.5.1}$$

In physical problems, integrals like this occur as *energy integrals*, e.g. as the kinetic energy of a vibrating surface. A natural problem is to try to minimize $D(u)$ among those functions on Ω that have a specified value f on the boundary Ω. This is a problem in the calculus of variations. Let $\mathscr{F}$ be the family of C^1 functions that have finite Dirichlet integral and equal to f on the boundary. If this family is not empty, then there is a sequence $\{u_n\} \subset \mathscr{F}$ such that

$$\lim_{n\to\infty} D(u_n) \;=\; \inf_{u\in\mathscr{F}} D(u).$$

The terms u_n can be viewed as elements of the Hilbert space $\mathbf{H}$ that has inner product

$$\langle u, v\rangle \;=\; \iint_{\Omega} [u_x v_x + u_y v_y]\, dx\, dy.$$

(Note that $\langle u, u\rangle = 0$ if and only if u is constant. Therefore elements of $\mathbf{H}$ are determined only up to an additive constant. Of course fixing the value at the boundary fixes the constant.) We might expect the sequence $\{u_n\}$ to converge to a unique element $u \in \mathbf{H}$.

A standard argument from the calculus of variations puts a constraint on such an element u. If w is any element of $C^1(\Omega)$ that vanishes on the boundary, then we should have, for all ε,

$$D(u) \;\le\; D(u+\varepsilon w) \;=\; \langle u+\varepsilon w, u+\varepsilon w\rangle \;=\; D(u) + 2\varepsilon\langle u, w\rangle + \varepsilon^2 D(w).$$

Differentiating the last expression with respect to ε at $\varepsilon = 0$, we see that the necessary condition is that $\langle u, v\rangle = 0$. Therefore, formally,

$$0 \;=\; \langle u, w\rangle \;=\; \iint_{\Omega} [u_x w_x + u_y w_y] \;=\; -\iint_{\Omega} (u_{xx} + u_{yy}) w, \tag{5.5.2}$$

where the (formal) integration by parts is (formally) justified by the assumption that $w = 0$ on the boundary. Since (5.5.1) is supposed to hold for *each* such w, the conclusion is that (in some sense) u is harmonic in Ω and is the solution of the Dirichlet problem for the pair (Ω, f).

To show that the sequence $\{u_n\}$ above does converge in $\mathbf{H}$ and that the limit u is continuous on Ω and equal to f on the boundary requires some assumptions on the nature of the boundary. If this is the case, though, one can show that u is indeed harmonic in Ω. In fact integration by parts in (5.5.2) can be done in the other direction:

$$0 = -\iint_{\Omega} u\,(w_{xx} + w_{yy}), \quad w \in C^2(\Omega),\ w = 0 \text{ on } \partial\Omega.$$

This says that u is a *weak solution* of $\Delta u = 0$, and is therefore an actual solution of $\Delta u = 0$: see Theorem 2.9.4.

This approach to the Dirichlet problem has roots in the work of Gauss, Green, Dirichlet, Riemann, Schwarz, and Hilbert, and was brought to fruition by Weyl [214]. (Weyl's work can be thought of as the beginning of the theory of distributions.) Some regularity of the boundary is needed. In fact Prym [174] gave an example of a pair (U, g) such that for any f that is continuous on $\overline{U}$ and equal to g on ∂U, the formal integral $\langle f, f\rangle$ is infinite.

Exercises

1. Prove that the harmonic function u and the function v of (1.9.2) satisfy the Cauchy–Riemann equations.
2. Suppose that $f : \Omega_1 \to \Omega_2$ is holomorphic and $u : \Omega_2 \to \mathbb{R}$ is harmonic. Show that $u \circ f$ is harmonic.
3. Prove that if u is harmonic on the half-disk $\mathbb{D}_+ = \{z : |z| < 1, \operatorname{Im} z > 0\}$, continuous on the closure of $\mathbb{D}_+$, and vanishes on $[-1, 1]$, then u can be continued as a harmonic function to all of $\mathbb{D}$, with $u(\bar{z}) = -u(z)$.
4. Fill in the details in the proof of Theorem 5.1.5.
5. Prove that the assumption in Theorem 1.7.2 that f is continuous up to I and $|f(z)| = 1$ on I can be replaced by the weaker assumption that $|f(z)| \to 1$ as z approaches I. This is the *Schwarz reflection principle*, which plays a major role in the study of conformal mapping.
6. Complete the proof of Theorem 5.1.1 by using the properties (i), (ii), (iii) of the Poisson kernel to prove that $u_n(re^{i\theta}) \to f(e^{i\theta})$ as $r \to 1$.
7. Use the Cayley transform and its inverse to find a solution to the problem: u harmonic in $\mathbb{H}$, $u = f$ on the boundary $\mathbb{R}$, where f is a bounded continuous function on $\mathbb{R}$. What is the value of u at $z = i$?
8. Show that the boundedness condition in Lemma 5.2.3 can be weakened to $\max\{u(z), 0\} = o(-\log|z|)$ as $|z| \to 0$.
9. Show that the boundedness condition in Corollary 5.2.4 can be weakened to $|u(z)| = o(-\log|z|)$ as $|z| \to 0$.
10. Suppose $K \subset \mathbb{C}$ is the union of finitely many disks. Show that the Dirichlet problem is solvable for K.
11. Use Corollary 5.2.4 to show that the Dirichlet problem on the punctured disk $U = \mathbb{D} \setminus \{0\}$ may not have a solution.
12. Suppose that p_0 is an isolated point of the boundary of a bounded domain. Show that p_0 is not a regular point.
13. (a) Suppose that $\Omega \subset \mathbb{C}$ is bounded and $p_0 \in \Omega$. Show (without using conformal mapping) that there is a function u, harmonic in $\Omega \setminus \{p_0\}$, such that

$$u + \log|z - p_0|$$

is harmonic near p_0.

(b) Show that if $\Omega = \mathbb{C}$, then there is no such harmonic function. Hint: let $\mathscr{F}$ be the largest set of subharmonic functions in $\Omega \setminus \{p_0\}$ that has the properties: (i) if $u, v \in \mathscr{F}$ then $u \vee v \in \mathscr{F}$; (ii) each harmonic regularization of an element of $\mathscr{F}$ belongs to $\mathscr{F}$; (iii) if $u \in \mathscr{F}$ then

$$u(z) + \log|z - p_0| \quad \text{is bounded in a neighborhood of } p_0.$$

(Here we take the branch of $\log|z - p_0|$ that is negative near p_0.)

(c) Prove (a) without the restriction on using conformal mapping.

Remarks and further reading

The Laplacian $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ in $\mathbb{R}^2$ has an obvious analogue in $\mathbb{R}^n$, and indeed in any Riemannian manifold, and solutions of $\Delta = 0$ are the associated harmonic functions. For more on harmonic functions in $\mathbb{R}^n$, see Axler, Bourdon, and Ramey [13]. For harmonic, subharmonic, and plurisubharmonic functions in one and several complex variables, see Hayman and Kennedy [99], Hayman [98], and texts on several complex variables, such as Krantz [125], Ohsawa [156], and Hörmander [109].

Maximum principles for solutions of general partial differential equations are treated by Protter and Weinberger [172] and Pucci and Serrin [173].

Chapter 6
General Riemann surfaces

In this chapter, we introduce the idea of an abstract Riemann surface S, construct a simply connected covering surface for S, and derive some consequences.

An example is the Riemann sphere $\mathbb{S} = \mathbb{C} \cup \{\infty\}$, the one-point compactification of $\mathbb{C}$. The complex structure at ∞ is transferred from that of $\mathbb{C}$ by the inversion $z \to 1/z$.

Another example of a Riemann surface is any domain U in $\mathbb{S}$: a two (real)-dimensional manifold with a conformal structure. Recall that, by definition, a domain U is connected. If $U \subset \mathbb{C}$ is also simply connected and omits at least two points, the Riemann mapping theorem says that U can be mapped conformally to the unit disk. If U omits only one point, then a linear fractional transformation with a pole at that point maps U to $\mathbb{C}$.

A third type of example is provided by the equation

$$z^2 + w^2 = 1. \tag{6.0.1}$$

This equation defines a *complex curve* $C \in \mathbb{C}^2$:

$$C \;\equiv\; \{(z, w) \in \mathbb{C}^2 \;:\; z^2 + w^2 = 1\}.$$

Note that as $z \to \infty$, the two choices for w are asymptotic to $\pm iz$. This suggests considering an appropriate extension of C to a subset $\widehat{C} \subset \mathbb{S} \times \mathbb{S}$. It can be shown that $\widehat{C}$ can be considered as a Riemann surface; it is equivalent to $\mathbb{S}$.

Remark. The terminology here unfortunately conflicts with the common usage of "curve" to refer to a map from a real interval into $\mathbb{C}$, or into a Riemann surface, or to the image of such a map. We rely on the context to make clear which use is intended.

As conceived originally by Riemann and Weierstrass, a Riemann surface was the appropriate domain of definition of a possibly multiple-valued function f, extended as far as possible by analytic continuation. This is connected to the third example, with $f(z) = \sqrt{z^2 - 1}$. The general intrinsic concept, due to Weyl [214], is treated in Section 6.1.

R. Beals and R. S. C. Wong, *More Explorations in Complex Functions*, Graduate Texts in Mathematics 298, https://doi.org/10.1007/978-3-031-28288-1_6

The remainder of the chapter leads part of the way to a general classification of Riemann surfaces as quotients of $\mathbb{S}$, $\mathbb{C}$, or $\mathbb{D}$ by certain equivalence relations. The first two steps are taken in Section 6.2. The first step is the construction of the universal cover S^u of a Riemann surface S. The second step is the identification of a group G of automorphisms of S^u that correspond to equivalence classes of curves in S that have a fixed endpoint. Then S itself can be identified with the quotient $\Gamma\backslash S^u$ of S^u by the equivalence relation induced by G.

In Section 6.3, we assume the *uniformization theorem*. This theorem, which is proved in Chapter 7, says that any simply connected Riemann surface is equivalent to one of $\mathbb{D}$, $\mathbb{C}$, or $\mathbb{S}$. The relevance to the preceding discussion is that any universal cover S^u is simply connected. Therefore the group G can be taken to be a (certain type of) group of linear fractional transformations. The possibilities when S^u is $\mathbb{C}$ or $\mathbb{S}$ are determined explicitly, and the much more variegated case $S^u \equiv \mathbb{D}$ is discussed.

6.1 Abstract Riemann surfaces

An (abstract) Riemann surface is a set S provided with a *conformal structure* by a collection of non-empty subsets U_α such that $S = \bigcup U_\alpha$ and:

(a) for each index α there is a bijection ϕ_α mapping U_α onto $\mathbb{D}$;

(b) if $U_\alpha \cap U_\beta$ is not empty, then $D_{\alpha\beta} \equiv \phi_a(U_\alpha \cap U_\beta)$ is an open subset of $\mathbb{D}$ and $\phi_\beta \circ \phi_\alpha^{-1} : D_{\alpha\beta} \to \mathbb{D}$ is holomorphic;

(c) if p and q are two points of S, then there is a sequence U_{α_j}, $1 \leq j \leq n$ such that $p \in U_{\alpha_1}$, $q \in U_{\alpha_n}$, and $U_{\alpha_j} \cap U_{\alpha_{j+1}}$ is not empty, $1 \leq j < n$.

Assumptions (a) and (b) provide S with a topology: the open sets of S are the subsets U with the property that each image $\phi_\alpha(U \cap U_\alpha)$ is open in $\mathbb{D}$. Assumption (c) implies that S is *pathwise connected*: given points z, w of S, there is a continuous curve that joins z to w. Note that a *domain*, i.e. a connected open subset U, of a Riemann surface is a Riemann surface, using the intersections $U \cap U_\alpha$ that are not empty to provide the complex structure.

Suppose that $U \subset S$ is a domain. A function $f : U \to \mathbb{C}$ is, by definition, holomorphic (resp. meromorphic) if each function $f \circ \phi_\alpha^{-1}$ is holomorphic (resp. meromorphic), where defined, on $\mathbb{D}$. In particular, each coordinate mapping ϕ_α is holomorphic on U_α. More generally, a map from U to a Riemann surface S' is defined to be holomorphic (resp. meromorphic) if $g \circ f$ is holomorphic (resp. meromorphic) on S' for each g that is holomorphic on $f(U) \subset S'$.

The sets U_α are referred to as *coordinate neighborhoods*, and the mappings ϕ_α as *coordinate charts* or simply as *coordinates*. More generally, any holomorphic map z that is defined on a simply connected open neighborhood U of $p \in S$ and maps U injectively into $\mathbb{C}$ is said to be a *coordinate at* p, and U is said to be a *coordinate neighborhood*.

As a first example, let us take the Riemann sphere $\mathbb{S} = \mathbb{C} \cup \{\infty\} = U_1 \cup U_2$, where

$$U_1 = \mathbb{C}, \quad U_2 = (\mathbb{C} \setminus \{0\}) \cup \{\infty\}.$$

with $\phi_1(z) = z$, $\phi_2(z) = 1/z$.

As a second example, consider the curve $\widehat{C}$ from the introduction to this chapter:

$$\widehat{C} = \{(z, w) \in \mathbb{C} \times C : z^2 + w^2 = 1\} \cup \{(\infty, \infty)\} \subset \mathbb{S} \times \mathbb{S}; \tag{6.1.1}$$

see Exercise 1

The definitions, results, and proofs in Section 1.8 carry over immediately to Riemann surfaces. In particular:

Theorem 6.1.1. *If S is simply connected, f is holomorphic in some coordinate neighborhood, and f can be continued along every curve in S, then f has a unique single-valued extension to S.*

Remark. With suitable modifications, we can define analytic continuation for meromorphic functions, and obtain an analogue of Theorem 6.1.1 for a meromorphic function defined on a coordinate neighborhood. Note that the Riemann sphere $\mathbb{S}$ carries no holomorphic functions, but carries many meromorphic functions: the rational functions.

Two Riemann surfaces S and S' are said to be *equivalent* if there is a holomorphic bijection ϕ from S onto S'. As shown in Chapter 1 for the case of domains in $\mathbb{C}$, in local coordinates injectivity implies that the derivative of ϕ is non-zero and the local inverse is holomorphic. Therefore the inverse map is also holomorphic, so this is indeed an equivalence relation.

The *uniformization theorem* says that any *simply connected* Riemann surface is equivalent to either the unit disk, the complex plane, or the Riemann sphere. Note that these three surfaces are themselves inequivalent: see Exercise 2. This theorem indicates the importance of the construction, for any S, of a simply connected covering surface. The proof of this theorem is the subject of Section 6.2.

6.2 The universal cover

We begin with some remarks about curves in a Riemann surface S. Again a *curve* in S is a continuous map γ from a real interval $I = [a, b]$ into S. The *endpoints* are the ordered pair $(\gamma(a), \gamma(b))$. Two curves γ_1 and γ_2 are taken to be the same if they differ only by parametrization: $\gamma_j : I_j \to S$ and $\gamma_2 = \gamma_1 \circ \phi$, where ϕ is an injective increasing map from I_2 onto I_1. Two curves γ_0 and γ_1 with the same domain I are said to be *homotopic* if there is a continuous mapping $\gamma : I \times [0, 1] \to S$ such that

$$\gamma(s, 0) = \gamma_0(s), \quad \gamma(s, 1) = \gamma_1(s), \qquad s \in I.$$

More generally, two curves are said to be homotopic if they are equivalent to a pair of curves that are homotopic in this sense. In other words, two curves are homotopic if one can be continuously deformed into the other. Canonical examples are curves in the punctured plane $S = \mathbb{C} \setminus \{0\}$:

$$\gamma_1(\theta) = e^{i\theta}, \quad \gamma_2(\theta) = 2e^{i\theta}, \quad \gamma_3(\theta) = 4 + e^{i\theta}, \qquad |\theta| \le \pi.$$

Then γ_1 and γ_2 are homotopic in S, but neither is homotopic in S to γ_3 – see Figure 6.1. We write $\gamma \sim \widetilde{\gamma}$ if the curves γ and $\widetilde{\gamma}$ are homotopic (in a specified domain), and we write $\gamma \nsim \widetilde{\gamma}$ otherwise.

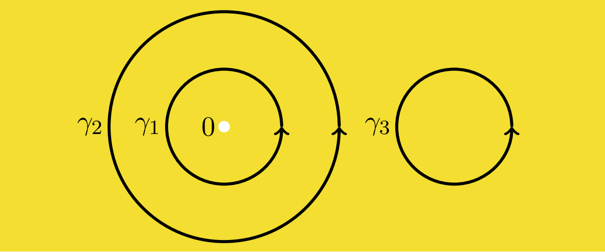

Fig. 6.1 In $\mathbb{C} \setminus \{0\}$, $\gamma_2 \sim \gamma_1$, but $\gamma_2 \nsim \gamma_3$.

In particular, any two constant curves, curves whose range is a single point, are equivalent; this is a consequence of pathwise connectedness.

The equivalence classes of curves in a Riemann surface S can be given a group structure. If γ_2 starts at the (second) endpoint of γ_1, then after a reparametrization, γ_1 followed by γ_2 is a curve $\gamma_1 \cdot \gamma_2$ from the first endpoint of γ_1 to the second endpoint of γ_2. If γ is a curve, let γ^{-1} denote the curve obtained by reversing the direction of travel. Then $\gamma \cdot \gamma^{-1}$ is homotopic to a constant curve; Exercise 4 It is easy to check that $(\gamma_1 \cdot \gamma_2) \cdot \gamma_3 \sim \gamma_1 \cdot (\gamma_2 \cdot \gamma_3)$ and that if $\gamma_j \sim \hat{\gamma}_j$, $j = 1, 2$, then $\gamma_1 \cdot \gamma_2 \sim \hat{\gamma}_1 \cdot \hat{\gamma}_2$. The equivalence classes form a (non-commutative) group, the *fundamental group* $H_1(S)$; Exercise 5.

Suppose that S and S' are Rieman surfaces. S' is said to be a *cover* of S if there is a mapping $\pi : S' \to S$ with the property that π is holomorphic, and for each $p \in S$, there is an open neighborhood U of p such that $\pi^{-1}(U)$ consists of disjoint open sets each of which is mapped bijectively by π onto U. For example, Figure 1.2 shows a portion of the domain of the logarithm as a cover of the punctured sphere $\mathbb{C} \setminus \{0\}$.

By a *lift* of a curve γ in S to a curve in a cover S', we mean a curve γ' such that $\pi \circ \gamma' = \gamma \circ \pi$, as in the commutative diagram

$$\begin{tikzcd}
S' \arrow[r, "\gamma'"] \arrow[d, "\pi"'] & S' \arrow[d, "\pi"'] \\
S \arrow[r, "\gamma"] & S.
\end{tikzcd}$$

Lemma 6.2.1. *Suppose that S' is a cover of S and suppose that γ is a curve in S that begins at p. For each $p' \in \pi^{-1}(p)$, there is a unique lift γ' in S' that starts at p'.*

Proof: Suppose that γ is parametrized by the interval $[0, 1]$. It is easily seen that for t close to zero there is a unique such lift of the restriction of γ to the interval $[0, t]$. It is also easily seen that the set of t such that γ has a unique lift from $[0, t]$ to S is both open and closed in $[0, 1]$. □

Lemma 6.2.2. *If curves γ_0 and γ_1 in S are homotopic, and γ_0', γ_1' are lifts to a cover S' that have the same starting point, then γ_0' and γ_1' are homotopic.*

Proof: Let γ_s, $0 \leq s \leq 1$ provide a continuous deformation from γ_0 to γ_1. Then the lifts γ_s' to S' provide a continuous deformation from γ_0' to γ_1'. □

A *universal cover* of S is a cover S^u that is simply connected. If S^u is a universal cover of S, then it is a cover for any cover S' of S. In particular, S^u is unique up to equivalence; see Exercise 6.

The notion of a universal cover, and the first constructions, go back to the work of Schwarz, Klein, Poincaré, and others on understanding and classifying the Riemann surfaces of algebraic functions. One line of attack is to build up from the original surface by cutting and pasting. A first step is illustrated schematically in Figure 6.2. Starting with a compact Riemann surface S of genus 2 (like the surface of a two-holed doughnut), S is cut along one curve that is not homotopic to a constant. Two copies of S are joined together along the two sides of the cut, to form a new surface $\widehat{S}$. Each point of S, such as p in the figure, corresponds naturally to two points, such as p' and p'' in $\widehat{S}$. This can be done in such a way as to preserve conformal structures, so that there is a two-to-one covering map π from $\widehat{S}$ onto S. Similar constructions can be carried out indefinitely, in such a way that the limiting manifold $\widetilde{S}$ is no longer compact, and such that curves like γ that are are not homotopic to constant curves in S become open curves in $\widetilde{S}$. Thus, $\widetilde{S}$ is simply connected.

The more modern approach to the construction of a universal cover is simpler and more conceptual.

Theorem 6.2.3. *A Riemann surface S has a universal cover.*

Proof: Fix a point p_0 in S. Consider the set $\widetilde{S}$ consisting of all pairs (p, γ), where p is a point of S and γ is a curve from p_0 to p. We define an equivalence relation $\sim$ in the set $\{(p, \gamma)\}$ by

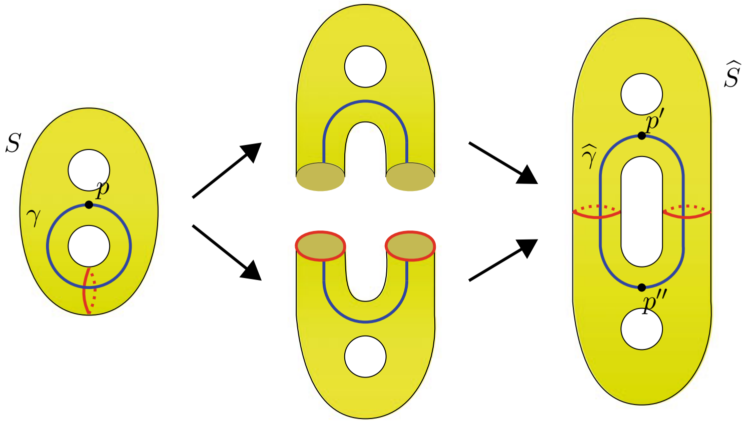

Fig. 6.2 The construction of a two-fold cover.

$$(p,\gamma_1) \sim (p,\gamma_2) \quad \text{if and only if } \gamma_1 \text{ is homotopic to } \gamma_2.$$

Let $[p,\gamma]$ denote the equivalence class of the pair (p,γ). Let S^u be the set of all equivalence classes $[p,\gamma]$. If p' lies in a coordinate neighborhood U of p, any curve from p_0 to p can be extended within U so as to reach p'. Extensions of two such curves will be homotopic if and only if the original curves are homotopic. Therefore we may sort the coordinate neighborhoods U of p into equivalence classes $[U,\gamma]$. Each equivalence class $[U,\gamma]$ can be considered as a coordinate neighborhood U of p, and any coordinate ϕ on U induces a coordinate on each $[U,\gamma]$. This gives us a covering of S^u and a corresponding set of mappings ϕ that satisfy the properties (a),(b),(c) stated at the beginning of Section 6.1; see Exercise 7. Therefore their union S^u is a Riemann surface. The map

$$\pi : [p,\gamma] \to p, \quad p \in S \tag{6.2.1}$$

is a covering map.

Finally, we need to show that S^u is simply connected. Let $\widetilde{p}_0$ be the equivalence class of (p_0,γ_0), where γ_0 is the constant curve at p_0. If γ is a curve from p_0 to p, denote by γ' the lift of γ to S^u that begins at $\widetilde{p}_0$. Then γ' ends at $[p,\gamma]$; Exercise 11. Suppose that γ_1' is a closed curve in S^u that begins and ends at $[p,\gamma]$. It is the lift to S^u, starting at $[p,\gamma]$ of the projection $\gamma_1 = \pi \circ \gamma_1'$ in S. Moreover, $\gamma_1' \cdot \gamma'$ is the lift to S^u of $\gamma_1 \cdot \gamma$. Since $\gamma_1' \cdot \gamma'$ and γ' have the same endpoint, it follows that $\gamma_1 \cdot \gamma$ and γ are homotopic. Therefore, $\gamma_1 \cdot \gamma \cdot \gamma^{-1}$ and $\gamma \cdot \gamma^{-1}$ are homotopic. The former of these last two curves is homotopic to γ_1 and the latter is homotopic to a constant. Therefore the lift γ_1' is homotopic to a constant, and we have shown that S^u is simply connected. □

Suppose that S^u is the universal cover of S as constructed. Choose a point p_0 of S. We shall associate to each closed curve γ starting at p_0 a map of S^u to itself. Given

$[p, \gamma_p]$ in S^u, let

$$A_\gamma([p, \gamma_p]) = [p, \gamma_p \cdot \gamma]. \tag{6.2.2}$$

Theorem 6.2.4. *(a)* $\pi \circ A = \pi$.
(b) A_γ *is an automorphism of* S^u.
(c) If γ *is the constant curve at* p_0, *then* A_γ *is the identity map of* S^u.
(d) $A_{\gamma_1} = A_{\gamma_2}$ *if and only if* γ_1 *and* γ_2 *are homotopic.*
(e) A_γ *has a fixed point if and only if* A_γ *is the identity map.*
(f) $A_{\gamma_1 \cdot \gamma_2} = A_{\gamma_2} A_{\gamma_1}$.

Proof: (a), (c), and (f) are immediate from the definitions.

(b) It follows readily from the definition that A_γ maps a a coordinate neighborhood of $[p, \gamma_p]$ holomorphically to the corresponding coordinate neighborhood of $[p, \gamma_p \cdot \gamma]$.

(d) This follows from the fact that γ_1 and γ_2 are homotopic if and only if $\gamma_1 \cdot \gamma$ and $\gamma_2 \cdot \gamma$ are homotopic.

(e) By (d) $A_\gamma([p, \gamma_p]) = [p, \gamma_p]$ if and only if γ_p and $\gamma_p \cdot \gamma$ are homotopic, which is the case if and only if γ is homotopic to a constant, which implies that A_γ is the identity map. Conversely, the identity map fixes every point of S^u. □

The automorphisms A_γ are called *cover transformations*, or *deck transformations* (from the German *decken*, to cover). A group G of automorphisms of a Riemann surface S is said to be *properly discontinuous* if, given two compact subsets C_1, C_2 of S, the intersection $g(C_1) \cap g(C_2)$, $g \in G$, is empty except for finitely many g.

Proposition 6.2.5. *The group G of cover transformations of* S^u *is properly discontinuous.*

Proof: Suppose that p_1 and p_2 are distinct points of S. Choose distinct connected neighborhoods U_1, U_2. Then $\widetilde{U}_j = \pi^{-1}(U_j)$ are disjoint open sets in S^u, each of which is invariant under G. The sets $\widetilde{U}_j$ themselves are unions of disjoint preimages $U_{j\alpha}$ of U_j, and any element of G permutes these preimages. Therefore the intersection under the action of an element of G on two such preimages is disjoint if they are distinct, and is disjoint for all but the identity element if they coincide. The extension to disjoint compact sets is immediate. □

As we shall see, $\mathrm{Aut}(S^u)$ has a natural topology. A consequence of Proposition 6.2.5 is that no sequence of non-identity cover transformations can have the identity transformation as limit:

Corollary 6.2.6. *The group of cover transformations of* S^u *is a discrete group.*

The importance of the group G is that it allows S to be recovered from its universal cover S^u. In fact, G induces an equivalence relation in S^u: two points are equivalent if some element of G takes one to the other. The quotient space, often written as $G \backslash S^u$, can be naturally identified with S: see Exercise 12.

6.3 Automorphism groups and cover transformations

We assume now the uniformization theorem of Chapter 7, so any simply connected Riemann surface can be taken to be one of $\mathbb{D} \cong \mathbb{H}$, $\mathbb{C}$, or $\mathbb{S}$. As noted in Chapter 2, in each case, the automorphism group is a subgroup of the group of linear fractional transformations

$$f(z) = \frac{az+b}{cz+d}, \qquad ad - bc \neq 0. \tag{6.3.1}$$

We may multiply numerator and denominator by $1/\sqrt{ad-bc}$ and reduce to

$$f(z) = \frac{az+b}{cz+d}, \quad ad - bc = 1. \tag{6.3.2}$$

This representation is still not unique: we can multiply both numerator and denominator by -1. In group theoretic terms, this means we are looking at $SL(2,\mathbb{C})/\{\pm\mathbf{1}\}$, the quotient of the group of 2×2 complex matrices divided by the subgroup consisting of $\pm\mathbf{1}$, where $\mathbf{1}$ is the identity matrix. This group is commonly written $PSL(2,\mathbb{C})$. It inherits a topology from $\mathbb{C}^4$.

Specifically, the results from Chapter 2 are the following.

Proposition 6.3.1. *(a) The automorphism group* $\mathrm{Aut}(\mathbb{C})$ *consists of the affine mappings $f(z) = az + b$, $a \neq 0$.*
(b) The automorphism group $\mathrm{Aut}(\mathbb{S})$ *consists of all linear fractional transformations (6.3.1).*
(c) The automorphism group $\mathrm{Aut}(\mathbb{H})$ *consists of the transformations (6.3.2) with real coefficients a, b, c, d and positive determinant.*

In view of Theorem 6.2.4 and the remarks which precede it, we would like to identify the candidates for cover transformations, as subgroups of the automorphism group of $\mathbb{D}$, $\mathbb{C}$ or $\mathbb{S}$ as the case may be.

Proposition 6.3.2. *If the universal cover S^u of S is equivalent to the Riemann sphere, then $S^u = S$.*

Proof: Any linear fractional transformation has at least one fixed point in $\mathbb{S}$. Therefore, by Theorem 6.2.4, there are no non-identity cover transformations and therefore no non-constant closed curves in S. Therefore the construction of S^u simply gives a bijection. □

Proposition 6.3.3. *Suppose that the universal cover of S is equivalent to $\mathbb{C}$. The group of cover transformations, carried over to $\mathbb{C}$, is generated by either one or two translations $z \to z + b$.*

Proof: Let G be the group of cover transformations. Every element of G is an affine transformation $f(z) = az + b$. Such a transformation has a fixed point in $\mathbb{C}$ unless $a = 1$. Thus, every non-identity cover transformation is a translation $T_b z = z + b$, $b \neq 0$. The group generated by a single translation is clearly discrete.

Suppose that a has minimum modulus among the values b such that T_b is in G. Let G_a denote the subgroup generated by T_a. Suppose that T_b is another element of G. If b lies on the line through 0 and a, then some translate $b' = [T_a]^n(b)$ lies in the interval from 0 to a. By the minimality assumption $b' = 0$ or $b' = a$, so T_b is in G_a.

If $G_a \neq G$, let b have minimum modulus among the c such that T_c is in $G \setminus G_a$, and let $G_{a,b}$ be the group generated by T_a and T_b. The image of 0 under $G_{a,b}$ is the lattice

$$\Lambda \;=\; \{ma + nb : m, n \in \mathbb{Z}\}. \tag{6.3.3}$$

Translations of the parallelepiped

$$\Pi \;=\; \{ra + sb : 0 \leq r, s \;<\; 1\}$$

form a partition of $\mathbb{C}$. Thus, given any point p of $\mathbb{C}$, some combination of T_a and T_b will move that point into Π, and further translation by $(T_a T_b)^{-1}$ will move p into the reflection $-\Pi = \{-z : z \in \Pi\}$. In particular, any point c' in the triangle with vertices a, b, $a + b$ maps to a point c'' in the triangle with vertices $-b$, 0, $-a$; see Figure 6.3.

Suppose now that T_c belongs to G. The preceding observations show that some element of $G_{a,b}$ will move c into a point c' that lies either in the triangle with vertices 0, a, b, or in the triangle with vertices 0, $-a$, $-b$: Figure 6.3. The minimality assumptions on a and b imply that c' or c'' must be one of the vertices 0, $\pm a$, $\pm b$, so T_c belongs to $G_{a,b}$. □

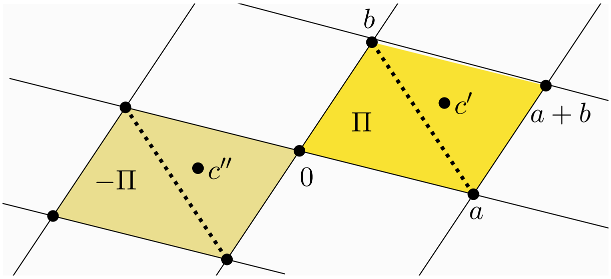

Fig. 6.3 Translation by $-a - b$.

Suppose, finally, that the covering manifold for S is the upper half-plane $\mathbb{H}$. If T is a linear fractional transformation with real coefficients, then the fixed points come in complex conjugate pairs. Therefore the only candidates for non-identity cover transformations are those whose fixed points are in $\mathbb{R} \cup \{\infty\}$. This is true of the fixed points if and only if $c = 0$ or $(a + d)^2 \geq 4$; see Exercise 15. Beyond this, it is not easy to give a simple description of a covering group. Proposition 6.2.5 gives a necessary condition. A subgroup of $\mathrm{Aut}(\mathbb{H})$ is said to be *Fuchsian* if it is properly discontinuous. Thus, any group of cover transformations is a Fuchsian group, with the additional constraint that non-identity elements have no fixed points.

Conversely, it can be shown that any such group G is the group of covering transformations of a Riemann surface $S = G\backslash\mathbb{H}$. The points of S are equivalence classes of points of $\mathbb{H}$, where two points $z_j \in \mathbb{H}$ are equivalent if and only if $z_2 = g(z_1)$ for some $g \in G$, i.e. they belong to the same orbit of G. The proof is left as Exercise 12.

Some standard examples of Fuchsian groups are $G_0 = SL(2, \mathbb{Z})$, the group of linear fractional transformations (2.1.1) having integer coefficients, and G_p, the subgroup of G_0 consisting of those transformations equal to the identity $\mathbf{1}$ modulo p, p a prime, i.e.

$$b, c = 0 \bmod p \quad \text{and} \quad a = d = 1 \bmod p \quad \text{or} \quad a = d = -1 \bmod p.$$

There are non-identity elements of G_0 that have fixed points in $\mathbb{H}$, but this is not the case for the non-identity elements of G_p; see Exercise 17.

Exercises

1. Show that the curve (6.1.1) has a natural structure as a Riemann surface. Hint: for each finite point (z_0, w_0), show that either $\phi(z, w) = \varepsilon(z - z_0)$ or $\phi(z, w) = \varepsilon(w - w_0)$ maps a neighborhood onto $\mathbb{D}$; this leaves (∞, ∞) to be considered.
2. Show that $\mathbb{D}$, $\mathbb{C}$, and $\mathbb{S}$ are inequivalent Riemann surfaces.
3. If γ is a curve, show that $\gamma \cdot \gamma^{-1}$ is homotopic to a constant curve.
4. If $\gamma_1, \gamma_2, \gamma_3$ are curves, show that $\gamma_1 \cdot (\gamma_2 \cdot \gamma_3) \sim (\gamma_1 \cdot \gamma_2) \cdot \gamma_3$.
5. Fill in the details to prove that the equivalence classes of curves in a Riemann surface S form a group.
6. Show that If S^u is a universal cover of S, then it is a cover for any cover S' of S. Show that S^u is unique up to equivalence.
7. Show that the set of equivalence classes $\{[U, \gamma]\}$ introduced in the proof of Theorem 6.2.3 has the properties (a), (b), (c) at the beginning of Section 6.1.
8. Show that the universal cover of the punctured plane $\mathbb{C} \setminus \{0\}$ can be taken to be the upper half-plane $\mathbb{H}$. Hint: start with the representation
$$\{z : z = re^{i\theta}, \quad r > 0, \quad \theta \in \mathbb{R}\}.$$
9. The universal cover constructed in Section 6.2 involves the choice of a point p_0. Show that the choice of any other point as starting point gives rise to the same equivalence classes and the same conformal structure.

10. (a) Suppose that S is a Riemann surface and $f : S \to \mathbb{C}$ is a meromorphic function. Show that f lifts to a meromorphic function $\widehat{f}$ on the universal cover S^u, i.e. $f \circ \pi = \pi \circ \widehat{f}$.

    (b) Conversely, suppose $g : S^u \to \mathbb{C}$ is meromorphic. Under what condition is g the lift $\widehat{f}$ of a meromorphic function on S?
11. Prove that if γ is a curve from p_0 to p, and γ' is the lift of γ to S^u that begins at $\widetilde{p}_0$, then γ' ends at $[p, \gamma]$.
12. Suppose that G is the group of cover transformations of a Riemann surface S. Show that the points of S are equivalence classes of points of S^u, where such points are equivalent if and only if they belong to the same orbit of G.
13. Prove that for any fixed-point-free Fuchsian group G, the space $G\backslash\mathbb{H}$ has a complex structure such that the projection π taking z to its equivalence class $[z]$, is locally conformal.
14. Using Proposition 6.3.3, discuss the determination of the equivalence classes of Riemann surfaces with universal cover $\mathbb{C}$, where "equivalent" means being related by a holomorphic bijection.
15. Verify that $T \in \mathrm{Aut}(\mathbb{H})$ with real coefficients has no fixed point in $\mathbb{H}$ if and only if $c = 0$ or $(a - d)^2 \geq 2$.
16. Verify that G_p, the subgroup of $\mathrm{Aut}(\mathbb{H})$ consisting of transformations with integer coefficients that are equal to $\mathbf{1}$ modulo the prime p, is a group.
17. Show that G_0, the subgroup of $\mathrm{Aut}(\mathbb{H})$ consisting of transformations with integer coefficients, has non-identity elements with fixed point in $\mathbb{H}$, but G_p does not.
18. Show that the group of conformal self-maps of a compact Riemann surface is finite.

Remarks and further reading

The definitive formulation of the general concept of a Riemann surface, and the basic theory of such surfaces, go back to Weyl [214]. There are many modern treatments, e.g. Donaldson [56], Schlag [185]. Farkas and Kra [79] is particularly comprehensive. Siegel [191], [192] has an efficient treatment of covering spaces and the basics of automorphic function theory.

Chapter 7
The uniformization theorem

This chapter is devoted to the proof of the uniformization theorem, and a discussion of its consequences. The theorem says that a simply connected Riemann surface is biholomorphically equivalent either to the unit disk $\mathbb{D}$ (or, equivalently, the upper half-plane $\mathbb{H}$), the complex plane $\mathbb{C}$, or the Riemann sphere $\mathbb{S}$. As shown in Chapter 6, every Riemann surface has a simply connected cover and is invariant under certain automorphisms of the cover. Therefore the uniformization theorem opens the way to a trove of information about general Riemann surfaces.

The first theorem of this type is the theorem known as the *Riemann mapping theorem*: a simply connected domain U in $\mathbb{C}$ whose boundary consists of more than one point is biholomorphically equivalent to the unit disk $\mathbb{D}$. Riemann's argument assumed that U was a Jordan domain – a domain bounded by a simple closed curve $\Gamma = \partial D$. On physical grounds, given a point $p_0 \in U$ there should be a point potential $g = g(p, p_0)$: a function harmonic in $U \setminus \{p_0\}$ that vanishes on Γ and has a singularity like $\log r$ near p_0, where $r(p) = |p - p_0|$. Then g is the real part of a function f that is holomorphic on U, and the function $F = \exp(-f)$ would map U biholomorphically onto $\mathbb{D}$. Riemann's argument was not a proof; see Section 5.5. However the idea can be made a proof by making more direct use of the solvability of the Dirichlet problem: see Exercises 13 – 15.

The proof of the Riemann mapping theorem that is usually presented now looks for F directly as the solution of a certain extremal problem. However one of the standard approaches to the general problem goes directly back to constructing a harmonic function u with a singularity – either like $\log r$ or like $\mathrm{Re}\,(1/z)$. The particular version we follow in this chapter is known as the *Perron method*. In the case of a singularity like $\log r$, finding a harmonic conjugate v to u and exponentiating $u + iv$ leads to a conformal map onto $\mathbb{D}$. In the case of a singularity like $1/z$, $u + iv$ itself leads to a conformal map onto $\mathbb{C}$ or $\mathbb{S}$.

Sections 7.1 and 7.2 treat the hyperbolic case: $S^u \cong \mathbb{H}$, singularity like $\log r$. Sections 7.3 and 7.4 treat the remaining cases: parabolic ($S^u \cong \mathbb{C}$) and elliptic ($S^u \cong \mathbb{S}$).

R. Beals and R. S. C. Wong, *More Explorations in Complex Functions*, Graduate Texts
in Mathematics 298, https://doi.org/10.1007/978-3-031-28288-1_7

7.1 Green's functions and harmonic measure

In this section we deal with a simply connected open Riemann surface S. It is convenient to introduce some terminology. If p_0 is a point of S, a coordinate map z defined in a neighborhood U of p_0 is said to be a *standard coordinate at* p_0 if $z(p_0) = 0$ and U contains the set $\{p : |z(p)| \leq 1\}$ as a compact subset.. With z understood, we let $D_r = D_r(p_0)$, $0 < r \leq 1$, be the disk $\{p : |z(p)| < r\}$. By a *coordinate disk* in S we mean any such $D_r(p_0)$ $0 < r < 1$, always with the assumption that the closure is compact in S.

A real-valued function u defined in an open subset $U \subset S$ is said to be *harmonic* if each point $p \in S$ has a neighborhood in which u is the real part of a holomorphic function. This is equivalent to saying that for any coordinate disk $D \subset U$, there is a real-valued function v such that $u + iv$ is holomorphic in D. Any such function v is called a *harmonic conjugate* of u.

As in the case $S = \mathbb{C}$ of Chapter 5, a real-valued function v defined in an open set $U \subset S$ is said to be *subharmonic* if for each coordinate disk D whose closure $\overline{D}$ is in U, if u is harmonic in D and continuous on the closure, and $v \leq u$ on ∂D, then $v \leq u$ in D.

We are now in a position to carry over from Section 5.3 the concept of a *Perron family*: a non-empty family $\mathscr{F}$ of subharmonic functions such that

(a) if u and v belong to $\mathscr{F}$ then so does the maximum $u \vee v$;

(b) if u belongs to $\mathscr{F}$, so does each harmonic regularization of u.

The first example that is of interest here is the following. Given $p_0 \in S$, let the family $\mathscr{F}(p_0)$ consist of all non-negative functions u, subharmonic in the punctured surface $S' = S \setminus \{p_0\}$, such that

(a) $u(p) = 0$ if p is in the complement of some compact set K (that depends on u);

(b) $u(p) + \log|z(p)|$ is bounded in a coordinate neighborhood centered at p_0.

It is easily seen that it if $\mathscr{F}(p_0)$ is non-empty, then it is a Perron family. But we have

Lemma 7.1.1. *For each $p_0 \in S$, the family $\mathscr{F}(p_0)$ is non-empty.*

Proof. Let z be a coordinate centered at p_0 such that the closure of $D_1(p_0)$ is compact in S. Define

$$u(p) = \begin{cases} -\log|z(p)|, & \text{if } |z(p)| \leq 1; \\ 0, & \text{otherwise.} \end{cases} .$$

Then u belongs to $\mathscr{F}(p_0)$. □

Remark. The preceding concepts are conformally invariant: see Exercises 1 – 4.

If the supremum $\widehat{u}$ of the Perron family $\mathscr{F}(p_0)$ is *finite*, then $\widehat{u}$ is called a *Green's function* for S with pole at p_0, and is denoted $g(p, p_0)$.

Theorem 7.1.2. *A Green's function $g(p, p_0)$ on S has the properties*

(a) $g(\cdot, p_0)$ *is harmonic on* $S \setminus \{p_0\}$*;*

(b) $g(p, p_0) > 0$.

(c) $g(p, p_0) + \log|z(p)|$ *has a harmonic extension to a neighborhood of* p_0*, where* z *is a standard coordinate at* p_0*;*

(d) $\inf_p\{g(p, p_0)\} = 0$.

Proof. Property (a) is Perron's principle, and property (b) follows from the fact that $u \equiv 0$ belongs to $\mathscr{F}(p_0)$, together with the strict maximum principle for $-g(\cdot, p_0)$.

(c) Let h denote the maximum value of $g(p, p_0)$ for $|z(p)| = 1$. Given p with $z(p) < 1$, choose a sequence $\{u_n\}$ in $\mathscr{F}(p_0)$ such that u_n is nondecreasing and $u_n(p) \to g(p, p_0)$. Let v_n be the harmonic function equal to $u_n + \log|z|$ at $|z| = 1$. By Lemma 5.2.3, $u_n + \log|z| \le v_n$ for $0 < |z| < 1$. By Theorem 5.2.2, the v_n converge to a function v, harmonic for $|z| < 1$ and $\le g$ for $|z| = 1$. Therefore $v \le \bar{v}$, where $\bar{v}$ is the harmonic function equal to g for $|z| = 1$. Moreover $v(p) = g(p, p_0) + \log|z(p)|$. Since p was arbitrary, $g(\cdot, p_0) + \log|z| \le \bar{v}$. But also $g(\cdot, p_0) \ge v$, so p_0 is a removable singularity.

(d) Let $c = \inf_p g(p, p_0)$ and suppose $u \in \mathscr{F}(p_0)$. By assumption, $u \le 0 \le g - c$ outside some compact subset. Moreover $u + \log|z|$ is bounded in a neighborhood of p_0. We know that $g + \log|z|$ is also bounded near p_0. Therefore, for each $\varepsilon > 0$, $(1 - \varepsilon)u \le g$ in a small enough neighborhood of p_0. Since g is harmonic and $(1 - \varepsilon)u$ is subharmonic, this implies that $(1 - \varepsilon)u \le u - c$ everywhere. Therefore $(1 - \varepsilon)g \le g - c$ everywhere, so $c = 0$. □

If K is a non-empty subset of S having compact closure, let $\mathscr{H}(K)$ denote the family of subharmonic functions u defined on $S \setminus K$ such that $0 \le u \le 1$, and u vanishes outside some compact set. Let $u_K = \sup\{u;\, u \in \mathscr{H}(K)\}$. Then $0 \le u_K \le 1$ on the complement of K. If u_K is not identically 0 or 1, then K is said to have *harmonic measure* u_K. Since u_K cannot attain its supremum or infimum, it follows that in this case $0 < u_K < 1$.

Lemma 7.1.3. *If* $\emptyset \ne K_1 \subset K_2$ *and the closure of* K_2 *in* S *is compact, then*

$$u_{K_1}\big|_{S\setminus K_2} \le u_{K_2}. \tag{7.1.1}$$

Proof. The restriction to $S \setminus K_2$ of an element of $\mathscr{H}(K_1)$ belongs to $\mathscr{H}(K_2)$, which implies (7.1.1). □

Lemma 7.1.4. *If* D_1 *is a standard coordinate neighborhood of* p_0 *and* $0 < r < 1$*, then* u_{D_r} *is not identically zero, and has limit 1 as* $p \to \partial D_r$*.*

Proof. Choose $r < s < 1$ and define

$$u(p) = \begin{cases} \dfrac{\log(s/|z(p)|}{\log(s/r)}, & p \in D_r \cap D_s; \\ 0, & p \notin D_s(p_0). \end{cases}$$

Then u belongs to $\mathscr{H}(D)$ and $u > 0$ on $D_s \setminus D_r$. Moreover, $u(p) \to 1$ as $p \to \partial D$. But $1 \geq u_{D_r} \geq u$. □

The following is an easy consequence; see Exercise 5.

Corollary 7.1.5. *If K has harmonic measure and $K' \subset K$ has non-empty interior, then K' has harmonic measure.*

Proposition 7.1.6. *Suppose that some coordinate disk D in S has harmonic measure. Then S has a Green's function with pole in D.*

Proof. Let D be $D_1(p_0)$ with respect to a local coordinate z. Choose $0 < r < s < 1$ and let $D_r = D_r(p_0)$, $D_s = D_s(p_0)$. By Corollary 7.1.5, D_r has harmonic measure u_r.

Suppose that v belongs to the Perron family $\mathscr{F}(p_0)$. Let c be the maximum value for v on ∂D_r. Then $v \leq c = cu_r$ on ∂D_r and $v \equiv 0$ at ∞, so

$$v \leq cu_r \quad \text{on the complement of } D_r. \tag{7.1.2}$$

Now choose $\varepsilon > 0$ and consider

$$w \ = \ v + (1+\varepsilon)\log|z| \quad \text{for} \quad |z| \leq s.$$

By assumption, $v + \log|z|$ is bounded as $z \to 0$. Since $\varepsilon \log|z| \to -\infty$ as $z \to 0$, it follows that the maximum of w is attained on ∂D_s. Therefore, on ∂D_s,

$$c + (1+\varepsilon)\log r \ \leq \ cu_r + (1+\varepsilon)\log s,$$

so

$$c \ \leq \frac{1+\varepsilon}{1-u_r}\log\frac{s}{r} \quad \text{on } \partial D_s.$$

Taking $\varepsilon \to 0$ we get

$$\max_{|z|\leq s}[v + \log|z|] \ \leq \ \left(1 - \max_{|z|=s} u_r\right)^{-1} \log\frac{s}{r}$$

on D_s. It follows that $\sup v$, $v \in \{\mathscr{F}(p_0)\}$ is finite for $0 < |z| < s$, hence finite everywhere. □

We shall want a partial converse.

Proposition 7.1.7. *Suppose that S has a Green's function g with pole at p_0 and suppose that D is a disk contained in the set*

$$\{z \ : \ 0 < \varepsilon < |z(p_0)| < r < 1\},$$

where z is a standard coordinate at p_0. Then D has harmonic measure.

Proof. Each function v in $\mathscr{H}$ is bounded near ∂D, and on the complement of some compact set, by the harmonic function $g/m + \varepsilon$, where m is the minimum of g on ∂D and $\varepsilon > 0$ is arbitrary. Therefore the supremum u is bounded by $g/m + \varepsilon$. The infimum ε of g/m is not attained, so there is some point $p \in S \setminus D$ where g/m is $< \varepsilon$. Therefore each v is $< 2\varepsilon$ at p. It follows that the supremum u is $\leq 2\varepsilon$ at p and therefore not identically 1. □

Theorem 7.1.8. *Suppose that S has a Green's function at a point p_0. Then S has a Green's function at every point of S.*

Proof. It follows from Proposition 7.1.7 and Proposition 7.1.6 that every point in a standard coordinate neighborhood of p_0 is the pole of a Green's function. Therefore the set of points that can serve as poles of Green's functions is both open and closed, hence is all of the connected set S. □

Remark. The reason for the terminology "Green's function" here is that if L is a linear differential operator, a Green's function for L is a function G such that

$$u(\mathbf{x}) \;=\; \iint_{\Omega} G(\mathbf{x}, \mathbf{y})\varphi(\mathbf{y})\, d\mathbf{y}$$

is a solution of $Lu = \varphi$ (subject to some conditions at the boundary of the domain of definition Ω). In particular, for $L = \Delta$, the Laplacian in $\mathbb{R}^2$, the Green's function is

$$G(\mathbf{x}, \mathbf{y}) = \frac{-\log|\mathbf{x} - \mathbf{y}|}{2\pi}, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^2.$$

7.2 Uniformization: the hyperbolic case

A Riemann surface that carries a Green's function is said to be *hyperbolic*.

Theorem 7.2.1. *(***Uniformization, part I***). If S is a simply connected hyperbolic Riemann surface, then there is a conformal map of S onto the disk $\mathbb{D}$.*

Proof. Let $g = g(\cdot, p_0)$ be the Green's function with pole at p_0, and let z be a standard coordinate at p_0. By Theorem 7.1.2 (c), $g + \log|z|$ is harmonic near $z = 0$. Let h_{p_0} be a harmonic conjugate defined in D_1, with $h_{p_0}(0) = 0$. Define f_{p_0} for $|z| < 1$ by

$$f_{p_0} \;=\; z\exp(-g - \log|z| - ih_{p_0}).$$

This function is bounded near p_0 and the exponential factor has a non-zero limit at p_0.

Given $p \neq p_0$, choose a standard coordinate at p, with $p_0 \notin D_1(p)$. Let h be a harmonic conjugate for g in D_1. Note that h is unique up to an additive constant, so $\exp(g + ih)$ is unique up to a multiplicative constant of modulus 1. If two such neighborhoods overlap, the second constant (say) can be adjusted so that $\exp(g + ih)$ is holomorphic on the union of the two neighborhoods. Similarly, if $D_1(p)$ and $D_1(p_0)$ overlap, then the modulus of the quotient f_{p_0}/f_p is

$$\left|\frac{z\exp(-g - \log|z| - ih_{p_0})}{\exp(-g - ih_p)}\right| = 1,$$

so the quotient is a constant of modulus 1.

It follows that f_{p_0} can be continued along each curve in S. Since S is assumed to be simply connected, the continuation is a single-valued holomorphic map $f(p, p_0)$ of S into $\mathbb{D}$. The next step is to show that $f(p, p_0)$ is injective. Suppose that $p_1 \neq p_0$ and $p \neq p_0$. Set

$$F(p) = T(f(p, p_0)) = \frac{f(p, p_0) - f(p_1, p_0)}{1 - \overline{f(p_1, p_0)}}.$$

The linear fractional transformation T maps $\mathbb{D}$ to itself, so F is holomorphic on S. Let z be a standard coordinate at p_1. Since $F(p_1) = 0$, it follows that $\log F(p) \sim \log|z|$ near p_1. Suppose now that v belongs to the Perron family $\mathscr{F}(p_1)$ that defines $g(p, p_1)$. Take $\varepsilon > 0$ and consider the subharmonic function

$$v + (1 + \varepsilon)\log|F(p)|. \tag{7.2.1}$$

Near p_1 this is similar to $\varepsilon \log|z|$, so it has limit $-\infty$. But by assumption v vanishes near ∞, so by the maximum principle

$$v + (1 + \varepsilon)\log|F(p)| \leq 0.$$

Taking the supremum over $v \in \mathscr{F}(p_1)$ and letting $\varepsilon \to 0$, we have

$$g(p, p_1) + \log|F(p)| \leq 0. \tag{7.2.2}$$

Exponentiating,

$$|F(p)| \leq |g(p, p_1)| = |f(p, p_1)|.$$

But $F(p_0) = -f(p_1, p_0)$, so

$$|f(p_1, p_0)| \leq |f(p_0, p_1)|.$$

Since p_0 and p_1 are interchangeable in this argument,

$$|f(p_0, p_1)| = |f(p_1, p_0)|.$$

Then (7.2.2) gives

$$g(p_0, p_1) + \log|F(p_0)| = 0.$$

The left-hand side is a harmonic function of p_0, so by the strict maximum principle it is constant, hence identically zero:

$$g(p, p_1) + \log|F(p)| \ = \ 0. \tag{7.2.3}$$

Now $F(p) = 0$ when $f(p, p_0) = f(p_1, p_0)$, and (7.2.3) shows that this implies that $p = p_1$. Thus $f(p, p_0)$ is single-valued.

We have now shown that $f(p) = f(p, p_0)$ is a conformal map from S to a (necessarily simply connected) subset of $\mathbb{D}$. By the Riemann mapping theorem, there is a conformal map from the image $f(S)$ onto $\mathbb{D}$. □

It follows from Theorem 7.2.1 that a simply connected open Riemann surface that is hyperbolic, i.e. carries a Green's function, also carries a non-constant bounded harmonic function. In the next section we will have use for the converse.

Proposition 7.2.2. *If S carries a non-constant real-valued bounded harmonic function, then S has a Green's function.*

Proof. Take u to be a harmonic function with $\sup u = 1$, $\inf u = 0$, and use the proof of Proposition 7.1.7 with u in place of g/m. □

7.3 An analogue of the Green's function

We assume throughout this section that S is a simply connected Riemann surface that is *not* hyperbolic, i.e. S has no Green's function. Such a surface is said to be *parabolic* if it is open, i.e. not compact. It is said to be *elliptic* if it is compact. As in the case of conic sections, these can be thought of as limiting cases of the hyperbolic case; see Exercise 12.

Proposition 7.3.1. *If S is parabolic then:*
(a) no coordinate disk in S has harmonic measure;
(b) S carries no non-constant bounded harmonic functions;
(c) for any non-empty compact subset K, the maximum principle holds in $S \setminus K$, in the sense that if u is bounded above and harmonic in $S \setminus K$, then

$$\sup u(z) \ = \ \limsup_{p \to \partial K} u(p).$$

Proof. Parts (a) and (b) follow from Proposition 7.1.6 and Proposition 7.2.2. For (c), as before we let $\mathscr{H}(K)$ be the Perron family consisting of subharmonic functions v on $S \setminus K$, such that $0 \le v \le 1$, v is not identically 0, and v vanishes outside some compact set. Suppose that u is harmonic in $S \setminus K$, $0 \le u \le 1$, and $\limsup_{p \to \partial K} u(p) = 0$. Then

$$\limsup_{p \to \partial K} [u(p) + v(p)] \ \le \ 1, \qquad \limsup_{p \to \infty} [u(p) + v(p)] \ \le \ 1.$$

Therefore $u + v \leq 1$. If K does not have harmonic measure, then v can be chosen to be arbitrarily close to 1. Therefore $u \leq 0$, and we have proved the maximum principle for $S \setminus K$. □

In place of a Green's function – a harmonic function with a pole like $\log(1/r)$ in some coordinate neighborhood of a point p_0 – we look for a harmonic function on S with a pole like $\operatorname{Re}(1/z)$.

We could take a corresponding Perron family to be the family of functions v that are subharmonic on $S \setminus \{p_0\}$, vanish outside some compact set, and such that $v - \operatorname{Re}(1/z)$ is bounded, where z is a standard coordinate at p_0. However it is far from obvious that there are any such functions. Instead, the basic idea of the proof is to construct a harmonic function with a singularity at a point p_0 by working with a family of functions defined outside successively smaller coordinate disks centered at p_0.

Lemma 7.3.2. *Let z be a standard coordinate at $p_0 \in S$. Given $0 < \rho < 1$, there is a unique bounded harmonic function u_ρ on $S \setminus D_\rho$ that is equal to* $\operatorname{Re}(1/z)$ *on ∂D_ρ.*

Proof. In the compact case, u_ρ is simply the solution of the corresponding Dirichlet problem. If S is not compact, we let $\mathcal{G}$ be the family of subharmonic functions on $S \setminus D_\rho$ that are continuous, bounded above by $\operatorname{Re}(1/z)$ at the boundary, and vanish outside some compact set. The supremum u_ρ is harmonic and bounded above by $\operatorname{Re}(1/z)$. On the other hand, $\mathcal{G}$ contains the function v obtained by solving the Dirichlet problem with value $\operatorname{Re}(1/z)$ on ∂D_ρ and 0 on ∂D_1, extended to be zero outside ∂D_1, so u_ρ is continuous and has the correct boundary value on ∂D_ρ. Boundedness and uniqueness follow from the maximum principle for $S \setminus \overline{D_\rho}$, applied to u and to $-u$. □

The next sequence of lemmas aims to estimate the behavior of u_ρ as $\rho \to 0$, starting with the *oscillation* on the circle $|z| = r$, $\rho \leq r \leq 1$:

$$M_r(u_\rho) \;=\; \max_{|z|=r} u_\rho - \min_{|z|=r} u_\rho$$

Lemma 7.3.3. *Let v be the solution to the Dirichlet problem on $\mathbb{D}$ with boundary values*

$$v(re^{i\theta}) \;=\; \begin{cases} 1, & 0 < \theta < \pi; \\ -1, & \pi < \theta < 2\pi. \end{cases}$$

Then

$$|v(r,\theta)| \;\leq\; c(r), \;\; \textit{where}\; c_0(r) \leq 1 \;\textit{and}\; c_0(r) = O(r) \;\textit{as}\;\; r \to 0. \tag{7.3.1}$$

Proof. Since $v(-z) + v(z) = 0$ and v is positive for $\operatorname{Re} z > 0$, it is enough to bound $v(z)$ for $\operatorname{Re} z > 0$. The Poisson integral formula 5.1.6 here becomes

$$\begin{aligned} v(re^{i\theta}) &= \frac{1}{2\pi}\int_0^{\pi}\left\{\frac{1-r^2}{1-2r\cos(\theta-\varphi)+r^2}-\frac{1-r^2}{1-2r\cos(\theta+\varphi)+r^2}\right\}d\varphi \\ &= \frac{1}{2\pi}\int_0^{2\pi}\frac{(1-r^2)\,4r\sin\theta\sin\varphi}{(1-2r\cos(\theta-\varphi)+r^2)(1-2r\cos(\theta+\varphi)+r^2}\,d\varphi. \end{aligned}$$

The integrand is bounded by $4(1+r)r(1-r)^{-3}\sin\theta\sin\varphi$. We also know from the maximum principle that $v(z) < 1$ for $\operatorname{Re} z > 0$, so integrating gives (7.3.1) with

$$c_0(r) \;=\; \min\left\{1,\, \frac{4r(1+r)|\sin\theta|}{(1-r)^3}\right\}. \tag{7.3.2}$$

□

Lemma 7.3.4. *Suppose that u is harmonic for $r_0 < |z| < 1$, continuous for $r_0 \le |z| \le 1$, and constant for $|z| = r_0$. Let*

$$M_r(u) \;=\; \max_{|z|=r} u(z) - \min_{|z|=r} u(z), \quad r_0 \le r \le 1.$$

Then

$$M_r(u) \;\le c(r)\, M_1(u), \tag{7.3.3}$$

where $c(r) = \pi c_0(r)$, with c_0 given by (7.3.2).

Proof. Up to a rotation and multiplication by a constant, we may assume that $M_1(u) = 1$ and, for a given value of r, that the maximum and minimum values of $u(z)$ for $|z| = r$ occur at complex conjugate points z_0 and $\bar{z}_0$ respectively. The function $\widetilde{u}(z) = u(z) - u(\bar{z})$ is harmonic in the intersection U of the annulus with the upper half-plane, continuous on the closure of U, equal to zero on the lower boundary and is ≤ 1 on the upper boundary. See Figure 7.1.

Let v be the function of Lemma 7.3.3. Then $v \ge 0$ on the lower boundary of U, so $\widetilde{u} \le v$ on U and (7.3.1) applies. □

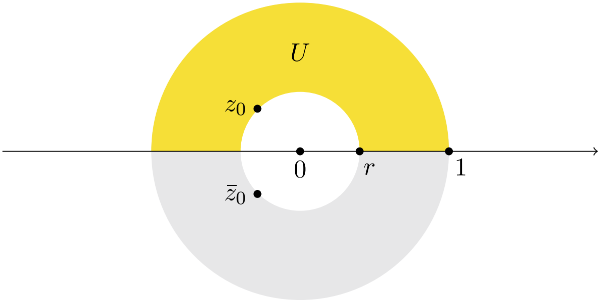

Fig. 7.1 The domain U.

The next several lemmas are aimed at estimating the mean value and the oscillation, and therefore the size, of $u_\rho - \mathrm{Re}\,(1/z)$. The goal is to show convergence of $u_\rho - \mathrm{Re}\,(1/z)$, as $\rho \to 0$, to a harmonic function with the desired singularity at p_0.

We begin with *Green's identity* for functions u, v that are smooth on the closure of bounded domain $U \subset S$ having smooth boundary:

$$\int_U [v\Delta u - u\Delta v]\, dm \;=\; \int_{\partial U} \left[v\frac{\partial u}{\partial n} - u\frac{\partial v}{\partial n}\right] ds, \tag{7.3.4}$$

where dm is the area measure, $\partial/\partial n$ is the outer normal derivative, and ds is arc-length measure on ∂U. If u and v are harmonic, this becomes

$$\int_{\partial U} \left[v\frac{\partial u}{\partial n} - u\frac{\partial v}{\partial n}\right] ds \;=\; 0. \tag{7.3.5}$$

In particular, suppose that u is harmonic and $v \equiv 1$, and U is an annulus, in a standard coordinate, bounded by the circles $|z| = r_1 < r_2$ and $|z| = r_2$. Then (7.3.5) implies

$$\int_0^{2\pi} \frac{\partial u}{\partial r}(r_1 e^{i\theta})\, d\theta \;=\; \int_0^{2\pi} \frac{\partial u}{\partial r}(r_2 e^{i\theta})\, d\theta. \tag{7.3.6}$$

Lemma 7.3.5. *The function u_ρ of Lemma 7.3.2 satisfies*

$$\int_0^{2\pi} \frac{\partial u_\rho}{\partial r}(re^{i\theta})\, d\theta \;=\; 0. \tag{7.3.7}$$

Proof. Suppose that $D_1 \subset K$, where K has compact closure and smooth boundary. Then each point of ∂K is regular, so D_r has harmonic measure v in K, namely the solution of the Dirichlet problem that is 1 on ∂D_r and 0 on ∂K. Then (7.3.5) specializes to

$$\int_{\partial D_\rho} \left[v\frac{\partial u_\rho}{\partial n} - u_\rho\frac{\partial v}{\partial n}\right] ds \;=\; \int_{\partial K} \left[v\frac{\partial u_\rho}{\partial n} - u_\rho\frac{\partial v}{\partial n}\right] ds. \tag{7.3.8}$$

By the maximum principle, $u_\rho \le 1/\rho$. Therefore (7.3.8) leads to

$$\left| \int_{\partial D_\rho} \frac{\partial u_\rho}{\partial n}\, ds \right| \;\le\; \frac{1}{\rho}\left| \int_{\partial D_\rho} \frac{\partial v}{\partial n}\, ds \right| + \frac{1}{\rho}\left| \int_{\partial K} \frac{\partial v}{\partial n}\, ds \right|. \tag{7.3.9}$$

We know that ∂D_r does not have harmonic measure, so as we take larger sets $K' \supset K$ the harmonic measures $v_{K'}$ for ∂D_r on K' increase to 1 uniformly on K. Replacing v by $v_{K'}$ we find that

$$\int_{\partial D_\rho} \frac{\partial u_\rho}{\partial n}\, ds \;=\; 0.$$

For the standard coordinate z, this is (7.3.7) at $r = \rho$. By (7.3.6), the integral in (7.3.7) is independent of r. □

Lemma 7.3.6. *The mean value of u_ρ over any circle $\{z : |z| = r\}$, $0 < r < 1$, is zero.*

Proof. By (7.3.7), it is enough to prove this at $r = \rho$. On ∂D_ρ, $u_\rho = \text{Re}\,(1/z)$, which has mean value zero. □

Lemma 7.3.7. *As $\rho \to 0$, u_ρ tends to a function u that is harmonic on $S \setminus \{p_0\}$, and bounded on the complement of each neighborhood of p_0. Moreover in a standard coordinate at p_0,*

$$\lim_{z\to 0}\left[u - \text{Re}\,\frac{1}{z}\right] = 0.$$

Proof. By Lemma 7.3.4 applied to $u - \text{Re}\,(1/z)$, the oscillation satisfies

$$M_r\left(u - \text{Re}\,\frac{1}{z}\right) \le c(r)\, M_1\left(u - \text{Re}\,\frac{1}{z}\right). \tag{7.3.10}$$

Now $M_r(\text{Re}\,(1/r)) = 2/r$, and the maximum principle implies that $M_1(u_\rho) \le M_r(u_\rho)$. Therefore

$$\begin{aligned} M_1(u_\rho) - \frac{2}{r} &\le M_r\left(u_\rho - \text{Re}\,\frac{2}{z}\right) \\ &\le c(r)\, M_1\left(u_\rho - \text{Re}\,\frac{2}{z}\right) \\ &\le c(r)\,[M_1(u_\rho) + 2]. \end{aligned}$$

Choose r_1 so that $c(r_1) < 1$. Then

$$M_1(u_\rho) \le 2\,\frac{c(r_1) + 1/r_1}{1 - c(r_1)} = C_1, \tag{7.3.11}$$

independent of $\rho < r_1$. Returning to Lemma 7.3.4, we see that by the scaling $z \to z/r_1$ we obtain a standard coordinate at p_0, and a corresponding version of (7.3.3) with $c_1(r) = c(r/r_1)$ as the multiplier. Therefore for $\rho < r < 1$,

$$\begin{aligned} M_r\left(u_\rho - \text{Re}\,\frac{1}{z}\right) &\le c\left(\frac{r}{r_1}\right) M_1\left(u_\rho - \text{Re}\,\frac{r_1}{z}\right) \\ &\le c_1(r)\left(C_1 + \frac{2r}{r_1}\right) \\ &= C_2\, c_1(r). \end{aligned} \tag{7.3.12}$$

By Lemma 7.3.6, the mean value of $u_\rho - \text{Re}\,(1/z)$ over a circle is zero. It follows that

$$\min_{|z|=r}\left(u_\rho - \text{Re}\,\frac{1}{z}\right) \le 0 \le \max_{|z|=r}\left(u_\rho - \text{Re}\,\frac{1}{z}\right).$$

Therefore (7.3.12) implies that

$$\max_{|z|=r}\left|u_\rho - \mathrm{Re}\,\frac{1}{z}\right| \;\le\; C_2\,c_1(r), \tag{7.3.13}$$

and so

$$\max_{|z|=r}|u_\rho - u_{\rho'}| \;\le\; C_2\,c_1(r), \quad \rho,\rho' < r < r_1. \tag{7.3.14}$$

By the maximum principle, (7.3.14) is also valid outside D_{r_1}. It follows that u_ρ has limit u as $\rho \to 0$, uniformly on the complement of any neighborhood of p_0. Moreover (7.3.13) and (7.3.14) imply that as $r \to 0$,

$$\max_{|z|=r}\left|u - \mathrm{Re}\,\frac{1}{z}\right| \;=\; o(r), \qquad \max_{|z|=r}\left|u_\rho - u\right| \;=\; o(r). \qquad \square$$

We are now in a position to complete the proof of the uniformization theorem.

7.4 Proof of the uniformization theorem, completed

We continue to assume that S is either parabolic or elliptic. We have established that for each point p_0 of S and each standard coordinate z at p_0, there is a function u, harmonic in $S \setminus \{p_0\}$, bounded outside any neighborhood of p_0, such that $u - \mathrm{Re}\,(1/z)$ has limit 0 at p_0 and $\liminf_{p\to\infty} u(p) = 0$. For each point of $S \setminus \{p_0\}$ and standard neighborhood $D_1 \subset S \setminus \{p_0\}$, we may choose a harmonic conjugate v, unique up to an additive constant, such that $f = u + iv$ is holomorphic in D_1. The same is true in $D_1(p_0)$, modulo the pole at p_0: choose w a harmonic conjugate to $u - \mathrm{Re}\,(1/z)$ with $v(0) = 0$, and let $v = w + \mathrm{Im}\,(1/z)$, so that $f = u + iv$ is meromorphic near p_0 with expansion

$$f(p) \;=\; \frac{1}{z} + az + \dots\,. \tag{7.4.1}$$

This function may be continued along any curve in S by adjusting the additive constant along an overlapping chain of coordinate neighborhoods that cover the curve. As in the hyperbolic case, simple connectedness tells us that, starting from $D_1(p_0)$, f has a unique analytic continuation to all of S.

In the previous standard neighborhood of p_0 we can take $\widetilde{z} = -iz$ as our standard coordinate and construct the analogous function $\widetilde{f}$, with an expansion

$$\widetilde{f}(p) = \frac{i}{z} + \widetilde{a}z + \dots\,. \tag{7.4.2}$$

We know that $u = \mathrm{Re}\,f$ is bounded outside each neigborhood of p_0. We do not yet know that this is true of $v = \mathrm{Im}\,f$ and, therefore of f itself. The following proposition settles this point.

Proposition 7.4.1. *Let f and $\widetilde{f}$ be the functions constructed above with expansions (7.4.1) and (7.4.2). Then $\widetilde{f} - if$ is constant.*

Proof. Near p_0 we work in a standard coordinate chart. Since f and $\widetilde{f}$ each have a simple pole at 0, it follows that if ρ is small enough, and $p_1 \in D_\rho$, then f takes the value $f(p_1)$ exactly once in D_r, and the same is true for $\widetilde{f}$. It will be useful to choose p_1 so that we also have $f(p) \neq f(p_1)$ for p in the complement of D_ρ.

To accomplish this, choose M so that

$$|\mathrm{Re}\, f(p)| \leq M, \quad |\mathrm{Re}\, \widetilde{f}(p)| \leq M, \qquad p \in S \setminus D_\rho,$$

and choose $p_1 = z_1 = (1+i)/\varepsilon$, where $\varepsilon > 0$ is small enough that

$$|\mathrm{Re}\, f(p_1)| > M, \quad |\mathrm{Re}\, \widetilde{f}(p_1)| > M.$$

Therefore we also have $f(p) \neq f(p_1)$ and $\widetilde{f}(p) \neq \widetilde{f}(p_1)$ if p is in the complement of D_ρ. It follows that the functions

$$F(p) = \frac{1}{\mathrm{Re}\, f(p) - M}, \qquad \widetilde{F}(p) = \frac{1}{\mathrm{Re}\, \widetilde{f}(p) - M} \tag{7.4.3}$$

are holomorphic except for simple poles at p_1, and vanish at 0. Therefore near p_1 they have expansions

$$F(p) = \frac{A}{z - z_1} + B + O(z - z_1), \qquad \widetilde{F}(p) = \frac{\widetilde{A}}{z - z_1} + \widetilde{B} + O(z - z_1). \tag{7.4.4}$$

Then $G = \widetilde{A}F - A\widetilde{F}$ is holomorphic on S. On the complement of D_ρ,

$$|G(p)| \leq C(|F(p)| + |\widetilde{F}(p)|) \leq \frac{C}{\mathrm{Re}\, F(p_1) - M} + \frac{C}{\mathrm{Re}\, \widetilde{F}(p_1) - M},$$

so G is a bounded holomorphic function on S, hence is a constant C_1. Therefore

$$\widetilde{A}[\widetilde{f}(p) - \widetilde{f}(p_1)] - A[f(p) - f(p_1) = C_1[f(p) - f(p_1)][\widetilde{f}(p) - \widetilde{f}(p_1)].$$

Since the left-hand side has at most a simple pole at p_0, it follows that $C_1 = 0$. The expansions (7.4.1) and (7.4.2) show that $\widetilde{A} = -iA$, so

$$\widetilde{f}(p) = if(p) + [\widetilde{f}(p_1) - if(p_1)].$$

But the expansions (7.4.1) and (7.4.2) show that the term in brackets is zero. □

Let us denote the function f with pole at p_0 by $f(p; p_0)$ and the corresponding function with pole at the point p_1 of Proposition 7.4.1 by $f(p; p_1)$.

Proposition 7.4.2. *The function $f(p; p_0)$ is injective.*

Proof. Let F be the function of (7.4.3). Then F and $f(p;\, p_1)$ are both meromorphic in S with a simple pole at p_1 and bounded outside any neighborhood of p_1. Therefore $F(p) = af(p;\, p_1) + b$, where a and b are constants. Since F is a linear fractional transformation of $f(p;\, p_0)$, it follows that $f(p;\, p_1)$ is a linear fractional transformation of $f(p;\, p_0)$. This is true for any p_1 in D_ρ. Continuing this argument along an overlapping chain of neighborhoods, we find that each $f(p;\, q)$ is a linear fractional transformation $T = T_q$ of $f(p;\, p_0)$.

Suppose now that $f(p_1;\, p_0) = f(p_2;\, p_0)$. Choose T so that $f(p;\, p_2) = Tf(p;\, p_0)$. Then

$$f(p_1;\, p_2) \;=\; Tf(p_1;\, p_0) \;=\; Tf(p_2;\, p_0) \;=\; f(p_2,\, p_2) \;=\; \infty.$$

But the only pole of $f(p;\, p_2)$ is at p_2, so $p_1 = p_2$. □

Theorem 7.4.3. (**Uniformization: the parabolic and elliptic cases**) *A simply connected parabolic or elliptic Riemann surface is biholomorphically equivalent to $\mathbb{C}$ or $\mathbb{S}$, respectively.*

Proof. The function $f(p;\, p_0)$ is an injective holomorphic map to an open subset U of $\mathbb{S}$. In the elliptic case the image must be compact, hence all of $\mathbb{S}$. Otherwise, if f omitted more than one point, then the Riemann mapping theorem would provide an equivalence with the disk and, therefore, a bounded holomorphic function. Therefore in the parabolic case f omits a single point $a \in \mathbb{C}$. Then $f(p) - a$ reaches every $z \in \mathbb{C}$, $z \neq 0$, so

$$g(p) \;=\; \frac{1}{f(p) - a},$$

which vanishes at p_0, is an equivalence of S and $\mathbb{C}$. □

Exercises

In the following exercises, S is a Riemann surface, U is a non-empty open subset of S, and Φ is a conformal map of S onto itself.

1. Suppose that $u : U \to \mathbb{R}$ is harmonic. Show that $u \circ \Phi^{-1}$ is harmonic on $\Phi(U)$.
2. Suppose that $v : U \to \mathbb{R}$ is subharmonic. Show that $v \circ \Phi^{-1}$ is subharmonic on $\Phi(U)$.
3. Suppose that $\mathscr{F}$ is a Perron family on U, Show that

$$\{u \circ \Phi^{-1} \;:\; u \in \mathscr{F}\}$$

   is a Perron family on $\Phi(U)$.
4. Given $p_0 \in S$, show that

$$\{u \circ \Phi^{-1} \;:\; u \in \mathscr{F}(p_0)\} \;=\; \mathscr{F}(\Phi(p_0)).$$

5. Prove Corollary 7.1.5.

6. Show that $\mathbb{D}$ has a Green's function with pole at 0.
7. Show that $\mathbb{H}$ has a Green's function with pole at i.
8. Given $R > 0$, show that there is a subharmonic function $u : \mathbb{C} \to \mathbb{R}$ such that $u + \log|z|$ is bounded for $0 < |z| < R$ and u is harmonic and positive for $0 < |z| < R$, $u = 0$ for $|z| \geq R$.
9. Show that $\mathbb{C}$ does not have a Green's function with pole at 0.
10. Show that $\mathbb{C}$ does not have any Green's function.
11. Show that $\mathbb{S}$ does not have any Green's function.
12. Show how elliptic and parabolic simply connected Riemann surfaces can be treated as limiting cases of hyperbolic simply connected Riemann surfaces.
13. Suppose that Ω is a Jordan domain in $\mathbb{C}$ with a barrier at each point of the boundary Γ, and $z_0 \in \Omega$. Let $u : \overline{\Omega} \to \mathbb{R}$ be the solution of the Dirichlet problem with $u = -\log|z - z_0|$ on Γ.
    (a) Show that the harmonic conjugate u^* is single-valued in Ω. Hint: Ω is simply connected. Choose u^* with $u^*(z_0) = 0$.
    (b) Show that $f = \exp(u + iu^*)(z - z_0)$ is a conformal map of Ω to $\mathbb{D}$ and $|f(z)| \to 1$ as $|z| \to \partial\Omega$.
    (c) Show that $f : \Omega \to \mathbb{D}$ is conformal. Hint: f has a unique zero in Ω.
14. (a) Suppose that $\Omega \subset \mathbb{C}$ is an arbitrary Jordan domain. Show that for each $n = 1, 2, \ldots$ there is a domain $\Omega_n \subset \Omega$ whose boundary Γ_n is a polygon contained in a $1/n$ neighborhood of $\Gamma = \partial\Omega$.
    (b) Show that there is a conformal map from Ω onto $\mathbb{D}$.
15. Suppose that $\Omega \subset \mathbb{C}$ is a domain whose boundary contains at least two points. Use the results of the preceding exercises to show that there is a conformal map of Ω onto $\mathbb{D}$. (Note that we may assume that Ω is bounded: see the first step in the proof of Theorem 2.4.1.)
16. Exercises 12 – 13 of Chapter 8 show that any domain $\Omega \in \mathbb{C}$ whose complement has two connected pieces, one bounded and one unbounded, can be mapped conformaly to a unique open annulus $A(1, r) = \{z : 1 < |z| < r\}$. The purpose of this and the following exercises are to prove a mapping theorem applicable to plane domains with any connectivity.
    Suppose that the complement of $\overline{\Omega} \subset \mathbb{C}$ consists of m disjoint connected sets, $\Omega_1, \ldots, \Omega_m$, Ω_m unbounded and the others bounded.
    Continuing, using inversions, show that we may assume inductively that each $\Gamma_j = \partial\Omega_j$ is an analytic curve, and that Γ_1 is a circle enclosing Ω, with the usual orientation.
17. Let z_0 be a point of Ω and let u be the harmonic function that is the solution of the Dirichlet problem with value $-\text{Re}\,(1/(z - z_0)$ on $\partial\Omega$.
    (a) Show that u has a harmonic extension to a neighborhood of $\partial\Omega$. Hint: use analyticity of the Γ_j.
    (b) Let u^* be a harmonic conjugate of u. Note that it may not be single-valued: it may have period a_k on Γ_k, i.e. a gain a_k as it is continued around Γ_k in the positive direction. Show that *locally*

$$u + iu^* + \frac{1}{(z - z_0)}$$

is holomorphic in $\Omega \setminus \{z_0\}$, with a simple pole at z_0, extends locally to be holomorphic in a neighborhood of $\partial\Omega$, and has real part zero on $\partial\Omega$.

18. With Ω as in Exercise 16, let u_j, $1 \leq j \leq m$ be the harmonic function on Ω that is the solution of the Dirichlet problem with value 1 on Γ_j and 0 on the other Γ_k. As in Exercise 16 (a), each u_j has a harmonic extension to a neighborhood of $\overline{\Omega}$.

    (a) Let a_{jk} be the period of u_j on Γ_k. Prove that the homogeneous system

    $$\sum_{j=1}^{m} \lambda_j a_{jk} = 0, \quad k = 1, 2, \dots, m$$

    has only the trivial solution all $\lambda_j = 0$. Hint: consider the real and imaginary parts separately.

    (b) Show that there is a linear combination of u^* above and the u_j such that

    $$u^* + \sum_{j=1}^{m} \lambda_j u_j$$

    has period zero on each Γ_k.

    (c) Show that

    $$f = u + iu^* + \sum_{j=1}^{m} \lambda_j u_j + \frac{1}{z - z_0}$$

    is single-valued and meromorphic in a neighborhood of $\overline{\Omega}$.

19. Show that the function f of Exercise 18 is a conformal map of Ω onto the complement of a set of disjoint horizontal slits $\{z : b_j \leq \operatorname{Re} z \leq c_j, \operatorname{Im} z = d_j\}$, $j = 1, \dots, m$.

Remarks and further reading

The formulation and proof of the uniformization theorem involved many of the leading analysts of the late 19th and early 20th centuries, including Schwarz, Klein, Poincaré and Koebe. This history is summarized, and an alternative proof is sketched, in Abikoff's Monthly article [2]; see also the discussion in §20 of Weyl [214]. Gray [92] has a detailed history of the proof, with discussion of the work of the previously mentioned authors as well as Osgood, Carathéodory, Bieberbach, and others. The theorem is covered in most texts on Riemann surfaces; see the references at the end of Chapter 6. Our presentation here mainly follows Ahlfors [7].

Chapter 8
Quasiconformal mapping

Conformal equivalence is a basic concept in complex analysis. In connection with simply connected proper domains in $\mathbb{C}$, it is very flexible, as shown by the Riemann mapping theorem. But, for domains that are not simply connected, it is much more rigid. As we shall see, two annuli $A_j = \{z : 1 < |z| < R_j\}$, $j = 1, 2$, are conformally equivalent if and only if $R_1 = R_2$. A more flexible, and very useful, concept is that of quasiconformal equivalence. This chapter covers the basic theory of quasiconformal mapping.

Section 8.1 introduces a general notion of a quadrilateral and the fundamental concept of the module of a quadrilateral. This provides the basis for the definition of a quasiconformal map in Section 8.2. Regular, i.e. C^1, conformal maps are characterized in Section 8.3.

Section 8.4 introduces ring domains, i.e. domains quasiconformally equivalent to annuli. The importance of ring domains comes from their separation property – separating the region surrounded by the ring from the region external to the ring. Of particular importance are *extremal* ring domains, the subject of Section 8.5. An extremal ring domain is one that has maximal module among all ring domains that have a specified separation property. As shown in Section 8.6, results on such domains are powerful tools for establishing such properties as Hölder continuity of quasiconformal mappings.

The question of the relation of a quasiconformal map of the closed upper half-plane to its restriction to the real line is treated in Section 8.7. Quasisymmetry, quasi-isometry, and the Beurling–Ahlfors extension of a map of $\mathbb{R} \to \mathbb{R}$ to a quasiconformal map $\mathbb{H} \to \mathbb{H}$ are covered.

Section 8.8 deals with the existence of quasiconformal maps with given dilatation via the Beltrami equation. The existence theory uses a special case of the Calderón-Zygmund inequality, which is proved in Section 8.9.

A principal motivation for the development of this theory was its use in the problem of finding a space of moduli to characterize Riemann surfaces – the subject of Chapter 9.

R. Beals and R. S. C. Wong, *More Explorations in Complex Functions*, Graduate Texts in Mathematics 298, https://doi.org/10.1007/978-3-031-28288-1_8

8.1 Quadrilaterals

By a *quadrilateral* in $\mathbb{C}$, we mean a Jordan domain Q with four distinguished points p_1, p_2, p_3, p_4 on the boundary, numbered in the positive direction and referred to as the *vertices* of Q. The boundary arcs between p_1 and p_2 and between p_3 and p_4 are referred to as the "a sides" of Q, and the other two arcs as the "b sides." As a first example, consider the unit disk $\mathbb{D}$ with four such points designated on the boundary. There is a unique conformal map in $f \in \operatorname{Aut}(\mathbb{D})$ that maps the ordered triple (p_1, p_1, p_3) to $(-1, -i, 1)$. Then $f(p_4)$ is uniquely determined; see Exercise 1. This shows that the conformal class of such a quadrilateral can be parametrized by a single real variable, e.g. $\operatorname{Re} f(p_4)$. A second natural example is a rectangle, with the usual vertices numbered in the positive direction from some choice of initial vertex. Here, as we shall see, a natural parametrization of the conformal class is the ratio of the a side lengths to the b side lengths.

We shall immediately broaden the definition to allow what can be thought of as Jordan domains in the Riemann sphere. The principal example is $\mathbb{H}$, with boundary $\mathbb{R} \cup \{\infty\}$, and vertices numbered in the positive direction. Usually, we shall take these vertices to be $(-1/k, -1, 1, 1/k)$, where $0 < k < 1$. Given any quadrilateral $Q(p_1, p_2, p_3, p_4)$, there is a unique choice of k such that

$$Q(p_1, p_2, p_3, p_4) \quad \text{is conformally equivalent to} \quad \mathbb{H}(-1/k, -1, 1, 1/k); \tag{8.1.1}$$

see Exercise 2. The quadrilateral $\mathbb{H}(-1/k, -1, 1, k)$ can be mapped to a rectangle with vertices $-K, K, K + iK', -K + iK'$, for suitable $K, K' > 0$, by the function

$$F(z) = \int_0^z \frac{d\zeta}{\sqrt{(1-\zeta^2)(1-k^2\zeta^2)}}; \tag{8.1.2}$$

see Exercise 3. In view of this, it is easily seen that any quadrilateral is conformally equivalent to a rectangle, as illustrated in Figure 8.1.

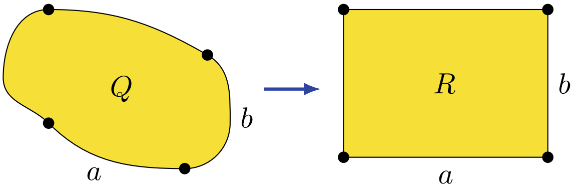

Fig. 8.1 A quadrilateral and a conformally equivalent rectangle.

Remark. Because we have assumed a quadrilateral to be a Jordan domain, such a conformal map to a rectangle extends to a homeomorphism (i.e. a continuous map with continuous inverse) from the closure of Q to the closure of R. This is really the necessary feature, so we may build it into the definition of a quadrilateral and of maps between quadrilaterals.

Two rectangles with specified a and b sides are conformally equivalent if and only if they are similar; see Exercise 5.

Proposition 8.1.1. *Two quadrilaterals Q, Q' in $\mathbb{C}$ are conformally equivalent if and only if the canonical images R, R' are similar: $a/b = a'/b'$.*

We have shown, in effect, that any quadrilateral Q has a *canonical image* with vertices $0, a, a + bi, bi$ for some choice of a, b. We will also refer to a canonical image of Q as a *model* of Q With Proposition 8.1.1 in mind, and abusing notation a bit, we note that for a rectangle whose a-sides have length a and whose b-sides have length b, the similarity class is determined by the ratio a/b. In general, the *module* $m(Q)$ of a quadrilateral Q is defined to be a/b, where a and b are the appropriate side lengths of a model of Q. The module is, by definition, a conformal invariant.

By changing the numbering of the vertices of Q, we can convert to a quadrilateral Q^* whose a-sides are the b-sides of Q. Clearly,

$$m(Q^*) \;=\; \frac{1}{m(Q)}. \tag{8.1.3}$$

It is useful to have a second characterization of $m(Q)$.

Proposition 8.1.2. *For any quadrilateral Q,*

$$m(Q) \;=\; \sup_{\rho} \frac{L(\rho)^2}{A(\rho)}, \tag{8.1.4}$$

where ρ runs through the non-zero functions that are non-negative on Q and have finite integral

$$A(\rho) \;=\; \iint_Q \rho(x+iy)^2\,dx\,dy,$$

and $L(\rho)$ is the infimum of

$$L_\gamma(\rho) \;=\; \int_\gamma \rho(z)|dz|$$

over rectifiable curves γ that join the two b-sides of Q.

Proof: It is enough to pass to a model R with side lengths a, b. Then

$$L(\rho) \;\le\; \int \rho(x+iy)\,dx,$$

so

$$b\,L(\rho)\ \le\ \iint_R \rho\,dx\,dy,$$

and the Cauchy–Schwarz inequality gives

$$b^2 L(\rho)^2 \le \iint_R^{\rho} \rho^2\,dx\,dy \iint_R dx\,dy\ =\ A(\rho)\,ab. \tag{8.1.5}$$

Thus, the expression on the right-hand side of (8.1.4) is $\le a/b = m(Q)$. Conversely, returning to R with $\rho = 1$, it is clear that we obtain equality in (8.1.4). □

Remarks. The quotient on the right in (8.1.4) is unchanged if ρ is multiplied by a positive constant. Moreover, equality occurs in (8.1.5) if and only if ρ is a constant multiple of 1. Thus, the choice $\rho = 1$ is essentially the only choice that gives equality in (8.1.4). As we shall see, this has important consequences.

It will be useful to check what happens if we split a quadrilateral into "vertical" or "horizontal" strips.

Proposition 8.1.3. *Suppose that the quadrilateral Q is divided into quadrilaterals Q_1, Q_2 by a curve κ that joins the two a sides of Q. Then*

$$m(Q)\ \ge\ m(Q_1) + m(Q_2). \tag{8.1.6}$$

Equality holds if and only if the image of κ in any model R of Q is a vertical line segment.

Proof: It is enough, once again, to work directly with a model with side lengths a, b. By the remark, we may work with $\rho = 1$. Let l_1 be the minimal distance from κ to the left side of R and l_2 the minimal distance to the right side of R. Then the area $A_j = A(Q_j)$ is $\ge l_j b$, with equality if and only if κ is a vertical segment. Then

$$m(Q_1) + m(Q_2)\ =\ \frac{l_1^2}{A_1} + \frac{l_2^2}{A_2} \le \frac{l_1^2}{l_1 b} + \frac{l_2^2}{l_2 b}\ =\ \frac{l_1 + l_2}{b}\ \le\ \frac{a}{b},$$

with equality if and only if κ is a vertical segment. □

The assumptions of the following useful lemma are illustrated in Figure 8.2.

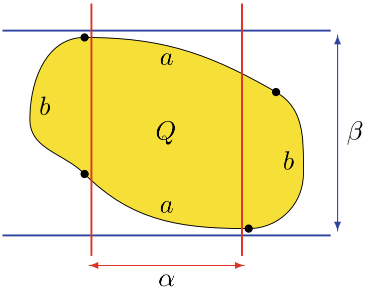

Fig. 8.2 Estimating the module of Q from below.

Lemma 8.1.4. *Suppose that the quadrilateral Q is contained in a horizontal strip of width β and the b sides of Q are separated by a vertical strip of width α. Then the module of Q is $\geq \alpha/\beta$.*

Proof: Define ρ on Q to be 1 between the vertical strips and zero elsewhere. Then $L(\rho) \geq \alpha$ and $A(\rho) \leq \alpha\beta$, so $m(Q) \geq \alpha^2/\alpha\beta$. □

We say that a sequence of quadrilaterals $\{Q_n\}$ *converges uniformly* to a bounded quadrilateral Q if for each $\varepsilon > 0$, $n \geq n(\varepsilon)$ implies that each a side of Q_n lies in an ε neighborhood of the corresponding a side of Q, and the same for the b sides. Consequently, Q_n lies in an ε neighborhood of Q and conversely.

Theorem 8.1.5. *If the sequence of quadrilaterals $Q_n \subset Q$ converges uniformly to the bounded quadrilateral Q, then*

$$\lim_{n\to\infty} m(Q_n) \;=\; m(Q).$$

Proof: Replace Q by a model and use Lemma 8.1.4 to estimate $m(Q_n)$. □

8.2 Quasiconformal mappings

We can now define the subject of this chapter. A homeomorphism f from a domain $\Omega \subset \mathbb{C}$ to a domain $\Omega' \subset \mathbb{C}$ is K*-quasiconformal*, $K < \infty$, if for each quadrilateral $Q \subset \Omega$ whose boundary in contained in Ω,

$$m(f(Q)) \;\leq\; K\, m(Q).$$

In view of (8.1.3), this implies that also $1/m(f(Q)) \leq 1/m(Q)$, so f is K-quasiconformal if and only if for each such quadrilateral $Q \subset \Omega$,

$$\frac{1}{K}\, m(Q) \;\leq\; m(f(Q)) \;\leq\; K\, m(Q). \tag{8.2.1}$$

Proposition 8.2.1. *(a) If f is K-quasiconformal, so is its inverse.*
(b) If f_j is K_j-quasiconformal, $j = 1, 2$, then $f_1 \circ f_2$ is K-quasiconformal, with $K \leq K_1 K_2$.
(c) If f is K-quasiconformal, then $K \geq 1$.
(d) f is 1-quasiconformal if and only if it is conformal.

Proof: (a) and (c) follow from (8.2.1), while (b) is obvious. The first part of (d) is clear, since a conformal map preserves modules.

Conversely, suppose that f is 1-quasiconformal. It is enough to show that it is conformal on each quadrilateral Q to $f(Q)$, and by composing with conformal maps, we may reduce to the case of a 1-quasiconformal map g that maps a model R to a model R'. Since R and R' have the same module, we may take $R' = R$ and assume that the lower side of R is the interval $[0, a]$. Given $0 < x < a$, let γ be the vertical segment from a to $a + ib$ and let $\kappa = g(\gamma)$. An application of Proposition 8.1.3 shows that κ must also be a vertical segment – in fact the same vertical segment. Thus, $\operatorname{Re} g(x + iy) = x$. The same argument applied to horizontal segments implies that $\operatorname{Im} g(x + iy) = y$. Thus, f, composed with certain conformal maps, is the identity. It follows that f is conformal. □

We now pass to some consequences of the approximation in Lemma 8.1.4 and Proposition 8.1.3.

Theorem 8.2.2. *Suppose that the quadrilateral Q is mapped to a quadrilateral Q' by a continuous map f that is K-quasiconformal on the interior of Q. Then f is K-quasiconformal on Q.*

Proof: Because of conformal invariance, we may replace Q and Q' by models R, R' with sides a, b and a', b'. Let $\widetilde{R}$ be a rectangle whose closure is contained in the interior of R, with sides $\widetilde{a}, \widetilde{b}$. By continuity, we may assume that the image $\widetilde{R}'$ is contained in a vertical strip in R' of width $\geq a' - \varepsilon$. By Lemma 8.1.4, the module

$$\frac{a' - \varepsilon}{b'} \leq m(\widetilde{R}') \leq Km(R).$$

As $\widetilde{R}$ increases to R, this shows that $a'/b' \leq K \cdot a/b$. □

Theorem 8.2.3. *Suppose that a map f is continuous on a domain Ω and K-quasiconformal on $\Omega \setminus \gamma$, where γ is an analytic arc. Then f is K-quasiconformal on Ω.*

Proof: From conformal invariance and the proof of Theorem 8.2.2, we may replace any quadrilateral in Ω by a smaller quadrilateral. The smaller quadrilateral intersects γ in a finite set of disjoint analytic arcs. Thus, we may reduce the problem to two rectangles R and R', and we may reverse the viewpoint and take γ' in R' to be analytic. It is enough to remove subarcs of $\gamma' \cap R'$ one at a time. If such a subarc is contained in a vertical line, then Proposition 8.1.3 allows us to remove that line and reduce the problem to consideration of each of the two resulting rectangles. Continuing this process, we may remove all such arcs and assume that $\gamma' \cap R'$ can be decomposed into arcs that intersect each vertical line at most once. Passing vertical lines through each endpoint of such an arc allow us to invoke Proposition 8.1.3 again, and reduce to the case that γ' runs from one vertical side of R' to the other.

Under this assumption, divide R' into vertical strips R'_j, which are divided into parts Q'_{j1}, Q'_{j2} by γ. These are the images of Q_{j1}, Q_{j2} in R, with moduli m_{j1}, m_{j2}. From Proposition 8.1.2

$$\frac{1}{m_{j1}} \geq \frac{b_{j1}^2}{A_{j1}}, \qquad \frac{1}{m_{j2}} \geq \frac{b_{j2}^2}{A_{j2}}, \tag{8.2.2}$$

where A_{j1}, A_{j2} are the areas and b_{j1}, b_{j2} are the shortest distances from γ to the horizontal sides of R. If the strips Q'_j are narrow enough, $b_{j1}^2 + b_{j2}^2 \geq (b-\varepsilon)^2$, where a, b are the horizontal and vertical side lengths for R. Then (8.1.6) gives

$$\begin{aligned}\frac{1}{m_{j1}} + \frac{1}{m_{j2}} &\geq \frac{b_{j_1}^2}{A_{j_1}} + \frac{b_{j_2}^2}{A_{j_2}} \\ &\geq \frac{b_{j_1}^2 + b_{j_2}^2}{A_{j_1} + A_{j_2}} \geq \frac{(b-\varepsilon)^2}{A_{j_1} + A_{j_2}}.\end{aligned}$$

But also

$$\frac{1}{m'_j} \geq \frac{1}{m'_{j1}} + \frac{1}{m'_{j2}} \geq \frac{1}{K}\left(\frac{1}{m_{j1}} + \frac{1}{m_{j2}}\right),$$

so

$$m'_j \leq K\,\frac{A_{j1} + A_{j2}}{(b-\varepsilon)^2}.$$

Therefore

$$m' = \sum_j m'_j < K \sum_j \frac{A_{j1} + A_{j2}}{(b-\varepsilon)^2} \leq K\frac{ab}{(b-\varepsilon)^2},$$

which, in the limit, gives $m' \leq Km$. □

As we shall see, quasiconformality, like continuity, is a local concept. We take the *maximal dilatation* of a quasiconformal map $f : \Omega \to \Omega'$ to be

$$K_f(\Omega) = K(\Omega) = \sup_{Q \subset \Omega} \frac{m(f(Q))}{m(Q)}. \tag{8.2.3}$$

The maximal dilatation of f at a point $p \in \Omega$ is

$$K_f(p) = \inf\{K(U) : U \text{ a neighborhood of } p\}. \tag{8.2.4}$$

Both these concepts are conformal invariants.

Proposition 8.2.4. *If $f : \Omega \to \Omega'$ is a K-quasiconformal map, then the maximal dilatation satisfies*

$$K(\Omega) \;=\; \sup_{p\in\Omega} K(p). \tag{8.2.5}$$

Proof: Clearly, $K_f(p) \leq K_f(Q)$. To prove the converse, it is enough to consider the case of a square Ω. We want to show that there is a point $p \in \Omega$ such that $K_f(p) \geq K_f(Q)$. Now Ω is the union of four disjoint subsquares and some line segments. In view of Theorem 8.2.3, at least one of these subsquares Ω_1 has $K_f(\Omega_1) = K_f(\Omega)$. Continuing, we get a nested sequence of squares Ω_n whose sides decrease by $1/2$ at each stage. The intersection $\bigcap \Omega_n$ is a single point p with $K_f(p) = K_f(\Omega)$. □

One more analogy of quasiconformal maps with conformal maps concerns convergence.

Theorem 8.2.5. *If $\{f_n\}$ is a sequence of K-quasiconformal maps from Ω to Ω' that converges uniformly on each compact subset of Ω, then the limit f is a homeomorphism, and is also K-quasiconformal.*

Proof: Given a quadrilateral Q contained in a compact subset of Ω, we may construct a sequence of quadrilaterals Q_n that converges uniformly to Q in the sense used in Theorem 8.1.5. For large enough k, $f_k(Q_n)$ is contained in $f(Q)$. Passing to a subsequence of $\{f_n\}$ and renumbering, we can obtain $f_n(Q_n) \subset f(Q)$. Since f is uniformly continuous on Q, the $f_n(Q_n)$ converge uniformly to $f(Q)$. Therefore Theorem 8.1.5 gives $m(f_n(Q_n)) \to m(f(Q))$, and the inequalities

$$\frac{1}{K}\, m(f_n(Q_n)) \;\leq\; m(Q_n) \;\leq\; K\, m(f_n(Q_n))$$

carry over to f. □

8.3 Regular quasiconformal maps

We say that a map $f : \Omega \to \Omega'$ is *regular* if it is an orientation-preserving homeomorphism that is of class C^1, i.e. the coordinate functions have continuous first partial derivatives. It will be useful to put this in terms of the complex derivatives of (1.2.3) and (1.2.4). Suppose that

$$f(x+iy) \;=\; u(x,y) + iv(x,y)$$

where u and v are real-valued and have continuous first partial derivatives. Let

$$p \;=\; \frac{\partial f}{\partial z} \;=\; \frac{1}{2}\left(\frac{\partial f}{\partial x} - i\frac{\partial f}{\partial y}\right), \qquad q \;=\; \frac{\partial f}{\partial \bar{z}} \;=\; \frac{1}{2}\left(\frac{\partial f}{\partial x} + i\frac{\partial f}{\partial y}\right).$$

Then some calculation shows that

$$\begin{aligned} u_x &= \operatorname{Re}(p+q), & u_y &= \operatorname{Im}(q-p), \\ v_x &= \operatorname{Im}(p+q), & v_y &= \operatorname{Re}(p-q). \end{aligned} \tag{8.3.1}$$

Some further calculation shows that the Jacobian of the map f,

$$J_f = \begin{vmatrix} u_x & u_y \\ v_x & v_y \end{vmatrix} = |p|^2 - |q|^2. \tag{8.3.2}$$

Since we are assuming that f preserves orientation, we have that $|q| < |p|$ at each point of Ω. Let us consider a uniform condition

$$|q| = \left|\frac{\partial f}{\partial \bar{z}}\right| \le k\,|p| = k\left|\frac{\partial f}{\partial z}\right| \quad \text{in } \Omega, \tag{8.3.3}$$

where k is a constant, $0 \le k < 1$. We want to relate this condition to quasiconformality.

For this purpose we consider directional derivatives

$$\partial_\alpha f = \cos\alpha \frac{\partial f}{\partial x} + i \sin\alpha \frac{\partial f}{\partial y} = e^{-i\alpha} p + e^{i\alpha} q. \tag{8.3.4}$$

The *dilatation quotient* of the map f at a point z of Ω is defined to be

$$D_f(z) = \frac{\sup_\alpha |\partial_\alpha f(z)|}{\inf_\alpha |\partial_\alpha f(z)|}. \tag{8.3.5}$$

Let us compute the numerator and denominator, assuming $p(z)q(z) \ne 0$. Let θ be the argument of q/p. Then (8.3.4) shows that

$$|\partial_\alpha f(z)| = |p + e^{2i\alpha} q|.$$

It follows easily that

$$D_f(z) = \frac{|p| + |q|}{|p| - |q|} = \frac{1 + |q|/|p|}{1 - |q|/|p|}. \tag{8.3.6}$$

Theorem 8.3.1. *Suppose that $f : \Omega \to \Omega'$ is a regular map.. Then f is K-quasiconformal if and only if the dilatation coefficient $D(z)$ is bounded. If so, then*

$$K_f(\Omega) = \sup_z D_f(z). \tag{8.3.7}$$

Proof: We show first that $K \le \sup D(z)$. The dilatation quotient is a conformal invariant, so we may consider a model rectangle R with side lengths m, 1 mapped by g onto a model rectangle R' with side lengths m', 1, where g has dilatation quotient $\le K$ at each point. By (8.3.2),

$$J_g \;=\; (|p|+|q|)(|p|-|q) \;\geq\; \frac{1}{K}(|p|+|q|)^2 \;\geq\; \frac{1}{K}|g_x|^2.$$

Therefore for the area of R' we have

$$m' \;=\; \iint_R J_g\,dx\,dy \;\geq\; \frac{1}{K}\int_0^1 \left(\int_0^m |g_x|^2\,dx\right) dy. \tag{8.3.8}$$

On the other hand

$$m' \;\leq\; \int_0^m |g_x(x+iy)|\,dx$$

since the integral is the length of a curve joining the b sides of R', so

$$m' \;\leq\; \frac{1}{m'}\left(\int_0^m |g_x(x+iy)|\,dx\right)^2 \;\leq\; \int_0^m |g_x(x+iy)|^2\,dx. \tag{8.3.9}$$

The inequalities (8.3.8) and (8.3.9) imply that $m' \leq Km$.

Conversely, by Proposition 8.2.4, to show that $\sup D_f(z) \leq K$, it is enough to show that $D_f(z) \leq K_f(z)$. Given z, let d_α and d_β be the directional derivatives where

$$d_\alpha f \;=\; |p|+|q|, \qquad d_\beta f \;=\; |p|-|q|.$$

A calculation shows that these directions are at right angles. As before, we may reduce to the case $z=0$, and $g(0)=0$, where g is the transplanted map, and we may take $d_\alpha = d_x$, $d_\beta \;=\; d_y$. Then for sufficiently small $\varepsilon > 0$, εR will lie in the domain of the transplanted map g. For any $z = x+iy \in \varepsilon R$,

$$g(x+iy) = g_x(0)x + g_y(0)(iy) + O(\varepsilon^2).$$

By assumption, $g_x(0)$ and $g_y(0)$ are positive, so for some constant c the strip

$$c\,\varepsilon^2 g_x(0) \leq x \leq g_x(0)\varepsilon - c\,\varepsilon^2$$

separates the sides of $g(\varepsilon R)$, while

$$-c\,\varepsilon^2 \leq \operatorname{Im} g(\varepsilon R) \leq g_y(0)\varepsilon + c\,\varepsilon^2.$$

It follows from Proposition 8.1.4 that

$$m(g(\varepsilon R)) \;\geq\; \frac{g_x(0)\varepsilon - 2c\varepsilon^2}{g_y(0)\varepsilon + 2c\varepsilon^2} \;=\; \frac{|p|+|q|-2c\varepsilon}{|p|-|q|+2c\varepsilon} \;=\; D_g(0) + O(\varepsilon)$$

Therefore $D_f(z) \leq K_f(z)$. □

Corollary 8.3.2. *If $f : \Omega \to \Omega'$ is a regular map with bounded maximal dilatation $K(\Omega)$, then*

$$K(\Omega) = \frac{1+k}{1-k}, \qquad k = \sup_{z\in\Omega} \frac{f_{\bar{z}}(z)}{f_z(z)}. \tag{8.3.10}$$

The following result could be deduced from Theorem 8.3.1 and earlier results, but a direct proof is simpler.

Proposition 8.3.3. *The dilatation quotient $D_f(z)$ of a regular map f is the same as the dilatation quotient of the inverse map at $f(z)$.*

Proof: Denote $f(z)$ by ζ. The identity $d\zeta = pdz + qd\bar{z}$ and its complex conjugate can be solved for dz to obtain

$$dz = \frac{\bar{p}d\zeta - qd\bar{\zeta}}{|p|^2 - |q|^2}, \tag{8.3.11}$$

which implies the stated result. □

Remark. The inequality $D_f(z) \le K$ holds if we simply assume that the K-quasiconformal map f is differentiable at the point z. The proof is the same as in the proof of Theorem 8.3.1, with $o(\varepsilon)$ in place of ε^2 in the remainder terms.

8.4 Ring domains

By a *ring domain* we mean a bounded domain B, the complement of whose closure consists of one (non-empty) bounded component and one unbounded component. For a suitable choice of r_1, r_2 there is a conformal map of B onto an annulus

$$A = A(r_1, r_2) = \{z : 0 < r_1 < |z| < r_2 < \infty\}; \tag{8.4.1}$$

see Exercises 12 – 14. The annulus A is called a *canonical image* of B. The annulus A itself is, up to the segment $[r_1, r_2]$, the bijective image under the exponential map of the rectangle

$$R = R(r_1, r_2) = \{(x + i\theta) : 0 < \theta < 2\pi,\ \log r_1 < x < \log r_2\}. \tag{8.4.2}$$

Note that we are considering the segments with fixed r to be joining the a sides of this rectangle, so the identity in Proposition 8.4.1 below takes the opposite form from Proposition 8.1.2.

The rectangle has module $m(R) = \log(r_2/r_1)/2\pi$. We follow custom and normalize by setting

$$m(B) = \log \frac{r_2}{r_1}. \tag{8.4.3}$$

Note that if $A = A(r_1, r_2)$ is the canonical image of B, then

$$m(B) = \int_{r_1}^{r_2} \frac{rdr}{r^2} = \frac{1}{2\pi} \iint_A \frac{dx\,dy}{|z|^2}. \tag{8.4.4}$$

We have the following analogue of Proposition 8.1.2:

Proposition 8.4.1. *The module of a ring domain B is*

$$m(B) \;=\; \frac{2}{\pi}\inf_{\rho}\frac{A(\rho)}{L(\rho)^2}, \tag{8.4.5}$$

where ρ runs through the non-zero functions that are non-negative on B and have finite integral

$$A(\rho) \;=\; \iint_B \rho(x+iy)^2\,dx\,dy,$$

and $L(\rho)$ is the infimum of

$$L_\gamma(\rho) \;=\; \int_\gamma \rho(z)|dz|$$

over closed rectifiable curves γ in B that separate the two components of the complement of $\overline{B}$.

Proof: It follows from the remarks above that there is a conformal map g from a rectangle (8.4.2) onto B, minus an analytic arc.

If γ is a curve as above, then $\widetilde{\gamma} = \gamma \circ g$ is a curve that joins the b sides of R, and conversely. With the convention that for $\zeta = u + iv \in R$ we have $g(\zeta) = z = x + iy \in B$, then $|dz| = |g'(\zeta)||d\zeta|$ and $dx\,dy = |g'(\zeta)|^2 du\,dv$. Thus, if we set $\widetilde{\rho}(\zeta) = \rho(f(\zeta))|g'(\zeta)|$, then

$$\int_{\widetilde{g}}\widetilde{\rho}(\zeta)|d\zeta| \;=\; \int_\gamma \rho(z)|dz|; \qquad \iint_R \widetilde{\rho}(\zeta)^2 du\,dv \;=\; \iint_B \rho(z)^2\,dx\,dy.$$

Therefore (8.4.5) follows from (8.1.4) applied to R. □

Remark. In the notation of the previous proof, equality holds in (8.1.4) only if $\widetilde{\rho}$ is constant. Thus, for equality we need $\rho(z) \;=\; 1/|g'(z)|$. If $f : \Omega \to A(r_1, r_2)$ is the canonical map, then g is the inverse of $\log f$, so $1/|g'| = |f'|/|f|$.

The next result gives another upper bound for the module of a ring domain.

Proposition 8.4.2. *If B is a ring domain that encloses the origin, then*

$$m(B) \;\le\; \frac{1}{2\pi}\iint_B \frac{dx\,dy}{|z|^2}. \tag{8.4.6}$$

Equality holds if and only if B is an annulus centered at the origin.

Proof: We use the notation of the proof of Proposition 8.4.1. Thus, $g : R \to B$ and

$$\iint_B \frac{dx\,dy}{|z|^2} \;=\; \frac{1}{2\pi}\iint_R \frac{|g'|^2}{|g|^2}du\,dv. \tag{8.4.7}$$

Now for fixed u, the integral of g'/g is the change in the argument of $g(u, \cdot)$, so

$$\int_0^{2\pi} \frac{g'(u,v)}{g(u,v)}\, dv \;=\; 2\pi. \tag{8.4.8}$$

Therefore

$$\iint_R \frac{g'}{g}\, du\, dv \;=\; 2\pi \log \frac{r_2}{r_1} \;=\; 2\pi\, m(B). \tag{8.4.9}$$

But also, by Cauchy–Schwarz,

$$\left|\iint_R \frac{g'}{g}\, du\, dv\right|^2 \;\le\; \iint_R \frac{|g'|^2}{|g|^2} du\, dv \cdot 2\pi\, m(B).. \tag{8.4.10}$$

Combining (8.4.7), (8.4.9), and (8.4.10), we obtain the inequality (8.4.6).

Suppose equality holds in (8.4.6). Then equality holds in (8.4.10). This implies that g'/g is constant. In view of (8.4.8), the constant is 1. Therefore $g(\zeta) = \zeta + c$. Recall that g is the inverse of $\log f$, where $f : B \to \mathbb{R}$. Therefore

$$f(z) \;=\; \exp(z - c) \;=\; e^{-c} e^z,$$

so B is the image under dilation by e^{-c} of the annulus R. □

Corollary 8.4.3. *If B is a ring domain that encloses 0 and $\{B_n\}$ is a sequence of disjoint ring domains contained in B that enclose 0, then*

$$\sum_n m(B_n) \le m(B).$$

Moreover, if B is an annulus centered at the origin, then equality holds only if the B_n are also annuli centered at the origin, and their union is dense in B.

In particular, the module of ring domains that enclose the origin is strictly increasing with respect to set inclusion.

Let us look at the connection with quasiconformal mapping.

Proposition 8.4.4. *If f is a homeomorphism from a domain Ω onto a domain Ω', then f is K-quasiconformal if and only if for each ring domain $B \subset \Omega$,*

$$\frac{1}{K} m(B) \;\le\; m(B') \;\le\; K\, m(B), \qquad B' = f(B). \tag{8.4.11}$$

Proof: Suppose that f is K-quasiconformal and $B \subset \Omega$ is a ring domain with image B'. We may pass to the canonical image and assume that B is an annulus $A = A(r_1, r_2)$. Removing the segment (r_1, r_2) from A leaves two quadrilaterals A_1, A_2. Taking the logarithm gives rectangles with modulus $2\pi/m(B)$. The images $B'_j = f(A_j)$ are disjoint quadrilaterals in B', so

$$\frac{2\pi}{m(B')} = m(B'_1) + m(B'_2) \leq K[m(A_1) + m(A_2)] = K\frac{2\pi}{m(B)}$$

so $m(B)/K \leq m(B')$. The other inequality in (8.4.11) follows, since f^{-1} is also K-quasiconformal.

The converse may be proved by reverse engineering: given a quadrilateral Q, divide Q (minus an analatic arc) into two quadrilateral Q_j with half the module, and use the canonical images to map conformally to a ring domain. Then use (8.4.11) to prove that

$$\frac{1}{K}m(Q) \leq m(Q') \leq K\,m(Q). \qquad \square \qquad (8.4.12)$$

8.5 Extremal ring domains

We consider now a question studied by Grötzsch [94]: what is the maximum module of a ring domain that separates a Jordan curve γ from two distinct points that lie on one component of the complement of γ. By a conformal map of the component that contains the two points, we may assume that γ is the boundary of the unit disk $\mathbb{D}$ and that the two points are $0, r$, $0 < r < 1$. Grötzsch showed that the maximal modulus is attained by *Grötzsch's extremal domain*: the complement in the unit disk of the segment $[0, r]$.

As we shall see, questions of this type are important for the understanding of the possible behaviour of a quasiconformal map. Questions about the extend to which a K-quasiconformal map can change the distance between two points can be attacked by looking at separation problems of this type, and using the maximum modulus of a separating ring in order to calculate an upper bound to the distortion or to a Hölder continuity norm.

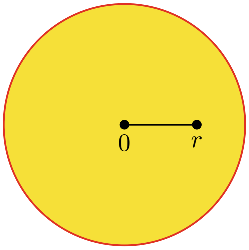

Fig. 8.3 Grötzsh's extremal domain.

In order to consider the domain in Figure 8.3 as a ring domain, we need to extend the definition to allow the complement of the domain to consist of sets that may have empty interior. Specifically, we allow any domain B_∞ that is the union of an increasing sequence of ring domains $\{B_n\}$, and take $m(B_\infty) = \lim_{n\to\infty} m(B_n)$.

The following result is known as *Grötzsch's module theorem*:

Theorem 8.5.1. *The domain B indicated in Figure 8.3 has the greatest module of any ring domain that separates the points* 0 *and* r, $0 < r < 1$, *from the unit circle* $\partial\mathbb{D}$.

Proof: Consider Q, the upper half of B, as a quadrilateral with vertices $(0, r, 1, -1)$. Given $R > 1$, consider the upper half of the annulus $A(1, R)$ to be a quadrilateral Q_R with vertices $(1, R, -R, -1)$. There is a unique conformal map f of B onto Q_R that takes the ordered triple $(0, r, 1)$ to $(1, R, -R)$. If we choose R so that $m(Q_R) = m(Q)$, then the map takes -1 to itself, and thus is a conformal map from Q to Q_R as quadrilaterals. Extend f to $\mathbb{D}$ by reflection across the real axis. This conformal image has the same modulus, so $R = e^{\mu}$, $\mu = \mu(B)$. Now suppose that γ is any curve that separate 0 and r from the unit circle. Then the portions of γ in the upper and lower halves of the disk must meet both intervals $(-1, 0)$ and $(r, 1)$, so the length of $f(\gamma)$ is at least 2π. The result follows from Proposition 8.4.1 and the remark that follows it. □

The module of Grötzsch's domain $B = B(r)$ is commonly denoted by $\mu(r)$.

We turn next to three similar problems, in the formulation given by Ahlfors [5]. Consider a ring domain A in the plane whose complement consists of a bounded region C_1 and an unbounded region C_2. What is the maximum modulus of A in the three cases?

I. (Grötzsch) $C_1 = \overline{\mathbb{D}}$, $C_2 = \{R\}$, $R > 1$.

In this case inversion $z \to 1/z$ sends the problem into the one in Theorem 8.5.1 with $r = 1/R$ and extremal domain

$$B_I = \{z : |z| > 1,\ z \notin [R, \infty)\}.$$

Therefore the module $m_I(R) = \mu(1/R)$.

II. (Teichmüller) C_1 contains -1 and 0; C_2 contains a point $P > 0$.

III. (Mori) $C_1 \cap \overline{\mathbb{D}}$ contains two points z_1, z_2 with $2 > |z_1 - z_2| \geq \lambda > 0$, C_2 contains the origin.

In case III there is an automorphism of the closed disk that moves $\{z_1, z_2\}$ to a pair of points $\{w, \bar{w}\}$ with $\operatorname{Re} w < 0$ and $|w - \bar{w}| = \lambda$; see Exercise 17. Therefore we may assume this configuration.

Extremal domains for these three cases are shown in Figure 8.4.

Theorem 8.5.2. (Teichmüller) *The extremal domain for Question II is B_{II} in Figure 8.4. The module is*

$$m_{II}(P) = 2m_I\left(\sqrt{P+1}\right) = 2\mu\left(\frac{1}{\sqrt{P+1}}\right). \tag{8.5.1}$$

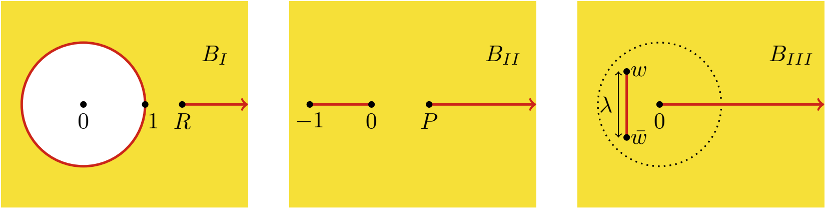

Fig. 8.4 Extremal domains of Grötzsch, Teichmüller and Mori.

Proof: Consider the circle Γ with center -1 and radius $\rho > 1$. Reflection through this circle maps 0 to $\rho^2 - 1$. Therefore if we choose $\rho = \sqrt{P+1}$, Γ separates the plane into two components, each of which is conformally equivalent to Grötzsch's domain in Figure 8.3, with $r = 1/\rho$. Map one of these components conformally onto an annulus centered at the origin. Reflection maps the other component onto an annulus centered at the origin, so altogether B_{II} is mapped conformally to the union of these two annuli, together with the circle that separates them. By Corollary 8.4.3, the module of the union is the sum of the moduli, which is $2\mu(1/\sqrt{P+1})$. This proves the statement about the module $m_{II}(P)$.

To show that B_{II} is extremal, suppose that A is a ring domain that separates $\{0, 1\}$ from $P > 0$. Let f be the conformal map of B_{II} onto the annulus that was constructed in the previous paragraph. As in the proof of Theorem 8.5.1, we conclude that the module of A is at most $m_{II}(P)$. □

Teichmüller considered the general problem of a ring domain that separates two distinct points of the sphere $\mathbb{S}$ from two other distinct points. We may normalize and consider the case of separating $\{0, z_1\}$ from $\{z_2, \infty\}$. The proof depends on two results from Chapter 4.

Theorem 8.5.3. *If the ring domain A separates* 0 *and* z_1 *from* z_2 *and* ∞, *then*

$$m(A) \le 2\mu\left(\frac{|z_1|}{\sqrt{|z_1| + |z_2|}}\right). \tag{8.5.2}$$

Proof: Let C_2 be the component of the complement of A that contains $\{z_2, \infty\}$ and let φ be the conformal map from the complement Ω of C_2 onto $\mathbb{D}$, for which $\varphi(0) = 0$ and $\varphi(z_1) = \zeta_1 > 0$. Then the function

$$g(z) = -\frac{4|z_2|\varphi(z)}{(1-\varphi(z))^2}$$

maps Ω conformally onto the plane, slit along the real axis from $|z_2|$ to ∞. The domain A is mapped onto a ring domain A' that separates $|z_2|$ and ∞ from 0 and $g(z_1)$,

$$g(z_1) = -\frac{4|z_2|\zeta_1}{(1-\zeta_1)^2} < 0.$$

By Theorem 8.5.2,

$$m(A) = m(A') = \mu\left(\sqrt{\frac{-4g(z_1)}{-g(z_1)+|z_2|}}\right). \tag{8.5.3}$$

Now μ is a decreasing function, so we want to show that $-g(z_1) \geq |z_1|$. Applying the Koebe distortion theorem, Theorem 4.1.7, to φ^{-1}, we get

$$|z_1| = |\varphi^{-1}(\zeta_1)| \leq \frac{|(\varphi^{-1})'(0)|\zeta_1}{(1-\zeta_1)^2}. \tag{8.5.4}$$

By Koebe's one-quarter theorem, Theorem 4.1.4,

$$|(\varphi^{-1})'(0)| \leq 4|z_2|. \tag{8.5.5}$$

The inequalities (8.5.4), (8.5.5) give the desired inequality $-g(z_1) \geq |z_1|$. □

We turn now to Mori's problem.

Theorem 8.5.4. (Mori) *The extremal domain for Question III is B_{III} in Figure 8.4. The module is*

$$m_{III}(\lambda) = \frac{1}{2}m_{II}\left(\frac{(2+\sqrt{4-\lambda^2})^2}{\lambda^2}\right) = m_I\left(\frac{\sqrt{4+2\lambda}+\sqrt{4-2\lambda}}{\lambda}\right). \tag{8.5.6}$$

Proof: Let A be a ring domain that separates $\{z_1, z_2\}$ from 0. The idea of the proof is to convert this to case (II).

There are two single-valued branches of the square root, $\pm\sqrt{z}$, defined in the complement of the unbounded component C_2. The pre-image in the ζ plane of C_2 separates the two components of the pre-image of C_1. We choose square roots $\zeta_1 = \sqrt{z_1}$ and $\zeta_2 = \sqrt{z_2}$. Let φ be the linear fractional transformation

$$\varphi(\zeta) = \frac{\zeta+\zeta_1}{\zeta-\zeta_1}\cdot\frac{\zeta_1+\zeta_2}{\zeta_1-\zeta_2},$$

and let $u = (\zeta_1+\zeta_2)/(\zeta_2-\zeta_1)$. Then

$$\varphi(-\zeta_1) = 0, \qquad \varphi(-\zeta_2) = -1, \qquad \varphi(\zeta_1) = \infty, \qquad \varphi(\zeta_2) = -u^2.$$

A simple calculation shows that u is imaginary, so $-u^2 > 0$. Also

$$u = \frac{(\zeta_1+\zeta_2)^2}{\zeta_1^2-\zeta_2^2} = \frac{z_1+z_2+2\zeta_1\zeta_2}{z_1-z_2}. \tag{8.5.7}$$

Since

$$|z_2+z_1|^2 = 2(|z_2|^2+|z_1|^2) - |z_2-z_1|^2 \leq 4-\lambda^2, \tag{8.5.8}$$

it follows from (8.5.7) that

$$|u| \;\le\; \frac{2+\sqrt{4-\lambda^2}}{\lambda}.$$

Therefore the ζ plane corresponds to case II, with

$$P \;=\; -u^2 \;\le\; \frac{\left(2+\sqrt{4-\lambda^2}\right)^2}{\lambda^2}. \tag{8.5.9}$$

Equality holds in (8.5.8) if and only if $|z_1| = |z_2| = 1$ and $|z_2 - z_1| = \lambda$. Note that in this case the $\pm\zeta_j$ all lie on the unit circle, so their images under φ lie on $\mathbb{R}$. Therefore φ maps the circle to $\mathbb{R}$. For equality to hold in (8.5.9) we need

$$\begin{aligned} 2 + |z_1 + z_2| = 2 + \sqrt{4-\lambda^2} \;&=\; |(\zeta_1 + \zeta_2)^2| \\ &= |z_1| + |z_2| + \zeta_1\overline{\zeta_2} + \overline{\zeta}_1\zeta_2. \end{aligned}$$

This is only possible if $z_1 + z_2$ and the term $\overline{\zeta}_1\zeta_2$ and its conjugate all lie on the same line, which must be $\mathbb{R}$. This, in turn, implies that $z_1 = \bar{z}_2$ and that each of the other terms is ± 1, according to the sign of $z_1 + z_2$. If $A = B_{III}$ and we take $\zeta_1 = -\bar{z}_2$, then all these conditions are fulfilled, equality holds in (8.5.9), and the image of A is B_{II}. Therefore m_{III} is extremal for Mori's problem and

$$\begin{aligned} m_{III}(\lambda) &= \frac{1}{2} m_{II}(P); \\ P &= \frac{(2+\sqrt{4-\lambda^2})^2}{\lambda^2} \;=\; \frac{(\sqrt{4+2\lambda}+\sqrt{4-2\lambda})^2}{\lambda^2} - 1. \end{aligned}$$

In view of (8.5.1), this proves (8.5.6). □

One standard notation for the modules of these domains is used in Künzi [127]. With our normalization of ring modules, it is

$$m(B_I) = \log \Phi(R); \tag{8.5.10}$$

$$m(B_{II}) = \log \Psi(P); \tag{8.5.11}$$

$$m(B_{III}) = \log X(\lambda). \tag{8.5.12}$$

The module calculations above can be translated into relations for these functions:

$$\Psi(P) \;=\; [\Phi(\sqrt{P+1})]^2; \qquad [\Phi(R)]^2 \;=\; \Psi(R^2 - 1) \tag{8.5.13}$$

and

$$X(\lambda) \;=\; \Psi\left(\frac{2+\sqrt{4-\lambda^2}}{\lambda}\right); \qquad \Psi(P) \;=\; X\left(\frac{4P}{\sqrt{P+1}}\right). \tag{8.5.14}$$

Another such relation is obtained by noting that

$$w(z) = \frac{1}{4}\left(z+\frac{1}{z}\right)-\frac{1}{2} = \frac{1}{4}\left(\sqrt{z}-\frac{1}{\sqrt{z}}\right)^2$$

is a conformal map of the complement of the Grötzsch domain B_I onto the Teichmüller domain B_{II} with $P = \frac{1}{4}(\sqrt{R}-1/\sqrt{R})^2$. Therefore

$$\Phi(R) = \Psi\left(\frac{(\sqrt{R}-1/\sqrt{R})^2}{4}\right). \tag{8.5.15}$$

Together with (8.5.13), (8.5.15) implies

$$\Phi(R) = \left[\Phi\left(\frac{\sqrt{R}}{2}+\frac{1}{2\sqrt{R}}\right)\right]^2. \tag{8.5.16}$$

This last identity can be converted to a functional equation for Grötzsch's module function μ:

$$\mu(r) = 2\mu\left(\frac{2\sqrt{r}}{1+r}\right). \tag{8.5.17}$$

This can be written in an equivalent form by solving $r_1 = 2\sqrt{r}/(1+r)$ for r:

$$\mu(r) = \frac{1}{2}\,\mu\left(\frac{(1-\sqrt{1-r^2})^2}{r^2}\right). \tag{8.5.18}$$

Estimates of $\mu(r)$ are important.

Proposition 8.5.5. *For* $0 < r < 1$,

$$\log\frac{(1+\sqrt{1-r^2})^2}{r} < \mu(r) < \log\frac{4}{r}. \tag{8.5.19}$$

Proof: The function $\varphi(z) = (k-z)/(kz-1)$ is an automorphism of $\mathbb{D}$ that takes $[0,r]$ to $[-k,k]$ if $k = (1-\sqrt{1-r^2})/r$. Then $k^{-1} = (1+\sqrt{1-r^2})/r$, so $\psi(z) = k^{-1}\varphi(z)$ maps B_r conformally onto B', the complement of $[-1,1]$ in the disk $D_R(0)$, where

$$R = k^{-1} = \frac{1+\sqrt{1-r^2}}{r}.$$

The function $\chi(z) = (z+z^{-1})/2$ maps the annulus $A = A(1,\rho)$ conformally onto the ellipse E_ρ with semi-axes $(\rho \pm \rho^{-1})/2$, slit along $[-1,1]$; see Exercise 19. This ellipse contains B' if $\rho - \rho^{-1} > 2R$, which is true if $\rho = 4/r$. The ellipse is contained in B' if $\rho + \rho^{-1} \le 2R$, and since $1/(2R-r) < r$, this is true if

$$\rho \;=\; 2R - r \;=\; \frac{(1+\sqrt{1-r^2})^2}{r}.$$

Since $\mu(r) = m(B')$, these considerations give us (8.5.19). □

Corollary 8.5.6. *The module μ satisfies*

$$\mu(r) \;\sim\; \log\frac{4}{r} \qquad as\;\; r \to 0. \tag{8.5.20}$$

We conclude this section with an exact formula for $\mu(r)$.

Proposition 8.5.7. *For $0 \le k \le 1$, let*

$$K(k) \;=\; \int_0^1 \frac{d\zeta}{(1-\zeta^2)(1-k^2\zeta^2)}. \tag{8.5.21}$$

Then

$$\mu(r) \;=\; \frac{\pi}{2}\frac{K(\sqrt{1-r^2})}{K(r)}. \tag{8.5.22}$$

Proof: As in the proof of Theorem 8.5.1, we begin with the map f from Q, the upper half of $\mathbb{D}$, to the upper half of the annulus $A(1, e^{\mu})$, where $\mu = \mu(r)$, that takes $(0, r, 1, -1)$ to $(1, e^{\mu}, -e^{-\mu})$. Let $g = e^{-\mu} f$, so g maps Q to the annulus $A(e^{-\mu}, 1)$. Then g can be continued across the upper boundary of Q, giving a map of $\mathbb{H}$ onto the annulus $A = A(e^{-\mu}, e^{\mu})$; see Figure 8.5.

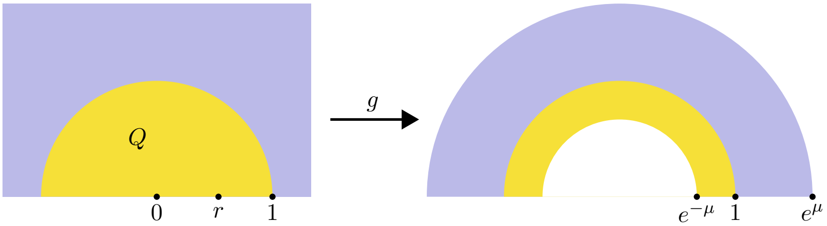

Fig. 8.5 Mapping the upper half plane to $A(e^{-\mu}, e^{\mu})$.

Thus, A is the image of $\mathbb{H}$ considered as a quadrilateral $R(0, r, 1/r, \infty)$. The image A has module $2\mu(r)/\pi$, so

$$\mu(r) \;=\; \frac{\pi}{2} m(Q_1). \tag{8.5.23}$$

For $z \in \mathbb{H}$, let

$$G(z) \;=\; \int_{-\infty}^{z} \frac{dt}{\sqrt{-t(1-rt)(1-r^{-1}t)}}.$$

As in the discussion in Exercise 3, G maps Q_1 to the rectangle

$$R(0, G(r), G(r) + iG(1/r), iG(1/r)). \tag{8.5.24}$$

The function $\varphi(z) = (1 + r^{-1})z(z+1)^{-1}$ is an automorphism of $\mathbb{H}$ that maps Q_1 to $Q_2 = R(0, 1, 1/r, \infty)$. As in Exercise 3 again, for $0 < k < 1$, the function

$$F(z) = \int_0^z \frac{d\xi}{\sqrt{(1-\xi^2)(1-k^2\xi^2)}}$$

maps Q_2 to the rectangle

$$\begin{aligned} Q_3 &= R(0, F(1), F(1) + iF(1/k); \\ &= (0, K, K + iK', iK'), \end{aligned} \tag{8.5.25}$$

where K is defined by (8.5.21) and

$$K' = \int_1^{1/k} \frac{d\xi}{\sqrt{(\xi^2-1)(1-k^2\xi^2)}}.$$

The change of variables $x \to \sqrt{1-k^2x^2}/\sqrt{1-k^2}$ shows that

$$K'(k) = K(k'), \qquad k' = \sqrt{1-k^2}. \tag{8.5.26}$$

Therefore the rectangles (8.5.24) and (8.5.25) will coincide if we take $k' = r$, so $k = \sqrt{1-r^2}$. Then

$$m(Q_1) = m(Q_2) = m(Q_3) = \frac{K(\sqrt{1-r^2})}{K(r)},$$

so (8.5.23) gives (8.5.22). □

Corollary 8.5.8. *For $0 < r \leq 1$, $\mu(r)$ is a continuous decreasing function of r.*

Combining (8.5.21) with (8.5.18), taking $r = 1/\sqrt{2}$, we find that

$$\mu\left(\frac{1}{\sqrt{2}}\right) = \frac{\pi}{2}. \tag{8.5.27}$$

The formula (8.5.22) gives an alternative way to derive the functional equation (8.5.17) and asymptotics like (8.5.19). For the functional equation, see Exercise 20; for asymptotics, see Exercises 21 – 23.

8.6 Distortion properties and Hölder continuity

Our first application of the results in Section 8.5 is to *circular distortion*.

Theorem 8.6.1. *Suppose that f is a K-quasiconformal homeomorphism of $\mathbb{C}$, and $f(0) = 0$. Then there is a constant $c(K)$ such that for each $r > 0$,*

$$\frac{\sup_\theta |f(re^{i\theta})|}{\inf_\theta |f(re^{i\theta})|} \le c(K). \tag{8.6.1}$$

Proof: For a given r, let z_1 and z_2 be the points on the circle of radius r centered at 0 at which f attains its minimum and maximum values, respectively. Let A' be the annulus $A(|z_1|, |z_2|)$, and let $A = f^{-1}(A')$. Then the annulus A separates the set $\{0, z_1\}$ from $\{z_2, \infty\}$. Theorem 8.5.2, the monotonicity of the function μ, and (8.5.27) imply

$$m(A) \le 2\mu\left(\frac{|z_1 - z_2|}{\sqrt{2(|z_1| + |z_2|)}}\right) \le 2\mu\left(\frac{1}{\sqrt{2}}\right) = \pi.$$

Therefore $m(B') \le K\pi$, and we may take $c(K) = e^{K\pi}$ in (8.6.1). □

In the remainder of this section we investigate properties of K-quasiconformal maps from $\mathbb{D}$ to $\mathbb{D}$. The culminating result is that such a map has a strong uniform continuity property.

Proposition 8.6.2. *Suppose that f is a K-quasiconformal map from $\mathbb{D}$ into itself, such that $f(0) = 0$. Then for $z \in \mathbb{D}$,*

$$|f(z)| \le \varphi_K(|z|), \quad \textit{where} \ \ \varphi_K(r) = \mu^{-1}(\mu(r)/K). \tag{8.6.2}$$

Proof: Slit the disk along the segment from 0 to z. The slit disk has module $\mu(|z|)$, and its image under f has module $\le \mu(|f(z)|)$. Therefore

$$\mu(z) \le K\,\mu(|f(z)|),$$

which is (8.6.2). □

It can be shown that equality in (8.6.2) can be attained if $f(\mathbb{D}) = \mathbb{D}$; see Exercise 24.

The next step is to estimate the distortion function φ_K.

Proposition 8.6.3. *The distortion function φ_K satisfies*

$$\varphi_K(r) \le 4^{1-1/K} \cdot r^{1/K}. \tag{8.6.3}$$

Proof: Suppose $0 < r < r' < 1$. The Grötzsch domain B_r contains the annulus $A = A(r/r', 1)$ and the ring domain

$$R = \{z : |z| < r/r'\} \setminus [0, r].$$

Since dilation by r'/r maps R onto $B_{r'}$, it follows that

$$m(A) + m(R) = \log\left(\frac{r}{r}\right) + \mu(r') \le \mu(r),$$

or

$$\log(1/r) - \mu(r) \le \log(1/r') - \mu(r').$$

But (8.5.19) shows that $\log(4/r) - \mu(r)$ is positive for $0 < r < 1$, so

$$\frac{1}{K}\left[\log\left(\frac{4}{r}\right) - \mu(r)\right] \le \log\left(\frac{4}{r'}\right) - \mu(r'). \tag{8.6.4}$$

Let $r' = \varphi_K(r)$. Then $\mu(r)/K = \mu(r')$, so (8.6.4) becomes (8.6.3). □

Proposition 8.6.4. *If $f : \mathbb{D} \to \mathbb{D}$ is K-quasiconformal, then for any $z_1, z_2 \in \mathbb{D}$,*

$$\left|\frac{f(z_2) - f(z_1)}{1 - \overline{f(z_1)}f(z_2)}\right| \le \varphi_K\left(\left|\frac{z_2 - z_1}{1 - \bar{z}_1 z_2}\right|\right). \tag{8.6.5}$$

Proof: The disk automorphisms

$$z \to \frac{z - z_1}{1 - \bar{z}_1 z}, \qquad w \to \frac{w - f(z_1)}{1 - \overline{f(z_1)}w}$$

map z_1 to 0 and $f(z_1)$ to zero, respectively, so the result follows from Proposition 8.6.3. □

The inequality (8.6.5) can be rephrased in terms of the hyperbolic metric on $\mathbb{D}$, (2.2.10):

$$\rho(z_1, z_2) = \frac{1}{2}\log\frac{|1 - \bar{z}_1\zeta_2| + |z_1 z_2|}{|1 - \bar{z}_1\zeta_2| - |z_1 z_2|} = \tanh^{-1}\frac{|z_1 - z_2|}{|\bar{z}_1 - z_2|}.$$

Thus, we may restate Proposition 8.6.4:

Proposition 8.6.5. *If $f : \mathbb{D} \to \mathbb{D}$ is K-quasiconformal, then for any $z_1, z_2 \in \mathbb{D}$, the hyperbolic distances satisfy*

$$\rho(f(z_1), f(z_2)) \le C(K)\,\rho(z_1, z_2), \tag{8.6.6}$$

for a suitable constant $C(K)$.

Combining Propositions 8.6.3 and 8.6.4 with a local change of scale, we get an important regularity result for quasiconformal maps.

Theorem 8.6.6. *If $f : \Omega \to \Omega'$ is K-quasiconformal, then f is locally Hölder contnous with exponent $1/K$, i.e. for each z_0 in Ω there are constants $\delta > 0$ and C such that if $|z_j - z_0| < \delta$, $j = 1, 2$, then*

$$|f(z_1) - f(z_2)| \ \le \ C\,|z_1 - z_2|^{1/K}. \tag{8.6.7}$$

The next step is a more precise but more specialized result on Hölder continuity.

Theorem 8.6.7. *Suppose that f is a quasiconformal map of $\mathbb{D}$ onto itself with maximal dilatation K, and $f(0) = 0$. Then f satisfies a Hölder continuity condition: if $z_1, z_2 \in \mathbb{D}$, then*

$$|f(z_1) - f(z_2)| \le 16\,|z_1 - z_2|^{1/K}. \tag{8.6.8}$$

Proof: The inequality (8.6.8) is automatic if $|z_1 - z_2| \ge 1/8$, so we assume that $|z_1 - z_2| < 1/8$. Suppose first that $|z_1 + z_2| \le 1$. Then

$$4|z_1 z_2| \ = \ |(z_1 + z_2)^2 + (z_1 - z_2)^2| \ \le \ |z_1 + z_1|^2 + |z_1 - z_2|^2 \ < \ 2,$$

so $|z_1 z_2| < 1/2|$, and $|1 - \bar{z}_1 z_2| > 1/2$. Also $|1 - \overline{f(z_1)}f(z_2)| \le 2$. The estimates (8.6.5) and (8.6.3) imply that

$$|f(z_1) - f(z_2)| \ \le 2\varphi_K(2|z_1 - z_2|) \ \le \ 8(2|z_1 - z_2|)^{1/K}.$$

Now suppose that $|z_1 + z_2| > 1$, and, for the moment, assume that f extends to the boundary of $\mathbb{D}$. Then we may extend f to $\mathbb{C}$ by reflection across the unit circle. The annulus

$$A \ = \ \left\{ z \ : \ \left|\frac{z_1 - z_2}{2}\right| < \left| z - \frac{z_1 + z + z_2}{2}\right| < \frac{1}{2}\right\}$$

has module $m(A) = \log(1/|z_1 - z_2|)$. Since $|z_1 + z_2| > 1$, both 0 and ∞ are in the unbounded component of the complement of $\overline{A}$. The points z_j belong to the closure $\overline{A}$, so $\widehat{A} = f(A)$ separates $f(z_1)$ and $f(z_2)$ from $0 = f(0)$ and $\infty = f(\infty)$. Let $\lambda = |f(z_1) - f(z_2)|$. By Theorem 8.5.4, we have

$$m(\widehat{A}) \ \le \ m_I\left(\frac{\sqrt{4 + 2\lambda} + \sqrt{4 - 2\lambda}}{\lambda}\right) \le m_I\left(\frac{4}{\lambda}\right). \tag{8.6.9}$$

Now $m_I(R) = \mu(1/R)$, so Proposition 8.5.5 implies that the estimate (8.6.9) gives $m(\widehat{A}) \le \log(16/|f(z_1) - f(z_2)|)$. Thus,

$$\log\left(\frac{1}{|z_1 - z_2|})\right) \ = \ m(A) \le m(\widehat{A}) \le \log\frac{16}{|f(z_1) - f(z_2)|},$$

which is (8.6.8).

To complete the proof, we must drop the assumption that the map f continues to the boundary. Let $f_r(z) = f(rz)$, $0 < r < 1$. The image $f_r(\mathbb{D})$ may be mapped conformally onto $\mathbb{D}$ by a unique map g_r that satisfies $g_r(0) = 0$, $g_r'(0) > 0$. The g_r are a family of holomorphic functions to $\mathbb{D}$ whose domains increase to fill $\mathbb{D}$. Some subsequence converges uniformly on compact subsets of $\mathbb{D}$ to an automorphism of $\mathbb{D}$. The conditions on g_r imply that this automorphism is the identity. Therefore the g_r themselves converge to the identity.

It follows from Theorem 8.6.6 that the boundary of $f_r(\mathbb{D})$ is a Jordan curve, so g_r is continuous to the boundary by Theorem 2.6.1. Therefore each K-quasiconformal map $g_r \circ f_r$ is continuous to the boundary and satisfies estimates (2.8.1). The f_r converges pointwise to f; it follows from this and the convergence of g_r that f itself satisfies (2.8.1). □

The Hölder exponent $\alpha = 1/K$ cannot be improved, in general; see Exercise 27.

Remark. The estimate (2.8.1) shows that f is uniformly continuous in $\mathbb{D}$. It follows that it extends continuously to the boundary. In other words, once we have *assumed* that it extends, then we have the means to *prove* that it does. Note also that f^{-1} is also K-quasiconformal and extends to the boundary, so the extensions are bijective from $\overline{\mathbb{D}}$ to itself.

Corollary 8.6.8. *A surjective K-quasiconformal map $f : \mathbb{D} \to \mathbb{D}$ extends to a K-quasiconformal map of $\mathbb{C}$ onto $\mathbb{C}$.*

Proof: Extend f by

$$f(z) = -\frac{1}{\overline{f(-1/\bar{z})}} = r \circ f \circ r(z), \qquad |z| > 1,$$

where r is the composition of the linear fractional transformation $z \to 1/z$ with the orientation-reversing map $z \to \bar{z}$. Therefore f is K-quasiconformal on the exterior region. It agrees with the original f on the unit circle. By Theorem 8.2.3, f is K-quasiconformal on $\mathbb{C}$. □

Corollary 8.6.9. *The family of K-quasiconformal maps f of $\mathbb{D}$ onto $\mathbb{D}$ such that $f(0) = 0$ is a complete normal family.*

8.7 Quasisymmetry and quasi-isometry

We know now that a quasiconformal homeomorphism of the disk extends to the boundary. In this section, we consider the properties of the boundary homeomorphism, and the relation between the boundary map and the map of the interior.

Suppose that $f_{\mathbb{D}}$ is a K-quasiconformal map of $\mathbb{D}$ onto $\mathbb{D}$. Up to composing with a rotation, we may normalize by setting $f(1) = 1$. It is convenient now to compose with the Cayley transform $C : \mathbb{H} \to \mathbb{D}$ and its inverse and examine $f = C^{-1} \circ f_{\mathbb{D}} \circ C$. Then f is a K-quasiconformal map of $\mathbb{H}$ onto $\mathbb{H}$ that extends by continuity to $h : \mathbb{R} \to \mathbb{R}$. Then h is a strictly increasing function from $\mathbb{R}$ onto $\mathbb{R}$. The goal here is to characterize the functions of this type that arise as boundary values of K-quasiconformal maps of $\mathbb{H}$ onto $\mathbb{H}$.

A strictly increasing function h from $\mathbb{R}$ onto $\mathbb{R}$ is said to be *quasisymmetric* if there is a constant M such that for every $x \in \mathbb{R}$ and $t > 0$,

$$\frac{1}{M} \leq \frac{h(x+t) - h(x)}{h(x) - h(x-t)} \leq M. \tag{8.7.1}$$

We shall see that this condition characterizes the boundary values in question. The first half of this characterization is:

Theorem 8.7.1. *If f is a K-quasiconformal map of $\mathbb{H} \to \mathbb{H}$ that fixes ∞, then the boundary value $h : \mathbb{R} \to \mathbb{R}$ is quasisymmetric.*

Proof: Suppose that $x_1 < x_2 < x_3$ are three points on $\mathbb{R}$, with images $\widehat{x}_j = h(x_j)$. Consider the quadrilaterals

$$Q = \mathbb{H}(\infty, x_1, x_2, x_3), \qquad \widehat{Q} = \mathbb{H}(\infty, \widehat{x}_1, \widehat{x}_2, \widehat{x}_3) = f(Q)$$

By assumption on f, $m(Q)/K \leq m(\widehat{Q}) \leq Km(Q)$.

The linear fractional transformation $\varphi(z) = (z - x_1)/(x_2 - x_1)$ maps (x_1, x_2, ∞) to $(0, 1, \infty)$ and takes x_3 to $1/k$, where

$$k = \frac{x_2 - x_1}{x_3 - x_1}.$$

As we know from earlier calculations, the module

$$m\left(\mathbb{H}(0, 1, 1/k, \infty)\right) = \frac{2}{\pi}\,\mu(k).$$

Therefore

$$m(Q) = \frac{2}{\pi}\,\mu(k); \quad m(\widehat{Q}) = \frac{2}{\pi}\,\mu(\widehat{k}), \quad \widehat{k} = \frac{\widehat{x}_2 - \widehat{x}_1}{\widehat{x}_3 - \widehat{x}_1}.$$

We now specialize to the case $x_1 = x - t < x_2 = x < x + t = x_3$, so that $k = 1/2$. Using the inequality (8.5.19), which implies $\log(1/r) < \mu(r) < \log(4/r)$, together with

$$\frac{1}{K}\,m(\widehat{Q}) \leq m(Q) = \frac{2}{\pi}\,\mu(1/2) \leq K\,m(\widehat{Q}),$$

we find that

$$\frac{1}{4}e^{-\mu(1/2)/K} \le \frac{1}{\widehat{k}} = \frac{h(x+t)-h(x-t)}{h(x)-h(x-t)} + 1 \le e^{K\mu(1/2)}. \tag{8.7.2}$$

Since

$$\frac{h(x+t)-h(x)}{h(x)-h(x-t)} = \frac{1}{\widehat{k}} - 1,$$

the estimate (8.7.2) implies an estimate of the form (8.7.1), where M depends only on K. □

We take $H(M)$ to be the set of quasisymmetric functions h that satisfy the normalization conditions

$$h(0) = 0, \qquad h(1) = 1.$$

Beurling and Ahlfors [26] showed that each h in $H(M)$ is the boundary value of a K-quasiconformal map of $\mathbb{H}$ onto $\mathbb{H}$. We follow the (much simpler) argument given in [132] and [131].

Lemma 8.7.2. *The family $H(M)$ is a complete normal family on $\mathbb{R}$.*

Proof: The left-hand inequality in (8.7.1) implies that

$$h(2^{1-n}) - h(2^{-n}) \ge \frac{h(2^{-n})}{M}, \qquad n = 0, 1, 2, \ldots.$$

Therefore

$$h(2^{-n}) \le \left(\frac{M}{M+1}\right)^n. \tag{8.7.3}$$

If $0 \le x < 2^{-n}$, then

$$0 \le h(a+x) - h(a) \le [h(a+1) - h(a)]\left(\frac{M}{M+1}\right)^n.$$

Thus, for $m \le a < m+1$,

$$\begin{aligned} h(a+1) - h(a) &\le M^m[h(a+1-m) - h(a-m)] \\ &\le M^m h(2) \le M^m(M+1). \end{aligned}$$

Thus, (8.7.3) implies equicontinuity. Therefore $H(M)$ is a normal family. If $\{h_n\}$ is a sequence that converges uniformly on each bounded interval, then clearly the limit f satisfies (8.7.1). □

Lemma 8.7.3. *For any bounded interval $[a, b] \subset \mathbb{R}$ and any $\varepsilon > 0$, there is a $\delta > 0$ such that if $h \in H(1+\delta)$, then $x \in [a, b]$ implies $|h(x) - x| < \varepsilon$.*

Proof: Suppose that for some $\varepsilon > 0$ there is a sequence of functions $h_n \in H(1 + 1/n)$ and a sequence of points $x_n \in [a, b]$ such that $|h_n(x_n) - x_n| \ge \varepsilon$. By Lemma 8.7.2, $H(2)$ is a normal family, so there is a subsequence of $\{h_n\}$ that converges uniformly on $[a, b]$. The limit is in $H(1)$, so it is the identity. □

The following important construction is the *Beurling–Ahlfors extension.*

Theorem 8.7.4. *If h belongs to $H(M)$, then there is a K-quasiconformal map of $\mathbb{H}$ onto $\mathbb{H}$ whose extension to the boundary is h, where K depends only on M, and $K \to 1$ as $M \to 1$. The map extends to a quasiconformal map of $\mathbb{C}$ onto $\mathbb{C}$.*

Proof: Define

$$f(x+iy) \;=\; \frac{1}{2}\int_0^1 [h(x+ty)+h(x-ty)]\,dt + \frac{i}{2}\int_0^1 [h(x+ty)-h(x-ty)]\,dt. \tag{8.7.4}$$

Then $f|_{\mathbb{R}} = h$. Note that if h is the identity map on $\mathbb{R}$, the extension is the identity map on $\mathbb{C}$.

For $y \neq 0$, let

$$\alpha(x,y) = \int_0^1 h(x+ty)\,dt \;=\; \frac{1}{y}\int_x^{x+y} h(\xi)\,d\xi;$$
$$\beta(x,y) = \int_0^1 h(x-ty)\,dt \;=\; \frac{1}{y}\int_{x-y}^{x} h(\xi)\,d\xi.$$

Then

$$f(x+iy) \;=\; u(x,y)+iv(x,y) \;=\; \tfrac{1}{2}(\alpha+\beta)+\tfrac{1}{2}i(\alpha-\beta), \qquad y \neq 0.$$

Clearly, α and β are C^1 on $\mathbb{C}\setminus\mathbb{R}$. Moreover, $\alpha(x,-y)=\beta(x,y)$, from which it follows that

$$f(\bar{z}) \;=\; \overline{f(z)}. \tag{8.7.5}$$

Since h is strictly increasing, f maps the upper half plane $\mathbb{H}$ into itself, and the lower half-plane into itself. We concentrate for now on $\mathbb{H}$: $y>0$.

Note that α and β represent the mean value of h over the intervals $[x, x+y]$ and $[x-y, x]$, respectively. Since h is strictly increasing, this implies that

$$\alpha \;>\; \beta; \quad \alpha_x \;>\; \beta_x \;>\; 0; \quad \alpha_y \;>\; -\beta_y \;>\; 0. \tag{8.7.6}$$

We know that f is single-valued from $\mathbb{R}$ to $\mathbb{R}$. It is also single-valued on $\mathbb{C}\setminus\mathbb{R}$. It is enough to prove this in $\mathbb{H}$. Suppose that $z_j = x_j + iy_j \in \mathbb{H}$, $j=1,2$, with $\alpha(x_1,y_1)=\alpha(x_2,y_2)$. Assume $x_1 \le x_2$. From the equality of the two mean values, we see that $y_1 \ge y_2$. Turning to β, we see that this implies that $x_1 \ge x_2$. Thus, $x_1 = x_2$ which implies that $y_1 = y_2$.

The inequalities (8.7.2) imply that the Jacobian of f in $\mathbb{H}$,

$$\begin{bmatrix} u_x & u_y \\ v_x & v_y \end{bmatrix} \;=\; \frac{1}{4}\begin{bmatrix} \alpha_x+\beta_x & \alpha_y+\beta_y \\ \alpha_x-\beta_x & \alpha_y-\beta_y \end{bmatrix} \;=\; \frac{\alpha_y\beta_x-\alpha_x\beta_y}{2} \;>\; 0. \tag{8.7.7}$$

Therefore f preserves orientation. It follows also that if a sequence $\{z_n\}\subset\mathbb{H}$ converges to a point $z \in U$, then $f(\mathbb{H})$ contains a neighborhood of $f(z)$. Therefore the boundary of $f(\mathbb{H})$ consists only of $\mathbb{R}$, so $f : \mathbb{H} \to \mathbb{H}$ is bijective. By (8.7.5), $f : \mathbb{C} \to \mathbb{C}$ is bijective.

To this point, we have not used the assumption of quasisymmetry, or of the normalization $f \in H(M)$; f could be any increasing homeomorphism of $\mathbb{R}$. As we know, the Jacobian of f can be expressed as $|f_z|^2 - |f_{\bar{z}}|^2$. The inequality (8.7.7) shows that f is quasiconformal on each compact subset of $\mathbb{C} \setminus \mathbb{R}$.

We now invoke the assumption $f \in H(M)$. Note that for any affine maps $A_j(z) = a_j z + b_j$ with $a_j > 0$ and $b_j \in \mathbb{R}$, $f_1 = A_1 \circ f \circ A_2$ is the Beurling–Ahlfors extension of $h_1 = A_1 \circ h \circ A_2$. Given any point $z_0 \in \mathbb{H}$, we may choose A_1 and A_2 in such a way that h_1 is normalized and $f(z_0) = f_1(i)$.

Suppose that f is not quasiconformal. Then there is a sequence of points $\{z_n\}$ in $\mathbb{H}$ such that the dilatation $D_f(z_n) \to \infty$. As just noted, we may normalize and obtain a sequence $\{f_n\}$ of normalized maps such that $f_n(i) = f(z_n)$. Passing to a subsequence and renumbering, we may assume that $\{h_n = f_n|_{\mathbb{R}}\}$ converges uniformly on bounded intervals in $\mathbb{R}$. This implies that the Jacobians of the Beurling–Ahlfors extensions converge uniformly in a neighborhood of i. Therefore the limit has finite dilatation at i, contradicting the assumption.

Finally, suppose that the last assertion in the statement of the theorem is not true. Then there is a sequence $h_n \in M(1/n)$ such that the maximal dilatation $K_{f_n} \geq 1 + \varepsilon > 1$, where f_n is the extension of h_n. Once again we may renormalize so that $D_{f_n}(i) \geq \varepsilon$, pass to a subsequence and assume uniform convergence on bounded intervals. The extension f of the limit function f has $D_f(i) \geq 1 + \varepsilon$. However, by Lemma 8.7.3, the limit function h is the identity, so the extension f is the identity, and $D_f(i) = 1$. □

Note that the Beurling–Ahlfors extension can be transferred to $\mathbb{D}$ by means of the Cayley transform. A different extension from the boundary of $\mathbb{D}$ to $\mathbb{D}$, with better invariance properties, is due to Douady and Earle [57].

An important property of the extension is its relation to the hyperbolic metric. A homeomorphism $f : \mathbb{H} \to \mathbb{H}$ is said to be a *quasi-isometry* if there is a constant $C > 0$ such that

$$\frac{1}{C}\,\frac{|dz|}{\operatorname{Im} z} \;\leq\; \frac{|df(z)|}{\operatorname{Im} f(z)} \;\leq\; C\,\frac{|dz|}{\operatorname{Im} z}. \tag{8.7.8}$$

Theorem 8.7.5. *The Beurling–Ahlfors extension f of a quasisymmetric function h on $\mathbb{R}$ is a quasi-isometry.*

Proof: The map f is K-quasiconformal for some K. Given $z \in \mathbb{H}$,

$$\frac{|df(z)|}{|dz|} \;\leq\; \sup_a |\partial_a f(z)| \;\leq\; \sqrt{K\,J_f(z)},$$

where J_f is the Jacobian. Therefore the right side of (8.7.8) will follow from

$$K\,J_f(z)(\operatorname{Im} z)^2 \;\leq\; C^2\,(\operatorname{Im} f(z))^2, \qquad z \in \mathbb{H}. \tag{8.7.9}$$

Once again we assume that (8.7.9) is false, and choose a sequence of normalized k-quasisymmetric functions h_n that converge uniformly on compact sets to a k-quasisymmetric function h, such that the Beurling–Ahlfors extensions f_n satisfy

$$\frac{J_{f_n}(i)}{(\operatorname{Im} f_n(i))^2} \to \infty. \tag{8.7.10}$$

But the Beurling–Ahlfors extension f of h satisfies $J_{f_n}(i) \to J_f(i)$ and $\operatorname{Im} f_n(i) \to \Im f(i)$, contradicting (8.7.10). The same kind of argument proves the other half of (8.7.8). □

8.8 Complex dilatation; the Beltrami equation

Throughout this section, we denote by C_0^m the space of functions $g : \mathbb{C} \to \mathbb{C}$ such that g belongs to C^m and g has compact support, i.e. g vanishes outside some bounded set.

In Section 8.1, we considered maps f that satisfy

$$\left|\frac{f_{\bar{z}}(z)}{f_z(z)}\right| \le k, \tag{8.8.1}$$

where $k < 1$ is a constant. The argument in Section 8.1 shows that at any point where the derivative of f exists and satisfies (8.8.1), the dilatation of f at that point is $K(z) \le (1+k)/(1-k)$. Therefore, if $f : \Omega \to \Omega'$ is a C^1 homeomorphism that satisfies (8.8.1) at each point, then f is K-quasiconformal, with $K \le (1+k)/(1-k)$.

Let us rewrite (8.8.1) in the form of a differential equation, of a type known as a *Beltrami equation*:

$$\frac{\partial f}{\partial \bar{z}} = \mu(z)\frac{\partial f}{\partial z}, \qquad |\mu(z)| \le k < 1. \tag{8.8.2}$$

In this section, we discuss the existence of a solution to (8.8.2), given some conditions on the function μ.

Remark. The function μ here is not to be confused with the Grötzsch module function of Section 8.5 and Section 8.6. (In both cases we are following standard usage.)

The strategy is to convert the Beltrami equation into a system

$$f_{\bar{z}} = g, \qquad f_z = Tg, \tag{8.8.3}$$

so that (8.8.2) becomes the 2-step process:

$$f = Pg, \qquad g = \mu T g.$$

The first step is then to solve $f_{\bar z} = g$ for some reasonably large class of functions g. This is accomplished by the (two-dimensional) *Cauchy transform* P:

$$\begin{aligned} Pg(z) &= -\frac{1}{\pi}\iint_{\mathbb{C}} g(w)\left(\frac{1}{w-z}-\frac{1}{z}\right)dx\,dy \\ &= \frac{1}{2\pi i}\iint_{\mathbb{C}} g(w)\left(\frac{1}{w-z}-\frac{1}{z}\right)dw\wedge d\bar w \qquad (8.8.4) \\ &= \frac{1}{2\pi i}\iint_{\mathbb{C}} g(z+w)\left(\frac{1}{w}-\frac{1}{z+w}\right)dw\wedge d\bar w. \qquad (8.8.5) \end{aligned}$$

As we shall see, for appropriate g, $f = Pg$ is a solution to $f_{\bar z} = g$. Moreover, $f_z = Tg$, where T is the *Hilbert transform* (in the plane),

$$\begin{aligned} Tg(z) &= -\frac{1}{\pi}\lim_{\varepsilon\to 0}\iint_{\varepsilon<|w-z|<1/\varepsilon}\frac{g(w)}{(w-z)^2}\,dx\,dy \\ &= \frac{1}{2\pi i}\lim_{\varepsilon\to 0}\iint_{\varepsilon<|w-z|<1/\varepsilon}\frac{g(w)}{(w-z)^2}\,dw\wedge d\bar w. \qquad (8.8.6) \end{aligned}$$

Transforming to polar coordinates centered at w shows that T vanishes on constant functions. Therefore, if g belongs to C_0^1 and R is large enough,

$$Tg(z) \;=\; T(g(z)-g(0)) \;=\; \frac{1}{2\pi i}\int_{|w|<2R}\frac{g(z+w)-g(z)}{w^2}\,dw\wedge d\bar w.$$

Moreover, in this case, Tg is easily seen to be continuous.

Lemma 8.8.1. *Suppose that g belongs to C_0^1. Then Pg is a C^1 function and*

$$(Pg)_{\bar z} = g; \qquad (Pg)_z \;=\; Tg. \qquad (8.8.7)$$

Proof: Differentiate under the integral sign to obtain

$$\begin{aligned} (Pg)_{\bar z} &= \frac{1}{2\pi i}\iint_{\mathbb{C}}\frac{g_{\bar z}(z+w)}{w}\,dw\wedge d\bar w \\ &= \frac{1}{2\pi i}\iint_{\mathbb{C}}\frac{g_{\bar w}(w)}{w-z}\,w\wedge d\bar w \\ &= -\frac{1}{2\pi i}\iint_{\mathbb{C}}\frac{dg(w)}{w-z}\wedge dw. \end{aligned}$$

Let $\Omega_\varepsilon \;=\; \{w : |w| > \varepsilon\}$. Applying Stokes's theorem, we obtain

$$\begin{aligned} (Pg)_{\bar z} &= -\lim_{\varepsilon\to 0}\frac{1}{2\pi i}\int_{\Omega_\varepsilon} d\left[\frac{g(w+z)}{w}\right]dw \\ &= \frac{1}{2\pi i}\lim_{\varepsilon\to 0}\int_{|w|=\varepsilon}\frac{g(w+z)}{w}\,dw \;=\; g(z). \end{aligned}$$

A similar calculation shows that

$$(Pg)_z = \lim_{\varepsilon\to 0}\left[-\frac{1}{2\pi i}\int_{|w|=\varepsilon}\frac{g(w-z)}{w}\,d\bar{w} + \frac{1}{2\pi i}\int_{|w|=\varepsilon}\frac{g(w-z)}{w^2}\,dw\wedge d\bar{w}\right]$$
$$= Tg(z).$$

□

Lemma 8.8.2. *Suppose that g belongs to C_0^2. Then*

$$P(g_z)(z) \;=\; Tg(z) - Tg(0). \tag{8.8.8}$$

Proof: Following the procedure in the proof of Lemma 8.8.1, we may write

$$\begin{aligned} P(g_z)(z) &= \frac{1}{2\pi i}\iint_{\mathbb{C}} g_w(w)\left(\frac{1}{w-z}-\frac{1}{w}\right)dx\,dy \\ &= \frac{1}{4\pi i}\iint_{\mathbb{C}}[g_x(w)-ig_y(w)]\left(\frac{1}{w-z}-\frac{1}{w}\right)dx\,dy. \end{aligned}$$

The last line gives two integrals. One can be integrated by parts in x so long as $y \neq 0$, and the other can be integrated by parts in y if $x \neq 0$. Integration over $\mathbb{C}$ does not see the exceptional lines, so integration by parts leads immediately to (8.8.8). □

Lemma 8.8.3. *Suppose g belongs to C_0^3. Then Tg is a C^1 function and*

$$\iint_{\mathbb{C}} |Tg(z)|^2\,dx\,dy \;=\; \iint_{\mathbb{C}} |g(z)|^2 dx\,dy. \tag{8.8.9}$$

Proof: Apply Lemmas 8.8.1 and 8.8.2 to g_z to find that

$$\begin{aligned} (Tg)_{\bar{z}} &= (Pg)_{z\bar{z}} \;=\; (Pg_z)_{\bar{z}} \;=\; g_z; \\ (Tg)_z &= (Pg_z)_z = T(g_z) \;=\; P(g_{zz})(z) + T(g_z)(0). \end{aligned}$$

The assumption implies that g_{zz} belongs to C_0^1. Therefore both $T(g)_{\bar{z}}$ and $T(g)_z$ are continuous. Thus, Tg is in C^1. The assumption that g has compact support implies that $Tg(z) = O(|z|^{-2})$ as $z \to \infty$. Therefore both sides of (8.8.9) are finite. Moreover, the following integrations-by-parts are justified:

$$\begin{aligned} \iint_{\mathbb{C}} Tg\,\overline{Tg} = \iint_{\mathbb{C}}(Pg)_z\,\overline{(Pg)_z} &\;=\; -\iint_{\mathbb{C}} Pg\,\overline{(Pg)}_{z\bar{z}} \\ = -\iint_{\mathbb{C}} Pg\,\bar{g}_{\bar{z}} &\;=\; \iint_{\mathbb{C}} g\,\bar{g}. \end{aligned}$$

□

Since C_0^2 is dense in $L^2(\mathbb{C})$, we have

Corollary 8.8.4. *T extends to an isometry of $L^2(\mathbb{C})$.*

On the other hand, as we shall see, the integral defining Pg is only guaranteed to converge if g belongs to $L^p(\mathbb{C})$ for $2 < p \leq \infty$. However, this difficulty can be overcome. The key is the *Calderón–Zygmund inequality*:

Theorem 8.8.5. *The operator T extends to a bijective map from L^p to L^p, $1 < p < \infty$, and*

$$||Tg||_p \;\leq\; C_p\,||g||_p, \tag{8.8.10}$$

where $1 \leq C_p < \infty$ and $C_p \to 1$ as $p \to 2$.

A proof is given in the next section. Let us turn to the operator P.

Theorem 8.8.6. *For $f \in L^p$, $2 < p < \infty$, Pf vanishes at $z = 0$ and satisfies a Hölder continuity condition*

$$|Pg(z_1) - Pg(z_2)| \;\leq K_p||g||_p\,|z_1 - z_2|^{1-2/p}. \tag{8.8.11}$$

Proof: Given $z \in \mathbb{C}$, the function $h_z(w) = |(w-z)^{-1} - w^{-1}| = |z||(w-z)w|^{-1}$ is $O(|w|^{-1} + |w-z|^{-1})$ near the singularities and is $O(|w|^{-2})$ as $w \to \infty$. Therefore it belongs to L^q for $1 < q < 2$. By Hölder's inequality, the integral $Pg(z)$ converges so long as $g \in L^p(\mathbb{C})$, $2 < p < \infty$. In fact

$$|Pg(z)| \;\leq\; ||g||_p\,||h_z||_q, \qquad \frac{1}{p} + \frac{1}{q} = 1, \quad p > 2,$$

i.e. $q = p/(p-1)$. Writing $w = x + iy = |z|\zeta = |z|(\xi + i\eta)$, we have

$$\begin{aligned} ||h_z||_q^q &= |z|^q \iint_{\mathbb{C}} \frac{1}{|(w-z)w)|^q}\,dx\,dy \\ &= |z|^{2-q} \iint_{\mathbb{C}} \frac{1}{|(\zeta-1)\zeta|^q}\,d\xi\,d\eta \\ &= |z|^{2-q}\,(K_p)^q, \end{aligned}$$

where K_p is constant. Since $(2-q)/q = 1 - 2/p$, we have

$$|Pg(z)| \;\leq\; |z|^{1-2/p}K_p||g||_p, \qquad p > 2, \quad q = p/(p-1). \tag{8.8.12}$$

Then

$$\begin{aligned} Pg(z_2) - Pg(z_1) &= -\frac{1}{\pi}\iint_{\mathbb{C}} g(w)\left(\frac{1}{w-z_2} - \frac{1}{w-z_1}\right)dx\,dy \\ &= -\frac{1}{\pi}\iint_{\mathbb{C}} g(w+z_1)\left(\frac{1}{w-(z_2-z_1)} - \frac{1}{w}\right)dx\,dy \\ &= P\widetilde{g}(z_2 - z_1), \qquad \widetilde{g}(z) = g(z+z_1). \end{aligned}$$

Now $||\widetilde{g}||_p = ||g||_p$, so (8.8.12) gives (8.8.11). □

We know now that for $g \in C_0^3$, $(Pg)_{\bar{z}} = g$, $(Pg)_z = Tg$, and Pg is Hölder continuous. These results can be carried over, in a certain sense, for $g \in L^p$: the "weak" sense, or the sense of distribution theory, as in Section 2.9.

Proposition 8.8.7. *If $g \in L^p$ for some $1 < p < \infty$, then $f = Pg$ satisfies the Hölder condition (8.8.11) and is a weak solution of equations (8.8.3).*

Proof: We know that these equations are true in the usual ("strong") sense if g is in C_0^1. The space C_0^1 is dense in L^p; Theorems 8.8.5 and 8.8.6 allow passage to the limit. □

We are now prepared to solve the Beltrami equation, in several steps. Throughout, μ will be a measurable function with $|\mu(z)| \le k < 1$, all $z \in \mathbb{C}$. Since (8.8.2) can only specify f up to an additive constant and a multiplicative constant, we shall normalize by requiring $f(0) = 0$ and $f(1) = 1$.

Given $k < 1$, we fix $p = p(k) > 2$ with $k\,C_p < 1$, where C_p is the constant in Theorem 8.8.5.

Theorem 8.8.8. *If μ has compact support, then (8.8.2) has a unique normalized solution f such that $f_z - 1$ belongs to L^p. Moreover f is Hölder continuous:*

$$|f(z_1) - f(z_2)| \;\le\; \frac{K_p}{1 - k\,C_p}||\mu||_p|z_1 - z_2|^{1-2/p} + |z_1 - z_2|. \tag{8.8.13}$$

Proof: Suppose first that f is a solution of (8.8.2). Then

$$f_{\bar{z}} \;=\; \mu f_z \;=\; \mu(f_z - 1) + \mu$$

belongs to L^p. The function $F = f - P(f_{\bar{z}})$ is a weak solution of $F_{\bar{z}} = 0$. By Theorem 2.9.3, F is an entire function. But

$$F' - 1 \;=\; (f_z - 1) + T(f_{\bar{z}}) \in L^p,$$

so $F(z) = z + c$. The normalization implies that the constant $c = 0$, so we must have

$$f(z) = P(f_{\bar{z}})(z) + z \;=\; P(\mu f_z) + z; \qquad f_z \;=\; T(\mu f_z) + 1. \tag{8.8.14}$$

If g is another solution, then $f_z - g_z = (f_z - 1) - (g_z - 1)$ belongs to L^p and

$$||f_z - g_z||_p \;=\; ||T(\mu(f_z - g_z))||_p \;\le\; k\,C_p||f_z - g_z||_p.$$

Therefore $g_z - f_z = 0$ a.e. Then the Beltrami equation implies $(f - g)_{\bar{z}} = 0$ a.e., and all this is true of $(\bar{f} - \bar{g})$ as well, so $f - g$ and $\bar{f} - \bar{g}$ are analytic. Therefore $f - g$ is constant, and the normalization gives $f = g$.

We have proved uniqueness, but the argument tells us how to prove existence. According to (8.8.14), we want

$$f_z - 1 \;=\; T(\mu(f_z - 1)) + T\mu.$$

The operator $Sg = T(\mu g)$ has a norm less than 1 as an operator in L^p. Therefore for $h \in L^p$, the series

$$T\mu + S(T\mu) + S^2(T\mu) + \cdots + S^n(T\mu) + \ldots \tag{8.8.15}$$

converges in norm to the unique solution $h \in L^p$ of $h = Sh + T\mu = T(\mu(h+1))$ (apply $I - S$ to the series). Since μ has compact support, $\mu(h+1)$ is also in L^p, and the construction shows that

$$||\mu(h+1)||_p \;\le\; \frac{1}{1 - k\,C_p}||\mu||_p. \tag{8.8.16}$$

Thus, we may define

$$f \;=\; P(\mu(h+1)) + z. \tag{8.8.17}$$

Then

$$f_{\bar z} \;=\; \mu(h+1), \qquad f_z \;=\; T(\mu(h+1)) + 1 \;=\; h + 1. \tag{8.8.18}$$

Therefore $f_z - 1 = h$ is in L^p and f is a (distribution) solution of (8.8.2). The estimate (8.8.13) follows from (8.8.16), (8.8.17), and (8.8.11). □

The solution f just constructed is termed the *normal* solution of the Beltrami equation. We want to show that the normal solution is a K-quasiconformal map, $K = (1+k)/(1-k)$. We begin by showing that if μ above has some regularity, then f is C^1.

Lemma 8.8.9. *Suppose that μ in Theorem 8.8.8 has a distribution derivative $\mu_z \in L^p$, $p > 2$. Then the normal solution f is a C^1 function.*

Proof: Consider the system $f_z = \lambda$, $f_{\bar z} = \lambda\mu$. By Theorem 2.9.5, this system has a C^1 solution f if λ has a weak derivative $\lambda_{\bar z} \in L^p$, $\lambda\mu$ has a weak derivative $(\lambda\mu)_z$ in L^p, and

$$\lambda_{\bar z} \;=\; (\mu\lambda)_z \;=\; \lambda_z\mu + \lambda\mu_z. \tag{8.8.19}$$

Dividing by λ, we want

$$(\log\lambda)_{\bar z} \;=\; \mu(\log\lambda)_z + \mu_z.$$

We can solve $q = T(\mu q) + T(\mu_z)$ for $q \in L^p$. Let

$$\sigma \;=\; P(\mu q + \mu_z) + c,$$

where the constant c is chosen so that $\lim_{z\to\infty}\sigma(z) = 0$. Then σ is continuous and

$$\sigma_{\bar z} \;=\; \mu q + \mu_z; \qquad \sigma_z \;=\; q.$$

Therefore $\lambda = e^\sigma$ satisfies (8.8.19), and $f_z = \lambda$, $f_{\bar z} = \lambda\mu$ has a C^1 solution. □

Lemma 8.8.10. *Suppose that the normal solution f in Theorem 8.8.8 is a C^1 function. Then f is K-quasiconformal, $K = (k+1)/(k-1)$.*

Proof: We know from (8.3.11) that f is locally invertible and that the local inverse satisfies

$$(f^{-1})_\zeta \;=\; \frac{\overline{f_z}}{J_f}, \quad (f^{-1})_{\bar\zeta} = -\frac{f_{\bar z}}{J_f},$$

where we have written $\zeta = f(z)$ and J_f is the Jacobian of f: $J_f = |f_z|^2 - |f_{\bar z}|^2$. Therefore f^{-1} is a (local) solution of the Beltrami equation with coefficient

$$\widetilde{\mu}(\zeta) \;=\; -\frac{\overline{f_{\bar z}(z)}}{f_z(z)}.$$

Then, writing $\zeta = s + it$ and using Hölder's inequality, we obtain

$$\begin{aligned}
\iint_C |\widetilde{\mu}(\zeta)|^p\, ds\, dy &= \iint_{\mathbb{C}} |\mu(z|^p |J_f(z)|\, dx\, dy \\
&= \iint_{\mathbb{C}} |\mu|^p (|f_z|^2 - |f_{\bar z}|^2)\, dx\, dy \\
&\le \iint_{\mathbb{C}} |\mu|^p |f_z|^2\, dx\, dy \;=\; \iint_{\mathbb{C}} |\mu_z|^{p-2} f_{\bar z}|^2\, dx\, dy \\
&\le \left(\int_{\mathbb{C}} |\mu|^p\, dx\, dy \right)^{(p-2)/p} \left(\iint_{\mathbb{C}} |f_{\bar z}|^p \right)^{2/p}.
\end{aligned}$$

Therefore

$$||\widetilde{\mu}||_p \;\le\; (1 - k\, C_p)^{-2/p} ||\mu||_p.$$

It follows that the normal solution to the Beltrami equation with coefficient $\widetilde{\mu}$ satisfies

$$|z_1 - z_2| \;\le\; K_p (1 - k\, C_p)^{-1-2/p} |f(z_1) - f(z_1)|^{1-p/2} + |f(z_1) - f(z_2)|.$$

Therefore f is globally invertible, hence a K-quasiconformal map. □

We now remove the assumption that the normal solution is a C^1 function.

Theorem 8.8.11. *The normal solution of the Beltrami equation is K-quasiconformal.*

Proof: Let $\{u_n\}$ be a sequence of functions with distribution derivatives $u_z \in L^p$, supported in some compact subset of $\mathbb{C}$, uniformly bounded by k, and such that $||\mu_n - u||_p \to 0$. Let $\{f_n\}$ and f be the corresponding normal solutions of the Beltrami equation. Then $||(f_n)_z - f_z||_p \to 0$. Consider the associated functions h_n and h constructed in the proof of Theorem 8.8.8: $(I - S)h_n = T\mu_n$. Then the series solution shows that

$$
\begin{aligned}
||h_n - h||_p &\le ||T(\mu_n - \mu)||_p + k\,C_p||T(\mu_n - \mu)||_p + (k\,C_p)^2||T(\mu_n - \mu)||_p + \dots \\
&\le \frac{C_p}{1 - k\,C_p}||\mu_n - \mu||_p \to 0.
\end{aligned}
$$

Therefore (8.8.17) and Theorem 8.8.6 imply uniform convergence on bounded sets. By Theorem 8.2.5, f is K-quasiconformal. □.

We would like to drop the requirement that μ has compact support. Let us start with an observation about composition of regular quasiconformal maps. Suppose that f and g are such maps, with *Beltrami coefficients* μ_f and μ_g. Write ζ for $f(z)$. Then

$$
\begin{aligned}
(g \circ f)_{\bar z} &= (g_\zeta \circ f)\, f_{\bar z} + (g_{\bar\zeta} \circ f)\, \overline{f_z}; \\
(g \circ f)_z &= (g_\zeta \circ f)\, f_z + (g_{\bar\zeta} \circ f)\, \overline{f_{\bar z}}.
\end{aligned}
$$

This system can be solved for $\mu_g \circ f$:

$$
\mu_g \circ f \;=\; \frac{f_z}{\overline{f_z}} \cdot \frac{\mu_{g\circ f} - \mu_f}{1 - \bar\mu_f \mu_{g\circ f}}. \tag{8.8.20}
$$

Theorem 8.8.12. *For any measurable function μ with $|\mu(z)| \le k < 1$ a.e., there is a unique K-quasiconformal map $f : \mathbb{C} \to \mathbb{C}$, $K = (1+k)/(1-k)$ such that f is a weak solution of the Beltrami equation (8.8.2), and $\mu(0) = 0$, $\mu(1) = 1$.*

Proof: We assume here that various coefficients μ are sufficiently regular that the solutions are C^1. The general result follows by an approximation argument.

If μ has compact support, we need only divide the normal solution f by $f(1)$. Suppose that $\mu = \mu_1 + \mu_2$, where μ_1 has compact support and μ_2 is $\equiv 0$ in a neighborhood of 0. Let

$$
\nu(z) \;=\; \mu_2\left(\frac{1}{z}\right)\frac{z^2}{\bar z^2}.
$$

Then ν has compact support, so there is a corresponding normalized solution f^ν. A computation shows that

$$
f^{\mu_2}(z) \;=\; \frac{1}{f^\nu(1/z)}.
$$

The next step is to assume that there is a solution f^μ and see how it can be written as a composition

$$
f^\mu \;=\; f^\lambda \circ f^{\mu_2}.
$$

The preceding computation for composite functions shows that the coefficient λ should be

$$
\lambda \circ f^{\mu_2} \;=\; \left[\frac{f_z^{\mu_2}}{\bar f_{\bar z}^{\mu_2}} \cdot \frac{\mu - \mu_2}{1 - \mu\bar\mu_2}\right] \;=\; \left[\frac{f_z^{\mu_2}}{\bar f_{\bar z}^{\mu_2}} \cdot \frac{\mu_1}{1 - \mu\bar\mu_2}\right]
$$

Then λ has compact support, so f^{μ} is well-determined. □

It is important for later use to consider quasiconformal maps of $\mathbb{H}$ to itself. Such a map can be extended to $\mathbb{C}$ by setting $\mu \equiv 0$ on the lower half-plane. By Theorem 8.2.3, the extension is quasiconformal on $\mathbb{C}$.

Theorem 8.8.13. *Given $\mu : \mathbb{H} \to \mathbb{C}$ with $||\mu||_{\infty} < 1$, there is a μ-quasiconformal map f^{μ} of $\mathbb{H}$ onto itself with $0, 1, \infty$ as fixed points.*

Proof: Extend μ to the lower half-plane by setting $\mu(z) = \overline{\mu(\bar{z})}$ if $\operatorname{Im} z < 0$. Then, by uniqueness, $f^{\mu}(\bar{z}) = \overline{f^{\mu}(z)}$. Therefore f^{μ} maps $\mathbb{R}$ onto $\mathbb{R}$. Starting with $\mu \in C_0^1$, so that $f(z) \sim z$ as $z \to \infty$, and proceeding by approximations, we see that f^{μ} must map $\mathbb{H}$ onto itself. □

The following result uses the same kind of construction to decompose a quasiconformal map.

Theorem 8.8.14. *Given $0 < t < 1$, the map $f = f^{\mu}$ can be decomposed as*

$$f = f_{1-t} \circ f_t, \qquad \textit{with} \quad K_f = K_{f_t}^{t} K_{f_{1-t}}^{1-t}.$$

Proof: Let L be the length of the hyperbolic geodesic in $\mathbb{D}$ from 0 to μ. Let μ_t be the point on that geodesic at hyperbolic distance tL from μ and let $f_t = f^{\mu_t}$. Define μ_{1-t} and f_{1-t} similarly. Then

$$K_{f_t} = \left|\frac{1+|\mu_t|}{1-|\mu_t|}\right| = \left|\frac{1+|\mu|}{1-|\mu|}\right|^{t} = K_f^{t},$$

and similarly for f_{1-t}. □

Any quasiconformal homeomorphism of $\mathbb{C}$ to $\mathbb{C}$ can be shown to be differentiable at a.e. point of $\mathbb{C}$, so that the quotient $\mu = f_{\bar{z}}/f_z$ is defined and < 1 almost everywhere. In fact, the converse of Theorem 8.8.12 is true.

Theorem 8.8.15. *Suppose that $f : \mathbb{C} \to \mathbb{C}$ is a K-quasiconformal homeomorphism, normalized so that $f(0) = 0$, $f(1) = 1$. Then $f = f^{\mu}$ for a unique μ in $L^{\infty}(\mathbb{C})$, with $||\mu||_{\infty} < 1$.*

For the proof, see [5] or [132].

8.9 The Calderón–Zygmund inequality

We know by Lemma 8.8.3 that T extends to an isometry of L^2. We only need (8.8.10) in the range $2 \le p < \infty$, but a duality argument, using the adjoint T^*, shows that the result extends to the remainder of the range $1 < p < \infty$.

The inequality (8.8.10) is a special case of a very general theory. However, the complex-variable context here allows some special arguments, as in Vekua [210]. The idea is that

$$Tf(z) \;\equiv\; -\frac{1}{\pi}\lim_{\varepsilon\to 0}\iint_{|w|>\varepsilon}\frac{f(z-w)}{w^2}\,dx\,dy \;=\; -T_1^2 f(z),$$

where T_1 is bounded from L^p to L^p, $1 < p < \infty$, and that this boundedness property of T_1 can be obtained easily from the corresponding boundedness in $L^p(\mathbb{R})$ of the one-dimensional Hilbert transform H. Starting with $f \in C_0^1(\mathbb{R})$,

$$Hf(x) \;=\; \frac{1}{\pi}\lim_{\varepsilon\to 0}\int_{|y|>\varepsilon}\frac{f(x-y)}{y}\,dy.$$

Notice that since $1/t$ is an odd function, we may rewrite this as

$$Hf(x) \;=\; \frac{1}{\pi}\lim_{\varepsilon\to 0}\int_{|y|>\varepsilon}\frac{f(x-y)-f(x)}{y}\,dy,$$

and conclude from the assumption on f that the limit exists, uniformly with respect to x.

We claim that

$$||Hf||_p \;\le\; A_p||f||_p, \qquad f \in L^p(\mathbb{R}), \quad 1 < p < \infty. \tag{8.9.1}$$

One proof of (8.9.1) starts with a complex transform of a real-valued function $f \in C_0^1(\mathbb{R})$. Define

$$F(z) \;=\; \frac{1}{\pi i}\int_{-\infty}^{\infty}\frac{f(t)\,dt}{t-z}, \qquad \operatorname{Im} z > 0.$$

A straightforward calculation shows that the real part

$$u(x,y) \;=\; \operatorname{Re} F(x+iy) \;=\; \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{f(x+ys)\,ds}{s^2+1},$$

so

$$\lim_{y\to 0} u(x,y) = f(x). \tag{8.9.2}$$

Similarly,

$$v(x,y) \;=\; \operatorname{Im} F(x+iy) \;=\; \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{f(x-s)s\,ds}{s^2+y^2}\,ds,$$

so

$$\lim_{y\to 0} v(x,y) \;=\; Hf(x). \tag{8.9.3}$$

Direct calculation, using the Cauchy–Riemann equations for u, v shows that

$$\Delta(|u|^p) = p(p-1)|u|^{p-2}(v_x^2+v_y^2);$$
$$\Delta(|v|^p) = p(p-1)|v|^{p-2}(v_x^2+v_y^2);$$
$$\Delta(|F|^p) = p^2|F|^{p-2}(v_x^2+v_y^2).$$

Therefore

$$\Delta\left(|F|^p - \frac{p}{p-1}|v|^p\right) \;=\; p^2(|F|^{p-2} - |v|^{p-2})(v_x^2 + v_y^2), \quad p \geq 2. \tag{8.9.4}$$

For $R > 1$, let γ_R be the curve consisting of the horizontal diameter and upper semicircle of the circle in the upper half-plane with center iy and radius R, with the usual orientation. Let Ω_R be the domain enclosed by γ_R. Applying Green's identity

$$\iint_{\Omega_R} (g\Delta h - h\Delta g) \;=\; \int_{\gamma_R}\left(g\frac{\partial h}{\partial n} - h\frac{\partial g}{\partial n}\right)$$

with $g = 1$ and $h = |F|^p - p/(p-1)\cdot|v|^p$, and letting $R \to \infty$ shows that

$$\frac{\partial}{\partial y}\int_{-\infty}^{\infty}\left(|F(x+iy)|^p - \frac{p}{p-1}|v(x,y)|^p\right)dx \;\leq 0.$$

The integral has limit 0 as $y \to +\infty$, so

$$\int_{-\infty}^{\infty}|F(x+iy|^p\,dx \;\geq\; \frac{p}{p-1}\int_{-\infty}^{\infty}|v(x,y)|^p\,dx, \qquad y > 0,\; p \geq 2.$$

Now

$$\left(\int_{-\infty}^{\infty}|F|^p\right)^{2/p} \;=\; |||F|^2||_{p/2} \;=\; ||u^2+v^2||_{p/2} \;\leq\; ||u^2||_{p/2} + ||v^2||_{p/2}.$$

Therefore

$$\left(\frac{p}{p-1}\right)^{2/p}||v^2||_{p/2} \leq ||u^2||_{p/2} + ||v^2||_{p/2};$$

$$||v^2||_{p/2} \leq \left[\left(\frac{p}{p-1}\right)^{2/p} - 1\right]^{-1}||u^2||_{p/2}.$$

In view of (8.9.2) and (8.9.3), raising this inequality to the $p/2$ power and taking the limit as $y \to 0$ gives (8.9.1) with

$$A_p \;=\; \left[\left(\frac{p}{p-1}\right)^{2/p} - 1\right]^{-p/2}.$$

(Note that $A_2 = 1$.) This proves (8.9.1) for $p \geq 2$. A duality argument using Hölder's inequality shows that it is also true for $1 < p \leq 2$ with $A_p = A_q$, where $1/p + 1/q = 1$.

In the following calculations, we write $z = x + iy$, $\zeta = \xi + i\eta$. Then we define T_1 for $f \in C_0^2$:

$$T_1 f(z) = \lim_{\varepsilon \to 0_+}\iint_{|\zeta|>\varepsilon}\frac{1}{2\pi}\frac{f(\zeta+z)}{\zeta|\zeta|}\,dm(\zeta)$$

$$= \frac{1}{2}\int_0^\pi \left(\frac{1}{\pi}\int_0^\infty \frac{f(z+re^{i\theta})-f(z-re^{i\theta})}{r}\,dr\right)e^{-i\theta}\,d\theta.$$

Therefore

$$\begin{aligned} ||T_1 f||_p &\le \frac{\pi}{2}\sup_\theta \left|\left|\frac{1}{\pi}\int_0^\infty \frac{f(z+re^{i\theta})-f(z-re^{i\theta})}{r}\right|\right|_p \\ &= \frac{\pi}{2}\sup_\theta ||Hf_\theta||_p, \qquad f_\theta(x) = f(xe^{i\theta}) \\ &\le \frac{\pi}{2}A_p||f||_p. \end{aligned}$$

The next step is to show that $-T_1^2 = T$. For $f \in C_0^1$,

$$\begin{aligned} T_1 f(z) &= -\frac{1}{\pi}\iint [f(z+\zeta)-f(z)]\frac{\partial}{\partial\zeta}\frac{1}{|\zeta|}\,d\xi\,d\eta \\ &= \frac{1}{\pi}\iint f_\zeta(z+\zeta)\frac{1}{|\zeta|}\,d\xi\,d\eta \qquad (8.9.5) \\ &= \frac{1}{\pi}\frac{\partial}{\partial z}\iint f(\zeta)\left(\frac{1}{|z-\zeta|}-\frac{1}{|\zeta|}\right)d\xi\,d\eta. \qquad (8.9.6) \end{aligned}$$

Therefore, for any test function $\phi \in C_0^1$,

$$\iint T_1 f(z)\,\phi(z)\,dx\,dy \;=\; -\frac{1}{\pi}\iint\iint f(z)\left(\frac{1}{|z-\zeta|}-\frac{1}{|z|}\right)\phi_\zeta(\zeta)\,d\xi\,d\eta\,dx\,dy.$$

The integral on the right converges for $f \in L^p$, so (8.9.6) is true in the weak sense for $f \in L^p$. Then

$$\begin{aligned} T_1^2 f(w) &= \frac{\partial}{\partial w}\iint \frac{1}{\pi}\iint T_1 f(z)\left(\frac{1}{|z-w|}-\frac{1}{|z|}\right)dx\,dy \\ &= \frac{1}{\pi^2}\frac{\partial}{\partial w}\left[\iint\left(\frac{1}{|\zeta-w|}-\frac{1}{|\zeta|}\right)dx\,dy\iint\frac{f_z\,dx\,dy}{|z-\zeta|}\right] \\ &= \frac{1}{\pi^2}\frac{\partial}{\partial w}\left[\iint f_z\iint\frac{1}{|z-\zeta|}\left(\frac{1}{|\zeta-w|}-\frac{1}{|\zeta|}\right)dm(w)\right]dm(z) \\ &= -\frac{1}{\pi^2}\frac{\partial}{\partial w}\left[\iint f\frac{\partial}{\partial z}\iint\frac{1}{|z-\zeta|}\left(\frac{1}{|\zeta-w|}-\frac{1}{|\zeta|}\right)dm(w)\right]dm(z). \end{aligned}$$

Differentiation and integration can be exchanged and the expression replaced by

$$\lim_{R\to\infty}\frac{\partial}{\partial z}\left\{\iint_{|\zeta-w|<R}\frac{d\xi\,d\eta}{|z-\zeta||\zeta-w|}-\iint_{|\zeta|<R}\frac{1}{|\zeta||z-\zeta|}\,d\xi\,d\eta\right\}. \qquad (8.9.7)$$

The first integral here gives

$$\begin{aligned}\frac{\partial}{\partial z}\iint_{|\zeta|<R/|z-w|}\frac{d\xi\,d\eta}{|\zeta||1-\zeta|} &= \frac{\partial}{\partial z}\int_0^{R/|z-w|}\int_0^{2\pi}\frac{dr\,d\theta}{|1-re^{i\theta}|}\\ &= -\frac{1}{2}\,\frac{R}{(z-w)|z-w|}\int_0^{2\pi}\frac{d\theta}{\left|1-\frac{Re^{i\theta}}{|z-w|}\right|}.\end{aligned}$$

The limit is clearly $-\pi/(z-w)$ and similarly the limit of the second integral in (8.9.7) is $-\pi/z$. Thus,

$$T_1^2 f(w) \;=\; \frac{1}{\pi}\frac{\partial}{\partial w}\iint\left[f(z)\left(\frac{1}{z-w}-\frac{1}{z}\right)\right]dx\,dy \;=\; -Tf(w).$$

The last step is to verify that the best constant C_p of (8.8.10) has limit 1 as $p \to 2$. By Corollary 8.8.4, $C_2 = 1$. The rest follows from a particular case of the *Riesz–Thorin convexity theorem.*

Theorem 8.9.1. *The constant C_p in the estimate*

$$||Tf||_p \;\le\; C_p||f||_p, \qquad 2 \le p < \infty$$

can be chosen so that for any $p_1 > 2$,

$$C_p \;=\; C_2^{1-t}C_{p_1}^t \qquad \frac{1}{p} \;=\; \frac{t}{2}+\frac{1-t}{p}{}_1, \quad 0 \le t \le 1. \tag{8.9.8}$$

Proof: We use Hölder's inequality (2.8.1) with dual exponents $2, q, q_1$,

$$1 \;=\; \frac{1}{2}+\frac{1}{2} \;=\; \frac{1}{p}+\frac{1}{q} \;=\; \frac{1}{p_1}+\frac{1}{q_1}.$$

It is enough to consider functions in C_0^0, since such functions are dense in each L^p. Given such functions f and g, define functions F_ζ and G_ζ for ζ in the strip $2 \le \operatorname{Re}\zeta \le p_1$ by $F_\zeta(z) = 0$ if $f(z) = 0$, $G_\zeta(z) \;=\; 0$ if $g(z) = 0$, and otherwise

$$F_\zeta \;=\; |f|^{a(\zeta)}\frac{f}{|f|}, \qquad G_\zeta \;=\; |g|^{b(\zeta)}\frac{g}{|g|},$$

where

$$a(\zeta) \;=\; (1-\zeta)\,\frac{p}{2}+\zeta\,\frac{p}{p_1}.$$

We may also normalize so that $||f||_p = 1 = ||g||_q$. We want to show that under these assumptions, necessarily

$$\left|\iint_{\mathbb{C}} f\,g\right| \;\le\; C_2^{1-t}C_{p_1}^t. \tag{8.9.9}$$

The function $\Phi(\zeta) \;=\; \iint F_\zeta G_\zeta$ is holomorphic in the strip. For $\operatorname{Re}(\zeta) = 0$, $\operatorname{Re} a(\zeta) \;=\; p/2, \operatorname{Re} b(\zeta) \;=\; q/2,$

$$||F_\zeta||_2 \;=\; (||f||_p)^{p/2} \;=\; 1 \;=\; ||G_\zeta||_2, \qquad \operatorname{Re}\zeta = 0.$$

Similarly,

$$||F_\zeta||_{p_1} \;=\; (||f||_p)^{p/p_1} \;=\; 1 \;=\; ||G_\zeta||_{q_1}, \qquad \operatorname{Re}\zeta = 1.$$

Therefore

$$|\Phi(\zeta)| \;\le\; C_2 = 1 \quad \text{if} \operatorname{Re}\zeta = 0; \qquad |\Phi(\zeta)| \;\le\; C_{p_1}, \quad \text{if } \operatorname{Re}\zeta = 1.$$

Since Φ is bounded in the strip, $\Phi_\varepsilon = e^{1-\varepsilon z^2}\Phi \to 0$ as $z \to \infty$ in the strip, for any $\varepsilon > 0$, so $|\Phi_\varepsilon|$ is bounded by $\max\{C_2, C_{p_1}\}$. Taking $\varepsilon \to 0$, it follows that $|\Phi|$ has the same bounds. But then $C_2^{\zeta-1}C_{p_1}^{-\zeta}\Phi(z)$ is bounded by 1. Finally

$$\left|\int f\,g\right| \;=\; |\Phi(t)| \;\le\; C_2^{1-t}C_{p_1}^{t}. \qquad \square$$

Remarks. Ahlfors cites the first edition of Zygmund's treatise on the Fourier Series for the proof of the boundedness of the one-dimensional Hilbert transform H. In fact, it was to get away from the very special complex-analytic arguments that are used in this section, and to gain an understanding of the real-variable nature of these integral transforms, that Zygmund and Calderón developed their far-reaching theory of "singular integrals;"[37]. The Calderón–Zygmund arguments apply to much more general integral transforms; see Stein [194], Christ [44], or Peyrière [167].

Exercises

1. Prove that given a triple (p_1, p_2, p_3) of points in the unit circle $\partial\mathbb{D}$, ordered in the positive direction, there is a unique $f \in \operatorname{Aut}(\mathbb{D})$ such that f takes (p_1, p_2, p_3) to $(-1, -i, 1)$.
2. Show that a quadrilateral can be mapped conformally to some $\mathbb{H}(q_1, q_2, q_3, q_4)$, such the vertices can be taken to be $-1/k, -1, 1, 1/k$ for some (unique) $k > 1$.
3. For $\operatorname{Re} z > 0$ and $0 < k < 1$, let

$$f(z) \;=\; \int_0^z \frac{dx}{\sqrt{(1-x^2)(1-k^2x^2)}}.$$

(Note that the integral is independent of the path of integration in the upper half-space $\mathbb{H}$. Note also that f extends to be continuous on the closure of $\mathbb{H}$.)

(a) Show that the image $f(\mathbb{H})$ is bounded.

(b) Show that f maps the interval $[0, 1]$ to an interval $[O, K]$, for some finite $K(k) > 0$ and that f maps the interval $[1, 1/k]$ to the vertical interval $[0, iK']$ for some finite $K'(k) > 0$. Show that the image of $[1/k, \infty]$ is a finite horizontal segment extending to the left from $K + iK'$ and the image of $\{iy : y > 0\}$ is a finite vertical segment extending upward from the origin. Conclude from this that the image of the upper half plane is the interior of a rectangle with vertices $0, K, K + iK', iK'$.

4. For what values of $a, b, c, d \in \mathbb{R}$, $ad - bc = 1$, is the map $z = x + iy \to \zeta = u + iv$ quasiconformal, and what is the dilatation quotient, if

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

5. Prove that two rectangles R and R' are conformally equivalent if and only if $a/b = a'/b'$. Hint: reduce to the case $R = R' = \mathbb{D}$.
6. Verify (8.3.2).
7. Let R_a be the rectangle $R_a = z = x + iy : 0 < x < a, 0 < y < 1$, and let R_b, $b > a$, be defined in the same way, so that $m(R_a) = a$, $m(R_b) = b$, $b > a$ Let $f(x + iy) = \rho x + iy$ with $\rho = b/a$, so that $f(Q_a) = Q_b$. Verify that f is b/a-quasiconformal.
8. (Grötzsch) Let R_a and R_b be as in Exercise 7 and let g be a K-quasiconformal. Adapt the proof of Theorem 8.3.1 to show directly that $K \geq b/a$. By considering carefully the inequalities involved, show that if $K = b/a$, then g is the map f of Exercise 7.
9. Verify (8.3.11).
10. Show that the maps $f_{nm}(z) = z^n|z|^m$, $n, m = 1, 2, 3, \ldots$ are quasiconformal.
11. Let $f(re^{i\theta}) = r^a e^{i\theta}$, with $a \geq 1$.
    (a) Show that f is a regular K-quasiconformal map, and compute K.
    (b) Show that for a suitable value of a, f_a maps the annulus $A(1, r)$ onto $A(r, s)$, where $s \geq r > 1$ are specified.
    (c) Show that f in part (b) has the smallest maximal dilatation of any quasiconformal map from $A(1, r)$ to $A(1, s)$.
12. Suppose that Ω is a domain in $\mathbb{C}$ with the property that the complement in $\mathbb{C}$ of the closure $\overline{\Omega}$ consists of one bounded component Ω_0 and one unbounded component Ω_∞. Use the Riemann mapping theorem and inversion to prove:
    (a) Up to conformal equivalence we may assume that Ω_∞ is the complement of $\mathbb{D}$.
    (b) Up to a further conformal equivalence we may assume that the inner boundary Γ_1 of Ω is the unit circle, and the outer boundary G_2 is an analytic curve.
13. Suppose that Ω is the domain in Exercise 12 (b). We want to construct a conformal map F from Ω to some annulus R_r, taking inner boundary to inner boundary. Since F has no zeros, it can be written as $F = \exp(f + ig)$, where f is a real harmonic function such that $f = 0$ on Γ_1 and f is a positive constant on Γ_2, while g is a harmonic conjugate of f that gains 2π around a curve in Ω that is homotopic to the Γ_j.
    (a) Let f_1 be the harmonic function on Ω that vanishes on the inner boundary Γ_1 and equals 1 on the outer boundary Γ_2. Locally, either boundary can be straightened, so f can be extended across. Thus, f_1 is harmonic in a neighborhood of the closure $\overline{\Omega}$. Show that the outer normal derivative $(f_1)_n = \partial f_1/\partial n$ is nonnegative.

(b) Let g_1 be a harmonic conjugate of f_1, defined first in a neighborhood of a point of Γ_2. Use the fact that $f + ig$ is a non-constant holomorphic function to show that $(f_1)_n$ is positive at all but finitely many points of Γ_2. Use the Cauchy–Riemann equations to show that where $(f_1)_n$ is positive, the tangential derivative $(g_1)_\tau = \partial g_1/\partial\tau$ in the positive direction on Γ_2 is positive. Thus, g_1 is strictly increasing in the positive direction on Γ_2.

(c) Conclude that there is a positive multiple $f = cf_1$ such that the harmonic conjugate g gains 2π on a full circuit of Γ_2 (or any curve in Ω homotopic to Γ_2.) Thus, $F = \exp(f + ig)$ is a (single-valued) holomorphic map of Ω onto R_r, $r = e^c$. Note that F has a holomorphic extension to a neighborhood of $\overline{\Omega}$.

14. The object here is to show that the map F of Exercise 13 (d) is injective, and thus conformal. The number of times $N(z_0)$ that F takes the value z_0 is given by the usual formula

$$N(z_0) \;=\; \frac{1}{2\pi i}\int_{\partial\Omega} \frac{F'(\zeta)\,d\zeta}{F(\zeta) - z_0} \;=\; N_2(z_0) - N_1(z_0),$$

where
$$N_j(z_{}) \;=\; \frac{1}{2\pi i}\int_{\Gamma_j} \frac{F'(\zeta)\,d\zeta}{F(\zeta) - z_0}, \qquad j = 1, 2.$$

Here, we take the positive (domain to the left) orientation for each of the curves Γ_j, so that, symbolically, $\partial\Omega = \Gamma_2 - \Gamma_1$.

(a) Apply the argument principle, Theorem 1.5.1, to conclude that $N_1(z_0)$ is constant for $|z_0| < 1$ and for $|z_0| > 1$ and that $N_2(z_0)$ is constant for $|z_0| < r$ and for $|z_0| > r$.

(b) Show by direct computation that $N_j(0) = 1$.

(c) Use (a) and (b) to complete the proof.

15. Suppose that S is any Riemann surface and $\Omega \subset S$ is a domain such that the complement of $\overline{\Omega}$ in S has two open components. Show that Ω is conformally equivalent to an annulus $A(r, 1)$.

16. (a) Show that the domains

$$\Omega_0 \;=\; A(0,2) \;=\; \{z \,:\, 0 < |z| < 2\}, \qquad \Omega_1 \;=\; A(1,2) \;=\; \{z \,:\, 1 < |z| < 2\}$$

are homeomorphic but not conformally equivalent.

(b) Show that Ω_0 and Ω_1 are not quasiconformally equivalent: there is no quasiconformal map of Ω_0 onto Ω_1.

17. Suppose that $z_{,}$ and z_2 are distinct points in $\mathbb{D}$ with $|z_1 - z_2| < 2$. Show that there is an automorphism of $\mathbb{D}$ that takes the pair $\{z_1, z_2\}$ to $\{w, \bar{w}\}$, with $\operatorname{Re} w < 0$ and $|w - \bar{w}| = |z_1 - z_2|$.

18. Prove the assertion about the function φ in the proof of Proposition 8.5.7.

19. Prove the assertion about the function χ in the proof of Proposition 8.5.5.

20. (a) Use the change of variables $\zeta = (1 + k)t/(1 + kt^2)$ in the integrand for $K(k_1)$, where $k_1 = 2\sqrt{k}/(1 + k)$, to prove that

$$K\left(\frac{2\sqrt{k}}{1+k}\right) \;=\; (1+k)K(k).$$

(b) Use (8.5.22) and part (a) to prove the functional equation (8.5.17).

21. The aim of this and the two following exercises is to prove a version of the estimate (8.5.19):

$$\mu(r) = O\left(r \log r\right)\quad \text{as}\;\; r \to 1; \tag{8.9.10}$$

$$\mu(r) = \log\frac{4}{r} + O\left(r\;\log r\right)\qquad \text{as}\;\; r \to 0. \tag{8.9.11}$$

Prove that

$$K(k) \;=\; \int_0^1 \frac{dx}{\sqrt{1-x^2}}[1+O(x^2)] \;=\; \frac{\pi}{2}[1+O(x^2)].$$

Use this fact to show that (8.9.11) is equivalent to the estimate

$$K(k) \;=\; \log\{\frac{2}{\sqrt{1-k}}\} + O\left(\sqrt{1-k}\log(\sqrt{1-k})\right). \tag{8.9.12}$$

22. (a) Let $\text{ch} = \frac{1}{2}(k+1/k)$ and $\text{sh} = \frac{1}{2}(k-1/k)$. Verify that $(1-x^2)(1-x^2)^2 = (\text{ch}-kx^2)^2 - \text{sh}^2$.
(b) Use (a) to verify

$$\frac{1}{(1-x^2)(1-x^2)^2} \;=\; \frac{1}{(\text{ch}-kx^2)^2}\left(1 - \frac{\text{sh}^2}{(1-x^2)(1-k^2x^2)}\right).$$

(c) Show that for $k \geq 1/2$ the last quotient in (b) is $\leq 4(1-k)^2/(1-x)^2$, so that for $1-x \leq \sqrt{1-k}$, this term is bounded by $4\sqrt{1-k}$.
(d) Set $x(k) = 1-\sqrt{1-k}$. Writing $\sqrt{1-k} = \varrho$, verify

$$\begin{aligned} \text{I}_k = \int_1^{x(k)} \frac{d\xi}{\sqrt{(1-\xi^2)(1-k^2\xi^2)}} \;&=\; \int_1^{x(k)} \frac{d\xi}{\text{ch}-t\xi^2}(1+\varrho) \\ = \log\frac{1+x(k)}{1-x(k)}[1+O(\varrho)] \;&=\; \log\frac{2}{\varrho} + O(\varrho\log\varrho). \end{aligned} \tag{8.9.13}$$

Let $t \;=\; \text{ch} - kx^2$, so $dx \;=\; -dt/2kx$. Then

$$\begin{aligned} \text{II}_k = \int_{x(k)}^1 \frac{d\xi}{\sqrt{(1-\xi^2)(1-k^2\xi^2)}} \;&=\; -\frac{1}{2k}\int_{x(k)}^1 \frac{dt}{(t^2-\text{sh}^2)x} \\ = -\frac{1}{2k}\int_{x(k)}^1 \frac{dt}{(t^2-\text{sh}^2)}[1+O(\varrho)]. \end{aligned}$$

23. The indefinite integral of the integrand in the last line of Exercise 22 is $\log(t + \sqrt{t^2-\text{sh}^2})$.

(a) Verify that the upper and lower limits of the integral satisfy

$$\begin{aligned} \mathrm{ch} - k = \mathrm{sh} &= \frac{1+k}{2k}(1-k) = (1-k) + O((1-k)^2) \\ \mathrm{ch} - kx^2 = \mathrm{ch} - k(1-\varrho)^2 &= 2\varrho[1 + 0(\varrho)]. \end{aligned}$$

(b) Use (a) to verify

$$\mathrm{II}_k = -\log\varrho + \frac{1}{2}\log\varrho + O(\varrho\log\varrho). \tag{8.9.14}$$

(c) Combine (8.9.13) and (8.9.14) to obtain

$$\begin{aligned} K(k) &= -\log\varrho + \frac{1}{2}\left[\log\varrho - \log\frac{\varrho}{2}\right] + O(\varrho\log\varrho) \\ &= -\log\sqrt{\varrho} + \frac{1}{2}\log 2 + O(\varrho\log\varrho), \end{aligned}$$

and check that this is (8.9.12).

24. Show that equality in (8.6.2) can be attained if $f(\mathbb{D}) = \mathbb{D}$.
25. Write out a proof of Theorem 8.6.6.
26. Use Theorem 8.6.6 to show that a quasiconformal map takes sets of measure zero to sets of measure zero.
27. Use the map $f(z) = z|z|^{(1-K)/K}$ to show that the Hölder exponent of continuity of a K-quasiconformal map cannot in general be increased.
28. Show that the set $H(M)$ of normalized quasisymmetric functions on $\mathbb{R}$ is closed under convex combination: if $h_0, h_1 \in H(M)$, then

$$h_t = (1-t)h_0 + th_1 \qquad \text{belongs to } H(M), 0 \le t \le 1.$$

A consequence is that for any two Beurling–Ahlfors extensions of elements of $H(M)$, there is a continuous deformation from one to the other.

29. Verify (8.8.20).
30. Carry through the proof of (8.9.1) in the case $p = 2$.

Remarks and further reading

The study of quasiconformal mappings in the plane was begun by Grötzsch [94] and continued by Morrey [143]. It was further developed by Teichmüller [201], [202], [203] in his study of the moduli problem for Riemann surfaces. A related subject is that of “generalized analytic functions,” where the Beltrami equation again plays a leading role; see Vekua [210] and Rodin [179].

The presentation here relies largely on two standard references for quasiconformal maps in the plane: the notes of Ahlfors [5], and the comprehensive and careful book of Lehto and Virtanen [132]. The Ahlfors notes, and most of the subsequent literature on quasiconformal maps in one complex variable, are focused on its use in Teichmüller theory. For this, we refer to Chapter 9 and the references at the end of that chapter. The additional chapters in the second edition of [5] give some overview of further development of the theory.

The theory has been generalized to higher dimensions, starting with the work of Loewner [137] and Gehring and Väisälä; see [87]. One striking application is the Mostow rigidity theorem [144] concerning the analogue of Teichmüller theory in higher dimensions. Anticipating Chapter 9, the homotopy class of a compact Riemann surface of genus $g > 2$ is characterized by $6g - 6$ real parameters. Mostow proved that, in the analogous case for higher dimension, the moduli space is a point: if M and M' are closed hyperbolic manifolds of (real) dimension ≥ 3 and $f : M \to M'$ is a homeomorphism, then f is homotopic to an isometry.

The multidimensional theory remains an active area of research. See, for example, Iwaniec and Martin [116], Heinonen et al. [101], Gehring, Martin, and Palka [88].

Chapter 9
Introduction to Teichmüller theory

We consider here the problem of classifying Riemann surfaces. The most obvious classification is topological. For example, a compact Riemann surface is characterized topologically by its *genus*: the number of holes in the doughnut. See Figure 9.1 for genus 0, genus 1, and genus 2.

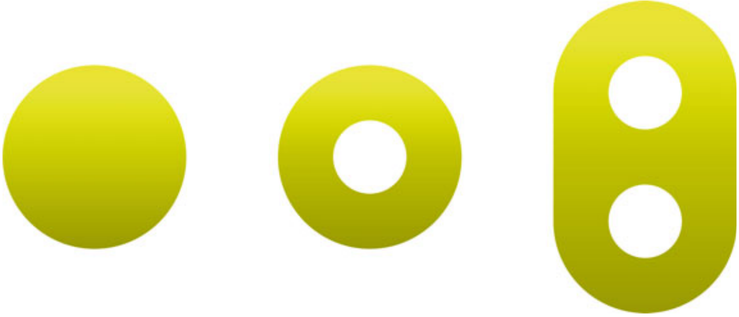

Fig. 9.1 Compact surfaces of genus 0, genus 1, and genus 2.

Conformal equivalence implies topological equivalence, so, for compact manifolds, the question is: given a topological surface of genus g, what inequivalent conformal structures can it carry?

For *simply connected* Riemann surfaces (genus 0), these questions are settled by the uniformization theorem, Theorems 7.2.1 and 7.4.3. Any such surface is conformally equivalent to the unit disk $\mathbb{D}$, the complex plane $\mathbb{C}$, or the Riemann sphere $\mathbb{S}$. (In place of $\mathbb{D}$ we shall usually consider the conformally equivalent upper half-space $\mathbb{H}$.) Now $\mathbb{C}$ and $\mathbb{D}$ are topologically—even real-analytically—equivalent ($re^{i\theta} \to \arctan(\pi r/2)e^{i\theta}$). But $\mathbb{C}$ and $\mathbb{D}$ are conformally distinct. For example, $\mathbb{D}$ carries non-constant holomorphic functions, but $\mathbb{C}$ does not. Neither $\mathbb{D}$ nor $\mathbb{C}$ is compact, but $\mathbb{S}$ is compact. Thus, in the simply connected case, the answer to the question

R. Beals and R. S. C. Wong, *More Explorations in Complex Functions*, Graduate Texts in Mathematics 298, https://doi.org/10.1007/978-3-031-28288-1_9

above is that a topological open disk or plane carries exactly two distinct conformal structures, while the sphere carries exactly one.

Restricting our attention to compact surfaces, we shall see that in the case of genus 1, there is a natural parametrization of the inequivalent complex structures by $\mathbb{C}$. Riemann [175] argued that for genus $g > 1$, these structures can be parametrized using $3g - 3$ complex parameters. However Riemann's argument does not provide the kind of geometric insight that the parameter space $\mathbb{C}$ provides for $g = 1$, for example—a measure of how close two structures are.

The problem of giving a good geometric description of Riemann's moduli space was revived in the 1940s by Teichmüller. Teichmüller's ideas have since been extended to the study of "bordered" Riemann surfaces, such as the closed disc $\overline{\mathbb{D}}$, and to non-compact Riemann surfaces, such as those obtained by removing closed sets (e.g. collections of isolated points and/or disjoint closed analytic disks) from a compact Riemann surface.

Teichmüller's fundamental idea was to start with maps from one Riemann surface to another. For example, if S and S' are compact Riemann surfaces that are homeomorphic, we shall see that there is a quasiconformal homeomorphism $f : S \to S'$. The maximal dilatation K_f is a measure of how far f deviates from a conformal map (for which $K_f = 1$). Therefore f provides an upper estimate $\log K_f$ on how far apart we should consider the conformal structures to be. We look for such f with minimal K_f, and consider $\log K_f$ as the relevant distance. Carrying this program through in a satisfactory way is not a simple matter. In particular, it relies heavily on the theory of quasiconformal maps as developed in Chapter 8.

In Section 9.1, we recall from Chapters 6 and 7 the basic machinery used to construct and classify Riemann surfaces: universal covers, the uniformization theorem, and groups of covering transformations. Moduli spaces are computed for spaces whose universal cover is conformally equivalent to $\mathbb{C}$.

Section 9.2 begins the study of maps from one Riemann surface to another, and their lifts to the universal cover. In Section 9.3, the developments in Section 9.2 are used to initiate the study of the Teichmüller space $T(S)$ of a Riemann surface S. This is a space of quasiconformal maps from S to topologically equivalent surfaces S'. More precisely, it is a space of *equivalence classes* of such maps. Minimizing the dilatation within the equivalence classes provides a natural metric. Section 4 examines $T(S)$ more closely for compact S.

In section 9.4, we turn to $T(\mathbb{H})$. Aside from the torus (genus 1), all the interesting Riemann surfaces are (up to conformal equivalence) quotients of $\mathbb{H}$ by a subgroup $G \subset \mathrm{Aut}(\mathbb{H})$. Quasiconformal maps of surfaces S, S' of the same genus lift to quasiconformal maps of $\mathbb{H}$ to itself. Conversely, if a quasiconformal homeomorphism of $\mathbb{H}$ to $\mathbb{H}$ is compatible with the action of the group of cover transformations G, then it induces a map of G to a group G', and projects to a map $S \to S'$. Therefore $T(\mathbb{H})$ is known as the *universal Teichmüller space*.

There are several different natural parametrizations of $T(\mathbb{H})$. One comes from considering the normalized quasiconformal maps f^{μ}. Another comes from considering certain modifications f_μ and their Schwarzian derivatives. This is the subject of Section 9.5. In Section 9.6, we indicate briefly how the theory proceeds from this point.

A very active area that has its roots in Teichmüller theory has come to be known as “higher Teichmüller theory.” A brief description is given in Section 9.8.

9.1 Coverings, quotients, and moduli of compact Riemann surfaces

Here, we summarize results from Chapter 6 and Chapter 7, and indicate the direction of further developments. We then sketch proofs that the moduli space for compact surfaces of genus 1 has complex dimension one, and that the moduli space for compact surfaces of genus $g \geq 1$ has real dimension $6g - 6$.

• As noted before, up to conformal equivalence, the only simply connected Riemann surfaces are the plane $\mathbb{C}$, the Riemann sphere $\mathbb{S}$, and the unit disk $\mathbb{D}$. In place of $\mathbb{D}$, we will work with the conformally equivalent upper half-plane $\mathbb{H}$.

• Any Riemann surface S has a simply connected universal covering surface: a Riemann surface S^u and a projection π of S^u onto S with the properties that π is holomorphic, and that each point $p \in S$ has an open neighborhood U such that $\pi^{-1}(U)$ consists of disjoint open sets, each of which is mapped conformally to U by π. Moreover, any closed curve γ beginning and ending at p lifts to a curve $\widetilde{\gamma}$ in S^u that begins and ends in $\pi^{-1}(p)$. The lift is closed if and only if γ is homotopic to a constant curve. A Riemann surface S is said to be *elliptic* if $S^u \cong \mathbb{S}$, *parabolic* if $S^u \cong \mathbb{C}$, and *hyperbolic* if $S^u \cong \mathbb{H}$.

• Let $\mathrm{Aut}(S^u)$ be the group of conformal self-maps of S^u. The group $\mathrm{Aut}(S^u)$ contains a subgroup G, the group of *cover transformations* (often called *deck transformations*, from *decken*, “to cover”). The automorphisms g belonging to G are characterized by the property that

$$\pi \circ g \;=\; \pi. \tag{9.1.1}$$

Thus, a *point* p of S corresponds to the *orbit* under G of any point in $\pi^{-1}(p)$, or, equivalently, to the quotient of S^u by the equivalence relation induced by G. This quotient is usually denoted $G\backslash\mathrm{Aut}(S^u)$.

Since S^u can be taken to be one of $\mathbb{C}$, $\mathbb{S}$, or $\mathbb{H}$, the automorphism group $\mathrm{Aut}(S^u)$ is isomorphic to a group of linear fractional transformations, and has a natural topology. The group G is a properly discontinuous subgroup of $\mathrm{Aut}(S^u)$, and $g \in G$ has no fixed point unless g is the identity. Such a group is called a *Fuchsian group*. Conversely, if $G \subset \mathrm{Aut}(S^u)$ is such a group, then $G\backslash S^u$ is a Riemann surface.

• The cover transformations can be generated as follows. The covering surface itself is constructed by choosing a point p_0 in S. The elements of S^u are equivalence classes $[\gamma_p]$ of curves γ_p from p_0 to p, two such curves being equivalent if they are homotopic. Then each closed curve γ begining and ending at p generates a cover transformation $g_{[\gamma]}$ by $g_{[\gamma]}([\gamma_p]) = [\gamma_p \cdot \gamma]$. This is a homomorphism from the fundamental group $H_1(S)$ into the automorphism group $\mathrm{Aut}(S^u)$.

- In the elliptic case, $S^u \cong \mathbb{S}$, each element of $\mathrm{Aut}(\mathbb{S})$ has a fixed point. It follows that G contains only the identity map, and $S = S^u$. In the parabolic case, $S^u \cong \mathbb{C}$, the fixed-point-free automorphisms are translations and the group of cover transformations is generated by either one or two translations: Proposition 6.3.3.

Theorem 9.1.1. *Suppose that $S_1 = G_1 \backslash S_1^u$ and $S_2 = G_2 \backslash S_2^u$.*
(a) If S_1 and S_2 are conformally equivalent, then any conformal equivalence $\phi : S_1 \to S_2$ can be lifted to a conformal equivalence $\widetilde{\phi}$ of the universal covers S_j^u.
(b) The lift $\widetilde{\phi}$ induces a isomorphism of the groups G_1 and G_2 of cover transformations.
(c) Any other lift of ϕ to a conformal map of S_1^u onto S_2^u has the form $h \circ \widetilde{\phi}$, for some $h \in G_2$.

Proof: The proof of Theorem 6.2.4 can be adapted to prove (a). In fact, fix points z_0 in S_1^u and w_0 in $\pi_2^{-1} \circ \phi \circ \pi_1(z_0)$. If γ is a path in S_1^u from z_0 to z, take $\widetilde{\phi}(z)$ to be the endpoint of the path obtained by lifting $\phi \circ \pi_1 \circ \gamma$ from w_0. Any other such path γ' is homotopic to γ, so $\phi \circ \pi_1 \circ \gamma'$ lifts to the same endpoint. The result can be illustrated by the commutative diagram

$$\begin{tikzcd}
S_1^u \arrow[r, "\widetilde{\phi}"] \arrow[d, "\pi_1"'] & S_2^u \arrow[d, "\pi_2"'] \\
S_1 \arrow[r, "\phi"] & S_2.
\end{tikzcd} \tag{9.1.2}$$

More concisely,

$$\phi \circ \pi_1 = \pi_2 \circ \widetilde{\phi}. \tag{9.1.3}$$

To prove (b), we let $\theta(g)$ be defined by

$$\theta(g) = \widetilde{\phi} \circ g \circ \widetilde{\phi}^{-1}, \qquad g \in G_1.$$

Then $\theta(g)$ is a conformal map of S_2^u onto S_2^u. We need to show that $\theta(g)$ belongs to G_2, i.e. that $\pi_2 \circ \theta(g) = \pi_2$. But (9.1.3) and (9.1.1) imply that

$$\begin{aligned} \pi_2 \circ \theta(g) = \pi_2 \circ \widetilde{\phi} \circ g \circ \widetilde{\phi}^{-1} &= \phi \circ \pi_1 \circ g \circ \widetilde{\phi}^{-1} \\ = \phi \circ \pi_1 \circ \widetilde{\phi}^{-1} &= \pi_2. \end{aligned}$$

Conversely, for $h \in G_2$, let $g = \widetilde{\phi}^{-1} \circ h \circ \widetilde{\phi}$. Then

$$\begin{aligned} \pi_1 \circ g = \pi_1 \circ \widetilde{\phi}^{-1} \circ h \circ \widetilde{\phi} &= \phi^{-1} \circ \pi_2 \circ h \circ \widetilde{\phi} \\ = \phi^{-1} \circ \pi_2 \circ \widetilde{\phi} &= \pi_1. \end{aligned}$$

Thus, θ is an isomorphism from G_1 to G_2.

For any $h \in G_2$, $h \circ \widetilde{\phi}$ is also a lift. Conversely, suppose f is a lift of ϕ. Let $h = f \circ \widetilde{\phi}^{-1}$. Then

$$\pi_2 \circ h \;=\; \phi \circ \pi_1 \circ \widetilde{\phi}^{-1} \circ h \;=\; \pi_2 \circ \widetilde{\phi} \circ \widetilde{\phi}^{-1} \circ h \;=\; \pi_2,$$

so h is in G_2. □

Note that one consequence is that conformally equivalent Riemann surfaces have conformally equivalent universal covers.

Corollary 9.1.2. *Discrete groups G_1, G_2 of fixed-point-free automorphisms of S^u, where S^u is $\mathbb{C}$ or $\mathbb{H}$, generate conformally equivalent Riemann surfaces if and only if they are related by an inner automorphism of* $\mathrm{Aut}(S^u)$*, i.e.*

$$g \in G_1 \;\to\; h \circ g \circ h^{-1} \tag{9.1.4}$$

for some h in $\mathrm{Aut}(S^\mu)$.

Proof: Suppose that $\phi : S_1 \to S_2$ is a conformal equivalence. We may identify both S^u_j with $\mathbb{C}$ or $\mathbb{H}$, as the case may be, and let $\widetilde{\phi}$ be the lift of ϕ. Then for $g \in G_1$

$$\pi_2 \circ \widetilde{\phi} \circ g \circ \widetilde{\phi}^{-1} \;=\; \phi \circ \pi_1 \circ \widetilde{\phi}^{-1} \;=\; \phi \circ \phi^{-1} \circ \pi_2 \;=\; \pi_2,$$

so $\widetilde{\phi} \circ g \circ \widetilde{\phi}^{-1}$ belongs to G_2. The mapping is clearly a homomorphism, and a similar calculation shows that the inverse $h \to \widetilde{\phi}^{-1} \circ h \circ \widetilde{\phi}$ maps G_2 to G_1.

Conversely, suppose that S_1 is defined by the automorphism group $G_1 \subset \mathrm{Aut}(S^u)$, and suppose that f belongs to $\mathrm{Aut}(S^u)$. Let $G_2 = h \circ G_1 \circ h^{-1}$ be the groups defined by the map $g \to h \circ g \circ h^{-1}$. This defines a Riemann surface S_2. The map $\phi = \pi_2 \circ h \circ \pi_2^{-1}$ is easily seen to be well defined and holomorphic, with inverse $\pi_2 \circ h \circ \pi_2^{-1}$. Therefore it is conformal. □

Let us return to the parabolic case. We may take $S = G\backslash\mathbb{C}$. We know that any element of G is a translation $T_a = z + a, a \neq 0$. As noted above, by Proposition 6.3.3 G has either one or two generators. Suppose first that G has the single generator T_a. It is easily verified that any other translation can be obtained as BTB^{-1} for some choice of $B \in \mathrm{Aut}(C)$, so all such surfaces are conformally equivalent. Note that they are not compact: topologically they are open cylinders obtained by identifying z and $z + a$.

Suppose now that G is generated by two translations, T_a and T_b. In the proof of Proposition 6.3.3, it was shown that b/a is not real. The resulting surface is a torus that is obtained by identifying opposite sides of a period parallelogram of a lattice, as in Figure 6.3.

The parameter space in this case can be taken to be the plane $\mathbb{C}$. An analytic proof of this fact takes some heavy machinery. Let us denote the generators of the lattice by ω_1, ω_2 rather than a, b, with the standard normalization $\mathrm{Im}\,(\omega_2/\omega_1) > 0$.

Suppose that A is an automorphism of $\mathbb{C}$ that takes one such period parallelogram Λ_1 bijectively to another, Λ_2. Then

$$A\omega_j \;=\; m_{j1}\,\omega_1' + m_{j2}\,\omega_2'; \qquad A^{-1}\omega_j' \;=\; n_{j1}\,\omega_1 + n_{j2}\,\omega_2,$$

$j = 1, 2$, where the m_{jk} and n_{jk} are integers. Thus, A corresponds to a matrix in the *modular group* $SL(2, \mathbb{Z})$ of invertible matrices with integer coefficients and determinant 1. More precisely, A belongs to the projective version $PSL(2, \mathbb{Z})$: the quotient of $SL(2, \mathbb{C})$ by the subgroup $\{\mathbf{1}, -\mathbf{1}\}$.

Thus, the moduli space for conformal structures on the torus is the quotient of $\mathbb{H}$ by $PSL(2, \mathbb{Z})$. Some elementary calculations show that, by choosing periods ω_1 and ω_2 with $|\omega_j|$ minimal, we can guarantee that $\tau = \operatorname{Im}(\omega_2/\omega_1)$ belongs to the region

$$\Delta \;=\; \{\tau : -1/2 < \operatorname{Re}\tau \le 1/2;\ |\tau| \ge 1;\ \operatorname{Re}\tau \ge 0 \ \text{if}\ |\tau| = 1\};$$

see Figure 2.2. A further argument shows that this choice of $\tau \in \Delta$ is unique. Therefore we may take Δ itself as the moduli space for conformal structures on a topological torus. Finally, the J-function, a holomorphic function on $\mathbb{H}$ that is invariant under the modular group (and is closely related to the elliptic modular function λ discussed in Section 2.5) maps Δ conformally onto $\mathbb{C}$. (For details, see Hille [107], Section 13.6, or [22], Chapter 17.)

To conclude this section, we sketch an argument concerning the moduli of a compact surface S of genus $g > 1$. As discussed in Section 11.2, cycles (simple closed curves) α_j, β_j, $j = 1, 2, \ldots g$, can be chosen in such a way that the only intersections are a single intersection of each pair α_j, β_j, and such that if S is cut along these cycles, the result is topologically a $4g$-sided polygon whose boundary consists of the sides in the order

$$\alpha_1, \beta_1, \alpha_1^{-1}, \beta_1^{-1}, \ldots, \alpha_g, \beta_g, \alpha_g^{-1}, \beta_g^{-1}.$$

By identifying the sides α_j, α_j^{-1} and the sides β_j, β_j^{-1} one reconstructs, topologically, the surface S. Figure 11.2 illustrates this in the case $g = 2$. Choose a base point in S, and g curves that join p_0 to the points of intersection of the cycles. The closed curve that starts at p_0, follows α_j and returns to p_0 determines an automorphism A_j of $S^u \cong \mathbb{H}$, and similarly for β_j and B_j: Theorem 6.2.4. The cycles α_j, β_j generate the fundamental group of S, so the A_j and B_j generate the group of cover transformations of $\mathbb{H}$ over S.

Each of these $2g$ transformations corresponds to a linear fractional transformation with real entries and determinant 1, unique up to multiplication by -1. Therefore the particular surface S has at most $6g$ "degrees of freedom." Looking at the boundary of the $4g$-gon, we see that

$$A_1B_1A_1^{-1}B_1^{-1} \cdots A_gB_gA_g^{-1}B_g^{-1} \;=\; I, \tag{9.1.5}$$

the identity transformation. The product on the left in (9.1.5) has determinant 1, so (9.1.5) puts 3 constraints on the A_j, B_j. Moreover, if B is any element of $\operatorname{Aut}(\mathbb{H})$, then the map $A \to B^{-1}A\mathrm{B}$, $A \in \Gamma$ is an automorphism of G, allowing us to eliminate 3 more degrees of freedom from the remaining $6g - 3$, leaving $6g - 6$.

To this point, the argument only gives $6g - 6$ as an upper bound for the dimension of the parameter space. However, the cycles α_j, β_j are determined only up to homotopy and intersection conditions, so the associated matrices have some room to move. Thus, the parameter space does actually have 6g-6 real dimensions as a subset of $\mathbb{R}^{6g}$ (modulo an equivalence relation corresponding to the freedom of multiplying any of the matrices by -1).

Riemann's argument, unlike this one, produced $3g - 3$ *complex* parameters. However, neither argument gives a clear geometric picture. For example, is there any natural parametrization that leads to an open ball in $\mathbb{R}^{6n-6}$ or in $\mathbb{C}^{3m-3}$ as the parameter set? This is the type of question that led to Teichmüller's work.

9.2 Homeomorphisms of Riemann surfaces

The case of annuli, considered in the light of Exercise 11 of Chapter 8, suggests that one way to relate inequivalent Riemann surfaces, and measure the extent to which they differ, is by the use of quasiconformal maps. In this section, we do not need to make use of quasiconformality, so we deal with general homeomorphisms (always assumed to be orientation-preserving).

The proof of Theorem 9.1.1 can be extended to give the following.

Theorem 9.2.1. *Suppose that $S_1 = G_1 \backslash S_1^u$ and $S_2 = G_2 \backslash S_2^u$.*
(a) If f is a homeomorphism of S_1 onto S_2, then f can be lifted to a homeomorphism $\widetilde{f}$ of the universal cover S_1^u onto the universal cover S_2^u. Moreover, if f is K-quasiconformal, the same is true of the lift $\widetilde{f}$.
(b) The map

$$\theta_f(g) = \widetilde{f} \circ g \circ \widetilde{f}^{-1}$$

is an isomorphism of G_1 onto G_2.
Any other lift of f has the form $h \circ \widetilde{f}$ for some $h \in \mathrm{Aut}(S_2^u)$.

Proof. Pictorially, we have again

$$\begin{tikzcd} S_1^u \arrow[r, "\widetilde{f}"] \arrow[d, "\pi_1"'] & S_2^u \arrow[d, "\pi_2"] \\ S_1 \arrow[r, "f"] & S_2, \end{tikzcd} \tag{9.2.1}$$

so again

$$\pi_2 \circ \widetilde{f} = f \circ \pi_1. \tag{9.2.2}$$

Since f maps closed curves in S_1 to closed curves in S_2, composition with $\widetilde{f}$ maps cover transformations to cover transformations. The map

$$\theta_f(g) \;=\; \widetilde{f} \circ g \circ \widetilde{f}^{-1}, \qquad g \in \mathrm{Aut}(S_1^u)$$

belongs to $\mathrm{Aut}(S_2^u)$.

Given $g \in G_1$, let $g_2 \;=\; \theta_f(g_1)$. Then, referring to (9.2.2), we see that again

$$\begin{aligned}\pi_2 \circ g_2 = \pi_2 \circ \widetilde{f} \circ g_1 \circ \widetilde{f}^{-1} \;&=\; f \circ \pi_1 \circ g_1 \circ \widetilde{f}^{-1}\\ = f \circ \pi_1 \circ \widetilde{f}^{-1} \;&=\; \pi_2.\end{aligned}$$

Thus, g_2 belongs to G_2. Clearly, θ_f is a homomorphism, and $\theta_{f^{-1}}$ inverts it. □

Since the universal covers of homeomorphic Riemann surfaces are homeomorphic and, in the interesting case, equivalent to $\mathbb{C}$ or $\mathbb{H}$, we may always take a Riemann surface S to be $G\backslash S^u$ with S^u equal $\mathbb{C}$ or $\mathbb{H}$, and with covering group G a subgroup of $\mathrm{Aut}(S^u)$.

Theorem 9.2.2. *Suppose that S and S' are homeomorphic Riemann surfaces. Two homomorphisms $f_0 : S \to S'$ and $f_1 : S \to S'$ induce the same isomorphism of the covering groups if and only if they are homotopic.*

Proof. Suppose that f_0 and f_1 induce the same isomorphism, and suppose that $S^u = \mathbb{H}$. We define a homotopy $\{\widetilde{f_t}\}$ from $\widetilde{f_0}$ to $\widetilde{f_1}$ by taking $\widetilde{f_t}(z)$, $z \in \mathbb{H}$, to be the point on the (hyperbolic) line segment from $\widetilde{f_0}(z)$ to $\widetilde{f_1}(z)$ such that

$$\rho_{\mathbb{H}}(\widetilde{f_0}(z), \widetilde{f_t}(z)) \;=\; t\, \rho_{\mathbb{H}}(\widetilde{f_0}(z), \widetilde{f_1}(z)),$$

where $\rho_{\mathbb{H}}$ is the hyperbolic distance. We want to show that the projection

$$f_t \;=\; \pi_2 \circ \widetilde{f_t} \circ \pi_1^{-1}$$

is a well-defined map. But $g \in G$ is an isometry, so if $w = g(z)$, then $f_t(w) = g(f_t(z))$, i.e. $\widetilde{f_t} \circ g = g \circ \widetilde{f_t}$. Therefore f_t is well defined and $\{f_t\}$ is a homotopy from f_0 to f_1. If $S^u = \mathbb{C}$, the same argument works, with the euclidean metric in place of the hyperbolic metric.

Conversely, suppose that $\{f_t\}$ is a homotopy from f_0 to f_1. It can be lifted to a homotopy $\{\widetilde{f_t}\}$ of $\widetilde{f_0}$ to $\widetilde{f_1}$. For $g \in G$ and $z \in \mathbb{H}$, the paths

$$t \to \widetilde{f_t} \circ g(z), \qquad t \to \widetilde{f_0} \circ g \circ \widetilde{f_0}^{-1}(\widetilde{f_t}(z))$$

have the same starting point $\widetilde{f_0}(g(z))$ in $\mathbb{H}$ and the same projections $\pi_2 \circ \widetilde{f_t}(z)$ on S_2. Therefore they agree at $t = 1$, giving

$$\widetilde{f_0} \circ g \circ \widetilde{f_0}^{-1} \;=\; \widetilde{f_1} \circ g \circ \widetilde{f_1}^{-1}, \qquad \text{all } g \in G.$$

Thus, f_0 and f_1 give the same isomorphism. □

The following is an equivalent way to state Theorem 9.2.2. We leave the proof as Exercise 1.

Corollary 9.2.3. *Two homeomorphisms $f_j : S \to S'$, $j = 0, 1$, induce the same isomorphisms $G \to G'$ if and only if $f_1^{-1} \circ f_0$ is homotopic to the identity map of S'*

In anticipation of the standard construction of Teichmüller spaces, let us push this idea one step further.

Corollary 9.2.4. *Suppose that S, S_1, and S_2 are homeomorphic compact Riemann surfaces with cover groups G, G_1, and G_2 in $\mathrm{Aut}(S^u)$. Suppose also that $f_j : S \to S_j$, $j = 1, 2$, are quasiconformal homeomorphisms.*
(a) Suppose that $f_1 \circ f_2^{-1} : S_2 \to S_1$ is homotopic to a conformal map ϕ. Then the lifts $\widetilde{f_1 \circ f_2^{-1}}$ and $\widetilde{\phi}$ induce the same isomorphism of G_2 to G_1.
(b) Suppose that $\phi : S_2 \to S_1$ is a conformal map. Then $\widetilde{f_1}$ and $\widetilde{\phi} \circ \widetilde{f_2}$ induce the same isomorphism of G to G_1 if and only if $f_1 \circ f_2^{-1}$ is homotopic to ϕ.

Proof: (a) Under this assumption, f_1 and $f = \phi \circ f_2$ are homotopic homeomorphisms from S to S_1, so $\widetilde{f_1}$ and $\widetilde{f}$ induce the same isomorphism from G to G_1.

(b) Under this assumption, $\widetilde{\phi}$ induces an isomorphism of G_2 to G_1. Again, let $f = f_2 \circ \phi$. Then $\widetilde{f_1}$ and $\widetilde{f}$ induce the same isomorphism if and only if the map

$$f_1 \circ f^{-1} = (f_1 \circ f_2^{-1}) \circ \phi^{-1}$$

is homotopic to the identity, which is true if and only if $f_1 \circ f^{-1}$ is homotopic to ϕ. □

There is a different way to phrase this result.

Proposition 9.2.5. *Suppose that S, S_1, and S_2 are Riemann surfaces, and $f_j : S \to S_j$ are homeomorphisms. Then S_1 and S_2 are conformally equivalent if and only if $f_1 \circ f_2^{-1} : S_2 \to S_2$ is homotopic to a conformal map.*

9.3 Homeomorphisms of compact Riemann surfaces

The discussion in the previous section shows that in considering homeomorphisms from one Riemann surface to another, we are principally interested in homotopy types. In the introduction to this chapter, we suggested that one should focus on quasiconformal maps. For compact surfaces, up to homotopy, we lose nothing by restricting to quasiconformal homeomorphisms.

In preparation for the proof of the last statement, we recall some facts about the Beurling–Ahlfors extension in Theorem 8.7.4, transplanted, via the Cayley transform, to an extension of an orientation-preserving homeomorphism h of $\partial\mathbb{D}$ (not necessarily normalized). The homeomorphism f of $\mathbb{D}$ constructed from h is C^1 on $\mathbb{D}$, with positive Jacobian, so it is quasiconformal on each compact subset of $\mathbb{D}$. (The maximal dilatation will grow indefinitely as one approaches some boundary point unless h satisfies a quasi-symmetry condition.)

Theorem 9.3.1. *Suppose that the Riemann surfaces S and S' are compact, and f is a homeomorphism of S onto S'. Then f is homotopic to a quasiconformal homeomorphism of S onto S'.*

Proof: We may choose domains $U_j, j = 1, 2, \dots n$, in S such that: (a) there are conformal maps $\psi_j : \mathbb{D}$ onto U_j; (b) these maps extend to the boundary; (c) for some $0 < r < 1$ the images $V_j = \psi_j(D_r(0))$ cover S and have the property that successive intersections $V_{j-1} \cap V_j$ are not empty. Define successive maps $f_0 = f$, and $f_j = f_{j-1}$ on $S \setminus U_j$. Extend f_j to U_j as the Beurling–Ahlfors extension of $f_{j-1}|_{\partial U_j}$. Then f_{j-1} and f_j agree except on U_j. We define a homotopy on U_j by

$$f_{t,j} = (1-t)f_{j-1} + tf_j.$$

Thus, $f = f_0$ is homotopic to each f_k. By construction, f_k is quasiconformal on $\bigcup_{j \le k} V_j$, so $f = f_0$ is homotopic to the quasiconformal map $f_n : S \to S'$. □

We assume throughout this section that all surfaces are hyperbolic: they have $\mathbb{H}$ as universal cover. One question is: how to recognize the compact case among all the cases $G\backslash\mathbb{H}$? We start by looking for a fundamental domain for G: a domain $\Omega \subset \mathbb{H}$ that is minimal with respect to the condition that $\mathbb{H}$ is covered by the closures (in $\mathbb{H}$) of the images $g(\Omega)$, $g \in G$. Equivalently, the condition is that $\pi(\overline{\Omega}) = S$ and Ω is minimal among domains with this property.

Suppose that $G \subset \mathrm{Aut}(\mathbb{H})$ is a properly discontinuous group, whose non-identity elements have no fixed points. A *Dirichlet domain* for G is defined by choosing a point $a \in \mathbb{H}$ and defining N_a to be set of points of $\mathbb{H}$ that are closer to a than to any of its images $g(a)$, for non-identity element $g \in G$:

$$N = N_a = \{z \in \mathbb{H} : \rho_{\mathbb{H}}(z,a) < \rho_{\mathbb{H}}(z, g(a)),\ \text{all } g \in G,\ g \neq \mathbf{1}\}, \qquad (9.3.1)$$

where $\rho_{\mathbb{H}}$ is again the hyperbolic metric in $\mathbb{H}$. The point a is called the *center* of N_a. Clearly, N_a is open and non-empty. If $b = g(a)$ for some $g \in G$, $g \neq \mathbf{1}$, then N_a and N_b are disjoint.

Lemma 9.3.2. *The closure $\overline{N}$ of N_a in $\mathbb{H}$ projects onto S.*

Proof: Every point of $\mathbb{H}$ either lies in some $N_{g(a)}$ for some $g \in G$, or lies on the boundary of (at least) two of these domains. Therefore the closures $\{\overline{N_{g(a)}}\}$ cover $\mathbb{H}$. Any two such closures project to the same set, so $\pi(\overline{N}) = S$. □

The elements of $\mathrm{Aut}(\mathbb{H})$ are isometries with respect to the hyperbolic metric, so the N_b, for $b = g(a)$, $g \in G$, are each congruent to $N = N_a$. In particular, either no N_b has finite diameter, or they all have the same finite diameter. This means that each has a boundary point (with respect to $\mathbb{C}$) on $\mathbb{R}$, or none do. This dichotomy is independent of the choice of starting point a, as shown by the following.

Theorem 9.3.3. *The Dirichlet domains of a Riemann surface $S = G\backslash\mathbb{H}$ are bounded (with respect to the hyperbolic metric) if and only if S is compact.*

Proof: Suppose that S is compact. Choose $a \in \mathbb{H}$ and let D_n be the hyperbolic disk

$$D_n = \{z \in \mathbb{H} : \rho_{\mathbb{H}}(a, z) < n\}.$$

The projections $\pi(D_n)$ are open sets that cover S, so there is some m such that $D_m = S$. Thus, for every $z \in \mathbb{H}$, there is a $g \in G$ such that $\rho_{\mathbb{H}}(a, g(z)) < m$. Equivalently, N_a has diameter $< 2m$. Conversely, if N_a is bounded, then its closure $\overline{N}$ is compact. Therefore $S = \pi(\overline{N})$ is also compact. □

We do not need the following description of Dirichlet domains, but it is easily established in the compact case. For convenience, we refer to the hyperbolic lines (geodesics) simply as "lines," and intervals on geodesics as "segments."

Proposition 9.3.4. *Any bounded Dirichlet domain N is a (hyperbolic) convex polygon, i.e. the boundary is the union of finitely many segments that meet at interior angles $< \pi$. The number of such sides is even, and there is a natural pairing of opposite sides.*

Proof: Any point on the boundary of $N = N_a$ is the midpoint of the line joining a to a point $b = g(a)$ some $g \in G$. Since there is a bound to the distance from a to the boundary of N, there is a bound to the distance from a of the b that can occur in this way. Therefore there are finitely many. The set of points equidistant from a and b is a line; see Exercise 2. Therefore the boundary of N is a union of finitely many segments E_j. Wherever two such edges meet, N lies in the intersection of the half-planes determined by the lines that contain these edges, so the interior angle is $< \pi$.

Finally, if E_j is associated to $g_j(a)$, $g_j \in G$, then reflecting through a takes E_j to the side associated to $g_j^{-1}(a)$. □

Corollary 9.3.5. *In the compact case, G is finitely generated.*

Proof: As noted in the preceding proof, each side E_j of a Dirichlet region N_a corresponds to a point $b_j = g_j(a)$, $g_j \in G$. We claim that these finitely many elements $\{g_j\}$ generate G. In fact, g_j maps N_a to N_{b_j}, $b_j = g_j(a)$. This image is a reflection through E_j. It follows that $g_k \circ g_j$ maps the the $g_j(E_j)$ to the corresponding side of N_{b_j}. Clearly, any of the images under G of N_a can be reached in this way, by means of an element of the group G' generated by the g_j. There is a bijective correspondence between these images and the elements of G, so $G' = G$. □

For an example of all this, set in the disk $\mathbb{D}$ rather than in $\mathbb{H}$, see Figure 9.2. The yellow region in the disk on the left is the Dirichlet region N_0 with center 0 for the Bolza curve, a compact genus 2 surface. The right side of the figure shows the decomposition of $\mathbb{D}$ into Dirichlet regions congruent to the region on the left.

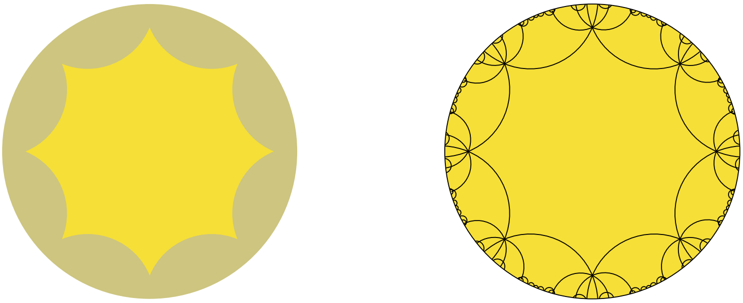

Fig. 9.2 Dirichlet regions for the Bolza curve.

Lemma 9.3.6. *Suppose that the Riemann surface $S = G\backslash\mathbb{H}$ is compact. Then the set of $x \in \mathbb{R} \cup \{\infty\}$ such that x is a fixed point of some $g \in G$, $g \neq \mathbf{1}$ contains at least three points.*

Proof: Suppose that there are at most two such points. Replacing G by hGh^{-1} for some $h \in \mathrm{Aut}(G)$ yields an equivalent surface. Assuming that there is only one such point, choose h so that the fixed point is the point at ∞. Then G consists of translations $g_b(z) = z + b, b \in \mathbb{R}$. Since G is discrete, there is a minimal such $b > 0$. Then $S = G\backslash\mathbb{H}$ is homeomorphic to the vertical strip

$$\{z \in \mathbb{H} \,:\, -b/2 \leq \mathrm{Re}\, z \leq b/2\}$$

with the edges $|\mathrm{Re}\, z| = b/2$ identified. Thus, S is an open cylinder, contradicting the assumption that S is compact.

Assuming that there are exactly two fixed points, we may take them to be 0 and ∞.Then each element of G is a map $z \to az$ for some $a > 0$. Since G is discrete, there is a smallest such $a > 1$. Then

$$G \;=\; \{g_n(z) = a^n z, \;\; n = 0, \pm 1, \pm 2, \ldots\}.$$

The associated surface $G\backslash\mathbb{H}$ can be taken to be the closure of the annulus $A(1, a)$ with the inner and outer boundaries identified. This is a torus, so the universal cover would be conformal to $\mathbb{C}$ rather than to $\mathbb{H}$. □

It is useful to be able to extend a homeomorphism $\mathbb{H} \to \mathbb{H}$ continuously to a homeomorphism of the closure $\mathbb{H} \cup \mathbb{R}$. This is not always possible; see Exercise 3. However, we know from Section 8.6 that it is possible for quasiconformal maps, and we also know that the lift of a K-quasiconformal homeomorphism of Riemann surfaces is a K-quasiconformal map of the covering spaces.

Theorem 9.3.7. *Suppose that $S = G\backslash\mathbb{H}$ and $S' = G'\backslash\mathbb{H}$ are compact. Two quasiconformal homeomorphism f_1 and f_2 from S to S' induce the same isomorphism of G and G' if and only if there are lifts $\widetilde{f_1}$, $\widetilde{f_2}$ that coincide on $\mathbb{R}$.*

Proof: Suppose first that such lifts induce the same homomorphism. The $\widetilde{f_j}$ map fixed points of G to fixed points of G'. Therefore for any $g \in G$, the maps

$$\widetilde{f_1} \circ g \circ \widetilde{f_1}^{-1}, \qquad \widetilde{f_2} \circ g \circ \widetilde{f_2}^{-1}$$

agree on $\mathbb{R}$. Since both maps belong to $\mathrm{Aut}(\mathbb{H})$, they must be identical.

Conversely, suppose that $\widetilde{f_1}$ and $\widetilde{f_2}$ induce the same isomorphism, i.e.

$$\widetilde{f_1} \circ g \circ \widetilde{f_1}^{-1} = \widetilde{f_2} \circ g \circ \widetilde{f_2}^{-1}, \qquad \text{all } g \in G.$$

Let $\psi = \widetilde{f_1}^{-1} \circ \widetilde{f_2}$, so that

$$g^n \circ \psi = \psi \circ g^n, \qquad g \in G, \quad n = 0, \pm 1, \pm 2, \dots.$$

If x is fixed by $g \in G$, then so is $\psi(x)$. Now $g^n(z) \to x$ either as $n \to \infty$ or as $n \to -\infty$ (or both), so, for the correct choice of sign, we have

$$\psi(x) = \lim \psi(g^n(z)) = \lim g^n(\psi(z)) = x.$$

Thus, each fixed point of G is a fixed point of ψ. Since there are at least three such points, $\psi = \mathbf{1}$. □

Remarks. Theorem 9.3.7 is true for a much wider class of hyperbolic Riemann surfaces, classified according to the properties of the covering group G. To be specific here, we need some definitions. The *limit set* of a Fuchsian group G is the set L of points $x \in \{\mathbb{R} \cup \{\infty\}$ with the property that there is a sequence of points $\{z_n\} \subset \mathbb{H}$ and a sequence of $\{g_n\}$ of distinct elements of G such that $g_n(z_n) \to x$. In particular, any fixed point of an element $g \in G$ belongs to L. The Fuchsian group G is said to be of the *first kind* if the limit set L is all of $\mathbb{R} \cup \{\infty\}$. Otherwise, G is said to be of the *second kind*. Theorem 9.3.7 carries over to any G of the first kind. It is not difficult to show that, in the compact case, G is of the first kind; see Exercise 5. For any such group, the set of fixed points is dense in $\mathbb{R}$.

For a full treatment of this topic, see Lehner [133].

9.4 The Teichmüller space of a Riemann surface

to consider only the hyperbolic case: Riemann surfaces that can be taken to be of the form $G\backslash\mathbb{H}$. Throughout this section, we also assume that the Riemann surfaces under consideration are of the form $G\backslash\mathbb{H}$ with covering group G that is of the first kind. As noted in the previous section, this includes all compact hyperbolic surfaces.

The *deformation space* $\mathrm{Def}(S)$ of a Riemann surface S is defined to be the collection of pairs (S', f), where S' is a Riemann surface and $f : S \to S'$ is a quasiconformal homeomorphism. The *Teichmüller space* $T(S)$ is defined to be the quotient of $\mathrm{Def}(S)$ by a certain equivalence relation $\sim$:

$$T(S) = \frac{\mathrm{Def}(S)}{\sim} \tag{9.4.1}$$

the relation $\sim$ is defined as follows:

$$\begin{aligned} &f_1 \sim f_2 \quad \text{if and only if} \quad f_2 \circ f_1^{-1} : S_1 \to S_2 \\ &\text{is homotopic to a conformal map } \phi : S_1 \to S_2. \end{aligned} \tag{9.4.2}$$

In light of Proposition 9.2.5, this is the same as saying that the images S_1 and S_2 are conformally equivalent.

Proposition 9.4.1. *If f belongs to* $\mathrm{Def}(S)$*, then the equivalence class* $[f]$ *contains an extremal: a map f_0 whose maximal dilatation is minimal:*

$$K_{f_0} = \inf_{g\sim f} K_g.$$

The proof is left as Exercise 4.

The *Teichmüller distance* D_T between two functions f, g in $\mathrm{Def}(S)$ is defined to be

$$D_T(f, g) = \frac{1}{2}\log K_{g\circ f^{-1}}.$$

Thus, $D_T(f, g) \geq 0$ and $D_T(f, g) = 0$ if and only if f and g are conformally equivalent. Moreover, $D_T(f, g) = D_T(g, f)$, since $K_h = K_{h^{-1}}$.

The *Teichmüller metric* on $T(S)$ is defined by

$$d_T([f_1], [f_2]) = \inf\{D_T(f, g) : f \sim f_1,\ g \sim f_2\}. \tag{9.4.3}$$

We show next that the term "metric" is justified.

Proposition 9.4.2. *The formula (9.4.3) defines a metric on $T(S)$.*

Proof: Clearly, $d_T([f], [g]) = d_T([g], [f]) \geq 0$. The triangle inequality follows from Proposition 8.2.1 (b):

$$K_{f\circ h^{-1}} \leq K_{f\circ g^{-1}} K_{g\circ h^{-1}}.$$

To complete the argument, we need to show that $d_T([f], [g]) = 0$ implies $[f] = [g]$. If $d_T([f], [g]) = 0$ then, as in the proof of Proposition 9.4.1, there is are sequences

$$\{f_n\} \subset [f], \quad \{g_n\} \subset [g], \quad K_{f_n \circ g_n^{-1}} \to 1$$

that can be lifted to $\mathbb{H}$ and yield limits $\widetilde{f_0}$, $\widetilde{g}_0$ that are lifts of $f_0 \in [f]$, $g_0 \in [g]$. Moreover, $K_{f_0 \circ g_0^{-1}} = 1$. Thus, $f_0 \circ g_0$ is conformal, so $[f_0] = [g_0]$. □

Suppose that f is a K-quasiconformal homeomorphism of S to S'. If $S = G\backslash\mathbb{H}$ and $S' = G'\backslash\mathbb{H}$, then the lift $\widetilde{f}$ is a K-quasiconformal homeomorphism of $\mathbb{H}$. If we normalize it, then by Theorems 8.8.13 and 8.8.15 is f^μ for some unique $\mu = \mu_{\widetilde{f}}$ in the unit ball B of $L^\infty(\mathbb{H})$:

$$B \;=\; \{\mu \in L^\infty(\mathbb{H}) \;:\; ||\mu||_\infty < 1\}. \tag{9.4.4}$$

Note: to simplify notation in the following, we write μ_f for $\mu_{\widetilde{f}}$.

We take advantage of the fact that the maximal dilatation of f is the same as the maximal dilatation of the normalized lift f^μ, and find a formula for $d_T(f, g)$ as follows. Given two quasiconformal homeomorphisms f and g from S to S', let $\widetilde{f}$ and $\widetilde{g}$ be the normalized lifts to $\mathbb{H}$. Let $h = g \circ f^{-1}$, so the normalized lift is $\widetilde{h} = \widetilde{g} \circ \widetilde{f^{-1}}$. The computation (8.8.20) can be rewritten, using $\widetilde{g} \circ \widetilde{f}^{-1}$ in place of g, as

$$\mu_h \circ \widetilde{f} \;=\; \left(\widetilde{f}_z / \overline{\widetilde{f}_z}\right) \cdot \frac{\mu_g - \mu_f}{1 - \overline{\mu_f}\mu_g}. \tag{9.4.5}$$

It follows that, in the distance calculation, we may replace the maximal dilatation of $h = g \circ f$ by that of $\widetilde{h}$, giving

$$\frac{1 + |\mu_h|}{1 - |\mu_h|} \;=\; \frac{|1 - \overline{\mu_f}\mu_g| + |\mu_g - \mu_f|}{|1 - \overline{\mu_f}\mu_g| - |\mu_g - \mu_f|}$$

Compare this to the hyperbolic distance between points $a, b \in \mathbb{D}$,

$$\rho(a, b) \;=\; \frac{1}{2} \log \frac{|1 - \bar{a}b| + |a - b|}{|1 - \bar{a}b| - |a - b|}.$$

Taking the supremum of $D_{g \circ f^{-1}}$ over $\mathbb{D}$ gives the maximal dilatation $K_{\widetilde{g} \circ \widetilde{f}^{-1}} = K_{g \circ f^{-1}}$:

$$D_T(f, g) \;=\; \rho_{\mathbb{D}}(\mu_f, \mu_g)). \tag{9.4.6}$$

In view of Proposition 9.4.1, we have proved the following:

Theorem 9.4.3. *If f and g belong to* Def(S)*, then*

$$d_T([f], [g]) \;=\; \rho_{\mathbb{D}}(\mu_{f_0}, \mu_{g_0}), \tag{9.4.7}$$

where f_0 and g_0 are extremal elements of $[f]$ *and* $[g]$.

These results show that we may study $T(S)$ as a metric space by studying the set of Beltrami coefficients that correspond to lifts from $S = G\backslash\mathbb{H}$. These are characterized among elements of the unit ball (9.4.4) by the condition

$$\mu \circ g \;=\; \mu, \quad \text{all } \; g \in G. \tag{9.4.8}$$

Such a coefficient μ is said to be *extremal* if $\mu = \mu_f$, where f is extremal.

Corollary 9.4.4. *The space $T(S)$ is pathwise connected.*

Proof: It is sufficient to show that every $f \in \mathrm{Def}(S)$ is homotopic to the the identity map. We may take f to be extremal in its equivalence class, and let $\mu = \mu_{\tilde{f}}$. In view of Theorem 9.4.3, it is natural to define μ_t, $0 \le t \le 1$ to be the point

$$\mu_t \;=\; \frac{(1+|\mu|)^t + (1-|\mu|)^t}{(1+|\mu|)^t - (|1-|\mu|)^t} \cdot \frac{\mu}{|\mu|} \tag{9.4.9}$$

on the geodesic from 0 to μ in $\mathbb{D}$. This is a homotopy from the identity map f^0 to f^μ. Since μ satisfies (9.4.8), it follows that μ_t does also. Therefore $\{f^{\mu_t}\}$ is the lift of a homotopy $\{f_t\}$ from the identity map $\mathbf{1}$ of S to the given map $f : S \to S'$. Then $K_{f_t} \;=\; K_f$ and $K_{f\circ f_t^{-1}} = K_f^{1-t}$. If $g \in [f_t]$, then $f_t \circ f_{1-t} \sim g \circ f_{1-t}$, so

$$K_f \;=\; K_f^t K_f^{1-t} \;\le\; K_{g\circ f_{1-t}} \;\le\; K_g K_{f_{1-t}} \;=\; K_g K_f^{1-t}.$$

Therefore $K_{f_t} \le K_g$, showing that f_t is extremal. Thus, $\{[f_t]\}$ is a path in $T(S)$ from the identity map to $[f]$. □

This argument shows that

$$d_T(\mathbf{1}, f) = t\, d_T(\mathbf{1}, f_t) + (1-t)\, d_T(\mathbf{1}, f). \tag{9.4.10}$$

Now it follows from the definition that

$$d_T(f, g) \;=\; d_T(\mathbf{1}, f \circ g^{-1}). \tag{9.4.11}$$

Combining this with the additive property (9.4.10), we can show that for any partition $\{t_j\}$ of the interval $[0, 1]$, we have

$$d_T(\mathbf{1}, f) \;=\; \sum_j d_T(f_{t_j}, f_{t_{j+1}}).$$

Therefore the path $\{[f_t]\}$ is a geodesic in $T(S)$.

The construction here can be generalized. Given extremal elements f_0 and f_1 of $\mathrm{Def}(S)$, let $\mu_t(z)$ be the point on the hyperbolic geodesic from $\mu_0(z) = \mu_{f_0}(z)$ to

$\mu_1(z) = \mu_{f_1}(z)$ such that

$$\rho_{\mathbb{D}}(\mu_0(z), \mu_t(z)) \;=\; t\, \rho_{\mathbb{D}}(\mu_0(z), \mu_1(z)). \tag{9.4.12}$$

Then again μ_t satisfies (9.4.8) and $\{[\pi \circ f^{\mu_t} \circ \pi^{-1}]\}$ is a geodesic path from $[f_0]$ to $[f_1]$ in $T(S)$. This sketch gives the following.

Theorem 9.4.5. *The space $T(S)$ is geodesically convex: for extremal elements f, g in* Def(S)*, the path $t \to \pi \circ f^{\mu_t}$ from $\mu_{\tilde{f}}$ to $\mu_{\tilde{g}}$ in $\mathbb{D}$, where μ_t is defined by (9.4.9), corresponds to a geodesic from $[f]$ to $[g]$ in* Def(S).

We complete this discussion of $T(S)$ with two more results about $T(S)$ as a metric space. The proof of the first of the two results is left as Exercise 7.

Theorem 9.4.6. *The space $T(S)$, with metric (9.4.3), is complete.*

Theorem 9.4.7. *If two surfaces in $\mathscr{S}$ are quasiconformally equivalent, then their Teichmüller spaces are isometric.*

Proof: Suppose $h : S \to S'$ is a quasiconformal homeomorphism. Then $f \to f \circ h^{-1}$ maps the family of quasiconformal self-maps of S to the corresponding family for S'. If $g_j = f_j \circ h^{-1}$, $j = 1, 2$, then

$$g_2 \circ g_1^{-1} \;=\; f_1 \circ f_2^{-1}. \tag{9.4.13}$$

Thus, $[f_1] = [f_2] \in T(S)$ if and only if $[g_1] = [g_2] \in T(S')$. Therefore $f \to f \circ h^{-1}$ is a bijection from $T(S)$ to $T(S')$. It follows from (9.4.13) lifted to $\mathbb{H}$ that this map is an isometry.

9.5 The universal Teichmüller space

We have noted that for all but some well-understood examples, the universal cover of a Riemann surface can be taken to be $\mathbb{H}$, and that any quasiconformal homomorphism of Riemann surfaces can be lifted to a quasiconformal homeomorphism of the covers. For this reason, $T(\mathbb{H})$ is called the *universal Teichmüller space*. The question is: how to define $T(\mathbb{H})$?

We need a stronger definition of equivalence of quasiconformal self-maps. In fact, with the definition (9.4.2), $T(\mathbb{H})$ would consist of a single point:

Proposition 9.5.1. *Any two quasiconformal homeomorphisms of $\mathbb{H}$ to $\mathbb{H}$ are homotopic.*

The proof is left as Exercise 8.

As noted above, each normalized K-quasiconformal homeomorphism of $\mathbb{H}$ to $\mathbb{H}$ is f^{μ} for a unique μ in the unit ball B of $L^{\infty}(\mathbb{H})$. In view of Theorem 9.3.7, it is natural to take the equivalence relation for quasiconformal homeomorphisms f, g of $\mathbb{H}$ to be: $f \sim g$ if there is a homotopy from f to g that is constant on $\mathbb{R}$. However, note that if we simply assume that $f|_{\mathbb{R}} = g|_{\mathbb{R}}$, then the homotopy constructed in (9.4.9) is constant on $\mathbb{R}$. Therefore we may define

$$f \;\sim\; g \quad \text{if and only if} \quad f|_{\mathbb{R}} = g|_{\mathbb{R}}. \tag{9.5.1}$$

Then $T(\mathbb{H})$ is the quotient of the family $\mathscr{F}$ of normalized quasiconformal homeomorphisms $f : \mathbb{H} \to \mathbb{H}$ by the equivalence relation (9.5.1):

$$T(\mathbb{H}) \;=\; \frac{\mathscr{F}}{\sim} \;=\; \frac{\{f^{\mu} \;:\; \mu \in B\}}{\sim} \tag{9.5.2}$$

The metric (9.4.7), as well as Theorems 9.4.5 and 9.4.6 carry over to $T(\mathbb{H})$:

$$d_T([f],[g]) \;=\; \rho_{\mathbb{D}}(\mu_{f_0}, \mu_{g_0}), \tag{9.5.3}$$

where $f_0 \in [f]$ and $g_0 \in [g]$ are extremal.

Theorem 9.5.2. *The space $T(\mathbb{H})$ is geodesically convex and complete.*

We may also identify $T(\mathbb{H})$ with the quotient of B by the an equivalence relation:

$$T(\mathbb{H}) \;\cong\; \frac{B}{\sim}, \qquad \mu \;\sim\; \nu \;\Leftrightarrow\; f^{\mu} \;\sim\; f^{\nu}. \tag{9.5.4}$$

Let QS denote the collection of normalized quasi-symmetric maps of $\mathbb{R}$. The equivalence class of any f^{μ} is uniquely determined by its restriction to $\mathbb{R}$, so we also have

$$T(\mathbb{H}) \;\cong\; QS. \tag{9.5.5}$$

The family $\mathscr{F} = \{f^{\mu}\}$ of normalized quasiconformal homeomorphism of $\mathbb{H}$ is a group under composition, as is QS, and the map $f \;\to\; f|_{\mathbb{R}}$ is a group homomorphism. We may also make B a group by defining

$$\mu \circ \nu \;=\; \mu_{f^{\mu} \circ f^{\nu}}. \tag{9.5.6}$$

After these general remarks, we pass to a construction that leads to a new and very fruitful way to parametrize $T(\mathbb{H})$. Given $\mu \in B$, we can define a new Beltrami coefficient on $\mathbb{C}$ by

$$\mu^*(z) \;=\; \begin{cases} 0, & z \in \mathbb{H} \\ \mu(\bar{z}), & z \in \mathbb{H}^*. \end{cases} \tag{9.5.7}$$

where $\mathbb{H}^*$ is the lower half-plane $\{z \in \mathbb{C} : \operatorname{Im} z < 0\}$. Let $f_{\mu} \;=\; f^{\mu^*}$. Then f_{μ} maps $\mathbb{H}$ conformally onto a domain bounded by the curve $L = f_{\mu}(\mathbb{R})$, and is a sense-reversing quasiconformal map of $\mathbb{H}^*$ onto the other component of the complement of L, with

$$\frac{(f_{\mu})_{\bar{z}}(z)}{(f_{\mu})_z(z)} \;=\; -\mu(z), \qquad z \in \mathbb{H}^*. \tag{9.5.8}$$

Theorem 9.5.3. *If μ and ν belong to B, then $f^\mu \sim f^\nu$ if and only if $f_\mu = f_\nu$ on $\mathbb{H}$.*

Proof: Suppose that $f_\mu = f_\nu$ on $\mathbb{H}$, and therefore on $\mathbb{R}$ as well. The maps of $\mathbb{H}$ defined by

$$g_\mu = f_\mu \circ (f^\mu)^{-1}|_{\mathbb{H}}, \qquad g_\nu \ = \ f_\nu \circ (f^\nu)^{-1}|_{\mathbb{H}} \tag{9.5.9}$$

are conformal and have the same image, so $g_\mu \circ g_\nu^{-1}$ belongs to $\operatorname{Aut}(\mathbb{H})$. This map fixes 0, 1, and ∞, so it is the identity. Therefore $f^\mu = g_\mu^{-1} \circ f_\mu$ and $f^\nu = g_\nu^{-1} \circ f_\nu$ agree on $\mathbb{R}$.

Conversely, suppose that $f^\mu \sim f^\nu$. Define a map $g : \mathbb{C} \to \mathbb{C}$ by

$$g(z) \ = \ \begin{cases} f_\mu \circ f_\nu{}^{-1}(z), & f_\nu(z) \in \mathbb{H} \cup \mathbb{R}; \\ f_\mu \circ (f^\mu)^{-1} \circ f^\nu \circ f_\nu^{-1}(z), & f_\nu(z) \in \mathbb{H}^*. \end{cases}$$

Since f_μ and f_ν agree on $\mathbb{R}$, g is continuous on $f_\nu(\mathbb{R})$ and thus is a homeomorphism of $\mathbb{C}$. Now $g|_{\mathbb{H}}$ is conformal. By (9.5.8),

$$\mu_{f_\mu}|_{\mathbb{H}^*} \ = \ \mu_{f^\mu}^{-1}|_{\mathbb{H}^*}, \qquad \mu_{f_\nu}|_{\mathbb{H}^*} \ = \ \mu_{f^\nu}^{-1}|_{\mathbb{H}^*}.$$

By (9.4.5), $f_\mu \circ (f^\mu)^{-1}$ and $f^\nu \circ f_\nu^{-1}$ are both conformal on $\mathbb{H}^*$. Therefore g is conformal, hence belongs to $\operatorname{Aut}(\mathbb{H})$. But g fixes 0, 1, ∞, so g is the identity. Thus, $f_\mu = f_\nu$ on $\mathbb{H}$. □

Theorem 9.5.3 gives us another way to characterize $T(\mathbb{H})$. The map from equivalence classes to functions

$$[f^\mu] \ \to \ f_\mu|_{\mathbb{H}}$$

is well defined, so

$$T(\mathbb{H}) \ \cong \ \{f_\mu|_{\mathbb{H}} \ : \ \mu \in B\}. \tag{9.5.10}$$

There is a conformal invariant associated with conformal homeomorphisms from $\mathbb{H}$ into $\mathbb{C}$, namely the *Schwarzian derivative*, or simply the *Schwarzian*. The Schwarzian $\{f, z\}$ of a holomorphic function f is defined to be

$$\{f, z\} \ = \ \left(\frac{f''(z)}{f'(z)}\right)' - \frac{1}{2}\left(\frac{f''(z)}{f'(z)}\right)^2. \tag{9.5.11}$$

To simplify the following statements, we assume that the functions in question are defined in a simply connected domain Ω in $\mathbb{S}$, which we generally take to be $\mathbb{H}$.

The formula (9.5.11) extends to meromorphic functions, and in any case defines a meromorphic function. In particular, we may compute the Schwarzian of a linear fractional transformation $f \in \operatorname{Aut}(\mathbb{S})$ and check that

$$\{f, z\} \equiv 0 \quad \text{if and only if} \quad f \in \operatorname{Aut}(\mathbb{S}). \tag{9.5.12}$$

Proposition 9.5.4. *(a) If f is holomorphic and f' has no zeros, then $\{f, z\}$ is holomorphic.*

(b) For meromorphic f and g,

$$\{g \circ f, z\} = \{g, f(z)\} f'(z)^2 + \{f, z\}. \tag{9.5.13}$$

(c) If g is a linear fractional transformation,

$$\{g \circ f, z\} = \{f, z\}. \tag{9.5.14}$$

(d) Schwarzians $\{f_1, z\}$, $\{f_2, z\}$ are identical if and only if

$$f_2 = h \circ f_1 \tag{9.5.15}$$

for some linear fractional transformation h.

We leave the proof of Proposition 9.5.4 as Exercise 9.

It is common in this context to change the standard notation by considering $z \to \{f, z\}$ as the definition of the function S_f. Then (9.5.13) and (9.5.14) are

$$\mathrm{S}_{g\circ f} = \mathrm{S}_g \circ f\,(f')^2 + \mathrm{S}_f; \tag{9.5.16}$$

$$\mathrm{S}_{g\circ f} = \mathrm{S}_f, \qquad g \in \mathrm{Aut}(\mathbb{S}). \tag{9.5.17}$$

The next result tells how to recover f from its Schwarzian.

Theorem 9.5.5. *If g is holomorphic and g' has no zeros, then solutions φ_1, φ_2 of the equation*

$$\varphi'' + \tfrac{1}{2} g\,\varphi = 0 \tag{9.5.18}$$

can be chosen in such a way that the quotient $f = \varphi_1/\varphi_2$ satisfies $S_f = g$.

Proof: We rely on some basic facts about linear differential equations. (Standard proofs for functions of a real variable carry over to the complex case, working with holomorphic coefficients and solutions.) Equation (9.5.18) has a two-dimensional space of solutions. If φ_1 and φ_1 are two solutions, then a simple computation shows that the Wronskian

$$\varphi_1'\varphi_2 - \varphi_1\varphi_2' \tag{9.5.19}$$

is constant. If the φ_j are chosen to be independent, then (9.5.19) is not zero, and we may normalize so that

$$\varphi_1'\varphi_2 - \varphi_1\varphi_2' = 1. \tag{9.5.20}$$

Then (9.5.20) implies that with $f = \varphi_1/\varphi_2$ we have $f' = 1/{\varphi_2}^2$. A further computation, again using (9.5.20), shows that the Schwarzian of f is g. □

Consider S_{f_μ} in $\mathbb{H}$. For any $h \in \mathrm{Aut}(\mathbb{H})$, $\mathrm{S}_h = 0$, so Proposition 9.5.4 (b) shows that under the change of coordinates $\zeta = h(z)$, the expression

$$\omega = \mathrm{S}_f(z)\,(dz)^2$$

transforms as

$$\omega(z) \;\to\; \mathrm{S}_f(h(z)\,(h')^2\,(dz)^2 \;=\; \mathrm{S}_f(\zeta)\,(d\zeta)^2 \;=\; \omega(\zeta).$$

Thus, ω is invariant under any conformal change of variables in $\mathbb{H}$. It is referred to as a *quadratic differential*. Since $\mathrm{S}f$ is holomorphic, ω is termed a *holomorphic* differential.

9.6 The Bers embedding

This section uses some results from Sections 2.2 and 4.1.

We saw in Section 9.5 that $T(\mathbb{H})$ can be identified with the family of holomorphic maps $\{f_\mu|_{\mathbb{H}}\}$. Each such map is a homeomorphism onto some domain $\Omega \subset \mathbb{C}$. Theorem 9.5.5 shows that $f_\mu|_{\mathbb{H}}$ can be reconstructed uniquely from its Schwarzian (taking into account the normalization of f_μ). Therefore we have one more identification: Let

$$s_\mu \;=\; \mathrm{S}_{f_\mu}|_{\mathbb{H}}. \tag{9.6.1}$$

Then $[f^\mu] \to s_\mu$ is well defined, and

$$T(\mathbb{H}) \;\cong\; \{s_\mu \;:\; \mu \in B\}. \tag{9.6.2}$$

This identification makes possible another natural choice of metric on $T(\mathbb{H})$. Recall that the hyperbolic distance element at a point z in $\mathbb{H}$ is $dz/2\mathrm{Im}\,(z)$. Therefore, if $h \in \mathrm{Aut}(\mathbb{H})$, then

$$\frac{dz}{2\,\mathrm{Im}\,z} \;=\; \frac{h'(z)\,dz}{2\,\mathrm{Im}\,h(z)}. \tag{9.6.3}$$

By (9.5.16) with $f = f_\mu|_{\mathbb{H}}$ and $h \in \mathrm{Aut}(\mathbb{H})$,

$$\mathrm{S}_{f\circ h} \;=\; \mathrm{S}_{f\circ h}\,(h')^2. \tag{9.6.4}$$

Combining (9.6.4) and (9.6.3), we get

$$4\,(\mathrm{Im}\,z)^2\,|s_\mu(z)| \;=\; 4\,(\mathrm{Im}\,h(z))^2|s_{\mu\circ h}(z)|. \tag{9.6.5}$$

In other words, the expression on the left in (9.6.5) is a conformal invariant in $\mathbb{H}$. Set

$$||\mathrm{S}_{f_\mu}||_{\mathbb{H}} \;=\; \sup_{z\in\mathbb{H}} 4\,(\mathrm{Im}\,z)^2|s_\mu(z)|.$$

It will be helpful to note that conformal equivalence of $\mathbb{H}$ and $\mathbb{D}$ implies that the invariance in (9.6.5) carries over to $\mathbb{D}$, where the hyperbolic distance element is $dz/(1-|z|^2)$. Thus, if $f : \mathbb{D} \to \mathbb{C}$ is conformal, then

$$(1-|z|^2)^2\,\mathrm{S}_f(z) \;=\; (1-|h(z)|^2)^2\,\mathrm{S}_f(h(z)), \qquad h\in \mathrm{Aut}(\mathbb{D}). \tag{9.6.6}$$

It is convenient to define

$$||\mathrm{S}f||_{\mathbb{D}} \;=\; \sup_{z\in\mathbb{D}} (1-|z|^2)^2\,|f(z)|.$$

Theorem 9.6.1. (Nehari) *Suppose that $f:\mathbb{D}\to\mathbb{C}$ is holomorphic.*
(a) If f is conformal, then $||\mathrm{S}_f||_{\mathbb{D}}\le 6$.
(b) If $||\mathrm{S}_f||_{\mathbb{D}}\le 2$, then f is conformal.

We prove part (a) now. Part (b) is a consequence of Lemma 9.6.3 below. Suppose that f is conformal and the maximum value of $(1-|z|^2)^2|\mathrm{S}_f(z)|$ is attained at z_0. We may take advantage of invariance under $z\to h(z)$, $h\in\mathrm{Aut}(\mathbb{S})$, to take $z=0$ and assume also that $f'(0)=1$. Thus,

$$f(z) \;=\; z+a_2z^2+a_3z^3+\ldots, \qquad S_f(0) \;=\; 6(a_3-a_2)^3.$$

The function

$$g(z) \;=\; \frac{1}{\mathrm{S}_f(1/z)} \;=\; z+\sum_{n=1}^{\infty} b_n z^{-n}.$$

satisfies the conditions of the Area Theorem, Theorem 4.1.1, so $|b_1|\le 1$. But a calculation shows that $b_1=a_3-a_2^2$, so we have $|S_f(0)|\le 6$. □

Remark. Part (a) was proved by Kraus [126] and rediscovered by Nehari [152]. Part (b) is deeper, and Nehari clearly considered it to be the principal result of his paper.

Theorem 9.6.1 (a) carries over to $\mathbb{H}$, using

$$||\mathrm{S}f||_{\mathbb{H}} \;=\; \sup_{\mathbb{H}} 4\,(\mathrm{Im}\,z)^2\,||\mathsf{f}||_{\mathbb{H}}.$$

We extend this expression to the space $Q(\mathbb{H})$ of holomorphic functions φ on $\mathbb{H}$:

$$||\varphi||_{\mathbb{H}} \;=\; \sup_{z\in\mathbb{H}} 4\,(\mathrm{Im}\,z)^2\,|\varphi(z)|. \tag{9.6.7}$$

It clearly has the properties of a norm:

$$||a\varphi||_{\mathbb{H}} \;=\; |a|\,||\varphi||_{\mathbb{H}},\;\; a\in\mathbb{C}; \qquad ||\varphi+\psi||_{\mathbb{H}} \;\le\; ||\varphi||_{\mathbb{H}}+||\psi||_{\mathbb{H}}.$$

Then $Q(\mathbb{H})$ is a complete normal family. In particular, it is complete with respect to the norm and is therefore a Banach space. Let Δ be the image of B in $Q(\mathbb{H})$:

$$\Delta \;=\; \{s_\mu\,:\,\mu\in B\}. \tag{9.6.8}$$

The rest of this section is devoted to the proof of the following result due to Ahlfors [4].

Theorem 9.6.2. *The image Δ of B under the map $\mu \to s_\mu$ is open in $Q(\mathbb{H})$.*

We begin with the proof of Theorem 9.6.1 (b).

Lemma 9.6.3. *If $\phi \in Q(\mathbb{H})$ and $||\phi||_{\mathbb{H}} < 2$, then $\phi \in \Delta$.*

Proof: We know from Theorem 9.5.5 that $\phi = \mathrm{S}_f$, where $f = v_1/v_2$ and the v_j are solutions of

$$v'' + \frac{1}{2}\phi v \ = \ 0 \tag{9.6.9}$$

such that

$$v_1'v_2 - v_2'v_1 \ = \ 1. \tag{9.6.10}$$

We want to extend f to the lower half-plane $\mathbb{H}^*$ by finding a function that coincides with f when $\operatorname{Im} z = 0$ and whose maximal dilatation is finite. We take the extension to be $F(\bar{z})$, where

$$F(z) \ = \ \frac{v_1(z) + (\bar{z} - z)v_1'(z)}{v_2(z) + (\bar{z} - z)v_2'(z)}, \qquad z \in \mathbb{H}.$$

By (9.6.10), the numerator multiplied by v_2, minus the denominator multiplied by v_2, gives $z - \bar{z}$. Therefore the numerator and denominator have no common zeros. Some computation shows that

$$F_{\bar{z}} = \frac{1}{[v_2(z) + (\bar{z} - z)v_2'(z)]^2}; \tag{9.6.11}$$

$$F_z = -\frac{\frac{1}{2}\phi(z)(z - \bar{z})^2}{[v_2(z) + (\bar{z} - z)v_2'(z)]^2}. \tag{9.6.12}$$

Therefore

$$|F_z/F_{\bar{z}}||_{\mathbb{H}} \ \leq \ \frac{1}{2}\,||\phi||_{\mathbb{H}} = k \ < 1.$$

It follows that the extension $f = F(\bar{z})$ is quasiconformal but sense reversing, with

$$\mu_f(z)^{-1} \ = \ -2\phi(\bar{z})(\operatorname{Im} z)^2, \qquad z \in \mathbb{H}.$$

If f extends continuously to the boundary from $\mathbb{H}$, then f is continuous on $\mathbb{C}$, and we may conclude that, after normalizing, $f = f_\mu$ with $\mu = \mu_f$ in $\mathbb{H}^*$.

Suppose first that ϕ is analytic on $\mathbb{R}$ and has a zero of order ≥ 4 at ∞. The function f is globally continuous on $\mathbb{S}$ and locally single-valued on $\mathbb{C}$. The vanishing assumption on ϕ implies that solutions of (9.6.9) have the form

$$v(\zeta) \ = \ \frac{a}{\zeta} + b + \zeta\psi(\zeta), \qquad \zeta = \frac{1}{z}$$

where ψ is analytic in ζ. In fact (9.6.9) becomes an equation

$$\zeta^2\psi'' + 2\zeta\psi' + \frac{\phi}{2\zeta^4}\psi \; = \; \zeta\, g(\zeta), \quad g \text{ holomorphic at } 0,$$

which has a regular singular point at $\zeta = 0$. Therefore

$$\varphi_j \; = \; a_j z + b_j + O(z^{-1}), \quad a_1 b_2 - a_2 b_1 \; = \; 1,$$

and

$$\lim_{z\to\infty} f(z) \; = \; \frac{a_1}{a_2} \; = \; \lim_{z\to\infty} F(z).$$

The fact that f is injective on $\mathbb{C}$ follows from the monodromy theorem. Composing with a linear transformation will normalize f.

To this point, we have shown that if $||\phi||_{\mathbb{H}} < 2$ and ϕ is analytic on $\mathbb{R}$ and has a zero of order ≥ 4 at ∞, then ϕ is in $\varDelta$. We pass to the general case by approximation, using a sequence of linear fractional transformations g_n with the property that the $g_n(\mathbb{H})$ expand to exhaust $\mathbb{H}$ and fix ∞. Such transformations are easily obtained in $\mathbb{D}$, fixing 1, in the form $h_n(z) = \rho_n z$ where $\varrho_n \uparrow 1$, and then transplanted to $\mathbb{H}$ by using the Cayley transform C. Let $\tau_n = C^{-1} \circ h_n \circ C$. If we set $\varrho_n = n - 1/2$, the result is

$$\tau_n(z) \; = \; \frac{2nz + i}{2n - iz}$$

Let

$$\phi_n \; = \; \phi \circ \tau_n \cdot (g_n')^2.$$

Now ϕ_n is analytic up to $\mathbb{R}$. Moreover, $|g_n'| \leq 1$ and $|\tau_n'(z)| = O(|z|^{-2})$ as $|z| \to \infty$. In addition,

$$||\phi_n||_{\mathbb{H}} \; \leq \; ||\phi||_{\mathbb{H}} \; < \; 2.$$

Therefore we can find $\{f_n\}$ with $\mathrm{S}_{f_n} = \phi_n$ in $\mathbb{H}$ and with a fixed bound for K_{f_n}. Normalizing, there is a subsequence that converges in $\mathbb{C}$ to a solution that is of the form f_μ, $\mu \in B$. □.

We turn now to the proof of Theorem 9.6.2. We begin with some remarks about the quasi-isometry property of the Beurling–Ahlfors extension, Theorem 8.7.8. If the K-quasiconformal map $\varphi : \mathbb{H} \to \mathbb{H}$ is such an extension, then

$$\frac{1}{c_1(K)}\,\frac{|dz|}{(\mathrm{Im}\, z)^2} \; \leq \; \frac{|d\varphi(z)|}{(\mathrm{Im}\,\varphi(z))^2} \; \leq \; c_1(K)\,\frac{|dz|}{(\mathrm{Im}\, z)^2}. \tag{9.6.13}$$

This may be rewritten in a conformally invariant form by noting that the density for the hyperbolic metric on $\mathbb{H}$ is $\eta_{\mathbb{H}}(z) = 1/(2\mathrm{Im}\, z)$. Therefore (9.6.13) is

$$\frac{1}{c_1(K)}\,\eta_{\mathbb{H}}(z)\,|dz| \; \leq \; \eta_{\mathbb{H}}(\varphi(z))\,|d\varphi(z)| \; \leq \; c_1(K)\,\eta_{\mathbb{H}}(z)\,|dz|. \tag{9.6.14}$$

We need a general fact about the hyperbolic density of a general simply connected domain $\varOmega$, from Proposition 2.2.3:

$$\frac{1}{4\,d(z,\partial\Omega)} \;\le\; \eta_\Omega(z) \;\le\; \frac{1}{d(z,\partial\Omega)}, \tag{9.6.15}$$

where $d(z,\partial\Omega)$ is the (euclidean) distance to the boundary.

The following theorem, due to Ahlfors, is key to the proof of Theorem 9.6.2. We use the argument in [131].

If L is a Jordan curve in $\mathbb{S}$, then a *K-quasiconformal reflection* in L is a K-quasiconformal homeomorphism $g:\mathbb{C}\to\mathbb{C}$ that is the identity on L, is sense-preserving on one component of the complement of L, sense-reversing on the other component, and is an *involution*, which means that $g\circ g$ is the identity.

Theorem 9.6.4. *Suppose that L is an unbounded Jordan curve in $\mathbb{S}$ that admits a K-quasiconformal reflection g. Let Ω_1 be one of the components of the complement of L. Then there is a $c(K)$-quasiconformal reflection λ in L that is C^1 on $\Omega_1\cup\Omega_2$ and satisfies*

$$|d\lambda(z)| \le\; C(K)\,|dz|,\quad z\in\Omega_1. \tag{9.6.16}$$

Proof: Let $h_1:\mathbb{H}\to\Omega_1$ and $h_2:\mathbb{H}^*\to\Omega_2$ be conformal maps. The assumptions imply that the Ω_j are Jordan domains in $\mathbb{S}$, so that h_j extend to homeomorphisms of $\mathbb{R}$ onto the closure of Ω_j. Assume that orientations are chosen so that $h_2\circ h_1^{-1}:\mathbb{R}\to\mathbb{R}$ is increasing. Let $j(z)=\bar z$ and

$$\psi \;=\; j\circ h_2\circ g\circ h_1,$$

Then $\psi|_{\mathbb{H}}$ is a K-quasiconformal homeomorphism of $\mathbb{H}$. Since g and j are the identity on $\mathbb{R}$, it follows that $h_2^{-1}h_1 \;=\; \psi$ on $\mathbb{R}$. Therefore $h=h_2\circ h_1^{-1}$ is quasisymmetric.

Let φ be the Beurling–Ahlfors extension of h to $\mathbb{H}$, and define

$$\lambda \;=\; \begin{cases} h_2^{-1}\circ j\circ\varphi\circ h_1 & \text{on }\; \Omega_1\cup L;\\ h_1^{-1}\circ\varphi\circ j\circ h_2 & \text{in }\; \Omega_2. \end{cases} \tag{9.6.17}$$

Then λ is a $c(K)$-quasiconformal reflection in L that is C^1 on $\Omega_1\cup\Omega_2$. The inequalities (9.6.14) carry over to give, in particular,

$$\frac{1}{c_1(K)}\eta_1(z)|dz| \;\le\; \eta_2(\lambda(z))|d\lambda(z)| \;\le\; c_1(K)\eta_1(z)|dz|,\quad z\in\Omega, \tag{9.6.18}$$

For the purpose of this proof, we shall abbreviate inequalities like (9.6.18), with a constant that depends only on K, as $\eta_1|dz|\sim\eta_2|d\lambda(z)|$. We want to show that $\eta_1(z)\sim\eta_2(\lambda(z))$. In view of (9.6.15), this amounts to showing that

$$d(z,L) \;\sim\; d(\lambda(z),L). \tag{9.6.19}$$

For this purpose, we use the circular distortion theorem, Theorem 8.6.1. Let $\psi=h_1$ on $\mathbb{H}\cup\mathbb{R}$ and $\psi=\lambda\circ h_1\circ j$ in $\mathbb{H}^*$. Then $\psi:\mathbb{C}\to\mathbb{C}$ is K-quasiconformal. Given

$z \in \Omega_1$ and $z_0 \in L$, let C be the circle in $\mathbb{C}$,

$$C = \{w \in \mathbb{C} : |w - h_1^{-1}(z_0)| = |h_1^{-1}(z) - h_1^{-1}(z_0)|\}.$$

Then $\lambda(C)$ passes through $\lambda(z)$ and $\lambda(z_0) = z_0$. It follows from Theorem 8.6.1 that

$$|z - z_0| \sim |\lambda(z) - z_0|.$$

Taking z_0 so that $|z - z_0| = d(z, L)$ and, separately, so that $|\lambda(z) - z_0| = d(\lambda(z), L)$, we obtain (9.6.19). □

Now we proceed to the proof of Theorem 9.6.2. Given $\phi_0 \in Q(\mathbb{H})$, with $\phi_0 = \mathrm{S}_{f_0}$, $f_0 = f_{\mu_0}$. Suppose that f_0 is K-quasiconformal. Let $\Omega = \phi_0(\mathbb{H})$, $L = \phi(\mathbb{R})$, $\Omega^* = \phi_0(\mathbb{H}^*)$. By Theorem 9.6.4, L admits a $c_0(K)$-quasiconformal reflection λ such that $|dz| \sim |d\lambda(z)|$. This implies inequalities

$$\frac{1}{c_1(K)} \leq |\varphi_{\bar{z}}| \leq c_1(K).$$

If $\phi \in Q(\mathbb{H})$, with $\phi = \mathrm{S}_f$, let $g = f \circ f_0^{-1}$. Then by (9.5.16)

$$\phi - \phi_0 = \mathrm{S}_{g \circ f_0} - \mathrm{S}_{f_0} = \mathrm{S}_g \circ f_0 (f_0')^2 = \mathrm{S}_f (f_0')^2.$$

The hyperbolic metric in Ω is given by

$$\eta_{\lambda(z)} |d\lambda(z)| = \frac{|dz|}{2\mathrm{Im}\, z}.$$

Therefore $||\phi - \phi_0||_{\mathbb{H}} \leq \varepsilon$ implies

$$|g(\zeta)| \leq \varepsilon \eta(\zeta)^2. \tag{9.6.20}$$

We want to show that for small ε, g has a quasiconformal extension. Let $\psi = \mathrm{S}_g$ let v_1, v_2 be normalized solutions of $v'' + \frac{1}{2}\psi v = 0$:

$$v_j'' + \frac{1}{2}\psi v_j = 0, \qquad v_1' v_2 - v_2' v_1 = 1. \tag{9.6.21}$$

Set

$$g(\zeta) = \frac{v_1(\zeta)}{v_2(\zeta)}, \qquad \zeta \in \Omega;$$
$$\widehat{g}(\zeta) = \frac{v_1(\zeta^*) + (\zeta - \zeta^*)v_1'(\zeta^*)}{v_2(\zeta*) + (\zeta - \zeta^*)v_2'(\zeta^*)}, \qquad \zeta \in \Omega^*, \quad \zeta^* = \lambda(\zeta).$$

Let $\Delta_j = v_j(\zeta^*) + (\zeta - \zeta^*)v_j'(\zeta^*)$. Then, using (9.6.21), we see that

$$\widehat{g}_{\bar\zeta} = \frac{(\zeta-\zeta^*)[v_1''\Delta_2 - v_2''\Delta_1]\lambda_{\bar\zeta}}{\Delta_2^2} = \frac{(\zeta-\zeta^*)^2\frac{1}{2}\psi(\zeta^*)\lambda_{\bar\zeta}(\zeta)}{\Delta_2^2} \tag{9.6.22}$$

$$\widehat{g}_{\zeta} = \frac{[v_1' + (\zeta-\zeta^*)v_1''\lambda_\zeta]\Delta_2 - [v_2' + (\zeta-\zeta^*)v_2''\lambda_\zeta]\Delta_1}{\Delta_2^2} \tag{9.6.23}$$

$$= \frac{1 + (\zeta-\zeta^*)^2\frac{1}{2}\psi(\zeta^*)\lambda_\zeta(\zeta)}{\Delta_2^2}. \tag{9.6.24}$$

Therefore

$$\mu_{\widehat{g}}(\zeta) \;=\; \frac{\frac{1}{2}(\zeta-\zeta^*)^2\psi(\zeta^*)\lambda_{\bar z}(\zeta)}{1+\frac{1}{2}(\zeta-\zeta^*)\psi(\zeta^*)^2\lambda_\zeta(\zeta)}, \qquad \zeta\in\Omega. \tag{9.6.25}$$

Now $|\lambda_z| < |\lambda_{\bar z}| < c(K)$ and $|\zeta-\zeta^*| < C/\eta(\zeta^*)$, so

$$|\mu_{\widehat{g}}| \;\le\; \frac{\varepsilon\, c(K)}{1-\varepsilon c(K)} \;<\; 1 \tag{9.6.26}$$

for sufficiently small ε. We need to show that $\widehat{g}$ is continuous and injective. Once again this is true if L is an analytic curve and ψ is analytic on L with a zero of order 4 at ∞. Therefore we proceed again by approximation. With τ_n as before, let $f_n = f_0\circ\tau_n$ and $L_n = f_n(\mathbb{R})$. Then L_n admits a quasiconformal reflection and ψ is analytic on L_n. Since $\tau_n(\mathbb{H}^*)\supset\mathbb{H}^*$, by Proposition 2.2.2 the hyperbolic density η_n of $f_n(\mathbb{H}^*)$ is $\ge \eta = \eta_{\mathbb{H}^*}$, so $|\psi|\le\varepsilon\eta$ implies $|\psi|\le\varepsilon\eta_n$. The associated maps $\widehat{g}_n$ with $\mathrm{S}_{g_n} = \psi$ in Ω_n satisfy (9.6.26) uniformly. Therefore a subsequence converges to a quasiconformal $\widehat{g}$ that equals g in $\mathbb{H}^*$. □

9.7 Further developments

Pushing these results much further requires many additional technical steps, and is beyond the scope of this book. In this section, we give a very brief look at some more of the theory.

Let us mention first what is referred to as *Teichmüller's theorem*. Opinions seem to differ about how close Teichmüller came to a rigorous proof of this, especially the statement of existence. The theorem—and the theory—have been generalized to many kinds of non-compact Riemann surfaces, and in other ways as well; see the references in the last section.

The exact statement of the theorem varies somewhat from monograph to monograph, but the following is the gist.

Theorem 9.7.1. *If S is a compact Riemann surface, then in every homotopy class of sense-preserving homeomorphism of S onto another Riemann surface, there is a unique map whose maximal dilatation k is the smallest. The Beltrami coefficient has the form*

$$\mu(z) \;=\; k\,\frac{\overline{\phi(z)}}{\phi(z)} \tag{9.7.1}$$

where

$$\phi\,(dz)^2 \tag{9.7.2}$$

is a holomorphic quadratic differential on S.

We encountered a holomorphic quadratic differential at the end of Section 9.5 in the form (9.7.2), with $\phi = \mathrm{S}_f$. The theory leading up to Teichmüller's theorem and its generalizations relies heavily on the study of such differentials, as is evident from some of the titles of the references for this chapter. The study focuses particularly on the "trajectories" of such differentials. Some neighborhood U of a point z_0 where ϕ is injective can be parametrized by

$$\zeta = \int_{z_0}^{z} \sqrt{\phi(t)}\,dt.$$

Then $(d\zeta)^2 = \phi(z)(dz)^2$. A curve $\gamma : [a, b] \to U$ is said to be a *horizontal* trajectory of ϕ if $\arg[\phi(\gamma(t))(\gamma'(t))^2] = 0$, a *vertical* trajectory if $\arg[\phi(\gamma(t))(g'(t))^2] = \pi$. Carefully chosen horizontal and vertical trajectories eventually provide the desired type of parametrization of the moduli space.

9.8 Higher Teichmüller theory

To put this subject in context, we follow [215] and begin with a bird's eye view of Teichmüller theory itself. Let S be a compact Riemann surface of genus $g > 1$. The Teichmüller space $T(S)$ consists of equivalence classes of pairs (S', f), where S' is a Riemann surface and $f : S \to S'$ is a quasiconformal orientation-preserving homeomorphism. The equivalence relation is:

$$(S_1', f_1) \sim (S_2', f_2) \;\Leftrightarrow\; f_2 \circ f_1^{-1}\text{is homotopic to a conformal map.}$$

The universal cover of S' can be taken to be $\mathbb{H}$. Choosing a base point in S', there is a homomorphism

$$H_1(S') \to \mathrm{Aut}(\mathbb{H}) = PSL(2, \mathbb{R})$$

from the fundamental group $H_1(S')$ to the group of deck transformations. Given (S', f) as above, there is an injective map f^*:

$$f^* : H_1(S) \to H_1(S') \to \mathrm{Aut}(\mathbb{H}) \cong PSL(2, \mathbb{R}).$$

This can be seen to induce an injective homomorphism, the *holonomy*, from $T(S)$ to the set of homomorphisms from

$$\mathrm{hol} : T(S) \to \mathrm{Hom}\left(H_1(S), PSL(2, \mathbb{R})/PSL(2, \mathbb{R})\right), \tag{9.8.1}$$

where $PSL(2,\mathbb{R})/PSL(2,\mathbb{R})$ denotes the space $\mathrm{Aut}(\mathbb{H}) = PSL(2,\mathbb{R})$ up to inner automorphisms.

The Teichmüller space $T(S)$ is a connected component of the target space on the right in (9.8.1). There is a second connected component, the image of $T(\bar{S})$, suitably constructed, where $\bar{S}$ is the topological surface S with the opposite orientation.

The idea behind higher Teichmüller theory is to replace $PSL(2,\mathbb{R})$ in this picture by a Lie group G of higher rank, in such a way that the associated (higher) Teuchmüller space of S is a union of connected components of

$$\mathrm{Hom}(\pi_1(S), G)/G$$

that consists of discrete faithful representations of $\pi_1(S)$. This happens only for special choices of G.

In retrospect, the theory seems to have begun in 1992 with results of Hitchin [108]. The fact that Hitchin's construction fits into the theory as described here was proved in 2006 by Fock and Goncharov [77] and by Labourie [128]. A second family of higher Teichmüller spaces was defined in a different way, and was shown to fit the definition above by Burger, Iozzi, Labourie, and Wienhard [36].

This history, various points of view, and further developments of the subject are described in Wienhard's survey article [215]. There are many interactions with other research areas, such as Lie theory, representation theory, and ergodic theory. However, as noted in [215],

> In classical Teichmüller theory complex analytic methods and the theory of quasiconformal mappings play a crucial role. These aspects are so far largely absent from higher Teichmüller theory.

An exception to this is Dumas and Sanders [62].

Exercises

1. Prove Corollary 9.2.3.
2. Show that the set of points in $\mathbb{H}$ that are equidistant from two distinct points of $\mathbb{H}$ with respect to the hyperbolic metric is a geodesic.
3. Construct a homeomorphism of $\mathbb{H}$ onto itself that cannot be extended continuously to a homeomorphism of the boundary.
4. Prove Proposition 9.4.1. Hint: transfer the problem to $\mathbb{D}$.
5. Show that every point of $\mathbb{R}$ is a limit point of G if $G\backslash\mathbb{H}$ is compact. Hint: the ratio between the euclidean and the hyperbolic diameter of a compact subset of $\mathbb{H}$ goes to zero as the set approaches $\mathbb{R}$.)
6. Show directly that the set of quasisymmetric functions $f : \mathbb{R} \to \mathbb{C}$, normalized by $f(0) = 0$, $f(1) = 1$, is a group under composition.
7. Prove Theorem 9.4.6.

8. Prove Proposition 9.5.1. Hint: suppose that f and g are normalized, and use (9.4.9).
9. Verify Proposition 9.5.4.
10. Fill in the details in the proof of Theorem 9.5.5.
11. Prove that every element of $Q(\mathbb{H})$ is the Schwarzian of some function that is meromorphic in $\mathbb{H}$.
12. The universal Teichmüller space $T(UH)$ is a group under the composition $[f][g] = [f \circ g]$. The aim of this exercise is to show that the group composition is not continuous with respect to the Teichmüller metric, following the proof in [84]. Normalize elements $[f] \in T(\mathbb{H})$ by requiring that $-1, 1, \infty$ are fixed by $f_{\mathbb{R}}$.

    Show that there is a sequence $\{f_n\} \subset T(\mathbb{H})$ of normalized maps such that the distance $d_T(f_n, \mathbf{1})$ to the identity map $\mathbf{1}$ converges to zero and such that each f_n is asymptotically conformal, but $d_T(g \circ f_n, g)$ does not converge to zero, where $g(z) = z|z|$. You may use the fact that to have $d_T(f_n, \mathbf{1}) \to 0$, it is enough to have the restrictions to $\mathbb{R}$ satisfy

    $$\sup \left\{ \frac{f_n(x+t) - f_n(x)}{f_n(x) - f_n(x-t)}, f_n(x) - f_n(x-t) f_n(x+t) - f_n(x) \right\} \to 0.$$

13. Verify (9.6.11) and (9.6.12).
14. Verify (9.6.22) and (9.6.24).

Remarks and further reading

The history of this subject exhibits gaps and then bursts of activity. Moreover, the subject seems to have inspired an unusual number of pithy comments. We cannot resist the temptation to summarize through some quotations. According to Weyl [214], footnote, p.176, Fricke [79], with the aid of his study of canonical polygons,

> succeeded in formulating and proving rigorously the statement of Riemann: the Riemann surfaces of genus p ($p > 1$) form a $6p - 6$-dimensional manifold.

The subject seems to have rested there until revived by Teichmüller in the late 1930s and the early 1940s. Ahlfors [3] attacked the problem a decade later and wrote

> In a systematic way the problem of extremal quasiconformal mapping was taken up by Teichmüller in a brilliant and unconventional paper ... . He formulates the general problem and, although unable to give a binding proof, is led by heuristic arguments to a highly elegant conjectured solution. Tbe paper contains numerous fundamental applications which clearly show the importance of the problem. In a later publication [8] Teichmüller has offered a proof of his main conjecture. In many respects this proof is an anticlimax when compared with the original article. It is based on the method of continuity, which of all classical methods is the least satisfactory, because of the nature of a posteriori verification.

Ahlfors goes on to give a variational proof of Teichmüller's theorem. This paper by Ahlfors led to an explosion of activity by Ahlfors, Bers, Ahlfors–Bers, and many others. See, in particular, the introduction of a complex structure in Teichmüller space by Bers [25]. For more on the history of the subject and on the two decades after [3], see the books by Abikoff [1], Gardiner [82], Lehto [131], and Nag [150].

Here is the classic beginning of Abikoff's review of [131], MR0867407:

> In the late 1930s the study of variational techniques in complex analysis received a tremendous boost from the work of M. M. Schiffer on the Bieberbach conjecture. Schiffer's approach emphasized the role of quadratic differentials as nonlinear differential equations satisfied by any extremal mapping. Previously Grötzsch had studied the most nearly conformal maps between plane domains. The deviation of these maps from conformality is measured by the L^∞-norm of the logarithm of the local distortion. The class of homeomorphisms of finite norm consists of the quasiconformal mappings. In a pair of brilliant, but marginally readable, papers, Teichmüller completely revised the deformation theory of Riemann surfaces. His methods were a brilliant merging and extension of the methods and ideas of Schiffer and Grötzsch. The deformation theory he obtained is now called Teichmüller theory.
>
> By now, Teichmüller theory has developed to a point of mathematical maturity. By this term I mean five distinct qualities. First, its major practitioners speak so vastly different languages that they can barely understand one another. It is being used as a tool in a wide variety of mathematical and physical disciplines. It is serving as a model for new areas of mathematical research. Several books have recently been written on the subject. Last, the discipline is probably named after the wrong person.

Our exposition here is based mainly on the rather terse notes of Ahlfors [5] and the expansive book of Lehto [131]. The state of the theory in the mid-1980s is described well by the books of Abikoff, Gardiner, Lehto, and Nag mentioned above. Some of the developments of the following two decades are contained in the books of Fletcher and Markovic [76], Hubbard [111], [112], and Gardiner and Lakic [83].

Work in several directions is summarized in the chapters that have been added in the second edition of [5]. Earle and Kra cover further work along the lines pioneered by Ahlfors and Bers. Shimakura describes the work of Sullivan, Thurston, and others relating quasiconformal mapping and complex dynamics. Hubbard outlines Thurston's remarkable work on 3-manifolds. For more on this and related topics, including higher Teichmüller theory, see the expository article of Wolpert [217].

For still more, see the *Handbook of Teichmüller Theory* [95]. (Volume V of the *Handbook* includes translations of Teichmüller's papers [202], [203].) For a glimpse into the different state of affairs in higher dimensions, see the remarks concerning the Mostow rigidity theorem at the end of Chapter 8.

Chapter 10
The Bergman kernel

If Ω is a domain in $\mathbb{C}$, the set $H(\Omega)$ of functions f that are holomorphic in Ω and square-integrable with respect to the area measure $dm(z) = dx\,dy = \frac{i}{2}dz\,d\overline{z}$,

$$\iint_{\Omega} |f(z)|^2 dm(z) \;<\; \infty$$

is a Hilbert space. The Bergman kernel K has the property that for f in $H(\Omega)$ and z in Ω,

$$f(z) \;=\; \iint_{\Omega} K(z,w)\, f(w)\, dm(w).$$

The kernel K itself is introduced in Section 10.1. Section 10.2 looks at its expansion with respect to an orthonormal basis.

The Bergman kernel is a conformal invariant of the domain Ω. As such it can be expected to be closely related to other such invariants. For simply connected Ω, one such invariant is the inverse of the Riemann map from $\mathbb{D}$ to Ω. Section 10.3 exhibits the relation of this inverse map to K.

The kernel function also has natural geometric significance. Section 10.4 covers conformal invariance and the Bergman metric. Conformal invariance suggests that the kernel K is closely related to other natural geometric features of a domain. We have already seen this in Section 10.3 in connection with the Riemann map, and we return to this theme in Section 10.5, in relation to conformal maps from domains that are not simply connected, and in Section 10.6 in connection with natural boundary problems for the Laplacian Δ in Ω.

R. Beals and R. S. C. Wong, *More Explorations in Complex Functions*, Graduate Texts in Mathematics 298, https://doi.org/10.1007/978-3-031-28288-1_10

10.1 The reproducing kernel

Suppose that Ω is a domain in $\mathbb{C}$ whose boundary contains more than one point. As in the introduction, we denote by $H(\Omega)$ the space of functions f, holomorphic on Ω, that are square-integrable on Ω. The corresponding inner product is

$$(f, g) = \iint_{\Omega} f(z)\overline{g(z)}\, dm(z). \qquad (10.1.1)$$

We assume throughout that Ω is such that $H(\Omega)$ contains non-zero functions. Note that this is not obvious if Ω is unbounded (and is not true if the boundary contains only one point; see Exercise 1).

Our first step here is to show that $H(\Omega)$ is a Hilbert space, i.e. that it is complete with respect to the metric induced by the norm

$$||f|| = (f, f)^{1/2} = \left(\iint_{\Omega} |f(z)|^2\, dm(z)\right)^{1/2}.$$

Lemma 10.1.1. *If f belongs to $H(\Omega)$ and z is a point of Ω, then*

$$|f(z)| \leq \frac{1}{\sqrt{\pi}\, R}||f|| \qquad (10.1.2)$$

where R is the distance from z to the boundary $\partial\Omega$.

Proof: For convenience, we translate coordinates so that $z = 0$. Then for $0 \leq r < R$,

$$|f(re^{i\theta})| = \left|\sum_{n=0}^{\infty} a_n r^n e^{in\theta}\right|, \qquad a_n = \frac{f^{(n)}}{n\,!}.$$

We square and average with respect θ to get

$$\frac{1}{2\pi}\int_0^{2\pi} |f(re^{i\theta})|^2 d\theta = \frac{1}{2\pi}\int_0^{2\pi} \sum_{n,m=0}^{\infty} a_n \bar{a}_m r^{n+m} e^{i(n-m)} d\theta$$
$$= \sum_{n=0}^{\infty} |a_n|^2 r^{2n},$$

since

$$\frac{1}{2\pi}\int_0^{2\pi} e^{i(n-m)\theta}\, d\theta = \begin{cases} 1, & m = n; \\ 0, & m \neq n. \end{cases}$$

The disk $D_R(0)$ is contained in Ω, so

$$
\begin{aligned}
||f||^2 &\geq \int_0^R \int_0^{2\pi} |f(re^{i\theta}|^2 r\, d\theta\, dr \\
&= 2\pi \int_0^R \sum_{n=0}^{\infty} |a_n|^2 r^{2n+1}\, dr \\
&= 2\pi \sum_{n=0}^{\infty} \frac{|a_n|^2}{2n+2} R^{2n+2} \\
&\geq 2\pi \frac{|a_0|^2}{2} R^2 \;=\; \pi |f(0)|^2 R^2.
\end{aligned}
$$ □

Note that the example $\Omega = \mathbb{D}$, $f = 1$, $z = 0$ shows that the estimate (10.1.2) is sharp.

Proposition 10.1.2. *$H(\Omega)$ is a Hilbert space.*

Proof: We must show that $H(\Omega)$ is complete. Suppose that $\{f_n\}$ is a Cauchy sequence in $H(\Omega)$. It follows from Lemma 10.1.1 that the functions $\{f_n\}$ converge uniformly on each compact subset of Ω. Therefore the limit function f is holomorphic. □

Proposition 10.1.3. *Given $z \in \Omega$, there is a unique element $k_z \in H(\Omega)$ such that for each f in $H(\Omega)$,*

$$f(z) \;=\; (f, k_z). \tag{10.1.3}$$

Proof: The estimate (10.1.1) shows that the linear map $f \to f(z)$ is bounded with respect to the Hilbert norm. Since $H(\Omega)$ is a Hilbert space, any such map can be written uniquely as an inner product with an element of $H(\Omega)$; this is Proposition 2.7.1. □

Proposition 10.1.4. *The element k_z satisfies the estimate*

$$||k_z|| \;\leq\; \frac{1}{\sqrt{\pi}R} \tag{10.1.4}$$

where R is the distance from z to $\partial\Omega$.

Proof. Apply (10.1.3) with $f(w) = k_z(w)$:

$$||k_z||^2 \;=\; (k_z, k_z) \;=\; k_z(z) \;\leq\; \frac{||k_z||}{\sqrt{\pi}R}.$$ □

The *Bergman kernel* for the domain Ω is defined by

$$K(z, w) \;=\; K(z, w; \Omega) \;=\; \overline{k_z(w)} \;=\; \overline{(k_w, k_z)} \;=\; (k_w, k_z), \tag{10.1.5}$$

where k_z is the element of $H(\Omega)$ that is defined by (10.1.3). By definition, the kernel K has the *reproducing property* for $H(\Omega)$: for each $f \in H(\Omega)$, we have

$$\iint_{\Omega} K(z,w) f(w)\, dm(w) \;=\; f(z). \tag{10.1.6}$$

In fact, the integral (10.1.6) is just (f, k_z), rewritten.

Proposition 10.1.5. *The Bergman kernel is holomorphic as a function of z, and anti-holomorphic as a function of w. Moreover, K has hermitian symmetry*

$$\overline{K(z,w)} \;=\; K(w,z), \tag{10.1.7}$$

and satisfies the estimate

$$|K(z,w)|^2 \;\leq\; |K(z,z)|\,|K(w,w)| \;\leq\; \frac{1}{\pi R_z R_w}, \tag{10.1.8}$$

where R_a is the distance from $a \in \Omega$ to $\partial\Omega$.

Proof: By definition, $k_z(w) \in H(\Omega)$ is holomorphic in w, so the complex conjugate $K(z,w)$ is anti-holomorphic in w. The assertion (10.1.7) will then imply that $K(z,w)$ is holomorphic with respect to z. The identity (10.1.7) follows from (10.1.5). The estimate (10.1.8) follows from (10.1.4) and (10.1.5). □

The kernel K has a certain extremal property that is important for applications, e.g. in Section 10.2.

Proposition 10.1.6. *Suppose z_0 is a point of Ω. The unique solution to the problem of minimizing $||f||$ for f in $H(\Omega)$ such that $f(z_0) = 1$ is given by the function*

$$f(z) \;=\; \frac{K(z,z_0)}{K(z_0,z_0)}. \tag{10.1.9}$$

Proof: We note first that the assumption that $H(\Omega) \neq (0)$, together with (10.1.8), implies that $K(z_0,z_0) > 0$, so f in (10.1.9) is well defined and satisfies $f(z_0) = 1$. Now f is a scalar multiple of k_{z_0}, so any $g \in H(\Omega)$ can be written as

$$g \;=\; cf + h,$$

where c is constant and h is orthogonal to k_{z_0}. Then $h(z_0) = (h, k_{z_0}) = 0$, $g(z_0) = 1$ implies that $c = 1$. Then

$$||g||^2 \;=\; (f+h, f+h) \;=\; ||f||^2 + ||h||^2.$$

Therefore the desired minimum is obtained where $h = 0$ and thus $g = f$. □

For a generalization of Proposition 10.1.6, see Exercise 2.

The assertion above is that K is separately holomorphic and anti-holomorphic in its arguments, but not that it is even continuous as a function of the two variables jointly. As we shall see, much more is true.

Lemma 10.1.7. *If f belongs to $H(\Omega)$ and z is a point of Ω, then there is a constant C such that derivatives of f satisfy*

$$|f^{(n)}(z)| \ \le \ \frac{n\,!2^n}{R^{n+1}}||f||, \tag{10.1.10}$$

where R is the distance from z to $\partial\Omega$.

Proof: Use the Cauchy integral formula to write $f^{(n)}(z)$ as an integral over the circle $\{\zeta : |\zeta - z| = R/2\}$, and use (10.1.2) to estimate $|f(\zeta)|$. □

Corollary 10.1.8. *The map $z \to k_z$ is continuous from Ω to $H(\Omega)$.*

Proof: Given z, z' in Ω,

$$||k_{z'} - k_z|| \ = \ \sup_{||f||=1} |(k_{z'} - k_z, f)| \ = \ \sup_{||f||=1} |f(z') - f(z)|$$

The estimate (10.1.10) with $n = 1$ gives local estimates $|f(z) - f(z')| = O(|z' - z|)$. □

Corollary 10.1.9. *$K(z, w)$ is jointly continuous in (z, w).*

Proof: This follows from Corollary 10.1.8 and the estimate (10.1.1). □

Corollary 10.1.9 allows us to combine separate Cauchy integral formulas for K with respect to z and to $\overline{w}$:

Proposition 10.1.10. *For z_0, $\overline{w}_0$ in Ω, and sufficiently small $r > 0$, $s > 0$, the kernel function K satisfies a double integral formula*

$$K(z_0, \overline{w}_0) \ = \ \frac{1}{(2\pi i)^2} \int_{|z-z_0|=r} \int_{|w-w_0|=s} \frac{K(z, \overline{w})\, dz\, d\overline{w}}{(z - z_0)(\overline{w} - \overline{w}_0)}. \tag{10.1.11}$$

Expanding the denominator of the integrand gives the power series expansion:

Corollary 10.1.11. *For z_0, $\overline{w}_0$ in Ω, and sufficiently small $r > 0$, $s > 0$, the kernel function K has a series expansion for $|z - z_0| < r$, $|\overline{w} - \overline{w}_0| < s$:*

$$K(z, \bar{w}) \ = \ \sum_{m,n=0}^{\infty} a_{mn}(z - z_0)^m(\overline{w} - \overline{w}_0)^n. \tag{10.1.12}$$

Remark. The Bergman kernel is one example of what was later termed a *reproducing kernel*; see the survey by Aronszajn [11]. An earlier example was examined by Szegő [199] in connection with the *Hardy space* H^2; see Exercise 14.

10.2 Orthonormal bases

Suppose that $\{\phi_n\}_{n=1}^{\infty}$ is an orthonormal basis for the Hilbert space $H(\Omega)$. Then each f in $H(\Omega)$ has an expansion

$$f = \sum_{n=1}^{\infty}(f, \phi_n)\phi_n,$$

and

$$||f||^2 = \sum_{n=1}^{\infty}|(f, \phi_n)|^2.$$

This series converges in norm, so by (10.1.1) it converges pointwise, uniformly on compact sets. Thus for z in Ω,

$$f(z) = \sum_{n=1}^{\infty}(f, \phi_n)\phi_n(z). \tag{10.2.1}$$

In particular,

$$k_z(w) = \sum_{n=1}^{\infty}(k_z, \phi_n)\phi_n(w) = \sum_{n=1}^{\infty}\overline{\phi_n(z)}\phi_n(w).$$

Since $K(z, w) = \overline{k_z(w)}$, we have proved

Proposition 10.2.1. *The Bergman kernel has an expansion*

$$K(z, w) = \sum_{n=1}^{\infty}\phi_n(z)\overline{\phi_n(w)}, \tag{10.2.2}$$

where $\{\phi_n\}_1^{\infty}$ *is any orthonormal basis for* $H(\Omega)$.

Let us consider two examples. The proofs are left as Exercise 4 and 5.

Proposition 10.2.2. *The functions*

$$\phi_n(z) = \frac{n^{1/2}}{\pi^{1/2}R^n}\, z^{n-1}, \qquad n = 1, 2, \ldots \tag{10.2.3}$$

are an orthonormal basis for $H(D_R)$, where $D_R = D_R(0)$ is the disk of radius R centered at the origin.

Corollary 10.2.3. *The kernel function K for the disk $D_R(0)$ is*

$$K(z,w) \;=\; \frac{1}{\pi R^2}\left(1-\frac{z\overline{w}}{R^2}\right)^{-2} \;=\; \frac{R^2}{\pi(R^2-z\overline{w})^2}. \tag{10.2.4}$$

Proposition 10.2.4. *The functions*

$$\psi_n(z) \;=\; \begin{cases} \left[\dfrac{|n|R^2}{\pi|R^{2n}-\varrho^{2n}|}\right]^{1/2} z^{n-1}, & |n|-1 = 0,1,2,\ldots; \\ [2\pi\,\log(R/\varrho]^{-1/2}\,z^{-1}, & n=0. \end{cases} \tag{10.2.5}$$

are an orthonormal basis for $H(A_{\varrho,R})$, where $A_{\varrho,R}$ is the annulus

$$A_{\varrho,R} \;=\; \{z : \varrho < |z| < R\}. \tag{10.2.6}$$

The basis (10.2.5) leads to the formula

$$K(z,w) \;=\; \frac{1}{2\pi\,\log(R/\rho)}(z\overline{w})^{-1} + \sum_{n\neq 0}^{\infty} \frac{R^2}{\pi(R^{2n}-\rho^{2n})}\,(z\overline{w})^{n-1}. \tag{10.2.7}$$

This series can be summed to give an explicit formula for the Bergman kernel for $A_{\varrho,R}$ in terms of the Weierstrass function $\wp$; see [24], Section 1.4.

We turn now to a result that will allow us to prove a monotonicity result for kernels. Suppose that $\Omega_1 \subset \Omega$. Let $(\;,\;)_1$ and $||\;||_1$ denote the inner product and norm in $H(\Omega_1)$. Given $u \in H(\Omega)$, the restriction (also denoted u) belongs to $H(\Omega_1)$ and $||u||_1 \le ||u||$.

Proposition 10.2.5. *If $\Omega_1 \subset \Omega$, there is an orthogonal basis $\{\phi_n\}$ for Ω such that the restrictions $\{\phi_n|_{\Omega_1}\}$ are orthogonal in $H(\Omega_1)$.*

Proof: We define an orthonormal set in $H(\Omega_1)$ inductively. Choose $\psi_1 \in H(\Omega)$ such that $||\psi_1||$ is minimal, subject to the constraint $||\psi_1||_1 = 1$. Having chosen $\psi_1, \ldots, \psi_{n-1}$, choose $\psi_n \in H(\Omega)$ such that $||\psi_n||$ is minimal, subject to the constraints

$$(\psi_n, \psi_m)_1 \;=\; 0, \quad m = 1,\ldots,n-1; \quad ||\psi_n||_1 \;=\; 1. \tag{10.2.8}$$

We claim that $\{\psi_n\}$ is an orthogonal set in $H(\Omega)$. Suppose that we have shown that $\{\psi_1, \ldots, \psi_{n-1}\}$ is an orthogonal set in $H(\Omega)$. Suppose that ψ_n is not orthogonal in $H(\Omega)$ to all the preceding ψ_k. Then we may choose some element g in the span of these ψ_k, normalized with

$$||g||_1 \ = \ 1, \quad (\psi_n, g) \ = \ -a \ < \ 0.$$

Given $0 < \varepsilon < 1$, let $f_\varepsilon = \sqrt{1-\varepsilon^2}\,\psi_n + \varepsilon g$. Then

$$\begin{aligned}||\psi_n||^2 \le ||f_\varepsilon||^2 \ &= \ (1-\varepsilon^2)||\psi_n||^2 - 2a\varepsilon\sqrt{1-\varepsilon^2} + \varepsilon^2||g||^2 \\ &= ||\psi_n||^2 - 2a\varepsilon + O(\varepsilon^2).\end{aligned}$$

For small ε, the right-hand side is less than $||\psi_n||^2$, a contradiction. Therefore the set $\{\psi_1, \ldots, \psi_n\}$ is orthogonal in $H(\Omega)$. It follows that the set

$$\{\phi_n = k_n^{-1}\psi_n\}_{n=1}^{\infty}, \quad k_n = ||\psi_n||$$

is orthogonal in $H(\Omega)$. Choosing a maximal such set and renumbering if necessary, we obtain the desired orthonormal basis. □

Corollary 10.2.6. *If $\Omega_1 \subset \Omega$, then the kernels K_1 for Ω_1 and K for Ω satisfy*

$$K_1(z,z) \ \ge K(z,z), \qquad z \in \Omega_1, \tag{10.2.9}$$

with strict inequality unless $H(\Omega_1) = H(\Omega)$.

Proof: Propositions 10.2.1 and 10.2.5 imply the inequality (10.2.9). Equality can only hold if the orthogonal basis for $H(\Omega)$ that is constructed in Proposition 10.2.5 is not only orthogonal in $H(\Omega_1)$ but also complete in $H(\Omega_1)$. □

For an example in which $\{\omega_m\}$ is orthogonal but not complete in $H(\Omega_1)$, see Exercise 8.

10.3 Conformal mapping, I

We have assumed throughout that Ω is a domain in $\mathbb{C}$ with the property that $\partial\Omega$ contains more than one point. In this section, we assume, in addition, that Ω is simply connected. The Riemann mapping theorem says that in this case, given a point z_0 of Ω, there is a unique conformal map f_1 of Ω onto the unit disk $\mathbb{D}$ having the properties

$$f_1(z_0) \ = \ 0, \qquad f_1'(z_0) \ > \ 0.$$

Let us renormalize by choosing $R = 1/f_1'(0)$, so that

$$f_R \ = \ Rf_1 : \Omega \to D_R = D_R(0) \tag{10.3.1}$$

is the unique conformal map of Ω onto the disk $D_R(0) = \{w : |w| < R\}$ that satisfies

$$f_R(z_0) = 0, \qquad f_R'(z_0) \;=\; 1. \tag{10.3.2}$$

We may use Proposition 10.4.1 with $\Omega' = \mathbb{D}$, together with Proposition 10.2.2 to construct an orthonormal basis for $H(\Omega)$ and compute the Bergman kernel K for Ω. The basis is

$$\omega_n(z) \;=\; \psi_n(f_R(z))f_R'(z) \;=\; \frac{n^{1/2}}{\pi^{1/2}R^n} f_R(z)^{n-1} f_R'(z), \qquad n = 1, 2, \dots,$$

and the formula for K is

$$K(z, w) \;=\; \sum_{n=1}^{\infty} \omega_n(z)\overline{\omega_n(w)}. \tag{10.3.3}$$

The conditions (10.3.2) tell us that

$$\omega_1(z_0) \;=\; \frac{1}{\pi^{1/2}R} f_R'(z_0); \qquad \omega_n(z_0) \;=\; 0, \quad n = 2, 3, \dots,$$

so, for $z = z_0$, the sum (10.3.3) collapses to

$$K(w, z_0) \;=\; \frac{1}{\pi R^2} f_R'(w). \tag{10.3.4}$$

In view of this, we have

$$K(z_0, z_0) \;=\; \frac{1}{\pi R^2}$$

and

$$f_R'(w) \;=\; \frac{1}{K(z_0, z_0)} K(w, z_0). \tag{10.3.5}$$

Since $f_R(z_0) = 0$, we may integrate (10.3.5) to obtain a formula for the conformal map in terms of the Bergman kernel.

Theorem 10.3.1. *Suppose that Ω is a simply connected domain in $\mathbb{C}$ such that $\partial\Omega$ has more than one point. Given a point z_0 in Ω, the function*

$$\Phi(z) \;=\; \frac{1}{K(z_0, z_0)} \int_{z_0}^{z} K(\zeta, z_0)\, d\zeta \tag{10.3.6}$$

is a conformal map of Ω onto the disk $D_R(0)$, where

$$R \;=\; \pi^{-1/2} K(z_0, z_0)^{-1/2}. \tag{10.3.7}$$

The map Φ is uniquely determined by the conditions

$$\Phi(z_0) = 0, \qquad \Phi'(z_0) = 1. \tag{10.3.8}$$

Three consequences of this result are the estimate

$$\left| \int_{z_0}^{z} K(\zeta, z_0)\, d\zeta \right| < \pi^{-1/2} K(z_0, z_0)^{1/2}, \tag{10.3.9}$$

the limit

$$\lim_{z \to \partial\Omega} \left| \int_{z_0}^{z} K(\zeta, z_0)\, d\zeta \right| = \pi^{-1/2} K(z_0, z_0)^{1/2}, \tag{10.3.10}$$

and the fact that the kernel function has no zeros:

Corollary 10.3.2. *The kernel function K has no zeros.*

Proof: Since f_R of (10.3.1) is conformal, its derivative has no zeros. □

For a simple, and much more direct, proof of this last result, see Exercise 3.

In light of Proposition 10.1.6 and the first area identity (1.2.9), we may look at Theorem 10.3.1 in a different way.

Theorem 10.3.3. *Suppose that z_0 is a point of Ω. The problem of finding a conformal map $\Phi : \Omega \to \mathbb{C}$ such that*

$$\Phi(z_0) = 0, \qquad \Phi'(z_0) = 1$$

and $\Phi(\Omega)$ has minimal area, has a unique solution

$$\Phi(z) = \frac{1}{K(z_0, z_0)} \int_{z_0}^{z} K(z_0, \zeta)\, d\zeta.$$

The image $f(\Omega)$ is the disk $D_R(0)$, where

$$R = \pi^{-1/2} K(z_0, z_0)^{-1/2}.$$

Remark. Theorem 10.3.3 is one of many examples in Bergman [24] recasting a classical problem as an extremal problem.

10.4 Conformal invariance and the Bergman metric

Suppose that Ω' and Ω are domains that are conformally equivalent, i.e. there is a bijective holomorphic map $\Phi : \Omega' \to \Omega$.

Proposition 10.4.1. *The map* $f \to \widetilde{f}$,

$$\widetilde{f}(z) = f(\Phi(z))\Phi'(z), \qquad z \in \Omega',$$

is a unitary map from $H(\Omega)$ *onto* $H(\Omega')$.

Proof: Note that the inverse map has the same form:

$$f(w) = \widetilde{f}(\Phi^{-1}(w))[\Phi^{-1}]'(w), \qquad w \in \Omega.$$

Distinguishing the inner products by subscripts, we only need to show that for f, g in $H(\Omega)$,

$$(\widetilde{f}, \widetilde{g})_{\Omega'} = (f, g)_{\Omega}.$$

This is true if and only if

$$|\Phi'(z')|^2 dx'\,dy' = dx\,dy \tag{10.4.1}$$

where we write $z' = x' + iy'$, $z' \in \Omega$, and $w = x + iy$, $w = \Phi(z) \in \Omega$. But this is the local form of the first area formula in (1.2.9). Thus, we have established (10.4.1). □

Proposition 10.4.2. *If* Φ *is a conformal map of* Ω' *onto* Ω*, then the Bergman kernels* K' *of* Ω' *and* K *of* Ω *are related by*

$$K'(z, w) = K(\Phi(z), \Phi(w))\Phi'(z)\overline{\Phi'(w)}. \tag{10.4.2}$$

Proof: Given z in Ω and w in Ω', let k_z and k'_w be the corresponding elements that evaluate functions at z and w, respectively. Then if f belongs to $H(\Omega')$, and $w = \Phi(z)$, then

$$(\widetilde{f}, k_z)_\Omega = \widetilde{f}(z) = f(\Phi(z))\Phi'(z) = (f, k'_w)_{\Omega'}\Phi'(z) = (f, \overline{\Phi'(z)}k'_w)_{\Omega'}.$$

Thus,

$$k_z = \widetilde{k'_w} \qquad \text{if } w = \Phi(z). \tag{10.4.3}$$

Then by (10.1.5) and (10.4.3), if $\Phi(z) = z'$ and $\Phi(w) = w'$, then

$$K(z, w) = (k_z, k_w)_\Omega = (\widetilde{k'_{z'}}, \widetilde{k'_{w'}})_{\Omega'} = K'(z', w')\Phi'(z)\overline{\Phi'(w)}. \qquad \square$$

Bergman introduced a Riemannian metric on the domain Ω by setting the distance element at z in Ω to be

$$ds^2 \;=\; K(z, \overline{z})(dx^2 + dy^2). \tag{10.4.4}$$

This is known as the *Bergman metric*.

We do not need to assume any knowledge of differential geometry to explain what this means, and to prove the conformal invariance of the metric. The point of an expression like (10.4.4) is to indicate how to compute distance, and one computes distance by first computing the length of smooth curves. Suppose that $\gamma : [a, b] \to \Omega$ is such a curve: $\gamma(t) = x(t) + iy(t)$, where x, y are differentiable functions of t. Then $\gamma' = x' + iy'$ and the length of γ with reference to the metric (10.4.4) is

$$L(\gamma) \;=\; \int_a^b K(\gamma(t), \gamma(t))^{1/2}[x'(t) + y'(t)]^{1/2}\, dt \;=\; \int_a^b K(\gamma(t), \gamma(t))^{1/2}|\gamma'(t)|\, dt.$$

The associated *distance* between two points z and w of Ω is the minimum length among such curves that begin at one point and end at the other. (Conversely, one can recover the metric from the distance function, using the observation that for points very close to one another, the distance is approximately a constant times the euclidean distance.)

Theorem 10.4.3. *The Bergman metric is a conformal invariant: suppose that Φ is a conformal map of Ω' onto Ω. Then a smooth curve γ' in Ω' and the induced curve $\gamma = \Phi \circ \gamma'$ in Ω have the same length as measured by the Bergman metrics on Ω' and Ω, respectively.*

Proof: The length of $\gamma : [a, b] \to \Omega$ is

$$\begin{aligned}&\int_a^b K(\Phi(\gamma'(t)), \Phi(\gamma'(t)))^{1/2}|[\Phi \circ \gamma]'(t)|\, dt\\ &\quad = \int_a^b \left\{K(\Phi(\gamma'(t)), \Phi(\gamma'(t)))^{1/2}|\Phi'(\gamma(t)|\right\} |\gamma'(t)|\, dt. \end{aligned} \tag{10.4.5}$$

By Proposition 10.4.2, the expression in braces is the square root of $K'(\gamma'(t), \gamma'(t))$, where K' is the Bergman kernel for Ω'. Therefore the right side of (10.4.5) is the length of the induced curve γ'. □

A *geodesic* for the Bergman metric is a curve γ that is locally of shortest length, meaning that for some $\delta > 0$, if z and w are points on γ and $|z - w| \le \delta$, then no curve from z to w has length shorter than the portion of γ from z to w. It is convenient to renormalize the parametrization of the curve so that $|\gamma'(t)| \equiv 1$, so that (with some new choice of a and b)

$$L(\gamma) \;=\; \int_a^b K(z(t), z(t))^{1/2}\, dt, \qquad |\gamma'(t)| \equiv 1. \tag{10.4.6}$$

With this parametrization, the necessary and sufficient conditions for γ to be a geodesic are the *Euler–Lagrange equations*

$$x''(t) = \frac{\partial}{\partial x}\left\{K(z(t), z(t))^{1/2})\right\}, \qquad y''(t) = \frac{\partial}{\partial y}\left\{K(z(t), z(t))^{1/2})\right\}, \tag{10.4.7}$$

see Exercise 15.

As an example, consider the unit disk $\mathbb{D}$, for which the Bergman kernel is

$$K(z, w) = \frac{1}{\pi(1 - z\overline{w})^2};$$

see Corollary 10.2.3. Thus the Bergman metric is

$$ds^2 = \frac{dx^2 + dy^2}{\pi(1 - r^2)^2}. \tag{10.4.8}$$

Up to a multiplicative constant, this is the Poincaré metric for the unit disk as a model of hyperbolic geometry; see Section 2.2.

The interval $(-1, 1)$ is a geodesic for this metric; see Exercise 16. As in Section 2.2, the remaining facts about the geometry follow by conformal invariance: rotations around the origin are conformal maps of $\mathbb{D}$ to itself, so each diameter of $\mathbb{D}$ is a geodesic. More generally, the linear fractional transformations

$$\Phi(z) = \omega \cdot \frac{z - a}{\bar{a}z - 1}, \qquad a \in \mathbb{D}, \quad |\omega| = 1,$$

are conformal maps of $\mathbb{D}$ to itself. Linear fractional transformations map any line to either a line or a circle. It can be deduced from these remarks that the remaining geodesics of $\mathbb{D}$ are precisely the circular arcs in $\mathbb{D}$ that meet the boundary at right angles. These are the "lines" for the hyperbolic geometry.

Remark. It follows from (10.4.8) that the Bergman metric for $\mathbb{D}$ blows up, as one approaches a boundary point z_0, like $1/2\sqrt{\pi}\rho$, where ρ is the distance to the boundary. This is true, not only qualitatively ($O(\rho^{-1})$) but also quantitatively, whenever the boundary is smooth in a neighborhood of z_0; see Exercise 10,

10.5 Conformal mapping, II

We saw in Section 10.3 that the Bergman kernel can be used to give an explicit formula for a conformal map of a simply connected domain onto a disk. A modified version of this statement is also true for domains that are not simply connected. In this case, the mapping takes such a domain to the plane minus a collection of horizontal

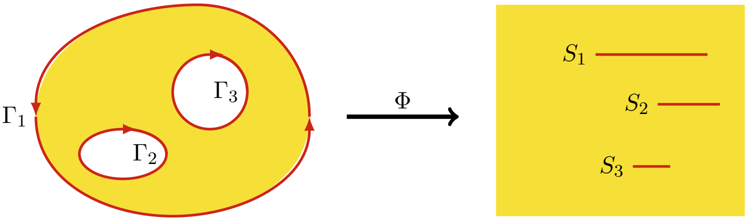

Fig. 10.1 Multiply connected domains.

slits. (For a different proof of this mapping theorem, see Exercises 16–19 of Chapter 7.)

In this section, we take Ω to be a bounded domain whose boundary $\partial\Omega$ consists of smooth simple closed curves Γ_1, Γ_1, … Γ_p, where Ω lies in the bounded domain enclosed by Γ_1; see the left side of Figure 10.1.

As we shall show,

Theorem 10.5.1. *There is a conformal map Φ of Ω onto $\mathbb{C} \setminus S$, where $S = S_1 \cup \cdots \cup S_p$ is a union of non-overlapping horizontal slits*

$$S_j \;=\; \{a_j + is \,:\, 0 \le s \le s_j\}, \qquad j = 1, 2, \ldots, p.$$

See the right side of Figure 10.1.

The conformal map Φ will be constructed using a certain modification of the Bergman kernel for Ω. We shall want a kernel with a single-valued integral in Ω.

Let $\widetilde{H}(\Omega)$ be the subspace of $H(\Omega)$ consisting of functions that are derivatives of functions that are holomorphic in Ω.

Lemma 10.5.2. *For any z_0 in Ω and f in $\widetilde{H}(\Omega)$, the integral*

$$I_f(z) \;=\; \int_{z_0}^{z} f(\zeta)\, d\zeta, \qquad z \in \Omega, \tag{10.5.1}$$

is independent of the path of integration.

Proof: Suppose that $f = g'$ in Ω. Then $I_f f - g$ is constant on any path that lies in a simply connected subdomain of Ω. It follows that $I_f - g$ is constant in each coordinate disk, hence constant. □

Corollary 10.5.3. *The subspace $\widetilde{H}(\Omega)$ is closed in $H(\Omega)$.*

Proof: If $\{f_n\}$ in $\widetilde{H}(\Omega)$ converges to $f \in H(\Omega)$, then the $\{f_n\}$ converge uniformly to f on compact subset of Ω. The corresponding integrals $\{I_{f_n}\}$, defined by (10.5.1), converge uniformly on compact subsets to a holomorphic function whose derivative is f. □

Corresponding to the closed subspace $\widetilde{H}(\Omega)$ is a Bergman kernel $\widetilde{K}$ defined as before via the reproducing property

$$\widetilde{K}(z,w) \;=\; \overline{\widetilde{k}_z(w)}, \tag{10.5.2}$$

where $\widetilde{k}_z$ is the unique element of $\widetilde{H}(\Omega)$ such that for each $f \in \widetilde{H}(\Omega)$,

$$f(z) \;=\; (f, \widetilde{k}_z).$$

Exactly as for K, we have $\widetilde{K}(z,z) > 0$ and

$$\overline{\widetilde{K}(z,w)} \;=\; \widetilde{K}(w,z); \qquad |\widetilde{K}(z,w)|^2 \le \widetilde{K}(z,z)\widetilde{K}(w,w). \tag{10.5.3}$$

Moreover, $\widetilde{k}_z$ is the orthogonal projection onto $\widetilde{H}(\Omega)$ of k_z in $H(\Omega)$, so

$$\widetilde{K}(z,z) \le K(z,z). \tag{10.5.4}$$

As with K, $\widetilde{K}(z,w)$ is holomorphic in z and anti-holomorphic in w.

The principle roles are played by

$$M(t,z) = M(t,t) + \int_t^z \widetilde{K}(\tau,z)\,d\tau \tag{10.5.5}$$

and a certain auxiliary function

$$N(t,z) \;=\; \frac{1}{\pi}\left(\frac{1}{t-z} + \lambda(z,t)\right),$$

where the function λ is holomorphic with respect to $z \in \Omega$. In fact, for any choice of z in Ω, the function

$$\Phi(t) \;=\; M(t,z) + N(t,z) \tag{10.5.6}$$

can be taken as the desired conformal map. The rest of this section is devoted to the proof of this statement.

Lemma 10.5.4. *There is a constant $\delta > 0$ such that at each point z of the boundary of Ω there is a disk of radius δ contained in Ω and tangent to $\partial\Omega$ at z, and also a disk of radius δ contained in the complement of the closure of Ω and tangent to $\partial\Omega$ at z.*

Proof: At any point of the boundary, there are two such disks of maximal radius $r_1(z)$, $r_2(z)$. Let $\varrho(z)$ be the smaller of the $r_j(z)$. Then $\varrho(z)$ is continuous on $\partial\Omega$, so it has a minimum $\delta > 0$. □

Lemma 10.5.5. *Given $t \in \Omega$, there are constants $C_1(t)$, $C_2(t)$ such that, for each z in Ω such that the distance $\rho(z)$ from z to $\partial\Omega$ is ≤ 1, we have*

$$|M(z,t)| \leq C_1(w) + C_2(w)|\log\rho(z)|. \tag{10.5.7}$$

Proof: In view of (10.5.5) and (10.5.3),

$$|M(t,z)| \leq |M(z,z)| + \widetilde{K}(z,z)^{1/2}\int_z^t \widetilde{K}(\tau,\tau)^{1/2}\,|d\tau|.$$

Therefore it is enough to estimate the integral, and, in view of (10.5.4), we may replace $\widetilde{K}(\tau,\tau)$ by $K(\tau,\tau)$. Let δ be as in Lemma 10.5.4. Let D be any disk of radius δ contained in Ω and tangent to the boundary. The set of points in Ω whose distance from the boundary is at least $\delta/2$ is compact, so it is enough to estimate the integral from the center of D along the radius ending at the boundary. By Corollary 10.2.6, $K(\tau,\tau)$ for z in D is dominated by the corresponding value of the Bergman kernel for D. Up to a translation and rotation, we may assume that $D = D_\delta(0)$. Then, by Corollary 10.2.3, we have

$$K(\tau,\tau) \;\leq\; \frac{\delta^2}{\pi(\delta^2-\tau^2)^2} \;=\; \frac{\delta^2}{\pi(\delta+\tau)^2(\delta-\tau)^2}, \qquad z \in D,\ |z| = t.$$

Therefore we want to estimate

$$\begin{aligned}\int_0^t K(s,s)^{1/2}\,ds &= \pi^{-1/2}\int_0^t \frac{\delta\,ds}{(\delta+s)(\delta-s)}\\ &= \frac{1}{2\pi^{1/2}}\log\frac{\delta+t}{\delta-t} \;=\; O(|\log(\delta-t)|).\end{aligned}$$

But $\delta - t$ is the distance $\rho(t)$ from t to the boundary. □

The construction of $\widetilde{K}$ shows that if f' belongs to $\widetilde{H}(\Omega)$ then

$$\iint_\Omega \widetilde{K}(z,w)\,f'(w)\,dm(w) \;=\; f'(z).$$

Given t not in the closure of Ω, we can take $f(z) = (t-z)^{-1}$ and

$$\iint_\Omega \frac{\widetilde{K}(z,w)}{t-w}\,dm(w) \;=\; \frac{1}{t-z}.$$

It follows that

$$\frac{d}{dt}\iint_\Omega \frac{\widetilde{K}(z,w)}{w-t}\,dm(w) \;=\; \frac{1}{(t-z)^2}.$$

Therefore

$$I(t,z) \equiv \iint_{\Omega} \frac{\widetilde{K}(z,w)}{w-t}\,dm(w) = \frac{1}{z-t} + c_k(z) \tag{10.5.8}$$

where c_k depends on the choice of z and on which component of the complement of the closure of Ω contains t, i.e. which of the curves Γ_k encloses t. (If Γ_k encloses the unbounded component, then by taking $t \to \infty$ we see that $c_k = 0$.)

Since $1/(w-t)$ is integrable, we may define $I(t,z)$ for $t \in \Omega$ by the same formula:

$$I(t,z) \equiv \iint_{\Omega} \frac{\widetilde{K}(z,w)}{w-t}\,dm(w) = \frac{1}{z-t}, \qquad t \in \Omega. \tag{10.5.9}$$

This converges (consider radial coordinates centered at t).

Lemma 10.5.6. *For t in Ω,*

$$I(t,z) = -\pi\overline{M(z,t)} + \lambda(z,t), \tag{10.5.10}$$

where λ is holomorphic in t.

Proof: Let Ω' be a domain with smooth boundary whose closure is contained in Ω, such that t is in Ω'. The integrand in (10.5.9) is regular in $\partial\Omega'$ except at $\zeta = t$. Let $D_\varepsilon = D_\varepsilon(t)$. For sufficiently small $\varepsilon > 0$, the closure $\overline{D_\varepsilon}$ is contained in Ω'. Let $\Omega'_\varepsilon = \Omega' \setminus \overline{D_\varepsilon}$.

By the Cauchy–Green formula (1.2.8),

$$2i\iint_{\Omega'} \frac{\widetilde{K}(z,w)}{w-t}\,dm(w) = \int_{\partial\Omega'_\varepsilon} \partial_{\bar w}\left\{\frac{\widetilde{K}(z,w)}{w-t}\right\} dz + 2i\iint_{D_\varepsilon} \frac{\widetilde{K}(z,w)}{w-t}\,dm(w).$$

Since the integrand is integrable over Ω, the second integral on the right goes to zero with ε. Now $\partial_{\bar w}\{(w-t)^{-1}\} = 0$ and

$$\partial_{\bar w}\widetilde{K}(z,w) = \overline{\partial_w \widetilde{K}(w,z)} = \overline{M(w,z)}.$$

Therefore

$$\begin{aligned} 2i\iint_{\Omega'} \frac{\widetilde{K}(z,w)}{t-w}\,dm(w) - 2i\iint_{D_\varepsilon} \frac{\overline{\widetilde{K}(z,w)}}{w-t}\,dm(w) \\ = \int_{\partial\Omega'} \frac{\overline{M(z,w)}}{t-w}\,dz + o(1). \end{aligned} \tag{10.5.11}$$

The second integral on the left is a holomorphic function of z, while the term on the right has limit $-\pi\overline{M}(z,t)$.

Suppose now that $\{\Omega_n\}$ is an increasing sequence of smoothly bounded domains, such that $\partial\Omega_n$ is in the $1/n$ neighborhood of $\partial\Omega$. For each such n, the preceding argument shows that

$$\iint_{\Omega_n} \frac{\widetilde{K}(z,w)}{w-t}\, dm(w) \;=\; -\pi\overline{M(z,t)} + \frac{1}{2i}\iint_{\partial\Omega_n} \frac{\widetilde{K}(z,w)}{w-t}|dw|.$$

The left side converges to $I(z,t)$, so the integrals on the right have a limit $\lambda(z,t)$ that is holomorphic in z. □

Lemma 10.5.7. *For fixed z in Ω, $I(z,t)$ is a continuous function of $t \in \mathbb{C}$.*

Proof: Let t_0 be a boundary point of Ω. Let δ be as in Lemma 10.5.4 and let D_1 contained in Ω and D_2 contained in the complement of the closure $\overline{\Omega}$ be disks of radius δ that are tangent to $\partial\Omega$ at t_0.

Write

$$\begin{aligned} I(t,z) &= \iint_{\Omega\setminus D_1} \frac{\widetilde{K}(z,w)}{w-t}\, dm(w) + \iint_{D_1} \frac{\widetilde{K}(z,w)}{w-t}\, dm(w) \\ &= I_1(t,z) + I_2(t,z). \end{aligned}$$

The Cauchy–Schwarz inequality shows that

$$|I_1(t_2,z) - I_1(t_1,z)|^2 \;\leq\; ||\widetilde{k}_z||^2 \iint_{\Omega\setminus D_1} \frac{|t_1-t_2|^2}{|w-t_1|^2|w-t_2|^2}\, dm(w). \tag{10.5.12}$$

For $t_1, t_2 \in D_1$, the integral is bounded by integration over a larger set, the complement of $D_1 \cup D_2$.

Equation (10.5.8) implies that $I(z,t)$ is uniformly continuous on the complement of Ω. Therefore it is enough to consider $I(z,t) - I(z,t_0)$ as $t \to t_0$ along the radius of D_j that ends at t_0. We change coordinates and take $\delta = 1$ and $t_0 = 0$, with the boundary of Ω vertical at 0; see Figure10.2.

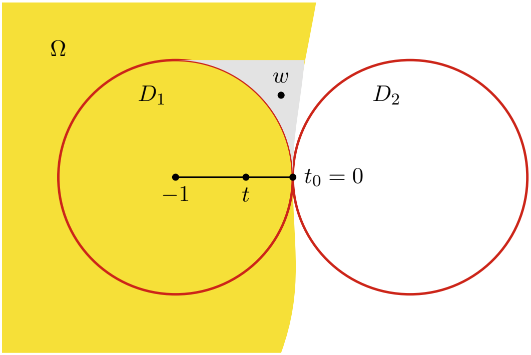

Fig. 10.2 Estimating $I_1(t,z) - I(t_0,z)$.

Note that the integrand in (10.5.12) is $O(|w|^{-4})$ as $|w| \to \infty$. It follows from this and symmetry that it is enough to consider the case $\operatorname{Im} w > 0$ and to estimate the integral over the region that is shaded gray in the figure.

With $w = x + iy$, we have

$$1 \;\le\; |w+1|^2 \;=\; (x+1)^2 + y^2, \qquad |x| \;\le\; 1 - \sqrt{1-y^2} \;=\; \frac{y^2}{2} + O(y^4).$$

At a given value of y, the range of x is contained in

$$-\frac{y^2}{2} + O(y^4) \;<x\; < \;\frac{y^2}{2} + O(y^4).$$

Therefore

$$\begin{aligned}|w-t|^2 = (x-t)^2 + y^2 \;&=\; (x^2 - 2xt) + t^2 + y^2 \ge \frac{y^2 t}{2} + t^2 + y^2 + O(y^4)\\ &\sim y^2 + t^2.\end{aligned}$$

Moreover, $|w|^2 \ge y^2$. Therefore the integral we are estimating is dominated by

$$\int_0^1 \frac{t^2 y^2\,dy}{t^2+y^2} \;=\; \int_0^{1/|t|} \frac{t\,ds}{1+s^2} \;\le\; t\int_0^{\infty} \frac{ds}{1+s^2} \;=\; \frac{t\pi}{2}.$$

This completes the argument for $I_1(z,t)$.

In the case of I_2, let us take D_1 to be the unit disk, and $t_0 = 1$, Here, it is enough to consider

$$I_2(t,z) - I_2(t^{-1},z) \;=\; \iint_{\mathbb{D}} \widetilde{K}(z,w)\left(\frac{1}{w-t} - \frac{1}{w-t^{-1}}\right) dm(w),$$

for $0 < t < 1$. We want to invoke the Cauchy–Green formula (10.5.11), which takes the form here

$$\begin{aligned}&2i \iint_{\mathbb{D}} \widetilde{K}(z,w)\left(\frac{1}{w-t} - \frac{1}{w-t^{-1}}\right) dm(w)\\ &= \int_{|z|=1} \overline{M(w,z)}\left(\frac{1}{w-t} - \frac{1}{w-t^{-1}}\right) dw. \qquad (10.5.13)\end{aligned}$$

To justify this, we need to consider the singularity of M at the boundary point 1. The formula is valid if we replace $\mathbb{D}$ by the portion of $\mathbb{D}$ that lies to the left of the arc of a circle of radius ε centered at 1. Lemma 10.5.5 implies that the integral over the arc is less than some constant times

$$\int_0^{\varepsilon} \log s\,ds \;=\; \varepsilon\log\varepsilon - \varepsilon$$

The limit as $\varepsilon \to 0$ is (10.5.13). To complete the proof, we take a small disk $D = D_\varepsilon(t)$ and use (10.5.13) to write

$$\begin{aligned} I_2(t,z) - I_2(t^{-1},z) = \frac{1}{2i}\int_{|w|=1} \overline{M(w,z)} \left(\frac{1}{w-t} - \frac{1}{w-t^{-1}}\right) dw \\ + \iint_{D_\varepsilon} \widetilde{K}(z,w) \left(\frac{1}{w-t} - \frac{1}{w-t^{-1}}\right) dm(w) \\ -\frac{1}{2i}\int_{|w|=\varepsilon} \overline{M}(w,z) \left(\frac{1}{w-t} - \frac{1}{w-t^{-1}}\right) dw \end{aligned}$$

As $\varepsilon \to 0$, the last two integrals on the right converge to 0 and to $-\pi\overline{M}(t,z)$, respectively.

On the unit circle $z = \bar{z}^{-1}$, so $dz = -\bar{z}^{-2}d\bar{z}$. Therefore, using the previous results and taking the complex conjugate, we have

$$\begin{aligned} &\overline{I_2(t,z) - I_2(t^{-1},z)} \\ &= -\frac{1}{2i}\int_{|w|=1} M(w,z) \left(\frac{1}{w^{-1}-t} - \frac{1}{w^{-1}-t^{-1}}\right) dw - \pi M(t,z). \end{aligned}$$

Since $M(w,z)$ is holomorphic with respect to $w \in \mathbb{D}$, and the residue at the pole $w = t^{-1}$ is $-\pi M(w,z)$, we have $I_2(t,z) = I_2(t^{-1},z)$. □

Let

$$N(t,z) = \frac{1}{\pi}\left(\frac{1}{t-z} + \lambda(z,t)\right), \tag{10.5.14}$$

where λ is the function in (10.5.10).

Lemma 10.5.8. *Let Γ_k be one of the curves that bound Ω. Then*

$$\lim_{t\to\Gamma_k} N(t,z) = \lim_{z\to\Gamma_k} \overline{M(t,z)} + \frac{c_k}{\pi}. \tag{10.5.15}$$

Proof: Since the function $I(t,z)$ is continuous at Γ_k, the formulas (10.5.8) and (10.5.10) must give the same value at any point of Γ_k. Thus,

$$\lim_{t\to\Gamma_k} \frac{1}{z-t} + c_k = \lim_{t\to\Gamma_k}\left[-\pi\overline{M(t,z)} + \lambda(z,t)\right]$$

which is the same as (10.5.15). □

We are now in a position to prove Theorem 10.5.1 in a more complete formulation. Fix some $z \in \Omega$ and let

$$\Phi(t) = M(t,z) + N(t,z). \tag{10.5.16}$$

Then (10.5.15) implies that

$$\lim_{t\to\Gamma_k} \operatorname{Im}\Phi(t) = \lim_{t\to C_k}\left[M(t,z)+N(t,z)-\overline{M(t,z)}-\overline{N(t,z)}\right]$$
$$= \frac{\operatorname{Im} c_k}{\pi}. \qquad (10.5.17)$$

Theorem 10.5.9. *For any choice of $z \in \Omega$, let $\Phi(t) = M(t,z) + N(t,z)$, where M is defined by (10.5.5) and N is defined by (10.5.14) and (10.5.10). Then Φ is a conformal map onto the complement in $\mathbb{C}$ of the union of disjoint horizontal slits $S_1,\ldots,S_p$.*

Proof: By construction, Φ is holomorphic on Ω, except for a simple pole at $t = z$. Lemma 10.5.8 shows that Φ is continuous up to the boundary. For any complex value a not in $\Phi(\partial\Omega)$, integrating $\Phi'(t)/[\Phi'(t) - a]$ over the boundary shows that Φ takes the value a exactly once in Ω. Therefore Φ is a conformal map of Ω onto the complement of the union of the images $\Phi(\Gamma_k)$. By (10.5.17), each such image is a horizontal slit.

To complete the proof, we need to show that distinct boundary curves Γ_k have distinct images $S_k = \Phi(\Gamma_k)$. If $j \neq k$, then we may find disjoint curves $\widetilde{\Gamma}_j$, $\widetilde{\Gamma}_k$ homotopic to Γ_j, Γ_k, respectively. The images are disjoint and enclose S_j and S_k, respectively. □

Remarks. 1. Theorem 10.5.9 contains the Riemann mapping theorem for domains with smooth boundaries. In fact, suppose that Ω has a single boundary curve, so that $\Phi(\Omega)$ is the plane with a single slit. This domain can be mapped onto the unit disk by an explicit conformal map; see Exercise 17.

2. The assumption that Ω has a smooth boundary in order for there to be a conformal map as described in Theorem 10.5.1 can be very much weakened; see Exercise 19.

10.6 The kernel function and partial differential equations

Suppose that Ω is a domain bounded by p analytic closed curves Γ_k, as in the left side of Figure 10.1. Two classic partial differential equations problems associated with such a domain are the *Dirichlet problem*: given a continuous function f on the boundary $\partial\Omega$, find a function u, continuous on the closure, such that

$$\Delta u \equiv u_{xx} + u_{yy} = 0 \text{ in } \Omega, \qquad u = f \text{ on } \partial\Omega, \qquad (10.6.1)$$

and the *Poisson problem* with *Dirichlet boundary condition*: given a bounded function g on Ω, find a function u, continuous on the closure, such that

$$\Delta u = f \text{ in } \Omega, \qquad u = 0 \text{ on } \partial\Omega. \qquad (10.6.2)$$

A function u that satisfies $\Delta u = 0$, as in (10.6.1), is said to be *harmonic*.

One useful tool here is the following form of *Green's identity* for functions that are smooth in Ω and continuous on the closure:

$$\int_{\partial\Omega}\left(\frac{\partial u}{\partial n}v - u\frac{\partial v}{\partial n}\right)|dz| \;=\; \iint_{\Omega}(v\Delta u, -u\Delta v)\;dm, \tag{10.6.3}$$

where $\partial/\partial n$ denotes differentiate in the direction of the outward normal vector.

By definition, a *Green's function* for Ω is a function $G(z, \zeta)$, $z, \zeta \in \Omega$, with the properties that as a function of z it is harmonic in $\Omega \setminus \{\zeta\}$, continuous and equal to zero at the boundary $\partial\Omega$, and has a logarithmic singularity at ζ:

$$G(z, \zeta) \;=\; \log\frac{1}{|z-\zeta|} + h(z, \zeta), \tag{10.6.4}$$

where h is harmonic with respect to z in all of Ω. The maximum principle for harmonic functions implies that the Green's function is unique.

A calculation shows that $G(z, \zeta)$ is harmonic with respect to z, apart from $z = \zeta$. It is also harmonic with respect to ζ, so by uniqueness

$$G(z, \zeta) \;=\; G(\zeta, z) \tag{10.6.5}$$

and $h(z, \zeta) = h(\zeta, z)$ is harmonic in each variable.

The existence of G can be established by solving the Dirichlet problem for $h(\cdot, \zeta)$, with boundary condition

$$h(z, \zeta) = -\log\frac{1}{|z-\zeta|}, \qquad z \in \partial\Omega.$$

For the existence of a solution, see Section 5.4.

The principal aim of this section is to connect Green's function for Ω and the Bergman kernel for Ω. As we shall see,

$$K(z, \zeta) \;=\; -\frac{2}{\pi}\frac{\partial^2 G}{\partial z\partial\overline{\zeta}}. \tag{10.6.6}$$

The importance of G is that it provides the solution to both the Dirichlet problem (10.6.1) and the Poisson problem (10.6.2).

Lemma 10.6.1. *G extends to be harmonic in a neighborhood of the boundary $\partial\Omega$.*

Proof: The assumption that the boundary curves Γ are analytic (i.e. $\Gamma(t)$ is an analytic function of t and $\Gamma' \neq 0$) implies that Γ extends to a coordinate chart in a neighborhood of any given boundary point $\Gamma(t_0)$. In this chart, the intersection with the nearby portion of $\partial\Omega$ becomes part of the real axis. Since the harmonic function G is zero on the boundary, the reflection principle says that it extends across. □

Proposition 10.6.2. *Given any continuous function f on $\partial\Omega$, the unique solution to the problem (10.6.1) is*

$$u(z) \;=\; \frac{1}{2\pi}\int_{\partial\Omega}\frac{\partial G}{\partial n}(\zeta)\, f(\zeta)|d\zeta|, \qquad z\in\Omega. \tag{10.6.7}$$

Proof: Let u be the solution of (10.6.1). For small $\varepsilon>0$, the closure $\overline{D_\varepsilon}$ of the disk $D_\varepsilon = D_\varepsilon(z)$ is contained in Ω. Let $\Omega_\varepsilon = \Omega\setminus\overline{D_\varepsilon}$. Then G and u are both harmonic in Ω_ε, so Green's formula (10.6.3) gives

$$\begin{aligned}\int_{\partial\Omega_\varepsilon}\frac{\partial G}{\partial n}(z,\zeta))\, u(\zeta)|d\zeta| &= \int_{\partial\Omega_\varepsilon} G(z,\zeta)\frac{\partial u}{\partial n}(\zeta)|d\zeta|\\ &= -\int_{\partial D_\varepsilon} G(z,\zeta)\frac{\partial u}{\partial n}(\zeta)|d\zeta|.\end{aligned} \tag{10.6.8}$$

As $\varepsilon\to 0$, on ∂D_ε, we may replace $\partial G/\partial n$ by $\partial h/\partial n = 1/\varepsilon$, so the left side of (10.6.8) has limit

$$\int_{\partial\Omega}\frac{\partial G}{\partial n}(z,\zeta) f(\zeta)|d\zeta| - 2\pi u(z).$$

Similarly, the right side of (10.6.7) is $O(\varepsilon|\log\varepsilon|)$. This proves (10.6.7). □

Proposition 10.6.3. *Given any bounded continuous function f on Ω, the unique solution to the Poisson problem (10.6.2) is*

$$u(z) \;=\; \frac{1}{2\pi}\iint_{\Omega} G(z,\zeta) f(\zeta)\, dm(\zeta). \tag{10.6.9}$$

Proof: With u defined by (10.6.9), the obstacle to a simple computation of Δu is behavior at the singularity at $\zeta = z$. We will approximate G by a family of smoother functions $\{G_\varepsilon\}$. To this end, we first choose a non-decreasing twice-differentiable function φ of one variable with the properties

$$\varphi(r) \;=\; 0, \;\text{ if } r\le 1/2; \qquad \varphi(r) = 1 \;\text{ if } r\ge 1.$$

Let

$$l(r) = \varphi(r)\log r; \qquad l_\varepsilon(r) \;=\; \varepsilon^2 l(r/\varepsilon).$$

Then with $s = r/\varepsilon$, we have

$$\begin{aligned}\Delta l_\varepsilon(r) &= \left\{\frac{d^2}{dr^2} + \frac{1}{r}\frac{d}{dr}\right\} l_\varepsilon(r)\\ &= [\varphi\ \log]''s + \tfrac{1}{s} l'(s).\end{aligned}$$

so the integral over $D_\varepsilon(0) \subset \mathbb{R}^2$ is

$$\begin{aligned} 2\pi \int_0^1 ds + 2\pi \int_0^1 [\varphi(s)\log s]'\, ds \ &= \ 2\pi \int_0^1 [s\varphi(s)\log s]'' ds \\ = \ 2\pi [s\varphi(s)\log s]' \Big|_0^1 \ &= \ 2\pi [s\phi'(s)\log s + \phi(s)] \Big|_0^1 \ = \ 2\pi\varphi(1) = 2\pi. \end{aligned}$$

With this in mind, we approximate G by

$$G_\varepsilon(z,\zeta) \ = \ -l_\varepsilon(|z-\zeta|) + h(z,\zeta).$$

Now G_ε is harmonic with respect to z for $|z-\zeta| > \varepsilon$ so with u given by (10.6.9),

$$\Delta u(z) \ = \ \frac{1}{2\pi} \int_{\Omega\cap D_\varepsilon(z)} \Delta l_\varepsilon(|z,\zeta|) f(\zeta)\, d\zeta. \tag{10.6.10}$$

For sufficiently small ε, the disk D_ε is contained in Ω. Taking into account the calculation of the integral of Δ_ε and the continuity of f, we see that the limit as $\varepsilon \to 0$ of the right side of (10.6.10) is $f(z)$. Since G_ε decreases to G, we obtain $\Delta u(z) = f(z)$. □

The principal step in getting to the identity (10.6.6) is to show that the function $-(2/\pi)\partial^2 G/\partial z \partial\bar{\zeta}$ has the reproducing property.

Proposition 10.6.4. *If f is holomorphic in Ω and continuous on the closure, then*

$$-\frac{2}{\pi} \iint_\Omega \frac{\partial^2 G}{\partial z \partial \bar{\zeta}}(z,\zeta)\, f(\zeta) dm(\zeta) \ = \ f(z). \tag{10.6.11}$$

Proof: Since

$$G(z,\zeta) \ = \ \log\frac{1}{|z-\zeta|} + h(z,\zeta),$$

it follows that

$$\begin{aligned} \frac{\partial G}{\partial z}(z,\zeta) &= -\frac{1}{z-\zeta} + \frac{\partial h}{\partial z}(z,\zeta); \\ \frac{\partial^2 G}{\partial z \partial\bar{\zeta}}(z,\zeta) &= \frac{\partial^2 h}{\partial z \partial\bar{z}}(z,\zeta). \end{aligned}$$

Therefore $\partial^2 G/\partial z \partial\bar{\zeta}$ has no singularities. Moreover, since $4\overline{\partial z}\partial\bar{z} = \Delta$, it follows that $\partial^2 G/\partial z \partial\bar{\zeta}(z,\zeta)$ is holomorphic with respect to z.

Given $z \in \Omega$, the Cauchy–Green formula (1.2.8) gives

$$\frac{1}{2i}\int_{\partial\Omega}\frac{\partial G}{\partial z}(z,\zeta)f(\zeta)\,dz \;=\; \iint_{\Omega}\frac{\partial^2 G}{\partial z\partial\overline{\zeta}}f(\zeta)\,dm(\zeta).$$

The left side is

$$=\frac{1}{2i}\int_{\partial\Omega}\left[-\frac{1}{z-\zeta}+h(z,\zeta)\right]f(\zeta)\,dz.$$

The integrand is meromorphic in Ω, so the value is π times the residue, i.e. $-\frac{\pi}{2}f(z)$.
□

Proposition 10.6.5. *The Bergman kernel $K(z,\zeta)$ for Ω is continuous up to the boundary in either variable.*

In fact, the kernel K extends analytically across the boundary. This uses again the assumption that the boundary curves are analytic arcs. For the lengthy proof, we refer to Bergman [24], Chapter 5, Section 3. This being the case, we can use K in place of f in (10.6.11) to prove the identity (10.6.6):

$$\begin{aligned} K(z,\eta) &= -\frac{2}{\pi}\iint_{\Omega}\frac{\partial^2 G}{\partial z\partial\overline{\zeta}}(z,\zeta)K(\zeta,\eta)\,dm(\zeta) \\ &= -\frac{2}{\pi}\iint_{\Omega}K(\zeta,\eta)\frac{\partial^2 G}{\partial z\partial\overline{\eta}}\,dm(\zeta). \end{aligned}$$

Taking the complex conjugate of the last integral gives

$$-\frac{2}{\pi}\iint K(\eta,\zeta)\frac{\partial^2 G}{\partial\overline{\zeta}\partial z}(z,\zeta)\,dm(\zeta) \;=\; -\frac{2}{\pi}\frac{\partial^2 G}{\partial\eta\partial\overline{z}}(\eta,z).$$

Remark. The assumption that the domain Ω is bounded by analytic curves is not, up to conformal equivalence, at all special; see Exercise 19.

Exercises

1. Suppose $\partial\Omega$ contains only one point. Show that if f is holomorphic on Ω and $\int_{\Omega}|f|^2<\infty$, then $f\equiv 0$.
2. Given points $z_1,\ldots,z_n$ in Ω and constants $a_1,\ldots a_j$, consider the interpolation problem: find a function f, holomorphic in Ω, such $f(z_j)=a_j$, $j=1,\ldots,n$ and $||f||$ is minimal among such functions. Show that the problem has a unique solution and describe its form.

3. Suppose that $H(\Omega)$ contains a non-zero function.
   (a) Show that for each $z \in \Omega$ there is a function in $H(\Omega)$ that does not vanish at z.
   (b) Show that the kernel K has no zeros.
4. Prove Proposition 10.2.2 (including the completeness of the set $\{\phi_n\}$.
5. Prove Proposition 10.2.4 (including the completeness of the set $\{\psi_n\}$.)
6. Find necessary and sufficient conditions that the function with Laurent expansion

$$f(z) = \sum_{k=-n}^{\infty} c_n z^n$$

belong to $H(A(r, R))$, where $A(r, R)$ is the annulus (10.2.6).

7. Use Corollary 10.2.3 to prove that for each disk, the Bergman kernel K has the boundary behavior

$$K(z, z) = \tfrac{1}{4\pi}\rho^{-2} + O(\rho)^{-1})$$

as the distance ρ from z to the boundary goes to zero. Note that this is independent of the radius of the disc.

8. Let Ω be the unit disk $\mathbb{D}$ and Ω_1 be $\mathbb{D}$ with the segment $[0, 1)$ removed. Show that each orthogonal basis for $H(\Omega)$ is an orthonormal set for $H(\Omega_1)$, but is not complete in $H(\Omega_1)$.
9. Suppose K is the kernel function for a domain Ω in $\mathbb{C}$. Show that for any set $\{z_1, z_2, \ldots, z_n\}$ of distinct points of Ω, and any set of constants $\{t_1, t_2, \ldots t_n\}$ in $\mathbb{C}$,

$$\sum_{1 \le j,k \le n} K(z_j, z_k) t_j \bar{t}_k \ge 0.$$

10. Suppose that the domain Ω has a smooth boundary. Use Corollary 10.2.6 and Exercise 7 to show that the kernel function $K(z, z)$ has the same boundary behavior, pointwise, as a disk. Hint: if $z_0 \in \partial\Omega$, then there are disks D_1 and D_2 such that D_1 is contained in Ω, D_2 is disjoint from Ω, and the boundaries meet at z_0. An inversion $\Phi(z) = 1/(z - z_1)$, where z_1 is not in the closure of Ω, is a conformal map with the property that $D_3 = \Phi(D_2)$ is a disk that contains $\Phi(\Omega)$ and is tangent to $\Phi(\Omega)$ at $\Phi(z_0)$.
11. A linear map U from a Hilbert space H to itself is said to be *unitary* if it is surjective and satisfies

$$(Uf, Ug) = (f, g), \qquad \text{all } f, g \in H.$$

   (a) Suppose that H has an orthogonal basis φ_n. Show that a linear map $U : H \to H$ is unitary if and only if $\{\psi_n\}$ is an orthomormal basis, where $\psi_n = U\phi_n$.
   (b) Suppose that $U : H(\Omega) \to H(\Omega)$ is unitary. Prove that U has a kernel K_U, i.e. a continuous function $K_U(z, w)$ such that for any $f \in H(U)$,

$$Uf(z) = \iint_{\Omega} K_U(z, w)\, f(w)\, dm(w).$$

12. An *orthogonal projection* in a Hilbert space H is a linear map $P : H \to H$ with the property that $P(H) = H_1$ is a closed subspace of H, and $Pf = f$ if $f \in H_1$ and $Pg = 0$ if g is orthogonal to H_1 (meaning that $(f, g) = 0$ for each $f \in H_1$. Prove that if $P : H(\Omega) \to H(\Omega)$ is a projection, then P has a kernel K_P:

$$Pf(z) = \iint_{\Omega} K_P(z, w)\, f(w)\, dm(w).$$

13. Compute the Bergman kernel for the upper half-plane $\mathbb{H}$.
14. The *Hardy space* H^2 is defined to be the space of functions f that are holomorphic in $\mathbb{D}$ and satisfy

$$\sup_{0 \le r < 1} \frac{1}{2\pi} \int_0^{2\pi} |f(re^{i\theta})|^2 \, d\theta < \infty. \tag{10.6.12}$$

(a) Show that this is a Hilbert space, and the square of the norm, $||f||_H^2$ is (10.6.12).

(b) Verify that if $f(z) = \sum_{n=0}^{\infty} a_n z^n$, then

$$||f||_H^2 = \sum_{n=0}^{\infty} |a_n|^2.$$

(c) Deduce that the inner product in H^2 of $f = \sum a_n z^n$ and $g = \sum b_n z^n$ is

$$(f, g)_H = \sum_{n=0}^{\infty} a_n \bar{b}_n.$$

(d) Use (c) to select an orthonormal basis $\{\phi_n\}_{n=0}^{\infty}$ for H^2, and compute

$$K_H(z, w) = \sum_{n=0}^{\infty} \phi_n(w)\, \overline{\phi_n(z)}.$$

(e) Show that K_H has the reproducing kernel property: if $f \in H^2$, then

$$f(z) = (f, K(z, \cdot))_H.$$

(f) Prove that

$$\overline{K}_H(z, w) = K(w, z); \qquad |K_H(z, w)|^2 \le K_H(z, z) K_H(w, w).$$

15. (a) Given a positive smooth function $F(x, y, \dot{x}, \dot{y})$, consider the problem of minimizing the integral

$$\int_0^1 F(x(t), y(t), x'(t), y'(t))\, dt$$

for curves $\gamma(t) = (x(t), y(t))$ in the plane that have fixed endpoints (x_0, y_0) and (x_1, y_1). If such a curve is minimal, and $\nu(t) = (\xi(t), \eta(t))$ is a smooth curve that begins and ends at (0, 0), then the value of the integral is not decreased by replacing (x, y) by $(x + \varepsilon\xi, y + \varepsilon\eta)$. Therefore a necessary condition on γ is that

$$\frac{d}{d\varepsilon}\bigg|_{\varepsilon=0} \left\{ \int_a^b F(x + \varepsilon\xi, y + \varepsilon\eta)\, dt \right\}$$

Deduce from this, integration by parts, and the boundary conditions, the Euler–Lagrange equations

$$\frac{d}{dt}\{F_{\dot{x}}\} = F_x; \qquad \frac{d}{dt}\left\{F_{\dot{y}}\right\} = F_y.$$

(b) Suppose that the function in part (a) has the form

$$F(x, y, \dot{x}, \dot{y}) = G(x, y)(\dot{x}^2 + \dot{y}^2)^{1/2}.$$

For the minimization problem we may, and shall, choose the parametrization of the curve $(x(t), y(t))$ to satisfy

$$[x'(t)^2] + [y'(t)]^2 = 1.$$

Find the form of the Euler–Lagrange equations for curves with this parametrization and verify (10.4.7).

16. Calculate the Euler–Lagrange equations (10.4.7) for $\Omega = \mathbb{D}$ and verify that the interval $(-1, 1)$ is a geodesic.
17. Map the plane minus a single (straight) slit conformally onto the unit disk. (One may as well take the slit to be the interval [-1,1].)
18. (a) Compute the Green's function for the unit disk. Hint: what is $G(z, 0)$ in this case?

    (b) Use the result in (a) to verify (10.6.6) for $\mathbb{D}$.
19. Suppose that Ω is a bounded domain in the plane, and suppose that the complement of Ω consists of an unbounded component Ω_1 and $p - 1$ bounded components Ω_1, ..., Ω_p. Show that Ω is conformally equivalent to a domain whose boundary consists of disjoint closed analytic curves. Hint: in case $p = 1$ there are two ways to proceed: (i) use the Riemann mapping theorem; (ii) choose a point $z_1 \in \Omega$, invert the plane by $z \to (z - z_1)^{-1}$, map the image of Ω to $\mathbb{D}$, and follow this by the inversion $w \to w^{-1}$.

20. Prove, by applying the Riemann mapping theorem p times, the converse of Theorem 10.5.1: If Ω is the complement in the plane of p disjoint finite vertical closed slits, then there is a conformal map of Ω to a domain bounded by p analytic Jordan curves.

Remarks and further reading

The classic exposition of this material is in Bergman [24]. It covers the material in this chapter much more completely. As we noted in the introduction, the Bergman kernel has close connections to other natural conformal invariants associated to a domain in $\mathbb{C}$. These connections are exhaustively investigated in [24]. Other topics there include further applications to partial differential equations, potential theory, and functions of two complex variables. Krantz [125] contains some more recent developments.

The Hilbert space $H(\mathbb{D})$ of square-integrable holomorphic functions in the disk is often denoted A^2. It has Banach space generalizations A^p, the holomorphic functions that are p-th power integrable, $1 \leq p < \infty$, as well as non-Banach space versions with $0 < p < 1$. These are known as *Bergman spaces*. There is now an extensive body of knowledge concerning their structure and properties; see Duren and Schuster [65].

Chapter 11
Theta functions

A polynomial $P(z, w)$ in two complex variables induces a complex curve

$$C_P = \{(z, w) \in \mathbb{C} \times \mathbb{C} : P(z, w) = 0\}.$$

This can generally be extended in a natural way to a subset of $\mathbb{S} \times \mathbb{S}$. If P is irreducible, then $C_P \subset \mathbb{S} \times \mathbb{S}$ is a compact Riemann surface. Conversely, every compact Riemann surface arises in this way; see [6] or [22], for the case of the Riemann surface of a function defined on a subset of $\mathbb{C}$. This surface carries a complex structure, so an object of natural interest is the associated *function field*: the field of meromorphic functions on C_P. The study of this function field leads naturally to the study of the associated theta functions.

This chapter is a brief introduction to a vast topic. The basic classification of the curves C_P is the topological invariant called the *genus*. The curve in $\mathbb{S}^2$ is a compact oriented smooth manifold of real dimension two. Therefore, topologically, it is either a sphere (genus 0), a torus (genus 1), or a torus-like figure with more than one hole (genus > 1). Figure 9.1 in Chapter 9 illustrates the cases genus 0, 1, and 2.

Genus 0 presents no particular difficulty. The case of genus 1 is the case of elliptic curves, which are treated in some detail in many presentations, including [21] and [22]. In this chapter, we discuss the general case, concentrating on genus ≥ 2. However, we discuss in detail only the case of hyperelliptic curves, where the necessary machinery can be calculated explicitly. This avoids long excursions into more general and more abstract issues while providing some insight into the overall picture.

R. Beals and R. S. C. Wong, *More Explorations in Complex Functions*, Graduate Texts in Mathematics 298, https://doi.org/10.1007/978-3-031-28288-1_11

11.1 Hyperelliptic curves

A *hyperelliptic curve* is one that has one of the two forms

$$w^2 = P_{2g+1}(z), \qquad P_{2g+1}(z) = \prod_{j=1}^{2g+1}(z - r_j) \tag{11.1.1}$$

or

$$w^2 = P_{2g+2}(z), \qquad P_{2g+2}(z) = \prod_{j=1}^{2g+2}(z - r_j), \tag{11.1.2}$$

where in either case the roots r_j are assumed to be distinct. Here, $g \geq 2$; the analogous curve with $g = 1$ is termed an *elliptic* curve. It is easy to construct the Riemann surface. The general process will be clear if we take the case $g = 2$ and specialize to the case of real roots. In the degree six case, we have six real roots

$$r_1 < r_2 < r_3 < r_4 < r_5 < r_6.$$

Slit the plane from r_{2j-1} to r_{2j} $j = 1, 2, 3$, and let $\mathbb{C}^+$ denote the plane with the slits $[r_{2j-1}, r_{2j}]$ removed. It is not difficult to see that a branch of the square root

$$\sqrt{P_6(z)}$$

can be chosen on $\mathbb{C}^+$. We choose the branch that is positive as $z \to +\infty$, $z \in \mathbb{R}$. We then take a second copy $\mathbb{C}^-$ of the slit plane, and choose the other branch of $\sqrt{P}$ on $\mathbb{C}^-$ (Figure 11.1). Extend these choices to the corresponding slit spheres $\mathbb{S}^\pm$. The value of $\sqrt{P}$ varies smoothly if we start in one copy of the slit sphere, cross one of the slits, and continue on from the corresponding slit on the other copy. Therefore it is natural to join the two copies across the slits and consider $\sqrt{P}$ as a continuous function on the resulting figure. This figure is, topologically, a surface with two holes; see Figure 11.1 (Compare this with the analogous construction for P of degree 2.) In the case (11.1.1) with $g = 2$ and five distinct real roots $z_1 < \cdots < z_5$, we set $z_6 = \infty$ and proceed in the same way.

When the roots are not assumed real, then in place of the slits we find three non-intersecting curves that join pairs of roots. For general g, this construction produces a surface with g holes, i.e. with genus g—often described as a "sphere with g handles attached."

To complete the picture, we want to have a complex structure on the resulting surface. At any point z_0 of $\mathbb{S}^\pm$, minus the slits, we have the usual choice of coordinate z in a neighborhood. A slit can be considered as having two sides, corresponding to the opening depicted in Figure 11.1. For a point on either side of the slit, there is a neighborhood that lies partially in $\mathbb{S}^+$ and partially in $\mathbb{S}^-$. The local coordinate z works in that neighborhood. Finally, near a root r_j that is the endpoint of a slit,

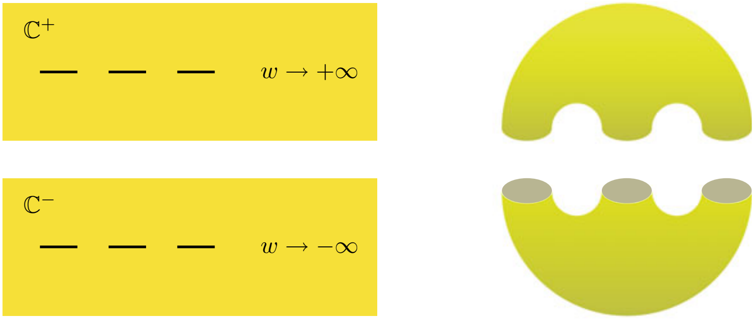

Fig. 11.1 Slit planes; connecting the slit spheres.

a determination of $w = \sqrt{P}$, continued around r_j, takes us from one copy of $\mathbb{S}$ onto the other and back, once again providing a coordinate. Covering the curve with appropriate coordinate neighborhoods shows that it is a Riemann surface in the sense of Chapter 6.

In general, if $P(z, w)$ is irreducible, the corresponding curve has the same topology as one of the hyperelliptic curves just described—a surface with g holes. However, the various curves for a given genus $g > 2$ may have different complex structures: see Farkas and Kra [67]. The point of Teichmüller theory is to construct a natural parametrization of the family of complex structures on a surface with a given topology; see Chapter 9.

11.2 Cycles and differentials

We return here to a hyperelliptic curve Γ_g of genus $g > 1$, associated to the polynomial P_{2g+1}. We state without proof a number of facts about the differential topology of Γ_g. Each will be illustrated in the case $g = 2$; the reader may construct similar illustrations to verify the statements in other cases.

Our terminology "curve" here conflicts with our previous use of "curve" as a mapping γ from a real interval into $\mathbb{C}$ or into a Riemann surface, or to the iamge of γ (with an orientation coming from the parametrization). In this chapter, we will term such a map γ or, its image, a *path*. A smooth, closed path that does not intersect itself will be called a *cycle*.

It is possible to find cycles $a_1, a_2, \ldots, a_g$ and $b_1, b_2, \ldots, b_g$ that have *intersection numbers*

$$a_j \cdot b_j = 1, \quad a_j \cdot a_k = a_j \cdot b_k = b_j \cdot b_k \ = \ 0 \ \text{ if } \ j \neq k. \tag{11.2.1}$$

The intersection number $a \cdot b$ of two cycles a and b is defined as follows: $a \cdot b = 0$ if the cycles do not meet; if a and b meet at a single point p, then it is assumed that the the tangents $\dot{a}$ and $\dot{b}$ in the forward direction along a and b meet at an angle, and $a \cdot b$ is ± 1 according to the sign of the angle from $\dot{a}$ to $\dot{b}$ (in local coordinates).

For a curve of genus 2, this is illustrated on the left in Figure 11.2. Such cycles are a *homology basis* for Γ_g, which means that every closed path in Γ_g is *homologous* to a path $\sum[m_j a_j + n_j b_j]$, where the integer coefficients indicate repeating the cycle that number of times (in the opposite direction if the coefficient is negative), and the addition refers to concatenating the paths— following one by another. Homologous means homotopic, cycle by cycle, but in homology, the order does not matter since we will be concerned only with integration of holomorphic functions along the path (or, more generally, closed 1-forms). Then homotopic paths yield the same integral, and addition is commutative.

Let us spell this out in the case illustrated in Figure 11.2. Cutting the curve Γ− along the cycles a_j, b_j, i.e. going to the complement of the union of the (images of) the a_j, b_j leads to a simply connected region $\widetilde{\Gamma}_g$ that is topologically a $4g$-sided polygon with boundary,

$$a_1 b_1 a_1^{-1} b_1^{-1} a_2 b_2 a_2^{-1} b_2^{-1} \cdots a_g b_g a_g^{-1} b_g^{-1}.$$

This is illustrated on the right in Figure 11.2, for the case $g = 2$. Points on a_j^{-1} are identified with points on a_j, and so on, to form Γ_g topologically.

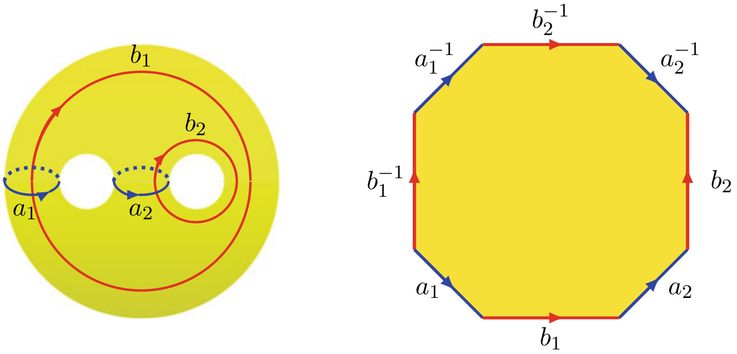

Fig. 11.2 A homology basis and the representation as a polygon.

Suppose that γ is a closed path in the curve Γ_g on the left in figure 11.2. Let Ω denote the interior of the octagon on the right. If the path γ stays in Ω, then it is homotopic to a constant map, and counts as zero. Otherwise, each connected part of the curve that lies in the interior is homotopic to a path that lies on the boundary. Thus, γ is homotopic to a path that lies on the union of the cycles. Starting at the intersection of two cycles, the path is homologous to a path that follows one of those cycles, or its inverse, an integer number of times. Thus, eventually, we have a unique homology representation of all of the path γ in the form $m_1 a_1 + m_2 a_2 + n_1 b_1 + n_2 b_2$ for some integers m_j, n_j, not necessarily positive.

A fundamental question concerning a curve Γ_g is to determine the field of functions that are meromorphic on Γ_g. A general picture is given by the following result.

Proposition 11.2.1. *If φ is a non-constant meromorphic function on Γ_g, then the number of zeros equals the number of poles (each counted according to multiplicity), and the number of poles is at least two.*

Proof: Since Γ_g is compact, φ must have at least one pole. We may assume that there are no zeros or poles on the cycles $\{a_j\}$ and $\{b_j\}$. (Otherwise, simply move the offending cycles slightly.) Integrate φ'/φ over the boundary $\partial\widetilde{\Gamma}_g$. The integral over a_j and the integral over a_j^{-1} cancel, and the same is true for the b_j. Therefore the number of zeros equals the number of poles. (In particular, there is at least one zero.) Integrate φ itself over the boundary to see that the sum of the residues is zero. Therefore the number of poles, counting multiplicity, must be greater than one. □

Replacing φ in the previous argument by $\varphi - c$, any $c \in \mathbb{C}$, we obtain

Corollary 11.2.2. *A non-constant meromorphic function takes each value (finite or infinite) the same number of times (counting multiplicity).*

We are still some distance from showing that non-constant meromorphic functions *exist*. For this, we need an excursion into differentials and theta functions.

In this context, the term used for a 1-form on Γ_g is a *differential*. In local coordinates, a differential ω has the form

$$\omega \;=\; f(z)\,dz + g(z)\,dz. \tag{11.2.2}$$

The differential ω is said to be *holomorphic*, or *Abelian of the first kind* if, in each such local representation, f is holomorphic and $g = 0$. In particular, as we shall see, the differentials

$$\eta_j \;=\; \frac{z^{j-1}}{\sqrt{P}}\,dz, \quad j = 1, 2, \ldots g, \tag{11.2.3}$$

$P = P_{2g+1}$ or $P = P_{2g+2}$, are holomorphic.

A holomorphic differential $\omega = f(z)dz$ is clearly closed: $d\omega = 0$; see Section 1.2. The same is true of its complex conjugate $\overline{f(z)dz}$.

Let us see that the differentials (11.2) are actually holomorphic, despite the apparent singularities at the zeros of $w = \sqrt{P(z)}$. Recall that, at these zeros, w itself can be taken as the local coordinate. Now

$$dz \;=\; \frac{dz}{dw}\,dw \;=\; \left(\frac{dw}{dz}\right)^{-1} dw \;=\; \frac{w}{\frac{1}{2}P'}\,dw.$$

Thus, near the zeros of P,

$$\eta_j \;=\; \frac{2z^{j-1}}{P'(z)}\,dw.$$

The roots are assumed to be distinct, so $P'(z)$ does not vanish near a root.

A natural question is: why stop at $j = g$? We need to examine what happens as $z \to \infty$. If $P = P_{2g+2}$, then as $z \to \infty$, $w = \pm z^{g+1} + \ldots$. Taking $\zeta = 1/z$ as an appropriate coordinate,

$$\eta_j = \frac{z^{j-1}}{w} dz = \pm \frac{\zeta^{1-j}[\zeta^{g+1} + \ldots]}{\zeta^2} d\zeta = \pm \zeta^{g-j}[1 + O(\zeta)]\, d\zeta,$$

so the condition $j \leq g$ is necessary for holomorphy at $z = \infty$. In the case $P = P_{2g+1}$, the point at ∞ is itself the endpoint of a slit; in this case, $\zeta = \sqrt{z}$ can be taken as a local coordinate, and a similar calculation shows that $j \leq g$ is precisely the necessary and sufficient condition for regularity at ∞; see Exercise 2.

Fix a point p_0 that does not lie on any of the cycles a_j, b_j. If ω is a holomorphic differential, we may define

$$f(p) = \int_{p_0}^{p} \omega. \tag{11.2.4}$$

Lemma 11.2.3. *Suppose that ω and ω' are closed differentials. Then*

$$\int_{\widetilde{\Gamma}_g} \omega \wedge \omega' = \int_{\partial \widetilde{\Gamma}_g} f(z)\omega' = \sum_{j=1}^{g} [A_j B'_j - A'_j B_j], \tag{11.2.5}$$

where f is defined by (11.2.4) and

$$A_j = \int_{a_j} \omega, \quad B_j = \int_{b_j} \omega, \quad A'_j = \int_{a_j} \omega', \quad B'_j = \int_{b_j} \omega'. \tag{11.2.6}$$

Proof: Since $\omega \wedge \omega' = d(f\omega')$, the first equality in (11.2.5) follows from Stokes's theorem. Next,

$$\int_{\partial \widetilde{\Gamma}_g} f\omega' = \sum_{j=1}^{g} \left[\int_{a_j} + \int_{a_j^{-1}} + \int_{b_j} + \int_{b_j^{-1}} \right] f\omega'. \tag{11.2.7}$$

Given corresponding points p on a_j and p' on a_j^{-1}, the segment from p to p' is a cycle in Γ_g that is homologous to b_j; see Figure (11.3), so

$$f(p_j) - f(p'_j) = -\int_{b_j} \omega = -B_j. \tag{11.2.8}$$

Similarly, if q_j and q'_j are corresponding points on b_j and b_j^{-1},

$$f(q_j) - f(q'_j) = \int_{a_j} \omega = A_j. \tag{11.2.9}$$

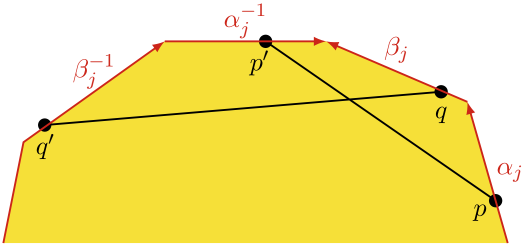

Fig. 11.3 Integrating $\omega \wedge \omega'$.

Therefore (11.2.7) is

$$\int_{\partial \widetilde{\Gamma}_g} f\omega' = \sum_{j=1}^{g} \left\{ -B_j \int_{a_j} \omega' + A_j \int_{b_j} \omega' \right\}$$
$$= \sum_{j=1}^{g} [A_j B'_j - A'_j B_j]. \qquad \square$$

The numbers A_j, B_j in (11.2.6) are called the *a periods* and *b periods* of ω, respectively.

Suppose that $\omega = f dz$ is holomorphic. Then $\omega \wedge \overline{\omega} = |f|^2 dz \wedge \overline{dz}$. In local coordinates $z = x + iy$,

$$dz \wedge \overline{dz} = (dx + i dy) \wedge (dx - i dy) = -2i dx \wedge dy.$$

Therefore (11.2.5) implies that if $\omega \neq 0$,

$$0 < \frac{i}{2} \int_{\widetilde{\Gamma}_g} \omega \wedge \overline{\omega} = \frac{i}{2} \sum_{j=1}^{g} [A_j \overline{B}_j - \overline{A}_j B_j]$$
$$= -\text{Im} \left[\sum_{j=1}^{g} A_j \overline{B}_j \right]. \tag{11.2.10}$$

We have proved

Corollary 11.2.4. *(a) If ω is a non-zero holomorphic differential on Γ_g, with a and b periods A_j, B_j, then*

$$\text{Im} \left[\sum_{j=1}^{g} A_j \bar{B}_j \right] < 0. \tag{11.2.11}$$

(b) If the a periods of a holomorphic differential ω all vanish, then $\omega = 0$.

Proposition 11.2.5. *The space of holomorphic differentials on Γ_g has dimension g.*

Proof: The g holomorphic differentials η_j of (11.2) are linearly independent, so the dimension is at least g. The matrix A with entries

$$A_{jk} \;=\; \int_{a_j} \eta_k \tag{11.2.12}$$

is non-singular, since otherwise some non-trivial linear combination of the η_j would have a periods zero, a contradiction. But this implies that given any holomorphic differential ω, there is a linear combination ω' of the η_j having the same a periods as ω, so $\omega = \omega'$. Thus, the η_j are a basis for the holomorphic differentials. □

We define a new basis $\{\omega_j\}$ of holomorphic differentials by setting

$$\omega_j \;=\; 2\pi i \sum_{k=1}^{g} (A^{-1})_{kj} \eta_k. \tag{11.2.13}$$

The $\{\omega_j\}$ are *canonically dual* to the basis of cycles $\{a_j\}$ in the sense that

$$\int_{a_j} \omega_j \;=\; 2\pi i; \qquad \int_{a_j} \omega_k \;=\; 0 \quad \text{if } j \neq k. \tag{11.2.14}$$

Having chosen this canonically dual basis, we let

$$B_{jk} \;=\; \int_{b_k} \omega_j. \tag{11.2.15}$$

Theorem 11.2.6. *The matrix $B = (B_{jk})$ satisfies*

(a) B is symmetric: $B_{jk} \;=\; B_{kj}$;

(b) $\operatorname{Re} B \;<\; 0$, i.e. $\operatorname{Re} B$ is negative definite.

Proof: (a) Both ω_j and ω_k are holomorphic so $\omega_j \wedge \omega_k = 0$. Therefore (11.2.5) becomes

$$0 \;=\; 2\pi i B_{kj} - 2\pi i B_{jk}.$$

(b) Suppose that $\eta = \sum_{j=1} c_j \omega_j \neq 0$, where the coefficients c_j are real. The a and b periods of η are

$$A_j \;=\; 2\pi i\, c_j, \qquad B_j \;=\; \sum_{k=1}^{g} c_k B_{kj}.$$

Therefore (11.2.10) for η, $\overline{\eta}$ becomes

$$0 > \operatorname{Im}\left[\sum_{k=1}^{g} A_k \overline{B}_k\right] = \operatorname{Im}\left[\sum_{j,k=1}^{g} 2\pi i c_k c_j B_{kj}\right] = 2\pi \operatorname{Re}\left[\sum_{j,k=1}^{g} B_{kj} c_j c_k\right] \quad \square$$

A symmetric matrix $B = (B_{jk})$ such that $\operatorname{Re} B < 0$ is called a *Riemann matrix.*

As noted above, holomorphic differentials are called *Abelian differentials of the first kind*. In addition, there are *Abelian differentials of the second kind* $\omega_p^{(n)}$, $n \geq 1$. These are meromorphic with a pole of order $n + 1$ at p and an expansion of the form

$$\left[\frac{1}{(z-p)^{n+1}} + \frac{a_n}{(z-p)^n} + \dots\right] dz.$$

if z is not a zero of P, and is not ∞ if P has odd degree. In the excluded cases, the expansion near p can be written in terms of w. Finally, there are *Abelian differentials of the third kind* ω_{pq}, where p and q are distinct points of Γ_g near which the coefficient has a simple pole with residues 1 and -1, respectively. For the existence of the differentials $\omega_p^{(n)}$ and ω_{pq}, see Exercises 3 and 4.

Each such differential is unique up to the addition of a holomorphic differential. We have already chosen a basis of differentials of the first kind, normalized by

$$\int_{a_j} \omega_k = \delta_{jk}. \tag{11.2.16}$$

Recall (11.2.15): the b-periods are

$$\int_{b_k} \omega_j = B_{jk} \tag{11.2.17}$$

We normalize the differentials of second and third kinds by

$$\int_{a_j} \omega_p^{(n)} = 0; \qquad \int_{a_j} \omega_{pq} = 0, \quad j = 1, 2, \dots, g. \tag{11.2.18}$$

Let $\omega_j = \varphi_j dz$ in a neighborhood of p. Then the b_j periods are

$$\int_{b_j} \omega_p^{(n)} = \frac{1}{n\,!}\frac{d^{n-1} f_j}{dz^{n-1}}\varphi_j(p); \qquad \int_{b_j} \omega_{pq} = \int_q^p \omega_j, \quad j = 1, 2, \dots, g. \tag{11.2.19}$$

Here, the basis forms ω_j are assumed to have the form $\omega_j = f_j(z)\,dz$ in a neighborhood of the point p. The proof is similar to the proof of Lemma 11.2.5; see Exercise 5

11.3 Theta functions and Abel's theorem

Let B be the Riemann matrix associated to the cycles $\{a_j\}$, $\{b_j\}$ on the hyperelliptic curve Γ_g. The associated *theta function* $\theta(z)$, $z \in \mathbb{C}^g$, is

$$\theta(z) \;=\; \theta(z|B) \;=\; \sum_{N\in\mathbb{Z}^g} \exp\left(\frac{1}{2}\langle BN, N\rangle + \langle N, z\rangle\right), \tag{11.3.1}$$

where

$$\langle BN, N\rangle \;=\; \sum_{j,k=1}^{g} B_{jk}N_kN_j, \qquad \langle N, z\rangle \;=\; \sum_{j=1}^{g} N_j z_j.$$

Let $-b < 0$ be the largest (i.e. least negative) eigenvalue of the real symmetric matrix $\operatorname{Re} B$. Then

$$|\langle BN, N\rangle| \le \langle \operatorname{Re} BN, N\rangle \;\le\; (-b)|N|^2 \;=\; (-b)\sum_{j=1}^{g} N_j^2;$$
$$|\langle N, z\rangle| \le |N||z|.$$

The sum in the definition (11.3.1) converges uniformly on compact sets in $\mathbb{C}^g$; see Exercise 6. Therefore θ is an entire function of z. Note that θ is an even function of z. Change N to $-N$ in (11.3.1):

$$\theta(-z) \;=\; \theta(z). \tag{11.3.2}$$

Let $\{e_j\}$ be the standard basis vectors in $\mathbb{C}^g$, and let $f_j = Be_j$.

Proposition 11.3.1. *The function θ satisfies*

$$\theta(z + 2\pi i e_k) = \theta(z); \tag{11.3.3}$$
$$\theta(z + f_k) = \exp(-\tfrac{1}{2}B_{kk} - z_k)\,\theta(z). \tag{11.3.4}$$

Proof: The identity

$$\langle N, z + 2\pi i e_j\rangle \;=\; \langle N, z\rangle + 2N_j\pi i$$

implies (11.3.3). Next,

$$\begin{aligned}
\tfrac{1}{2}\langle BN, N\rangle + \langle N, z + f_k\rangle \;&=\; \frac{1}{2}\langle BN, N\rangle + \langle N, z + Be_k\rangle \\
&=\; \tfrac{1}{2}\langle B(N + 2e_k), N\rangle + \langle N, z\rangle \\
&=\; \tfrac{1}{2}\langle B(N + e_k), N + e_k\rangle - \tfrac{1}{2}\langle Be_k, e_k\rangle + \langle N + e_k, z\rangle - \langle e_k, z\rangle \\
&=\; \left[\tfrac{1}{2}\langle B(N + e_k), N + e_k\rangle + \langle N + e_k, z\rangle\right] - \tfrac{1}{2}B_{kk} - z_k.
\end{aligned}$$

Since summing a function of N over N and summing over $N + e_k$ give the same result, this proves (11.3.4). □

These equations show that the theta function has *periods* $\{2\pi i e_k\}$ and *quasiperiods* $\{f_k\}$. More generally, any element of the *period lattice*

$$\Lambda = \Lambda(B) = \left\{2\pi i N + BM \,:\, M, N \in \mathbb{Z}^g\right\} \tag{11.3.5}$$

is a period or quasiperiod of θ. The transformation laws (11.3.3), (11.3.4) generalize to

$$\theta(z + 2\pi i N + BM) = \exp(-\tfrac{1}{2}\langle BM, M\rangle - \langle N, z\rangle, \quad M, N \in \mathbb{Z}^g. \tag{11.3.6}$$

The *Jacobi variety*, or *Jacobian* $J(\Gamma_g)$ of the curve G_g is the $2g$ torus that is the quotient of C^g by the lattice (11.3.5):

$$J(\Gamma_g) = C^g/\Lambda.$$

Put differently, $J(\Gamma_g)$ is the set of equivalence classes of elements of $\mathbb{C}^{2g}$ under the equivalence relation

$$w \equiv w' \text{ if and only if } w - w' \text{ belongs to } \Lambda. \tag{11.3.7}$$

More generally, we may consider a *theta function with characteristics* $\alpha, \beta \in \mathbb{R}^g$:

$$\theta[\alpha, \beta](z) = \exp\left(\frac{1}{2}\langle B\alpha, \alpha\rangle + \langle \alpha, z + 2\pi i\beta\rangle\right)\theta(z + 2\pi i\beta + B\alpha). \tag{11.3.8}$$

This has a representation similar to (11.3.1):

$$\theta[\alpha, \beta](z) = \sum_{N\in\mathbb{Z}^g} \exp\left(\frac{1}{2}\langle B(N + \alpha, N + \alpha\rangle + \langle N + \alpha, z + 2\pi i\beta\rangle\right). \tag{11.3.9}$$

The transformation law generalizes (11.3.6):

$$\begin{aligned} &\theta[\alpha, \beta](z + 2\pi i N + BM) \\ &= \exp\left(-\tfrac{1}{2}\langle BM, M\rangle - \langle M, z\rangle + 2\pi i[\langle N, \alpha\rangle - \langle M, \beta\rangle]\right)\theta[\alpha, \beta](z). \end{aligned} \tag{11.3.10}$$

The proof is left as Exercise 7.

A case of particular interest is that of a *half-period*: when each α_j and β_j is either 0 or 1/2 but not all are 0. A half-period is said to be *even* or *odd* if $4\langle\alpha, \beta\rangle$ is even or odd.

Proposition 11.3.2. *If* (α, β) *is a half-period, then* $\theta[\alpha, \beta]$ *is even (resp. odd) if* $4\langle\alpha, \beta\rangle$ *is even (resp. odd).*

Proof: The summands of (11.3.9) depend on z by the factor $\exp(2\pi i\langle\alpha,\beta\rangle)$. □

Fix a point $p_0 \in \Gamma_g$ and define $A(p)$ for $p \in \Gamma_g$ by

$$A_j(p) \;=\; \int_{p_0}^{p} \omega_j, \quad j = 1, \ldots, g. \tag{11.3.11}$$

We take the path of integration to be the same for each index $j = 1, 2, \ldots, g$. The *Abel map* $A : \Gamma_g \to J(C_g)$ is the map

$$A(p) \;=\; \big(A_1(p), A_2(p), \ldots, A_g(p)\big)\,. \tag{11.3.12}$$

This is independent of the chosen path (so long as the path is independent of the index j). In fact, any two paths differ by a path that is homologous to some path

$$c = \sum_{j=1}^{g}[n_j a_j + m_j b_j],$$

and

$$\int_c \omega_k \;=\; 2n_k\pi i + \sum_{j=1}^{g} B_{jk} m_j,$$

which is the k-th component of a point of the lattice Λ.

We are now in a position to determine the meromorphic functions on Γ_g.

Theorem 11.3.3. (Abel) *The distinct points $p_1, p_2, \ldots p_n$ and $q_1.q_2, \ldots q_n$ in Γ_g are, respectively, the (simple) zeros and poles of a meromorphic function on Γ_g if and only if*

$$\sum_{j=1}^{n} A(p_j) - \sum_{j=1}^{n} A(q_j) \quad \textit{belongs to } \Lambda. \tag{11.3.13}$$

Proof: Suppose that f is meromorphic on Γ_g with the prescribed zeros and poles. Then $\Omega = d\log f = df/f$ is a meromorphic differential with simple poles and zeros. It has residue 1 at each zero and residue -1 at each pole. Therefore it has an expansion

$$\Omega \;=\; \sum_{j=1}^{g} m_j \omega_{p_j q_j} + \sum_{n=1}^{g} c_j \omega_j. \tag{11.3.14}$$

We are assuming that f is single-valued on Γ_g, so the integral over any closed cycle is an integer multiple of $2\pi i$:

$$\int_{a_j} \Omega \;=\; 2n_j\pi i, \qquad \int_{b_j} \Omega \;=\; 2m_j\pi i. \tag{11.3.15}$$

Taking into account (11.3.14), (11.3.15), and the normalizations (11.2.17), (11.2.18), it follows that

$$\begin{aligned} 2\pi i n_k &= \int_{\alpha_k} \Omega \;=\; 2\pi i c_k; \\ 2\pi i m_k &= \int_{\beta_k} \Omega \;=\; \sum_{j=1}^{n} \int_{q_j}^{p_j} \omega_k + \sum_{j=1}^{g} n_j B_{jk}. \end{aligned}$$

Therefore

$$\begin{aligned} \sum_{j=1}^{n} \big[A(p_j) - A(q_j)\big] &= -\sum_{j=1}^{n} \int_{q_j}^{p_j} \omega_k \\ &= -2\pi i m_k + \sum_{j=1}^{n} n_j B_{jk}. \end{aligned} \tag{11.3.16}$$

The right-hand side belongs to the lattice Λ, so (11.3.13) is true.

Conversely, suppose that (11.3.13) is true. Then there are integers $\{n_k\}$, $\{m_k\}$ such that (11.3.16) is true. Let $c_k = n_k$ and use these coefficients to define Ω in (11.3.14). Then

$$F(p) \;=\; \exp \int_{p_0}^{p} \Omega$$

is single-valued on Γ_g and has the $\{p_j\}$ and $\{q_j\}$ as its zeros and poles, respectively. □

11.4 Jacobi inversion

Let $S^g(\Gamma_g)$ be the g-th symmetric product of the curve Γ_g. This means that its elements are the unordered g-tuples $(p_1, \ldots, p_g)$ of points of Γ_g. The Abel map extends to a map

$$A : S^g(\Gamma_g) \;\to\; J(\Gamma_g) \;=\; \mathbb{C}^g/\Lambda$$

defined by

$$A(p_1, \ldots, p_g) \;=\; A(p_1) + \cdots + A(p_q). \tag{11.4.1}$$

The problem of inverting the Abel map is known as the *Jacobi inversion problem.* Thus, given

$$z \;=\; (z_1, z_2, \ldots, z_g) \in J(\Gamma_g) \;=\; \mathbb{C}/\Lambda$$

we want to find points p_1, p_2, ...,p_g in Γ_g such that

$$\sum_{j=1}^{g} \int_{p_0}^{p_j} \omega_k \;\equiv\; z_k, \quad k = 1, \ldots, g$$

where the ω_k are the standard a-cycles of Γ_g, and the equivalence relation is (11.3.7): $b \equiv c$ means that $b - c$ belongs to Λ.

Let $\overrightarrow{e} = (e_1, \ldots, e_g)$ be a fixed element of $\mathbb{C}^g$ and set

$$F(p) \;=\; \theta(A(p) - \overrightarrow{e}). \tag{11.4.2}$$

This function is holomorphic on the cut surface $\widetilde{\Gamma}_g$. Changing $\overrightarrow{e}$ slightly, if necessary, we may assume that F is not identically zero. Recall that $\partial\widetilde{\Gamma}_g$ is the union of cycles a_k, b_k, a_k^{-1}, b_k^{-1}. Changing the cycles slightly, if necessary, we may assume that F has no zeros on the boundary $\partial\widetilde{\Gamma}_g$.

Lemma 11.4.1. *F defined by (11.4.2) has g zeros, counting multiplicity, on $\widetilde{\Gamma}_g$.*

Proof: The number of zeros is

$$\frac{1}{2\pi i}\int_{\partial\widetilde{\Gamma}_g} \frac{F'}{F} \;=\; \frac{1}{2\pi i}\int_{\partial\widetilde{\Gamma}_g} d\log F. \tag{11.4.3}$$

Let F^+ denote F on the union of the a_k and b_k, and let F^- denote F on the union of the inverses a_k^{-1}, b_k^{-1}, and similarly for $A^{\pm}$. Then (11.4.3) is

$$\frac{1}{2\pi i}\sum_{k=1}^{g}\left(\int_{a_k} + \int_{b_k}\right)\left[d\log F^+ - d\log F^-\right]. \tag{11.4.4}$$

It follows from (11.2.8) and (11.2.9) that if p is a point of a_k, then

$$A_j^-(p) \;=\; A_j^+(p) + B_{jk} \tag{11.4.5}$$

and if p is a point of b_k then

$$A_j^+(p) \;=\; A_j^-(p) + 2\pi i\delta_{jk}. \tag{11.4.6}$$

From these equations and (11.3.3), (11.3.4), it follows that

$$\log F^-(p) = -\tfrac{1}{2}B_{kk} - A_k(p) + e_k + \log F_j^+(p), \qquad p \in a_k; \tag{11.4.7}$$

$$\log F^-(p) = \log F^+(p), \qquad p \in b_k. \tag{11.4.8}$$

But $dA_k(p) \;=\; \omega_k$, so

$$d\log F^-(p) = d\log F^+(p) - \omega_k, \qquad p \in a_k; \tag{11.4.9}$$
$$d\log F^-(p) = d\log F^+(p), \qquad p \in b_k. \tag{11.4.10}$$

Therefore (11.4.4) is

$$\frac{1}{2\pi i}\int_{\partial\widetilde{\Gamma}_g} d\log F \;=\; \frac{1}{2\pi i}\sum_{k=1}^{g}\int_{a_k}\omega_k \;=\; g.$$

□

Theorem 11.4.2. *Suppose that the zeros of F on Γ_g are $p_1, \ldots, p_g$. Then*

$$A(p_1, \ldots, p_g) \;\equiv\; \overrightarrow{e} - \overrightarrow{K}, \tag{11.4.11}$$

where

$$K_j \;=\; \frac{2\pi i + B_{jj}}{2} - \frac{1}{2\pi i}\sum_{k\neq j}\int_{a_k}\left(\omega_k(p)\int_{p_o}^{p}\omega_j\right). \tag{11.4.12}$$

Proof: Let $\overrightarrow{\zeta}$ be defined by

$$\zeta_j \;=\; A_j(p_1) + \cdots + A_j(p_g). \tag{11.4.13}$$

The sum on the right can be viewed as the sum of residues

$$\zeta_j \;=\; \frac{1}{2\pi i}\int_{\partial\widetilde{\Gamma}_g} A_j(p)\frac{F'(p)}{F(p)} \;=\; \frac{1}{2\pi i}\int_{\partial\widetilde{\Gamma}_g} d\log F(p). \tag{11.4.14}$$

In view of the calculations in the proof of Lemma 11.4.1, this integral is

$$\begin{aligned}
\zeta_j \;&=\; \frac{1}{2\pi i}\sum_{k=1}^{g}\left(\int_{a_k}+\int_{b_k}\right)[A_j^+ d\log F^+ - A_j^- d\log F^-]\\
&=\; \frac{1}{2\pi i}\sum_{k=1}^{g}\int_{a_k}[A_j^+ d\log F^+ - (A_j^+ + B_{jk})(d\log F^+ - \omega_k)]\\
&\quad+\frac{1}{2\pi i}\sum_{k=1}^{g}\int_{b_k}[A_j^+ d\log F^+ - (A_j^+ - 2\pi i\delta_{jk})\, d\log F^+]\\
&=\; \frac{1}{2\pi i}\sum_{k=1}^{g}\left(\int_{a_k}A_j^+\omega_k - B_{jk}\int_{a_k}d\log F^+ + 2\pi i B_{jk}\right) + \int_{b_k}d\log F^+.
\end{aligned}$$

Because of the way that the a_k are chosen, F takes the same value at the ends, so

$$\int_{a_k} d\log F^+ \;=\; 2\pi i n_k, \quad n_k \in \mathbb{Z}. \tag{11.4.15}$$

Let q_j and $\widetilde{q}_j$ be the initial and final points of b_j. Then

$$\begin{aligned}\int_{b_j} d\log F^+ &= \log F^+(\widetilde{q}_j) - \log F^+(q_j) + 2\pi i m_j, \quad m_j \in \mathbb{Z}\\ &= \log\theta(A(q_j) + f_j - \overrightarrow{e}) - \log\theta(A(q_j) - \overrightarrow{e}) + 2\pi i m_j\\ &= -\tfrac{1}{2}B_{jj} + e_j - A_j(q_j) + 2\pi i m_j.\end{aligned} \tag{11.4.16}$$

As before, $f_j \;=\; (B_{j1}, \ldots, B_{jg})$ belongs to the lattice Λ. Therefore

$$\begin{aligned}\zeta_j &\equiv e_j - \tfrac{1}{2}B_{jj} - A_j(q_j) + \frac{1}{2\pi i}\sum_{k=1}^{g}\int_{a_k} A_j\omega_k\\ &\quad +2\pi i m_j + \sum_k B_{jk}(-n_k + 1)\\ &\equiv e_j - \tfrac{1}{2}B_{jj} - A_j(q_j) + \frac{1}{2\pi i}\sum_{k=1}^{g}\int_{a_k} A_j\omega_k.\end{aligned} \tag{11.4.17}$$

Now q_j, the beginning of b_j, is the end of a_j, so

$$\int_{a_j} A_j\omega_j \;=\; \int_{a_j} d[\tfrac{1}{2}A_j^2 \;=\; \tfrac{1}{2}[A_j^2(q_j) - A_j^2(r_j)],$$

where r_j is the beginning of a_j, and $A_j(q_j) - A_j(r_j) \;=\; 2\pi i$. Therefore

$$\int_{a_j} A_j\omega_j \;=\; \tfrac{1}{2}2\pi i[2A_j(q_j) - 2\pi i],$$

and

$$-A_j(q_j) + \frac{1}{2\pi i}\sum_{k=1}^{g}\int_{a_k} A_j\omega_k \;=\; -\pi i + \frac{1}{2\pi i}\sum_{k\neq j}\int_{\alpha_j} A_j\omega_k. \tag{11.4.18}$$

Combining (11.4.15) – (11.4.17), we obtain (11.4.11). □

The constants K_j of (11.4.12) are known as the *Riemann constants* associated to the given cycles and the choice of p_0.

For the following result, we refer to Farkas and Kra [67].

Theorem 11.4.3. *The function $\theta(A(p) - \overrightarrow{e})$ is identically zero on Γ_g if and only if $\overrightarrow{e}$ can be written as*

$$\overrightarrow{e} \;=\; A(q_1) + \ldots A(q_g) + \overrightarrow{K}, \tag{11.4.19}$$

where the points q_j are the unique poles (counting multiplicity) of a meromorphic function on Γ_g.

Corollary 11.4.4. *If $\zeta \in \mathbb{C}^g$ has the property that $F(p) = \theta(A(p) - \overrightarrow{\zeta} - \overrightarrow{K})$ does not vanish identically on γ, then F has g zeros p_j on Γ_g that are the solution of the Jacobi inversion problem*

$$A(p_1) + \cdots + A(p_g) \;\equiv\; \overrightarrow{\zeta}\,.$$

Moreover, the p_j are uniquely determined by these equations.

Proof: The first statement is just Theorem 11.4.2. Suppose that $\{q_1, \ldots q_g\}$ is disjoint from $\{p_1, \ldots, p_g\}$ and

$$A(q_1) + \cdots + A(q_n) \;\equiv\; \zeta.$$

Then by Theorem 11.3.3, there is a meromorphic function with poles precisely at the p_j and zeros at the q_j. This contradicts Theorem 11.4.3. □

Corollary 11.4.5. *The zeros $\overrightarrow{e}$ of θ can be parametrized by S^{g-1}:*

$$\overrightarrow{e} \;=\; A(p_1) + \cdots + A(p_{g-1}) + \overrightarrow{K}, \tag{11.4.20}$$

where $p_1, \ldots, p_{g-1}$ (counting multiplicity) are any points of Γ_g.

Proof: If $\theta(\overrightarrow{e}) = 0$, let $F(p) \;=\; \theta(A(p) - \overrightarrow{e})$, and suppose first that $F(p)$ is not identically 0. Let p_0 be the lower limit of the integration that defines A, so $A(p_0) = 0$. It follows from the definition (11.3.1) that θ is an even function, so $F(p_0) = \theta(-\overrightarrow{e}) = 0$. Lemma 11.4.1 and Theorem 11.4.2 imply that there are unique points $p_1, \ldots, p_g$ such that

$$\overrightarrow{e} \;=\; A(p_1) + \cdots + A(p_g) + \overrightarrow{K}. \tag{11.4.21}$$

We know that one of these points, say p_g is p_0, so that (11.4.21) reduces to (11.4.20).

If $F(p)$ is identically zero, then by Theorem 11.4.3,

$$\overrightarrow{e} \;=\; A(q_1) + \cdots + A(q_g) + \overrightarrow{K}, \tag{11.4.22}$$

where the q_j are the unique poles of a function f, meromorphic on Γ_g, We may choose the integration limit p_0 to be one of the zeros of f. Let $p_1, \ldots, p_{g-1}$ be the remaining zeros. Again $A(p_0) = 0$. By Theorem 11.3.3,

$$\overrightarrow{e} \;=\; A(q_1) + \cdots + A(q_g) + K \;=\; A(p_1) + \cdots + A(p_{g-1}) + \overrightarrow{K},$$

so again we obtain (11.4.20). □

The approach to the Jacobi inversion problem that we have just described is due to Riemann. A second approach is due to Weierstrass. We illustrate the Weierstrass approach for genus $g = 2$, with defining equation

$$w^2 \;=\; P_5(z).$$

We work with the original differentials

$$\eta_1 \;=\; \frac{dz}{w}, \qquad \eta_2 \;=\; \frac{z\,dz}{w}.$$

The corresponding modified Abel map is

$$A(z) \;=\; \left(\int_{z_0}^{z} \eta_1, \int_{z_0}^{z} \eta_2\right)$$

so

$$A(z_1, z_2) \;=\; A(z_1) + A(z_2) \;=\; (\zeta_1, \zeta_2)$$

with

$$\zeta_1 = \int_{z_0}^{z_1} \eta_1 + \int_{z_0}^{z_2} \eta_1; \tag{11.4.23}$$

$$\zeta_2 = \int_{z_0}^{z_1} \eta_2 + \int_{z_0}^{z_2} \eta_2. \tag{11.4.24}$$

Thus,

$$\frac{d\zeta_1}{dz}(z_j) \;=\; \frac{1}{w(z_j)}, \qquad \frac{d\zeta_2}{dz}(z_j) \;=\; \frac{z_j}{w(z_j)}. \tag{11.4.25}$$

The idea is to relate two systems of differential equations for $(z_1, z_2) \in \Gamma_2 \times \Gamma_2$ to the corresponding system of equations for the image (ζ_1, ζ_2) under the Abel map (11.4.23), (11.4.24). The systems for (z_1, z_2) are

$$\frac{dz_1}{ds} = \frac{w(z_1)}{z_1 - z_2}, \qquad \frac{dz_2}{ds} \;=\; \frac{w(z_2)}{z_2 - z_1}, \tag{11.4.26}$$

$$\frac{dz_1}{dt} = \frac{z_2 w(z_1)}{z_1 - z_2}, \qquad \frac{dz_2}{dt} \;=\; \frac{z_1 w(z_2)}{z_2 - z_1},. \tag{11.4.27}$$

Proposition 11.4.6. *Under the Abel map $A : S^2\Gamma_2 \to J(\Gamma_2)$, the systems (11.4.26), (11.4.27) become*

$$\frac{d\zeta_1}{ds} = 0, \qquad \frac{d\zeta_2}{ds} = 1; \tag{11.4.28}$$

$$\frac{d\zeta_1}{dt} = -1, \qquad \frac{d\zeta_2}{dt} = 0; . \tag{11.4.29}$$

Proof: The equations (11.4.28) and (11.4.29) follow readily from the equations (11.4.23)–(11.4.27); see Exercise 9.

We cannot resist closing this section with a remark of Weyl, [214], footnote, p. 144:

> The principal significance of the inversion problem to us today lies primarily, not in its intrinsic value, but in the splendid development created by Riemann and Weierstrass in their efforts to solve the problem.

Exercises

1. (a), (b), (c), (d): section by section, work out the results in the case of a torus: $g = 1$.
2. Show that $j \leq g$ is also the necessary and sufficient condition that η_j be holomorphic at ∞ in the case when P has degree $2g + 1$.
3. Prove the existence of the differentials of the second kind. Hint: look for

$$\omega_p^{(n)} = \frac{g(w)\,dz}{(z-p)^{n+1}}.$$

   What conditions are needed on $g(w)$ to guarantee that $\omega_p^{(n)}$ has the correct behavior at the point p on each sheet $\mathbb{C}_\pm$?
4. Prove the existence of the differentials of the third kind ω_{pq}. Hint: look at Exercise 3.)
5. Prove (11.2.19).
6. Prove that (11.3.1) converges uniformly on compact sets in $\mathbb{C}$.
7. Prove the transformation law (11.3.10).
8. Show that θ has $2^{g-1}(2^{g+1})$ even periods and $2^{g-1}(2^g - 1)$ odd periods.
9. Prove (11.4.28) and (11.4.29).
10. Use the Weierstrass system of differential equations to show that map to $J(G_2)$ is surjective and can be inverted by following trajectories of two systems. (Start from p_0.)

Remarks and further reading

The theory of theta functions was initiated by Jacobi and advanced by Riemann and Weierstrass. It is still a large and active area of research, with connections to algebraic geometry, analytic number theory, representation theory, algebraic topology, nonlinear partial differential equations, and quantum physics.

The classical theory of theta functions is treated exhaustively by Baker in [17]. Baker's monograph has been reissued, with a foreword by Krichever that outlines the theory and delves into its application to the study of "completely integrable" nonlinear partial differential equations, such as the Korteweg–deVries (KdV) equation for waves in a channel. Our presentation is based on Dubrovin's exposition [60], which goes on to treat these applications in detail. These applications are also among the (very many) topics treated in Mumford's lectures [145], [146], [147].

Various versions of theta functions are associated with the names of Ramanujan [48] and Siegel and Ruelle [35]. There are connections with modular forms [48], [68], quantum field theory [208], moduli spaces [113], eta functions [124], all of the above [148], and knot theory [89].

Chapter 12
Padé approximants and continued fractions

The Taylor series of a function that is holomorphic in a neighborhood of a given point provides an approximation of the function by polynomials: the successive partial sums of the Taylor series. For both theoretical and practical reasons, it can be useful to approximate instead by general rational functions, i.e. quotients of polynomials. Systematic use of this idea goes back at least to Frobenius [80] in 1881. Padé [162] treated some exceptional cases in his thesis a decade later. Both theory and practice – and the theory behind the practice – have developed greatly since then. In this chapter we touch on the main theoretical questions and practical methods, and exhibit some interesting examples.

In Section 12.1 we introduce the general terminology, notation, and concepts, as well as the basic existence theorems. Sections 12.2 and 12.3 give three connections of Padé theory to continued fractions.

The connection of Padé approximants to the Stieltjes transform and to orthogonal polynomials is introduced in Section 12.4. Stieltjes transforms and related functions are characterized in Section 12.5. The Padé approximants of these transforms are examined in Section 12.6.

Two continued fraction expansions of the exponential function are examined in Section 12.9. Section 12.8 contains several examples illustrating the theory, with some specific numerical results.

Section 12.7 describes the basic theory behind practical methods of computation: Shanks' method and generalizations.

R. Beals and R. S. C. Wong, *More Explorations in Complex Functions*, Graduate Texts in Mathematics 298, https://doi.org/10.1007/978-3-031-28288-1_12

12.1 Padé approximants and Taylor series

Suppose that f is holomorphic in a neighborhood of 0:

$$f(z) \;=\; \sum_{k=0}^{\infty} a_k z^k. \tag{12.1.1}$$

Summing the series exactly is often not practical, but one can resort to the partial sums as approximations to the value $f(z)$. The partial sums are the Taylor polynomials

$$T_m(z) \;=\; \sum_{k=0}^{m} a_k z^k,$$

which are uniquely determined by the property that $f(z) = T_N(z) + O(|z|^{N+1})$ as $z \to 0$. One can obtain this same approximation property with rational functions. For a pair of integers $m \geq 0$, $n \geq 0$, the $[m, n]$ *Padé approximant to* f at 0 is the quotient

$$\frac{P_m(z)}{Q_n(z)} \;=\; \frac{\sum_{k=0}^{m} p_k z^k}{\sum_{k=0}^{n} q_k z^k}$$

with the property that

$$f(z) - \frac{P_m(z)}{Q_m(z)} \;=\; O(|z|^{m+n+1}) \quad \text{as} \;\; z \to 0. \tag{12.1.2}$$

For a given m and n, (12.1.2) is equivalent to

$$T_{m+n}(z) - \frac{P_m(z)}{Q_n(z)} \;=\; O(|z|^{m+n+1}) \quad \text{as} \;\; z \to 0. \tag{12.1.3}$$

This is unique if we normalize by taking $Q(0) = q_0 = 1$.

The quotient P_m/Q_n is sometimes denoted $[m, n]_f$. The *Padé table* of f is the infinite matrix $([m, n]_f)_{m,n=0}^{\infty}$. The leftmost column $\{[m, 0]_f\}$ consists of the Taylor polynomials T_m. If $a_0 \neq 0$, then the first row $\{[0, n]_f\}$ consists of the Taylor polynomials for $1/f$. In applications one often uses the main diagonal, i.e. the approximants $[n, n]_f$, or some nearby diagonal, such as $[n, n \pm 1]_f$.

As we shall see, there are a number of reasons for going beyond the Taylor polynomials for the purpose of approximation. The Taylor polynomials can only converge uniformly on disks $D_r(0)$ with r less than the radius of convergence of the series (12.1.1). If f is holomorphic in a larger region, rational approximations may converge in larger subsets of that region; see Exercises 1 and 2. Even in disks where the Taylor series converges, some Padé approximants may converge more rapidly than the Taylor polynomials. Furthermore, Padé approximation is, in a number of ways, more flexible and adaptable to special circumstances, such as dealing with asymptotic series or simultaneous convergence near two or more points.

The approximation property (12.1.2) or (12.1.3) can be put in a linear form:

$$T_{n+m}(z)Q_m(z) - P_m(z) \;=\; O(|z|^{n+m+1}) \qquad \text{as } z \to 0. \tag{12.1.4}$$

This is a system of $m+n+1$ linear equations for the $m+1$ coefficients of P_m and the n coefficients q_k, $k > 0$ of Q_n. In fact

$$T_{n+m}Q_n(z) - P_m(z) \;=\; \sum_{k=0}^{\infty}\left[\sum_{j=0}^{k} a_{k-j}q_j - p_k\right] z^k, \tag{12.1.5}$$

where $a_k = 0, k > m+n, q_j = 0, j > n, p_k = 0, k > m$. We set the terms in brackets in (12.1.5) equal to zero for $k \le m+n$. As we show next, this system always has a solution. However, the solution may not be unique; see Exercise 1.

Theorem 12.1.1. *The problem (12.1.2) has a solution with P_m a polynomial of degree $\le m$ and Q_n a polynomial of degree $\le n$.*

Proof. The idea is to use the extended Euclidean algorithm, starting with the polynomials $A(z) = z^{n+m+1}$, $B(z) = T_{n+m}(z)$. The algorithm constructs polynomials R_k, S_k:

$$R_0 \;=\; A, \quad R_1 \;=\; B, \quad R_{k-1} \;=\; S_kR_k + R_{k+1}, \quad \deg R_{k+1} < \deg R_k. \tag{12.1.6}$$

The third equation in (12.1.6) determines polynomials S_k and R_{k+1} up to a constant multiple. The extension of the algorithm uses the S_k to compute polynomials U_k, V_k:

$$\begin{aligned} U_0 = 1, \quad U_1 = 0, \quad U_{k+1} \;&=\; U_{k-1} - S_kU_k; \\ V_0 = 0, \quad V_1 = 1, \quad V_{k+1} \;&=\; V_{k-1} - S_kV_k. \end{aligned} \tag{12.1.7}$$

It follows by induction that we have the Bezout identities

$$AU_k + BV_k \;=\; R_k. \tag{12.1.8}$$

It follows from (12.1.6) that

$$\deg S_k \;=\; \deg R_{k-1} - \deg R_k$$

and from (12.1.7) that

$$\deg V_{k+1} \;=\; \deg S_k + \deg S_{k-1} + \cdots + \deg S_1 \;=\; \deg A - \deg R_k. \tag{12.1.9}$$

Let us stop the process as soon as $\deg V_{k+1} > n$. Then $\deg V_k \le n$ and (12.1.9) implies that $\deg R_k \le m$. The Bezout identity (12.1.8) gives us

$$z^{m+n+1}U_k(z) + T_{m+n}(z)V_k(z) \;=\; R_k(z).$$

Therefore taking $P_m = R_k$, $Q_n = V_k$ gives us our approximant. □

Remarks. 1. In the previous discussion we made no use of the *convergence* of the series (12.1.1), but only that it is an asymptotic series for f:

$$f(x) - \sum_{k=0}^{n} a_k z^k \;=\; O(|z|^{n+1}). \tag{12.1.10}$$

2. We might also assume that (12.1.10) is valid only in some subset of a neighborhood of 0, e.g. in the sector $\{z : \arg z < \alpha\}$.

3. A typical situation in which one might have an asymptotic expansion valid in some sector concerns behavior as $z \to \infty$:

$$g(z) \sim \; b_0 + \frac{b_1}{z} + \frac{b_2}{z^2} + \ldots \quad \text{as} \;\; z \to \infty \tag{12.1.11}$$

in some sector.

4. If g satisfies (12.1.11) with $b_0 \neq 0$, then $f(z) = g(z^{-1})$ has an expansion (12.1.10). More generally, if g has an expansion with $b_0, \ldots b_{k-1} = 0$ and $b_k \neq 0$, then $f(z) = z^{-k} g(z^{-1})$ has an expansion (12.1.10).

Theorem 12.1.2. *Suppose that g has an asymptotic expansion (12.1.10) in some sector. Then for any non-negative integers m and n there is a $[m, n]$-Padé approximant P_m/Q_n such that*

$$g(z) - \frac{P_m(z)}{Q_n(z)} \;=\; O(z^{-m-n-1}) \quad \text{as} \;\; z \to \infty \;\; \textit{in the sector.}$$

Proof. We may assume that not every term in the expansion (12.1.11) is zero. Then for some k, $f(z) = z^{-k} g(1/z)$ has an expansion (12.1.10) in a sector at 0. By Theorem 12.1.1 and the preceding remarks, f has a Padé approximant $[n-k, m]_f = R(z)$, where R is a rational function with at most $n-k$ zeros and at most m poles. Then

$$g(z) \;=\; z^{-k} f(z^{-1}) \;=\; z^{-k} R(z^{-1}) + O(z^{-k-(n-k+m)-1}),$$

and $z^{-k} R(z^{-1})$ is a rational function of z with at most $n-k$ zeros and at most m poles. Therefore $z^{-k} R(z^{-1})$ is the quotient of polynomials P and Q of degrees $\leq m$ and n, respectively. □

The method as described so far could be called "one-point" Padé approximation. It relies on information about behavior of a function f at a single point, which we have taken to be the origin or the point at infinity. This can easily be generalized.

Suppose that $f(z)$ has asymptotic expansion at distinct points z_0, z_1:

$$f(z) \sim \sum_{n=0}^{\infty} a_n (z - z_0)^n, \qquad \text{as} \;\; z \to z_0, \tag{12.1.12}$$

$$f(z) \sim \sum_{n=0}^{\infty} b_n (z - z_1)^n, \qquad \text{as} \;\; z \to z_1. \tag{12.1.13}$$

A two-point Padé approximation to $f(z)$ is a rational function $R(z) = P_m(z)/Q_n(z)$, where $Q_m(z_0) = 1$, while $P_m(z)$ and $Q_n(z)$ are polynomials of degrees m and n, respectively. The $n + m + 1$ coefficients (after the normalization $Q_n(z_0) = 1$) are chosen so that

$$f(z) = \frac{P_m(z)}{Q_n(z)} + O((z - z_0)^k) \qquad \text{as } z \to z_0;$$
$$f(z) = \frac{P_m(z)}{Q_n(z)} + O((z - z_0)^l) \qquad \text{as } z \to z_1,$$

where $k + l = m + n + 1$. We leave the formulation of the general equations for the coefficients of the polynomials $P_n(z)$ and $Q_m(z)$, as well as the development of efficient numerical techniques, to Exercises 5 and 6.

12.2 Padé approximation and continued fractions

Often Padé approximations are made for functions whose asymptotic behavior at ∞ is known, say

$$F(z) \sim \frac{c_0}{z} + \frac{c_1}{z^2} + \cdots + \frac{c_n}{z^{n+1}} + \ldots \tag{12.2.1}$$

as $z \to \infty$ in some sector.

Suppose $c_0 \neq 0$. Then we may write (12.2.1) as

$$F(z) \sim \frac{c_0}{z}\left[1 + \frac{c_1}{c_0 z} + \frac{c_2}{c_0 z^z} + \ldots\right].$$

A key observation is that this implies that

$$\frac{1}{F(z)} \sim \frac{z}{c_0} - \frac{b_1}{c_0} + \frac{F_1(z)}{c_0}, \qquad b_1 = \frac{c_1}{c_0},$$

where F_1 has an expansion similar to the expansion (12.2.1) of F. Thus

$$F(z) \sim \frac{a_1}{z - b_1 + F_1(z)}, \qquad a_1 = c_0. \tag{12.2.2}$$

If the leading coefficient in the expansion of F_1 is not zero, this can be continued:

$$F(z) \sim \cfrac{a_1}{z - b_1 + \cfrac{a_2}{z - b_2 + F_2(z)}},$$

where F_2 has an expansion of the form (12.2.1). So long as leading coefficients do not vanish we get a continued fraction expansion

$$F(z) \sim \cfrac{a_1}{z - b_1 + \cfrac{a_2}{z - b_2 + \cfrac{a_3}{z - b_3 + \dots}}}. \tag{12.2.3}$$

Suppose we truncate this, so that the last denominator is $z - b_n$. The resulting expression R_n is called the n-th *convergent* of the continued fraction (12.2.3). It is easily seen that R_n is a rational function that approximates F to degree z^{-n-1}. As we shall see, $R_n = P_n/Q_n$, where P_n and Q_n are polynomials and deg $P_n = n - 1$, deg $Q_n = n$. We shall show that R_n is a Padé approximant to f at ∞.

In discussing the theory further, it will be convenient to ease notation by looking first at the numerical case of continued fractions:

$$\cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{b_3 + \cfrac{a_4}{b_4 + \dots}}}}. \tag{12.2.4}$$

This is sometimes written

$$\frac{a_1}{b_1+}\,\frac{a_2}{b_2+}\,\frac{a_3}{b_3+}\,\frac{a_4}{b_4+}\,\dots\,. \tag{12.2.5}$$

We assume throughout this discussion that the coefficients $\{a_n\}$, $\{b_n\}$ are each nonzero.

Let us look at the successive *convergents*: the successive truncations t_n that end at a denominator b_n:

$$t_1 = \frac{a_1}{b_1}, \qquad t_2 = \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2}} = \frac{b_2 a_1}{b_2 b_1 + a_2},$$

$$t_3 = \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{b_3}}} = \frac{b_3 b_2 a_1 + a_3 a_1}{b_3 b_2 b_1 + b_3 a_2 + b_1 a_3}.$$

Let us write these as

$$\frac{p_1}{q_1} = \frac{a_1}{b_1}, \quad \frac{p_2}{q_2} = \frac{b_2 a_1}{b_2 b_1 + a_2}, \quad \frac{p_3}{q_3} = \frac{b_3 b_2 a_1 + a_3 a_1}{b_3(b_2 b_1 + a_2) + a_3 b_1}.$$

Identifying numerators with numerators and denominators with denominators in each equation here, we see that

$$p_2 = b_2 p_1, \qquad q_2 = b_2 q_1 + a_2;$$
$$p_3 = b_3 p_2 + a_3 p_1, \; q_3 = b_3 q_2 + a_3 q_1.$$

If we set $p_{-1} = 1$, $p_0 = 0$, $q_{-1} = 0$, and $q_0 = 1$, then these equations take the form of three-term recursions:

$$p_k = b_k p_{k-1} + a_k p_{k-2}, \qquad q_k = b_k q_{k-1} + a_k q_{k-2}, \tag{12.2.6}$$

$k = 1, 2, 3$. Replacing b_k in these equations by $b_k + a_{k+1}/b_{k+1}$ in the quotient p_k/q_k leads to the corresponding equations for p_{k+1} and q_{k+1}; see Exercise 7

Proposition 12.2.1. *The equations (12.2.6), extended for all values of k, produce the convergents $r_k = p_k/q_k$ of the continued fraction (12.2.4).*

The equations (12.2.6) have powerful consequences.

Lemma 12.2.2. *The convergents of the continued fraction (12.2.4) satisfy the relation*

$$\frac{p_n}{q_n} - \frac{p_{n-1}}{q_{n-1}} = \frac{(-1)^{n-1} a_1 a_2 \cdots a_n}{q_n q_{n-1}}. \tag{12.2.7}$$

Proof. Multiplying the left side of (12.2.7) by $q_n q_{n-1}$, and using the equations (12.2.6), we obtain

$$\begin{aligned} p_n q_{n-1} - p_{n-1} q_n &= (b_n p_{n-1} + a_n p_{n-2}) q_{n-1} - p_{n-1}(b_n q_{n-1} + a_n q_{n-2}) \\ &= -a_n [p_{n-1} q_{n-2} - p_{n-2} q_{n-1}]. \end{aligned}$$

The term in square brackets is the numerator of $p_{n-1}/q_{n-1} - p_{n-2}/q_{n-2}$. Therefore we may continue the calculation back to the numerator $p_0 q_1 - p_1 q_0 = 1$. Dividing by $q_n q_{n-1}$ gives (12.2.7). □

Now

$$\frac{p_n}{q_n} = \left[\frac{p_n}{q_n} - \frac{p_{n-1}}{q_{n-1}}\right] + \cdots + \left[\frac{p_1}{q_1} - \frac{p_0}{q_0}\right] + \frac{p_0}{q_0}.$$

Combining this with (12.2.7) gives

Corollary 12.2.3. *The n-th convergent of the continued fraction (12.2.4) can be written as a sum*

$$\frac{p_n}{q_n} = (-1)^{n-1} \frac{a_1 a_2 \cdots a_n}{q_n q_{n-1}} + (-1)^{n-2} \frac{a_1 a_2 \cdots a_{n-1}}{q_{n-1} q_{n-2}} + \cdots + \frac{a_1}{q_1 q_0}. \tag{12.2.8}$$

Returning now to the continued fraction (12.2.3), we will assume that the coefficients $\{a_k\}$ are non-zero.

Proposition 12.2.4. *The $[n, n]$ Padé approximant for the continued fraction (12.2.3) is the rational function R_n obtained by truncating at denominator b_n. It has the form*

$$R_n(z) = \frac{P_n(z)}{Q_n(z)}, \tag{12.2.9}$$

where P_n and Q_n are polynomials of degree $n-1$ and n, respectively. They are defined recursively by

$$\begin{aligned} P_{-1} = 1, \quad P_0 &= 0, \quad P_k(z) = (z - b_k)P_{k-1}(z) + a_k P_{k-2}(z); \quad (12.2.10) \\ Q_{-1} = 0, \quad Q_0 &= 1, \quad Q_k(z) = (z - b_k)Q_{k-1}(z) + a_k Q_{k-2}(z). \end{aligned}$$

Proof. The equations (12.2.10) are an immediate consequence of Proposition 12.2.1. The determination of the degrees of P_n and Q_n follows by induction, using the assumption that the a_n are non-zero. It follows from (12.2.8) that the first omitted term in the expansion of (12.2.3) has degree $(n+1)+n$ so R_n agrees with the formal expansion of (12.2.3) through terms of order z^{-2n}. □

Multiplying numerator and denominator by $\lambda_1 = 1/a_1$ produces the same formal fraction, with a_1, b_1 and a_2 replaced by 1, b_1/a_1, and a_2/a_1. Then multiplying both terms in the fraction with numerator a_2 by $\lambda_2 = a_1/a_2$ changes the numerator to 1 and b_2 to $\lambda_2 b_2$. Continuing this process leads to

$$\cfrac{1}{\hat{b}_1 + \cfrac{1}{\hat{b}_2 + \cfrac{1}{\hat{b}_3 + \dots}}}, \tag{12.2.11}$$

where

$$\hat{b}_n = \lambda_n b_n, \qquad \lambda_{2m-1} = \frac{a_2 a_4 \cdots a_{2m-2}}{a_1 a_3 \cdots a_{2m-1}}, \qquad \lambda_{2m} = \frac{a_1 a_3 \cdots a_{2m-1}}{a_2 a_4 \cdots a_{2m}}.$$

The basic analytic question concerning continued fractions is: do the "convergents" actually converge? One of the most elegant results is the following theorem of Seidel [188].

Theorem 12.2.5. *Suppose $\hat{b}_n > 0$ for $n = 1, 2, \cdots$. The continued fraction (12.2.11) converges if and only if the series $\sum \hat{b}_n$ diverges.*

Proof. It follows from (12.2.8) that the convergents of (12.2.11) are the partial sums of the series

$$\frac{1}{q_1} - \frac{1}{q_1 q_2} + \dots + (-1)^n \frac{1}{q_{n-1} q_n} + \dots \tag{12.2.12}$$

where the q_n are the denominators of the convergents. They are all positive. By definition $q_0 = 1$, and clearly $q_1 = \hat{b}_1$. We claim that for all n,

$$q_n < \prod_{k=1}^{n} (1 + \hat{b}_k).$$

This is clearly true for $n = 0$ and $n = 1$. If it is true for $n-1$ and $n-2$, then by (12.2.6)

$$q_n < \hat{b}_n \prod_{k=1}^{n-1}(1+\hat{b}_k) + \prod_{k=1}^{n-2}(1+\hat{b}_k) < (1+\hat{b}_n)\prod_{k=1}^{n-1}(1+\hat{b}_k).$$

Suppose first that $\sum_{n=1}^{\infty} \hat{b}_k < \infty$. Then the infinite product $\prod_{n=1}^{\infty}(1+\hat{b}_n)$ has a finite limit. (See Section 1.6, or use the inequality $1+\hat{b}_n < e^{\hat{b}_n}$). This implies that the terms in the series (12.2.12) are bounded away from zero, so the convergents fail to converge.

Suppose instead that $\sum_{n=1}^{\infty} \hat{b}_n$ diverges. We claim that for all $n \geq 1$,

$$q_n \;\geq\; \rho q_{n-1}, \qquad \rho = \min\{1, \hat{b}_1\}.$$

By definition $q_{-1} = 0$, so this is true for $n = 1$ and $n = 2$. If it is true for $n-1$ and $n-2$, then

$$q_n \;=\; \hat{b}_n q_{n-1} + q_{n-2} \;\geq\; \rho \hat{b}_n \rho + \rho \;>\; \rho.$$

Now

$$q_n q_{n-1} \;=\; (\hat{b}_n q_{n-1} + q_{n-2}) q_{n-1} \;\geq\; \hat{b}_n \rho^2 + q_{n-1} q_{n-2},$$

so

$$\begin{aligned} q_n q_{n-1} &= (q_n q_{n-1} - q_{n-1} q_{n-2}) + \cdots + (q_2 q_1 - q_1 q_0) + q_1 q_0 \\ &\geq \hat{b}_n \rho^2 + \cdots + \hat{b}_2 \rho^2 + \hat{b}_1 \rho^2. \end{aligned}$$

Therefore $q_n q_{n-1} \to \infty$, so the terms of (12.2.12) decrease to zero, and the series converges. □

As an example, consider

$$\cfrac{1}{2x + \cfrac{1}{2x + \cfrac{1}{2x + \cfrac{1}{2x + \ldots}}}}, \qquad x > 0.$$

By Theorem 12.2.5 the convergents converge to some limit $L = L(x) > 0$. Clearly

$$L \;=\; \frac{1}{2x + L}, \tag{12.2.13}$$

so

$$L^2 + 2xL - 1 \;=\; 0, \qquad L = -x + \sqrt{x+1}. \tag{12.2.14}$$

Conversely, (12.2.14) leads to (12.2.13), which identifies the continued fraction expansion of the function $-x + \sqrt{x+1}$. It is a general fact that if the sequence

$\{\hat{b}_n\}$ is eventually periodic, then the limit of the convergents of (12.2.11) is the solution of a quadratic with coefficients that are rational functions of the $\{\hat{b}_n\}$; see [120].

We conclude this section with another method for determining a partial fractions expansion.

Theorem 12.2.6. *For any $n \geq 1$, we have*

$$\sum_{k=1}^{n} \frac{1}{D_k} = \cfrac{1}{D_1 - \cfrac{D_1^2}{D_1 + D_2 - \cfrac{D_2^2}{D_2 + D_3 - \cdots \cfrac{D_{n-1}^2}{D_{n-1} + D_n}}}}. \tag{12.2.15}$$

Proof. The identity (12.2.15) is true for $n = 1$. For $n = 2$ we have

$$\cfrac{1}{D_1 - \cfrac{D_1^2}{D_1 + D_2}} = \frac{D_1 + D_2}{(D_1 + D_2)D_1 - D_1^2} = \frac{D_1 + D_2}{D_1 D_2} = \frac{1}{D_1} + \frac{1}{D_2}.$$

If (12.2.15) is true for n, then changing the last denominator $D_{n-1} + D_n$ to

$$D_{n-1} + D_n - \frac{D_n^2}{D_n + D_{n+1}}$$

amounts to replacing the last term in the sum $1/D_1 + \cdots + 1/D_n$ by

$$\cfrac{1}{D_n - \cfrac{D_n^2}{D_n + D_{n+1}}} = \frac{D_n + D_{n+1}}{D_n D_{n+1}} = \frac{1}{D_n} + \frac{1}{D_{n+1}}.$$

□

12.3 Another view of Padé approximants and continued fractions

Let us take another look at the Padé approximants of a function f with Taylor expansion

$$f(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \dots . \tag{12.3.1}$$

Assuming that $a_0 \neq 0$, set $f_0 = f$ and define f_1 by

$$\frac{a_0}{f_0(z)} = 1 + z f_1(z).$$

Then f_1 has a Taylor expansion similar to (12.3.1), and

$$f_0(z) \;=\; \frac{c_0}{1+zf_1(z)}; \qquad f_1(z) \;=\; \left[\frac{c_0}{f_0} - 1\right] \cdot \frac{1}{z} = c_1 + O(z),\, c_0 = a_0.$$

If the constant term of f_1, $c_1 = -a_1/a_0$, is not zero, then this process can be continued:

$$f_1(z) \;=\; \frac{c_1}{1+zf_2(z)}.$$

For brevity, we say that f is *normal* if this iterative process

$$f_{n+1} \;=\; \left[\frac{c_n}{f_n} - 1\right] \cdot \frac{1}{z} \tag{12.3.2}$$

leads to $c_n = f_n(0) \neq 0$ for all n, and thus to the continued fraction representation

$$f(z) \sim \cfrac{c_0}{1+\cfrac{c_1 z}{1+\cfrac{c_2 z}{1+\ddots\cfrac{c_{2n} z}{1+\dots}}}} \tag{12.3.3}$$

The coefficients c_k can be determined by expanding the convergents of (12.3.3) into power series and comparing the coefficients with those of the power series to be approximated. This procedure is very similar to the Padé approximation as discussed in Section 12.1. As we shall see, in this case there is a simple algorithm.

Proposition 12.3.1. *Suppose that f is normal. The n-th convergent of the continued fraction (12.3.3) is the $[n, n]$ Padé approximant of f if n is odd, and the $[n-1, n]$ approximant if n is even.*

Proof. As we showed in Section 12.2, the n convergent is P_n/Q_n, where

$$P_{-1} = 0, \quad P_0 \;=\; c_0, \quad P_n(z) = P_{n-1}(z) + c_n z P_{n-2}(z); \tag{12.3.4}$$

$$Q_{-1} = 1, \quad Q_0 \;=\; 1, \quad Q_n(z) = Q_{n-1}(z) + c_n z Q_{n-2}(z). \tag{12.3.5}$$

It follows inductively that for $n \geq 0$, P_{2m+1} and Q_{2m+1} have degrees n and $n+1$, respectively, while P_{2m+2} and Q_{2m+2} have degrees $n+1$ and $n+2$, respectively. Thus $Q_n Q_{n-1}$ has degree n, and the argument in Proposition 12.2.4 shows that the n-th convergent agrees with f to $O(z^{-2n})$. □

For convenience, we introduce the notation $R_n^m(z)$ for the $[m, n]$ Padé approximant to the function (12.3.1).

$$R_n^m(z) = \frac{P_m(z)}{Q_n(z)} = \frac{\sum_{k=0}^m p_k z^k}{\sum_{k=0}^n q_k z^k}. \tag{12.3.6}$$

Consider the Padé approximants $R^m_m(z)$ and $R^m_{m+1}(z)$. The Padé sequence $R^0_0(z)$, $R^0_1(z)$, $R^1_1(z)$, $R^1_2(z)$, $R^2_2(z)$, $P^2_3(z)$, $\dots$, is said to be *normal* if every member of the sequence exists and no two members are identically equal. This sequence is normal if and only if the function f is normal; see Exercise 8. The coefficients $\{c_k\}$ are the same for every term of the Padé sequence. Thus, R^m_{m+1}, $m \geq 1$, is obtained from R^m_m by simply replacing $c_n z$ by $c_n z/(1 + c_{n+1}z)$ where $n = 2m$ and R^{m+1}_{m+1}, $m \geq 0$, is obtained from R^m_{m+1} by replacing $c_n z$ by $c_n z/(1 + c_{n+1}z)$, where $n = 2m + 1$.

Note that when Padé approximants are written as ratios of polynomials, every coefficient in the rational fraction must be recomputed as we go from one member of the normal sequence to the next. However, the entire normal sequence may be rewritten as a simple continued fraction and only *one* new coefficient needs to be computed as we go from one member to the next. Moreover, as we mentioned above, there is a simple algorithm for computing the c_n. Then the iterations (12.3.4), (12.3.5) yield the polynomials P_n, Q_n.

The algorithm for computing the c_n proceeds as follows. Let us write the functions f_n in the construction (12.3.2) as quotients:

$$f_k(z) = \frac{a^{(k)}}{b^{(k)}} = \frac{\sum_{n=0}^{\infty} a^{(k)}_n z^n}{\sum_{n=0}^{\infty} b^{(k)}_n z^n}.$$

In particular we take

$$a^{(0)} = f_0, \qquad b^{(0)} = 1. \tag{12.3.7}$$

Given any such power series $d(z) = \sum_{n=0}^{\infty} d_n z^n$ with leading coefficient $d(0) = d_0$, let us write $\widetilde{d}(z) = d(z)/d(0)$ for the normalized series with leading term 1. Then the algorithm (12.3.2) can be written as

$$\begin{aligned} \frac{a^{(k+1)}}{b^{(k+1)}} &= \left[\frac{\widetilde{b}^{(k)}}{\widetilde{a}^{(k)}} - 1\right] \cdot \frac{1}{z} \\ &= \frac{\widetilde{b}^{(k)} - \widetilde{a}^{(k)}}{\widetilde{a}^{(k)}} \cdot \frac{1}{z}. \end{aligned} \tag{12.3.8}$$

Thus we may define $b^{(k)}$ and $a^{(k)}$ recursively from (12.3.8) by

$$b^{(k+1)}(z) = \widetilde{a}^{(k)}(z), \qquad a^{(k+1)} = \frac{\widetilde{b}^{(k)} - \widetilde{a}^{(k)}}{z}.$$

Then $\widetilde{b}^(k) = b^{(k)}$ for all k. Since c_k is the constant term of $f_k = a^{(k)}/b^{(k)}$, we have

$$c_k = a^{(k)}(0). \tag{12.3.9}$$

In fact we may eliminate the $b^{(k)}$ entirely by setting

$$a^{(-1)} = 1, \quad a^{(0)} = f, \quad a^{(k+1)}(z) = \frac{\widetilde{a}^{(k-1)}(z) - \widetilde{a}^{(k)}(z)}{z}, \quad k \geq 0, \tag{12.3.10}$$

where again $\widetilde{a}^{(k)}(z) = a^{(k)}(z)/a^{(k)}(0)$.

This method can be extended to derive a continued fraction expansion for other Padé approximants $R_m^n(z)$ for f. Given $J > 0$, the Padé approximants R_k^{J+k}, R_{k+1}^{J+k} of the function f of (12.3.1) can be represented as convergents of

$$\sum_{k=0}^{J-1} a_k z^k + \cfrac{c_0 z^J}{1+\cfrac{c_1 z}{1+\cfrac{c_2 z}{1+\ddots\cfrac{c_{2n} z}{1+\dots}}}} \tag{12.3.11}$$

provided that the Padé sequence R_0^J, R_1^J, R_1^{J+1}, R_2^{J+1}, ... is normal, i.e. no two members of the sequence are identically equal, so that all the coefficients c_n are non-zero. Any convergent of the form (12.3.11) is the ratio of a polynomial of degree $J+m$ to a polynomial of degree $J+m$ or $J+m+1$. The coefficients of z^p in the expansion of the convergent whose last denominator is $c_n z$ are a_p for $p = 0, 1, \cdots, J-1$. The coefficients c_p involve only the coefficients of z^k for $k = J, J+1, \cdots, J+p$. Thus, the coefficients c_p can be determined by a slight modification of the formulas in (12.3.9).

To represent the members of the Padé sequence R_J^0, R_J^1, R_{J+1}^1, R_{J+1}^2, R_{J+2}^2, $\cdots$ with $J \geq 0$ as continued fractions, we only need to observe that the R_m^n Padé approximant to $1/f(z)$ is identical to the R_n^m Padé approximant to f, evaluated at $1/z$; see Exercise 7. Therefore, assuming normality, the desired sequence of Padé approximants can be represented as the inverse of the expressions (12.3.11) with the coefficients a_k of $f(z)$ replaced by the expansion coefficients of $1/f(z)$.

Let us consider the question of convergence of the convergents of the continued fraction (12.3.3) in the special case when the c_n are equal:

$$f(z) \sim \cfrac{c}{1+\cfrac{cz}{1+\cfrac{cz}{1+\ddots\cfrac{cz}{1+\dots}}}} \tag{12.3.12}$$

If there is convergence, then

$$f(z) = \frac{c}{1+zf(z)}. \tag{12.3.13}$$

This gives a quadratic equation in $f(z)$, and considering $z \to 0$ identifies the solution as

$$f(z) = \frac{\sqrt{1+4cz}-1}{2z}. \tag{12.3.14}$$

We know that the convergents have the form P_n/Q_n, where

$$\begin{aligned} P_{-1} &= 0, \quad P_0 \;=\; c, \quad P_n(z) = P_{n-1}(z) + czP_{n-2}(z);\\ Q_{-1} &= 1, \quad Q_0 \;=\; 1, \quad Q_n(z) = Q_{n-1}(z) + czQ_{n-2}(z). \end{aligned}$$

Writing S for the shift operator on sequences $\mathbf{x} = \{x_n\}$, $S\mathbf{x} = \{x_{n+1}\}$, the sequences $\{P_n\}$ and $\{Q_n\}$ are solutions of $(S^2 - S - cz)\mathbf{x} = 0$. Now

$$S^2 - S - cz\mathbf{1} \;=\; (S - \lambda_+\mathbf{1})(S - \lambda_-\mathbf{1}), \quad \lambda_\pm = \frac{1}{2} \pm \frac{1}{2}\sqrt{1+4cz}.$$

(Here we take the branch of the square root that is positive for $cz > 0$ and holomorphic on the complement of $\{z : \operatorname{Re} cz \le 0\}$.) Therefore the sequences $\{P_n\}$ and $\{Q_n\}$ are linear combinations of the sequences $\{\lambda_+^{n+1}\}$, $\{\lambda_-^{n+1}\}$. Taking into account the initial conditions at $n = -1$, $n = 0$, we find that

$$P_n(z) \;=\; \frac{c}{2\sqrt{1+4cz}}[\lambda_+^{n+1} - \lambda_1^{n+1}]; \qquad Q_n(z) \;=\; \frac{1}{2\sqrt{1+4cz}}[\lambda_+^{n+2} - \lambda_-^{n+2}].$$

Now throughout the sector on which we have defined $\lambda_\pm$, the principal branch of $\log \lambda_\pm$ is positive for λ_+ and negative for λ_-, so λ_-^n is exponentially small as $n \to \infty$. Therefore, since $\lambda_+\lambda_- = -cz$,

$$\frac{P_n(z)}{Q_n(z)} \sim \frac{c}{\lambda_+} = -\frac{\lambda_-}{z} = \frac{\sqrt{1+4cz} - 1}{2z}.$$

Thus we have proved convergence and verified (12.3.14).

12.4 The Stieltjes transform, Padé approximants, and orthogonal polynomials

Suppose that μ is a measure defined on an interval $I = (a, b) \subset \mathbb{R}$. For our purpose here we may assume that μ is defined by a non-negative density function $w : I \to [0, \infty)$. The associated Hilbert space $L^2(I, w(t)\,dt)$ is the completion of the space of continuous functions φ such that $\int_a^b |\varphi(t)|^2 w(t)\,dt < \infty$ with respect to the norm $||\varphi|| = (u, u)_w^{1/2}$ that corresponds to the inner product

$$(\varphi, \psi)_w \;=\; \int_a^b \varphi(t)\overline{\psi(t)}w(t)\,dt.$$

Two other objects associated to w are the *Stieltjes transform*

$$F(z) \;=\; \int_a^b \frac{w(t)\,dt}{z - t}, \quad z \notin [a, b], \tag{12.4.1}$$

and the set of *moments* of w. We assume that $\int_a^b x^{2n}\, w(x)\, dx,\ n = 0, 1, 2, \ldots$ is finite; then the moments are the constants

$$c_n \;=\; \int_a^b x^n\, w(x)\, dx. \tag{12.4.2}$$

It is convenient to assume that the measure is normalized with

$$c_0 \;=\; \int_a^b w(x)\, dx \;=\; 1. \tag{12.4.3}$$

Let us look for the [n-1,n] Padé approximants $P - n/Q_n$ to the Stieltjes transform F, where deg $P_n = n - 1$, deg $Q_n = n$. Since we have assumed $c_0 = 1$, we may take P_n and Q_n to be *monic*: having leading coefficients equal to 1. The Padé condition is

$$Q_n(z)F(z) \;=\; P_n(z) + O(z^{-n-1}). \tag{12.4.4}$$

For any monic polynomial Q of degree n,

$$Q(z)F(z) \;=\; \int_a^b \frac{Q(z)}{z-t} w(t)\, dt \;=\; P(z) + R(z),$$

where

$$P(z) \;=\; \int_a^b \frac{Q(z) - Q(t)}{z-t}\, w(t)\, dt \tag{12.4.5}$$

has degree $n - 1$. Expanding $(z - t)^{-1}$ as $z \to \infty$ along a non-real ray,

$$\begin{aligned} R(z) &= \int_a^b \frac{Q(t)}{z-t} w(t)\, dt \\ &= \sum_{k=0}^{n-1} \left\{ \int_a^b Q(t) t^k w(t)\, dt \right\} z^{-k-1} + O(z^{-n-1}). \end{aligned}$$

It follows that $Q_n = Q$ satisfies (12.4.4) if and only if Q_n is orthogonal to every polynomial of lower degree:

$$(Q_n, P)_w \;=\; 0 \qquad \text{if} \quad \deg P < n.$$

If so, then $P_n = P$ is defined by (12.4.5).

It is not difficult to see that there is a unique sequence $\{Q_n\}$ of monic polynomials that are mutually orthogonal:

$$(Q_n, Q_m)_w \;=\; 0, \qquad n \neq m.$$

In fact the n linear equations $(Q_n, t^k)_w = 0, 0 \leq k < n$ determine the n lower degree coefficients of Q_n. Note that $Q_0 = 1,\ Q_1 = z - b$, where $b = c_1 - 1$. For $n \geq 2$, $Q_n - zQ_{n-1}$ has degree $n - 1$ and is orthogonal to all polynomials of degree $\leq n - 3$, so it is a linear combination of Q_{n-1} and Q_{n-2}:

$$Q_n = (z - b_n)Q_{n-1} + a_n Q_{n-2}, \qquad n \geq 2. \tag{12.4.6}$$

It follows from (12.4.5) that P_n satisfies the same three-term recursion, with

$$P_0 = 0, \qquad P_1 = 1.$$

Since $(Q_n, Q_{n-2})_w = 0 = (Q_{n-1}, Q_{n-2})_w$, (12.4.6) implies

$$\begin{aligned} a_n||Q_{n-2}||_w^2 = (a_n Q_{n-2}, Q_{n-2})_w &= -(zQ_{n-1}, Q_{n-2})_w \\ = -(Q_{n-1}, zQ_{n-2})_w &= -||Q_{n-1}||_w^2, \end{aligned}$$

since zQ_{n-2} differs from Q_{n-1} by a polynomial of lower degree. Therefore $a_n < 0$. Comparing this with the discussion at the beginning of Section 12.2, we see that if we take $a_1 = 1$, the $[n-1, n]$ Padé approximant of F is the n-th convergent of the continued fraction

$$F(z) \sim \cfrac{1}{z - b_1 + \cfrac{a_2}{z_2 - b_2 + \cfrac{a_3}{z - b_3 + \dots}}}. \tag{12.4.7}$$

Let us estimate the remainder term $R_n = F - P_n/Q_n$:

$$\begin{aligned} F(z) - \frac{P_n(z)}{Q_n(z)} &= \int_a^b \frac{w(t)\,dt}{z-t} - \frac{1}{Q_n(z)} \int_a^b \frac{Q_n(z) - Q_n(t)}{z-t}\, w(t)\,dt \\ &= \frac{1}{Q_n(z)} \int_a^b Q_n(t) w(t)\,dt. \end{aligned} \tag{12.4.8}$$

Since we have assumed that $\int_a^b w(t)\,dt = 1$, the Cauchy–Schwarz inequality yields

$$\int_a^b |Q_n(t)| w(t)\,dt \leq ||Q_n||_w. \tag{12.4.9}$$

In this connection, we may use the following to make more explicit estimates.

Lemma 12.4.1. *The monic orthogonal polynomial Q_n has minimal L^2 norm $|| \, ||_w$ among all monic polynomials of degree n.*

Proof. Suppose that Q is a monic polynomial of degree n. Then $P = Q - Q_n$ has degree $< n$, so $(Q_n, P)_w = 0$. Therefore

$$||Q||_w^2 = (Q, Q)_w = (Q_n + P, Q_n + P)_w = (Q_n, Q_n)_w + (P, P) = ||Q_n||_w^2 + ||P||_w^2,$$

so $||Q||_w > ||Q_n||_w$ unless $Q = Q_n$. □

Using the monic polynomials x^n in the general case, and the polynomials $(x - \frac{1}{2}[b-a])^n$ in the case of a finite interval, we see that (12.4.8), (12.4.9), and Lemma 12.4.1 imply the following estimates.

Theorem 12.4.2. *Let $d(z)$ be the distance from $z \in \mathbb{C}$ to the interval $I = (a, b)$. Then $d(z) \ge |\mathrm{Im}\, z|$ and*

$$\left| F(z) - \frac{P_n(z)}{Q_n(z)} \right| \le \frac{1}{d(z)^n} \left\{ \int_a^b t^{2n} w(t)\, dt \right\}^{1/2} = \frac{\sqrt{c_{2n}}}{d(z)^n}. \tag{12.4.10}$$

If the interval is finite, then $d(z) \sim |z|$ as $z \to \infty$, and

$$\left| F(z) - \frac{P_n(z)}{Q_n(z)} \right| \le \frac{1}{d(z)^n} \cdot \left(\frac{b-a}{2} \right)^n. \tag{12.4.11}$$

Remark. The discussion and the results of this section are unchanged if we consider a general Borel measure μ in place of one determined by a weight function w, so long as all the moments $\int_I t^{2n}\, d\mu(t)$ are finite. The most general type, in the case of $\mathbb{R}$, is a *Riemann–Stieltjes integral.*

Such an integral is determined by a bounded, non-decreasing function ψ on $\mathbb{R}$. We may normalize it by assuming that $\lim_{x\to-\infty} \psi(x) = 0$ and that ψ is continuous from the left. The measure of an interval (a, b) is $\psi(b) - \psi(a)$. The integral of a continuous function that vanishes outside a bounded interval $[a, b]$ is the limit of the sums

$$\sum_{k=1}^{n} f(x_k)[\psi(x_k) - \psi(x_{k-1})]$$

over partitions $x_0 < a < x_1 < \cdots < x_n = b$ as $\sup\{|x_k - x_{k-1}|\} \to 0$. The integral is extended to more general functions f by taking appropriate limits.

As an example, consider the step function

$$\phi(x) = 0, \qquad x < 0. \qquad \phi(x) = 1, \quad x \ge 0.$$

For any continuous function $f : \mathbb{R} \to \mathbb{C}$,

$$\int_{-\infty}^{\infty} f(x-t)\, d\phi(t) = f(x).$$

The corresponding measure μ is often denoted $\delta(t)dt$, where δ is the *Dirac delta "function"*, which vanishes outside the origin and is thought of as being infinite at the origin to just the correct amount so that $\int_{\mathbb{R}} \delta(t)\, dt = 1$. (This is the reason for taking partitions that start to the left of a, above, so that one can have a jump "at" $x = a$ even if the interval in question starts at a.)

12.5 Characterization of Stieltjes transforms

In general, the question of convergence of a Padé sequence can be difficult. In this section we consider a class of functions for which the theory is quite complete. As we showed in Section 12.4, if F is the Stieltjes transform of a positive measure $d\psi$ on the line, all of whose moments are finite, then the Padé approximants converge to F on the complement of the support of the measure. In this section we ask how to characterize such Stieltjes transforms F, and how to determine the measure from the moments

$$c_k = \int t^k \, d\psi(t), \qquad k = 0, 1, 2, \ldots . \tag{12.5.1}$$

We make some changes from the discussion in Section 12.4. First, we assume that the interval $I = (a, b)$ is not all of $\mathbb{R}$, so up to a translation and, if necessary, a change of orientation, we assume $I \subset (0, \infty)$. Second, we change some signs in the Stieltjes transform, and characterize functions

$$f(z) = T(\mu) = \int_0^\infty \frac{d\mu(t)}{z+t}, \qquad z \notin (-\infty, 0], \tag{12.5.2}$$

where μ is a measure all of whose moments are finite. Let us note the particular case: μ supported at the origin, with $\mu(\{0\}) = c > 0$. Then $c_0 = c$, $c_n = 0$ for $n > 0$, and

$$f(z) = \frac{c}{z}. \tag{12.5.3}$$

It will be convenient to consider (12.5.2) as a Riemann–Stieltjes integral with respect to the non-decreasing function

$$\psi(t) = 0, \quad t < 0, \qquad \psi(t) = \mu([0, t)), \quad t \geq 0.$$

Thus

$$f(z) = \int_0^\infty \frac{d\psi(t)}{z+t}, \qquad z \notin (-\infty, 0]. \tag{12.5.4}$$

The function ψ can be recovered from f. We leave the following as Exercise 10.

Proposition 12.5.1. *The function (12.5.4) is the limit*

$$\psi(x) = \lim_{\varepsilon \downarrow 0} \frac{1}{2\pi i}\left[f(-x - i\varepsilon) - f(-x + i\varepsilon)\right].$$

The characterization theorem is a consequence of a classic result of Herglotz [104] and Riesz [176]:

Theorem 12.5.2. *Suppose that $g : \mathbb{D} \to \mathbb{C}$ is holomorphic and $g(0) > 0$. Then g has positive real part if and only if there is a positive measure μ on $[0, 2\pi]$ such that*

$$g(z) = \int_0^{2\pi} \frac{e^{i\theta} + z}{e^{i\theta} - z} \, d\mu(\theta). \tag{12.5.5}$$

Proof. Suppose that g is defined by (12.5.5), with μ such a measure. Then g is holomorphic in $\mathbb{D}$. For $z = 0$ we get $g(0) = \int_0^{2\pi} d\mu(t) > 0$. Taking the real part of (12.5.5) gives

$$\operatorname{Re} g(z) = \int_0^{2\pi} \frac{1-|z|^2}{|e^{i\theta}-z|^2}\, d\mu(\theta), \qquad 0 \le r < 1. \tag{12.5.6}$$

The integrand is positive, so $\operatorname{Re} g > 0$.

Conversely, suppose that g is holomorphic in $\mathbb{D}$, with positive real part. Let $g_\varepsilon(z) = g(z/(1+\varepsilon))$, $0 < \varepsilon < 1$. Then g_ε is continuous on the closure of $\mathbb{D}$. By Theorem 5.1.6,

$$g_\varepsilon(z) = \int_0^{2\pi} \frac{e^{i\theta}+z}{e^{i\theta}-z} h_\varepsilon(e^{i\theta})\, d\theta, \qquad h_\varepsilon(z) = \operatorname{Re} g_\varepsilon(z).$$

By assumption, each h_ε is positive. Let us use h_ε to define a non-decreasing function and a corresponding Riemann–Stieltjes measure

$$\phi_\varepsilon(\theta) = \int_0^\theta h_\varepsilon(e^{is})\, ds, \quad 0 \le \theta \le 2\pi; \qquad \mu_\varepsilon = d\phi_\varepsilon.$$

Then $\mu_\varepsilon(\partial\mathbb{D}) = \phi_\varepsilon(2\pi) > 0$. Note that $\phi_\varepsilon(2\pi) = g(0)$.

By the usual diagonal process we may find a subsequence $\{\varepsilon_n\}$ of the sequence $\{1/n\}$ such that each limit

$$a_m = \lim_{n\to\infty} \int_0^{2\pi} e^{im\theta}\, d\phi_{\varepsilon_n}, \quad m \in \mathbb{Z} \tag{12.5.7}$$

exists. By the Weierstrass approximation theorem, Corollary 5.1.2, linear combinations of the $\{e^{im\theta}\}$ are dense in $C(\partial\mathbb{D})$. Since each μ_ε has total mass $g(0)$, it follows that the linear maps $\lambda_n : C(\partial\mathbb{D}) \to \mathbb{C}$ defined by

$$\lambda_n(u) = \int_0^{2\pi} u(\theta)\, d\phi_{\varepsilon_n}(\theta)$$

converge to a linear map λ such that $\lambda(u) \ge 0$ if $u \ge 0$, and

$$|\lambda(u)| \le g(0) \sup_\theta |u(\theta)|, \qquad \lambda(u) = g(0) \text{ if } u \equiv 1.$$

This is one characterization of a measure μ of total mass $g(0)$ on $\partial\mathbb{D}$:

$$\int_0^{2\pi} u(\theta)\, d\mu(\theta) = \lambda(u).$$

On any given compact subset of $\mathbb{D}$, the g_{ε_n} converge uniformly to g, so

$$g(z) = \lim_{n\to\infty} g_{\varepsilon_n}(z) = \lim_{n\to\infty} \int_0^{2\pi} \frac{e^{i\theta}+z}{e^{i\theta}-z}\, d\mu_{\varepsilon_n}(\theta) = \int_0^{2\pi} \frac{e^{i\theta}+z}{e^{i\theta}-z}\, d\mu(\theta). \qquad \square$$

Theorem 12.5.3. *Let f be a function defined on the complement of the real interval* $(-\infty, 0]$. *Then $f(-z)$ has the form (12.5.2), where μ is a positive measure on* $[0, \infty)$, *all of whose moments are finite, if and only the following conditions hold:*

(i) f is holomorphic;

(ii) $f(x) > 0$ for $x > 0$;

(iii) for any $x \in \mathbb{R}$ and $y \neq 0$, $y \operatorname{Im} f(x+iy) < 0$;

(iv) there is a sequence of constants c_0, c_1, c_2, ... such that for each $\varepsilon > 0$, the function $f(z) \sim \sum_{n=0}^{\infty} c_n z^{-n-1}$ as $z \to \infty$ in the sector $|\arg z| \leq \pi - \varepsilon$.

Proof. Suppose that f is given by (12.5.2), where μ is such a measure. It is easily checked that f satisfies conditions (i), (ii), and (iii). To verify condition (iv), we expand

$$\frac{1}{z+t} = \sum_{k=0}^{n-1} (-1)^k \frac{t^k}{z^{k+1}} + (-1)^n \frac{t^n}{z^n} \frac{1}{z+t}. \tag{12.5.8}$$

Therefore as $z \to \infty$ in the sector,

$$f(z) = \sum_{k=0}^{n-1} (-1)^k \frac{c_k}{z^{k+1}} + O(z^{-n-1}),$$

where the c_k are the moments of μ.

Conversely, suppose that f satisfies conditions (i), (ii), and (iii). We use the inverse of the Cayley transform, $C^{-1} : \mathbb{D} \to \mathbb{H}$, to transfer f to the function

$$g(w) = if \circ C^{-1}(w) = if\left(i \frac{1+w}{1-w}\right).$$

Then g is holomorphic on $\mathbb{D}$ and has positive real part. Moreover, f is holomorphic across $(0, \infty)$ and is real on this half-line. Theorem 12.5.2 tells us that there is a positive measure $\nu = d\phi$ on $\partial\mathbb{D}$ such that

$$g(w) = ia + \int_{-\pi}^{\pi} \frac{e^{i\theta} + w}{e^{i\theta} - w} \, d\phi(\theta),$$

where a is the imaginary part of $g(0)$. Then

$$\begin{aligned} f(z) = -ig\left(\frac{z-i}{z+i}\right) &= a - i \int_{-\pi}^{\pi} \frac{e^{i\theta}(z+i) + (z-i)}{e^{i\theta}(z+i) - (z-i)} \, d\phi(\theta) \\ &= a - i \int_{-\pi}^{\pi} \frac{e^{i\theta/2}(z+i) + e^{-i\theta/2}(z-i)}{e^{i\theta/2}(z+i) - e^{-i\theta/2}(z-i)} \, d\phi(\theta) \\ &= a - i \int_{-\pi}^{\pi} \frac{z\cos(\theta/2) - \sin(\theta/2)}{iz\sin(\theta/2) + i\cos(\theta/2)} \, d\phi(\theta). \end{aligned}$$

Let $t = \cot(\theta/2)$ and $\psi(t) = \frac{1}{2}\phi(2\cot^{-1} t)$. Then

$$d\phi(\theta) = \frac{d\psi(t)}{1+t^2},$$

so

$$f(z) = a + \int_0^\infty \frac{1-tz}{t+z}\,\frac{d\psi(t)}{1+t^2}. \tag{12.5.9}$$

By assumption, $f(z) \to 0$ as $z \to \infty$ in any sector $|\arg z| < \pi$. Therefore (12.5.9) implies that

$$a = \int_0^\infty \frac{t\,d\psi(t)}{1+t^2}.$$

Since $t + (1-tz)/(t+z) = (1+t^2)/(t+z)$, (12.5.9) becomes

$$f(z) = \int_0^\infty \left[t + \frac{1-tz}{t+z}\right] \frac{d\psi(t)}{1+t^2} = \int_0^\infty \frac{d\psi(t)}{z+t}. \tag{12.5.10}$$

We have proved this for $\operatorname{Im} z > 0$, but continuation across $(-\infty, 0)$ establishes it for all z in the complement of $(-\infty, 0]$.

It remains to show that the moments of $\mu = d\psi$ are finite. This is a consequence of assumption (iii); see Exercise 11. □

12.6 Stieltjes functions and Padé approximants

Given a modified Stieltjes transform

$$f = T(\mu) = \int_0^\infty \frac{d\mu(t)}{z+t},$$

we define the associated *Stieltjes function* $F = F_\mu$ in the complement of $(-\infty, 0]$ by

$$F(z) = \frac{1}{z} f\left(\frac{1}{z}\right) = \int_0^\infty \frac{d\mu(t)}{1+zt} \tag{12.6.1}$$

where again all moments of μ are assumed to be finite. An example is a constant function $F(t) \equiv c > 0$. In fact this corresponds to f of (12.5.2), where the measure μ has total mass c supported on $\{0\}$.

The relation (12.6.1) is reciprocal: $f(z) = z^{-1}F(z^{-1})$. This relation between F and f, and the integral formula (12.6.1), makes it easy to verify the following characterization; see Exercise 12.

Theorem 12.6.1. *A function F, defined for z in the complement of $(-\infty, 0)$, is the Stieltjes function for a positive measure on $[0, \infty)$ if and only if it satisfies the properties*

(i) F is holomorphic on the complement of $(-\infty, 0]$;

(ii) $F(x) > 0$ for $x \geq 0$;

(iii) $\operatorname{Im}(z\,F(z))$ *and* $\operatorname{Im} z$ *have the same sign, for* $\operatorname{Im} z \neq 0$*;*

(iv) there is a sequence of constants $a_0 > 0$*,* a_1*,* a_2*, ... such that for any* $\varepsilon > 0$*,* $F(z) \sim \sum_{n=0}^{\infty}(-1)^n a_n z^n$ *as* $z \to 0$*,* $|\arg z| \leq \pi - \varepsilon$*.*

It is easily checked that the coefficients in the expansion (iv) of a Stieltjes function are the moments

$$a_k = \int_0^{\infty} t^k \, d\mu(t).$$

Let us turn to the Padé approximants of Stieltjes functions.

Lemma 12.6.2. *Suppose that* μ *is a positive measure on* $[0, \infty]$ *and* $\int_0^{\infty} d\mu(t)$ *is finite. Then the functions*

$$f(z) = \int_0^{\infty} \frac{d\mu(t)}{z+t}, \qquad F(z) = \int_0^{\infty} \frac{d\mu(t)}{1+z\,t}, \qquad \operatorname{Im} z \neq 0$$

satisfy

$$\operatorname{Im} z \operatorname{Im} f(z) < 0, \qquad \operatorname{Im} z \operatorname{Im} F(z) < 0; \tag{12.6.2}$$

$$\operatorname{Im} z \operatorname{Im}(zf(z)) > 0, \qquad \operatorname{Im} z \operatorname{Im}(zF(z)) > 0. \tag{12.6.3}$$

Proof. This follows immediately from

$$\frac{1}{z+t} = \frac{\bar{z}+t}{|z+t|^2}, \qquad \frac{z}{z+t} = \frac{|z|^2 + tz}{|z+t]^2}$$

and the analogous identities for the integrand of F, since these integrands are bounded functions of t and therefore integrable with respect to μ. □

Lemma 12.6.3. *Suppose that the function* F *is a Stieltjes function. Then the same is true of the functions* G_1 *and* G_2 *defined by*

$$\frac{F(z)}{F(0)} = \frac{1}{1+zG_1(z)}; \qquad G_2(z) = \frac{c}{1+zF(z)}, \quad c > 0. \tag{12.6.4}$$

Proof. First

$$G_1(z) = \left[\frac{F(0)}{F(z)} - 1\right] z^{-1}. \tag{12.6.5}$$

It follows from (12.6.5) and from the second equation in (12.6.4) that, since F is a Stieltjes function, G_1 and G_2 satisfy conditions (i) and (iv) of Theorem 12.6.1. Condition (ii) is immediate for G_2, and for G_1 it follows from the fact that for $x > 0$,

$$F(0) - F(x) = \int_0^{\infty} \frac{x\,t}{1+x\,t}\, d\mu(t) > 0.$$

As for condition (iv),

$$z\,G_1(z) \;=\; \frac{F(0)}{F(z)} - 1; \qquad z\,G_2(z) \;=\; \frac{c}{z^{-1} + F(z)},$$

and Lemma 12.6.2 shows that the imaginary part of $zG_k(z)$ has the same sign as $\operatorname{Im} z$. □

Theorem 12.6.4. *If F is a Stieltjes function then all the constants c_n in the continued fraction expansion*

$$\cfrac{c_0}{1 + \cfrac{c_1 z}{1+ \ddots \cfrac{c_{2n} z}{1 + \dots}}} \tag{12.6.6}$$

of F are positive.

Proof. Let $G_0 = F$ and $c_0 = F(0)$, and define G_n inductively:

$$G_k(z) \;=\; \frac{c_k}{1 + zG_{k+1}(z)}, \qquad c_k = G_k(0), \quad k = 1, 2, 3, \dots.$$

By Lemma 12.6.5, these functions are Stieltjes functions. Therefore (12.6.6) with coefficients

$$c_k \;=\; G_k(0) = \int_0^\infty d\mu_k(t) \;>\; 0$$

is the continued fraction associated with $G_0 = F$.

Conversely, suppose that the coefficients in the continued fraction (12.6.6) are all positive. For any given positive integer let $G_0^{(n)} \equiv c_n$. This is a Stieltjes function. Then by Lemma 12.6.3 so are the functions defined iteratively by

$$G_k^{(n)}(z) \;=\; \frac{c_{n-k}}{1 + zG_{k-1}^{(n)}(z)}, \qquad k = 1, 2, \dots, n.$$

Then $F_n = G_n^{(n)}$ is the n-th convergent of (12.6.6), and has a representation

$$F_n(z) \;=\; \int_0^\infty \frac{d\mu_n(t)}{1 + z\,t},$$

where μ_n is a positive measure on $[0, \infty)$. It has total mass

$$\int_0^\infty d\mu_n(t) \;=\; F_n(0) \;=\; c_0.$$

Therefore, as in the proof of Theorem 12.5.2, some subsequence $\{\mu_{2n_k}\}$ of $\{\mu_{2n}\}$ converges to a positive measure μ with total mass c_0.

We know that $F_n = P_n/Q_n$ where P_n and Q_n are polynomials that satisfy

$$P_n = P_{n-1} + c_n z P_{n-2}, \qquad P_{-1} = 0,\ P_0 = c_0; \tag{12.6.7}$$

$$Q_n = Q_{n-1} + c_n z Q_{n-2}, \quad Q_{-1} = Q_0 = 1. \tag{12.6.8}$$

In particular, positivity of the c_k implies that for $x > 0$

$$1 = Q_0(x) < Q_1(x) = 1 + c_1 x < Q_2(x) < \dots . \tag{12.6.9}$$

As in (12.2.7), the recursion (12.6.8) leads to

$$F_n(z) - F_{n-1}(z) = \frac{(-1)^n c_0 c_1 \cdots c_n z^n}{Q_n(z) Q_{n-1}(z)}.$$

The same kind of calculation leads to

$$F_{n+1}(z) - F_{n-1}(z) = \frac{(-1)^n c_0 c_1 \cdots c_n z^n}{Q_{n+1}(z) Q_{n-1}(z)}.$$

It follows from these identities and (12.6.9) that for $x > 0$,

$$\frac{c_0}{1 + c_1 x} = F_1(x) < F_3(x) < F_5(x) < \cdots < F_4(x) < F_2(x) < F_0(x) = c_0.$$

Now

$$\lim_{k\to\infty} F_{2n_k}(z) = \lim_{k\to\infty} \int_0^\infty \frac{d\mu_{2n_k}}{1 + z\,t} = \int_0^\infty \frac{d\mu(t)}{1 + zt}.$$

Thus the limiting function F is Stieltjes. It has the continued fraction expansion whose $2n$-th convergent is F_{2n}. Now for each fixed $x > 0$, $F_{2n}(x)$ decreases, so

$$\lim_{n\to\infty} F_{2n}(x) = \lim_{k\to\infty} F_{2n_k}(x) = \int_0^\infty \frac{d\mu(t)}{1 + xt}.$$

Thus F has continued fraction expansion (12.6.6). □

We know from Proposition 12.3.1 that the convergent F_n is the $[n-1, n]$ Padé approximant $[n, n-1]_F$ if n is even, and is $[n, n]_F$ if n is odd. Therefore Theorem 12.6.4 and its proof give us the following.

Corollary 12.6.5. *For any Stieltjes function F, the Padé approximants satisfy*

$$\begin{aligned} 0 < [1,0]_F(x) < [3,2]_F(x) < [5,4]_F(x) < \dots \\ < \dots [4,4]_F(x) < [2,2]_F(x) < [0,0]_F(x) = F(0) \end{aligned}$$

for $x \geq 0$.

Remarks. Recall that the coefficients in the expansion

$$F(z) \sim \sum_{n=0}^{\infty}(-1)^n a_n z^n, \qquad F(z) = \int_0^{\infty} \frac{d\mu(t)}{1+z\,t}$$

are the moments $a_n = \int_0^{\infty} t^n \, d\mu(t)$. These can be computed from the convergents F_n of the continued fraction (12.6.6), and conversely. One question is: do the moments determine μ uniquely? The answer is no, in general; see [9], [186]. For example, the measures with density functions

$$w(t) = \exp(-t^{1/4})[1 - a\sin(t^{1/4})], \qquad 0 \le a < 1$$

all have the same moments $a_n = 4 \cdot (4n+3)\,!$. One positive result is due to Carleman [40]

Theorem 12.6.6. *If the moments a_n of the positive measure μ on* $[0, \infty)$ *satisfy*

$$\sum_{n=0}^{\infty} a_n^{-1/2n} = \infty,$$

then μ is uniquely determined.

12.7 Generalized Shanks Transformation

There is no general method for transforming the terms of a convergent series

$$A = \sum_{n=0}^{\infty} a_n = \lim_{n\to\infty} A_n, \qquad A_n = \sum_{k=1}^{n} a_k,$$

so that the transformed series converges more rapidly. However if something is known about the (approximate) form of the remainder terms $A - A_n$, such a transformation may be possible. Let us write $A_n = A + \varepsilon_n$. By assumption, $\varepsilon_n \to 0$. The system

$$A_{n+1} = A + \varepsilon_{n+1}, \quad A_n = A + \varepsilon_n, \quad A_{n-1} = A + \varepsilon_{n-1}$$

gives

$$A = \frac{\Delta_n A_{n-1} - A_n \Delta_{n-1}}{\Delta_n - \Delta_{n-1}} - \frac{\varepsilon_n^2 - \varepsilon_{n+1}\varepsilon_{n-1}}{\Delta_n - \Delta_{n-1}}, \qquad \Delta_n = A_{n+1} - A_n. \qquad (12.7.1)$$

This produces the limit A as an explicit function of any three successive A_n, provided the denominator does not vanish and the $\varepsilon_n \neq 0$ satisfy $\varepsilon_n^2 - \varepsilon_{n-1}\varepsilon_{n+1} = 0$, i.e.

$$\frac{\varepsilon_{n+1}}{\varepsilon_n} = \frac{\varepsilon_n}{\varepsilon_{n-1}} = \lambda \neq 0.$$

If this is the case, then

$$\varepsilon_n = \frac{\varepsilon_{n-1}}{\varepsilon_{n-2}} \cdots \frac{\varepsilon_1}{\varepsilon_0}\varepsilon_0 = \lambda^n \varepsilon_0.$$

The *Aitkens Δ^2 process* [8] converts the sequence $\{A_n\}$ to the sequence $\{T(A)\}$,

$$\begin{aligned} T(A)_n &= \frac{A_{n+1}A_{n-1} - A_n^2}{A_{n+1} - 2A_n + A_{n-1}} \\ &= A_{n+1} - \frac{(A_{n+1} - A_n)^2}{(A_{n+1} - A_n) - (A_n - A_{n-1})} \\ &= A_{n+1} - \frac{\Delta_n^2}{\Delta_n - \Delta_{n-1}} \end{aligned} \tag{12.7.2}$$

(Note that if the denominator is 0 for $n \geq 1$, then $\{A_n\}$ is constant.) Thus, if $A_n = \alpha\lambda^n$ for some (possibly unknown) constants α, λ, with $0 < |\lambda| < 1$, then the sequence (12.7.2) converges and the transformed sequence is constant and immediately gives the limit A. Conversely, if the transformed sequence is constant, then $A_n = \alpha\lambda^n$ for some α and λ, $0 < |\lambda| < 1$. More generally, if $A_n = \alpha\lambda^n + \varepsilon_n$, where ε_n/λ^n is small, then the sequence $\{T(A)_n\}$ can be expected to converge more rapidly than $\{A_n\}$; see Exercise 15.

The Aitken's process (12.7.2) is sometimes called the *Shanks transformation*, because of Shanks's generalization to the case of more than one "transient," as in the case of the series

$$A = \sum_{n=0}^{\infty} [\alpha_1\lambda_1^n + \alpha_2\lambda_2^n]. \tag{12.7.3}$$

To improve the convergence of series like (12.7.3), we look for a generalization of the nonlinear transformation T.

We call a term in the remainder of the power series a *transient*, if it decays like λ^n for some $0 < \lambda < 1$ and other parts of the remainder decay more rapidly. If A_n has k distinct transient terms

$$A_n = A + \sum_{j=1}^{k} \alpha_j \lambda_j^n,$$

then A can be determined by the $(2k+1)$ terms $A_{n-k}, A_{n-k+1}, \cdots, A_{n+k}$. The solution A of this system of $2k+1$ equations with $2k+1$ unknowns is the *kth-order Shanks transformation*, given by a ratio of determinants

$$A = S_k(A)_n = \frac{\begin{vmatrix} A_{n-k} & \cdots & A_{n-1} & A_n \\ \Delta A_{n-k} & \cdots & \Delta A_{n-1} & \Delta A_n \\ \Delta A_{n-k+1} & \cdots & \Delta A_n & \Delta A_{n+1} \\ \vdots & & \vdots & \vdots \\ \Delta A_{n-1} & \cdots & \Delta A_{n+k-2} & \Delta A_{n+k-1} \end{vmatrix}}{\begin{vmatrix} 1 & \cdots & 1 & 1 \\ \Delta A_{n-k} & \cdots & \Delta A_{n-1} & \Delta A_n \\ \Delta A_{n-k+1} & \cdots & \Delta A_n & \Delta A_{n+1} \\ \vdots & & \vdots & \vdots \\ \Delta A_{n-1} & \cdots & \Delta A_{n+k-2} & \Delta A_{n+k-1} \end{vmatrix}},$$

where $\Delta A_p = A_{p+1} - A_p$. Note that $S_1(A)_n = T(A)_n$.

The Taylor series for the function $f(z) = 1/(z+1)(z+2)$ is a very slowly convergent series. The nth partial sum of this Taylor series is

$$\begin{aligned} f_n(z) &= \sum_{k=0}^{n} (-1)^k \left(1 - \frac{1}{2^{k+1}}\right) z^k \\ &= \frac{1}{(z+2)(z+1)} - \frac{(-z)^{n+1}}{z+1} - \frac{(-z/2)^{n+1}}{z+2}. \end{aligned}$$

The poles of $f(z)$ at $z = -1$ and $z = -2$ affect the rate of convergence of $f_n(z)$ to $f(z)$, and they are the origin of the two transients of $f_n(z)$. When S_k is applied to a sequence of partial sums whose convergence is governed by two transients, the result is exact, that is, $S_k[f_n(z)] = f(z)$ for all $k \geq 2$. If the function $f(z)$ has p simple poles, then its partial sums have p transient terms; see Exercise 16. Moreover S_k applied to the partial sums is exact for $k \geq p$.

The higher order Shanks transformations S_k are closely related to Padé approximants. In fact, this treatment can be regarded as an alternative derivation of the Padé approximants. It can be shown that if $k \leq n$, then $S_k(A_n)$ is identical to the Padé approximant $P_n(z)/Q_k(z)$ of the series $A_1 + \sum_{j=1}^{\infty}(A_{j+1} - A_j)z^j$ evaluated at $z = 1$; see Exercise 17.

For large k, the determinants in the transformations S_k are not easily computed. The ε-*algorithm* developed by Wynn [221], with techniques for implementation due to Wynn and others, is used to make computation efficient.

12.8 Examples

The Stieltjes function

$$f(z) = \int_0^\infty \frac{e^{-t}\,dt}{1 + z\,t} \qquad |\arg z| < \pi$$

has asymptotic expansion

$$f(z) \sim \sum_{n=0}^{\infty} \left\{ \int_0^{\infty} t^n e^{-t}\, dt \right\} (-z)^n = \sum_{n=0}^{\infty} n\,!\,(-z)^n; \tag{12.8.1}$$

see Exercise 18. The series diverges for $z \neq 0$. The standard way to obtain approximate values is by truncating the series after the minimal term. Since

$$\frac{(n+1)!\, z^{n+1}}{n!\, z^n} = (n+1)z,$$

this means summing to $n \sim |z|^{-1}$.

Table 12.1 illustrates convergence of the diagonal Padé approximants for this series.

Table 12.1 Stieltjes Series

n	$P_n(1)/Q_n(1)$	$P_n(10)/Q_n(10)$
6	0.59682	0.24256
7	0.59657	0.23284
8	0.59646	0.22593
9	0.59641	0.22086
10	0.59638	0.21706
50	0.59635	0.20156
∞	0.59635	0.20146

The next example exhibits a two-point expansion. Consider the function $f(z)$ given by the integral

$$f(z) = \frac{1}{2\sqrt{z}} e^{-z} \int_0^z \frac{e^t}{\sqrt{t}} dt, \tag{12.8.2}$$

which is a solution to the differential equation $2zf'(z) = -(1+2z)f(z) + 1$. This solution has power series expansions at both $z = 0$ and $z = \infty$. The differential equation allows one to compute the coefficients recursively, by plugging a proposed expansion into (12.8.2) and looking at the coefficient of z^k or z^{-k}. At $z = 0$ we obtain the Taylor series

$$f(z) = \sum_{n=1}^{\infty} a_n z^n, \qquad a_n = \frac{(-4)^n n!}{(2n+1)!}, \tag{12.8.3}$$

which converges for all finite z. At $z = \infty$, we obtain the divergent asymptotic expansion

$$f(z) \sim \sum_{n=1}^{\infty} \frac{b_n}{z^n}, \qquad b_n = \frac{2(2n-2)!}{4^n (n-1)!}, \qquad n \geq 1. \tag{12.8.4}$$

Here we discuss only the diagonal Padé sequence $P_n(z)/Q_n(z)$, and use as input $k = n+1$ terms of the Taylor series (12.8.3) at $z = 0$ and $l = n$ terms of the asymptotic series (12.8.4) at $z = \infty$. Write

$$P_n(z) \;=\; \sum_{k=0}^{n} A_k z^k, \quad Q_n(z) = 1 + \sum_{k=1}^{n} B_k z^k.$$

The approximation property (12.1.2) at $z_0 = 0$ becomes the system of equations

$$A_k \;=\; \sum_{j=0}^{k} a_{k-j} B_j, \qquad 0 \le k \le n. \tag{12.8.5}$$

Similarly, the approximation property (12.1.13) at $z_1 = \infty$ becomes, by looking at powers z^{-j}, the system

$$A_k \;=\; \sum_{j=n-k}^{n} B_j b_{j-k}, \qquad 1 \le k \le n. \tag{12.8.6}$$

In Table 12.2 we take $z = x$ to be real. The values of the diagonal Padé R_n/S_n are given for the two-point approximants about $x = 0$ and $x = \infty$, the one-point approximant P_n/Q_n about $x = 0$, and x^{-1} times the one-point approximant to $xf(x)$ at $x = \infty$. Note that for small $x(< 5)$ the two-point Padé is significantly more accurate than the one-point Padé at $x = \infty$, while it is only slightly less accurate than the one-point Padé at $x = 0$. For large x (> 50), the two-point Padé is significantly more accurate than the one-point Padé at $x = 0$, while it is only slightly less accurate than the one-point Padé at ∞. In general, the two-point approximant gives a more uniform approximation to $f(x)$.

Table 12.2 Two-point Padé approximation of f

		$x = 1$	
n	Two-point Padé	One-point Padé at 0	One-point Padé at ∞
5	0.538045407	0.538079506	–1.436
6	0.538069836	0.538079506	1.783
7	0.538078314	0.538079506	0.973
8	0.538079573	0.538079506	0.745
	Exact value of $f(1) = 0.538079506$		
		$x = 16$	
7	0.03237	0.03203	0.032336
8	0.03069	0.03239	0.032337
9	0.03240	0.03233	0.032336
10	0.03235	0.03234	0.032337
	Exact value at $f(16) = 0.032337000$		
		$x = 256$	
5	0.001956983	–0.284	0.001956962
6	0.001956964	0.242	0.001956962
7	0.001956962	–0.198	0.001956962
8	0.001956963	0.167	0.001956962
	Exact value of $f(256) = 0.001956962$		

The next example is Stirling series. It is well known that the gamma function has the Stirling series expansion

$$\Gamma(z) \sim e^{-z} z^z \left(\frac{2\pi}{z}\right)^{1/2} \left(1 + \frac{1}{12z} + \frac{1}{288z^2} + \cdots\right) \tag{12.8.7}$$

as $z \to \infty$ in $|\arg z| \le \pi - \delta < \pi$; see [160], p.294]. By transforming this series into a sequence of Padé approximants, the applicability of the Stirling series is increased. It can now be used to compute $\Gamma(z)$ to greater accuracy than can be obtained by the optimal truncation of the Stirling series. Let $P_n(x)/Q_n(x)$ denote the diagonal Padé approximants of the Stirling series in (12.8.7). Table 12.3 provides some numerical values of the function $(x/e)^x \sqrt{\frac{2\pi}{x}} \frac{P_n(1/x)}{Q_n(1/x)}$.

Table 12.3 Stirling Series

n	$x = 0.2$	$x = 0.5$
10	4.46010	1.77180
11	4.69419	1.77297
12	4.47753	1.77199
13	4.68203	1.77283
14	4.49052	1.77211
15	4.67269	1.77274
$\Gamma_{\text{opt}}(x)$	4.71183	1.76224
$\Gamma(x)$	4.59084	1.77245

Our final example here is the Bessel function

$$J_0(2x) = 1 - x^2 + \frac{x^4}{4} - \frac{x^6}{36} + \cdots, \tag{12.8.8}$$

and its Padé approximation $P_n(x)/Q_m(x)$. When $n = m = 1$, the inequality (12.8.9) written out in full gives

$$\left|(1 - x^2 + \frac{x^4}{4} - \frac{x^6}{36} + \cdots)(1 + q_1 x) - (p_0 + p_1 x)\right| \le M|x|^\rho, \tag{12.8.9}$$

where we have as usual $q_0 = 1$. After regrouping, we obtain

$$|(1 - p_0) + (q_1 - p_1)x - x^2 + \cdots| \le M|x|^\rho.$$

Taking $p_0 = 1$ and $p_1 = q_1$, the left-hand side is $O(x^2)$ for small x. Hence, the exponent ρ on the right-hand side can be at most 2, and is one less than the expected value $n + m + 1 = 3$. Furthermore, the Padé approximation in this case simply reduces to $J_0(2x) \approx 1$.

Let us write the quantity inside the absolute value sign on the left-hand side of (12.8.9) as

$$f(z)Q_m(z) - P_n(z) = \sum_{k=0}^{\infty} c_k z^k,$$

which is possible as long as $f(z)$ has a power series expansion. The inequality in (12.8.9) requires $c_0 = c_1 = \cdots = c_{\rho-1} = 0$ with ρ as big as possible. This new representation gives an indication of the error in the Padé approximation. Once the coefficients in the power series $\sum c_k z^k$ are known, we have

$$f(z) - \frac{P_n(z)}{Q_m(z)} = \frac{c_\rho z^\rho + \cdots}{Q_m(z)}. \tag{12.8.10}$$

In our case, $f(z) = J_0(2z)$ and we take $(n, m) = (2, 4)$. The last equation becomes

$$\left(1 - x^2 + \frac{x^4}{4} - \frac{x^6}{36} + \frac{x^8}{576} - \cdots\right)(1 + q_1x + q_2x^2 + q_3x^3 + q_4x^4)$$
$$-(p_0 + p_1x + p_2x^2) = c_0 + c_1x + c_2x^2 + \cdots + c_{\rho-1}x^{\rho-1} + c_\rho x^\rho + \cdots .$$

Collecting terms on the left-hand side gives

$$(1 - p_0) + (q_1 - p_1)x + (q_2 - 1 - p_2)x^2 + (q_3 - q_1)x^3$$
$$+ \left(q_4 - q_2 + \frac{1}{4}\right)x^4 + \left(\frac{q_1}{4} - q_3\right)x^5 + \left(-\frac{1}{36} + \frac{q_2}{4} - q_4\right)x^6$$
$$+ \left(-\frac{q_1}{36} + \frac{q_3}{4})x^7 + (\frac{1}{576} - \frac{q_2}{36} + \frac{q_4}{4}\right)x^8 + \cdots .$$

Taking $q_0 = p_0 = 1$, $q_1 = p_1 = q_3 = 0$, $q_2 = 8/27$, $p_2 = -19/27$, $q_4 = 5/108$ results in coefficients of $x^0, x^1, \cdots, x^7$ all vanishing, and we have

$$c_8 = \frac{1}{576} - \frac{q_2}{36} + \frac{q_4}{4} = \frac{79}{15,552}.$$

In fact, we can compute as many c_k as we wish. The Padé approximant $P_2(x)/Q_4(x)$ of the function $J_0(2x)$ is thus given by

$$J_0(2x) = \frac{1 - \frac{19}{27}x^2}{1 + \frac{8}{27}x^2 + \frac{5}{108}x^4} + \frac{\frac{79}{15,552}x^8 + \cdots}{1 + \frac{8}{27}x^2 + \frac{5}{108}x^4}.$$

As an application of this approximation, let us try to approximate the first zero of $J_0(2x)$, which is known to be ±1.202. By solving the simple equation

$$1 - \frac{19}{27}x^2 = 0,$$

we obtain $x = \pm 1.192$, whereas the first positive zero obtained from the first three terms of the power series in (12.8.8) gives 1.414.

12.9 Continued fraction expansions of e^x

At the end of Section 12.2 we established the identity

$$\sum_{k=1}^{n} \frac{1}{D_k} = \cfrac{1}{D_1 - \cfrac{D_1^2}{D_1 + D_2 - \cfrac{D_2^2}{D_2 + D_3 - \cdots \cfrac{D_{n-1}^2}{D_{n-1} + D_n}}}}.$$

It can be used to compute a continued fraction expansion for any power series. As an example, let us compute such an expansion of the exponential function:

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n\,!} = 1 + \frac{1}{x^{-1}} + \frac{1}{2x^{-2}} + \dots .$$

By Theorem 12.2.6, we have

$$e^x = \cfrac{1}{1 - \cfrac{1}{1 + x^{-1} - \cfrac{x^{-2}}{x^{-1} + 2x^{-2} - \cdots - \cfrac{(n\,!\,x^{-n})^2}{n\,x^{-n} + (n+1)\,!\,x^{-n-1} - \dots}}}}.$$

Multiplying the first denominator by x/x changes it to

$$\cfrac{x}{x + 1 - \cfrac{1/x}{1/x + 2/x^2 - \cfrac{4/x^4}{2/x^2 + 6/x^3 - \dots}}} = \cfrac{x}{x + 1 - \cfrac{x}{x + 2 - \cfrac{4/x^2}{2/x^2 + 6/x^3 - \dots}}}.$$

Continuing to simplify in this way leads to

$$e^x = \cfrac{1}{1 - \cfrac{x}{x + 1 - \cfrac{x}{x + 2 - \cfrac{2x}{x + 3 - \cfrac{3x}{x + 4 - \cdots}}}}} \tag{12.9.1}$$

This converges for each $z \in \mathbb{C}$; see Exercise 19.

Consider now the expansion of the form (12.3.3) for the exponential function. Computing the constants $\{c_k\}$ as above, it can be shown that

$$c_0 = -c_1 = 1, \quad c_{2n} = \frac{1}{4n-2}, \quad c_{2n+1} = -\frac{1}{4n+2}, \quad n = 1, 2, 3 \dots .$$

For these coefficients, (12.3.5) is

$$Q_{2m+1} = Q_{2m} - \frac{z}{4m+2} Q_{2m-1}; \tag{12.9.2}$$

$$Q_{2m} = Q_{2m-1} + \frac{z}{4m-2} Q_{2m-2}, \quad m \geq 1. \tag{12.9.3}$$

Replacing m by $m+1$ in (12.9.3) gives

$$Q_{2m+1} = Q_{2m+2} - \frac{z}{4m+2} Q_{2m}.$$

Substituting this equation into (12.9.2), we obtain

$$Q_{2m+2} - Q_{2m} = \frac{z^2}{16m^2 - 4} Q_{2m-2}, \qquad m \geq 1, \tag{12.9.4}$$

which is a second-order linear difference equation in $T_m = Q_{2m}$. We look for an asymptotic solution of the form

$$Q_{2m} \sim 1 + \sum_{k=1}^{\infty} \frac{a_k}{m^k}. \qquad m \to \infty.$$

The coefficients a_k can be determined by a recurrence relation; see Exercise 21. One can find a_1 by asymptotic matching. Since

$$Q_{2m+2} - Q_{2m} \sim -\frac{a_1}{m^2} + \cdots,$$

it follows from (12.9.4) that $a_1 = -z^2/16$. Hence, we have

$$Q_{2m} = 1 - \frac{z^2}{16m} + O\left(\frac{1}{m^2}\right).$$

For any non-zero function $C(z)$, the product $C(z)Q_{2m}$ is also a solution of the difference equation (12.9.4). Therefore

$$Q_{2m}(z) = C(z)\left[1 - \frac{z^2}{16m} + O\left(\frac{1}{m^2}\right)\right], \qquad m \to \infty. \tag{12.9.5}$$

Coupling (12.9.5) and (12.9.3) gives

$$Q_{2m-1} = C(z)\left[1 - \frac{z}{4m} - \frac{z^2}{16m} + O\left(\frac{1}{m^2}\right)\right], \qquad m \to \infty. \tag{12.9.6}$$

From (12.9.5), (12.9.6), and the identity (12.2.7), and with the aid of Stirling's approximation (2.10.5) and the duplication formula for the gamma function:

$$(n-1)! = \Gamma(n) \sim \left(\frac{n}{e}\right)^n \left(\frac{2\pi}{n}\right)^{1/2}; \qquad \Gamma(2n) = \frac{2^{2n-1}}{\sqrt{\pi}}\,\Gamma(n+\tfrac{1}{2})\,\Gamma(n),$$

we obtain

$$\frac{P_n}{Q_n} - \frac{P_{n-1}}{Q_{n-1}} \sim D(z)\,\frac{\sigma_n z^n \sqrt{n}}{2^n n!}, \qquad n \to \infty, \tag{12.9.7}$$

where $D(z)$ is a function of z (independent of n) and $\sigma_{4n} = \sigma_{4n+1} = 1$, $\sigma_{4n+2} = \sigma_{4n+3} = -1$. In comparison, if $T_n(z)$ is the nth partial sum of the Taylor series of e^z,

$$T_n - T_{n-1} = \frac{z^n}{n\,!}. \tag{12.9.8}$$

There is an extra factor of 2^{-n} in (12.9.7) relative to (12.9.8), so the Padé approximants $R_n = P_n/Q_n$ converge to their limit much faster than the Taylor series $T_n(z)$. However, we have not shown that the limit of $R_n(z)$ is indeed e^z; this problem is left as Exercise 22.

Exercises

1. Suppose $f(z) = 1/(1-z)$. (a) Prove that, for $n \geq 1$, the $[m, n]$ Padé approximant of f at 0 is f.
   (b) Find all solutions of (12.1.4) in the case $m = n = 2$.
2. Suppose f is a rational function. Prove that for m and n sufficiently large, the $[m, n]$ Padé approximant of f at 0 is f.
3. Let $f(z) \sim \sum_{n=0}^{\infty} a_n z^n$ as $z \to 0$, with $a_0 \neq 0$. Let $P_m^n(z)$ be the Padé approximant to $f(z)$. Prove that $1/P_m^n(z)$ is the Padé approximant to $1/f(z)$.
4. Verify the steps in the proof of Theorem 12.1.1.
5. Formulate a set of equations for the coefficients of a two-point Padé approximation $P_n(z)/Q_n(z)$ to a function $f(z)$ having the asymptotic expansions (12.1.12), (12.1.13). Hint: The result is analogous to (12.1.5).
6. Derive an efficient numerical method for computing two-point Padé approximants as in Exercise 5). Hint: Modify the continued fraction development of the one-point Padé approximation in Section 12.3.
7. Verify Proposition 12.2.1.
8. Prove that the function f of (12.3.1) is normal if and only if the sequence of Padé approximants $\{R_m^m, R_{m+1}^m\}$ is normal.
9. Provide a proof for Theorem 12.4.2.
10. Prove Proposition 12.5.1.
11. Prove that the moments μ_n in Theorem (12.5.3) are finite and given by $\mu_n = (-1)^n c_{n+1}$ for each $n = 0, 1, 2, \cdots$, where c_n are the coefficients in the asymptotic expansion of $f(z)$ in that theorem. Hint: Use induction.
12. Verify the characterization of Stieltjes functions, Theorem 12.6.1.
13. Verify the inequalities (12.6.2) and (12.6.3).

14. Verify (12.7.1).
15. Suppose that $\varepsilon_n/\lambda \to 0$. Estimate the rate of convergence of (12.7.2) compared to that of A_n.
16. Show that if the only singularities of $f(z)$ in the finite z-plane are l simple poles, then the remainder in the Taylor series for $f(z)$ has l transient terms.
17. Show that if $k \leq n$, the kth-order Shanks transform $S_k(A_n)$ is identical to $P_k^n(1)$, where $P_k^n(z)$ is the Padé approximant of the series $A_1 + \sum_{j=1}^{\infty}(A_{j+1} - A_j)z^j$.
18. Verify the asymptotic expansion (12.8.1).
19. Prove convergence of (12.9.1).
20. Find an asymptotic expansion for the solutions to the linear difference equation in (12.9.4).
21. Use (12.9.4) to derive a recurrence relation for the coefficients of Q_{2m}.
22. (a) Show that the Padé approximants $P_n^n(z)$ and $P_{n+1}^n(z)$ of e^z converge to e^z as $n \to \infty$. Hint: Let $F_n(z) = R_n(z)/S_n(z)$ be the nth member of the Padé sequence $P_0^0, P_1^0, P_1^1, \cdots$. Since $S_n(z)e^z - R_n(z) = O(z^n)$ as $z \to 0$ and S_n and R_n are polynomials, use Cauchy's theorem to show that

$$e^z - F_n(z) = \frac{z^n}{2\pi i S_n(z)} \int_L \frac{S_n(t)e^t}{(t-z)t^n} dt,$$

where L is any contour on which $|z| < |t|$. Then, use (12.9.6) to show that $S_n(z) \to C(z)$ as $n \to \infty$, where $C(z)$ is a finite function. Use this result to prove that $F_n(z) \to e^z$ as $n \to \infty$, provided that $C(z) \neq 0$.

(b) Show that $C(z) = e^{z/2}$ in (12.9.5).

Remarks and further reading

The standard treatise on Padé approximants is Baker and Graves-Morris [16]. For a full, up-to-date account of the computational aspects of the subject, its history, and related developments, see Brezinski and Redivo-Zaglia [31], [32]. A version of Padé approximants was developed by Hermite to prove that e is transcendental, and further adapted by Lindemann to prove that π is transcendental. See Van Assche [209] for a discussion of Padé and Hermite–Padé approximation.

The topics in this chapter have many applications. For more on orthogonal polynomials, see Khrushchev [121] and Ismail [114]. Khinchine [120] is the standard introduction to continued fractions. A more contemporary reference is Hensley [103]. Sauer [183] treats continued fractions and signal processing. Cuyt et al. [50] contain continued fraction expansions of many functions, and related information important for applications.

The classic text on the moment problem and its ramifications is Akhiezer [9]. Schmüdchen [186] contains recent developments.

Chapter 13
Riemann–Hilbert problems

In his thesis, Riemann considered the following problem: given a Jordan curve Γ in $\mathbb{C}$ that bounds a domain Ω, and given real-valued functions a, b, c on Γ, find a function $W = U + iV$ holomorphic in Ω and continuous to the boundary Γ, such that

$$aU = bV + c \tag{13.0.1}$$

on Γ. Hilbert later generalized the problem by allowing the functions a, b, c to be complex-valued.

A related problem is known as the *Riemann–Hilbert factorization problem*: given a matrix-valued function V on Γ, find M_+ holomorphic on the unbounded component of the complement of Γ and M_- holomorphic on the bounded component, such that on Γ we have

$$M_+ = V M_-. \tag{13.0.2}$$

In this chapter we focus on the classical Riemann–Hilbert problem (13.0.1), and its relation to integral transforms and integral equations. The key ingredient is the Cauchy transform

$$F(z) = \frac{1}{2\pi i} \int_\Gamma \frac{f(t)}{t-z}\, dt, \qquad z \notin \Gamma,$$

and its limits $F_\pm(t)$ as z approaches the curve Γ from one side or the other. The formulas of Sokhotski and Plemelj for these limits are proved in Section 13.1. Section 13.2 covers Carleman's approach to the Riemann–Hilbert problem. The remaining sections give some applications of the Riemann–Hilbert problem: to integral transforms in Section 13.3, and to integral equations in Sections 13.4, 13.5, 13.6, and 13.7.

This Chapter could as well have been titled "*A* Riemann–Hilbert problem." A rather different problem, which is also commonly referred to as "the Riemann–

R. Beals and R. S. C. Wong, *More Explorations in Complex Functions*, Graduate Texts in Mathematics 298, https://doi.org/10.1007/978-3-031-28288-1_13

Hilbert problem" comes up as number 21 in Hilbert's famous list of 23 problems [106]. In Section 13.8 we describe this other problem.

13.1 The Sokhotski–Plemelj formula

Consider the Cauchy integral

$$F(z) = \frac{1}{2\pi i}\int_\Gamma \frac{f(t)}{t-z}\,dt, \tag{13.1.1}$$

where Γ is a finite simple oriented C^1 curve in the complex plane, and $f : \Gamma \to \mathbb{C}$ is piecewise continuous. The curve Γ may be either a finite arc or a closed contour. Then F defined by (13.1.1) is holomorphic on the complement of Γ. Let t_0 be a point on Γ, but not an end point. We wish to determine the value of the limit of $F(z)$ as $z \to t_0$. The answer was found by Sokhotski [193] in 1873. It was rediscovered by Plemelj [168] in 1908, in his work on the Riemann–Hilbert problem.

Specifically, let $\mathbf{n} = \mathbf{n}(t_0)$ be a vector normal to Γ at t_0 and pointing to the left with respect to the direction along the curve. We start by considering $F(t_0 \pm \varepsilon\mathbf{n})$ as $\varepsilon \downarrow 0$.

Theorem 13.1.1. *Under the preceding conditions, suppose that f is Hölder continuous at t_0, i.e. there are constants $C > 0$ and $0 < \alpha < 1$ such that for $t \in \Gamma$,*

$$|f(t) - f(t_0)| \le C|t-t_0|^\alpha.$$

Then the limits $F_\pm = \lim_{\varepsilon\to 0} F(t_0 \pm \varepsilon\mathbf{n})$ exist, and

$$F_\pm(t_0) = \pm\frac{1}{2}f(t_0) + \frac{1}{2\pi i}\,p.v.\int_\Gamma \frac{f(t)}{t-t_0}\,dt. \tag{13.1.2}$$

The principal value integral here is defined by

$$p.v.\int_\Gamma \frac{f(t)}{t-t_0}\,dt = \lim_{\delta\to 0}\int_{t\in\Gamma,|t-t_0|>\delta} \frac{f(t)}{t-t_0}\,dt.$$

Proof. For convenience we translate and rotate the coordinate system so that $t_0 = 0$ and Γ is tangent to the real axis in the positive direction. The unit normal is $\mathbf{n} = i$. It follows that

$$F(t_0 \pm \varepsilon\mathbf{n}) = F(\pm\varepsilon i) = \frac{1}{2\pi i}\int_\Gamma \frac{f(t)}{t \mp i\varepsilon}\,dt.$$

A simple calculation gives

$$F(\varepsilon i) - F(-\varepsilon i) = \frac{1}{\pi}\int_\Gamma \frac{\varepsilon}{t^2+\varepsilon^2}f(t)\,dt.$$

Given $\varepsilon > 0$, let $\Gamma_\varepsilon = \Gamma \cap \{t : |t - t_0| < \varepsilon^{1/4}\}$. Then on $\Gamma \setminus \Gamma_\varepsilon$, the last integrand is dominated by $\varepsilon^{1/2}$, so we may concentrate on Γ_ε. For small ε, $t \in \Gamma_\varepsilon$ implies that $\operatorname{Im} t = o(t)$, so

$$\frac{1}{\pi}\int_{\Gamma_\varepsilon} \frac{\varepsilon f(t)\,dt}{t^2+\varepsilon^2} \sim \frac{1}{\pi}\int_{-\varepsilon^{1/4}}^{\varepsilon^{1/4}} \frac{\varepsilon f(t)\,dt}{t^2+\varepsilon^2} = \frac{1}{\pi}\int_{-\varepsilon^{-3/4}}^{\varepsilon^{-3/4}} \frac{f(\varepsilon x)}{x^2+1}\,dx$$
$$\sim \left[\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{dx}{x^2+1}\right] f(0) = f(0)$$

as $\varepsilon \to 0$. (Note that here we used only continuity at t_0, not the Hölder condition.)

Similarly

$$F(\varepsilon i) + F(-\varepsilon i) = \frac{1}{i\pi}\int_\Gamma \frac{t\,f(t)\,dt}{t^2+\varepsilon^2}.$$

Here we set $\Gamma_\delta = \{t \in \Gamma : |t| < \delta\}$. Clearly

$$\lim_{\varepsilon\to 0} \frac{1}{i\pi}\int_{\Gamma\setminus\Gamma_\delta} \frac{t\,f(t)\,dt}{t^2+\varepsilon^2} = \frac{1}{\pi i}\int_{\Gamma\setminus\Gamma_\delta} \frac{f(t)\,dt}{t}.$$

Now

$$\int_{\Gamma_\delta} \frac{t\,f(t)\,dt}{t^2+\varepsilon^2} = \int_{\Gamma_\delta} \frac{t[f(t)-f(0)]\,dt}{t^2+\varepsilon^2} + \left[\int_{\Gamma_\delta} \frac{t\,dt}{t^2+\varepsilon^2}\right] f(0).$$

Because of the Hölder continuity assumption, the first integral on the right has a limit $l(\delta)$ as $\varepsilon \to 0$, and $l(\delta) \to 0$ as $\delta \to 0$. In the second integral on the right, we note that integrand is the derivative of $\log\sqrt{t^2+\varepsilon^2}$. For small δ the endpoints of the path of integration are $\pm\delta + r_\pm$, where $r_\pm = o(\delta)$. Therefore as $\varepsilon \to 0$ the second integral is

$$\log(\delta + r_+) - \log(\delta + r_-) = \log\frac{\delta + r_+}{\delta + r_-} = \log(1 + o(\delta)) = o(\delta).$$

At this point we have proved that the difference and the sum of $F(t_0 \pm \varepsilon)$ have limits. Therefore the individual limits $F_\pm(t_0)$ exist and satisfy

$$F_+(t_0) - F_-(t_0) = f(t_0), \quad F_+(t_0) + F_-(t_0) = \frac{1}{\pi i}\,p.v.\int_\Gamma \frac{f(t)\,dt}{t - t_0}. \tag{13.1.3}$$

Solving (13.1.3) for F_+ and F_- gives (13.1.2). □

If we assume that f is uniformly Hölder continuous on the curve,

$$|f(t) - f(s)| \le C|t-s|^\alpha, \qquad \text{all } t, s \in \Gamma, \tag{13.1.4}$$

where again $C, \alpha > 0$, then the pointwise result above can be converted to a uniform result.

Theorem 13.1.2. *Under the assumptions of Theorem 13.1.1 and the additional assumption (13.1.4), if Γ is a closed curve then F is continuous from either side of Γ up to Γ itself. If Γ is an arc, then F is continuous up to Γ minus the endpoints. The formulas (13.1.3) hold at each point $t_0 \in \Gamma$ that is not an endpoint of Γ.*

Proof. Suppose that $t_0 \in \Gamma$ is not an endpoint. For some sufficiently small disk $D_{2r}(t_0)$, and for $\varepsilon < r$, the sets

$$I_\pm(\varepsilon) = \{t \pm \varepsilon \mathbf{n}(\mathbf{t}) : t \in \Gamma, \ |t - t_0| < r\}$$

are arcs that are approximately parallel to Γ at distance ε. The function F is uniformly continuous along each such arc. The previous argument shows that F converges at a uniform rate along each of the normal vectors $\pm n(t)$, so F is continuous up to $\{t \in \Gamma : |t - t_0| < r\}$. □

Let us consider the behavior of (13.1.1) at the endpoint a of a C^1 arc Γ.

Theorem 13.1.3. *Suppose that Γ is a simple C^1 arc from a to b and $f : \Gamma \to \mathbb{C}$ is piecewise continuous on Γ and continuous at the endpoints. Then*

$$F(z) \sim \begin{cases} -\dfrac{1}{2\pi i}\log(a-z)\cdot f(a), & \text{as } z \to a, \\ \dfrac{1}{2\pi i}\log(b-z)\cdot f(b), & \text{as } z \to b. \end{cases} \tag{13.1.5}$$

Proof. Consider the endpoint a. Let Γ' be an extension of Γ past b to a simple C^1 curve from a to ∞. We choose a branch of the logarithm on the complement of Γ' and extend it to Γ by taking the limit from the left. Then

$$F(z) = \frac{1}{2\pi i}\int_\Gamma \frac{f(t)}{t-z}\,dt = \frac{1}{2\pi i}\int_\Gamma \frac{dt}{t-z}\,f(a) + \frac{1}{2\pi i}\int_\Gamma \frac{f(t)-f(a)}{t-z}\,dt.$$

The first integral on the right is

$$\frac{1}{2\pi i}\log\frac{b-z}{a-z}\cdot f(a) + O(1)$$

for any choice of the branch of the logarithm. To estimate the second integral, we ease notation by assuming that the coordinates were chosen so that $a = 0$. The integral is

$$\frac{1}{2\pi i}\int_\Gamma \frac{g(t)}{t-z}\,dt, \qquad g(t) = f(t) - f(0).$$

Then g is also Hölder continuous, and $g(0) = 0$. Given $z \in \mathbb{C}$, if $|z - t| > |t|/3$ for all $t \in \Gamma$, then Hölder continuity implies that

$$\left|\int_\Gamma \frac{g(t)\,dt}{t-z}\right| < \int_\Gamma \frac{C\,|t|^\alpha}{|t|/3}|dt| = C_1,$$

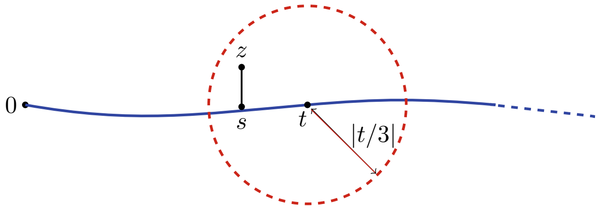

Fig. 13.1 $a = 0, |z - s| \leq |t|/3$

a constant. Otherwise there is a z such that $|z - t| \leq |t|/3$. Let $s = s(z)$ be the point of Γ that is closest to z. Then $|s| \geq |t| - |t|/3$ and

$$|z - s| \leq |z - t| \leq \frac{|t|}{3} \leq \frac{|s|}{2}; \tag{13.1.6}$$

see Figure 13.1

Suppose for now that Γ is a straight line segment, which we may take to be $[0, c] \subset \mathbb{R}$. Then $t \in \Gamma$ implies $|z - t|^2 = |t - s|^2 + |z - s|^2$, so $|z - t| \geq |t - s|$ and

$$\left|\int_\Gamma \frac{g(t)\,dt}{t - z}\right| \leq \left|\int_\Gamma \frac{|g(t) - g(s)|}{|t - s|}\,|dt|\right| + \left|\int_\Gamma \frac{g(s)\,dt}{t - z}\right|.$$

Again, Hölder continuity of g yields a bound for the first integral that is independent of s. The second integral is

$$\log(t - z)\Big|_0^c g(s) = O(s^\alpha \log(-z)) = O(|z|^\alpha \log|z|),$$

since $|z| \sim s$, by (13.1.6).

In the general case, it is enough to restrict attention to a small neighborhood of a in which Γ is sufficiently close to an interval so that we can conclude that $t \in \Gamma$ implies $|z - t| \geq \frac{1}{2}|t - s|$, where s is again the point of Γ that is closest to z. Then the previous argument goes through. Let us note explicitly that this argument allows for $|z - s| = 0$, i.e. $z \in \Gamma$, $z \neq a$.

The argument for the endpoint b is the same: simply reverse the direction of travel on Γ and extend Γ past a to select a branch of the logarithm. □

If we allow Γ to be an infinite contour, then some restriction on f needs to be made to ensure that F is defined on the complement of Γ, such as

$$\int_\Gamma \frac{|f(t)|}{1 + |t|}\,|dt| < \infty. \tag{13.1.7}$$

With such a restriction, the previous results hold. In (13.1.4) we may allow the constant C to grow as $|t| \to \infty$.

Remark. As another generalization, we can permit the curve Γ to have a corner at z_0; see Figure 13.2. If the angle is θ, then

$$F_+(t_0) = \left(1 - \frac{\theta}{2\pi}\right) f(t_0) + \frac{1}{2\pi i}\, p.v. \int_\Gamma \frac{f(t)}{t - t_0}\, dt;$$
$$F_-(t_0) = -\frac{\theta}{2\pi} f(t_0) + \frac{1}{2\pi i}\, p.v. \int_\Gamma \frac{f(t)}{t - t_0}\, dt. \tag{13.1.8}$$

See Exercise 2.

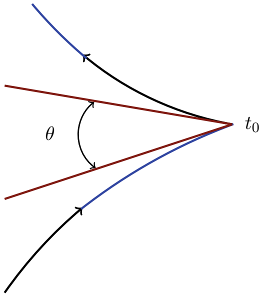

Fig. 13.2 Corner at t_0

A problem sometime encountered is to find a function G that is holomorphic on the complement of a finite curve Γ, such that the discontinuity $G_+(t) - G_-(t)$, $t \in \Gamma$, t not an endpoint, is a prescribed function f. If f is continuous on Γ, the proof of Theorem 13.1.1 shows that the Cauchy integral F is a solution. A natural question is that of uniqueness of the solution. Clearly $F(z) \to 0$ as $z \to \infty$. If G is another solution, then $G - F$ is continuous on Γ except possibly at the endpoints. If f is continuous at the endpoints, and G has at most the same kind of logarithmic growth as F at the endpoints, then $G - F$ has removable singularities at the endpoints, and is entire. More generally, if f is such that at an endpoint a of Γ,

$$F(z) \;=\; O(|z-a|^r), \qquad r > -1 \;\text{ as }\; z \to a, \;\; z \notin \Gamma, \tag{13.1.9}$$

and G is required to obey the same estimate then the singularities of $G - F$ at a are removable.

We have proved one version of the *discontinuity theorem* [43]:

Theorem 13.1.4. *Suppose that Γ is a finite simple C^1 curve and $f : \Gamma \to \mathbb{C}$ is continuous except possibly at the endpoints. Suppose that G is holomorphic on the complement of Γ, and $G_+ - G_- = f$ on G. Suppose finally that at the endpoints, if* *any, both the Cauchy integral (13.1.1) and G satisfy estimates of the form (13.1.9). Then $G - F$ is an entire function.*

Remark. If $G = O(z^n)$ as $|z| \to \infty$ for some integer $n \geq 0$, it follows that $G - F$ is a polynomial of degree $\leq n$.

Theorem 13.1.4 is one example of solving a problem by turning the Sokhotski–Plemelj formula around. The following is a different example. The problem is to evaluate the principal value integral

$$I(x) = p.v. \int_{-1}^{1} \frac{(1-t)^{\alpha-1}}{(1+t)^{\alpha}(t-x)} dt, \qquad |x| < 1, \tag{13.1.10}$$

where $0 < \alpha < 1$. Consider the function

$$G(t) = \frac{(t-1)^{\alpha-1}}{(t+1)^{\alpha}} \tag{13.1.11}$$

with the branch cut $(-1, 1)$ and the branch chosen to correspond to principal values for t real and $t > 1$. For $t_0 \in (-1, 1)$, it follows from (13.1.11) that

$$G_{+}(t_0) = \frac{(1-t_0)^{\alpha-1}}{(1+t_0)^{\alpha}} e^{i(\alpha-1)\pi}$$

and

$$G_{-}(t_0) = \frac{(1-t_0)^{\alpha-1}}{(1+t_0)^{\alpha}} e^{-i(\alpha-1)\pi}.$$

Thus, $G_{+}(t_0) - G_{-}(t_0) = (1-t_0)^{\alpha-1}(1+t_0)^{-\alpha}(-2i \sin \alpha\pi)$. In view of Theorem 13.1.4, we obtain

$$G(z) = (-2i \sin \alpha\pi) \frac{1}{2\pi i} \int_{-1}^{1} \frac{(1-t)^{\alpha-1}}{(1+t)^{\alpha}} \frac{dt}{t-z}.$$

From (13.1.3) it follows that

$$I(x) = \pi \cot \alpha\pi \, (1-x)^{\alpha-1}(1+x)^{-\alpha}. \tag{13.1.12}$$

(Note the interesting special case $\alpha = \frac{1}{2}$.)

We end this section with an extension of Theorem 13.1.3 to the case of singularities at the endpoints.

Theorem 13.1.5. *Suppose that Γ is a finite simple C^1 arc from a to b. Suppose that $f : \Gamma \to \mathbb{C}$ is continuous except possibly at the endpoints, and suppose that at the endpoint $c = a$ or $c = b$ it has the form*

$$f(t) = \frac{\widetilde{f}(t)}{(t-c)^{\sigma}}, \qquad \sigma = \alpha + i\beta \neq 0. \qquad 0 \leq \alpha < 1. \tag{13.1.13}$$

Here α and β are real, and $\widetilde{f}(t)$ is continuous. Then the Cauchy integral 13.1.1 satisfies

(a) as $z \to c$, with z not on the arc,

$$F(z) = \pm \frac{e^{\pm\sigma\pi i}}{2i \sin \sigma\pi} \frac{\widetilde{f}(c)}{(z-c)^{\sigma}} + \delta(z); \tag{13.1.14}$$

(b) as $t \to c$, with t on the arc,

$$F(t) = \pm \frac{\cot \sigma\pi}{2i} \frac{\widetilde{f}(c)}{(t-c)^{\sigma}} + \rho(t), \tag{13.1.15}$$

where the positive and negative signs correspond to $c = a$ and $c = b$, respectively. If $\alpha = Re\,\sigma = 0$, then $\sigma(z)$ and $\rho(t)$ are bounded functions with limits at c. If $\alpha > 0$, then

$$|\delta(z)| \le \frac{M_0}{|z-c|^{\alpha_0}}, \qquad |\rho(t)| \le \frac{\widetilde{\rho}(t)}{|t-c|^{\alpha_0}}, \qquad \alpha_0 < \alpha,$$

where $\widetilde{\rho}(t)$ is continuous near c. The function $(z-c)^{\sigma}$ is any branch that is single-valued near c with the branch cut taken along the arc with the value $(t-c)^{\sigma}$ on the left side of the curve.

Proof. We only present a sketch of the proof by using the Sokhotski–Plemelj formula, and refer the readers to [149] for details. Consider the case $c = a$. Take the branch cut of $(z-a)^{\sigma}$ from the endpoint a to ∞ going through b; see Figure 13.3. Select the branch that tends to $(t-a)^{\sigma}$ on the left side of the cut, i.e.

$$(t-a)^{\sigma} = (t-a)^{\sigma}_{+}. \tag{13.1.16a}$$

To find the value of $(t-a)^{\sigma}$ on the right of the cut, we follow the contour in Figure 13.3.

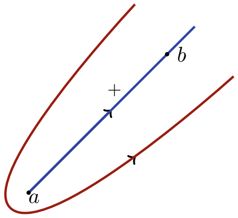

Fig. 13.3 The cut from a to ∞ through b

Thus

$$(t-a)^{\sigma}_{-} = e^{2\pi\sigma i}(t-a)^{\sigma}_{+}. \tag{13.1.16b}$$

Equations (13.1.16a) and (13.1.16b) can be written as

$$(t-a)_+^{-\sigma} - (t-a)_-^{-\sigma} = (1-e^{-2\pi\sigma i})(t-a)^{-\sigma}$$

or equivalently

$$\frac{e^{i\pi\sigma}}{2i\sin\pi\sigma}(t-a)_+^{-\sigma} - \frac{e^{i\pi\sigma}}{2i\sin\pi\sigma}(t-a)_-^{-\sigma} = (t-a)^{-\sigma}. \qquad (13.1.17)$$

Equation (13.1.17) shows that the function $(t-a)^{-\sigma}$ can be written as a difference function of a "+" and a "−" function. Since it is expected that the major contribution to the Cauchy integral (13.1.1) will come from the locations where $f(t)$ is singular (i.e. the endpoints), it follows from (13.1.13) that as $z \to a$,

$$F(z) \sim \frac{\widetilde{f}(a)}{2\pi i}\int_a^b \frac{(t-a)^{-\sigma}}{t-z}\,dt.$$

On account of (13.1.17), we obtain

$$F(z) \sim \widetilde{f}(a)\frac{e^{i\pi\sigma}}{2i\sin\sigma\pi}\left[\frac{1}{2\pi i}\int_a^b \frac{(t-a)_+^{-\sigma}}{t-z}\,dt - \frac{1}{2\pi i}\int_a^b \frac{(t-a)_-^{-\sigma}}{t-z}\,dt\right].$$

From (13.1.13), it follows that

$$F(z) \sim \frac{e^{i\pi\sigma}}{2i\sin\sigma\pi}[F_+(z) - F_-(z)].$$

By the Sokhotski–Plemelj formulas we have, for z not on the curve of integration,

$$F(z) \sim \frac{e^{i\pi\sigma}}{2i\sin\sigma\pi}\frac{\widetilde{f}(a)}{(z-a)^{\sigma}}.$$

In view of (13.1.16a) and (13.1.16b), for $z = t$ on the path of integration we have

$$\begin{aligned} F(t) &= \frac{1}{2}[F_+(t) + F_-(t)] \\ &\sim \frac{e^{i\pi\sigma}}{2i\sin\sigma\pi}\frac{\widetilde{f}(a)}{2}[(t-a)_+^{-\sigma} + (t-a)_-^{-\sigma}] \\ &= \frac{\cot\sigma\pi}{2i}\frac{\widetilde{f}(a)}{(t-a)^{\sigma}}. \qquad \square \end{aligned}$$

13.2 Riemann–Hilbert Problems

As we noted in the introduction, the problem originally posed by Riemann was to find a function $W = U + iV$, holomorphic inside a bounded domain Ω and continuous to the boundary, that satisfies a linear relation between the boundary values of its real and imaginary parts. Up to conformal equivalence we may take $\Omega = \mathbb{D}$ and look for

$$a(\zeta)U(\zeta) + b(\zeta)V(\zeta) = c(\zeta), \qquad |\zeta| = 1, \tag{13.2.1}$$

where a, b, and c are given real-valued functions. If we set

$$W_-(z) = \overline{W}_+\left(\frac{1}{\overline{z}}\right), \qquad |z| > 1, \tag{13.2.2}$$

then $\overline{W}_- = W_+$ on $\Gamma = \partial\mathbb{D}$. Therefore we may rewrite (13.2.1) as

$$\frac{a(\zeta) - ib(\zeta)}{2} W_+(\zeta) + \frac{a(\zeta) + ib(\zeta)}{2} W_-(\zeta) = c(\zeta). \tag{13.2.3}$$

Thus, we can reformulate Riemann's problem in the form: find two functions $W_+(z)$ and $W_-(z)$, holomorphic inside and outside of the unit circle, respectively, such that their boundary values on the unit circle satisfy the linear relation (13.2.3). With this formulation, W is unique only up to multiplication by an entire function, so we also specify the behavior of $W_-(z)$ at ∞; for instance, from (13.2.2), we require $W_-(z) \to \overline{W}_+(0)$ as $z \to \infty$].

As a generalization, Hilbert posed the problem of finding a function $W(z)$, holomorphic on the complement of a closed curve Γ such that for all $\zeta \in \Gamma$,

$$W_+(\zeta) = g(\zeta)W_-(\zeta) + f(\zeta), \tag{13.2.4}$$

where $g(\zeta)$ and $f(\zeta)$ are two given complex-valued functions. In Hilbert's original problem, Γ is a closed curve, the general problem (13.2.4), whether Γ is open or closed, has become known as the *Riemann–Hilbert problem*. Again, for uniqueness, the behavior of $W(z)$ at ∞ is required. If Γ is an open arc, then the endpoint behavior should also be prescribed.

In his work on singular integral equations (see Section 13.7), Carleman [41] found an effective method of attack. First find a function $L(z)$ that satisfies

$$L_+(\zeta) = g(\zeta)\, L_-(\zeta), \tag{13.2.5}$$

where $L_+(\zeta)$, $L_-(\zeta)$, and $L(\zeta)$ have no zeros. Substituting (13.2.5) into (13.2.4) yields

$$\frac{W_+(\zeta)}{L_+(\zeta)} - \frac{W_-(\zeta)}{L_-(\zeta)} = \frac{f(\zeta)}{L_+(\zeta)}. \tag{13.2.6}$$

Note that the function $W(z)/L(z)$ is holomorphic for z not on Γ, since $L(z) \neq 0$. Hence the conditions in Theorem 13.1.4 are met, and the general function satisfying

(13.2.6) is given; see the remark following Theorem 13.1.4. Hence, if $L(z)$ is known then $W(z)$ has been found.

Before proceeding to find $L(z)$, we observe that solving equation (13.2.6) is equivalent to solving

$$\frac{f(\zeta)}{L_+(\zeta)} = F_+(\zeta) - F_-(\zeta)$$

and then defining

$$F(z) = \frac{1}{2\pi i}\int_\Gamma \frac{f(\zeta)/L_+(\zeta)}{\zeta - z}d\zeta; \tag{13.2.7}$$

see (13.1.2.

Equation (13.2.6) can be written as

$$\frac{W_+(\zeta)}{L_+(\zeta)} - F_+(\zeta) = \frac{W_-(\zeta)}{L_-(\zeta)} - F_-(\zeta).$$

The function

$$\frac{W(z)}{L(z)} - F(z) \tag{13.2.8}$$

has the same boundary values on each side of Γ, so it is an entire function. The function $W(z)$ is thus determined, up to addition of an entire function. In the case when Γ is an infinite straight line parallel to the real axis, this method is known as the Wiener–Hopf technique; see [22].

We now return to the problem of finding a function $L(z)$ that satisfies (13.2.5). Assuming that $g(\zeta) \neq 0$ for $\zeta \in \Gamma$, we take logarithms on both sides of (13.2.5). This gives

$$\log L_+(\zeta) - \log L_-(\zeta) = \log g(\zeta). \tag{13.2.9}$$

For now we assume that Γ is an arc, and that $g(\zeta)$ is continuous to the end points a, b of the arc. By the discontinuity theorem, Theorem 13.1.4, a particular solution of (13.2.9) is

$$G(z) = \log L(z) = \frac{1}{2\pi i}\int_\Gamma \frac{\log g(\zeta)}{\zeta - z}d\zeta. \tag{13.2.10}$$

Thus, $L(z) = e^{G(z)}$ and $L(z)$ is non-zero. Furthermore,

$$\frac{L_+(\zeta)}{L_-(\zeta)} = e^{[G_+(\zeta)-G_-(\zeta)]} = e^{\log g(\zeta)} = g(\zeta),$$

i.e. (13.2.5) is satisfied. Here, the second equality again follows from Theorem 13.1.4. From (13.2.10), we have $L(z) \to 1$ as $z \to \infty$. The behavior of $L(z)$ as $z \to a$ or b may not be appropriate for the application of Theorem 13.1.4. Fortunately, we can adjust the behavior of $L(z)$ by incorporating an integral power of $z - a$ or $z - b$ into $L(z)$. For instance, we know from Theorem 13.1.5 that

$$G(z) \sim -\frac{1}{2\pi i} \log g(a) \log (z-a)$$

as $z \to a$; see (13.1.5). Hence

$$\begin{aligned} L(z) &\sim (z-a)^{-\log g(a)/2\pi i} \\ &\sim (z-a)^{\alpha+i\beta}, \end{aligned}$$

where α and β are real numbers. In this case, we can revise $L(z)$ by multiplying it by a factor $(z-a)^{p_1}$, where p_1 is an integer, with $-1 < \alpha + p_1 < 0$. A similar factor can be incorporated to yield the desired behavior at the other endpoint.

If Γ is a closed curve, equation (13.2.9) is usually not useful, since $\log g(z)$ will not in general return to its initial value after a complete circuit. Thus the function $\log g(\zeta)$ in the integral defining $G(z)$ in (13.2.10) has a discontinuity, and the Sokhotski–Plemelj formulas are not valid. Let $\log g(\zeta)$ increase by $2\pi ni$, n an integer, during a circuit of Γ. We can avoid this difficulty by defining

$$g_0(\zeta) = (\zeta - z_0)^{-n} g(\zeta),$$

where z_0 is a point inside Γ. Now, define

$$N(z) = \begin{cases} L(z) & \text{for } z \text{ inside } G \\ (z-z_0)^n L(z) & \text{for } z \text{ outside.} \end{cases} \tag{13.2.11}$$

Our problem is now to solve

$$N_+(\zeta) = g_0(\zeta) N_-(\zeta),$$

where $g_0(\zeta)$ is single-valued, and the procedure for the arc can be used.

Example. Find a function $W(z)$ satisfying

$$W_+(\zeta) + W_-(\zeta) = f(\zeta) \tag{13.2.12}$$

for ζ on an arc Γ, with $W(z)$ being of finite degree at ∞ and having singularities near endpoints a and b which are no worse than algebraic with degree > -1. The function $f(\zeta)$ may have integrable singularities at the endpoints a and b.

From (13.2.4) with $g(\zeta) = -1$, we obtain one solution, namely,

$$\log L(z) = \frac{1}{2\pi i} \int_a^b \frac{i\pi}{\zeta - z} d\zeta,$$

i.e.

$$L(z) = \sqrt{\frac{z-b}{z-a}}. \tag{13.2.13}$$

It is easily shown that equation (13.2.5) is satisfied. For $W(z)/L(z)$ not to grow too fast as $z \to a$ or b, we need to make $L(z)$ grow algebraically (with exponent > -1) as $z \to a, b$. Therefore we use for $L(z)$ a function obtained by multiplying the right-hand side of (13.2.13) by $1/(z - b)$. That is, we choose

$$L(z) = \frac{1}{\sqrt{(z-a)(z-b)}} \tag{13.2.14}$$

and the branch cut along the arc, with $L(z) \sim z^{-1}$ as $z \to \infty$. For $\zeta \in \Gamma$, $L_+(\zeta)$ and $L_-(\zeta)$ can easily be calculated from (13.2.14). For instance, if Γ is the line segment $(-1, 1)$ of the real line, then

$$L_+(\zeta) = \frac{-i}{\sqrt{1-\zeta^2}} \quad \text{and} \quad L_-(\zeta) = \frac{i}{\sqrt{1-\zeta^2}}. \tag{13.2.15}$$

Equation (13.2.6) now gives

$$\frac{W(z)}{L(z)} = \frac{1}{2\pi i} \int_\Gamma \frac{f(\zeta)}{L_+(\zeta)(\zeta - z)} d\zeta + p_n(z), \tag{13.2.16}$$

where $p_n(z)$ is a polynomial of degree $< n$. The function given in (13.2.16) is the most general solution for which $W(z)/z^n \to 0$ as $z \to \infty$.

13.3 The Radon Transform and the Fourier transform

The Radon transform is defined by

$$Q(k, p) = \int_L q(x_1, x_2)\, d\tau,$$

where the integral is taken along a line L with direction determined by the unit vector $\mathbf{k} = \left(\frac{1}{\sqrt{1+k^2}}, \frac{k}{\sqrt{1+k^2}}\right)$, at a distance p from the origin, and τ is a parameter on this line; see Figure 13.4. This transform plays a fundamental role in the mathematical formulation of computerized tomography (CT): the reconstruction of a function from the knowledge of its line integrals, irrespective of the particular field of application. The most prominent application of CT is in diagnostic radiology. Here a cross section of the human body is scanned by a thin X-ray beam whose intensity loss is recorded by a detector and processed by a computer to produce a two-dimensional image that in turn is displayed on a screen.

A simple physical model is as follows; see Figure 13.5. Let $f(x_1, x_2)$ be the X-ray attenuation coefficient of the tissue at the point $\mathbf{x} = (x_1, x_2)$. This means that X-ray traversing a small distance $\Delta\tau$ along the line L suffers the relative intensity loss

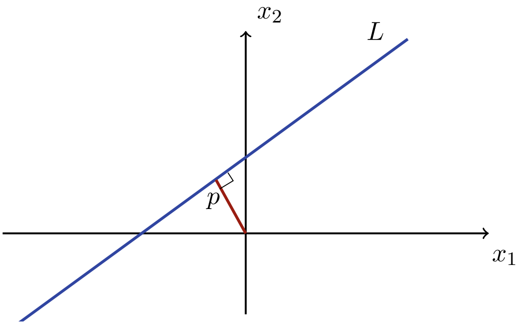

Fig. 13.4 Line L, distance p

$$\frac{\Delta I}{I} = -f(x_1, x_2)\Delta\tau.$$

Let I_0 and I_1 be the initial and final intensities of the beam, before and after leaving the body, respectively. In the limit $\Delta\tau \to 0$, it follows from the above equation that

$$\frac{I_1}{I_0} = e^{-\int_L f(x_1,x_2)\,d\tau},$$

that is, the scanning process determines an integral of the function $f(x_1, x_2)$ along each line L. Given all these integrals, one wishes to reconstruct the function f.

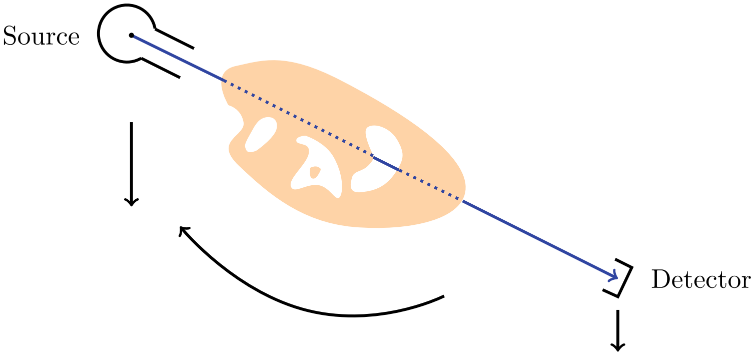

Fig. 13.5 Simple physical model of CT

Let $\mathbf{k} = (1/\sqrt{1+k^2}, k/\sqrt{1+k^2})$ be a unit vector along L and let $\mathbf{k}^\perp$ be the unit vector orthogonal to $\mathbf{k}$, that is, $\mathbf{k}^\perp = (-k/\sqrt{1+k^2}, 1/\sqrt{1+k^2})$. Then any point $\mathbf{x} = (x_1, x_2)$ can be written as $\mathbf{x} = p\mathbf{k}^\perp + \tau\mathbf{k}$. For fixed k and p, we write

$$x_1(\tau) = \frac{\tau - pk}{\sqrt{1+k^2}}; \qquad x_2(\tau) = \frac{\tau k + p}{\sqrt{1+k^2}}.$$

Note that as τ varies, $\mathbf{x} = (x_1(\tau), x_2(\tau))$ moves along the line as depicted in Figure 13.5. Therefore the Radon transform can also be written as

$$\widetilde{q}(k, p) = \int_{-\infty}^{\infty} q\left(\frac{\tau - pk}{\sqrt{1+k^2}}, \frac{\tau k + p}{\sqrt{1+k^2}}\right) d\tau. \tag{13.3.1}$$

Along the line of integration in (13.3.1), the derivative of a function $\mu(\tau)$ is

$$\frac{d\mu}{d\tau} = \frac{1}{\sqrt{1+k^2}}\left[\frac{\partial\mu}{\partial x_1} + k\frac{\partial\mu}{\partial x_2}\right],$$

so in order to calculate $\widetilde{q}$ in (13.3.1), we are led to the partial differential equation

$$\frac{\partial\mu}{\partial x_1} + k\frac{\partial\mu}{\partial x_2} = q(x_1, x_2). \tag{13.3.2}$$

Then

$$\widetilde{q}(k, p) = \sqrt{1+k^2}\cdot \mu(x_1(t), x_2(\tau))\Big|_{-\infty}^{\infty}. \tag{13.3.3}$$

As we shall see, equation (13.3.2) leads naturally to a Riemann–Hilbert problem. Let us begin with a simpler model, the differential equation

$$\frac{d\mu}{dx}(x) - ik\mu = \sqrt{1+k^2}\, q(x), \qquad -\infty < x < \infty, \;\; k \in \mathbb{C}. \tag{13.3.4}$$

Assuming that q and q_x belong to L^1, we have the following particular solutions:

$$\begin{aligned} \mu^+(x, k) &= \int_{-\infty}^{x} q(\xi)e^{ik(x-\xi)}d\xi, \\ \mu^-(x, k) &= -\int_{x}^{\infty} q(\xi)e^{ik(x-\xi)}d\xi. \end{aligned} \tag{13.3.5}$$

We define a solution of (13.3.4) by

$$\mu(x, k) = \begin{cases} \mu^+(x, k), & k_I \geq 0, \\ \mu^-(x, k), & k_I \leq 0, \end{cases} \qquad k = k_R + ik_I. \tag{13.3.6}$$

Taking into account (13.3.5), it is readily seen that μ^+ is holomorphic in the upper half-plane ($k_I > 0$) and μ^- is holomorphic in the lower half-plane ($k_I < 0$). Furthermore, the large x behavior of both μ^+ and μ^- is uniquely determined by $\widehat{q}(k)$, which is defined by

$$\widehat{q}(k) = \int_{-\infty}^{\infty} q(x)e^{-ikx}dx, \qquad k \in \mathbb{R}. \tag{13.3.7}$$

Indeed,

$$\lim_{x\to-\infty}\left(e^{-ikx}\mu^{-}\right) = -\widehat{q}(k), \qquad \lim_{x\to\infty}\left(e^{-ikx}\mu^{+}\right) = \widehat{q}(k). \tag{13.3.8}$$

Equation (13.3.7) defines $\widehat{q}$ in terms of q. To invert this relationship we will formulate the problem as a Riemann–Hilbert problem. Taking the difference of the two equations in (13.3.5), we have

$$\mu^{+}(x,k) - \mu^{-}(x,k) = e^{-ikx}\widehat{q}(k), \qquad k \in \mathbb{R}. \tag{13.3.9}$$

By integrating by parts, it can be seen from (13.3.5) that

$$\mu = O\left(\frac{1}{k}\right) \qquad \text{as } k \to \infty.$$

Then equation (13.3.9), with $\mu \to 0$ as $k \to \infty$, defines a Riemann–Hilbert problem for the function $\mu(x,k)$; see (13.2.4) and the following remark. The solution is given by

$$\mu(x,k) = \frac{1}{2\pi i}\int_{-\infty}^{\infty}\frac{e^{ixl}\widehat{q}(l)}{l-k}dl, \qquad k \in \mathbb{C}. \tag{13.3.10}$$

Given $\widehat{q}(l)$, equation (13.3.10) yields $\mu(x,k)$, which then gives $q(x)$ through equation (13.3.4). An elegant formula for q can be obtained by comparing the large k asymptotics of equations (13.3.4) and (13.3.10). Equation (13.3.4) implies $q = -i\lim_{k\to\infty}(k\mu)$, while (13.3.10) yields

$$\lim_{k\to\infty} k\mu = -\frac{1}{2\pi i}\int_{-\infty}^{\infty} e^{ixl}\widehat{q}(l)dl.$$

Hence

$$q(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{ixk}\widehat{q}(k)\,dk. \tag{13.3.11}$$

Equations (13.3.7) and (13.3.11) are the usual formulas for the direct and inverse *Fourier transform*.

Let us now turn this argument around. To solve (13.3.4), write the proposed solution μ and the right-hand side in terms of their Fourier transforms:

$$\mu(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{ilx}\widehat{\mu}(l)\,dl. \tag{13.3.12}$$

Then the differential equation (13.3.4) becomes

$$\frac{d\mu}{dx}(x) - ik\mu(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} i(l-k)e^{ilx}\widehat{\mu}(l)\,dl = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{ilx}q(l)\,dl.$$

Thus we expect $\widehat{\mu}(l) = \widehat{q}/(il - ik)$. With this choice, the inversion formula gives (13.3.10).

Let us take a second look at this, writing the solution in terms of a Green's function G for the operator $d/dx - ik$, i.e. we want to obtain the solution μ as an integral

$$\mu(x) = \int_{-\infty}^{\infty} G(x-y)\, q(y)\, dy.$$

(The form $G(x, y) = G(x - y)$ reflects the fact that the operator is invariant under translation.) A simple computation shows that taking the Fourier transform gives

$$\widehat{\mu} = \widehat{G} \cdot \widehat{q}.$$

In view of this and (13.3.12), we want $\widehat{G}(l) = 1/i(l - k)$, so

$$G(x,k) = \frac{1}{2\pi i} \int_{-\infty}^{\infty} e^{ixl} \frac{1}{l-k}\, dl, \qquad k \notin \mathbb{R}.$$

The limits as $\pm\text{Im}\, k \downarrow 0$ can be computed (see Exercise 3), and we recover (13.3.5).

Making use of the analogy with (13.3.4), let us return to equation (13.3.2), with k allowed to be complex:

$$\frac{\partial \mu}{\partial x_1} + k\frac{\partial \mu}{\partial x_2} = q, \qquad -\infty < x_1, x_2 < \infty, \quad k \in \mathbb{C}. \tag{13.3.13}$$

As in the case of (13.3.4), we make use of the Fourier transform, this time in two dimensions. Treating one variable at a time, it is easy to see that under appropriate assumptions on $q(x) = q(x_1, x_2)$ we have the relation between q and its Fourier transform $\widehat{q}$:

$$\widehat{q}(l) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-i(l_1x_1+l_2x_2)}\, q(x)\, dx_1\, dx_2;$$
$$q(x) = \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{i(l_1x_1+l_2x_2)}\, \widehat{q}(l)\, dl_1\, dl_2.$$

In analogy with the argument given above with respect to (13.3.4), we look for a Green's function G for the equation (13.3.13):

$$\mu(x) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(x-y) q(y) \mu(y)\, dy_1\, dy_2, \tag{13.3.14}$$

and derive the equation

$$G(x_1, x_2, k) = \frac{1}{i(2\pi)^2} \int_{\mathbb{R}^2} \frac{e^{i(x_1\xi_1 + x_2\xi_2)}}{\xi_1 + k\xi_2} d\xi_1 d\xi_2. \tag{13.3.15}$$

The above integral can be evaluated by using contour integration (Exercise 3), and we have

$$G(x_1, x_2, k) = \frac{\operatorname{sgn}(\operatorname{Im} k)}{2\pi i(x_1 - kx_2)}, \qquad \operatorname{Im} k \neq 0. \tag{13.3.16}$$

Putting (13.3.16) into (13.3.14) we obtain

$$\mu_\pm(x_1, x_2, k) = \pm\frac{1}{2\pi i}\int_{\mathbb{R}^2} \frac{q(y_1, y_2)}{[(x_2 - y_2) - k(x_1 - y_1)]}\, dy_1\, dy_2, \quad k \in \mathbb{C}^\pm, \operatorname{Im} k \neq 0.$$

Applying the Sokhotski–Plemelj formulas, we obtain

$$\begin{aligned}\mu_\pm(x_1, x_2, k) = &\pm \frac{1}{2\pi i}\int_{-\infty}^{\infty}\left(p.v.\int_{-\infty}^{\infty} \frac{q(y_1, y_2)}{(x_2 - y_2) - k(x_1 - y_1)}dy_2\right) dy_1\\ &+ \frac{1}{2}\left(\int_{-\infty}^{x_1} - \int_{x_1}^{\infty}\right) q\left(y_1, x_2 - k(x_1 - y_1)\right) dy_1, \qquad k \in \mathbb{R};\end{aligned}$$

see Exercise 4. The difference of these two equations gives

$$(\mu_+ - \mu_-)(x_1, x_2, k) = \frac{1}{\pi i}\int_{-\infty}^{\infty} p.v.\int_{-\infty}^{\infty} \frac{q(y_1, y_2)}{(x_2 - y_2) - k(x_1 - y_1)}\, dy_2\, dy_1, \quad k \in \mathbb{R}. \tag{13.3.17}$$

The right-hand side of this equation can be written in terms of the Radon transform of the function $q(x_1, x_2)$ defined by

$$\widetilde{q}(k, p) = \int_{-\infty}^{\infty} q\left(\frac{\tau - pk}{\sqrt{1 + k^2}}, \frac{\tau k + p}{\sqrt{1 + k^2}}\right) d\tau. \tag{13.3.18}$$

Indeed, changing variables from (y_1, y_2) to (p', τ') where

$$y_1 = \frac{\tau' - p'k}{\sqrt{1 + k^2}}, \qquad y_2 = \frac{\tau' k + p'}{\sqrt{1 + k^2}}$$

and using equation (13.3.17) and the Jacobian of the transformation

$$J = \left|\det\begin{pmatrix} \frac{\partial y_1}{\partial \tau'} & \frac{\partial y_2}{\partial \tau'} \\ \frac{\partial y_1}{\partial p'} & \frac{\partial y_2}{\partial p'} \end{pmatrix}\right| = 1,$$

it follows that

$$\begin{aligned}&\mu_+(x_1, x_2, k) - \mu_-(x_1, x_2, k)\\ &\quad = \frac{1}{i\pi} p.v.\int_{-\infty}^{\infty} \frac{\widetilde{q}(k, p')}{x_2 - kx_1 - p'\sqrt{1 + k^2}}dp', \quad k \in \mathbb{R}.\end{aligned} \tag{13.3.19}$$

Equation (13.3.14) implies that

$$\mu = O\left(\frac{1}{k}\right), \qquad k \to \infty, \tag{13.3.20}$$

so equations (13.3.19) and (13.3.20) define a Riemann–Hilbert problem for the function $\mu(x_1, x_2, k)$. Its unique solution, for $\operatorname{Im} k \neq 0$, is

$$\mu(x_1, x_2, k) = \frac{1}{2\pi i} \int_{-\infty}^{\infty} \left(\frac{1}{\pi i} p.v. \int_{-\infty}^{\infty} \frac{\widetilde{q}(k, p')dp'}{x_2 - kx_1 - p'\sqrt{1+k^2}} \right) \frac{dk'}{k' - k};$$

see (13.3.19). Comparing the large-k asymptotics of equations (13.3.13), (13.3.14) and (13.3.20), it follows that

$$q = \lim_{k \to \infty} \frac{\partial}{\partial x_2}(k\mu)$$

or

$$q(x_1, x_2) = \frac{1}{2\pi^2} \frac{\partial}{\partial x_2} \int_{-\infty}^{\infty} \left(p.v. \int_{-\infty}^{\infty} \frac{\widetilde{q}(k, p)}{x_2 - kx_1 - p\sqrt{1+k^2}} dp \right) dk. \tag{13.3.21}$$

Equations (13.3.18) and (13.3.21) are the usual formulas for the direct and inverse Radon transform.

13.4 Integral Equations with Cauchy Kernels

A typical integral equation in one variable has the form

$$m(x)u(x) = \lambda \int_{-\infty}^{\infty} K(x, y)u(y)\, dy,$$

where m is a given function, λ is a complex parameter, and various assumptions are made about the *kernel* K, such as

$$K(x, y) = K(x - y), \quad K(y, x) = -K(x, y), \quad \text{or} \quad K(x, y) = 0 \text{ if } y > x.$$

In this and subsequent sections we examine some cases where the problem can be treated by Riemann–Hilbert methods.

In this section we consider the case

$$m(x)u(x) = \lambda\, p.v. \int_{-1}^{1} \frac{u(\tau)}{\tau - x} d\tau + k(x), \qquad |x| < 1, \tag{13.4.1}$$

where λ is real and positive, and where $m(x)$, $k(x)$ are given real-valued functions. Define

$$U(z) = \frac{1}{2\pi i}\int_{-1}^{1}\frac{u(\tau)}{\tau - z}\,d\tau. \tag{13.4.2}$$

(Here we allow $U(z)$ to have an algebraic singularity of degree > -1 at the endpoints -1 and 1.) From the Sokhotski–Plemelj formulas (13.1.2), we have

$$[m(x) - \lambda\pi i]U_+(x) = [m(x) + \lambda\pi i]U_-(x) + k(x), \tag{13.4.3}$$

which is of the form discussed in Section 13.2; see (13.2.3).

First we look for a non-zero function $L(z)$ such that

$$\frac{L_+(x)}{L_-(x)} = \frac{m(x) + \lambda\pi i}{m(x) - \lambda\pi i}.$$

A suitable choice is given by

$$L(z) = \frac{1}{z-1}e^{G(z)}, \tag{13.4.4}$$

where

$$G(z) = \frac{1}{2\pi i}\int_{-1}^{1}\frac{1}{\tau - z}\log\frac{m(\tau) + \lambda\pi i}{m(\tau) - \lambda\pi i}\,d\tau; \tag{13.4.5}$$

see (13.2.10). Note that

$$\frac{1}{2\pi i}\log\frac{m(\tau) + \lambda\pi i}{m(\tau) - \lambda\pi i}$$

is purely real, and we take it to lie in the range (0, 1). The factor $(z-1)^{-1}$ in (13.4.4) has been inserted to make sure that $L(z)$ grows algebraically, with index between -1 and 0, as z tends to either endpoint -1 or 1.

For x in $(-1, 1]$, equation (13.4.3) gives

$$\frac{U_+(x)}{L_+(x)} - \frac{U_-(x)}{L_-(x)} = \frac{k(x)}{L_+(x)[m(x) - \lambda\pi i]}, \tag{13.4.6}$$

where

$$L_+(x) = \frac{1}{x-1}\sqrt{\frac{m(x) + \lambda\pi i}{m(x) - \lambda\pi i}}\,e^{w(x)}, \tag{13.4.7a}$$

$$L_-(x) = \frac{1}{x-1}\sqrt{\frac{m(x) - \lambda\pi i}{m(x) + \lambda\pi i}}\,e^{w(x)}, \tag{13.4.7b}$$

and

$$w(x) = \frac{1}{2\pi i}\,p.v.\int_{-1}^{1}\frac{1}{\tau - x}\log\frac{m(\tau) + \lambda\pi i}{m(\tau) - \lambda\pi i}\,d\tau. \tag{13.4.8}$$

(To derive these formulas, first write (13.4.3) in the form of (13.2.4), and then follow the steps leading to equation (13.2.4)–(13.2.8).) On account of the behavior of $U(z)$ and $L(z)$ as $z \to \infty$, and by Theorem 13.1.4, the most general solution of (13.4.6) is

$$u(x) = \frac{m(x)k(x)}{m^2(x)+\lambda^2\pi^2} + \frac{\lambda e^{w(x)}}{\sqrt{m^2(x)+\lambda^2\pi^2}}\, p.v. \int_{-1}^{1} \frac{k(\tau)e^{-w(\tau)}}{(\tau - x)\sqrt{m^2(\tau)+\lambda^2\pi^2}}\, d\tau + \frac{C e^{w(x)}}{(1-x)\sqrt{m^2(x)+\lambda^2\pi^2}}, \qquad (13.4.9)$$

where C is constant and $w(x)$ is given in (13.4.8); see Exercise 4. The singularity at $x = 1$ in the last term of (13.4.9) is offset by the factor $e^{w(x)}$, so that the last term is integrable. In the case when $k(x) = 0$ in (13.4.1), the resulting homogeneous equation has a solution for all λ, i.e. the *spectrum* is continuous.

In the case when $m(x) = 0$ and $k(x) = -\lambda l(x)$, equation (13.4.1) reduces to

$$p.v. \int_{-1}^{1} \frac{u(\tau)}{\tau - x}\, d\tau = l(x), \qquad (13.4.10)$$

and its solution is given by

$$u(x) = -\frac{1}{\pi^2}\sqrt{\frac{1-x}{1+x}}\, p.v. \int_{-1}^{1} \frac{l(\tau)\sqrt{1+\tau}}{\sqrt{1-\tau}(\tau - x)}\, d\tau + \frac{C}{\sqrt{1-x^2}}; \qquad (13.4.11)$$

see Exercise 6. If $l(x) = 1$ in (13.4.10), then the solution further simplifies to

$$u(x) = -\frac{1}{\pi}\sqrt{\frac{1-x}{1+x}} + \frac{C_1}{\sqrt{1-x^2}},$$

where C_1 is a new constant.

13.5 Integral Equations with Algebraic Kernels

Consider the Abel-type integral equation

$$\int_{0}^{1} \frac{u(\tau)}{|\tau - x|^\alpha}\, d\tau = k(x) \qquad (13.5.1)$$

for $x \in (0, 1)$, where $0 < \alpha < 1$. This equation is not of Cauchy type but Carleman [41] showed that it is still useful to introduce a function

$$U(z) = \int_{0}^{1} \frac{u(\tau)}{(z - \tau)^\alpha}\, d\tau, \qquad (13.5.2)$$

analogous to that used for the Cauchy type in Section 13.4; cf. (13.4.1)-(13.4.2). This function is defined for all $z \notin (-\infty, 1)$; for z real and $z > 1$, we use principal values in (13.5.2). For $x \in (0, 1)$, it is easily seen that

$$U_+(x) = \int_0^x \frac{u(\tau)}{(x-\tau)^\alpha}\,d\tau + e^{-i\alpha\pi} \int_x^1 \frac{u(\tau)}{(\tau-x)^\alpha}\,d\tau, \tag{13.5.3}$$

$$U_-(x) = \int_0^x \frac{u(\tau)}{(x-\tau)^\alpha}\,d\tau + e^{i\alpha\pi} \int_x^1 \frac{u(\tau)}{(\tau-x)^\alpha}\,d\tau. \tag{13.5.4}$$

Here, as before, $U_+(x)$ and $U_-(x)$ denote the limits of $U(z)$ as $z \to x$ from above or below, respectively. Equations (13.5.3) and (13.5.4) may be viewed as the appropriate replacement for the Sokhotski–Plemelj formulas for Cauchy integrals. Since

$$e^{i\alpha\pi} U_+(x) - e^{-i\alpha\pi} U_-(x) = 2i \sin\alpha\pi \int_0^x \frac{u(\tau)}{(x-\tau)^\alpha}\,d\tau, \tag{13.5.5}$$

the function $u(x)$ can be determined from the knowledge of $U_+(x)$ and $U_-(x)$, by using the solution of a conventional Abel equation; see [207].

Solving (13.5.3) and (13.5.4), we obtain

$$\int_0^x \frac{u(\tau)}{(x-\tau)^\alpha}\,d\tau, \qquad \int_x^1 \frac{u(\tau)}{(\tau-x)^\alpha}\,d\tau$$

in terms of $U_+(x)$ and $U_-(x)$, and use (13.5.1) to obtain

$$U_+(x) = -e^{-i\alpha\pi} U_-(x) + (1 + e^{-i\alpha\pi})k(x) \tag{13.5.6}$$

for $x \in (0, 1)$. For $x \in (-\infty, 0)$, equation (13.5.2) gives

$$U_+(x) = e^{-2i\alpha\pi} U_-(x). \tag{13.5.7}$$

This is again a Riemann–Hilbert problem, but it involves two arcs $(-\infty, 0)$ and $(0, 1)$, and the above-mentioned method no longer works. Fortunately, the coefficients in (13.5.6) and (13.5.7) are constants. Trying a factor of the form $z^\nu(z-1)^\mu$, we find that the new function

$$V(z) = z^{(\alpha-1)/2}(z-1)^{(\alpha-1)/2} U(z)$$

reduces (13.5.7) to

$$V_+(x) = V_-(x) \tag{13.5.8}$$

for $x \in (-\infty, 0)$. Furthermore, equation (13.5.6) becomes

$$V_+(x) = V_-(x) - 2i \cos\frac{\alpha\pi}{2} x^{(\alpha-1)/2}(1-x)^{(\alpha-1)/2} k(x) \tag{13.5.9}$$

for x in (0, 1), with

$$\begin{aligned} V_+(x) &= x^{(\alpha-1)/2}(1-x)^{(\alpha-1)/2}e^{i\pi(\alpha-1)/2}U_+(x), \\ V_-(x) &= x^{(\alpha-1)/2}(1-x)^{(\alpha-1)/2}e^{-i\pi(\alpha-1)/2}U_-(x) \end{aligned} \tag{13.5.10}$$

for $x \in (0, 1)$.

The solution of (13.5.9) is

$$V(z) = -\frac{1}{\pi}\cos\frac{\alpha\pi}{2}\int_0^1 \frac{[\tau(1-\tau)]^{(\alpha-1)/2}k(\tau)}{\tau - z}\,d\tau, \tag{13.5.11}$$

where we have allowed $U(z)$ to have algebraic singularities near the points 0, 1 of order not greater than $-\frac{1}{2}(\alpha+1)$, which is equivalent to allowing $u(\tau)$ to have nothing worse than an integrable algebraic singularity at each point. Computing $V_+(x)$ and $V_-(x)$ and using equations (13.5.10), (13.5.5) and Exercise 7, we obtain

$$\begin{aligned} u(x) =& \frac{\sin\alpha\pi}{2\pi}\frac{d}{dx}\int_0^x \frac{k(t)}{(x-t)^{1-\alpha}}dt - \frac{\cos^2\alpha\pi/2}{\pi^2}\cdot \\ & \frac{d}{dx}\int_0^x \left[\frac{[\tau(1-\tau)]^{(1-\alpha)/2}}{(x-\tau)^{1-\alpha}}p.v.\int_0^1 \frac{k(t)[t(1-t)]^{(\alpha-1)/2}}{t-\tau}dt\right]d\tau. \end{aligned}$$

13.6 Integral Equations with Logarithmic Kernels

Consider the integral equation

$$\int_{-1}^1 \log|x-t|\,u(t)\,dt = k(x), \qquad x \in (0, 1). \tag{13.6.1}$$

To solve this equation, we define the function

$$U(z) = \int_{-1}^1 \log(z-t)u(t)\,dt. \tag{13.6.2}$$

For $x < -1$, we have

$$\begin{aligned} U_+(x) &= \int_{-1}^1 \log|x-t|u(t)\,dt + i\pi\int_{-1}^1 u(t)\,dt, \\ U_-(x) &= \int_{-1}^1 \log|x-t|u(t)\,dt - i\pi\int_{-1}^1 u(t)\,dt. \end{aligned} \tag{13.6.3}$$

But, for $x \in (-1, 1)$, there is a discontinuity and we have

$$\begin{aligned} U_+(x) &= \int_{-1}^{1} \log|x-t|u(t)\,dt + i\pi\int_x^1 u(t)\,dt, \\ U_-(x) &= \int_{-1}^{1} \log|x-t|u(t)\,dt - i\pi\int_x^1 u(t)\,dt; \end{aligned} \tag{13.6.4}$$

Exercise 8. To avoid the discontinuity, we can use $U'(z)$ instead of $U(z)$. Indeed, for $x \in (-1, 1)$, we have

$$U_+'(x) + U_-'(x) = 2k'(x);$$

Exercise 9. In terms of the function

$$V(z) = U'(z)\sqrt{z^2-1},$$

the last equation becomes

$$V_+(x) - V_-(x) = 2i\sqrt{1-x^2}\ k'(x)$$

for $x \in (-1, 1)$. The solution is

$$V(z) = \frac{1}{\pi}\int_{-1}^{1} \frac{\sqrt{1-t^2}\ k'(t)}{t-z}\,dt + \int_{-1}^{1} u(t)\,dt.$$

Note that the last term is a constant; cf. Theorem 13.1.4 and the following remark. (In considering the behavior of $V(z)$ near -1 and $+1$, we have allowed $u(t)$ to have an integrable singularity at each end point.) Since $U_+'(x) - U_-'(x) = -2\pi i u(x)$ by (13.6.4), it follows that

$$u(x) = \frac{1}{\sqrt{1-x^2}}\left[\frac{1}{\pi^2}\,p.v.\int_{-1}^{1}\frac{\sqrt{1-t^2}\,k'(t)}{t-x}\,dt + \frac{1}{\pi}\int_{-1}^{1} u(t)\,dt\right]. \tag{13.6.5}$$

To obtain an expression for the second integral in (13.6.5), we first note that if $k(x) \equiv 1$, then (13.6.1) can be used to show that the integral

$$\int_{-1}^{1} \frac{\log|x-t|}{\sqrt{1-t^2}}\,dt$$

is a constant. Setting $x = 0$ shows that the value of the integral is $-\pi \log 2$; Exercise 10. Multiplying (13.6.1) by $(1-x^2)^{-1/2}$ and integrating from -1 to 1, we obtain

$$\int_{-1}^{1} u(t)\,dt = -\frac{1}{\pi\log 2}\int_{-1}^{1}\frac{k(t)}{\sqrt{1-t^2}}\,dt.$$

Inserting this into (13.6.5) yields Carleman's formula

$$u(x) = \frac{1}{\pi^2\sqrt{1-x^2}}\left[p.v.\int_{-1}^{1}\frac{\sqrt{1-t^2}k'(t)}{t-x}\,dt - \frac{1}{\log 2}\int_{-1}^{1}\frac{k(t)}{\sqrt{1-t^2}}\,dt\right]. \tag{13.6.6}$$

As an extension of equation (13.6.1), we now consider the more general equation

$$\int_{-1}^{1}\left[\log|x-t|p(x-t) + q(x-t)\right]u(t)\,dt = k(t) \tag{13.6.7}$$

for $x \in (-1, 1)$, where $p(x)$ and $q(x)$ are polynomials. As in the previous case, we first define the function

$$U(z) = \frac{1}{\sqrt{z^2-1}}\int_{-1}^{1}\left[p(z-t)\log\frac{z-t}{z+1} + q(z-t)\right]u(t)\,dt, \tag{13.6.8}$$

which is single-valued in the z-plane with a cut along the real axis from -1 to 1; see (13.6.2). For $x \in (-1, 1)$, we have

$$\begin{aligned} U_+(x) &= \frac{-i}{\sqrt{1-x^2}}\left[k(x) - \log(x+1)\int_{-1}^{1}p(x-t)u(t)\,dt\right. \qquad (13.6.9)\\ &\qquad\qquad \left. + i\pi\int_{x}^{1}p(x-t)u(t)dt\right],\\ U_-(x) &= \frac{i}{\sqrt{1-x^2}}\left[k(x) - \log(x+1)\int_{-1}^{1}p(x-t)u(t)\,dt\right. \qquad (13.6.10)\\ &\qquad\qquad \left. - i\pi\int_{x}^{1}p(x-t)u(t)\,dt\right], \end{aligned}$$

and hence

$$U_+(x) - U_-(x) = \frac{-2i}{\sqrt{1-x^2}}\left[k(x) - \log(x+1)\int_{-1}^{1}p(x-t)u(t)\,dt\right]. \tag{13.6.11}$$

Examining the behavior of $U(z)$ at ∞ as well as at the endpoints -1 and 1, we conclude that

$$\begin{aligned} U(z) = -\frac{1}{\pi}\int_{-1}^{1}\frac{1}{\sqrt{1-t^2}}\left[k(t) - \log(1+t)\int_{-1}^{1}p(t-r)u(r)dr\right]\frac{dt}{t-z}\\ +R(z); \qquad (13.6.12) \end{aligned}$$

cf. (13.1.1) and (13.6.12). Here, $R(z)$ is that part of the Laurent series for $U(z)$, in the region outside the unit circle, which does not involve negative powers of z. From (13.6.8), $R(z)$ may be expressed in terms of a finite number of unknown constants c_n defined by

$$c_n = \int_{-1}^{1}t^n u(t)dt, \qquad n \geq 0. \tag{13.6.13}$$

These same constants c_n also occur in the term coming from $\int_{-1}^{1} p(t-r)u(r)dr$ in (13.6.12). Hence, except for a finite number of these c_n, $U(z)$ is known. From (13.6.9) and (13.6.10), we also have

$$U_+(x) + U_-(x) = \frac{2\pi}{\sqrt{1-x^2}} \int_x^1 p(x-t)u(t)\,dt. \tag{13.6.14}$$

By using Laplace transforms, one can show that

$$u(x) = \frac{d}{dx} \int_x^1 M(t-x) \left\{ \frac{1}{2\pi} \sqrt{1-x^2} \left[U_+(t) + U_-(t) \right] \right\}' dt, \tag{13.6.15}$$

where $M(t)$ is the inverse transform of $[s^2 P_1(s)]^{-1}$, $P_1(s)$ being the transform of $p(-t)$; Exercise 11. From (13.6.12), simple calculation shows that

$$\begin{aligned} U_+(x) + U_-(x) = & -\frac{2}{\pi} p.v. \int_{-1}^{1} \frac{1}{\sqrt{1-t^2}} \Big[k(t) \\ & - \log(1+t) \int_{-1}^{1} p(t-r)u(r)dr \Big] \frac{dt}{t-x} \\ & + 2R(x). \end{aligned} \tag{13.6.16}$$

Thus, the solution is complete, except for the evaluation of the constants c_n. A set of linear algebraic equations for the c_n may also be obtained by multiplying equation (13.6.15) by appropriate powers of t and integrating over $(-1, 1)$.

For the special case $p(t) = 1$ and $q(t) = 0$, the result is

$$u(t) = \frac{d}{dx} \left[\frac{1}{\pi^2} \sqrt{1-x^2} p.v. \int_{-1}^{1} \frac{k(t) - c_0 \log(1+t)}{\sqrt{1-t^2}(t-x)} dt \right], \tag{13.6.17}$$

where c_0 is a constant. The value of the constant c_0 can be determined from the condition that $U(z)$, as given in (13.6.12), has no terms of $\frac{1}{z}$ as $z \to \infty$ [cf. (13.6.8) with $p(t) = 1$ and $q(t) = 0$]. This yields

$$c_0 = -\frac{1}{\pi \log 2} \int_{-1}^{1} \frac{g(t)}{\sqrt{1-t^2}} dt \tag{13.6.18}$$

as before; Exercise 14.

13.7 Singular Integral Equations

We conclude this chapter with a discussion of the singular integral equation

$$a(x)u(x) + \frac{b(x)}{\pi i}\, p.v. \int_L \frac{u(t)}{t-x} dt = c(x), \tag{13.7.1}$$

where $a(x), b(x), c(x)$ satisfy a Hölder condition on L, and $a \pm b \neq 0$ on L. Solving this equation is equivalent to finding the function defined by the Cauchy integral

$$U(z) = \frac{1}{2\pi i} \int_L \frac{u(t)}{t-z}\, dt \tag{13.7.2}$$

associated with the Riemann–Hilbert problem

$$U_+(t) = g(t)U_-(t) + f(t), \qquad t \in L; \qquad U_-(\infty) = 0, \tag{13.7.3}$$

where

$$g(t) \equiv \frac{a(t)-b(t)}{a(t)+b(t)}, \qquad f(t) \equiv \frac{c(t)}{a(t)+b(t)}. \tag{13.7.4}$$

To show that finding a solution to equation (13.7.1) reduces to solving the Riemann–Hilbert problem (13.7.3), one can use the Sokhotski–Plemelj formulas for $U(z)$, that is,

$$U_+(t) - U_-(t) = u(t), \qquad U_+(t) + U_-(t) = \frac{1}{\pi i}\, p.v. \int_L \frac{u(\tau)}{\tau - t}\, d\tau. \tag{13.7.5}$$

Substituting these equations into (13.7.1), we obtain (13.7.3). The converse is also true, that is, if the Cauchy integral $U(z)$ in (13.7.2) is the solution of the Riemann–Hilbert problem (13.7.3) with boundary condition $U_-(\infty) = 0$, then the function $u(t)$ in (13.7.5) is a solution of the integral equation (13.7.1); see Muskhelishivili [149].

Singular integral equations of the form (13.7.1) play an important role in studying the more general equation

$$a(t)u(t) + \frac{1}{\pi i}\, p.v. \int_L \frac{K(t,\tau)u(\tau)}{\tau - t}\, d\tau = c(t). \tag{13.7.6}$$

Writing $K(t,\tau) = K(t,t) + [K(t,\tau) - K(t,t)]$, and letting $b(t) \equiv K(t,t)$ and $F(t,\tau) \equiv \frac{1}{i\pi}[K(t,\tau) - K(t,t)]/(\tau - t)$, we get

$$a(t)u(t) + \frac{b(t)}{i\pi}\, p.v. \int_L \frac{u(\tau)}{\tau - t}\, d\tau + \int_L F(t,\tau)u(\tau)\, d\tau = c(t). \tag{13.7.7}$$

Equations of the type (13.7.7) are much more complicated to study than equation (13.7.1). Here we only note that if $F(t,\tau)$ is degenerate, i.e. if $F(t,\tau) = \sum_1^n H_j(t)H_j(\tau)$, then equation (13.7.7) can also be solved in closed form.

Example. Consider the singular integral equation

$$\begin{aligned}(t + t^{-1})u(t) &+ \frac{t - t^{-1}}{\pi i}\, p.v. \int_\Gamma \frac{u(\tau)}{\tau - t}\, d\tau \\ &- \frac{1}{2\pi i} \int_\Gamma (t + t^{-1})(\tau + \tau^{-1})u(\tau)\, d\tau = 2t^2,\end{aligned} \tag{13.7.8}$$

where Γ is the unit circle. The kernel $(t+t^{-1})(\tau+\tau^{-1})$ is degenerate. Hence, according to the remark above, it is expected that equation (13.7.8) is solvable in closed form. Let

$$A = \frac{1}{2\pi i}\int_\Gamma (\tau+\tau^{-1})u(\tau)\,d\tau.$$

Equation (13.7.8) can then be written as

$$(t+t^{-1})u(t) + \frac{t-t^{-1}}{\pi i}\; p.v.\int_\Gamma \frac{u(\tau)}{\tau - t}\,d\tau = 2t^2 + (t+t^{-1})\cdot A.$$

By the Sokhotski–Plemelj formula in (13.7.5), the above equation is equivalent to the Riemann–Hilbert problem.

$$\begin{aligned}(t+t^{-1})\left[U_+(t) - U_-(t)\right] + (t-t^{-1})\left[U_+(t)+U_-(t)\right]&\\ = 2t^2 + (t+t^{-1})\cdot A,&\end{aligned}$$

which can be reduced to

$$U_+(t) = t^{-2}U_-(t) + t + \frac{1}{2}(1+t^{-2})\cdot A, \qquad U_-(\infty) = 0; \tag{13.7.9}$$

see (13.7.3).

We now return to the homogeneous Riemann–Hilbert problem (13.2.5):

$$L_+(\zeta) = g(\zeta)L_-(\zeta) \tag{13.7.10}$$

and the nonhomogeneous Riemann–Hilbert problem (13.2.4):

$$W_+(\zeta) = g(\zeta)W_-(\zeta) + f(\zeta). \tag{13.7.11}$$

In our case, $g(\zeta) = \zeta^{-2}$ and

$$f(\zeta) = \zeta + \frac{1}{2}(1+\zeta^{-2})A. \tag{13.7.12}$$

With $g(\zeta) = \zeta^{-2}$, the homogeneous problem is simply

$$L_+(\zeta) = \zeta^{-2}L_-(\zeta). \tag{13.7.13}$$

By inspection we can take

$$L_+(\zeta) = 1, \qquad L_-(\zeta) = \zeta^2. \tag{13.7.14}$$

Substituting (13.7.13) into (13.7.11) (i.e. replacing $g(\zeta)$ by $L_+(\zeta)/L_-(\zeta)$) gives

$$\frac{W_+(\zeta)}{L_+(\zeta)} - \frac{W_-(\zeta)}{L_-(\zeta)} = \frac{f(\zeta)}{L_+(\zeta)}; \tag{13.7.15}$$

see (13.2.6). From (13.1.2) it follows that the function

$$F(\zeta) = \frac{1}{2\pi i} \int_{\Gamma} \frac{f(\tau)/L_+(\tau)}{\tau - \zeta} \, d\tau \tag{13.7.16}$$

can be written as

$$\frac{f(\zeta)}{L_+(\zeta)} = F_+(\zeta) - F_-(\zeta). \tag{13.7.17}$$

Coupling (13.7.15) and (13.7.17) yields

$$\frac{W(\zeta)}{L(\zeta)} = F(\zeta) + p_{n-1}(\zeta), \tag{13.7.18}$$

where $p_n(\zeta)$ is a polynomial; see Theorem 13.1.4 and the following remark. Note that in our case, the W in (13.7.11) is just the U in (13.7.9). Thus, the boundary condition $U_-(\infty) = 0$ and the function $L_-(\zeta) = \zeta^2$ in (13.7.14) imply that the left-hand side of (13.7.18) is of the order $o(\zeta^{-2})$. From (13.7.16), it is easily seen that the function $F(\zeta)$ on the right-hand side of (13.7.18) has the asymptotic expansion

$$\begin{aligned} F(\zeta) &\sim \frac{i}{2\pi} \int_{\Gamma} \frac{f(\tau)}{L_+(\tau)} \left(\frac{1}{\zeta} + \frac{\tau}{\zeta^2} + \frac{\tau^2}{\zeta^3} + \cdots \right) d\tau \\ &\sim \sum_{s=0}^{\infty} \frac{c_s}{\zeta^{s+1}}, \qquad \zeta \to \infty. \end{aligned} \tag{13.7.19}$$

Balancing the terms on both sides, it follows readily that the polynomial $p_{n-1}(\zeta)$ in (13.7.18) is zero and the coefficients c_0 and c_1 in (13.7.19) must vanish, i.e.

$$\int_C \left[\tau + \frac{A}{2}(1 + \tau^{-2}) \right] d\tau = 0, \qquad \int_{\Gamma} \left[\tau + \frac{A}{2}(1 + \tau^{-2}) \right] \tau d\tau = 0;$$

see (13.7.12) and (13.7.14). The first equation automatically holds, but the second equation requires that $A = 0$. Thus, from (13.7.12), we have $f(\zeta) = \zeta$. With $p_{n-1}(\zeta) = 0$, $f(\tau) = \tau$ and $W(\zeta) = U(\zeta)$, we obtain from (13.7.18), (13.7.16), and (13.7.14)

$$U(\zeta) = \frac{L(\zeta)}{2\pi i} \int_{\Gamma} \frac{\tau}{\tau - \zeta} \, d\tau = \begin{cases} \zeta, & \zeta \text{ inside the circle,} \\ 0, & \zeta \text{ outside the circle.} \end{cases}$$

Returning to (13.7.9) and (13.7.12), we have $U_+(t) = t$, $U_-(t) = 0$ and $f(\zeta) = \zeta$. Therefore, we conclude from (13.7.5) that equation (13.7.8) has the unique solution $u(t) = t$ if the constant A defined above is zero, which is indeed the case since

$$A = \frac{1}{2\pi i} \int_{\Gamma} (\tau + \tau^{-1}) u(\tau) \, d\tau = \frac{1}{2\pi i} \int_{\Gamma} (\tau + \tau^{-1}) \tau \, d\tau = 0.$$

13.8 The other Riemann–Hilbert problem

There is another problem that was studied in various forms by Riemann and, later, Hilbert. It is (at least) equally well known by the term *Riemann–Hilbert problem.* Consider a linear differential equation

$$f^{(p)}(z) + q_1(z)f^{(p-1})(z) + \cdots + q_p(z)f(z) \;=\; 0, \tag{13.8.1}$$

where the coefficients $\{q_k\}$ are rational functions. Let P be the set of poles of the $\{q_k\}$ in $\mathbb{S}$. Fix a coordinate disk D in $\Omega = \mathbb{S} \setminus P$, centered at a point z_0. There is a basis $f_1, f_2, \ldots f_p$ of solutions of (13.8.1) defined in D. If γ is a closed curve in Ω that begins and ends at z_0, then each f_j can be continued along γ to another solution $\widetilde{f}_j$, giving a second basis of solutions defined in D, related to the original set by a matrix A_γ in the group $GL(n, \mathbb{C})$ of $n \times n$ invertible complex matrices. This gives a homomorphism χ from the fundamental group to the $n \times n$ matrices:

$$\chi : H_1(\Omega) \to \; GL(n). \tag{13.8.2}$$

The image is called the *monodromy group* of the equation.

The equation (13.8.1) can be reformulated as a system of equations of first order. If the resulting singular points (including the point at ∞, if it is singular) are simple poles, then (13.8.1) is said to be of *Fuchsian type*. After changing coordinates by an element of $\mathrm{Aut}(\mathbb{S})$ if necessary, we may assume that ∞ is a regular point. Then the first-order system has the form

$$f_j'(z) \;=\; \sum_{k=1}^{n} B_{jk}\frac{1}{z - a_k}, \qquad 1 \le j \le p, \tag{13.8.3}$$

where the a_k are distinct, the B_{jk} are constant, and

$$\sum_{k=1}^{n} B_{jk} \;=\; 0. \tag{13.8.4}$$

Then Problem XXI in Hilbert's famous list of problems [106] can be formulated as follows:

> Let the representation (13.8.2) be given. Prove that there is always a system (13.8.3), (13.8.4) with the given monodromy (13.8.2).

As it turns out, this can be done for equations of degree ≤ 3 or with ≤ 3 singularities. Bolibrukh [29] showed that otherwise there are counter-examples, so the problem, in Hilbert's formulation, has a negative solution. For a full treatment, see Anosov and Bolibrukh [10].

Exercises

1. Prove that the Hölder continuity condition at t_0 in Theorem 13.1.1 can be replaced by the weaker condition
$$\int_{\Gamma} \frac{|f(t) - f(t_0)|}{|t - t_0|} |dt| < \infty.$$
2. Prove (13.1.8).
3. Let $\Gamma_y = \{x + iy : -\infty < x < \infty\}$, $\Gamma_+ = \lim_{y \to 0^+} \Gamma_y$ and $\Gamma_- = \lim_{y \to 0^-} \Gamma_y$. Consider the contour integrals
$$I_{\pm} = \int_{\Gamma_{\pm}} \frac{e^{iz}}{z} dz.$$
   (a) Show that
$$I_{\pm} = \int_{\Gamma_{\pm}} \frac{e^{iz}}{iz^2} dz,$$
   thus proving that the contour integrals are convergent and well defined.
   (b) Use Cauchy's integral formula to show that
$$I_+ - I_- = 2\pi i,$$
   and
$$I_+ = \lim_{R \to \infty} \int_{\pi}^{0} \frac{e^{iR(\cos\theta + i \sin\theta)}}{Re^{i\theta}} d\theta = 0,$$
   hence $I_- = 2\pi i$.
   (c) For any $\pm \mathrm{Re}\, w > 0$, we have
$$\int_{\mathbb{R}} \frac{e^{ix}}{x + iw} dx = \int_{\mathbb{R}} \frac{e^{i(z - iw)}}{z} dz = \begin{cases} 0, & \mathrm{Re}\, w > 0, \\ 2\pi i e^{w}, & \mathrm{Re}\, w < 0. \end{cases}$$
4. (a) Use the results in Exercise 3 to show that if $x_1 > 0$ then
$$\int_{\mathbb{R}} \frac{e^{ix_1\xi_1}}{\xi_1 + i\xi_2} d\xi_1 = \begin{cases} 0, & \mathrm{Re}\, \xi_2 > 0, \\ 2\pi i e^{x_1\xi_2}, & \mathrm{Re}\, \xi_2 < 0. \end{cases}$$
   and if $x_1 < 0$ then
$$\int_{\mathbb{R}} \frac{e^{ix_1\xi_1}}{\xi_1 + i\xi_2} d\xi_1 = \begin{cases} -2\pi i e^{x_1\xi_2}, & \mathrm{Re}\, \xi_2 > 0, \\ 0, & \mathrm{Re}\, \xi_2 < 0. \end{cases}$$
   (b) Show that for $a \in \mathbb{R}$ and $b > 0$,
$$\int_{\mathbb{R}^2} \frac{e^{i(x_1\xi_1 + x_2\xi_2)}}{\xi_1 + (a \pm ib)\xi_2} d\xi_1\xi_2 = \frac{2\pi}{x_2 - (a \pm ib)x_1}.$$

(c) Use (b) to conclude that

$$\int_{\mathbb{R}^2} \frac{e^{i(x_1\xi_1+x_2\xi_2)}}{\xi_1+k\xi_2} d\xi_1\xi_2 = \frac{2\pi \operatorname{sgn}(\operatorname{Im} k)}{x_2-kx_1}, \quad \operatorname{Im} k \neq 0,$$

which proves (13.3.16).

5. (a) Show that equation (13.3.14) can be written as

$$\mu_\pm(x_1,x_2,k) = \mp \int_{-\infty}^{\infty} \left[\frac{1}{2\pi i} \int_{-\infty}^{\infty} \frac{q(y_1,y_2)}{y_2-[x_2-k(x_1-y_2)]} dy_2 \right] dy_1.$$

(b) By applying the Sokhotski–Plemelj formula to the equations in (a), prove that for $k \in \mathbb{R}$,

$$\mu_\pm(x_1,x_2,k) = \pm\frac{1}{2\pi i} \int_{-\infty}^{\infty} p.v. \int_{-\infty}^{\infty} \frac{q(y_1,y_2)}{(x_2-y_2)-k(x_1-y_1)} dy_2 dy_2$$
$$+\frac{1}{2}\left(\int_{-\infty}^{x_1} - \int_{x_1}^{\infty}\right) q(y_1,x_2-k(x_1-y_1))dy_1,$$

which gives the formula in (13.3.17).

6. Prove (13.4.11).
7. Prove the two equations in (13.6.4).
8. Prove the identity

$$U_+'(x)+U_-'(x) = 2k'(x), \qquad x \in (-1,1),$$

where $U(z)$ is defined in (13.6.2).

9. (a) By taking $k=1$ in (13.6.1) and (13.6.5), show that the integral

$$\int_{-1}^{1} \frac{\log|x-t|}{\sqrt{1-t^2}} dt$$

is a constant.

(b) By setting $x=0$ in (a), show that the value of the integral in (a) is $-\pi \log 2$.

10. When the parameters are positive, the usual *Laplace transform* is defined by

$$F(x) = \mathscr{L}_f(s) = \int_0^\infty f(t)e^{-st}\, dt, \qquad s>0. \tag{1}$$

When the parameters are negative, one can use a different notation. For instance, for $\theta<0$, we define

$$F(\theta) = L_f(\theta) = \int_{-\infty}^{0} f(x)e^{-\theta x} dx. \tag{2}$$

Define

$$\hat{f}(y) = f(-y). \tag{3}$$

Show that

$$L_f(\theta) \;=\; \mathscr{L}_{\hat{f}}(-\theta). \tag{4}$$

11. Let $g(x) \;=\; \frac{1}{2\pi}\sqrt{1-x^2}[U_+(x)+U_-(x)]$, so that equation (13.6.14) becomes

$$g(x) \;=\; \int_x^1 p(x-t)u(t)\,dt. \tag{5}$$

Define

$$\tilde{u}(t) \;=\; u(t+1) \quad \text{and} \quad \tilde{g}(x-1) \;=\; g(x). \tag{6}$$

(a) Show that

$$\tilde{g}(x-1) \;=\; \int_{x-1}^0 p(x-1-t)\tilde{u}(t)\,dt, \tag{7}$$

and

$$\tilde{g}(0) \;=\; g(1) = 0. \tag{8}$$

(b) With the Laplace transforms defined in Exercise 10, show that

$$L_{\tilde{g}}(\theta) \;=\; L_{\tilde{u}}(\theta)\cdot L_p(\theta); \tag{9}$$

equivalently,

$$\tilde{G}(\theta) \;=\; \tilde{U}(\theta)P(\theta). \tag{10}$$

(c) Using integration by parts, show that

$$L_{\tilde{g}}(\theta) \;=\; \frac{1}{\theta}L_{\tilde{g}'}(\theta). \tag{11}$$

If $I(x)$ denotes the integral

$$I(x) \;=\; \int_x^0 \tilde{u}(t)\,dt, \tag{12}$$

then show

$$L_I(\theta) \;=\; -\frac{1}{\theta}L_{\tilde{u}}(\theta). \tag{13}$$

(d) Use equation (9) to conclude that

$$L_I(\theta) \;=\; -\frac{1}{\theta^2}\frac{L_{\tilde{g}'}(\theta)}{L_p(\theta)}. \tag{14}$$

12. Recall equation (13.6.15): $M(t)$ is the inverse Laplace transform of $[s^2 P_i(x)]^{-1}$, where $P_1(s)$ is the Laplace transform of $p(-t)$, i.e.

$$\mathscr{L}_M(s) \overset{(a)}{=} \frac{1}{s^2 P_1(s)}, \qquad P_1(s) \overset{(b)}{=} \mathscr{L}_{p(-t)}(s). \tag{15}$$

(a) Show that

$$\mathscr{L}_{\hat{p}}(-\theta) = P_1(\theta).$$

(b) Use (4), (14) and (15) to show that

$$L_I(\theta) = -L_{\hat{M}}(\theta)L_{\tilde{g}'}(\theta), \tag{16}$$

where $\hat{M}(x) = M(-x)$.

13. (a) By interchanging the order of integration, show that

$$\int_{-\infty}^{0} e^{-x\theta} \int_{x}^{0} \hat{M}(x-t)\tilde{g}(t)\,dt\,dx = L_{\hat{M}}(\theta)L_{\tilde{g}'}(\theta)$$

(b) Use the results in (a) and Exercise 12 (b) to conclude

$$\int_{x}^{0} \tilde{u}(t)\,dt = -\int_{x}^{0} M(t-x)\tilde{g}'(t)\,dt.$$

(c) Now prove the formula in (13.6.15).

14. Prove (13.6.18). Hint: use the result in Exercise 9).

Remarks and further reading

The Riemann–Hilbert problem and applications to singular integral equations in $\mathbb{C}$ are treated in depth in Vekua [210], [211]. An important generalization of the Riemann–Hilbert factorization problem takes the function to be factored to be a matrix-valued function. This makes it possible to treat matrix-valued singular integral equations; see Clancey and Gohberg [45]. Calderón and Zygmund [37] developed a far-reaching generalization of the theory of singular integral equations in the plane; see Stein [194], Christ [44], or Peyrière [167]. The Riemann–Hilbert problem plays a crucial part in several areas of asymptotic analysis, including random matrices; see Deift [53].

An active area of application of both Riemann–Hilbert problems is the study of inverse scattering and integrable systems of nonlinear partial differential equations. For use of the first version of Riemann–Hilbert, see Beals, Deift, and Zhou [19] and Deift and Zhou [54]. For use of the second version of Riemann–Hilbert, see the expository article by Its [115].

Chapter 14
Asymptotics and Darboux's method

Suppose that f is holomorphic in a domain that includes the unit disk $\mathbb{D}$. Then its Maclaurin expansion

$$f(z) = \sum_{n=0}^{\infty} a_n z^n \tag{14.0.1}$$

converges in $\mathbb{D}$. The coefficients a_n are determined by f: on any smaller circle centered at the origin we have the integral representation

$$a_n = \frac{1}{2\pi i} \int_{|z|=r} \frac{f(z)}{z^{n+1}}\, dz.$$

A problem that arises frequently in number theory [96], combinatorics [75] and orthogonal polynomials [114] is to determine the asymptotic behavior of the a_n.

One such problem that we treat in this chapter involves the Legendre polynomials $\{P_n\}$ that play a role in Chapter 4. The generating function for these polynomials can be written as

$$f_\theta(z) = \sum_{n=0}^{\infty} P_n(\cos\theta) z^n = \frac{1}{(e^{i\theta} - z)^{1/2}(e^{-i\theta} - z)^{1/2}}, \tag{14.0.2}$$

where the branches are chosen such that $(e^{\pm i\theta} - z)^{1/2} \to e^{\pm \pi i \theta/2}$ as $z \to 0$. For fixed $\theta, 0 < \theta < \pi$, $f_\theta(z)$ is holomorphic in $\mathbb{D}$ and the restriction of f to $\Gamma = \partial\mathbb{D}$ has two algebraic singularities that coalesce as $\theta \to 0$. Thus the asymptotics of the Maclaurin coefficients of f_θ are the asymptotics of $P_n(\theta)$.

As this example suggests, we might want to extract information about the a_n from f on $\Gamma = \partial\mathbb{D}$, under the assumption that the singularities of f on Γ are somehow manageable. Darboux [51] was the first to consider problems of this nature. Darboux considered the case of finitely many distinct algebraic singularities. This work is described in Section 14.1. Recent extensions of the Darboux method are the subject of remaining sections: logarithmic singularities in Section 14.2 and coalescing singularities in Section 14.3. Section 14.4 is devoted to showing that the result on

R. Beals and R. S. C. Wong, *More Explorations in Complex Functions*, Graduate Texts
in Mathematics 298, https://doi.org/10.1007/978-3-031-28288-1_14

coalescing singularities gives an asymptotic expansion for the coefficients. In Section 14.5, these results are applied to the case of the Heisenberg polynomials.

14.1 Algebraic singularities

Suppose that f is holomorphic on $\mathbb{D}$ and on $\Gamma = \partial\mathbb{D}$ except at finitely many distinct points s_j where, in a neighborhood,

$$f(z) = (s_j - z)^{\alpha_j} g_j(z), \quad 1 \le j \le l, \tag{14.1.1}$$

where $g_j(z)$ is holomorphic at $z = \sigma_j$, s_j is a complex number, and the branch of $(s_j - z)^{\alpha_j}$ is holomorphic on $\mathbb{D}$.

For simplicity, we look first at the case of a single singularity at $a \in \Gamma$. The associated function g has an expansion

$$g(z) = \sum_{r=0}^{\infty} c_r (a - z)^r. \tag{14.1.2}$$

The mth *Darboux approximant* of $f(z)$ is defined by

$$f_m(z) = \sum_{r=0}^{m} c_r (a - z)^{r+\alpha}. \tag{14.1.3}$$

Since $f_m(z)$ is holomorphic in $\mathbb{D}$, it has a Maclaurin expansion

$$f_m(z) = \sum_{n=0}^{\infty} b_{mn} z^n. \tag{14.1.4}$$

From (14.1.3), a simple calculation gives

$$b_{mn} = \frac{1}{n!} f_m^{(n)}(0) = (-1)^n \sum_{r=0}^{m} c_r \binom{r+\alpha}{r} a^{r+\alpha-n}. \tag{14.1.5}$$

By Cauchy's theorem, we have from (14.1.4) and (14.0.1)

$$a_n - b_{mn} = \frac{1}{2\pi i} \int_{\Gamma} \frac{f(z) - f_m(z)}{z^{n+1}} \, dz, \tag{14.1.6}$$

where Γ is any contour that contains the origin and lies in $\mathbb{D}$. For convenience, we let

$$\varepsilon_m(z) = f(z) - f_m(z) \tag{14.1.7}$$

and

$$\delta_m(n) = \frac{1}{2\pi i} \int_{\Gamma} \varepsilon_m(z) z^{-n-1} dz. \tag{14.1.8}$$

In view of (14.1.5), equation (14.1.6) can be written as

$$a_n = (-1)^n \sum_{r=0}^{m} c_r \binom{r+\alpha}{n} a^{r+\alpha-n} + \delta_m(n), \tag{14.1.9}$$

where

$$\binom{r+\alpha}{n} = \frac{(r+\alpha)(r+\alpha-1)\cdots(r+\alpha-n+1)}{n\,!}.$$

The functional equation for the gamma function is $\Gamma(z+1) = z\Gamma(z)$, so (14.1.9) can be rewritten as

$$a_n = \sum_{r=0}^{m} c_r \frac{\Gamma(n-\alpha-r)}{n\,!\,\Gamma(-\alpha-r)} a^{r+\alpha-n} + \delta_m(n). \tag{14.1.10}$$

equation

We claim that

$$\delta_m(n) = o(n^{-\alpha-m-1}), \qquad \text{as } n \to \infty. \tag{14.1.11}$$

Integration by parts N times gives

$$\delta_m(n) = \frac{(n-N)!}{n!} \frac{1}{2\pi i} \int_\Gamma \varepsilon_m^{(N)}(z) z^{-(n-N+1)} dz. \tag{14.1.12}$$

Since

$$\varepsilon_m(z) = c_{m+1}(a-z)^{m+\alpha+1} + c_{m+2}(a-z)^{m+\alpha+2} + \cdots$$

in a neighborhood of $z = a$, we have

$$\varepsilon_m^{(N)}(z) = O((a-z)^{m+\alpha+1-N}) \tag{14.1.13}$$

as $z \to 1$. As long as N satisfies

$$m + \operatorname{Re}\alpha + 1 \le \mathrm{N} < \mathrm{m} + \operatorname{Re}\alpha + 2, \tag{14.1.14}$$

then $(a-z)^{m+\alpha+1-N}$ is integrable on Γ, so the the contour Γ can be expanded so that (14.1.12) becomes

$$\delta_m(n) = \frac{1}{2\pi} \frac{(n-N)!}{n!} \int_0^{2\pi} \varepsilon_m^{(N)}(e^{i\theta}) e^{-i(n-N)\theta} d\theta. \tag{14.1.15}$$

Let us emphasize here that $N = N(m) \sim m$. Since the last integral is absolutely integrable, it follows from the Riemann–Lebesgue lemma (Exercise 2) that

$$\delta_m(n) = o\left(\frac{(n-N)!}{n!}\right) = o(n^{-N}) \qquad \text{as } n \to \infty, \tag{14.1.16}$$

which in view of (14.1.14) establishes our claim in (14.1.11). Thus, (14.1.9) gives

$$a_n \sim \sum_{r=0}^{\infty} c_r a^{r+\alpha-n} \frac{\Gamma(n-\alpha-r)}{n!\,\Gamma(-\alpha-r)}, \qquad n \to \infty. \tag{14.1.17}$$

Let us remark that as $n \to \infty$,

$$\frac{\Gamma(n-\alpha-r)}{n!} = \frac{\Gamma(n-\alpha-r)}{\Gamma(n+1)} \sim n^{-\alpha-r-1}; \tag{14.1.18}$$

see Exercise 1.

Let us return to the case (14.1.1) with finitely many singularities $\{s_j\}$. We can apply the same derivation to each s_j. The end result is a sum of asymptotic expansions.

Theorem 14.1.1. *Suppose that f is holomorphic on $\mathbb{D}$ and has distinct singularities $s_1, s_2, \cdots, s_l$, on $\partial\mathbb{D}$, and that in a neighborhood of s_j,*

$$f(z) = \sum_{r=0}^{\infty} c_{jr}(s_j - z)^{\alpha_j + r}. \tag{14.1.19}$$

Then the Maclaurin coefficients a_n in (14.0.1) have the asymptotic expansion

$$\begin{aligned} a_n &\sim \sum_{r=0}^{\infty}\sum_{j=1}^{l} c_{jr}(-1)^n \binom{r+\alpha_j}{n} s_j^{\alpha_j+r-n} \\ &\sim \sum_{r=0}^{\infty}\sum_{j=1}^{l} c_{jr} s_j^{\alpha_j+r-n} \frac{\Gamma(n-\alpha_j-r)}{n\,!\,\Gamma(-\alpha_j-r)} \end{aligned} \tag{14.1.20}$$

as $n \to \infty$.

Now let us return to the example in the introduction:

$$f_\theta(z) = \sum_{n=0}^{\infty} P_n(\cos\theta) z^n = \frac{1}{(e^{i\theta}-z)^{1/2}(e^{-i\theta}-z)^{1/2}},$$

where the branches are chosen such that $(e^{\pm i\theta} - z)^{-\frac{1}{2}} \to e^{\mp\frac{1}{2}i\theta}$ as $z \to 0$. The algebraic singularities are at $a_1 = e^{i\theta}$ and at $a_2 = e^{-i\theta}$. Since

$$\frac{1}{(e^{i\theta}-z)^{1/2}(e^{-i\theta}-z)^{1/2}} = \frac{e^{i\pi/4}}{\sqrt{2\sin\theta}} \sum_{r=0}^{\infty} \binom{-\frac{1}{2}}{r} \frac{(e^{i\theta}-z)^{r-\frac{1}{2}}}{(-2i\sin\theta)^r} \tag{14.1.21}$$

for $|e^{i\theta} - z| < 2\sin\theta$, and

$$\frac{1}{(e^{i\theta}-z)^{1/2}(e^{-i\theta}-z)^{1/2}} = \frac{e^{-i\pi/4}}{\sqrt{2\sin\theta}} \sum_{r=0}^{\infty} \binom{-\frac{1}{2}}{r} \frac{(e^{-i\theta}-z)^{r-\frac{1}{2}}}{(2i\sin\theta)^r} \tag{14.1.22}$$

for $|e^{-i\theta} - z| < 2\sin\theta$, the constants c_{jk} and α_j in (14.1.19) are given by $\alpha_1 = \alpha_2 = -\frac{1}{2}$ and

$$c_{1r} = \frac{e^{i\pi/4}}{\sqrt{2\sin\theta}}\binom{-\frac{1}{2}}{r}\frac{1}{(-2i\sin\theta)^r};$$
$$c_{2r} = \frac{e^{-i\pi/4}}{\sqrt{2\sin\theta}}\binom{-\frac{1}{2}}{r}\frac{1}{(2i\sin\theta)^r}.$$

From (14.1.20), it now follows that

$$P_n(\cos\theta) \sim \left(\frac{2}{\sin\theta}\right)^{\frac{1}{2}}\sum_{r=0}^{\infty}\binom{\frac{1}{2}}{r}\binom{r-\frac{1}{2}}{n}\frac{\cos\theta_{n,r}}{(2\sin\theta)^r} \tag{14.1.23}$$

as $n \to \infty$, where $\theta_{n,r} = (n - r + \frac{1}{2})\theta + (n - \frac{1}{2}r - \frac{1}{4})\pi$.

Olver [157] pointed out an interesting paradox associated to (14.1.23). It is easily verified that the series on the right-hand side of (14.1.23) converges when $2\sin\theta > 1$; that is, $\frac{1}{6}\pi < \theta < \frac{5}{6}\pi$. Thus, it is natural to expect that the sum is $P_n(\cos\theta)$. But from (14.1.22), we have

$$\frac{1}{\sqrt{1-2z\cos\theta+z^2}} = \frac{e^{-i\pi/4}}{\sqrt{2\sin\theta}}\sum_{r=0}^{\infty}\binom{-\frac{1}{2}}{r}\frac{(e^{-i\theta}-z)^{n-\frac{1}{2}}}{(2i\sin\theta)^r}, \tag{14.1.24}$$

which converges uniformly when $|e^{-i\theta} - z| \leq 2\sin\theta - \delta, \delta > 0$. If $2\sin\theta > 1$, then $z = 0$ lies inside the region of uniform convergence. According to (14.0.2), $P_n(\cos\theta)$ is the nth Maclaurin coefficient of the function on the left-hand side of (14.1.24). Hence, differentiating (14.1.24) n times, setting $z = 0$, and equating real parts, we obtain

$$P_n(\cos\theta) = \frac{1}{\sqrt{2\sin\theta}}\sum_{r=0}^{\infty}\binom{-\frac{1}{2}}{r}\binom{r-\frac{1}{2}}{n}\frac{\cos\theta_{n,r}}{(2\sin\theta)^r}. \tag{14.1.25}$$

However compare (14.1.23) with (14.1.25), and note that

$$P_n(\cos\theta) \sim 2P_n(\cos\theta), \qquad \frac{1}{6}\pi < \theta < \frac{5}{6}\pi, \quad n \to \infty. \tag{14.1.26}$$

Example. In [177], Robinson considered the following problem: "Let there be n straight lines in a plane, no three of which meet at a point. Determine the number, g_n, of groups of n of their points of intersection such that no three of the points of the group be on one of the straight lines."

Although Robinson did not find an explicit form for g_n, he showed that g_n satisfies the recurrence relation

$$g_{n+1} = ng_n + \binom{n}{2}g_{n-2}, \qquad n \geq 3, \tag{14.1.27}$$

where $g_1 = g_2 = 0$ and $g_3 = 1$. He also proved that the limit

$$\lim_{n\to\infty}\frac{g_n}{n^n e^{-n}} = B \tag{14.1.28}$$

exists, and conjectured that B might be a new geometric constant. Although several solutions were given to show that the conjecture is false, none of these used the method of Darboux; cf. an editorial note in [178].

If we multiply (14.1.27) by $z^n/n!$ and sum from $n = 3$ to ∞, we obtain

$$\sum_{n=3}^{\infty} g_{n+1}\frac{z^n}{n!} = \sum_{n=2}^{\infty} g_{n+1}\frac{z^{n+1}}{n!} + \frac{1}{2}\sum_{n=3}^{\infty} g_n\frac{z^{n+2}}{n!}.$$

Hence, if we define $f(z) = \sum_{n=3}^{\infty} g_n z^n/n!$, we obtain

$$(1-z)f'(z) - \frac{1}{2}z^2 f(z) = \frac{1}{2}z^2. \tag{14.1.29}$$

The solution of this first-order equation is

$$f(z) = \frac{c}{\sqrt{1-z}}\exp\left\{-\left(\frac{z^2}{4}+\frac{z}{2}\right)\right\} - 1.$$

Since $f(0) = 0$, $c = 1$, and

$$f(z) = \frac{1}{\sqrt{1-z}}\exp\left\{-\left(\frac{z^2}{4}+\frac{z}{2}\right)\right\} - 1. \tag{14.1.30}$$

From the Darboux result (14.1.17) with $a = 1$, $\alpha = -\frac{1}{2}$, and $c_0 = e^{-3/4}$, we have

$$\frac{g_n}{n^n e^{-n}} \sim \sqrt{2}e^{-3/4}, \tag{14.1.31}$$

and hence the constant B in (14.1.28) is given by $B = \sqrt{2}e^{-3/4}$.

14.2 Logarithmic singularities

In this section we consider an extension of Darboux's method to deal with a general version of the type of singularity that occurs in the following example:

$$L(z) = \frac{1}{\log(1-z)} + \frac{1}{z} = \sum_{n=0}^{\infty} l_n z^n. \tag{14.2.1}$$

A second form of the same example is

$$M(z) = \frac{z}{\log(1+z)} = \sum_{n=0}^{\infty} A_n\frac{z^n}{n!}, \qquad |z| < 1; \tag{14.2.2}$$

in fact

$$zL(-z) = M(-z) + 1,$$

so $A_n/n\,! = (-1)^{n-1} l_{n-1}$, $n \geq 1$.

In 2004 Donald Knuth asked Frank Olver about the asymptotics of the coefficients l_n in (14.2.1). Polya [169], pp. 8-9 gives the first few coefficients A_N in (14.2.2):

$$\begin{aligned} &A_0 = 1,\ A_1 = 1,\ A_2 = 1,\ A_3 = 2,\ A_4 = 4, \\ &A_5 = 14,\ A_6 = 38,\ A_7 = 216,\ A_8 = 600,\ A_9 = 6240, \end{aligned} \tag{14.2.3}$$

and asks for a conjecture on A_n. In the solution section, after noting that (14.2.3) makes it reasonable to conjecture that A_n is positive and increasing, Pólya points out that asymptotically

$$\frac{A_n}{n!} \sim (-1)^{n-1} \frac{1}{n \log^2 n}. \tag{14.2.4}$$

The extension that we discuss now, taken from [220], is apparently considered useful in combinatorics; see a remark in [75], p. 438, line 12.

Let $f(z)$ be holomorphic, with Maclaurin expansion

$$f(z) = \sum_{n=0}^{\infty} a_n z^n, \qquad |z| < 1. \tag{14.2.5}$$

Assume that $f(z)$ has a singularity at $z = 1$, and is holomorphic within and on the contour Γ shown in Figure 14.1), for some $\delta > 0$. In a neighborhood of $z = 1$, $f(z)$ is assumed to have the form

$$f(z) = (1-z)^{\lambda-1} (\log(1-z))^{\mu} g(z), \tag{14.2.6}$$

where λ and μ are complex numbers, $g(z)$ is holomorphic at $z = 1$, and $\log(1-z)$ has its principal value, which is real when z is real and < 1.

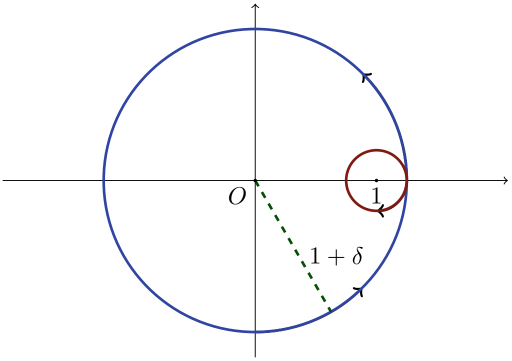

Fig. 14.1 Contour Γ; $\delta > 0$

From (14.2.5) we have

$$\begin{aligned} 2\pi i a_n &= \int_\Gamma f(z) z^{-n-1} dz \\ &= \int_{|z|=1+\delta} f(z) z^{-n-1} dz + \int_{|z-1|=\delta} f(z) z^{-n-1} dz. \end{aligned} \tag{14.2.7}$$

We will vary δ, with

$$\delta = \delta_n = n^{-\frac{1}{2}}, \tag{14.2.8}$$

and assume that on the larger circle $|z| = 1 + \delta_n$, f satisfies

$$f(z) = O(n^s) \qquad n \to \infty, \tag{14.2.9}$$

for some fixed s. A simple estimation then gives

$$\begin{aligned} \int_{|z| = 1+\delta_n} f(z) z^{-n-1} dz &= O\left(\frac{n^s}{\left(1 + 1/\sqrt{n}\,\right)^n} \right), \qquad n \to \infty, \\ &= O\left(\exp(-\varepsilon\sqrt{n}\,) \right), \qquad n \to \infty, \end{aligned}$$

for some fixed $\varepsilon > 0$. Since this integral is exponentially small, it is clear that the asymptotic behavior of a_n will be determined by the asymptotic behavior of the integral

$$I_n = \frac{i}{2\pi} \int_{|z-1|=\delta_n} f(z) z^{-n-1} dz, \tag{14.2.10}$$

where the path of integration on $|z - 1| = \delta_n$ is now oriented in the positive direction.

Before we begin the study of behavior of the integral I_n, we shall digress briefly to discuss the function

$$M(\lambda, \mu; n) = \frac{i}{2\pi} \int_\infty^{(0^+)} (-z)^{\lambda-1} (\log(-z))^\mu e^{-(n+1)z} dz, \tag{14.2.11}$$

where the loop contour of integration and the cuts in the z-plane are illustrated in Figure 14.2. If $\mu = 0$, then the integral in (14.2.11) can be expressed in terms of the gamma function; see Section 2.10.

$$\frac{1}{\Gamma(1-\lambda)} = \frac{i}{2\pi} \int_\infty^{(0^+)} (-u)^{\lambda-1} e^{-u}\, du, \qquad |\arg(-u)| \le \pi. \tag{14.2.12}$$

Differentiating both sides with respect to λ gives

$$D^k \left[\frac{1}{\Gamma(1-\lambda)} \right] = \frac{i}{2\pi} \int_\infty^{(0^+)} (-u)^{\lambda-1} (\log(-u))^k e^{-u}\, du, \tag{14.2.13}$$

where $D^k = d^k/d\lambda^k$.

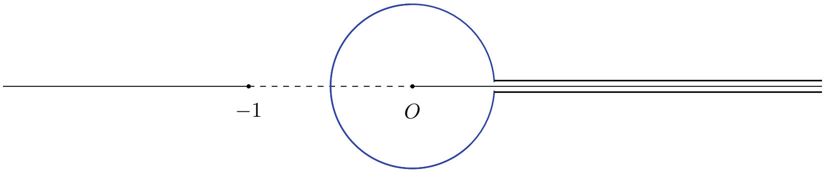

Fig. 14.2 Loop contour

Lemma 14.2.1. *For any fixed integer $N \geq 0$,*

$$M(\lambda, \mu; n) = \frac{(-\log(n+1))^{\mu}}{(n+1)^{\lambda}} \left[\sum_{k=0}^{N} \binom{\mu}{k} \frac{D^k[1/\Gamma(1-\lambda)]}{(-\log(n+1))^k} + O\left(\frac{1}{(\log(n+1))^{N+1}} \right) \right] \tag{14.2.14}$$

as $n \to \infty$.

Proof. In (14.2.11), we make the change of variable $u = (n+1)z$ and obtain

$$M(\lambda, \mu; z) = \frac{1}{(n+1)^{\lambda}} \frac{i}{2\pi} \int_{\infty}^{(0+)} (-u)^{\lambda-1} \left[\log\left(\frac{-u}{n+1} \right) \right]^{\mu} e^{-u}\, du. \tag{14.2.15}$$

Divide the loop path of integration into two parts A and B, where A is the portion contained in $|u| \leq (n+1)^{\rho}$ for some fixed ρ in $(0, 1)$, and B is the remaining portion of the path (i.e. two half-lines extending to ∞). Since $\arg(-u) = \pm\pi$ on B, $\log(-u/n+1)$ satisfies the inequalities

$$\pi \leq \left| \log\left(-\frac{u}{n+1} \right) \right| \leq \left| \log\left| \frac{u}{n+1} \right| \right| + \pi. \tag{14.2.16}$$

Hence, $|\log(-u/(n+1))|$ is uniformly bounded away from zero. Although this function becomes unbounded on B, it is bounded by the larger of $\log|u|$ and $\log(n+1)$. An easy estimation shows that for $|u| \geq (n+1)^{\rho}$, there must exist an $\varepsilon > 0$ such that

$$\int_B (-u)^{\lambda-1} \left[\log\left(\frac{-u}{n+1} \right) \right]^{\mu} e^{-u}\, du = O\left(\exp(-\varepsilon n^{\rho}) \right) \tag{14.2.17}$$

as $n \to \infty$, with λ and μ unrestricted, and the order relation holds uniformly.

On the part A of the loop, we have

$$\left[1 - \frac{\log(-u)}{\log(n+1)} \right]^{\mu} = \sum_{k=0}^{N} \binom{\mu}{k} \frac{(\log(-u))^k}{(-\log(n+1))^k} + O\left(\frac{(\log u)^{N+1}}{(\log(n+1))^{N+1}} \right) \tag{14.2.18}$$

as $n \to \infty$, for every fixed integer $N \geq 0$. Since $\int_A (-u)^{\lambda-1}(\log(-u))^k e^{-u}\,du$ exists as an absolutely convergent integral for each fixed $k \geq 0$, it follows that

$$\begin{aligned}&\int_A (-u)^{\lambda-1}\left[\log\left(\frac{-u}{n+1}\right)\right]^{\mu} e^{-u}\,du \\ &= (-\log(n+1))^{\mu}\left[\sum_{k=0}^{N}\binom{\mu}{k}\frac{1}{(-\log(n+1))^k}\int_A (-u)^{\lambda-1}(\log(-u))^k e^{-u}\,du \right.\\ &\qquad \left. + O\left(\frac{1}{(\log(n+1))^{N+1}}\right)\right], \quad \text{as } n \to \infty. \qquad (14.2.19)\end{aligned}$$

By the argument used to obtain (14.2.17), we also have

$$\begin{aligned}&\int_A (-u)^{\lambda-1}(\log(-u))^k e^{-u}\,du \\ &= \int_{\infty}^{(0^+)} (-u)^{\lambda-1}(\log(-u))^k e^{-u}\,du + O(\exp(-\varepsilon n^{\rho})). \qquad (14.2.20)\end{aligned}$$

Hence, on account of (14.2.13),

$$\frac{i}{2\pi}\int_A (-u)^{\lambda-1}(\log(-u))^k e^{-u}\,du = D^k\left[\frac{1}{\Gamma(1-\lambda)}\right] + O(\exp(-\varepsilon n^{\rho})) \qquad (14.2.21)$$

as $n \to \infty$. A combination of the results (14.2.15), (14.2.17), (14.2.19), and (14.2.21) yields the desired result (14.2.14). □

The result in Lemma 14.2.1 will be used in a slightly different form. First, we note that

$$\begin{aligned}E(w,u) &= \exp\{-(n+1)[\log(1+u) - u]\} \\ &= \exp\left\{-\frac{1}{2}wu\left[\frac{2(\log(1+u)-u)}{u^2}\right]\right\}, \qquad (14.2.22)\end{aligned}$$

where

$$w = (n+1)u. \qquad (14.2.23)$$

(The first equality will be used later in (14.2.35).) Now, let $P_m(w)$ be the polynomials defined by

$$g(u+1)E(w,u) = \sum_{m=0}^{\infty} P_m(w)u^m, \qquad (14.2.24)$$

where $g(z)$ is the function given in (14.2.6). Thus we have

$$P_m(w) = \frac{1}{m!}\frac{d^m}{du^m}[g(u+1)E(w,u)]\bigg|_{u=0}. \qquad (14.2.25)$$

Consider the integral

$$J_m(n) \;=\; \frac{i}{2\pi}\int_{\gamma_n}(-u)^{\lambda+m-1}(\log(-u))^{\mu}P_m((n+1)u)e^{-(n+1)u}\,du, \tag{14.2.26}$$

where γ_n is the contour which traverses the circle $|u| = \delta_n$ in the positive direction, and begins and ends on the positive half of the real axis. The polynomial $P_m((n+1)u)$ may be written as

$$P_m((n+1)u) = \sum_{s=0}^{m} p_s(n+1)^s u^s, \tag{14.2.27}$$

where p_s is a fixed number. Hence,

$$J_m(x) \;=\; \sum_{s=0}^{m}(-1)^s p_s(n+1)^s\frac{i}{2\pi}\int_{\gamma_n}(-u)^{\lambda+m+s-1}(\log(-u))^{\mu}e^{-(n+1)u}d\,u. \tag{14.2.28}$$

Since the error incurred by extending the circular paths of integration to infinite loops is exponentially small, we have

$$J_m(n) \;=\; \sum_{s=0}^{m}(-1)^s p_s(n+1)^s M(\lambda+m+s,\mu;n) + O\left(\exp(-\varepsilon n^{\frac{1}{2}})\right) \tag{14.2.29}$$

as $n \to \infty$; see (14.2.17). By Lemma 14.2.1,

$$J_m(n) \sim \frac{(-\log(n+1))^{\mu}}{(n+1)^{\lambda+m}}\sum_{k=0}^{\infty}\binom{\mu}{k}A_k(\lambda,m)(-\log(n+1))^{-k}, \tag{14.2.30}$$

where

$$A_k(\lambda,m) \;=\; \sum_{s=0}^{m}(-1)^s p_s D^k\left[\frac{1}{\Gamma(1-\lambda-m-s)}\right]. \tag{14.2.31}$$

Returning to (14.2.10), we replace $z-1$ by u and obtain

$$I_n \;=\; \frac{i}{2\pi}\int_{\gamma_n} f(u+1)(u+1)^{-n-1}\,du, \tag{14.2.32}$$

where γ_n is the contour described in (14.2.26).

Theorem 14.2.2. *If the function in (14.2.5) is holomorphic within and on the contour Γ shown in Figure 14.1, and if $f(z)$ satisfies the conditions in (14.2.6) and (14.2.9), then for any fixed integers $N \geq 0$ the Maclaurin coefficients of $f(z)$ have the asymptotic expansion*

$$a_n \;=\; \sum_{m=0}^{N}(-1)^m J_m(n) + O\left(\frac{(\log n)^{\mu}}{n^{\lambda+N+1}}\right) \tag{14.2.33}$$

as $n \to \infty$, where $J_m(n)$ is defined in (14.2.26) and its asymptotic behavior is given in (14.2.30).

Proof. Substituting (14.2.6) into (14.2.32) gives

$$I_n = \frac{i}{2\pi} \int_{\gamma_n} (-u)^{\lambda} (\log(-u))^{\mu} g(u+1)(u+1)^{-n-1} du. \tag{14.2.34}$$

By (14.2.22) and (14.2.24),

$$g(u+1) \exp\{-(n+1)[\log(u+1) - u]\} = \sum_{m=0}^{N} P_m(w) u^m + R_N(n,u), \tag{14.2.35}$$

where $N \geq 0$ is any fixed number. The error term $R_N(n,u)$ in (14.2.35) can be expressed as

$$R_N(n,u) = \left(\frac{1}{2\pi i} \int_{|\zeta| = 2K/n^{\frac{1}{2}}} g(\zeta+1) E(w,\zeta) \frac{d\zeta}{\zeta^{N+1}(\zeta - u)} \right) u^{N+1},$$

using Taylor's formula with remainder, where K is a positive constant and $E(w,\zeta)$ is given in (14.2.22). A simple estimation gives

$$R_N(n,u) = O(n^{(N+1)/2} u^{N+1}) \qquad \text{as } n \to \infty, \tag{14.2.36}$$

provided $|u| \leq K/n^{\frac{1}{2}}$. Coupling (14.2.26) and (14.2.35), we obtain

$$I_n = \sum_{m=0}^{N} (-1)^m J_m(n) + E_N(n), \tag{14.2.37}$$

where

$$E_N(n) = \frac{i}{2\pi} \int_{\gamma_n} (-u)^{\lambda-1} (\log(-u))^{\mu} R_N(n,u) e^{-(n+1)u} du. \tag{14.2.38}$$

Now, choose N large enough so that $\mathrm{Re}\,(\lambda + \mathrm{N} - 1) > 0$. The circular path of integration can then be replaced by two line segments joining $u = 0$ to $u = \delta_n$, one on the upper side of the cut in the u-plane, and the other on the lower side of this cut. Hence,

$$E_N(n) = O\left(n^{(N+1)/2} \int_L \left| (-u)^{\lambda+N} (\log(-u))^{\mu} e^{-(n+1)u}\, du \right| \right), \tag{14.2.39}$$

where L is the integration path shown in Figure 14.3. By the argument given in Lemma 14.2.1, it can be shown that the integral in (14.2.39) is $O\left((\log n)^{\mu}/n^{\lambda+N+1}\right)$. Thus,

$$E_N(n) = O\left((\log n)^{\mu}/n^{\lambda+(N+1)/2}\right) \quad \text{as } n \to \infty. \tag{14.2.40}$$

From (14.2.37), it follows that

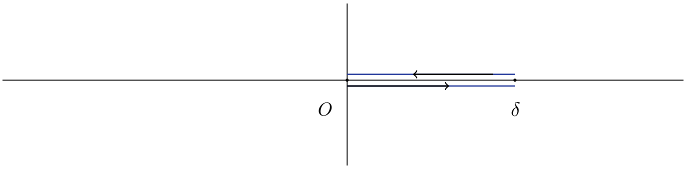

Fig. 14.3 Integration path L

$$I_n = \sum_{m=0}^{N}(-1)^m J_m(n) + O\left((\log n)^\mu / n^{\lambda+(N+1)/2}\right). \tag{14.2.41}$$

This is short of the claim in (14.2.33). However, the order of the terms $J_m(n)$, given in (14.2.30), indicates that the result in (14.2.41) can be improved to read

$$I_n = \sum_{m=0}^{N}(-1)^m J_m(n) + O\left((\log n)^\mu / n^{\lambda+N+1}\right) \tag{14.2.42}$$

as $n \to \infty$, for any fixed integer $N \geq 0$. This is essentially the statement of the theorem, on account of (14.2.7) and (14.2.10). □

When $\mu = 0$, the canonical form (14.2.5) reduces to the Darboux condition (14.1.1) with α replaced by $\lambda - 1$, and our expansion (14.2.33) is equivalent to the result given in (14.1.17).

From (14.2.30), we have

$$J_0(n) \sim \frac{(-\log(n+1))^\mu}{(n+1)^\lambda} \sum_{k=0}^{\infty} \binom{\mu}{k} (-\log(n+1))^{-k}\, p_0 D^k \left[\frac{1}{\Gamma(1-\lambda)}\right]$$

and

$$J_1(n) \sim \frac{(-\log(n+1))^\mu}{(n+1)^{\lambda+1}} \sum_{k=0}^{\infty} \binom{\mu}{k} (-\log(n+1))^{-k} \sum_{s=0}^{1} p_s D^k \left[\frac{1}{\Gamma(1-\lambda-s)}\right].$$

Hence, for any integer $N \geq 0$,

$$\begin{aligned} J_0(n) - J_1(n) = g(1) &\frac{(-\log(n+1))^\mu}{(n+1)^\lambda} \\ &\times \left[\sum_{k=0}^{N} \binom{\mu}{k} (-\log(n+1))^{-k}\, D^k \left[\frac{1}{\Gamma(1-\lambda)}\right] \right. \\ &\left. + O\left((\log(n+1))^{-N-1}\right) + O\left(n^{-1}\right) \right] \end{aligned}$$

as $n \to \infty$. Clearly, none of the terms of $J_1(n)$ can contribute to the asymptotic expansion for a_n unless the infinite asymptotic expansion for $J_0(n)$ terminates after a finite number of terms (e.g. when μ is a positive integer). The same will be true for $J_m(n)$ for $m \geq 1$. Hence, the general situation is

$$a_n \sim g(1) \frac{(-\log(n+1))^\mu}{(n+1)^\lambda} \sum_{k=0}^{\infty} \binom{\mu}{k} D^k \left[\frac{1}{\Gamma(1-\lambda)} \right] (-\log(n+1))^{-k} \quad (14.2.43)$$

as $n \to \infty$.

Returning to (14.2.1), we note that

$$zL(z) - 1 = f(z),$$

where $f(z)$ is given in (14.2.6) with $\lambda = 1$, $\mu = -1$ and $g(z) = z$. Thus,

$$l_n = a_{n+1}, \qquad n = 0, 1, 2, \cdots$$

Since $\Gamma(1-\lambda)\Gamma(\lambda) = \pi / \sin \pi z$, a straightforward calculation gives

$$l_n \sim \frac{1}{(n+2)\log^2(n+2)} \left[1 - \frac{2\gamma}{\log(n+2)} + \cdots \right], \quad (14.2.44)$$

where $\gamma = -\Gamma'(1)$ is the Euler constant.

14.3 Two coalescing singularities

Returning to (14.0.2), we note that the generating function for the Legendre polynomial has two algebraic singularities, one at $z = e^{i\theta}$ and the other at $z = e^{-i\theta}$. As $\theta \to 0^+$, these two singularities coalesce at $z = 1$ and the asymptotic expansion of the Legendre polynomials, given in (14.1.23), breaks down; that is, Darboux's method fails when two or more singularities coalesce.

Fields [71] in 1967 presented a uniform treatment of Darboux's method when two or three singularities coalesce. He considered the case

$$f(z, \theta) = (1-z)^{-\lambda} [(e^{i\theta} - z)(e^{-i\theta} - z)]^{-\alpha} g(z; \theta) = \sum_{n=0}^{\infty} a_n(\theta) z^n, \quad (14.3.1)$$

where the Maclaurin expansion converges for $|z| < 1$ uniformly for $\theta \in [0, \pi]$, the branch of $(1-z)^{-\lambda}$ and $[(e^{i\theta} - z)(e^{-i\theta} - z)]^{-a}$ are chosen such that each is holomorphic on $\mathbb{D}$ and equals 1 as $z = 0$, and $g(z, \theta)$ is holomorphic in $|z| \leq e^{\eta}$ $(\eta > 0)$.

Fields derived an expansion which is uniform in certain θ-intervals depending on n. However his result seems too complicated for practical applications; see, e.g. Erdélyi [66], p.167, Olver [159], pp.112–113, and Wong [218], p.145. In response to the comments by Erdélyi and Olver, Wong and Zhao [219] found a way to derive a simpler form of uniform asymptotic expansion for the Maclaurin coefficients $a_n(\theta)$ in (14.3.1) when two or three algebraic singularities on the circle of convergence coalesce. (Neither the series in (14.1.17) given by Darboux nor Field's result is an asymptotic (power) expansion in the usual sense.)

To begin, we start with the simple case of two singularities, namely,

$$f(z,\theta) = [(e^{i\theta}-z)(e^{-i\theta}-z)]^{-\alpha}g(z,\theta) = \sum_{n=0}^{\infty} a_n(\theta)z^n, \tag{14.3.2}$$

where $g(z,\theta)$ is holomorphic in $|z| \le e^{\eta}$, $\eta > 0$; this is (14.3.1) with $\lambda = 0$. Contribution to the large n behavior of $a_n(\theta)$ still comes from the singular points $z = e^{\pm i\theta}$, which are now allowed to vary as $\theta \to 0^+$. We shall show that the approximants in this case are

$$T_1(x) = \frac{1}{2\pi i}\int_{\Gamma_0}(s^2+1)^{-\alpha}e^{xs}ds, \quad T_2(x) = \frac{1}{2\pi i}\int_{\Gamma_0} s(s^2+1)^{-\alpha}e^{xs}ds, \tag{14.3.3}$$

where Γ_0 is a Hankel-type loop which starts and ends at $-\infty$, and encircles $s = \pm i$ in the positive sense. It is easily verified that $(d/dx)T_1(x) = T_2(x)$, and it can also be shown that

$$T_1(x) = \frac{\sqrt{\pi}}{\Gamma(\alpha)}\left(\frac{x}{2}\right)^{\alpha-\frac{1}{2}} J_{\alpha-\frac{1}{2}}(x), \quad T_2(x) = \frac{\sqrt{\pi}}{\Gamma(\alpha)}\left(\frac{x}{2}\right)^{\alpha-\frac{1}{2}} J_{\alpha-\frac{3}{2}}(x), \tag{14.3.4}$$

where $J_\nu(x)$ is the Bessel function; see Exercise 3. Our ultimate goal here is to establish that the Maclaurin coefficients in (14.3.2) have an asymptotic expansion of the form

$$a_n(\theta) \sim \frac{\sqrt{\pi}}{\Gamma(\alpha)}\left(\frac{n}{2\theta}\right)^{\alpha-\frac{1}{2}}\left[J_{\alpha-\frac{1}{2}}(n\theta)\sum_{k=0}^{\infty}\frac{\alpha_k(\theta)}{n^k} + J_{\alpha-\frac{3}{2}}(x)\sum_{k=0}^{\infty}\frac{\beta_k(\theta)}{n^k}\right] \tag{14.3.5}$$

as $n \to \infty$, holding uniformly for $\theta \in [0, \pi-\delta]$, $\delta > 0$, with coefficients $\alpha_k(\theta)$ and $\beta_k(\theta)$ determined recursively.

To prove (14.3.5), we start with the Cauchy formula

$$a_n(\theta) = \frac{1}{2\pi i}\int_{\Gamma} g(z,\theta)(1-2z\cos\theta+z^2)^{-\alpha}\frac{dz}{z^{n+1}}, \tag{14.3.6}$$

where Γ is a simple closed contour which encloses $z = 0$ but not $z = e^{\pm i\theta}$ and lies in the domain of z-holomorphy of $f(z;\theta)$. We may choose Γ so that it consists of two portions Γ_I and Γ_E, where Γ_I is a curve starting from $z = e^{-0i}e^{\eta}$, enclosing $z = e^{\pm i\theta}$ but not $z = 0$ in clockwise orientation, and ending at $z = e^{0i}e^{\eta}$, while Γ_E is the circle $|z| = e^{\eta}$, oriented anticlockwise; see Figure 14.4.

We first show that the contribution from Γ_E is exponentially small. Indeed, let us define

$$\mathscr{A}_n(\theta) = \frac{1}{2\pi i}\int_{\Gamma_I} g(z,\theta)(1-2z\cos\theta+z^2)^{-\alpha}\frac{dz}{z^{n+1}} \tag{14.3.7}$$

and

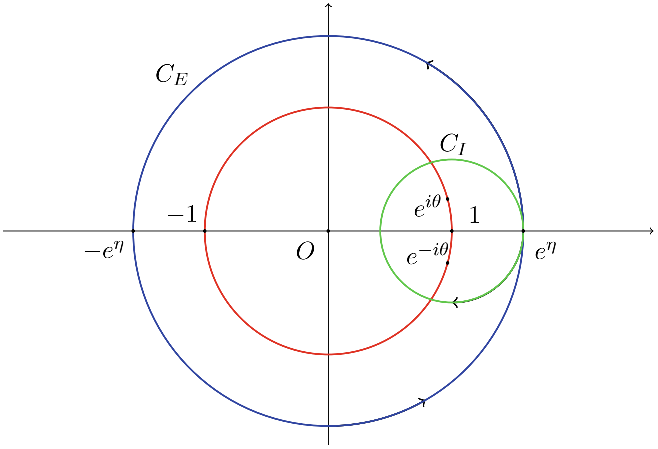

Fig. 14.4 Contour in (14.3.6)

$$\varepsilon_E(\theta) = \frac{1}{2\pi i}\int_{\Gamma_E} g(z,\theta)(1-2z\cos\theta+z^2)^{-\alpha}\frac{dz}{z^{n+1}}. \tag{14.3.8}$$

On Γ_E, we have

$$(e^{\eta}-1)^2 \le |1-2z\cos\theta+z^2| \le (e^{\eta}+1)^2.$$

From (14.3.8), it follows

$$|\varepsilon_E(\theta)| \le c(g,\eta)e^{-\eta n}, \tag{14.3.9}$$

where $c(g,\eta)$ is a positive constant. In fact, one may choose

$$c(g,\eta) = \max_{|z|=e^{\eta}}\{|g(z;\theta)|\}\cdot\max\{(e^{\eta}-1)^{-2\alpha},\ (e^{\eta}+1)^{-2\alpha}\}.$$

From (14.3.6) to (14.3.9), we obtain

$$a_n(\theta) = \mathscr{A}_n(\theta)+\varepsilon_E(\theta), \tag{14.3.10}$$

where $|\varepsilon_E| \le c(g,\theta)e^{-\eta n}$.

Now we consider the behavior of $\mathscr{A}_n$. The change of variable

$$z = e^{-\theta s} \tag{14.3.11}$$

in (14.3.7) gives

$$\mathscr{A}_n(\theta) = \frac{\theta^{1-2\alpha}}{2\pi i}\int_{\Gamma} h_0(s,\theta)(s^2+1)^{-\alpha}e^{n\theta s}ds, \tag{14.3.12}$$

where

$$h_0(s,\theta) \;=\; g(e^{-\theta s},\theta)\left[\left(\frac{e^{-s\theta}-e^{i\theta}}{(-s-i)\theta}\right)\left(\frac{e^{-s\theta}-e^{-i\theta}}{(-s+i)\theta}\right)\right]^{-\alpha} \tag{14.3.13}$$

is holomorphic in s for Re s $\geq -\eta/\theta$ and $|s \pm i| \leq 2\pi/\theta$. In (14.3.12), Γ is the image of Γ_I under the transformation (14.3.11). That is, Γ is the positively oriented curve in the s-plane which starts at $e^{-i\pi}\eta/\theta$, ends at $e^{i\pi}\eta/\theta$, and encloses both $s = \pm i$.

To pick up the first-level contribution from the integral in (14.3.12), we write

$$h_0(s,\theta) \;=\; \alpha_0(\theta) + s\beta_0(\theta) + (s^2+1)g_0(s,\theta), \tag{14.3.14}$$

where the coefficients $\alpha_0(\theta)$ and $\beta_0(\theta)$ are determined by setting $s = \pm i$. More precisely, we have

$$\alpha_0(\theta) \;=\; \frac{1}{2}[h_0(i,\theta)+h_0(-i,\theta)], \quad \beta_0(\theta) \;=\; \frac{1}{2i}[h_0(i,\theta)-h(-i,\theta)]. \tag{14.3.15}$$

Note that $g_0(s,\theta)$ in (14.3.14) has the same domain of s-holomorphy as $h_0(s,\theta)$. Inserting (14.3.14) into (14.3.12) and integrating the last term by parts give

$$\mathscr{A}_n(\theta) \;=\; \theta^{1-2\alpha}\alpha_0(\theta)[T_1(n\theta)-\varepsilon_{T_1}] + \theta^{1-2\alpha}\beta_0(\theta)[T_2(n\theta)-\varepsilon_{T_2}] + \frac{1}{n}\varepsilon_1, \tag{14.3.16}$$

where $T_1(x)$ and $T_2(x)$ are given in (14.3.3),

$$\varepsilon_{T_l} \;=\; \int_{e^{i\pi}\infty}^{e^{-i\pi}\eta/\theta} s^{l-1}(s^2+1)^{-\alpha}e^{n\theta s}ds + \int_{e^{i\pi}\eta/\theta}^{e^{i\pi}\eta/\theta} s^{l-1}(s^2+1)^{-\alpha}e^{n\theta s}ds, \tag{14.3.17}$$

$l = 1, 2$, and

$$\varepsilon_1 \;=\; \Sigma_1 + \varepsilon_{1,E}. \tag{14.3.18}$$

In (14.3.18),

$$\varepsilon_{1,E} \;=\; \theta^{-2\alpha}\cdot\frac{1}{2\pi i}\left[g_0(s,\theta)(s^2+1)^{1-\alpha}e^{n\theta s}\right]\Bigg|_{s=e^{-i\pi}\eta/\theta}^{s=e^{i\theta}\eta/\theta} \tag{14.3.19}$$

represents the endpoint contribution and

$$\Sigma_1 \;=\; \frac{\theta^{1-2\alpha}}{2\pi i}\int_\Gamma h_1(s,\theta)(s^2+1)^{-\alpha}e^{n\theta s}ds, \tag{14.3.20}$$

where

$$\begin{aligned} h_1(s,\theta) &= -\frac{1}{\theta}(s^2+1)^{\alpha}\frac{d}{ds}\left[g_0(s,\theta)(s^2+1)^{1-\alpha}\right] \\ &= -\frac{1}{\theta}\left[(s^2+1)\frac{d}{ds}+2(1-\alpha)s\right]g_0(s,\theta). \end{aligned} \tag{14.3.21}$$

It can be seen from (14.3.21) that $h_1(s,\theta)$ has the same domain of s-holomorphy as $g_0(s,\theta)$, and hence as $h_0(s,\theta)$. It can also be seen that the integral representation for Σ_1 is of the same form as (14.3.12) for $\mathscr{A}_n(\theta)$. Thus, the procedure can be repeated.

Define inductively

$$h_k(s,\theta) \;=\; \alpha_k + s\beta_k + (s^2+1)g_k(s,\theta), \quad k = 0, 1, 2, \cdots, \tag{14.3.22}$$

and

$$h_{k+1}(s,\theta) \;=\; -\frac{1}{\theta}\left[(s^2+1)\frac{d}{ds} + 2(1-\alpha)s\right]g_k(s,\theta), k = 0, 1, 2, \cdots \tag{14.3.23}$$

Repeated application of integration by parts as above gives the expansion

$$\begin{aligned}\alpha_n(\theta) = \theta^{1-2\alpha}T_1(n\theta)\sum_{k=0}^{m-1}\frac{\alpha_k(\theta)}{n^k}&\\ +\theta^{1-2\alpha}T_2(n\theta)\sum_{k=0}^{m-1}\frac{\beta_k(\theta)}{n^k} + \varepsilon(\theta,m)&\end{aligned} \tag{14.3.24}$$

for $m = 1, 2, \cdots$, where

$$\begin{aligned}\varepsilon(\theta,m) = \varepsilon_E + \sum_{k=1}^{m}\frac{\varepsilon_{k,E}}{n^k} - \theta^{1-2\alpha}\sum_{k=0}^{m-1}\frac{\alpha_k(\theta)\varepsilon_{T_1} + \beta_k(\theta)\varepsilon_{T_2}}{n^k}&\\ +\frac{1}{n^m}\Sigma_m.)&\end{aligned} \tag{14.3.25}$$

In (14.3.25), ε_E is given in (14.3.8),

$$\varepsilon_{k,E} \;=\; \theta^{-2\alpha}\frac{1}{2\pi i}\left[g_{k-1}(s,\theta)(s^2+1)^{1-\alpha}e^{ns\theta}\right]\Big|_{s=e^{-i\pi}\eta/\theta}^{s=e^{i\pi}\eta/\theta}, \; k = 1, 2, \cdots \tag{14.3.26}$$

and

$$\Sigma_m \;=\; \frac{\theta^{1-2\alpha}}{2\pi i}\int_{\Gamma} h_m(s,\theta)(s^2+1)^{-\alpha}e^{n\theta s}ds, \quad m = 1, 2, \cdots \tag{14.3.27}$$

One can see from (14.3.22) and (14.3.23) that $h_k(s,\theta)$ and $g_k(s,\theta)$ have the same domain of s-holomorphy as $h_0(s,\theta)$.

To conclude this section, we show that ε_{T_1} and ε_{T_2} in (14.3.17) are exponentially small. Set

$$I \;=\; \int_{\eta/\theta}^{\infty}(s^2+1)^{-\alpha}e^{-\eta\theta s}ds, \tag{14.3.28}$$

and make the change of variables $s = (t+1)\eta/\theta$. The integral in (14.3.28) becomes

$$I \;=\; \eta^{1-2\alpha}\theta^{2\alpha-1}e^{-\eta n}\int_0^{\infty}\left[(t+1)^2 + \frac{\theta^2}{\eta^2}\right]^{-\alpha}e^{-\eta n t}\,dt. \tag{14.3.29}$$

Note that $\theta^2/\eta^2 \geq 0$, and

$$\left[(t+1)^2 + \frac{\theta^2}{\eta^2}\right]^{-\alpha} \leq (t+1)^{-2\alpha}$$

for $\theta \in [0, 2\pi]$ and $t \geq 0$. Hence

$$|I| \leq C(\eta)\theta^{2\alpha-1}e^{-\eta n}\int_0^\infty (t+1)^{-2\alpha}e^{-\eta n t}\,dt \;\leq\; C(\eta)\theta^{2\alpha-1}\frac{1}{n}e^{-\eta n}, \tag{14.3.30}$$

where we have used $C(\eta)$ as a generic symbol to denote positive constants, independent of θ and n, the value of which may differ in different places. From (14.3.17) and (14.3.30), it follows that

$$\theta^{1-2\alpha}|\varepsilon_{T_1}| \leq C(\eta)\frac{1}{n}e^{-\eta n} \tag{14.3.31}$$

and

$$\theta^{1-2\alpha}|\varepsilon_{T_2}| \leq C(\eta)e^{-\eta n} \tag{14.3.32}$$

for $\theta \in [0, \pi]$. The last inequality is obtained by combining (14.3.17) with (14.3.30) and integrating by parts in both integrals in (14.3.17).

14.4 Asymptotic nature of the expansion (14.3.24)

In the previous section we derived the expansion (14.3.24) for the Maclaurin coefficients $a_n(\theta)$ in (14.3.2). To show that (14.3.24) is an asymptotic expansion, we must estimate the remainder term $\varepsilon(\theta, m)$ and demonstrate that the coefficients $\alpha_k(\theta)$ and $\beta_k(\theta)$ are bounded. In this section we do this step by step.

Theorem 14.4.1. *Assume that $g(z,\theta)$ in (14.3.2) is uniformly bounded for $\theta \in [0, \pi]$, and is z-holomorphic in $|z| \leq e^\eta$ ($\eta > 0$). For any integers $m \geq 1$, we have*

$$a_n(\theta) = \theta^{1-2\alpha}T_1(n\theta)\sum_{k=0}^{m-1}\frac{\alpha_k(\theta)}{n^k} + +\theta^{1-2\alpha}T_2(n\theta)\sum_{k=0}^{m-1}\frac{\beta_k(\theta)}{n^k} + \varepsilon(\theta, m), \tag{14.4.1}$$

where

$$|\alpha_k(\theta)| \;\leq\; M_k, \qquad |\beta_k(\theta)/\theta| \;\leq\; M_k \tag{14.4.2}$$

for $k = 0, 1, 2, \cdots$, *and*

$$|\varepsilon(\theta, m)| \;\leq\; M_m\frac{\theta^{1-2\alpha}}{n^m}\Big[|T_1(n\theta)| + |T_2(n\theta)|\Big] \tag{14.4.3}$$

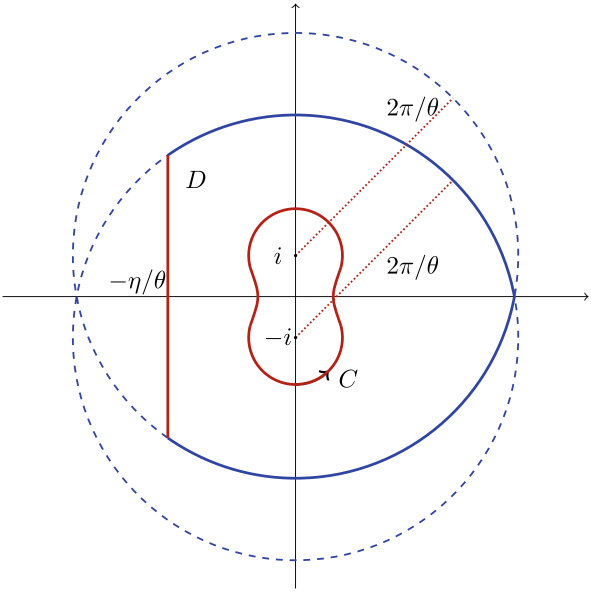

Fig. 14.5 Domain D and contour Γ

for $m = 1, 2, 3, \cdots$. The positive constants M_k, $k = 0, 1, 2, \cdots$, are independent of $\theta \in [0, \pi - \delta]$, $\delta > 0$, the coefficients $\alpha_k(\theta)$ and $\beta_k(\theta)$ are defined successively by (14.3.13), (14.3.22), and (14.3.23). The remainder term $\varepsilon(\theta, m)$ in (14.4.1) involves ε_E, ε_{T_l}, $\varepsilon_{k,E}$, and Σ_m, which are explicitly given in (14.3.8), (14.3.17), (14.3.26), and (14.3.27), respectively.

Step 1. *Proof of (14.4.2).* Define the region

$$D = \left\{ s \,|\, \mathrm{Re\, s} \ \geq \ -\frac{\eta}{\theta},\ |\mathrm{s} \pm \mathrm{i}| < \frac{2\pi}{\theta} \right\}, \tag{14.4.4}$$

and let Γ be a contour in D which encloses $s = \pm i$ in the positive sense; see Figure 14.5. Using Cauchy's integral formula and the fact that $h_0(s, \theta)$ is s-holomorphic in the region D, we have from (14.3.15)

$$\begin{aligned} \alpha_0(\theta) &= \frac{1}{2\pi i} \int_\Gamma A_0(s, \theta) h_0(s, \theta)\, ds, \\ \beta_0(\theta) &= = \frac{1}{2\pi i} \int_\Gamma B_0(s, \theta) h_0(s, \theta)\, ds, \end{aligned} \tag{14.4.5}$$

where

$$A_0(s, \theta) = \frac{s}{s^2 + 1} \quad \text{and} \quad B_0(s, \theta) = \frac{1}{s^2 + 1}. \tag{14.4.6}$$

Define inductively

$$A_k(s,\theta) = \frac{1}{\theta}(1+s^2)^{-1}\left\{(s^2+1)\frac{d}{ds}+2\alpha s\right\}A_{k-1}(s,\theta) \tag{14.4.7}$$

and

$$B_k(s,\theta) = \frac{1}{\theta}(1+s^2)^{-1}\left\{(s^2+1)\frac{d}{ds}+2\alpha s\right\}B_{k-1}(s,\theta) \tag{14.4.8}$$

for $k = 1, 2, 3, \cdots$. The differential operator in (14.4.7) and (14.4.8) can be written as

$$\frac{1}{\theta}(s^2+1)^{-\alpha}\frac{d}{ds}\Big[(s^2+1)^{\alpha}A_{k-1}(s,\theta)\Big]. \tag{14.4.9}$$

In terms of these rational functions, we obtain

Lemma 14.4.2. *For $\theta \in [0,\pi]$ and $k = 0, 1, 2, \cdots$, we have*

$$\alpha_k(\theta) = (1-2\alpha)\frac{\beta_{k-1}(\theta)}{\theta} + \frac{1}{2\pi i}\int_{\Gamma} A_k(s,\theta)h_0(s,\theta)\,ds \tag{14.4.10}$$

and

$$\beta_k(\theta) = \frac{1}{2\pi i}\int_{\Gamma} B_k(s,\theta)h_0(s,\theta)\,ds, \tag{14.4.11}$$

where Γ is the same contour given in (14.4.5) and for convenience, we have set $\beta_{-1} = 0$.

Proof. We demonstrate only the result in (14.4.10). The corresponding result in (14.4.11) can be established in a similar manner. The case $k = 0$ is already given in (14.4.5). For $k \geq 1$, we have from (14.3.22) and (14.3.23)

$$\begin{aligned}
\alpha_k(\theta) &= \frac{1}{2\pi i}\int_{\Gamma} A_0(s,\theta)h_k(s,\theta)\,ds\\
&= \frac{1}{2\pi i}\int_{\Gamma} A_0(s,\theta)\left\{-\frac{1}{\theta}(s^2+1)^{\alpha}\frac{d}{ds}\Big[g_{k-1}(s,\theta)(s^2+1)^{1-\alpha}\Big]\right\}ds\\
&= \frac{1}{2\pi i}\int_{\Gamma} A_1(s,\theta)(s^2+1)g_{k-1}(s,\theta)\,ds\\
&= \frac{1}{2\pi i}\int_{\Gamma} A_1(s,\theta)[h_{k-1}(s,\theta)-\alpha_{k-1}(\theta)-s\beta_{k-1}(\theta)]ds\\
&= \frac{1}{2\pi i}\int_{\Gamma} A_1(s,\theta)h_{k-1}(s,\theta)\,ds + (1-2\alpha)\frac{\beta_{k-1}(\theta)}{\theta}\\
&\cdots\cdots\\
&= \frac{1}{2\pi i}\int_{\Gamma} A_k(s,\theta)h_0(s,\theta)\,ds + (1-2\alpha)\frac{\beta_{k-1}(\theta)}{\theta},
\end{aligned}$$

thus proving (14.4.10). Here we have repeatedly used

$$\frac{1}{2\pi i}\int_{\Gamma} A_k(s,\theta)ds = 0, \qquad k = 1, 2, \cdots,$$

$$\frac{1}{2\pi i}\int_{\Gamma} \theta s A_1(s,\theta)ds = 2\alpha - 1,$$

and

$$\frac{1}{2\pi i}\int_{\Gamma} \theta s A_k(s,\theta)\,ds = 0, \qquad k = 2, 3, \cdots,$$

which follows from (14.4.6), (14.4.7), and (14.4.13) below. □

Lemma 14.4.3. *For $\theta \in (0, \pi]$, there exists a constant $M_k > 0$, independent of θ, such that*

$$|A_k(s,\theta)| \le M_k\theta \ \text{ and } \ |B_k(s,\theta)| \le M_k\theta^2 \tag{14.4.12}$$

for $|s| \le M/\theta$, $|s - i| \ge L/\theta$ and $|s + c| \ge L/\theta$, where L and M are positive constants.

Proof. By induction, one can use (14.4.6) and (14.4.7) to write

$$A_k(s,\theta) = \frac{1}{\theta^k}\frac{p_{k+1}(s)}{(1+s^2)^{k+1}} \tag{14.4.13}$$

for $k = 0, 1, 2, \cdots$, where $p_{k+1}(s)$ is a polynomial of degree $k + 1$, with coefficients independent of θ. It can also be shown from (14.4.6) and (14.4.8) that

$$B_k(s,\theta) = \frac{1}{\theta^k}\frac{q_k(s)}{(1+s^2)^{k+1}} \tag{14.4.14}$$

for $k = 0, 1, 2, \cdots$, where $q_k(s)$ is a polynomial of degree k, with coefficients independent of θ. The two inequalities in (14.4.12) now follow from (14.4.13) and (14.4.14), respectively. □

To estimate the function $h_0(s,\theta)$ in (14.3.13), we first recall that $g(e^{-\theta s},\theta)$ is uniformly bounded for $\theta \in [0, \pi]$ and $\operatorname{Re} s \ge -\eta/\theta$. Hence there exists a constant M_g, independent of θ and s, such that

$$|g(e^{-\theta s},\theta)| \le M_g \qquad \text{for} \quad \operatorname{Re} s \ge -\eta/\theta. \tag{14.4.15}$$

Next, since $(e^z - 1)/z$ has no zero and is bounded on the circle $|z| = b$ for $0 < b < 2\pi$, there exist positive constants m_b and M_b such that

$$m_b \le \left|\frac{e^z - 1}{z}\right| \le M_b \qquad \text{for} \quad |z| \le b.$$

Hence, for $0 < b < 2\pi$, we have

$$m_b \le \left|\frac{e^{-s\theta} - e^{i\theta}}{(-s - i)\theta}\right| \le M_b \qquad \text{for} \quad |s + i| \le \frac{b}{\theta} \tag{14.4.16a}$$

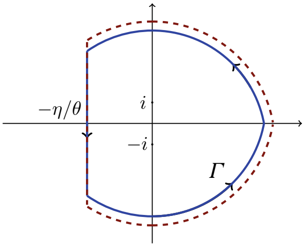

Fig. 14.6 The deformed contour Γ

and

$$m_b \;\leq\; \left|\frac{e^{-s\theta}-e^{-i\theta}}{(-s+i)\theta}\right| \;\leq\; M_b \qquad \text{for} \quad |s-i| \leq \frac{b}{\theta}. \tag{14.4.16b}$$

Combining (14.4.15), (14.4.16), and (14.3.13) gives

Lemma 14.4.4. *For $\theta \in (0,\pi]$, there exists a constant $M_D > 0$, independent of s and θ, such that*

$$|h_0(s,\theta)| \;\leq\; M_D \;\; \textit{for}\, \mathrm{Re\,s} \;\geq\; -\frac{\eta}{\theta},\;\; |\mathrm{s}+\mathrm{i}| \;\leq\; \frac{\mathrm{b}}{\theta} \;\textit{and}\; |\mathrm{s}-\mathrm{i}| \;\leq\; \frac{\mathrm{b}}{\theta}. \tag{14.4.17}$$

Now, for $\theta \in [0,\pi-\delta]$, one may specify $b = 2\pi-\delta$ in (14.4.17). Without loss of generality, we may always assume that $\eta < \sqrt{\pi(3\pi-2\delta)}$. The contour Γ in (14.4.5) may be deformed so that it consists of

(i) $|s+i| = b/\theta$, $\mathrm{Im\,s} \geq 0$ and $\mathrm{Re\,s} \geq -\eta/\theta$;
(ii) $|s-i| = b/\theta$, $\mathrm{Im\,s} \leq 0$ and $\mathrm{Re\,s} \geq -\eta/\theta$; and
(iii) the segment of $\mathrm{Re\,s} = -\eta/\theta$ joining (i) and (ii); see Figure 14.6.

The constants M and L in Lemma 14.4.3) may be chosen to be $M = \max\{\eta,\ 3\pi - 2\delta\}$ and $L = \min\{\eta,\ \delta\}$. A combination of Lemmas 14.4.2–14.4.4) gives the boundedness of the coefficients $\alpha_k(\theta)$ and $\beta_k(\theta)/\theta$, thus proving (14.4.2).

Step 2. *Bounds for ε_E and $\varepsilon_{k,E}$.* To describe the behavior of $T_1(n\theta)$ and $T_2(n\theta)$, we make use of (14.3.4). From the behavior of $J_{\alpha-\frac{1}{2}}(n\theta)$ and $J_{\alpha-\frac{3}{2}}(n\theta)$, when $n\theta$ is small, we have

$$T_1(n\theta) \sim \frac{1}{\Gamma(2\alpha)}(n\theta)^{2\alpha-1} \qquad \text{as} \quad n\theta \to 0^+,$$

and

$$T_2(n\theta) \sim \frac{1}{\Gamma(2\alpha-1)}(n\theta)^{2\alpha-1} \qquad \text{as} \quad n\theta \to 0^+;$$

see, e.g. [22], p.225. Hence

$$|T_1(n\theta)| + |T_2(n\theta)| \;\geq\; C(n\theta)^{2\alpha-1} \tag{14.4.18}$$

for $n\theta \in [0, \varepsilon]$ and ε small, where C depends only on ε. The interval of validity for (14.4.18) can of course be extended to $n\theta \in [0, B]$ for a finite B, since $J_{\alpha-\frac{1}{2}}(\tau)$ and $J_{\alpha-\frac{3}{2}}(\tau)$ have no common zero. The constant C may then depend on B.

In view of the behavior of the Bessel function (see, e.g. [158], p.133), we again have from (14.3.4)

$$T_1(n\theta) \sim \frac{1}{\Gamma(\alpha)}\left(\frac{n\theta}{2}\right)^{\alpha-1}\cos\left(n\theta - \frac{1}{2}\alpha\pi\right) \qquad \text{as} \quad n\theta \to \infty,$$

and

$$T_2(n\theta) \sim \frac{1}{\Gamma(\alpha)}\left(\frac{n\theta}{2}\right)^{\alpha-1}\cos\left(n\theta - \frac{1}{2}\alpha\pi + \frac{\pi}{2}\right) \qquad \text{as } n\theta \to \infty.$$

Hence

$$|T_1(n\theta)| + |T_2(n\theta)| \ \geq\ C(n\theta)^{\alpha-1} \tag{14.4.19}$$

for $n\theta \in [B, \infty)$, where B is a large but fixed number.

To estimate the error terms, we note from (14.3.31) that

$$|\varepsilon_{T_1}| \ \leq\ \frac{C(\eta)}{n}\theta^{2\alpha-1}e^{-\eta n} \ =\ C(\eta)\left\{n^{m-2\alpha}e^{-\eta n}\right\}\frac{(n\theta)^{2\alpha-1}}{n^m} \ \leq\ C\frac{(n\theta)^{2\alpha-1}}{n^m}$$

for $n\theta \in [0, B]$, and from (14.3.32) that

$$\theta|\varepsilon_{T_2}| \ \leq\ C\frac{(n\theta)^{2\alpha-1}}{n^m}$$

also for $n\theta \in [0, B]$. Here, $C(\eta)$ is the generic symbol for positive constants, independent of θ and n, used in (14.3.30), (14.3.31), and (14.3.32).

When $n\theta \in [B, \infty)$, and hence for $\theta \in [B/n, \pi]$, it follows from (14.3.31)

$$|\varepsilon_{T_1}| \ \leq\ C\left\{\frac{\theta^\alpha}{n^\alpha}e^{-\eta n}\right\}(n\theta)^{\alpha-1} \ \leq\ C\{n^{m+|\alpha|-\alpha}e^{-\eta n}\}\frac{(n\theta)^{\alpha-1}}{n^m} \ \leq\ C\frac{(n\theta)^{\alpha-1}}{n^m}.$$

Similarly, from (14.3.32) we have

$$\theta|\varepsilon_{T_2}| \ \leq\ C\frac{(n\theta)^{\alpha-1}}{n^m}.$$

Summarizing the last four inequalities, we obtain, in view of (14.4.18) and (14.4.19)

$$\theta^{l-1}|\varepsilon_{T_l}| \ \leq\ C_m\frac{1}{n^m}\{|T_1(\theta)| + |T_2(\theta)|\}, \qquad l \ =\ 1, 2,$$

where C_m is a constant independent of n and θ. Accordingly,

$$\left|\theta^{1-2\alpha}\sum_{k=0}^{m-1}\frac{\alpha_k(\theta)\varepsilon_{T_1}+\beta_k(\theta)\varepsilon_{T_2}}{n^k}\right| \le C\frac{\theta^{1-2\alpha}}{n^m}\{|T_1(\theta)|+|T_2(\theta)|\} \qquad (14.4.20)$$

for all n and θ, where use has been made of the estimates in (14.4.2). Note that the quantity on the left-hand side of this inequality is exactly the third member in the remainder $\varepsilon(\theta, n)$ given in (14.3.25) and (14.4.1).

An estimate for ε_E can be obtained by comparing (14.3.9) with (14.4.18) and (14.4.19). Since

$$e^{-\eta n} = \{n^{m-2\alpha+1}e^{-\eta n}\}\frac{\theta^{1-2\alpha}(n\theta)^{2\alpha-1}}{n^m} \le C\frac{\theta^{1-2\alpha}(n\theta)^{2\alpha-1}}{n^m}$$

for $n\theta \in [0, B]$, and

$$e^{-\eta n} \le C\{n^{m-\alpha+|\alpha|+1}e^{-\eta n}\}\frac{\theta^{1-2\alpha}(n\theta)^{\alpha-1}}{n^m} \le C\frac{\theta^{1-2\alpha}(n\theta)^{\alpha-1}}{n^m}$$

for $n\theta \in [B, \infty)$ (and hence $\theta \in [B/n, \pi]$), it follows that

$$|\varepsilon_E| \le M_m\frac{\theta^{1-2\alpha}}{n^m}[|T_1(\theta)|+|T_2(\theta)|]. \qquad (14.4.21)$$

Note that ε_E is the first member in the remainder term $\varepsilon(\theta, m)$ given in (14.3.25).

To investigate $\varepsilon_{k,E}$, we first analyze $h_k(s,\theta)$ and $g_k(s,\theta)$. Analogous to the sequences $\{A_k(s,\theta)\}$ and $\{B_k(s,\theta)\}$ defined inductively in (14.4.7) and (14.4.8), we now introduce another sequence of rational functions associated with (14.3.22) and (14.3.23). By Cauchy's theorem,

$$h_0(s,\theta) = \frac{1}{2\pi i}\int_{C_u}\frac{h_0(u,\theta)}{u-s}du,$$

where the integration path C_u is a contour that lies in the domain D of the u-holomorphy (see Figure 14.5)), and encloses $u = s$ and $u = \pm i$ in the anticlockwise direction. Set

$$Q_0(u,s,\theta) = \frac{1}{u-s}. \qquad (14.4.22)$$

Then

$$h_0(s,\theta) = \frac{1}{2\pi i}\int_{C_u} Q_0(u,s,\theta)h_0(u,\theta)\,du. \qquad (14.4.23)$$

We further define

$$Q_k(u,s,\theta) = \frac{1}{\theta}\left[\frac{d}{du}+2\alpha\frac{u}{u^2+1}\right]Q_{k-1}(u,s,\theta), \qquad k = 1,2,3,\cdots; \qquad (14.4.24)$$

see the comment following (14.4.8).

In view of (14.4.22) and (14.4.24), it can be shown by induction that

$$Q_k(u,s,\theta) = \frac{1}{\theta^k}\sum_{l=0}^{k} \frac{P_l(u)}{(u-s)^{k-l+1}(1+u^2)^l}, \tag{14.4.25}$$

where $P_l(u)$ is a polynomial of degree l whose coefficients are independent of s and θ. The last equation suggests that

$$\frac{1}{2\pi i}\int_{C_u} Q_k(u,s,\theta)\,du = 0, \qquad k = 1,2,3,\cdots,$$
$$\frac{1}{2\pi i}\int_{C_u} \theta u Q_1(u,s,\theta)\,du = 2\alpha - 1,$$
$$\frac{1}{2\pi i}\int_{C_u} u Q_k(u,s,\theta)\,du = 0, \qquad k = 2,3,4,\cdots.$$

(These are similar to the last three equations in the proof of Lemma 14.4.2; see also the proof of Lemma 14.4.3.)

Similar to the derivation of (14.4.10) and (14.4.11), we have

Lemma 14.4.5. *For $\theta \in [0,\pi]$ and $k = 0,1,2,\cdots$, we have*

$$h_k(s,\theta) = (1-2\alpha)\frac{\beta_{k-1}(\theta)}{\theta} + \frac{1}{2\pi i}\int_{C_u} Q_k(u,s,\theta)h_0(u,\theta)\,du, \tag{14.4.26}$$

where, for convenience, we have set $\beta_{-1}(\theta) = 0$.

From (14.4.25), one can also see that the following estimates hold:

Lemma 14.4.6. *For $\theta \in (0,\pi]$, $|u| \le M/\theta$, $|s| \le M/\theta$, $|u-s| \ge L/\theta$, $|u-i| \ge L/\theta$, and $|u+i| \ge L/\theta$, there exist constants M_k, $k = 0,1,2,\cdots$, such that*

$$|Q_k(u,s,\theta)| \le M_k\theta. \tag{14.4.27}$$

Choose an s-contour Γ_s similar to Γ, described in the paragraph following Lemma 14.4.4, Γ_s consists of

(i) $|s+i| = b/\theta$, $\mathrm{Im\,s} \ge 0$ and $\mathrm{Re\,s} \ge -(\eta-\varepsilon)/\theta$;
(ii) $|s-i| = b/\theta$, $\mathrm{Im\,s} \le 0$ and $\mathrm{Re\,s} \ge -(\eta-\varepsilon)/\theta$; and
(iii) the segment of $\mathrm{Re\,s} = -(\eta-\varepsilon)/\theta$ joining (i) and (ii), where ε is a positive number which is sufficiently small so that Γ_s encloses $\pm i$; see Figure 14.7.

Similarly, we define Γ_u, consisting of

(i) $|u+i| = (b+\varepsilon)/\theta$, $\mathrm{Im\,u} \ge 0$ and $\mathrm{Re\,u} \ge -\eta/\theta$;
(ii) $|u-i| = (b+\varepsilon)/\theta$, $\mathrm{Im\,u} \le 0$ and $\mathrm{Re\,u} \ge -\eta/\theta$; and
(iii) segment of $\mathrm{Re\,u} = -\eta/\theta$ joining (i) and (ii).

Denote by D_s the domain bounded by Γ_s. If $s \in D_s$ and $u \in \Gamma_u$, then Lemma 14.4.4 holds with s replaced by u, and Lemmas 14.4.5 and 14.4.6 hold since Γ_u encloses $u = s$ and $u = \pm i$, and lies in D, and since it follows from the previous description that $|u-s| \ge \varepsilon/\theta$.

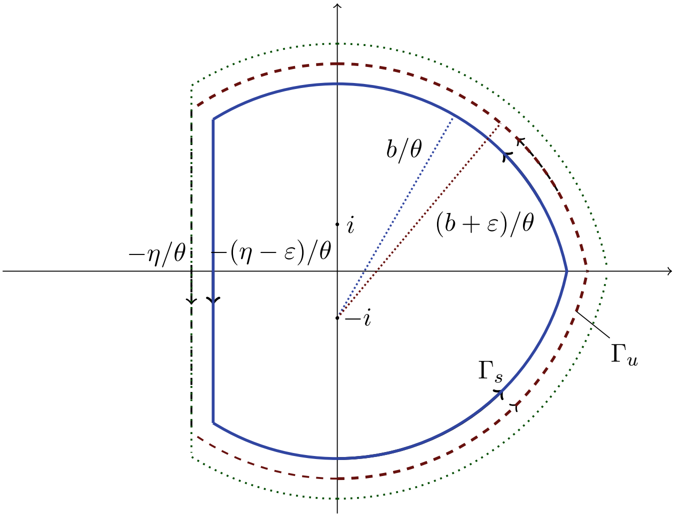

Fig. 14.7 Contours Γ_u and Γ_s

We notice that in the previous derivation leading to (14.3.24) and the estimation leading to (14.4.2), (14.4.20), and (14.4.21), we only require that η be a fixed positive number. Hence, in these cases we can replace η by a smaller number, say $\eta' = \eta - \varepsilon$, and the validity of these previous results will remain. For convenience, we continue to denote the small η' by η. With this understanding, one obtains the following result by combining Lemmas 14.4.4–14.4.6 and using the fact that $\int_{\Gamma_u} |du| = O(1/\theta)$.

Lemma 14.4.7. *For $\theta \in (0, \pi]$, and $k = 0, 1, 2, \cdots$, we have*

$$|h_k(s,\theta)| \le M_k, \qquad s \in D_s, \tag{14.4.28}$$

where D_s is the domain bounded by

(i) $|s+i| = b/\theta$, $\operatorname{Re} s \ge -\eta/\theta$ *and* $\operatorname{Im} s \ge 0$*;*
(ii) $|s-i| = b/\theta$, $\operatorname{Re} s \ge -\eta/\theta$ *and* $\operatorname{Im} s \le 0$*; and*
(iii) $\operatorname{Re} s = -\eta/\theta$, $|\operatorname{Im} s| \le \sqrt{\mathrm{b}^2 - \eta^2}/\theta$.

We are now ready to consider the term $\varepsilon_{k,E}$ given in (14.3.26). By (14.3.22),

$$g_{k-1}(s,\theta)(s^2+1)^{1-\alpha} e^{ns\theta} = \left[h_{k-1}(s,\theta) - \alpha_{k-1}(\theta) - s\frac{\beta_{k-1}(\theta)}{\theta}\theta\right] \times (s^2+1)^{-\alpha} e^{ns\theta}.$$

Since

$$\eta^2/\theta^2 < \eta^2/\theta^2 + 1 < (\eta^2 + \pi^2)/\theta^2,$$

it follows that $(s^2+1)^{-\alpha}\Big|_{s=e^{\pm i\pi}\eta/\theta}$ is bounded by $C(\eta)\theta^{2\alpha}$. In view of the boundedness of h_{k-1}, $\alpha_{k-1}(\theta)$, and $\beta_{k-1}(\theta)/\theta$, it follows that

$$|\varepsilon_{k,E}| \ \leq\ C(\eta, M_{k-1})e^{-\eta n}. \tag{14.4.29}$$

Using the inequalities preceding (14.4.21), one can show that the estimate for ε_E in (14.4.21) also holds for $\varepsilon_{k,E}$, $k=1,2,\cdots$. Hence, we have

$$\left|\sum_{k=1}^{m}\frac{\varepsilon_{k,E}}{n^k}\right| \ \leq\ M_m\frac{\theta^{1-2\alpha}}{n^m}[|T_1(n\theta)|+|T_2(n\theta)|]. \tag{14.4.30}$$

Note that this is an estimate for the second member in the error term $\varepsilon(\theta,m)$ given in (14.3.24) and (14.4.1).

Step 3. *An Estimate for $\Sigma_m(\theta)$.* The only remaining task in this section is to estimate Σ_m given in (14.3.27). For $n\theta\in[0,B]$, we deform the integration path Γ so that it starts from $e^{-i\pi}\eta/\theta$ and ends at $e^{i\pi}\eta/\theta$, and that there are positive constants L and M such that $|s\pm i|\ \geq\ L/\theta$ and $|s|\ \leq\ M/\theta$ along Γ; for an example of such a path, see the paragraph following Lemma 14.4.4. Now make the change of variable $n\theta s\ =\ t$, and denote the image of s-curve Γ by $\tilde{\Gamma}_t$. It is readily seen that $\tilde{\Gamma}_t$ is a curve which starts at $e^{-i\pi}\eta n$ and ends at $e^{i\pi}\eta n$; along $\tilde{\Gamma}_t$, we have $|t\pm in\theta|\geq nL$, $|t|\leq nM$, and

$$\Sigma_m\ =\ \frac{\theta^{1-2\alpha}}{2\pi i}(n\theta)^{2\alpha-1}\int_{\tilde{\Gamma}_t}h_m(s,\theta)(t^2+(n\theta)^2)^{-\alpha}e^t\,dt. \tag{14.4.31}$$

We further deform $\tilde{\Gamma}_t$ so that it traverses from $e^{-i\pi}\eta n$ to $e^{-i\pi}(2B)$ along the lower edge of the negative real line, moves to $e^{i\pi}(2B)$ on the circle $|t|\ =\ 2B$ in the anticlockwise direction, and then along the upper edge of the negative real line to $e^{i\pi}\eta n$. The deformed curve will still be denoted by $\tilde{\Gamma}_t$. Along this new curve, $|(t^2+(n\theta)^2)^{-\alpha}|\ \leq\ C(B)t^{-2\alpha}$ and $|h_m(s,\theta)|\ \leq\ M_m$; see (14.4.28). Thus

$$\begin{aligned}|\Sigma_m| &\leq \theta^{1-2\alpha}C(M_m,B)(n\theta)^{2\alpha-1}\int_{\tilde{\Gamma}_t}|t^{-2\alpha}e^t||\,dt|\\ &\leq\ C(M_m,B)\theta^{1-2\alpha}(n\theta)^{2\alpha-1}\end{aligned} \tag{14.4.32}$$

for $n\theta\in[0,B]$. In view of (14.4.18), we obtain

$$|\Sigma_m|\ \leq\ M_m\theta^{1-2\alpha}[|T_1(n\theta)|+|T_2(n\theta)|] \tag{14.4.33}$$

for $n\theta\in[0,B]$, $m\ =\ 1,2,3,\cdots$.

Finally we consider the case when $n\theta\to+\infty$. First, we introduce a curve Γ_c depending on $n\theta$, which starts at $e^{-i\pi}\eta/\theta$, moves to $e^{-i\pi}/n\theta$ along the lower edge of the negative real axis, encircles the origin along the circle $|s|\ =\ 1/n\theta$ in the positive sense, and then proceeds from $e^{i\pi}/n\theta$ to $e^{i\pi}\eta/\theta$ along the upper edge of

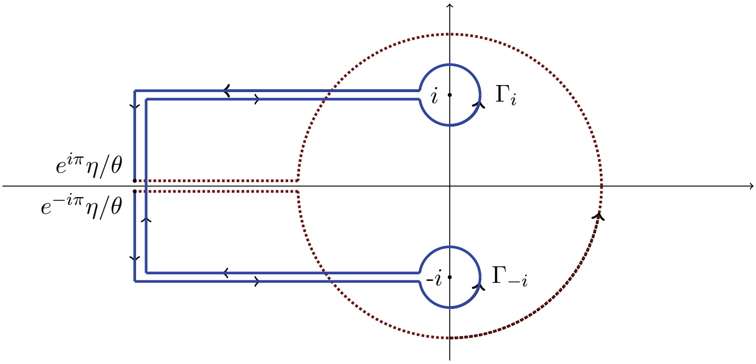

Fig. 14.8 Contour $\Gamma = \Gamma_i \cup \Gamma_{-i} \cup \Gamma_r$

the negative real line. We now deform the path of integration in (14.3.27), and split it into three parts: $\Gamma_i = \Gamma_c + i$, $\Gamma_{-i} = \Gamma_c - i$, and Γ_r, where Γ_r consists of three segments on $\mathrm{Re}\, s = -\eta/\theta$ connecting:

(i) $e^{i\pi}\eta/\theta - i$ and $e^{-i\pi}\eta/\theta + i$;
(ii) $e^{i\pi}\eta/\theta + i$ and $e^{i\pi}\eta/\theta$;
(iii) $e^{-i\pi}\eta/\theta - i$ and $e^{-i\pi}\eta/\theta$; see Figure 14.8.

We know from Lemma 14.4.7 that $h_m(s,\theta)$ is bounded on $\Gamma = \Gamma_i \cup \Gamma_{-i} \cup \Gamma_r$, and that the bound is uniform in $\theta \in [0, \pi - \delta]$. Consider

$$\begin{aligned} I_i &\equiv \frac{1}{2\pi i}\int_{\Gamma_i} h_m(s,\theta)(s^2+1)^{-\alpha} e^{n\theta s}\, ds \\ &= \frac{e^{in\theta}}{2\pi i}\int_{\Gamma_c} \{h_m(s+i,\theta)(s+2i)^{-\alpha}\} s^{-\alpha} e^{n\theta s}\, ds. \end{aligned}$$

In the last integral we put $v(s) \equiv h_m(s+i,\theta)(s+2i)^{-\alpha}$ and make the change of variable $t = n\theta s$. Since $v(s)$ is uniformly bounded on $\{\Gamma_c : \mathrm{Re}\, s \geq -3\}$, we have

$$\begin{aligned} &\left| \frac{e^{in\theta}}{2\pi i}\int_{\{\Gamma_c:\, \mathrm{Re}\, s\, \geq\, -3\}} v(s) s^{-\alpha} e^{n\theta s}\, ds \right| \\ &\qquad = (n\theta)^{\alpha-1} \left| \frac{e^{in\theta}}{2\pi i}\int_{-3n\theta}^{(0^+)} v(s(t)) t^{-\alpha} e^t\, dt \right| \\ &\qquad \leq \left\{ C\int_{-\infty}^{(0^+)} |t|^{-\alpha} e^{\mathrm{Re}\, t} |\, dt| \right\} (n\theta)^{\alpha-1}. \end{aligned}$$

On the other hand,

$$\left|\frac{e^{in\theta}}{2\pi i}\int_{\{\Gamma_c:\ \mathrm{Re}\, s\,\le\,-3\}} w(s)e^{n\theta s}\,ds\right| \le \frac{M_m C(\alpha)}{\pi}\int_3^{\eta/\theta} t^{-2\alpha}e^{-n\theta t}\,dt$$
$$\le Ce^{-3n\theta},$$

which is in turn bounded by $(n\theta)^{\alpha-1}$ for $n\theta \ge B$. The second to last inequality follows from the fact that $w(s) \equiv h_m(s+i,\theta)(s+2i)^{-\alpha}s^{-\alpha}$ is uniformly bounded by $C|s|^{-2\alpha}$ on that part of Γ_c. Hence

$$|I_i| \le C(n\theta)^{\alpha-1}. \tag{14.4.34}$$

Similarly, we have

$$|I_{-i}| \le C(n\theta)^{\alpha-1}. \tag{14.4.35}$$

The estimate of the integral I_r over Γ_r can be obtained by taking the absolute value of the integrand. Indeed, we have

$$|I_r| \le Ce^{-\eta n}\theta^{-2\alpha} \le C(n\theta)^{\alpha-1}. \tag{14.4.36}$$

To obtain the last inequality, we have used the fact that $\theta \in [B/n, \pi]$ for $n\theta \in [B, \infty)$. A combination of (14.4.34), (14.4.35), (14.4.36), and the fact that $\Sigma_m = \theta^{1-2\alpha}(I_i + I_{-i} + I_r)$ gives

$$|\Sigma_m(\theta)| \le C\theta^{1-2\alpha}(n\theta)^{\alpha-1} \le C\theta^{1-2\alpha}\{|T_1(n\theta)| + |T_2(n\theta)|\} \tag{14.4.37}$$

for $n\theta \in [B, +\infty)$, B being sufficiently large, $m = 1, 2, 3, \cdots$; see (14.4.19). The results in (14.4.33) and (14.4.37) imply that there exists a constant C such that

$$|\Sigma_m(\theta)| \le C\theta^{1-2\alpha}\{|T_1(n\theta)| + |T_2(n\theta)|\} \tag{14.4.38}$$

for all n and θ. The desired result (14.4.3) now follows from (14.4.20), (14.4.21), (14.4.30), (14.4.38), and (14.3.25).

Remark. To conclude this section, we note that in addition to the expression in (14.3.4), the approximant $T_2(x)$ in (14.4.1) can also be expressed as

$$T_2(n\theta) = \frac{(\alpha-\frac{1}{2})\sqrt{\pi}}{\Gamma(\alpha)}\left(\frac{n\theta}{2}\right)^{\alpha-\frac{3}{2}} J_{\alpha-\frac{1}{2}}(n\theta) - \frac{\sqrt{\pi}}{\Gamma(\alpha)}\left(\frac{n\theta}{2}\right)^{\alpha-\frac{1}{2}} J_{\alpha+\frac{1}{2}}(n\theta). \tag{14.4.39}$$

As a consequence, we have the following useful corollary.

Corollary 14.4.8. *Under the same assumption as Theorem 14.4.1, the following holds:*

$$a_n(\theta) = \left(\frac{n}{2\theta}\right)^{\alpha-\frac{1}{2}} J_{\alpha-\frac{1}{2}}(n\theta)\sum_{k=0}^{m-1}\frac{\tilde{\alpha}_k(\theta)}{n^k}$$
$$+\left(\frac{n}{2\theta}\right)^{\alpha-\frac{1}{2}} J_{\alpha+\frac{1}{2}}(n\theta)\sum_{k=0}^{m-1}\frac{\tilde{\beta}_k(\theta)}{n^k} + \tilde{\varepsilon}(\theta, m). \tag{14.4.40}$$

With α_k, β_k, and $\varepsilon(\theta, m)$ replaced by $\tilde{\alpha}_k$, $\tilde{\beta}_k$, and $\tilde{\varepsilon}(\theta, m)$, respectively, the estimates in (14.4.2 and (14.4.3) remain valid.

Indeed, inserting (14.3.4) and (14.4.39) into (14.4.1) gives, explicitly,

$$\begin{aligned}
\tilde{\alpha}_0(\theta) &= \sqrt{\pi}\alpha_0(\theta)/\Gamma(\alpha),\\
\tilde{\alpha}_k(\theta) &= (\sqrt{\pi}/\Gamma(\alpha))[\alpha_k(\theta) + (2\alpha - 1)\beta_{k-1}(\theta)/\theta], \qquad k = 1, 2, 3, \cdots,\\
\tilde{\beta}_k(\theta) &= -\sqrt{\pi}\beta_k(\theta)/\Gamma(\alpha), \qquad k = 0, 1, 2, \cdots, \qquad (14.4.41)
\end{aligned}$$

and

$$\tilde{\varepsilon}(\theta, m) = \varepsilon(\theta, m) + \frac{(2\alpha - 1)\sqrt{\pi}}{\Gamma(\alpha)} \left(\frac{n}{2\theta}\right)^{\alpha - \frac{1}{2}} \frac{\beta_{m-1}(\theta)}{\theta} \frac{1}{n^m} J_{\alpha - \frac{1}{2}}(n\theta). \qquad (14.4.42)$$

The above result is immediately applicable to the ultraspherical polynomials $P_n^{(\lambda)}(x)$ defined by

$$[(e^{i\theta} - z)(e^{-i\theta} - z)]^{-\lambda} = \sum_{n=0}^{\infty} P_n^{(\lambda)}(\cos\theta) z^n. \qquad (14.4.43)$$

With $\alpha = \lambda$ and $a_n(\theta) = P_n(\cos\theta)$, one can immediately write down the asymptotic expansion

$$\begin{aligned}
P_n^{(\lambda)}(\cos\theta) \sim \left(\frac{n}{2\theta}\right)^{\lambda - \frac{1}{2}} J_{\lambda - \frac{1}{2}}(n\theta) \sum_{k=0}^{\infty} \frac{\tilde{\alpha}_k(\theta)}{n^k}&\\
+ \left(\frac{n}{2\theta}\right)^{\lambda - \frac{1}{2}} J_{\lambda + \frac{1}{2}}(n\theta) \sum_{k=0}^{\infty} \frac{\tilde{\beta}_k(\theta)}{n^k}& \qquad (14.4.44)
\end{aligned}$$

uniformly for $\theta \in [0, \pi - \delta]$, $\delta > 0$. Here

$$\tilde{\alpha}_0(\theta) = \frac{\sqrt{\pi}}{\Gamma(\lambda)} \left(\frac{\sin\theta}{\theta}\right)^{-\lambda} \cos\lambda\theta, \quad \tilde{\beta}_0(\theta) = -\frac{\sqrt{\pi}}{\Gamma(\lambda)} \left(\frac{\sin\theta}{\theta}\right)^{-\lambda} \sin\lambda\theta,$$

$$\begin{aligned}
\tilde{\alpha}_1(\theta) = \frac{\sqrt{\pi}}{\Gamma(\lambda)}\lambda \left(\frac{\sin\theta}{\theta}\right)^{-\lambda} \Bigg\{ \frac{\lambda - 1}{2}\Bigg[-\frac{\theta\cos\theta - \sin\theta}{\theta\sin\theta} \sin\lambda\theta&\\
+2\cos\lambda\theta\Bigg] + \frac{\sin\lambda\theta}{\theta}\Bigg\},&
\end{aligned}$$

and

$$\tilde{\beta}_1(\theta) = -\frac{\sqrt{\pi}}{\Gamma(\lambda)}\frac{1}{2}\lambda(\lambda-1)\left(\frac{\sin\theta}{\theta}\right)^{-\lambda}\left[\frac{\theta\cos\theta-\sin\theta}{\theta\sin\theta}\cos\lambda\theta+2\sin\lambda\theta\right].$$

Setting $\lambda = \frac{1}{2}$, the above result reduces to a uniform asymptotic expansion for the Legendre polynomials defined in (14.0.2).

14.5 Heisenberg polynomials

The Heisenberg polynomials are polynomials in z and $\bar{z}$, defined by

$$C_n^{(\alpha,\beta)}(z) = \sum_{j=0}^{n}\frac{(\alpha)_j(\beta)_{n-j}}{j!(n-j)!}\bar{z}^j z^{n-j}, \qquad n = 0,1,2,\cdots, \tag{14.5.1}$$

where α and β are real numbers, and $(\gamma)_k$ is the Pochhammer symbol defined by $(\gamma)_0 = 1$ and $(\gamma)_k = \gamma(\gamma+1)\cdots(\gamma+k-1)$. This representation can be derived from the generating function

$$(1-w\bar{z})^{-\alpha}(1-wz)^{-\beta} = \sum_{n=0}^{\infty}C_n^{(\alpha,\beta)}(z)w^n, \qquad |wz| < 1. \tag{14.5.2}$$

The notation $C_n^{(\alpha,\beta)}(z)$ was used by Gasper [86] in the sense of (14.5.1) and (14.5.2), but the term Heisenberg polynomials was first used by Dunkl [63].

From (14.5.2), it is readily seen that the Heisenberg polynomials have the property $C_n^{(\alpha,\beta)}(\rho e^{i\theta}) = \rho^n C_n^{(\alpha,\beta)}(e^{i\theta})$. Hence, to study the behavior of these polynomials as $n \to \infty$, it suffices to consider the polynomials on the unit circle. The generating function in (14.5.2) now takes the form

$$(e^{i\theta}-z)^{-\alpha}(e^{-i\theta}-z)^{-\beta}e^{i\theta(\alpha-\beta)} = \sum_{n=0}^{\infty}C_n^{(\alpha,\beta)}(e^{i\theta})z^n. \tag{14.5.3}$$

Note that the exponents of the two factors on the left-hand side of the above equation are different; hence, the result of Theorem 14.4.1 is not directly applicable to Heisenberg polynomials. However, the arguments in the last two sections can be modified to deal with the current situation. To this end, we define

$$T_l(x) = \frac{1}{2\pi i}\int_{\Gamma_0}s^{l-1}(s-i)^{-\beta}(s+i)^{-\alpha}e^{xs}\,ds, \quad l = 1,2, \tag{14.5.4}$$

where Γ_0 is a Hankel-type loop which starts and ends at $-\infty$ and encircles $s = \pm i$ in the positive sense; cf. (14.3.3). We further introduce an auxiliary function

$$h_0(s,\theta) = e^{i\theta(\alpha-\beta)}\left[\frac{e^{i\theta}-e^{-\theta s}}{(s+i)\theta}\right]^{-\alpha}\left[\frac{e^{-i\theta}-e^{-\theta s}}{(s-i)\theta}\right]^{-\beta} \tag{14.5.5}$$

and a sequence $\{h_k(s,\theta)\}_{k=0}^{\infty}$ defined by

$$h_k(s,\theta) = \alpha_k(\theta)+s\beta_k(\theta)+(s^2+1)g_k(s,\theta), \tag{14.5.6a}$$

$$h_{k+1}(s,\theta) = \frac{s^2+1}{\theta}\left\{\frac{\alpha-1}{s+i}+\frac{\beta-1}{s-i}-\frac{d}{ds}\right\}g_k(s,\theta) \tag{14.5.6b}$$

for $k=0,1,2,\cdots$; cf. (14.3.13), (14.3.14), (14.3.22), and (14.3.23). The coefficients $\alpha_k(\theta)$ and $\beta_k(\theta)$ are determined by requiring all $h_k(s,\theta)$ and $g_k(s,\theta)$ to be holomorphic in $D=\{s : \mathrm{Re\,s} \ge -\frac{\eta}{\theta}, |\mathrm{s}\pm \mathrm{i}| < \frac{2\pi}{\theta}\}$; see Fig. 4.5.

Theorem 14.5.1. *For $\theta \in [0,\pi-\delta]$ with arbitrary $\delta > 0$, we have*

$$C_n^{(\alpha,\beta)}(e^{i\theta}) = \theta^{1-\alpha-\beta}T_1(n\theta)\sum_{k=0}^{m-1}\frac{\alpha_k(\theta)}{n^k} + \theta^{1-\alpha-\beta}T_2(n\theta)\sum_{k=0}^{m-1}\frac{\beta_k(\theta)}{n^k}+\varepsilon_{\theta,m}, \tag{14.5.7}$$

where $|\alpha_k(\theta)| \le M_k$, $|\beta_k(\theta)/\theta| \le M_k$ for $k=0,1,2,\cdots$, and

$$|\varepsilon_{\theta,m}| \le M_m\theta^{1-\alpha-\beta}n^{-m}\{|T_1(n\theta)|+|T_2(n\theta)|\}, \tag{14.5.8}$$

for $m = 1,2,\cdots$. The positive constants M_k, $k=1,2,\cdots$, are independent of θ for $\theta \in [0,\pi-\delta]$. The coefficients $\alpha_k(\theta)$ and $\beta_k(\theta)$ are given by (14.5.6), with

$$\alpha_0(\theta) = \frac{e^{i\theta\alpha}}{2}\left(\frac{\sin\theta}{\theta}\right)^{-\alpha}+\frac{e^{-i\theta\beta}}{2}\left(\frac{\sin\theta}{\theta}\right)^{-\beta},$$
$$\beta_0(\theta) = \frac{e^{i\theta\alpha}}{2i}\left(\frac{\sin\theta}{\theta}\right)^{-\alpha}-\frac{e^{-i\theta\beta}}{2i}\left(\frac{\sin\theta}{\theta}\right)^{-\beta}.$$

By expanding the slowly varying factor in the integrand of (14.5.4) in a uniformly convergent power series of $1/s$ and integrating term by term, we obtain

$$T_1(x) = x^{\alpha+\beta-1}\sum_{k=0}^{\infty}\sum_{l=0}^{k}\frac{i^k(-1)^l(\alpha)_l(\beta)_{k-l}}{l!(k-l)!\Gamma(\alpha+\beta+k)}x^k, \tag{14.5.9}$$

where $x^{\alpha+\beta-1}$ is positive for real positive x. It is easily seen that $x^{-\alpha-\beta+1}T_1(x)$ is an entire function. From (14.5.4), it is also readily verified that $T_2(x) = T_1'(x)$ in the cut plane $\mathbb{C}\setminus(-\infty,0]$. Moreover, with $a=2-\alpha-\beta$ and $b=\beta-\alpha$, $T_1(x)$ satisfies the differential equation

$$xT_1''+aT_1'+(x-bi)T_1 = 0. \tag{14.5.10}$$

Furthermore, making the change of variables

$$T_1(x) = x^{1-\alpha} e^{ix} y(x), \qquad z = -2ix \tag{14.5.11}$$

yields the confluent hypergeometric equation

$$z\frac{d^2y}{dz^2} + [(2-a)-z]\frac{dy}{dz} - \frac{2-a-b}{2}y = 0. \tag{14.5.12}$$

Taking into account the first two terms of the infinite series in (14.5.9), one obtains

$$T_1(x) = \frac{1}{\Gamma(\alpha+\beta)} x^{\alpha+\beta-1} e^{ix} M(\alpha, \alpha+\beta, -2ix), \tag{14.5.13}$$

where M is the Kummer function [22], p.201. Since $T_2(x) = T_1'(x)$, it also follows

$$\begin{aligned} T_2(x) = & \frac{(\alpha+\beta-1+ix)x^{\alpha+\beta-2}e^{ix}}{\Gamma(\alpha+\beta)} M(\alpha, \alpha+\beta, -2ix) \\ & - \frac{2ix^{\alpha+\beta-1}e^{ix}}{\Gamma(\alpha+\beta)} M'(\alpha, \alpha+\beta, -2ix), \end{aligned} \tag{14.5.14}$$

where $M'(\gamma, \delta, z) = \frac{d}{dz}M(\gamma, \delta, z)$. Substituting (14.5.13) and (14.5.14) into (14.5.7), we obtain an asymptotic expansion of the Heisenberg polynomials in terms of the Kummer function.

Theorem 14.5.2. *Assume that α and β are real and fixed and that $z = \rho e^{i\theta}$ with $\rho > 0$ and θ real. Then we have the compound asymptotic expansion [158], p.118:*

$$\begin{aligned} C_n^{(\alpha,\beta)}(z) \sim n^{\alpha+\beta-1} z^n \Bigg\{ & M(\alpha, \alpha+\beta, -2in\theta) \sum_{k=0}^{\infty} \frac{c_k(\theta)}{n^k} \\ & + M'(\alpha, \alpha+\beta, -2in\theta) \sum_{k=0}^{\infty} \frac{d_k(\theta)}{n^k} \Bigg\} \end{aligned} \tag{14.5.15}$$

as $n \to \infty$, uniformly with respect to $\rho \in (0, \infty)$ and $\theta \in [0, \pi - \delta]$, where $0 < \delta \leq \pi$. The coefficients are given by

$$c_k(\theta) = \frac{\alpha_k(\theta) + i\beta_k(\theta)}{\Gamma(\alpha+\beta)} + \frac{\beta_k(\theta)/\theta}{\Gamma(\alpha+\beta-1)}, \quad d_k(\theta) = -\frac{2i\beta_k(\theta)}{\Gamma(\alpha+\beta)} \tag{14.5.16}$$

for $k = 0, 1, 2, \cdots$, $\alpha_k(\theta)$ and $\beta_k(\theta)$ being defined in (14.5.6). In particular,

$$\begin{aligned} c_0(\theta) &= \frac{e^{i\theta\alpha}}{\Gamma(\alpha+\beta)}\left(\frac{\sin\theta}{\theta}\right)^{-\alpha}, \\ d_0(\theta) &= \frac{e^{-i\theta\beta}}{\Gamma(\alpha+\beta)}\left(\frac{\sin\theta}{\theta}\right)^{-\beta} - \frac{e^{i\theta\alpha}}{\Gamma(\alpha+\beta)}\left(\frac{\sin\theta}{\theta}\right)^{-\alpha}. \end{aligned} \tag{14.5.17}$$

When the parameters α and β in the above theorem are non-positive integers, the coefficients c_k and d_k in (14.5.15) all vanish; see (14.5.16). However the asymptotic relation remains valid, since the polynomials also vanish for large value of n, i.e. it is a trivial result.

The results in this section are taken from [135]. For a proof of Theorem 14.5.1, we refer to [135].

Exercises

1. Use Stirling's approximation (2.10.5) to prove (Re 4.1.19a).
2. Prove the Riemann–Lebesgue lemma: if $\int_{-\infty}^{\infty} |f(\theta)|\,d\theta < \infty$, then

   $$\lim_{n\to\infty} e^{in\theta} f(\theta)\,d\theta = 0.$$

   Hint: by a density argument, it is enough to verify this for the characteristic function of a bounded interval.
3. Prove the two identities in (14.3.4).
4. For $n = 1, 2, 3, \ldots$, let

   $$a_n = 1 + \frac{1-n}{1} + \frac{(2-n)^2}{2} + \cdots + \frac{(-1)^{n-1}}{n-1}.$$

   Find the generating function of the sequence $\{a_k\}$ and show that $a_n \to 0$ as $n \to \infty$.
5. Let q_n denote the probability that in n tosses of an ideal coin, no run of three consecutive heads appears. Clearly $q_0 = q_1 = q_2 = 1$, and in probability theory it is established that $q_n = \frac{1}{2}q_{n-1} + \frac{1}{4}q_{n-2} + \frac{1}{8}q_{n-3}$. Show that the $\{q_n\}$ have generating function

   $$\sum_{n=0}^{\infty} q_n t^n = \frac{2t^2 + 4t + 8}{8 - 4t - 2t^2 - t^3}$$

   and deduce the asymptotic formula

   $$q_n \sim \frac{1.2368398446}{(1.0873780254)^{n+1}} \quad \text{as } n \to \infty.$$

   (See Feller [70], p.278.)
6. The *Charlier polynomials* $C_n^{(a)}(x)$ can be defined by the generating function

$$e^{-aw}(1+\omega)^x = \sum_{n=0}^{\infty} C_n^{(a)}(x)\frac{\omega^n}{n\,!}.$$

For fixed $x > 0$, find an asymptotic expansion for $C_n^{(a)}(x)$ as $n \to \infty$. Write out the first two terms of the expansion. (See [20], equations (5.3.6) and (5.3.12).)

7. The *Meixner polynomials* $M_n(x;\,\beta, c)$ have generating function

$$\left(1-\frac{\omega}{c}\right)^x(1-\omega)^{-x-\beta} = \sum_{n=0}^{\infty} m_n(x;\,\beta, c)\,\frac{\omega^n}{n\,!}.$$

For fixed $x > 0$, find an asymptotic expansion for $m_n(x;\,\beta, c)$ as $n \to \infty$. Calculate the coefficients of the two leading terms. (See [20], equation (5.5.7) and Exercise 10.21, pp.262–264.)

8. The *Stirling numbers of the first kind* S_n^m, $m, n = 0, 1, 2, \ldots$, have generating function

$$[\log(1+t)]^m = \sum_{n=0}^{\infty} m!\,S_n^m\,\frac{t^n}{n\,!}.$$

Use the results in Section 14.2 to write out an asymptotic expansion for S_n^m as $n \to \infty$. (This formula was first given in Jordan [117].)

9. The *Jacobi polynomials* $P_n^{(\alpha,\beta)}$ have the generating function

$$\sum_{n=0}^{\infty} P_n^{(\alpha,\beta)}\,t^n = 2^{\alpha+\beta}R^{-1}(1-t+R|^{-\alpha}(1+t+R)^{-\beta},$$

where $R = R(z, t) = (1 - 2z\,t - t^2)^{-1/2}$. The branch of the square root is chosen so that $R(z, 0) = 1$. Use Darboux's method to show that:

(a) For $\theta \in [\varepsilon, \pi - \varepsilon]$, $\varepsilon > 0$, we have

$$P_n^{(\alpha,\beta)}(\cos\theta) = n^{-1/2}k(\theta)\cos(N\theta+\gamma) + O(n^{-3/2}),$$

where

$$k(\theta) = \pi^{-1/2}(\sin\tfrac{1}{2})^{-\alpha-\frac{1}{2}}(\cos\tfrac{1}{2}\theta)^{-\beta-\frac{1}{2}}$$
$$N = n + \tfrac{1}{2}(\alpha+\beta+1), \qquad \gamma = -\tfrac{1}{2}\pi(\alpha+\tfrac{1}{2}),$$

and the $O(n^{-3/2})$ term is uniform on the interval $[-\varepsilon, \theta - \varepsilon]$. (See [20], equations (4.6.7) and (4.6.11).)

(b) For fixed $x > 1$ we have

$$P_n^{(\alpha,\beta)}(x) \sim (x-1)^{-\alpha/2}(x+1)^{-\beta/2}[\sqrt{x-1}+\sqrt{x+1}]^{\alpha+\beta}$$
$$+(x^2-1)^{-1/4}(2\pi n)^{-1/2}[x+\sqrt{x^2-1}]^{n+\frac{1}{2}}.$$

10. The Heisenberg polynomials $C_n^{(\alpha,\beta)}$ have the generating function given in (14.5.2):

$$(1-\omega\,\bar{z})^{-\alpha}(1-\omega\,z)^{-\beta} = \sum_{n=0}^{\infty} C_n^{(\alpha,\beta)}(z)\,\omega^n, \qquad |\omega z| < 1.$$

As noted at the beginning of Section 14.5, the homogeneity property of these polynomials allows us, in studying asymptotics, to restrict attention to the unit circle and thus to the restricted generating function (14.5.3):

$$(e^{i\theta}-z)^{-\alpha}(e^{-i\theta}-z)^{-\beta}e^{i\theta(\alpha-\beta)} = \sum_{n=0}^{\infty} C_n^{(\alpha,\beta)}(e^{i\theta})\,z^n.$$

By modifying the arguments given in this chapter, prove the result stated in Theorem 14.5.1.

Remarks and further reading

The original presentation of Darboux's method was in Darboux [51]. The discussion at the start of this chapter is mainly based on Carrier, Krook, and Pearson [43].

Since its inception, Darboux's method has been used extensively to compute asymptotics of orthogonal polynomials. An excellent source for results of this type is Ismail [114].

For some recent applications, see Bai and Zhao [15] and Wang and Zhao [212]. For other recent developments, in addition to the work described in this chapter, see Flagolet et al. [74], Temme [204], Boyd [30].

References

1. Abikoff, W.: The Real Analytic Theory of Teichmüller Spaces. Springer, New York, Heidelberg, Berlin (1980)
2. Abikoff, W.: The uniformization theorem. Am. Math. Mon. **88**, 574–592 (1981)
3. Ahlfors, L.V.: On quasiconformal mapping. J. d'Analyse **3**, 1–58 (1953/54)
4. Ahfors, L.V.: Quasiconformal reflections. Acta Math. **109**, 291–301 (1963)
5. Ahlfors, L.V.: Lectures on Quasiconformal Mappings, 2nd edn. American Mathematical Society, Providence (2006)
6. Ahlfors, L.V.: Complex Analysis, 3rd edn. McGraw-Hill, New York (1978)
7. Ahlfors, L.V.: Conformal Invariants. McGraw-Hill, New York (1973)
8. Aitken, A.C.: On Bernoulli's numerical solution of algebraic equations. Proc. R. Soc. Edinb. **46**, 289–305 (1925)
9. Akhiezer, N.I.: The Classical Moment Problem and Some Related Questions in Analysis. Reprint of the 1965 edition. SIAM, Philadelphia (2021)
10. Anosov, D.V., Bolibruch, A.A.: The Riemann-Hilbert Problem. Vieweg, Braunschweig, Wiesbaden (1994)
11. Aronszajn, N.: Theory of reproducing kernels. Trans. Am. Math. Soc. **68**, 337–404 (1950)
12. Askey, R., Gasper, G.: Positive Jacobi polynomial sums II. Am. J. Math. **98**, 709–737 (1976)
13. Axler, S., Bourdon, P., Ramey, W.: Harmonic Function Theory, 2nd edn. Springer, New York, Heidelberg, Berlin (2001)
14. Baerenstein, A., et al.: The Bieberbach Conjecture. American Mathematical Society, Providence (1986)
15. Bai, X., Zhao, Y.: A uniform asymptotic expansion for Jacobi polynomials via uniform treatment of Darboux's method. J. Approx. Theory **148**, 1–11 (2007)
16. Baker, G.A., Jr., Graves-Morris, P.: Padé Approximants, 2nd edn. Cambridge University Press, Cambridge (1996)
17. Baker, H.F.: Abelian Functions. Abel's Theorem and the Allied Theory of Theta Functions. Cambridge University Press, Cambridge (1897, reissued 1995)
18. Bazilevich, I.E.: On distortion theorems in the theory of univalent functions. Mat. Sb. **28**, 147–164 (Russian) (1951)
19. Beals, R., Deift, P., Zhou, X.: The inverse scattering transform on the line. Important Developments in Soliton Theory, pp. 7–32. Springer, Berlin (1993)
20. Beals, R., Wong, R.S.C.: Special Functions: A Graduate Text. Cambridge University Press (2010)

R. Beals and R. S. C. Wong, *More Explorations in Complex Functions*, Graduate Texts
in Mathematics 298, https://doi.org/10.1007/978-3-031-28288-1

21. Beals, R., Wong, R.S.C.: Special Functions and Orthogonal Polynomials. Cambridge University Press (2016)
22. Beals, R., Wong, R.S.C.: Explorations in Complex Functions. Springer, New York, Heidelberg, Berlin (2020)
23. Beardon, A.F.: Iteration of Rational Functions. Springer, New York, Heidelberg, Berlin (1991)
24. Bergman, S.: The Kernel Function and Conformal Mapping, 2nd edn. American Mathematical Society, Providence (1970)
25. Bers, L.: Quasiconformal mappings and Teichmüller's theorem, Analytic Functions, pp. 89–119. Princeton University Press, Princeton (1960)
26. Beurling, A., Ahlfors, L.V.: The boundary correspondence under quasiconformal mapping. Acta Math. **96**, 125–142 (1956)
27. Bieberbach, L.: Über die Koeffizienten derjenigen Potenzreihen, welche eine schlichte Abbildung des Einheitskreise vermitteln. S.-B. Preuss. Akad, Wiss., 940–955 (1916)
28. Boettcher, L.E.: The principal laws of convergence of iterates and their applications to analysis (Russian). Izv. Kazan. Fiz-Mat. Obshch. **14**, 155–234 (1904)
29. Bolibrukh, A.A.: The Riemann-Hilbert problem on the complex projective line. (Russian) Mat. Zametki **46**, 118–120 (1989)
30. Boyd, J.P.: The breakdown of Darboux's principle and natural boundaries for a function periodised from a Ramanujan Fourier transform pair. East Asian J. Appl. Math. **9**, 409–423 (2019)
31. Brezinski, C., Redivo-Zaglia, M.: Extrapolation and Rational Approximation. The Works of the Main Contributors. Springer, New York, Heidelberg, Berlin (2020)
32. Brezinski, C., Redivo-Zaglia, M.: A survey of Shanks' extrapolation methods and their applications. Comput. Math. Math. Phys.**61** (2021)
33. Brooks, R., Matelski, J.P.: The dynamics of 2-generator subgroups of $PSL(2,\mathbb{C})$. In: Riemann Surfaces and Related Topics: Proceedings of the 1978 Stony Brook Conference. Ann. Math. Stud. **97**. Princeton University Press, Princeton (1978)
34. Brjuno, A.D.: On convergence of transforms of differential equations to the normal form. (Russian) Dokl. Akad. Nauk SSSR **165**, 987–989 (1965)
35. Bunke, U., Olbrich, M.: Selberg Theta and Eta Functions. A Differential Operator Approach. Akademie-Verlag, Berlin (1995)
36. Burger, M., Iozzi, A., Labourie, A., Wienhard, A.: Maximal representations of surface groups: symplectic Anosov structures. Pure Appl. Math. Q. 1. Special Issue: In memory of Armand Borel, 543590 (2005)
37. Calderón, A.P., Zygmund, A.: On the existence of certain singular integrals. Acta Math. **88**, 85–1339 (1952)
38. Carathéodory, C.: Untersuchungen über die konformen Abbildungen von festen und veränderlichen Gebieten. Math. Ann. **72**, 107–144 (1912)
39. Carathéodory, C.: Zur Rñadezuordnung bei konformer Abbildung. Nachr. Königl. Ges. Wiss. Göttingen, Math.-Phys. Kl. 509–518 (1913)
40. Carleman, T.: Sur la résolution de certaines équations intégrales. Ark. Mat. Astronom. Fys. **16**, 1–19 (1921)
41. Carleman, T.: La théorie des équations intégrales singuliéres et ses applications. Ann. Inst. Henri Poincaré **1**, 401–430 (1930)
42. Carleson, L., Gamelin, T.W.: Complex Dynamics. Springer, New York, Heidelberg, Berlin (1993)
43. Carrier, G.F., Krook, M., Pearson, C.E.: Functions of a Complex Variable. McGraw-Hill, New York (1966)
44. Christ, M.: Lectures on Singular Integral Operators. American Mathematical Society, Providence (1990)
45. Clancey, K.F.: Gohberg: Factorization of Matrix Functions and Singular Integral Operators. Birkhüser, Boston (1981)
46. Cohn, H.: Conformal Mapping on Riemann Surfaces. Reprint of the 1967 edition. Dover, New York (1980)

47. Conway, J.B.: Functions of One Complex Variable II. Springer, New York, Heidelberg, Berlin (1995)
48. Cooper, S.: Ramanujan's Theta Functions. Springer, Cham (2017)
49. Cremer, H.: Über die Häufigkeit der Nichzentren. Math. Ann. **115**, 573–580 (1938)
50. Cuyt, A., Petersen, V.B., Verdonk, B., Waadeland, H., Jones, W.B.: Handbook of Continued Fractions for Special Functions. Springer, New York, Heidelberg, Berlin (2008)
51. Darboux, G.: Mémoire sur l'approximation des fonctions de très grands nombres, et sur une classe étendue de développements en série. J. Math. Pures Appl. **4**(5–56), 377–416 (1878)
52. de Branges, L.: A proof of the Bieberbach conjecture. Acta Math. **154**, 137–152 (1985)
53. Deift, P.A.: Orthogonal polynomials and random matrices: a Riemann-Hilbert approach. Courant Lecture Notes in Mathematics, 3. New York University, Courant Institute of Mathematical Sciences, New York, American Mathematical Society, Providence (1999)
54. Deift, P., Zhou, X.: A steepest descent method for oscillatory Riemann-Hilbert problems. Asymptotics for the MKdV equation. Ann. Math. **137**, 295–368 (1993)
55. Dieudonné, J.: Sur les fonctions univalentes. C. R. Acad. Sci. Paris **192**, 1148–1150 (1931)
56. Donaldson, S.: Riemann Surfaces. Oxford University Press, Oxford (2011)
57. Douady, A., Earle, C.: Conformally natural extensions of homeomorphisms of the circle. Acta Math. **157**, 23–48 (1986)
58. Douady, A., Hubbard, J.H: Itération des polynômes quadratiques complexes. Ann. Sci. École Norm. Sup. **18** (1982)
59. Douady, A., Hubbard, J.H.: The dynamics of polynomial-like mappings. Ann. Sci. École Norm. Sup. Paris **18**, 287–343 (1982)
60. Dubrovin, B.A.: Theta functions and nonlinear equations. Uspekhi Mat. Nauk **36**(2), 11–80 (1981), Russ. Math. Surv. **36**(2), 11–92 (1981)
61. Duistermaat, J.J., Kolk, J.A.C.: Distributions. Theory and applications. Birkhüser, Boston (2010)
62. Dumas, D., Sanders, A.: Geometry of compact complex manifolds associated to generalized quasi-Fuchsian representations. Geom. Topol. **24**, 1615–1693 (2020)
63. Dunkl, C.F.: The Poisson kernel for Heisenberg polynomials on the disk. Math. Z. **187**, 527–547 (1984)
64. Duren, P.: Univalent Functions. Springer, New York, Heidelberg, Berlin (1983)
65. Duren, P., Schuster, A.: Bergman Spaces. American Mathematical Society, Providence (2004)
66. Erdélyi, A.: Uniform asymptotic expansion of integrals. Analytic Methods in Mathematical Physics, pp. 149–168. Gordon and Breach, New York (1970)
67. Farkas, H.M., Kra, I.: Riemann Surfaces, 2nd edn. Springer, New York, Heidelberg, Berlin (1992)
68. Farkas, H.M., Kra, I.: Theta Constants, Riemann Surfaces, and the Modular Group. Springer, New York, Heidelberg, Berlin (2001)
69. Fatou, P.: Sur les équations fonctionelles. Bull. Doc. Math. Fr. **47** 161–271, **48**, 33–94, 208–314 (1919-20)
70. Feller, W.: An Introduction to Probability Theory and Its Applications, vol. I, 3rd edn. Wiley, New York, London, Sydney (1968)
71. Fields, L.: A uniform treatment of Darboux's method. Arch. Rat. Mech. Anal. **27**(1967), 289–305 (1967)
72. FitzGerald, C.H.: Quadratic inequalities and coefficient estimates for Schlicht functions. Arch. Rat. Mech. Anal. **46**, 356–368 (1972)
73. FitzGerald, C.M., Pommerenke, C.: The de Branges theorem on univalent functions. Trans. Am. Math. Soc. **290**, 683–690 (1985)
74. Flajolet, P., Fusy, E., Gourdon, X., Panario, D., Pouyanne, N.: A hybrid of Darboux's method and singularity analysis in combinatorial asymptotics. Electron. J. Comb. **13**(1), Research Paper 103, 35 pp (2006)
75. Flajolet, P., Sedgewick, R.: Analytic Combinatorics. Cambridge University Press, Cambridge (2009)

76. Fletcher, A., Markovic, V.: Quasiconformal Maps and Teichmüller Theory. Oxford University Press (2007)
77. Fock, V., Goncharov, A.: Moduli spaces of local systems and higher Teichmüller theory. Publ. Math. I. H. É. S. **103**, 1–211 (2006)
78. Folland, G.B.: Real Analysis: Modern Techniques and Their Applications, 2nd edn. Wiley, New York (1999)
79. Fricke, R., Klein, F.: Vorlesung über die Theorie der automorphen Funktionen. Teubner, Stuttgart (1926)
80. Frobenius, G.: Über Relationen zwischen den Näherungsbruchen von Potenzreihen. J. für Math. **90**, 1–17 (1881)
81. Garabedian, P.R., Schiffer, M.: A proof of the Bieberbach conjecture for the fourth coefficient. J. Rat. Mech. Anal. **4**, 427–465 (1955)
82. Gardiner, F.: Teichmüller Theory and Quadratic Differentials. Wiley, New York (1987)
83. Gardiner, F., Lakic, N.: Quasiconformal Teichmüller Theory. American Mathematical Society, Providence (2000)
84. Gardiner, F., Sullivan, D.: Symmetric structures on a closed curve. Am. J. Math. **114**, 683–736 (1992)
85. Garnett, J., Marshall, D.: Harmonic Measure. Cambridge University Press, Cambridge (2005)
86. Gasper, G.: Orthogonality of certain functions with respect to complex valued weights. Can. J. Math. **33**, 1261–1270 (1981)
87. Gehring, F.W.: Quasiconformal mappings in space. Bull. Am. Math. Soc. **69**, 146–164 (1963)
88. Gehring, F., Marten, G., Palka, B.: An Introduction to the Theory of Higher-Dimensional Quasiconformal Mappings. American Mathematical Society, Providence (2017)
89. Gelca, R.: Theta Functions and Knots. World Scientific, Hackensack, N. J (2014)
90. Georgiev, S.G.: Theory of Distributions, 2nd edn. Springer, Cham (2021)
91. Goluzin, G.M.: On the coefficients of univalent functions. Mat. Sb. **22**, 373–380 (1948)
92. Gray, J.: On the history of the Riemann mapping theorem. Rend. Circ. Mat. Palermo **2**(Suppl. 34), 47–94 (1994)
93. Gronwall, T.H.: Some remarks on conformal representation. Ann. Math. **16**, 72–76 (1914–1915)
94. Grötzsch, H.: Über möglichst konforme Abbildungen von schlichten Bereichen. Ner. Verh. Sächs. Akad. Wiss. Leipzig **84** (1932)
95. Handbook of Teichmüller Theory, A. Papadopoulos, ed. Eur. Math. Soc. Zürich, vols.1–VI (2007–2016)
96. Hardy, G.H., Wright, E.M.: An Introduction to the Theory of Numbers, 6th edn. Oxford University Press, Oxford (2008)
97. Hayman, W.K.: The asymptotic behavior of p-valent functions. Proc. Lond. Math. Soc. **5**, 257–284 (1955)
98. Hayman, W.K.: Subharmonic Functions, vol. 2. Academic Press, London (1989)
99. Hayman, W.K., Kennedy, P.B.: Subharmonic Functions, vol. 1. Academic Press, London, New York (1976)
100. Hedenmalm, H., Korenblum, B.: Theory of Bergman Spaces. Springer, New York, Heidelberg, Berlin (2000)
101. Heinonen, J., Koskela, P., Shanmugalingam, N., Tyson, J.T.: Sobolev Spaces on Metric Measure Spaces. An Approach Based on Upper Gradients. Cambridge University Press, Cambridge (2015)
102. Henrici, P.: Applied and Computational Complex Analysis, vol. III. Wiley, New York (1986)
103. Hensley, D.: Continued Fractions. World Scientific, Hackensack, NJ (2006)
104. Herglotz, G.: Über Potenzreihen mit positive reellen Teil im Einheitskreisen. Ber. Verh. Sachs. Akad. Wiss. Leipzig **63**, 501–511 (1911)
105. Herman, M.: Exemples de fractions rationelles ayant une orbite dense dans la sphère du Riemann. Bull. Soc. Math. Fr. **112**, 93–142 (1984)
106. Hilbert, D.: Mathematical problems. Bull. Am. Math. Soc. **8**, 437–479 (1902). Reprinted in Bull. Am. Math. Soc. **37**, 407–436 (2000)

107. Hille, E.: Analytic Function Theory, vol. II. Ginn & Co., Boston (1962)
108. Hitchin, N.: Lie groups and Teichmüller space. Topology **30**, 449–473 (1992)
109. Hörmander, L.: An Introduction to Complex Analysis in Several Variables, 3rd edn. North-Holland, Amsterdam (1990)
110. Horowitz, D.: A further refinement for coefficient estimates of univalent functions. Proc. Am. Soc. **71**, 217–221 (1978)
111. Hubbard, J.: Teichmüller Theory and Applications to Geometry, Topology and Dynamics, vol. 1. Matrix Editions, Ithaca, N.Y. (2006)
112. Hubbard, J.: Teichmüller Theory and Applications to Geometry, Topology and Dynamics, vol. 2. Matrix Editions, Ithaca, N.Y. (2016)
113. Hulek, K., Kahn, C., Weintraub, S.: Moduli Spaces of Abelian Surfaces: Compactification, Degenerations, and Theta Functions. De Gruyter, Berlin (1993)
114. Ismail, M.E.H.: Classical and Quantum Orthogonal Polynomials in One Variable. Cambridge University Press, Cambridge (2005)
115. Its, A.R.: The Riemann-Hilbert problem and integrable systems. Not. Am. Math. Soc. **50**, 1389–1400 (2003)
116. Iwaniec, T., Martin, G.: Geometric Function Theory and Non-linear Analysis. Oxford University Press, New York (2001)
117. Jordan, C.: The Calculus of Finite Differences, 2nd edn. Chelsea, New York (1947)
118. Julia, G.: Mémoire sur l'itération des fonctions rationlles. J. Math. Pures Appl. **8**, 47–245 (1918)
119. Kemppainen, A.: Schramm-Loewner Evolution. Springer, Cham (2017)
120. Khinchin, A.Y.: Continued Fractions. University of Chicago Press (1964)
121. Khrushchev, S.: Orthogonal Polynomials and Continued Fractions. From Euler's point of view. Cambridge University Press, Cambridge (2008)
122. Koebe, P.: Über die Uniformisierung beliebiger analytischen Kurven. Göttinger Nachr. 191–210 (1907)
123. Kœnigs, G.: Recherches sur les intégrales de certaines équations fonctionelles. Ann. Sci. ÉNS Paris **1**, supplém. 1–4 (1884)
124. Köhler, G.: Eta Products and Theta Series Identities. Springer, Heidelberg (2011)
125. Krantz, S.: Geometric Analysis of the Bergman Kernel. Springer, New York, Heidelberg, Berlin (2013)
126. Kraus, W.: Über der Zusammenhang einige Characteristiken eines einfach zusammenhängenden Bereiches mit der Kreisabbildung. Mitt. Math. Sem. Giessen **21**, 1–28 (1932)
127. Künzi, H.P.: Quasikonforme Abbildungen. Springer, Berlin (1960)
128. Labourie, F.: Anosov flows, surface groups and curves in projective space. Inven. Math. **165**, 51–114 (2006)
129. Lawler, G.F., Limic, V.: Random Walk: A Modern Introduction. Cambridge University Press, Cambridge (2010)
130. Leau, L.: Études sur les équations fonctionelles à une ou plusieurs variables. Ann. Fac. Sci. Toulouse **11**, E1–E110 (1897)
131. Lehto, O.: Univalent Functions and Teichmüller Spaces. Springer, New York, Heidelberg, Berlin (1987)
132. Lehto, O., Virtanen, K.I.: Quasiconformal Mappings in the Plane. Springer, New York, Heidelberg, Berlin (1973)
133. Lehner, J.: Discontinuous Groups and Automorphic Functions. American Mathematical Society, Providence (1964)
134. Littlewood, J.E.: On inequalities in the theory of functions. Proc. Lond. Math. Soc. **23**, 481–519 (1925)
135. Liu, S.-Y., Wong, R., Zhao, Y.-Q.: Uniform treatment of Darboux's method and the Heisenberg polynomials. Proc. Am. Math. Soc. **141**, 2683–2691 (2013)
136. Löwner, K.: Untersuchungen über schlichte konforme Abbildungen des Einheitskreises I. Math. Ann. **89**, 103–121 (1923)
137. Loewner, C.: On the conformal capacity in space. J. Math. Mech. **8**, 411–414 (1959)

138. Lyubich, M.Y.: Dynamics of rational transformations (Russian). Uspehi Mat. Nauk **41**, 35–95, 235, Russian Math. Surv. **41**, 43–117 (1986)
139. Mandelbrot, B.: Fractal aspects of the iteration of $z \to \zeta(\lambda - z)$ for complex λ, z. Ann. N. Y. Acad. Sci. **357**, 249–259 (1980)
140. Milin, I.M.: Estimation of coefficients of univalent functions. Dokl. Akad. Nauk SSSR **160**, 769–771 (1965), Soviet Math. Dokl. **6**, 196–198 (1965)
141. Milnor, J.: Dynamics in One Complex Variable. Princeton University Press, Princeton and Oxford (2006)
142. Mori, A.: On an absolute constant in the theory of quasiconformal mappings with prescribed complex dilatation. J. Math. Soc. Jpn. **8**, 156–166 (1956)
143. Morrey, C.B.: On the solution of quasilinear elliptic partial differential equations. Trans. Am. Math. Soc. **43**, 156–166 (1938)
144. Mostow, G.D.: Strong Rigidity of Locally Symmetric Spaces. Princeton University Press, Princeton, N.J., University of Tokyo Press, Tokyo (1973)
145. Mumford, D.: Tata Lectures on Theta. I. Reprint of the 1983 edition. Birkhüser, Boston (2007)
146. Mumford, D.: Tata Lectures on Theta. II. Reprint of the 1984 edition. Jacobian Theta Functions and Differential Equations. Birkhüser, Boston (2007)
147. Mumford, D.: Tata Lectures on Theta. III. Reprint of the 1991 original. Birkhüser, Boston (2007)
148. Murty, M.R. (ed.): Theta Functions: From the Classical to the Modern. American Mathematical Society, Providence (1993)
149. Muskhelishvili, N.I.: Singular Integral Equations. Nordhoff, NV, Groningen (1953)
150. Nag, S.: The Complex Analytic Theory of Teichmüller Spaces. Wiley, New York (1988)
151. Nehari, Z.: The Schwarzian derivative and schlicht functions. Bull. Am. Math. Soc. **55**, 545–551 (1949)
152. Nehari, Z.: Conformal Mapping. McGraw Hill, London (1952), Dover, New York (1975)
153. Neuenschwander, E.: Studies in the history of complex function theory II. Interactions among the French school, Riemann, and Weierstrass. Bull. Am. Math. Soc. **5**, 87–105 (1981)
154. Nevanlinna, R.: Über die konforme Abbildung von Sterngebieten. Översikt Fin = ska Vetenskaps-Soc. Förh **63A**(6), 1–21 (1920-21)
155. Newman, M.A.: Elements of the Topology of Plane Sets of Points. Cambridge University Press, Cambridge (1961)
156. Ohsawa, T.: Analysis of Several Complex Variables. American Mathematical Society, Providence (2002)
157. Olver, F.W.J.: A paradox in asymptotics. SIAM J. Math. Anal. **1**, 533–534 (1970)
158. Olver, F.W.J.: Asymptotics and Special Functions. Harcourt Brace Jovanovich, New York-London (1974)
159. Olver, F.W.J.: Unsolved problems in the asymptotic estimation of special functions. In: Askey, R. (ed.) Theory and Applications of Special Functions, pp. 99–142. Academic Press, New York (1975)
160. Olver, F.W.J., Lozier, D.W., Boisvert, R.F., Clark, C.W.: NIST Handbook of Mathematical Functions. Cambridge University Press, Cambridge (2010)
161. Ozawa, M.: On the Bieberbach conjecture for the sixth coefficient. Kōdai Math. Sem. Rep. **21**, 97–128 (1969)
162. Padé, H.: Sur la représentation approchée d'une fonction par des fractions rationelles, Thesis, Ann. École Norm. (3), **9**, 1–93 supplement (1892)
163. Pederson, R.N.: A proof of the Bieberbach conjecture for the sixth coefficient. Arch. Rat. Mech. Anal. **31**, 331–351 (1968-69)
164. Pederson, R.N., Schiffer, M.: A proof of the Bieberbach conjecture for the fifth coefficient. Arch. Rat. Mech. Anal. **45**, 161–193 (1972)
165. Pfeifer, G.A.: On the conformal mapping of curvilinear angles. The functional equation $\phi[f(x)] = \alpha_1 \phi(x)$. Trans. Am. Math. Soc. **18**, 185–198 (1917)
166. Perron, O.: Eine neue Behandlung der ersten Randwertaufgabe für $\Delta u = 0$. Math. Z. **18**, 42–54 (1923)

167. Peyrière, J.: An Introduction to Singular Integrals. Soc. Ind. Appl. Mathematics, Philadelphia, PA (2018)
168. Plemelj, J.: Problems in the Sense of Riemann and Klein. Interscience, New York (1964)
169. Pólya, G.: Mathematics and Plausible Reasoning, Vol. 1: Induction and Analogy in Mathematics. Princeton University Press, Princeton (1954)
170. Pommerenke, C.: Univalent Functions. Vandenhoeck und Ruprecht, Göttingen (1975)
171. Pommerenke, C.: Boundary Behavior of Conformal Maps. Springer, New York, Heidelberg, Berlin (1992)
172. Protter, M.H., Weinberger, H.F.: Maximum Principles in Differential Equations. Springer, New York, Heidelberg, Berlin (1984)
173. Pucci, P., Serrin, J.: The Maximum Principle. Birkhäuser, Basel (2007)
174. Prym, F.E.: Zur Integration der Differentialgleichung $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$. J. Math. **73**, 340–364 (1871)
175. Riemann, B.: Theorie der Abel'schen Functionen. J. Reine Angew. Math. **54**, 115–155 (1857). Collected Works, 2nd edn, pp. 88–144. Dover, New York (1953)
176. Riesz, F.: Sur certains systèmes singuliers d'équations inténtegrales. Ann. Sci. Éc. Norm. Sup. **28**, 33–62 (1911)
177. Robinson, R.: A new absolute geometric constant? Am. Math. Mon. **58**, 442–469 (1951)
178. Robinson's constant. Am. Math. Mon. **59**, 296–297 (1952)
179. Rodin, Y.L.: Generalized Analytic Functions on Riemann Surfaces. Springer, New York, Heidelberg, Berlin (1987)
180. Rogosiniski, W.: Über positive harmonische Entwicklungen und typisch-reelle Potenzreihen. Math. Z. **35**, 93–121 (1932)
181. Rosenblum, M., Rovnyak, J.: Topics in Hardy Classes and Univalent Functions. Birkhäuser, Basel (1994)
182. Rudin, W.: Functional Analysis, 2nd edn. McGraw-Hill, New York (1991)
183. Sauer, T.: Continued Fractions and Signal Processing. Springer, New York, Heidelberg, Berlin (2021)
184. Schiff, J.L.: Normal Families. Springer, New York (1993)
185. Schlag, W.: A Course in Complex Analysis and Riemann Surfaces. American Mathematical Society, Providence (2014)
186. Schmüdgen, K.: The Moment Problem. Springer, New York, Heidelberg, Berlin (2017)
187. Schramm, O.: Scaling limits of loop-erased random walks and uniform spanning trees. Isr. J. Math. **118**, 221–288 (2000)
188. Seidel, L.: Untersuchungen über die Konvergenz und Divergenz der Kettenbrüche. Habilschrift München (1846)
189. Shanks, D.: Nonlinear transformations of divergent and slowly convergent sequences. J. Math. Phys. **34**, 1–42 (1955)
190. Siegel, C.L.: Iteration of analytic functions. Ann. Math **43**, 607–612 (1942)
191. Siegel, C.L.: Topics in Complex Function Theory, vol. I. Wiley-Interscience, New York (1969)
192. Siegel, C.L.: Topics in Complex Function Theory, vol. II. Wiley-Interscience, New York (1971)
193. Sokhotski, Y.W.: On definite integrals and functions used in series expansions. Doctor thesis, Saint Petersburg (1873)
194. Stein, E.M.: Singular Integrals and Differentiability Properties of Functions. Princeton University Press, Princeton (1970)
195. Stein, E.M., Shakarchi, R.: Real Analysis. Integration, and Hilbert Spaces. Princeton University Press, Princeton, Measure Theory (2005)
196. Stein, E.M., Shakarchi, R.: Functional Analysis. Introduction to Further Topics in Analysis. Princeton University Press, Princeton (2011)
197. Steinmetz, N.: Rational Iteration. Complex Analytical Dynamical Systems. deGruyter, Berlin (1993)
198. Sullivan, D.: Conformal dynamical systems. Geometic Dynamics. Lecture Notes in Mathematics, vol. 1007, pp. 725–752. Springer, New York, Heidelberg, Berlin (1983)

199. Szegő, G.: Über orthogonale Polynome, die zu einer gegebenen Kurve der komplexen Ebene gehören. Math. Z. **9**(1921), 218–270 (1921)
200. Tazzioli: Schwarz's critique and interpretation of the Riemann representation theorem. (Italian) Rend. Circ. Mat. Palermo (2) Suppl. **34**, 95–132 (1994)
201. Teichmüller, O.: Untersuchung über konforme und quasiconforme Abbildung. Dtsch. Math. **3**, 621–678 (1938)
202. Teichmüller, O.: Extremale quasiconformale Abbildungen und quadratische Differentiale. Abh. Preuss. Akad. Math.-Nat. Kl. 1939, no. 22
203. Teichmüller, O.: Bestimmung der extremalen quasiconformalen Abbildungen bei geschlossenen orientierten Riemannschen Flächen. Abh. Preuss. Akad. Math.-Nat. Kl. 1943, no. 4
204. Temme, N.M.: Asymptotic Methods for Integrals. World Scientific, Hackensack, N.J. (2015)
205. Thomas, D., Tunseki, N., Vasudevarao, A.: Univalent Functions. A Primer. De Gruyter, Berlin (2018)
206. Titchmarsh, E.C.: The Theory of Functions, 2nd edn. Oxford University Press, Oxford (1939)
207. Titchmarsh, E.C.: Introduction to the Theory of Fourier Integrals, 3rd edn. Chelsea, New York (1986)
208. Tyurin, A.: Quantization. Classical and Quantum Field Theory and Theta Functions. American Mathematical Society, Providence (2003)
209. Van Assche, W.: Padé and Hermite-Padé approximation and orthogonality. Surv. Approx. Theory **2**, 61–91 (2006)
210. Vekua, I.N.: Generalized Analytic Functions. Addison-Wesley, Reading, Mass (1962)
211. Vekua, N.P.: Systems of Singular Integral Equations. P. Noordhoff Ltd., Groningen (1967)
212. Wang, H., Zhao, Y.: Uniform asymptotics and zeros of a system of orthogonal polynomials defined via a difference equation. J. Math. Anal. Appl. **369**, 453–472 (2010)
213. Weinstein, L.: The Bieberbach conjecture. Duke Math. J. **64**, 61–64 (1991)
214. Weyl, H.: Die Idee der Riemannschen Flächen. Teubner, Stuttgart,: Translation: The Concept of a Riemann Surface, 3rd edn. Dover, New York (1913). (2009)
215. Wienhard, A.: An invitation to higher Teichmüller theory. In: Proceedings of the International Congress of Mathematicians-Rio de Janeiro, vol. 2, pp. 1031–1058 (2018)
216. Wilf, H.: A footnote on two proofs of the Bieberbach-de Branges theorem. Bull. Lond. Math. Soc. **26**, 61–63 (1994)
217. Wolpert, S.: Counting geodesics, Teichmïller space, and random hyperbolic surfaces. Notices, Am. Math. Soc. **68**, 1890–1899 (2021)
218. Wong, R.S.C.: Asymptotic Approximations of Integrals. Academic Press, Boston (1989)
219. Wong, R.S.C., Zhao, Y.-Q.: On a uniform treatment of Darboux's method. Constr. Approx. **21**, 225–255 (2005)
220. Wong, R.S.C., Wyman, M.: The method of Darboux. J. Approx. Theory **10**, 159–171 (1974)
221. Wynn, P.: On a device for computing the em(Sn) transformation. Math. Table Aids Comput. **10**, 91–96 (1956)
222. Yoccoz, J.-C.: Linéarisation des germes de diffeomorphismes holomorphes de (C,0). C. R. Acad. Sci. Paris Sér. I Math. **306**, 55–58 (1988)

Index

R. Beals and R. S. C. Wong, *More Explorations in Complex Functions*, Graduate Texts
in Mathematics 298, https://doi.org/10.1007/978-3-031-28288-1