

Complex Analysis

Course Notes — Harvard University — Math 213a

Fall 2000, 2006, 2010

C. McMullen

January 7, 2011

Contents

1	Basic complex analysis	1
2	The simply-connected Riemann surfaces	27
3	Entire and meromorphic functions	39
4	Conformal mapping	56
5	Elliptic functions and elliptic curves	79

Forward

Complex analysis is a nexus for many mathematical fields, including:

1. Algebra (theory of fields and equations);
2. Algebraic geometry and complex manifolds;
3. Geometry (Platonic solids; flat tori; hyperbolic manifolds of dimensions two and three);
4. Lie groups, discrete subgroups and homogeneous spaces (e.g. $\mathbb{H}/\mathrm{SL}_2(\mathbb{Z})$);
5. Dynamics (iterated rational maps);
6. Number theory and automorphic forms (elliptic functions, zeta functions);
7. Theory of Riemann surfaces (Teichmüller theory, curves and their Jacobians);
8. Several complex variables and complex manifolds;
9. Real analysis and PDE (harmonic functions, elliptic equations and distributions).

This course covers some basic material on both the geometric and analytic aspects of complex analysis in one variable.

Prerequisites: Background in real analysis and basic differential topology (such as covering spaces and differential forms), and a first course in complex analysis.

1 Basic complex analysis

We begin with a quick review of elementary facts about the complex plane and analytic functions.

Some notation. The complex numbers will be denoted \mathbb{C} . We let Δ, \mathbb{H} and $\widehat{\mathbb{C}}$ denote the unit disk $|z| < 1$, the upper half plane $\mathrm{Im}(z) > 0$, and the Riemann sphere $\mathbb{C} \cup \{\infty\}$. We write $S^1(r)$ for the circle $|z| = r$, and S^1 for the unit circle, each oriented counter-clockwise. We also set $\Delta^* = \Delta - \{0\}$ and $\mathbb{C}^* = \mathbb{C} - \{0\}$.

Algebra of complex numbers. The complex numbers are formally defined as the field $\mathbb{C} = \mathbb{R}[i]$, where $i^2 = -1$. They are represented in the

Euclidean plane by $z = (x, y) = x + iy$. There are two square-roots of -1 in \mathbb{C} ; the number i is the one with positive imaginary part.

An important role is played by the Galois involution $z \mapsto \bar{z}$. We define $|z|^2 = N(z) = z\bar{z} = x^2 + y^2$. (Compare the case of a real quadratic field, where $N(a + b\sqrt{d}) = a^2 - db^2$ gives an *indefinite* form.) Compatibility of $|z|$ with the Euclidean metric justifies the identification of \mathbb{C} and \mathbb{R}^2 . We also see that z is a field: $1/z = \bar{z}/|z|$.

It is also convenient to describe complex numbers by polar coordinates

$$z = [r, \theta] = r(\cos \theta + i \sin \theta).$$

Here $r = |z|$ and $\theta = \arg z \in \mathbb{R}/2\pi\mathbb{Z}$. (The multivaluedness of $\arg z$ requires care but is also the ultimate source of powerful results such as Cauchy's integral formula.) We then have

$$[r_1, \theta_1][r_2, \theta_2] = [r_1 r_2, \theta_1 + \theta_2].$$

In particular, the linear maps $f(z) = az + b$, $a \neq 0$, of \mathbb{C} to itself, preserve angles and orientations.

This formula should be proved *geometrically*: in fact, it is a consequence of the formula $|ab| = |a||b|$ and properties of similar triangles. It can then be used to derive the addition formulas for sine and cosine (in Ahlfors the reverse logic is applied).

Algebraic closure. A critical feature of the complex numbers is that they are *algebraically closed*; every polynomial has a root. (A proof will be reviewed below).

Classically, the complex numbers were introduced in the course of solving *real* cubic equations. Starting with $x^3 + ax + b = 0$ one can make a Tschirnhaus transformation so $a = 0$. This is done by introducing a new variable $y = cx^2 + d$ such that $\sum y_i = \sum y_i^2 = 0$; even when a and b are real, it may be necessary to choose c complex (the discriminant of the equation for c is $27b^2 + 4a^3$.) It is negative when the cubic has only one real root; this can be checked by looking at the product of the values of the cubic at its max and min.

Polynomials and rational functions. Using addition and multiplication we obtain naturally the polynomial functions $f(z) = \sum_0^n a_n z^n : \mathbb{C} \rightarrow \mathbb{C}$. The ring of polynomials $\mathbb{C}[z]$ is an integral domain and a unique factorization domain, since \mathbb{C} is a field. Indeed, since \mathbb{C} is algebraically closed, fact every polynomial factors into linear terms.

It is useful to add the allowed value ∞ to obtain the Riemann sphere $\hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$. Then rational functions (ratios $f(z) = p(z)/q(z)$) of relatively

prime polynomials, with the denominator not identically zero) determine rational maps $f : \mathbb{C} \rightarrow \mathbb{C}$. The rational functions $\mathbb{C}(z)$ are the same as the field of fractions for the domain $\mathbb{C}[z]$. We set $f(z) = \infty$ if $q(z) = 0$; these points are called the *poles* of f .

Analytic functions. Let U be an open set in \mathbb{C} and $f : U \rightarrow \mathbb{C}$ a function. We say f is *analytic* if

$$f'(z) = \lim_{t \rightarrow 0} \frac{f(z+t) - f(z)}{t}$$

exists for all $z \in U$. It is crucial here that t approaches zero through arbitrary values in \mathbb{C} . Remarkably, this condition implies that f is a smooth (C^∞) function. For example, polynomials are analytic, as are rational functions away from their poles.

Note that any *real linear* function $\phi : \mathbb{C} \rightarrow \mathbb{C}$ has the form $\phi(v) = av + b\bar{v}$. The condition of analytic says that $Df_z(v) = f'(z)v$; in other words, the \bar{v} part is absent.

To make this point systematically, for a general C^1 function $F : U \rightarrow \mathbb{C}$ we define

$$\frac{dF}{dz} = \frac{1}{2} \left(\frac{dF}{dx} + \frac{1}{i} \frac{dF}{dy} \right) \quad \text{and} \quad \frac{dF}{d\bar{z}} = \frac{1}{2} \left(\frac{dF}{dx} - \frac{1}{i} \frac{dF}{dy} \right).$$

We then have

$$DF_z(v) = \frac{dF}{dz}v + \frac{dF}{d\bar{z}}\bar{v}.$$

We can also write complex-valued 1-form dF as

$$dF = \partial F + \bar{\partial} F = \frac{dF}{dz} dz + \frac{dF}{d\bar{z}} d\bar{z}$$

Thus F is analytic iff $\bar{\partial} F = 0$; these are the *Cauchy-Riemann* equations.

We note that $(d/d\bar{z})\bar{z}^n = n\bar{z}^{n-1}$; a polynomial $p(z, \bar{z})$ behaves as if these variables are independent.

Sources of analytic functions.

Algebraic functions. Beyond the rational and polynomial functions, the analytic functions include *algebraic* functions such that $f(z) = \sqrt{z^2 + 1}$. A general algebraic function $f(z)$ satisfies $P(f) = \sum_0^N a_n(z)f(z)^n = 0$ for some rational functions $a_n(z)$; these arise, at least formally, when one forms algebraic extension of $\mathbb{C}(z)$. Such functions are generally *multivalued*, so we must choose a particular branch to obtain an analytic function.

Differential equations. Analytic functions also arise when one solves *differential equations*. Even equations with constant coefficients, like $y'' + y =$

0, can give rise to transcendental functions such as $\sin(z)$, $\cos(z)$ and e^z . A special case of course is integration. While $\int (x^2 + ax + b)^{1/2} dx$ can be given explicitly in terms of trigonometric functions, already $\int (x^3 + ax + b)^{1/2} dx$ leads one into elliptic functions; and higher degree polynomials lead one to hyperelliptic surfaces of higher genus.

Power series. Analytic functions can be given concretely, locally, by power series such as $\sum a_n z^n$. Conversely, suitable coefficients determine analytic functions; for example, $e^z = \sum z^n/n!$.

Riemann surfaces and automorphy. A third natural source of complex analytic functions is functions that satisfy invariant properties such as $f(z + \lambda) = f(z)$ for all $\lambda \in \Lambda$, a lattice in \mathbb{C} ; or $f(g(z)) = f(z)$ for all $g \in \Gamma \subset \text{Aut}(\mathbb{H})$.

The *elliptic modular functions* $f : \mathbb{H} \rightarrow \mathbb{C}$ have the property that $f(z) = f(z + 1) = f(1/z)$, and hence $f((az + b)/(cz + d)) = f(z)$ for all $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \text{SL}_2(\mathbb{Z})$.

Geometric function theory. Finally a complex analytic function can be specified by a domain $U \subset \mathbb{C}$; we will see that for simply-connected domains (other than \mathbb{C} itself), there is an essentially unique analytic homeomorphism $f : \Delta \rightarrow U$. When U tiles \mathbb{H} or \mathbb{C} , this is related to automorphic functions; and when ∂U consists of lines or circular arcs, one can also give a differential equation for f .

Example. We can also define analytic functions by taking limits of polynomials or other known functions. For example, consider the formula:

$$e^z = \lim (1 + z/n)^n.$$

The triangle with vertices 0, 1 and $1 + i\theta$ has a hypotenuse of length $1 + O(1/n^2)$ and an angle at 0 of $\theta + O(1/n^2)$. Thus one finds *geometrically* that $z_n = (1 + i\theta/n)^n$ satisfies $|z_n| \rightarrow 1$ and $\arg z_n \rightarrow \theta$; in other words,

$$e^{i\theta} = \cos \theta + i \sin \theta.$$

In particular, $e^{\pi i} = -1$ (Euler).

Exponential and trigonometric functions in \mathbb{C} . Here are some useful facts about these familiar functions when extended to \mathbb{C} :

$$\begin{aligned} |\exp(z)| &= \exp \operatorname{Re} z \\ \cos(iz) &= \cosh(z) \\ \sin(iz) &= i \sinh(z) \\ \cos(x + iy) &= \cos(x) \cosh(y) - i \sin(x) \sinh(y) \\ \sin(x + iy) &= \sin(x) \cosh(y) + i \cos(x) \sinh(y). \end{aligned}$$

In particular, the apparent *boundedness* of $\sin(z)$ and $\cos(z)$ fails badly as we move away from the real axis, while $|e^z|$ is actually very small in the halfplane $\operatorname{Re} z \ll 0$.

Complex integration; Cauchy's theorem. Now suppose \overline{U} is a compact, *connected*, smoothly bounded region in \mathbb{C} , $f : \overline{U} \rightarrow \mathbb{C}$ is continuous and $f : U \rightarrow \mathbb{C}$ is analytic. We then have:

Theorem 1.1 (Cauchy) *For any analytic function $f : U \rightarrow \mathbb{C}$, we have $\int_{\partial U} f(z) dz = 0$.*

Remark. It is critical to know the *definition* of such a path integral. (For example, $f(z) = 1$ is analytic, its average over the circle is 1, and yet $\int_{S^1} 1 dz = 0$; why is this?)

If $\gamma : [a, b] \rightarrow \mathbb{C}$ parameterizes an arc, then we define

$$\int_{\gamma} f(z) dz = \int_a^b f(\gamma(t)) \gamma'(t) dt.$$

Alternatively, we choose a sequence of points z_1, \dots, z_n close together along γ , and then define

$$\int_{\gamma} f(z) dz = \lim \sum_1^{n-1} f(z_i)(z_{i+1} - z_i).$$

(If the loop is closed, we choose $z_n = z_1$).

This should not be confused with the integral with respect to arclength:

$$\int_{\gamma} f(z) |dz| = \lim \sum f(z_i) |z_{i+1} - z_i|.$$

Note that the former *depends* on a choice of orientation of γ , while the latter does not.

Proof of Cauchy's formula: (i) observe that $d(f dz) = (\overline{\partial} f) d\overline{z} dz = 0$ and apply Stokes' theorem. (ii — Goursat). Cut the region U into small squares, observe that on these squares $f(z) \approx az + b$, and use the fact that $\int_{\partial U} (az + b) dz = 0$. ■

Aside: distributions. The first proof implicitly assumes f is C^1 , while the second does not. (To see where C^1 is used, suppose $\alpha = u dx + v dy$ and $d\alpha = 0$ on a square S . In the proof that $\int_{\partial S} \alpha = 0$, we integrate $v dy$ over the vertical sides of S and observe that this is the same as integrating $dv/dx dx dy$ over the square. But if α is not C^1 , we don't know that dv/dx is integrable.)

Notes. More on the Cauchy–Riemann equations and with minimal smoothness assumptions can be found in [GM].

More generally, we say a distribution (e.g. an L^1 function f) is (weakly) analytic if $\int f \bar{\partial} \phi = 0$ for every $\phi \in C_c^\infty(U)$. By convolution with a smooth function (a *mollifier*), any weakly analytic function is a limit of C^∞ analytic functions. We will see below that uniform limits of C^∞ analytic functions are C^∞ , so even weakly analytic functions are actually smooth.

Cauchy's integral formula: Differentiability and power series. Because of Cauchy's theorem, only one integral has to be explicitly evaluated in complex analysis (hence the forgetability of the definition of the integral). Namely, setting $\gamma(t) = e^{it}$ we find, for any $r > 0$,

$$\int_{S^1(r)} \frac{1}{z} dz = 2\pi i.$$

By integrating between ∂U and a small loop around $p \in U$, we then obtain *Cauchy's integral formula*:

Theorem 1.2 *For any $p \in U$ we have*

$$f(p) = \frac{1}{2\pi i} \int_{\partial U} \frac{f(z) dz}{z - p}.$$

Now the integrand depends on p only through the rational function $1/(z - p)$, which is infinitely differentiable. (It is the *convolution* of $f|_{\partial U}$ and $1/z$.) So we conclude that $f(p)$ itself is infinitely differentiable, indeed, it is approximated by a sum of rational functions with poles on ∂U . In particular, differentiating under the integral, we obtain:

$$\frac{f^{(k)}(p)}{k!} = \frac{1}{2\pi i} \int_{\partial U} \frac{f(z) dz}{(z - p)^{k+1}}. \quad (1.1)$$

If $d(p, \partial U) = R$, the length of ∂U is L and $\sup_{\partial U} |f| = M$, then this gives the bound:

$$|a_k| = \left| \frac{f^{(k)}(p)}{k!} \right| \leq \frac{ML}{2\pi R^{k+1}}.$$

In particular, $\sum a_k(z-p)^k$ has radius of convergence at least R , since $\limsup |a_k|^{1/k} < 1/R$. This suggests that f is represented by its power series, and indeed this is the case:

Theorem 1.3 *If f is analytic on $B(p, R)$, then $f(z) = \sum a_k(z-p)^k$ on this ball.*

Proof. We can reduce to the case $z \in B(0, 1)$. Then for $w \in S^1$ and fixed z with $|z| < 1$, we have

$$\frac{1}{w-z} = \frac{1}{w} (1 + (z/w) + (z/w)^2 + \cdots),$$

converging uniformly on the circle $|w| = 1$. We then have:

$$f(z) = \frac{1}{2\pi i} \int_{S^1} \frac{f(w) dw}{w-z} = \sum z^k \frac{1}{2\pi i} \int_{S^1} \frac{f(w) dw}{w^{k+1}} = \sum a_k z^k$$

as desired. ■

Corollary 1.4 *An analytic function has at least one singularity on its circle of convergence.*

That is, if f can be extended analytically from $B(p, R)$ to $B(p, R')$, then the radius of convergence is at least R' . So there must be some obstruction to making such an extension, if $1/R = \limsup |a_k|^{1/k}$.

Example: Fibonacci numbers. Let $f(z) = \sum a_n z^n$, where a_n is the n th Fibonacci number. We have $(a_0, a_1, a_2, a_3, \dots) = (1, 1, 2, 3, 5, 8, \dots)$. Since $a_n = a_{n-1} + a_{n-2}$, except for $n = 0$, we get

$$f(z) = (z + z^2)f(z) + 1$$

and so $f(z) = 1/(1 - z - z^2)$. This has a singularity at $z = 1/\gamma$ and thus $\limsup |a_n|^{1/n} = \gamma$, where $\gamma = (1 + \sqrt{5})/2 = 1.618\dots$ is the golden ratio (slightly more than the number of kilometers in a mile).

Theorem 1.5 *A power series represents a rational function iff its coefficients satisfy a recurrence relation.*

Aside: Pisot numbers. The golden ratio is an example of a Pisot number; it has the property that $d(\gamma^n, \mathbb{Z}) \rightarrow 0$ as $n \rightarrow \infty$. It is an unsolved problem to show that if $\alpha > 1$ satisfies $d(\alpha^n, \mathbb{Z}) \rightarrow 0$, then α is an algebraic number.

Kronecker's theorem asserts that $\sum a_i z^i$ is a rational function iff determinants of the matrices $a_{i,i+j}$, $0 \leq i, j \leq n$ are zero for all n sufficiently large [Sa, §I.3]

Question: why are 10:09 and 8:18 such pleasant times? [Mon].

Isolation of zeros. If $f(z) = \sum a_n(z-p)^n$ vanishes at p but is not identically zero, then we can factor out the leading term and write:

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

where g is analytic and $g(p) \neq 0$. This is the simplest case of the *Weierstrass preparation theorem*: it shows germs of analytic functions behave like polynomials times units (invertible functions).

In particular, we find:

Theorem 1.6 *The zeros of a nonconstant analytic function are isolated.*

Warning: we are assuming the domain is connected!

Proof. Let $U_0 \subset U$ be the largest open set with $f|_{U_0} = 0$. Let $U_1 = U - U_0$. Then by the factorization theorem above, the zeros of f in U_1 are isolated. Thus U_1 is open as well. So either $U = U_0$ — in which case f is constant — or $U = U_1$ — in which case f has isolated zeros. ■

Corollary 1.7 *An analytic function which is constant along an arc, or even on a countable set with an accumulation point, is itself constant.*

Corollary 1.8 *The extension of a function $f(x)$ on $[a, b] \subset \mathbb{R}$ to an analytic function $f(z)$ on a connected domain $U \supset [a, b]$ is unique — if it exists.*

Mean value and maximum principle. On the circle $|z| = r$, with $z = re^{i\theta}$, we have $dz/z = i d\theta$. Thus Cauchy's formula gives

$$f(0) = \frac{1}{2\pi i} \int_{S^1(r)} f(z) \frac{dz}{z} = \frac{1}{2\pi} \int_{S^1(r)} f(z) d\theta.$$

In other words, analytic functions satisfy:

Theorem 1.9 (The mean-value formula) *The value of $f(p)$ is the average of $f(z)$ over $S^1(p, r)$.*

Corollary 1.10 (The Maximum Principle) *A nonconstant analytic function does not achieve its maximum in U .*

Proof. Suppose $f(z)$ achieves its maximum at $p \in U$. Then $f(p)$ is the average of $f(z)$ over a small circle $S^1(p, r)$. Moreover, $|f(z)| \leq |f(p)|$ on this circle. The only way the average can agree is if $f(z) = f(p)$ on $S^1(p, r)$. But then f is constant on an arc, so it is constant in U . ■

Corollary 1.11 *If \overline{U} is compact, then $\sup_U |f| = \sup_{\partial U} |f|$.*

Cauchy's bound and algebraic completeness of \mathbb{C} . Suppose f is analytic on $B(p, R)$ and let $M(R) = \sup_{|z-p|=R} |f(z)|$; then Cauchy's bound (1.1) becomes:

$$\frac{|f^{(n)}(p)|}{n!} \leq \frac{M(R)}{R^n}.$$

On the other hand, if $U = \mathbb{C}$ — so f is an *entire* function — then the bound above forces $M(R)$ to grow unless some derivative vanishes identically. Thus we find:

Theorem 1.12 *A bounded entire function is a constant. More generally, if $M(R) = O(R^n)$, then f is a polynomial of degree at most n .*

Corollary 1.13 *Any polynomial $f \in \mathbb{C}[z]$ of degree 1 or more has a zero in \mathbb{C} .*

Otherwise $1/f(z)$ would be a nonconstant, bounded entire function. Alternatively, $1/f(z) \rightarrow \infty$ as $|z| \rightarrow \infty$, so we obtain a violation of the maximum principle.

Corollary 1.14 *There is no conformal homeomorphism between \mathbb{C} and Δ .*

(An analytic map with no critical points is said to be *conformal*, because it preserves angles.)

Aside: quasiconformal maps. A diffeomorphism $f : U \rightarrow V$ between domains in \mathbb{C} is *quasiconformal* if $\sup_U |\overline{\partial}f/\partial f| < \infty$. Many qualitative theorems for conformal maps also hold for quasiconformal maps. For example, there is no quasiconformal homeomorphism between \mathbb{C} and Δ .

Parseval's theorem. The power series of an analytic function on the ball $B(0, R)$ also contains information about its L^2 -norm on the circle $|z| = R$: namely if $f(z) = \sum a_n z^n$, then we have:

$$\sum |a_n|^2 R^{2n} = \frac{1}{2\pi} \int_{|z|=R} |f(z)|^2 d\theta.$$

This comes from the fact that the functions z^n are orthogonal in $L^2(S^1)$. It also gives another important perspective on holomorphic functions: they are the elements in $L^2(S^1)$ with *positive* Fourier coefficients, and hence give a half-dimensional subspace of this infinite-dimensional space.

Compactness of bounded functions. Cauchy's bound on a disk also implies that if f is small, then f' is also small, at least if we are not too near the edge of U .

Theorem 1.15 *Let $f(z)$ be analytic on U and bounded by M . Then $|f'(z)| \leq M/d(z, \partial U)$.*

Corollary 1.16 *If f_n are analytic functions and $f_n \rightarrow f$ uniformly, then f' exists and $f'_n \rightarrow f'$ locally uniformly.*

Corollary 1.17 *A uniform limit of analytic functions is analytic.*

Note: *Many fallacies in real analysis become theorems in complex analysis.*

Note that $f_n(z) = z^n/n$ tends to zero uniformly on Δ , but $f'_n(z) = z^{n-1}$ only tends to zero *locally* uniformly.

Corollary 1.18 *Let f_n be a sequence of functions on U with $|f_n| \leq M$. Then after passing to a subsequence, there is an analytic function g such that $f_n \rightarrow g$ locally uniformly on U .*

Proof. For any compact set $K \subset U$, the restrictions $f_n|_K$ are equicontinuous. Apply the Arzela–Ascoli theorem. ■

Laurent series. Using again the basic series $1/(1-z) = \sum z^n$ and Cauchy's formula over the two boundary components of the *annulus* $A(r, R) = \{z : r < |z| < R\}$, we find:

Theorem 1.19 *If $f(z)$ is analytic on $A(r, R)$, then in this region we have*

$$f(z) = \sum_{n=-\infty}^{\infty} a_n z^n.$$

The positive terms converge for $|z| < R$, and the negative terms converge for $|z| > r$.

Corollary 1.20 *An analytic function on the annulus $r < |z| < R$ can be expressed as the sum of a function analytic on $|z| < R$ and a function analytic on $|z| > r$.*

Isolated singularities. As a special case, if f is analytic on $U - p$, with $p \in U$, then near p we have a Laurent series expansion $f(z) = \sum_{n=-\infty}^{\infty} a_n (z - p)^n$. We say f has an *isolated singularity* at p .

We write $\text{ord}(f, p) = n$ if $a_n \neq 0$ but $a_i = 0$ for all $i < n$. The values $n = -\infty$ and $+\infty$ are also allowed. If $n \geq 0$, the apparent singularity at p is removable and f has a *zero* of order n at p . If $-\infty < n < 0$, we say f has a *pole* of order $-n$ at p . In either of these cases, we can write

$$f(z) = (z - p)^n g(z),$$

where $g(p) \neq 0$ and $g(z)$ is analytic near p (so $\mathcal{O}(g, p) = 0$).

If $\text{ord}(f, p) > -\infty$, then f has a *finite Laurent series*

$$f(z) = \frac{a_{-n}}{(z - p)^n} + \cdots + \frac{a_{-1}}{z - p} + \sum_{n=0}^{\infty} a_n z^n$$

near p . The germs of functions at p with finite Laurent tails form a local field, with $\text{ord}(f, p)$ as its discrete valuation. (Compare \mathbb{Q}_p , where $v_p(p^n a/b) = n$.)

If $\text{ord}(f, p) = -\infty$ we say f has an *essential singularity* at p . (Example: $f(z) = \sin(-1/z)$ at $z = 0$.)

The residue. A critical role is played by the *residue* of $f(z)$ at p , defined by $\text{Res}(f, p) = a_{-1}$. It satisfies

$$\int_{\gamma} f(z) dz = 2\pi i \text{Res}(f, p)$$

for any small loop encircling the point p in U . Thus the residue is intrinsically an invariant of the 1-form $f(z) dz$, *not* the function $f(z)$. (If we regard $f(z) dz$ as a 1-form, then its residue is invariant under change of coordinates.)

Theorem 1.21 (The residue theorem) *Let $f : U \rightarrow \mathbb{C}$ be a function which is analytic apart from a finite set of isolated singularities. We then have:*

$$\int_{\partial U} f(z) dz = 2\pi i \sum_{p \in U} \text{Res}(f, p).$$

We will develop two applications of the residue theorem: the argument principle, and the evaluation of definite integrals.

The argument principle. The previous argument for algebraic completeness of \mathbb{C} is clever and nonconstructive. One way to make the proof more transparent and constructive is to employ the *argument principle*.

We first observe that if f has at worst a pole at p , then its *logarithmic derivative*

$$d \log f = f'(z)/f(z) dz$$

satisfies

$$\text{Res}(f'/f, p) = \text{ord}(f, p).$$

We thus obtain:

Theorem 1.22 (Argument principle) *If f is analytic in U and $f|_{\partial U}$ is nowhere zero, then the number of zeros of f in U is given by*

$$N(f, 0) = \sum_{p \in U} \text{ord}(f, p) = \frac{1}{2\pi i} \int_{\partial U} \frac{f'(z) dz}{f(z)}.$$

Corollary 1.23 (Rouché's Theorem) *If $|f| > |g|$ along ∂U , then f and $f + g$ have the same number of zeros in U .*

Proof. The continuous, integer-valued function $N(f + tg, U)$ is constant for $t \in [0, 1]$. ■

Example. Consider $p(z) = z^5 + 14z + 1$. Then all its zeros are inside $|z| < 2$, since $|z|^5 = 32 > |14z + 1|$ when $|z| = 2$; but only one inside $|x| < 3/2$, since $|z^5 + 1| \leq 1 + (3/2)^5 < 9 < |14z|$ on $|z| = 3/2$. (Intuitively, the zeros of $p(z)$ are close to the zeros of $z^5 + 14z$ which are $z = 0$ and otherwise 4 points on $|z| = 14^{1/4} \approx 1.93$.)

Corollary 1.24 *Let $p(z) = z^d + a_1 z^{d-1} + \dots + a_d$, and suppose $|a_i| < R^i/d$. Then $p(z)$ has d zeros inside the disk $|z| < R$.*

Proof. Write $p(z) = f(z) + g(z)$ with $f(z) = z^d$, and let $U = B(0, R)$; then on ∂U , we have $|f| = R^d$ and $|g| < R^d$; now apply Rouché's Theorem. ■

Corollary 1.25 *A nonconstant analytic function is an open mapping.*

Proof. Suppose $f(p) = q$. Then p is an isolated zero of $f(z) - q$. Choose a ball $B(p, r)$ so there are no other zeros inside this ball, and let s be the minimum of $|f(z) - q|$ over $\partial B(p, r)$. Then if $|q - q'| < s$, we find $f(z) - q'$ also has a zero in $B(p, r)$, and thus $f(B(p, r))$ contains $B(q, s)$. This shows $f(U)$ is open whenever U is open. ■

The open mapping theorem gives an alternate proof of the maximum principle *and* its strict version:

Corollary 1.26 *If $|f|$ achieves its maximum in U , then f is a constant.*

Geometric picture. Suppose for simplicity that U is a disk, and $p \notin f(\partial U)$. The the number of solutions to $f(z) = p$ in U , counted with multiplicity, is:

$$N(f, p) = \frac{1}{2\pi i} \int_{\partial U} d \log(f(z) - p) = \frac{1}{2\pi} \int_{\partial U} d \arg(f(z) - p).$$

This is nothing more than the *winding number* of $f(\partial U)$ around p . Observe that $N(f, p)$ is a locally constant function on $\mathbb{C} - f(\partial U)$, zero on the noncompact component. Summing up:

Theorem 1.27 *For any $p \in \mathbb{C} - f(\partial U)$, the number of solutions to $f(z) = p$ in U is the same as the winding number of $f(\partial U)$ around p .*

Using isolation of zeros, we have:

Corollary 1.28 *If $f'(p) \neq 0$, then f is a local homeomorphism at p .*

In fact for r sufficiently small and q close to p , we have

$$f^{-1}(q) = \frac{1}{2\pi} \int_{B(p, r)} \frac{zf'(z) dz}{f(z) - q}.$$

Aside: the smooth case. These results also hold for *smooth mappings* f once one finds a way to count the number of solutions to $f(z) = p$ correctly. (Some may count negatively, and the zeros are only isolated for *generic* values of p .)

Aside: Linking numbers and intersection multiplicities of curves in \mathbb{C}^2 . Counting the number of zeros of $y = f(x)$ at $x = 0$ is the same as counting the multiplicity of intersection between the curves $y = 0$ and $y = f(x)$ in \mathbb{C}^2 , at $(0, 0)$.

In general, to an analytic curve C defined by $F(x, y) = 0$ passing through $(0, 0) \in \mathbb{C}^2$, we can associate a *knot* by taking the intersection of C with the boundary $S^3(r)$ of a small ball centered at the origin.

A pair of distinct, irreducible, curves C_1 and C_2 passing through $(0, 0)$ (say defined by $f_i(x, y) = 0$, $i = 1, 2$), have a *multiplicity* of intersection $m(C_1, C_2)$. This can be defined geometrically by intersection with a small sphere $S^3(r)$ in \mathbb{C}^2 centered at $(0, 0)$. Then we get a pair of *knots*, and their *linking number* is the same as this multiplicity.

Examples. It is often simpler to use $S^3 = \partial\Delta^2$ instead of $|x|^2 + |y|^2 = 1$. Then S^3 is the union of two solid tori, $\Delta \times S^1$ and $S^1 \times \Delta$. The axes $x = 0$ and $y = 0$ are the core curves of these tori; they have linking number one. The line $y = x$ is a $(1, 1)$ curve on $S^1 \times S^1$; more generally, for $\gcd(a, b) = 1$, $y^a = x^b$ is an (a, b) curve, parameterized by $(x, y) = (e^{ita}, e^{itb})$. In particular, the cusp $y^2 = x^3$ meets S^3 in a trefoil knot. It links $x = 0$ twice and $y = 0$ three times.

Problem. Show that the figure-eight knot cannot arise from an analytic curve.

Multivalued functions. It is useful to have a general discussion of the sometimes confusing notion of ‘branch cuts’ and ‘multivalued functions’. Here are 2 typical results.

Theorem 1.29 *Let $U \subset \mathbb{C}^*$ be a simply-connected region, and suppose $e^a = b \in U$. Then there is a unique analytic function $L : U \rightarrow \mathbb{C}$ such that $L(b) = a$ and $\exp L(z) = z$ for all $z \in U$.*

This function is a ‘branch’ of $\log(z)$. To define it, we simply set

$$L(z) = a + \int_b^z dt/t.$$

The integral is over any path in U connecting b to z . Since U is simply-connected and does not contain zero, this integral is path independent, and we have $L'(z) = 1/z$ and $L(b) = a$. Thus $f(z) = e^{L(z)}/z$ satisfies $f'(z) = 0$, and hence $f(z)$ is a constant. Since $f(b) = e^a/b = 1$, we conclude that $f(z) = \log z$.

Corollary 1.30 *If $a^n = b$ then there is a unique analytic function $R : U \rightarrow \mathbb{C}$ such that $R(b) = a$ and $R(z)^n = z$ for all $z \in U$.*

Of course near $z = 1$ with the usual normalizations these functions can be written down explicitly. Integrating $1/(1+z)$, we find:

$$\log(1+z) = z - z^2/2 + z^3/3 - \dots$$

and, by the binomial theorem,

$$(1+z)^\alpha = 1 + \alpha z + \alpha(\alpha-1)z^2/2 + \dots$$

Functional factorization. Here is an alternate proof of the open mapping theorem. It is clear that $p_n(z) = z^n$ is an open map, even at the origin; and (by the inverse function theorem) that an analytic function is open at any point p where $f'(p) \neq 0$. The general case follows from these via:

Theorem 1.31 *Let f be an analytic map at p with $\text{ord}(f', p) = n \geq 0$. Then up to a local change of coordinates in domain and range, $f(z) = z^{n+1}$. That is, there exist analytic diffeomorphisms with $h_1(0) = p$, and $h_2(0) = f(p)$, such that*

$$h_2 \circ f \circ h_1^{-1} = z^{n+1}.$$

Proof. We may assume $p = f(p) = 0$ and $f(z) = z^n g(z)$ where $g(0) \neq 0$. Then $h(z) = g(z)^{1/n}$ is a well-defined analytic function near $z = 0$, once we have chosen a particular value for $g(0)^{1/n}$. It follows that $f(z) = (zh(z))^n = p_n(h_1(z))$ where $h'_1(z) \neq 0$. Thus h_1 is a local diffeomorphism, and $f \circ h_1^{-1}(z) = z^n$. ■

Bounded functions and essential singularities.

Theorem 1.32 *Every isolated singularity of a bounded analytic function is removable.*

Proof. Suppose the singularity is at $z = 0$, and write $f(z)$ as a Laurent series $\sum a_n z^n$. Then for any $r > 0$ we have

$$\text{Res}(f, 0) = a_{-1} = \frac{1}{2\pi i} \int_{S^1(r)} f(z) dz.$$

But if $|f| \leq M$ then this integral tends to zero as $r \rightarrow 0$, and hence $a_{-1} = 0$. Similarly $a_{-(n+1)} = \text{Res}(z^n f(z), 0) = 0$ for all $n \geq 0$. Thus the Laurent power series gives an extension of f to an analytic function at $z = 0$. ■

Corollary 1.33 (Weierstrass-Casorati) *If $f(z)$ has an essential singularity at p , then there exist $z_n \rightarrow p$ such that $f(z_n)$ is dense in \mathbb{C} .*

Proof. Otherwise there is a neighborhood U of p such that $f(U - \{p\})$ omits some ball $B(q, r)$ in \mathbb{C} . But then $g(z) = 1/(f(z) - q)$ is bounded on U , and hence analytic at p , with a zero of finite order. Then $f(z) = q + 1/g(z)$ has at worst a pole at p , not an essential singularity. ■

Aside: several complex variables. Using the same type of argument and some basic facts from several complex variables, it is easy to show that if $V \subset U \subset \mathbb{C}^n$ is an analytic hypersurface, and $f : U - V \rightarrow \mathbb{C}$ is a bounded analytic function, then f extends to all of U .

More remarkably, if V has codimension two or more, then *every* analytic function extends across V . For example, an isolated point is always a removable singularity in \mathbb{C}^2 .

Residue calculus and definite integrals. The residue theorem can be used to systematically evaluate various definite integrals.

Definite integrals 1: rational functions on \mathbb{R} . Whenever a rational function $R(x) = P(x)/Q(x)$ has the property that $\int_{\mathbb{R}} |R(x)| dx$ is finite, we can compute this integral via residues: we have

$$\int_{-\infty}^{\infty} R(x) dx = 2\pi i \sum_{\text{Im } p > 0} \text{Res}(R, p).$$

(Of course we can also compute this integral by factoring $Q(x)$ and using partial fractions and trig substitutions.)

Example. Where does π come from? It emerges naturally from rational functions by integration — i.e. it is a *period*. Namely, we have

$$\int_{-\infty}^{\infty} \frac{dx}{1+x^2} = 2\pi i \text{Res}(1/(1+z^2), i) = 2\pi i(-i/2) = \pi.$$

Of course this can also be done using the fact that $\int dx/(1+x^2) = \tan^{-1}(x)$. More magically, for $f(z) = 1/(1+z^4)$ we find:

$$\int_{-\infty}^{\infty} \frac{dx}{1+x^4} = 2\pi i (\text{Res}(f, (1+i)/\sqrt{2}) + \text{Res}(f, (1+i)/\sqrt{2})) = \frac{\pi}{\sqrt{2}}.$$

Both are obtained by closing a large interval $[-R, R]$ with a circular arc in the upper halfplane, and then taking the limit as $R \rightarrow \infty$.

We can even compute the general case, $f(z) = 1/(1 + z^n)$, with n even. For this let $\zeta_k = \exp(2\pi i/k)$, so $f(\zeta_{2n}) = 0$. Let P be the union of the paths $[0, \infty)\zeta_n$ and $[0, \infty)$, oriented so P moves positively on the real axis. We can then integrate over the boundary of this pie-slice to obtain:

$$(1 - \zeta_n) \int_0^\infty \frac{dx}{1 + x^n} = \int_P f(z) dz = 2\pi i \operatorname{Res}(f, \zeta_{2n}) = 2\pi i/(n\zeta_{2n}^{n-1}),$$

which gives

$$\int_0^\infty \frac{dx}{1 + x^n} = \frac{2\pi i}{n(-\zeta_{2n}^{-1} + \zeta_{2n}^{+1})} = \frac{\pi/n}{\sin \pi/n}.$$

Here we have used the fact that $\zeta_{2n}^n = -1$. Note that the integral tends to 1 as $n \rightarrow \infty$, since $1/(1 + x^n)$ converges to the indicator function of $[0, 1]$.

Definite integrals 2: rational functions of $\sin(\theta)$ and $\cos(\theta)$. Here is an even more straightforward application of the residue theorem: for any rational function $R(x, y)$, we can evaluate

$$\int_0^{2\pi} R(\sin \theta, \cos \theta) d\theta.$$

The method is simple: set $z = e^{i\theta}$ and convert this to an integral of an analytic function over the unit circle. To do this we simply observe that $\cos \theta = (z + 1/z)/2$, $\sin \theta = (z - 1/z)/(2i)$, and $dz = iz d\theta$. Thus we have:

$$\int_0^{2\pi} R(\sin \theta, \cos \theta) d\theta = \int_{S^1} R\left(\frac{1}{2i}\left(z - \frac{1}{z}\right), \frac{1}{2}\left(z + \frac{1}{z}\right)\right) \frac{dz}{iz}.$$

For example, for $0 < a < 1$ we have:

$$\int_0^{2\pi} \frac{d\theta}{1 + a^2 - 2a \cos \theta} = \int_{S^1} \frac{iz}{(z - a)(az - 1)} = 2\pi i(i/(a^2 - 1)) = \frac{2\pi}{1 - a^2}.$$

Definite integrals 3: fractional powers of x . $\int_0^\infty x^a R(x) dx$, $0 < a < 1$, R a rational function.

For example, consider

$$I(a) = \int_0^\infty \frac{x^a}{1 + x^2} dx.$$

Let $f(z) = z^a/(1 + z^2)$. We integrate out along $[0, \infty)$ then around a large circle and then back along $[0, \infty)$. The last part gets shifted by analytic continuation of x^a and we find

$$(1 - 1^a)I(a) = 2\pi i(\operatorname{Res}(f, i) + \operatorname{Res}(f, -i))$$

and $\text{Res}(f, i) = i^a/(2i)$, $\text{Res}(f, -i) = (-i)^a/(-2i)$ (since $x^a/(1+x^2) = x^a/(x-i)(x+i)$). Thus, if we let $i^a = \omega = \exp(\pi ia/2)$, we have

$$I(a) = \frac{\pi(i^a - (-i)^a)}{(1 - 1^a)} = \pi \frac{\omega - \omega^3}{1 - \omega^4} = \frac{\pi}{\omega + \omega^{-1}} = \frac{\pi}{2 \cos(\pi a/2)}.$$

For example, when $a = 1/3$ we get

$$I(a) = \pi/(2 \cos(\pi/6)) = \pi/\sqrt{3}.$$

Residues and infinite sums. The periodic function $f(z) = \pi \cot(\pi z)$ has the following convenient properties: (i) It has residues 1 at all the integers; and (ii) it remains bounded as $\text{Im } z \rightarrow \infty$. From these facts we can deduce some remarkable properties: by integrating over a large rectangle $S(R)$, we find for $k \geq 2$ even,

$$0 = \lim_{R \rightarrow \infty} \frac{1}{2\pi i} \int_{S(R)} \frac{f(z) dz}{z^k} = \text{Res}(f(z)/z^k, 0) + 2 \sum_1^{\infty} 1/n^k.$$

Thus we can evaluate the sum $\sum 1/n^2$ using the Laurent series

$$\begin{aligned} \cot(z) &= \frac{\cos(z)}{\sin(z)} = \frac{1 - z^2/2! + z^4/4! - \dots}{z(1 - z^2/3! + z^4/5! - \dots)} \\ &= z^{-1}(1 - z^2/2! + z^4/4! - \dots)(1 + z^2/6 + 7z^4/360 + \dots) \\ &= z^{-1} - z/3 - z^3/45 - \dots \end{aligned}$$

(using the fact that $(1/(3!)^2 - 1/5! = 7/360)$. This shows $\text{Res}(f(z)/z^2, 0) = -\pi^2/3$ and hence $\sum 1/n^2 = \pi^2/6$. Similarly, $2\zeta(2k) = -\text{Res}(f(z)/z^{2k}, 0)$. For example, this justifies $\zeta(0) = 1 + 1 + 1 + \dots = -1/2$.

Little is known about $\zeta(2k+1)$. Apéry showed that $\zeta(3)$ is irrational, but it is believed to be transcendental.

We note that $\zeta(s) = \sum 1/n^s$ is analytic for $\text{Re } s > 1$ and extends analytically to $\mathbb{C} - \{1\}$ (with a simple pole at $s = 1$). In particular $\zeta(0)$ is well-defined. Because of the factorization $\zeta(z) = \prod (1 - 1/p^s)^{-1}$, the behavior of the zeta function is closely related to the distribution of prime numbers. The famous *Riemann hypothesis* states that any zero of $\zeta(s)$ with $0 < \text{Re } s < 1$ satisfies $\text{Re } s = 1/2$. It implies a sharp form of the prime number theorem, $\pi(x) = x/\log x + O(x^{1/2+\epsilon})$.

The zeta function also has *trivial zeros* at $s = -2, -4, -6, \dots$

Hardy's paper on $\int \sin(x)/x \, dx$. We claim

$$I = \int_0^{\infty} \frac{\sin x \, dx}{x} = \frac{\pi}{2}.$$

Note that this integral is improper, i.e. it does not converge absolutely. Also, the function $f(z) = \sin(z)/z$ has no poles — so how can we apply the residue calculus?

The trick is to observe that

$$-2iI = \lim_{r \rightarrow 0} \int_{r < |x| < 1/r} \frac{e^{ix} dx}{x}.$$

We now use the fact that $|e^{ix+iy}| \leq e^{-y}$ to close the path in the upper halfplane, and conclude that

$$2iI = \lim_{r \rightarrow 0} \int_{S^1(r)_+} \frac{e^{iz} dz}{z}.$$

(Here the semicircle is oriented counter-clockwise, as usual.) Since $\text{Res}(e^{iz}/z, 0) = 1$, we find $2iI = (2\pi i)(1/2)$ and hence $I = \pi/2$.

Harmonic functions. A C^2 real-value function $u(z)$ is *harmonic* if

$$\Delta u = \frac{d^2 u}{dx^2} + \frac{d^2 u}{dy^2} = 4 \frac{d^2}{dz d\bar{z}} = 0.$$

Equivalently, we have

$$d * du = 0,$$

where $*$ is the Hodge star operator (given by $*dx = dy$ and $*dy = -dx$). In terms better adapted to complex analysis, u is harmonic iff

$$\bar{\partial} \partial u = 0.$$

In physical terms, $\nabla \cdot \nabla u = 0$, and $d^2 u = 0$ implies $\nabla \times \nabla u = 0$; i.e., u generates a *volume preserving flow* with *zero curl*.

Note. The operators ∂ and $\bar{\partial}$ extend naturally to maps from 1-forms to 2-forms; they satisfy $\partial(f(z) d\bar{z}) = (\partial f) d\bar{z}$, $\partial(f(z) dz) = 0$. Then $d = \partial + \bar{\partial}$ on 1-forms as well.

Basic facts:

1. If $f = u + iv$ is analytic, then f, \bar{f}, u and v are harmonic.
2. A function u is harmonic iff du/dz is holomorphic.
3. Any real-valued harmonic function is locally the real part of a holomorphic function.

(Integrate ∂u to obtain an analytic function with $\partial f = \partial u$. Then $\bar{\partial} \bar{f} = \bar{\partial} u$; thus $d(f + \bar{f}) = du$ and so $u = f + \bar{f}$ up to an additive constant.)

4. Thus any C^2 harmonic function is actually infinitely differentiable.
5. A harmonic function satisfies the mean-value theorem: $u(p)$ is the average of $u(z)$ over $S^1(z, p)$. This implies:
6. A harmonic function satisfies the maximum principle.
7. If u is harmonic and f is analytic, then $u \circ f$ is also harmonic.
8. If $f = u + iv$ is analytic, we say v is a *harmonic conjugate* of u . Then $-u$ is a harmonic conjugate of v .
9. A uniform limit of harmonic functions is harmonic.

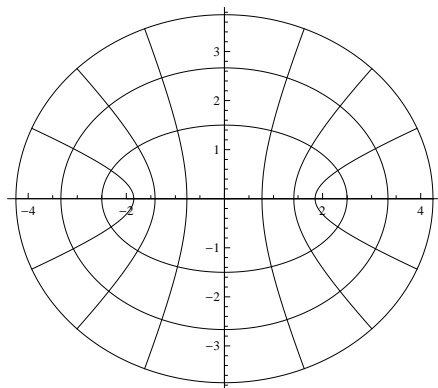


Figure 1. Orthogonal level sets.

Examples. The function $\operatorname{Re}(z^3) = x^3 - 3xy^2$ is a harmonic polynomial. The function $\arg z$ is the harmonic conjugate of $\log |z|$. This shows the harmonic conjugate may be multivalued.

Flows. The level sets of u and v are orthogonal. Thus the area-preserving flow generated by ∇u follows the level sets of v , and vice-versa. A simple example is provided by polar coordinates, which give the level sets of u and v where $\log |z| = u + iv$. The area-preserving flows are rotation around the origin, and the radial flow at rate $1/r$ through circles of radius r .

See Figure 1 for another example, this time on $U = \mathbb{C} - \{-2, 2\}$, where the level sets are conics. These conics are the images of radial lines and circles under $f(z) = z + 1/z$, so they are also locally level sets of harmonic functions.

Harmonic extension. Here is one of the central existence theorems for harmonic functions.

Theorem 1.34 *There is a unique linear map $P : C(S^1) \rightarrow C(\overline{\Delta})$ such that $u = P(u)|_{S^1}$ and $P(u)$ is harmonic on Δ .*

Proof. Uniqueness is immediate from the maximum principle. To see existence, observe that we must have $P(\bar{z}^n) = \bar{z}^n$ and $P(z^n) = z^n$. Thus P is well-defined on the span S of polynomials in z and \bar{z} , and satisfies there $\|P(u)\|_\infty = \|u\|_\infty$. Thus P extends continuously to all of $C(S^1)$. Since the uniform limit of harmonic functions is harmonic, $P(u)$ is harmonic for all $u \in C(S^1)$. ■

Poisson kernel. The map P can be given explicit by the *Poisson kernel*. For example, $u(0)$ is just the average of u over S^1 . We can also say $u(p)$ is the expected value of $u(z)$ under a random walk starting at p that exits the disk at z .

To find the Poisson kernel explicitly, suppose we have a δ -mass at $z = 1$. Then it should extend to a positive harmonic function u on Δ which vanishes along S^1 except at 1, and has $u(0) = 0$. In turn, u should be the real part of an analytic function $f : \Delta \rightarrow \mathbb{C}$ such that $f(0) = 1$ and $\operatorname{Re} f|_{S^1} = 0$ and f has a pole at $z = 1$. Such a function is given simply by the Möbius transformation $f : \Delta \rightarrow U = \{z : \operatorname{Re} z > 0\}$:

$$f(z) = \frac{1+z}{1-z}.$$

Convolving, we find the analytic function with $\operatorname{Re} f = u$ for a given $u \in C(S^1)$ is given by

$$F(z) = \frac{1}{2\pi} \int_{S^1} f(z/t)u(t)|dt|,$$

and thus

$$u(r, \alpha) = \frac{1}{2\pi} \int_{S^1} P_r(\alpha - \theta)u(\theta) d\theta,$$

where, for $z = re^{i\theta}$, we have

$$P_r(\theta) = \operatorname{Re} f(z) = \frac{1 - |z|^2}{|1 - z|^2} = \frac{1 - r^2}{1 - 2r \cos \theta + r^2}.$$

Relation to Fourier series. The above argument suggests that, to define the harmonic extension of u , we should just write $u(z) = \sum_{-\infty}^{\infty} a_n z^n$ on S^1 ,

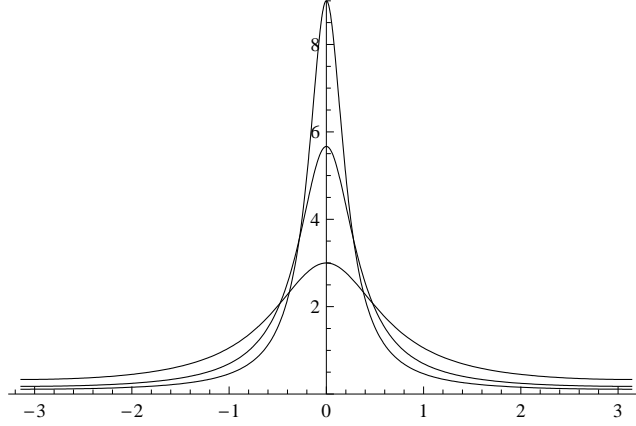


Figure 2. The Poisson kernel $P_r(\theta)$ for $r = 0.5, 0.7, 0.8$.

and then replace z^{-n} by \bar{z}^n to get its extension to the disk. This actually works, and gives another approach to the Poisson kernel.

Given $f \in C(S^1)$, it is not true, in general, that its Fourier series $\sum a_n z^n$ converges for all $z \in S^1$. However, it *is* true that this series converges in the unit disk and defines a harmonic function there. As we have seen, this harmonic function provides a continuous extension of f . This shows:

Theorem 1.35 *If $f \in C(S^1)$ has Fourier coefficients a_n , then for all $z \in S^1$ we have*

$$f(z) = \lim_{r \rightarrow 1^-} \sum_{n=-\infty}^{\infty} a_n r^{|n|} z^n.$$

(Here $a_n = (1/2\pi) \int_{S^1} f(z) \bar{z}^n |dz|$.)

Abel summation. In general, we say S is the *Abel sum* of the series $\sum b_n$ if $S = \lim_{r \rightarrow 1^-} \sum b_n r^n$. For example,

$$1 - 2 + 3 - 4 + \cdots = \lim(1 - 2r + 3r^2 - \cdots) = \lim -1/(1+r)^2 = -1/4.$$

The result above shows the Fourier series of f is *Abel summable* to the original function f .

Laplacian as a quadratic form, and physics. Suppose $u, v \in C_c^\infty(\mathbb{C})$ – so u and v are smooth, real-valued functions vanishing outside a compact set. Then, by integration by parts, we have

$$\int_{\Delta} \langle \nabla u, \nabla v \rangle = - \int_{\Delta} \langle u, \Delta v \rangle = - \int_{\Delta} \langle v, \Delta u \rangle.$$

To see this using differential forms, note that:

$$0 = \int_{\Delta} d(u * dv) = \int_{\Delta} (du)(*dv) + \int_{\Delta} u(d * dv).$$

In particular, we have

$$\int_{\Delta} |\nabla u|^2 = - \int_{\Delta} u \Delta u.$$

Compare this to the fact that $\langle Tx, Tx \rangle = \langle x, T^*Tx \rangle$ on any inner product space. Thus $-\Delta$ defines a positive-definite quadratic form on the space of smooth functions.

The extension of u from S^1 to Δ is a ‘minimal surface’ in the sense that it minimizes $\int_{\Delta} |\nabla u|^2$ over all possible extensions. Similarly, minimizing the energy in an electric field then leads to the condition $\Delta u = 0$ for the electrical potential.

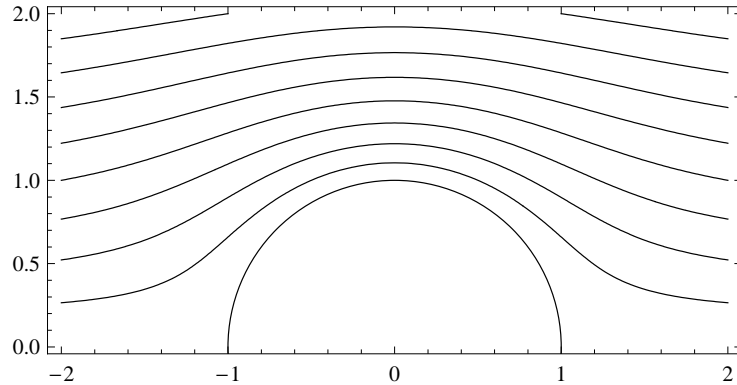


Figure 3. Streamlines around a cylinder.

Probabilistic interpretation. Brownian motion is a way of constructing random paths in the plane (or \mathbb{R}^n). It leads to the following simple interpretation of the extension operator P . Namely, given $p \in \Delta$, one considers a random path p_t with $p_0 = p$. With probability one, there is a first $T > 0$ such that $|p_T| = 1$; and then one sets $u(p) = E(u(p_T))$. In other words, $u(p)$ is the expected value of $u(p_T)$ at the moment the Brownian path exits the disk.

Using the Markov property of Brownian motion, it is easy to see that $u(p)$ satisfies the mean-value principle, which is equivalent to it being harmonic.

It is also easy to argue that $|p_0 - p_T|$ tends to be small when p_0 is close to S^1 , and hence $u(p)$ is a continuous extension of $u|_{S^1}$.

The Poisson kernel $(1/2\pi)P_r(\theta)d\theta$ gives the *hitting density* on S^1 for a Brownian path starting at $(r, 0)$.

Hyperbolic geometry interpretation. Alternatively, $u(z)$ is the expected value of $u(p)$ and the endpoint of a random hyperbolic geodesic ray γ in Δ with one vertex at z . (The angle of the ray in $T_z\Delta$ is chosen at random in S^1 .)

Example: fluid flow around a cylinder. We begin by noticing that $f(z) = z + 1/z$ gives a conformal map from the region $U \subset \mathbb{H}$ where $|z| > 1$ to \mathbb{H} itself, sending the circular arc to $[-2, 2]$. Thus the level sets of $\text{Im } f = y(1 - 1/(x^2 + y^2))$ describe fluid flow around a cylinder. Note that we are modeling incompressible fluid flow with *no rotation*, i.e. we are assuming the curl of the flow is zero. This insures the flow is given by the gradient of a function.

Harmonic functions and the Schwarz reflection principle. Here is an application of harmonic functions that will be repeatedly used in geometric function theory.

Let $U \subset \mathbb{C}$ be a region invariant under $z \mapsto \bar{z}$. Suppose $f : U \rightarrow \mathbb{C}$ is continuous,

$$f(\bar{z}) = \overline{f(z)}, \quad (1.2)$$

and f is analytic on $U_+ = U \cap \mathbb{H}$. We can then conclude that f is analytic on U . In particular, f is analytic at each point of $U \cap \mathbb{R}$.

Here is a stronger statement:

Theorem 1.36 *Suppose $f : U_+ \rightarrow \mathbb{C}$ is analytic and $\text{Im } f(z) \rightarrow 0$ as $\text{Im } z \rightarrow 0$. Then f extends to an analytic function on U satisfying (1.2).*

Proof. The statement is local, so we can assume $U = B(p, r)$ where $p \in \mathbb{R}$. Let $f(z) = u(z) + iv(z)$; then v is harmonic on U_+ and v extends continuously to the real axis, with $v(x) = 0$. Extend v to U by $v(\bar{z}) = -v(z)$.

Now by the Poisson integral, $v|_{S^1(p, r)}$ extends to a unique harmonic function h on $B(p, r)$. By uniqueness, $h(\bar{z}) = -h(z)$, and hence h also vanishes on the real axis. Thus $h = v$ on ∂U_+ . By the maximum principle, $h = v$ on U_+ , and hence on U .

We are now done: by the existence of harmonic conjugates, there is *some* analytic function on U with $\text{Im } F = v$. But then $\text{Im } F = \text{Im } f$ on U , so F and f differ by a real constant, which can be normalized to be zero. ■

Remark. One can replace $z \mapsto \bar{z}$ with reflection through any circle, or more generally with local reflection through a real-analytic arc.

Example. Suppose $f(z)$ is analytic on \mathbb{H} and $f(z) \rightarrow 0$ along an interval in \mathbb{R} . Then f is identically zero. This extends the ‘isolated zero’ principle to the boundary of a region.

Additional topics.

The Phragmen–Lindelöf Theorems. These theorems address the following question. Suppose $f(z)$ is an analytic function on the horizontal strip $U = \{x + iy : a < y < b\}$, and continuous on \bar{U} . Can we assert that $\sup_U |f| = \sup_{\partial U} |f|$?

The answer is no, in general. However, the answer is yes if $f(x + iy)$ does not grow too rapidly as $|x| \rightarrow \infty$. In fact, this is a property of harmonic functions $u(z)$. The point is that if we truncate the strip to a rectangle by cutting along the lines where $|x| = R$, then the harmonic measure of the ends (as seen from a fixed point $z \in U$) tends to zero exponentially fast. Thus if $|u(x + iy)| = O(|x|^n)$ for some n , we get the desired control.

Runge’s theorem. Here is an interesting and perhaps surprising application of Cauchy’s formula.

Let $K \subset \mathbb{C}$ be a compact set, let $C(K)$ denote the Banach space of continuous functions with the sup-norm, and let $A(K)$, $R(K)$ and $P(K) \subset C(K)$ denote the closures of the analytic, rational and polynomial functions. *Note that all three of these subspaces are algebras.*

Example. For $K = S^1$ we have $R(K) = A(K) = C(K)$ by Fourier series, but $P(K) \neq C(K)$, since $\int_{S^1} p(z) dz = 0$ for any polynomial. In particular, $1/z \notin P(S^1)$.

Remarkably, if we remove a small interval from the circle to obtain an arc $K = \exp[0, 2\pi - \epsilon]$, then $1/z$ can be approximated by polynomials on K .

Theorem 1.37 (Runge) *For any compact set $K \subset \mathbb{C}$ we have $R(K) = A(K)$, and $P(K) = A(K)$ provided $\mathbb{C} - K$ is connected.*

Proof. For the first result, suppose $f(z)$ is analytic on a smoothly bounded neighborhood U of K . Then we can write

$$f(z) = \frac{1}{2\pi i} \int_{\partial U} \frac{f(t) dt}{t - z} = \int_{\partial U} F_z(t) dt.$$

Since $d(z, \partial U) \geq d(K, \partial) > 0$, the functions $\{F_z\}$ range in a compact subset of $C(\partial U)$. Thus we can replace this integral with a finite sum at the cost

of an error that is small independent of z . But the terms $f(t_i)/(t_i - z)$ appearing in the sum are rational functions of z , so $R(K) = A(K)$.

The second result is proved by pole-shifting. By what we have just done, it suffices to show that $f_p(z) = 1/(z - p) \in P(K)$ for every $p \notin K$. Let $E \subset \mathbb{C} - K$ denote the set of p for which this is true.

Clearly E contains all p which are sufficiently large, because then the power series for $f_p(z)$ converges uniformly on K . Also E is closed by definition. To complete the proof, it suffices to show E is open.

The proof that E is open is by ‘pole shifting’. Suppose $p \in E$, $q \in B(p, r)$ and $B(p, r) \cap K = \emptyset$. Note that $f_q(z)$ is analytic on $\mathbb{C} - B(p, r)$, and tends to zero as $|z| \rightarrow \infty$. Thus $f_q(z)$ can be expressed as a power series in $1/(z - p)$:

$$(z - q)^{-1} = \sum_0^\infty a_n(z - p)^{-n} = \sum a_n f_p(z)^n,$$

convergent for $|z - p| > |z - q|$, and converging uniformly on K . (Compare the expression

$$\frac{1}{z - 1} = \sum_{n=1}^\infty \frac{1}{z^n},$$

valid for $|z| > 1$.) Since $(z - q)^{-1} \rightarrow 0$ as $|z| \rightarrow \infty$, only terms with $n \geq 0$ occur on the right. But $f_p \in A(K)$ by assumption, and $A(K)$ is an algebra, so it also contains f_p^n . Thus $f_q \in A(K)$ as well. ■

Aside: Lavrentiev’s Theorem. It can be shown that if $\mathbb{C} - K$ is connected and K has no interior, then $A(K) = C(K)$. This is definitely false for a fat Swiss cheese: if ∂K is rectifiable and of finite length, then $\int_{\partial K} f(z) dz = 0$ for all f in $A(K)$, and so $A(K) \neq C(K)$. For more details see [Gam].

Applications: pointwise convergence. Runge’s theorem can be used to show easily that there is a sequence of polynomials $f_n(z)$ that converge pointwise, but whose limit is not even continuous. Indeed, let $A_n = [0, n]$ and let $B_1 \subset B_2 \subset \dots$ be an increasing sequence of compact sets such that $\bigcup B_n = \mathbb{C} - [0, \infty)$. Then every $z \in \mathbb{C}$ eventually belongs to A_n or B_n . Let $f_n(z)$ be a polynomial, whose existence is guaranteed by Runge’s theorem, such that $|f_n(z)| < 1/n$ on A_n and $|f_n(z) - 1| < 1/n$ on B_n . Then clearly $\lim f_n(z) = 0$ on $[0, \infty)$ and $\lim f_n(z) = 1$ elsewhere.

Applications: embedding the disk into affine space. Runge’s theorem can also be used to show there is a *proper* embedding of the unit disk into \mathbb{C}^3 . See [Re, §12.3].

2 The simply-connected Riemann surfaces

Riemann surfaces. A *Riemann surface* X is a *connected* complex 1-manifold. This means X is a Hausdorff topological space equipped with charts (local homeomorphisms) $f_i : U_i \rightarrow \mathbb{C}$, and the transition functions $f_{ij} = f_i \circ f_j^{-1}$ are analytic where defined.

It then makes sense to discuss analytic functions on X , or on any open subset of X . Technically we obtain a *sheaf of rings* \mathcal{O}_X with $\mathcal{O}_X(U)$ consisting of the analytic maps $f : U \rightarrow \mathbb{C}$.

Aside from \mathbb{C} and *connected* open sets $U \subset \mathbb{C}$, the first interesting Riemann surface (and the basic example of a *compact* Riemann surface) is the Riemann sphere $\hat{\mathbb{C}}$. The map $f : \hat{\mathbb{C}} - \{0\} \rightarrow \mathbb{C}$ given by $f(z) = 1/z$ provides a chart near infinity.

Meromorphic functions. There is also a natural notion of holomorphic (analytic) maps *between* Riemann surfaces. A *meromorphic function* on X is an analytic map $f : X \rightarrow \hat{\mathbb{C}}$ that is *not* identically ∞ .

Since X is *connected*, the ring $K(X)$ of all meromorphic functions on X forms a *field*, and the ring $\mathcal{O}(X)$ of all analytic functions forms an *integral domain*.

We will see that $K(\hat{\mathbb{C}}) = \mathbb{C}(z)$, while $\mathcal{O}(\hat{\mathbb{C}}) = \mathbb{C}$, so not every meromorphic function is a quotient of holomorphic functions. We will see that $\mathcal{O}(\mathbb{C})$ is much wilder (it contains $\exp(\exp z)$, etc.), and yet $K(\mathbb{C})$ is the field of fractions of $\mathcal{O}(\mathbb{C})$.

Classification. In principle the classification of Riemann surfaces is completed by the following result:

Theorem 2.1 (The Uniformization Theorem) *Every simply-connected Riemann surface is isomorphic to \mathbb{H} , \mathbb{C} or $\hat{\mathbb{C}}$.*

Then by the theory of covering spaces, an arbitrary Riemann surface satisfies $X \cong \hat{\mathbb{C}}$, $X = \mathbb{C}/\Gamma$, or $X = \hat{\mathbb{C}}/\Gamma$, where Γ is a group of automorphisms. Such Riemann surfaces are respectively *elliptic*, *parabolic* and *hyperbolic*. Their natural metrics have curvatures 1, 0 and -1 .

In this section we will discuss each of the simply-connected Riemann surfaces in turn. We will discuss their geometry, their automorphisms, and their proper endomorphisms.

2.1 The complex plane

An analytic map $f : X \rightarrow Y$ is *proper* if f^{-1} sends compact sets to compact sets. This is equivalent to the condition that $f(z_n) \rightarrow \infty$ whenever $z_n \rightarrow \infty$.

Proper analytic maps $f : X \rightarrow Y$ are the ‘tamest’ maps between Riemann surfaces. For example, they have the following properties:

1. If f is not constant, it is surjective.
2. If f' never vanishes, then f is a covering map.
3. If f' never vanishes and Y is simply-connected, then f is an isomorphism.

Example. The entire function $\exp : \mathbb{C} \rightarrow \mathbb{C}$ is not proper, since it omits the point 0.

Theorem 2.2 *An analytic function $f : \mathbb{C} \rightarrow \mathbb{C}$ is proper iff $f(z)$ is a non-constant polynomial.*

Proof. Clearly a polynomial of positive degree is proper. Conversely, if f is proper, then f has finitely many zeros, and so no zeros for $|z| > \text{some } R$. Then $g(z) = 1/f(1/z)$ is a nonzero, bounded analytic function for $0 < |z| < R$. Consequently $g(z)$ has a zero of finite order at $z = 0$. This gives $|g(z)| \geq \epsilon|z|^n$, and hence $|f(z)| \leq M|z|^n$. By Cauchy’s bound, f is a polynomial. ■

To say $f : \mathbb{C} \rightarrow \mathbb{C}$ is proper is the same as to say that f extends to a continuous function $F : \widehat{\mathbb{C}} \rightarrow \widehat{\mathbb{C}}$. The same argument shows, more generally, that if $f : X \rightarrow Y$ is a continuous map between Riemann surfaces, and f is analytic outside a discrete set $E \subset X$, then f is analytic.

Corollary 2.3 *The automorphisms of \mathbb{C} are given by the affine maps of the form $f(z) = az + b$, where $a \in \mathbb{C}^*$ and $b \in \mathbb{C}$.*

Thus $\text{Aut}(\mathbb{C})$ is a *solvable group*. If $f \in \text{Aut}(\mathbb{C})$ is fixed-point free, then it must be a translation. This shows:

Corollary 2.4 *Any Riemann surface covered by \mathbb{C} has the form $X = \mathbb{C}/\Lambda$, where Λ is a discrete subgroup of $(\mathbb{C}, +)$.*

Example. We have $\mathbb{C}/(z \mapsto z + 1) \cong \mathbb{C}^*$; the isomorphism is given by $\pi(z) = \exp(2\pi iz)$.

Degree. For a proper map $f : X \rightarrow Y$, the number of points in $f^{-1}(z)$, counted with multiplicity, is independent of z . This number is called the

degree of f . In the case $X = Y = \mathbb{C}$, it is the same as the degree of the polynomial f .

Metrics. A *conformal metric* on a Riemann surface is given in local coordinates by $\rho = \rho(z) |dz|$. We will generally assume that $\rho(z) \geq 0$ and ρ is continuous, although metrics with less regularity are also useful.

A conformal metric allows one to measure lengths of arcs, by

$$L(\gamma, \rho) = \int_{\gamma} \rho = \int_a^b \rho(\gamma(t)) |\gamma'(t)| dt;$$

and areas of regions, by

$$A(U, \rho) = \int_U \rho^2 = \int_U \rho^2(z) |dz|^2 = \int_U \rho^2(z) dx dy.$$

Metrics pull back under analytic maps by the formula:

$$f^*(\rho) = \rho(f(z)) |f'(z)| |dz|,$$

and satisfy natural formulas such as

$$L(\gamma, f^* \rho) = L(f(\gamma), \rho)$$

(so long as $f(\gamma)$ is understood as the parameterized path $\gamma \circ f$).

A metric allows us to intrinsically measure the size of the derivative a map $f : (X_1, \rho_1) \rightarrow (X_2, \rho_2)$: it is given by

$$\|Df\| = |f'|_{\rho} = \frac{|f'(z)| \rho_2(f(z))}{\rho_1(z)} = \frac{f^* \rho_2}{\rho_1}.$$

We have $|f'| = 1$ iff f is a local isometry.

Flat metrics. The Euclidean metric on the plane is given by $\rho = |dz|$. Since $|dz|$ is Λ -invariant, we find:

Theorem 2.5 *Every Riemann surface covered by \mathbb{C} admits a complete flat metric, unique up to scale.*

For example, on $\mathbb{C}^* \cong \mathbb{C}/\mathbb{Z}$ we have the *cylindrical metric* $\rho = |dz|/|z|$. Every circle $|z| = r$ has length 2π in this metric.

Cone metrics. Consider the metric $\rho = |z|^\alpha |dz|/|z|$ on \mathbb{C} . We claim this is a flat metric, making the origin into a cone point of total angle $\theta = 2\pi\alpha$.

To see this is plausible, note that the unit ball $B(0, 1)$ has radius $R = \int_0^1 t^\alpha dt/t = 1/\alpha$, and circumference $C = 2\pi$, so $C/R = 2\pi\alpha$.

Alternatively, let $f(z) = z^n$. Then we find

$$f^*(\rho) = |z|^{n\alpha} n |dz| / |z| = n |dz|$$

if $\alpha = 1/n$. Thus the case $\alpha = 1/n$ gives the quotient metric on $(\mathbb{C}, n|dz|)/\langle \zeta_n \rangle$, where $\zeta_n = \exp(2\pi i/n)$.

Orbifold quotients. We can also take the quotient of \mathbb{C} by the infinite dihedral group, $D_\infty = \langle z+1, -z \rangle \subset \text{Aut } \mathbb{C}$. The result is \mathbb{C} itself, which the quotient map given by $\cos(2\pi z)$.

Closely related is the important map

$$\pi : \mathbb{C}^* \rightarrow \mathbb{C}$$

given by $\pi(z) = z + 1/z$. This degree two map gives the orbifold quotient of \mathbb{C}^* by $z \mapsto 1/z$. It gives an intermediate covering space to the one above, if we regard \mathbb{C}^* as \mathbb{C}/\mathbb{Z} .

The map π sends circles to ellipse and radial lines to hyperboli, with foci $[-2, 2]$. In fact all conics with these foci arise in this way. (See Figure 1.)

By attaching cone angles to ± 2 , one turns \mathbb{C} into the $(2, 2, \infty)$ orbifold X . A loop around a cone point of order n has order n in the orbifold fundamental group, so $\pi_1(X) = \mathbb{Z}/2 * \mathbb{Z}/2$. In general, by broadening the scope of covering spaces and deck groups to include orbifolds and maps with fixed points, we enrich the supply of Euclidean (and other) Riemann surfaces. For example, the $(3, 3, 3)$ orbifold is also Euclidean — it is the double of an equilateral triangles.

Chebyshev polynomials. Now let $S_n(z) = z^n$. Since $S_n(1/z) = 1/S_n(z)$, there is a sequence of polynomials $P_1(z) = z$, $P_2(z) = z^2 - 2$, $P_3(z) = z^3 - 3z$, etc. satisfying

$$P_n(z + 1/z) = (z + 1/z)^n,$$

or equivalently

$$P_n(\pi(z)) = \pi(S_n(z)).$$

These *Chebyshev* polynomials are related to multiple angle formulas for cosine, since for $z = e^{i\theta}$ we have $\pi(z) = 2 \cos \theta$, and thus:

$$P_n(2 \cos \theta) = 2 \cos(n\theta).$$

Since $S_n(z)$ preserves circles and radial lines, we have:

Theorem 2.6 *The Chebyshev polynomials preserve the hyperbolas and ellipses with foci at ± 2 .*

In particular $P_n : [-2, 2] \rightarrow [-2, 2]$ by degree n , and $P_n^k(z) \rightarrow \infty$ for all other $z \in \mathbb{C}$. (Thus the Julia set of P_n is $[-2, 2]$.)

Solving the cubic. Complex algebra finds its origins in the work of Cardano et al on solving cubic polynomial equations. Remarkably, complex numbers intervene even when the root to be found is real.

One can always make a simple transformation of the form $x \mapsto x + c$ to reduce to the form

$$x^3 + ax + b = 0.$$

One can further replace x with cx to reduce to the form

$$x^3 - 3x = b.$$

Thus the solution to the cubic involves inverting a *single* cubic function $P_3(z) = z^3 - 3z$.

But to solve $P_n(x) = a$, we just write $a = y + 1/y$ (by solving a quadratic equation), and then we have $x = y^{1/n} + y^{-1/n}$. In particular, this method can be used to solve $x^3 - 3x = b$.

Classification of polynomials. Let us say $p(z)$ is equivalent to $q(z)$ if there are $A, B \in \text{Aut}(\mathbb{C})$ such that $Bp(Az) = q(z)$. Then every polynomial is equivalent to one which is monic and centered (the sum of its roots is zero). Every quadratic polynomial is equivalent to $p(z) = z^2$.

The reasoning above shows, every cubic polynomial with distinct critical points is equivalent to $P_3(z) = z^3 - 3z$. Otherwise it is equivalent to z^3 . But for degree 4 polynomials we are in new territory: the cross-ratio of the 3 critical points, together with infinity, is an invariant.

It is a famous fact (proved using Galois theory) that a general quintic polynomial (with integral coefficients) cannot be solved by radicals.

2.2 The Riemann sphere

We now turn to the Riemann sphere $\widehat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$. We will examine the sphere from several perspectives: as a Riemann surface, as a round sphere, as the projectively line, and as the boundary of hyperbolic space.

The projective line. We have a natural projection

$$\pi : \mathbb{C}^2 - \{(0, 0)\} \rightarrow \widehat{\mathbb{C}} \cong \mathbb{P}^1$$

given by $\pi(z_0, z_1) = z_0/z_1$; it records the slope of each line. This gives a natural identification of $\widehat{\mathbb{C}}$ with the *projective line* \mathbb{P}^1 , i.e. the space of lines in \mathbb{C}^2 .

Aside: Some topology of projective spaces. The real projective plane \mathbb{RP}^2 is the union of a disk and a Möbius band. The natural map $p : \mathbb{C}^2 - \{(0, 0)\} \rightarrow \widehat{\mathbb{C}}$ factors through the Hopf map $S^3 \rightarrow S^2$, whose fibers have linking number one. This map generates $\pi_3(S^2)$.

Möbius transformations. So long as $ad - bc \neq 0$, the map $f(z) = (az + b)/(cz + d)$ defines an automorphism of $\widehat{\mathbb{C}}$. Its inverse can be found by inverting the matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$. In fact, the map π above transports the linear action of $\mathrm{SL}_2(\mathbb{C})$ on \mathbb{C}^2 to the fractional linear action of $\mathrm{PSL}_2(\mathbb{C})$ on $\widehat{\mathbb{C}}$.

Theorem 2.7 *We have $\mathrm{Aut}(\widehat{\mathbb{C}}) = \mathrm{PSL}_2(\mathbb{C})$.*

Proof. Given $f \in \mathrm{Aut}(\widehat{\mathbb{C}})$, pick $g \in \mathrm{PSL}_2(\widehat{\mathbb{C}})$ so $g(f(\infty)) = \infty$. Then $g \circ f \in \mathrm{Aut}(\mathbb{C}) \subset \mathrm{PSL}_2(\mathbb{C})$, so $f \in \mathrm{PSL}_2(\mathbb{C})$. ■

Corollary 2.8 *The action of $\mathrm{Aut}(\widehat{\mathbb{C}})$ is uniquely triply-transitive: any distinct triple of points can be sent to $(0, 1, \infty)$ by a unique Möbius transformation.*

Corollary 2.9 *The group $\mathrm{PSL}_2(\mathbb{C})$ is isomorphic, as a complex manifold, to the space of distinct triples of points on $\widehat{\mathbb{C}}$.*

Classification of automorphisms up to conjugacy. There is a natural map $\mathrm{tr} : \mathrm{PSL}_2(\mathbb{C}) \rightarrow \mathbb{C}/\langle \pm 1 \rangle$. This is clearly a class function (it is constant on conjugacy classes), and in fact the conjugation class is almost determined by this map. Namely we have the following classes:

- (1) The identity map, $\mathrm{tr}(A) = \pm 2$.
- (2) Parabolics, $A(z) = z + a$; $\mathrm{tr}(A) = \pm 2$.
- (3) Elliptics: $A(z) = e^{i\theta}z$; $\mathrm{tr}(A) = 2 \cos \theta/2 \in [-2, 2]$.
- (4) Hyperbolics: $A(z) = e^t z$, $\mathrm{Re} t \neq 0$; $\mathrm{tr}(A) = 2 \cosh(t/2)$.

Note that a Möbius transformation (other than the identity) either has two simple fixed points, or a single fixed point of multiplicity two. The fixed points of A correspond to its eigenvectors on \mathbb{C}^2 .

Note also that all these elements, except for irrational elliptics, generate discrete subgroups of $\mathrm{Aut}(\widehat{\mathbb{C}})$.

If A has eigenvalues $\lambda^{\pm 1}$, then $\mathrm{tr}(A) = \lambda + 1/\lambda$. Our previous analysis of this map shows, for example, that the traces of matrices a given value for $|\lambda|$ correspond to an ellipse with foci ± 2 .

Cross-ratios. The *cross-ratio* is an invariant of ordered 4-tuples of points, characterized by the conditions that $[a : b : c : d] = [ga : gb : gc : gd]$ for all $g \in \text{Aut}(\widehat{\mathbb{C}})$, and $[0 : 1 : \infty : \lambda] = \lambda \in \widehat{\mathbb{C}} - \{0, 1, \infty\}$. The cross-ratio gives an explicit isomorphism between the *moduli space* $\mathcal{M}_{0,4}$ and the triply-punctured sphere.

Stereographic projection. There is a geometric identification between the unit sphere $S^2 \subset \mathbb{R}^3$ and the Riemann sphere $\widehat{\mathbb{C}}$ in which the north pole N becomes ∞ and the rest of the sphere is projected linearly to $\mathbb{C} = \mathbb{R}^2 \times \{0\}$.

Theorem 2.10 *Stereographic projection is a conformal map that sends circles to circles.*

Proof of conformality. Consider two vectors at a point $p \neq N$ on S^2 . Construct a pair of circles tangent to these vectors at p and passing through N . Then these circles meet in the same angles at p and N . On the other hand, each circle is the intersection of the sphere with a plane. These planes meet \mathbb{C} in the same angle they meet a plane tangent to the sphere at the north pole N . Thus stereographic projection preserves angles. ■

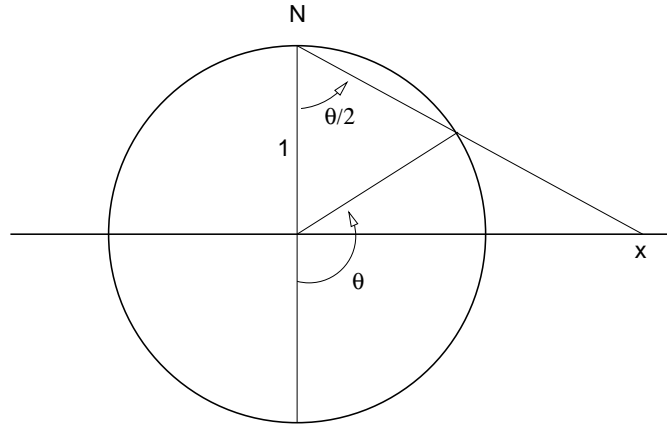


Figure 4. Stereographic projection on S^1 .

Theorem 2.11 *The Euclidean metric on S^2 is transported, by stereographic projection, to the metric $2|dz|/(1 + |z|^2)$ on $\widehat{\mathbb{C}}$.*

Proof. Reducing dimensions by one, we obtain the stereographic projection $p : S^1 - N \rightarrow \mathbb{R}$. If $\theta \in [-\pi, \pi]$ is the angle on the sphere normalized so

$\theta(N) = \pi$, then we find $x = p(\theta) = \tan(\theta/2)$. Thus $dx = 2 \sec^2(\theta/2) d\theta = 2(1+x^2)$, which gives $d\theta = 2dx/(1+x^2)$. The case of S^2 follows by conformality and rotation invariance. ■

The unitary point of view. Here is an alternative perspective on the spherical metric. Let $\langle v, w \rangle = v_0 \bar{w}_0 + v_1 \bar{w}_1$ be the usual Hermitian metric on \mathbb{C}^2 . Define the angle θ between two lines $\mathbb{C}v, \mathbb{C}w \subset \mathbb{C}^2$ by

$$\cos(\theta/2) = \frac{|\langle v, w \rangle|}{|v||w|}.$$

This θ agrees with the spherical distance on $\widehat{\mathbb{C}}$. (To check this, consider the points 1 and $\exp(i\alpha)$ in S^1 . Their spherical distance is α , and if we set $v = [1, 1]$ and $w = [e^{i\theta/2}, e^{-i\theta/2}]$, then $|v||w| = 2$ and $\langle v, w \rangle = 2\cos(\alpha/2)$.)

Given this agreement, it is now clear that the group $\mathrm{SU}(2) \subset \mathrm{SL}_2(\mathbb{C})$ acts *isometrically* on $\widehat{\mathbb{C}}$. In fact this group is the full group of orientation-preserving isometries.

Theorem 2.12 *The isometry group of $\widehat{\mathbb{C}}$ in the spherical metric is given by*

$$\mathrm{Isom}^+(S^2) = \left\{ \begin{pmatrix} a & b \\ -\bar{b} & \bar{a} \end{pmatrix} : |a|^2 + |b|^2 = 1 \right\} \subset \mathrm{PSL}_2(\mathbb{C}).$$

Note that these matrices satisfy $AA^* = I$.

Exercise. What is the distance, in the spherical metric, from 1 to $1+i$? For $[1 : 1]$ and $[1+i : 1]$ we get

$$\cos \theta/2 = \sqrt{5}/\sqrt{6},$$

and hence $\theta = 2 \arccos(\sqrt{5/6}) \approx 0.841069 \dots$

Circles and lines. A *circle* $C \subset \widehat{\mathbb{C}}$ is either a line through ∞ , or an ordinary Euclidean circle in \mathbb{C} .

Theorem 2.13 *Möbius transformations send circles to circles, and any two circles are equivalent.*

Proof 1. A circle $x^2 + y^2 + Ax + By + C = 0$ is also given by $r^2 + r(A \cos \theta + B \sin \theta) + C = 0$, and it is easy to transform the latter under $z \mapsto 1/z$, which replaces r by $1/r$ and θ by $-\theta$.

Proof 2. A circle is the same, projectively, as the set of null vectors for a Hermitian form on \mathbb{C}^2 of signature $(1, 1)$. Any two such forms are, up to scale, related by an element of $\mathrm{SL}_2(\mathbb{C})$. ■

Area of triangles. The geodesics on S^2 are arcs of great circles. Their angle sum always exceeds π , and in fact we have:

Theorem 2.14 *The area of a spherical triangle is equal to its excess angle, $\alpha + \beta + \gamma - \pi$.*

Proof. Since the sphere has area 4π , a lune of angle θ has area 4θ . The three lunes coming from a given triangle T cover the whole sphere, with points inside T and its antipode triply covered and the rest simply covered. Thus $4\pi + 4\text{area}(T) = 4(\alpha + \beta + \gamma)$. ■

This result is a special case of the Gauss-Bonnet formula:

$$2\pi\chi(X) = \int_X K + \int_{\partial X} k.$$

The general formula can be deduced from the constant curvature formulas (for $K = \pm 1$) by subdivision.

2.3 The hyperbolic plane

We now turn to a discussion of the hyperbolic plane $\mathbb{H} = \{z \in \mathbb{C} : \text{Im}(z) > 0\}$. This is the most versatile of the simply-connected Riemann surfaces. Note that \mathbb{H} is isomorphic to Δ , e.g. by the Möbius transformation $z \mapsto \frac{z-i}{z+i}$ which sends $(i, 0, \infty)$ to $(0, -1, 1)$. One often uses these two models interchangeably.

First, observe that the Schwarz reflection principle implies:

Theorem 2.15 *Every automorphism of \mathbb{H} or Δ extends to an automorphism of $\hat{\mathbb{C}}$.*

Corollary 2.16 *$\text{Aut}(\mathbb{H})$ corresponds to the subgroup $\text{SL}_2(\mathbb{R}) \subset \text{SL}_2(\mathbb{C})$.*

Proof. Suppose $g(z) = (az + b)/(cz + d)$ preserves \mathbb{H} . Then g preserves $\partial\mathbb{H} = \hat{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$, so we can assume the coefficients of g are real. Then $\text{Im } g(i) = \text{Im}(ai + b)(-ci + d)/|ci + d|^2 = (ad - bc)/|ci + d|^2 > 0$, so $\det(g) > 0$. Thus we can be further rescale by $1/\sqrt{\det(g)}$ so that $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \text{SL}_2(\mathbb{R})$. Conversely, every Möbius transformation represented by a matrix in $\text{SL}_2(\mathbb{R})$ preserves \mathbb{H} . ■

Corollary 2.17 $\text{Aut}(\Delta)$ corresponds to the subgroup $\text{SU}(1, 1)$ of isometries of the form $|Z|^2 = |Z_0|^2 - |Z_1|^2$.

Proof. This can be deduced from the preceding result by a change of coordinates. Note that the vectors in \mathbb{C}^2 with $|Z|^2 = 0$ correspond to $\partial\Delta$, and Δ itself corresponds to the cone $|Z|^2 < 0$. ■

Corollary 2.18 The automorphism group of the disk is given by the subgroup

$$\text{Aut}(\Delta) = \left\{ \begin{pmatrix} a & b \\ \bar{b} & \bar{a} \end{pmatrix} : |a|^2 - |b|^2 = 1 \right\} \subset \text{PSL}_2(\mathbb{C}).$$

Note that these matrices satisfy $AQA^* = Q$, where $Q = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$.

Hyperbolic geometry. The *hyperbolic metric* on \mathbb{H} is given by $\rho = |dz|/\text{Im } z = |dz|/y$. On the unit disk Δ , the hyperbolic metric becomes $\rho = 2|dz|/(1 - |z|^2)$.

Theorem 2.19 Every automorphism of \mathbb{H} is an isometry for the hyperbolic metric.

Proof. This is clear for automorphisms of the form $g(z) = az + b$, and can be easily checked by $g(z) = -1/z$. These two types of automorphisms generate $\text{Aut}(\mathbb{H})$. ■

The *geodesics* for the hyperbolic metric are given by circles orthogonal to the boundary (of \mathbb{H} or Δ). For example, in the case of the imaginary axis $\gamma = i\mathbb{R}_+ \subset \mathbb{H}$, it is easy to see that the projection $\pi : \mathbb{H} \rightarrow \gamma$ given by $\pi(x, y) = (0, y)$ is distance-decreasing, and so γ gives a shortest path between any two of its points. The general case follows from this one, since any circle orthogonal to the boundary of \mathbb{H} is equivalent, under $\text{Aut}(\mathbb{H})$, to γ .

These geodesics satisfy all of Euclid's postulates except the fifth. Thus if we declare them to be straight lines, we find:

Theorem 2.20 Euclid's fifth postulate cannot be deduced from the other axioms of geometry.

Triangles. Gauss-Bonnet for hyperbolic triangles reads:

$$\text{area}(T) = \pi - \alpha - \beta - \gamma.$$

For example, an *ideal triangle* in \mathbb{H} has vertices at infinity and internal angles of 0. Its area is π .

Curvature. We remark that the Gauss curvature of a conformal metric $\rho(z) |dz|$ is readily computed in terms of the Laplacian of ρ : we have

$$K(\rho) = -\rho^{-2} \Delta \log \rho.$$

Note that this expression is invariant under change of coordinates, since an analytic function satisfies $\Delta \log |f'| = 0$. It is particularly easy to check that $K(1/y) = -1$, so $(\mathbb{H}, |dz|/y)$ has constant curvature -1 .

Classification of isometries. Let $f : \mathbb{H} \rightarrow \mathbb{H}$ be an automorphism. Then f is an isometry, and we can define its *translation distance* by

$$\tau(f) = \inf_{\mathbb{H}} d(x, f(x)).$$

This invariant is useful for the classification of isometries. Here are the possibilities:

1. (Elliptic) $\tau(f) = 0$, achieved. Then f is conjugate to a rotation fixing i , of the form

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}.$$

(Note that when $\theta = \pi$ above, the matrix is $-I$ and so $f(z) = z$.) In the disk model, we can simply conjugate so $f(z) = e^{2i\theta} z$.

2. (Hyperbolic) $\tau(f) > 0$. In this case $\tau(f)$ is achieved along a unique geodesic $\gamma \subset \mathbb{H}$ which is translated distance $\tau(f)$ by f . Up to conjugacy, $f(z) = e^t z$ where $t = \tau(f)$.
3. (Parabolic) $\tau(f) = 0$, not achieved. Then f is conjugate to $f(z) = z + 1$.

The Schwarz Lemma. The following fundamental fact is central to the theory of analytic maps between hyperbolic Riemann surfaces.

Theorem 2.21 *Let $f : \Delta \rightarrow \Delta$ be an analytic function such that $f(0) = 0$. Then $|f'(0)| \leq 1$ and $|f(z)| \leq |z|$ for all z .*

Equality holds (for some $z \neq 0$) iff $f(z) = e^{i\theta}$ is a rotation.

Proof. The maximum principle implies that the analytic function $g(z) = f(z)/z$ satisfies $|g(z)| \leq 1$. If equality holds, then g is constant. ■

Corollary 2.22 *A holomorphic map $f : \Delta \rightarrow \Delta$ is either an isometry, or it is distance decreasing for the hyperbolic metric. (This means $|f'|_\rho < 1$ and $d(f(x), f(y)) < d(x, y)$ if $x \neq y$.)*

Remark. The Schwarz lemma, instead of Schwarz reflection, can also be used to show every automorphism of \mathbb{H} or Δ extends to $\widehat{\mathbb{C}}$.

The hyperbolic metric on other Riemann surfaces. Now suppose, more generally, that $X = \mathbb{H}/\Gamma$ is a *hyperbolic Riemann surface*. (Almost all Riemann surfaces are of this type — only $\widehat{\mathbb{C}}, \mathbb{C}, \mathbb{C}^*$ and \mathbb{C}/Λ not hyperbolic. In particular, every region $U \subset \mathbb{C}$ other than \mathbb{C} itself is hyperbolic.)

Since $\rho = |dz|/y$ is invariant under $\text{Aut}(\mathbb{H})$, it *descends* to give a *hyperbolic metric* on X itself. (Also called the Poincaré metric.)

Example. The hyperbolic metric on the strip $S = \{z = x+iy : 0 < |y| < \pi\}$ is given by $\rho = |dz|/\sin(y)$. Indeed, the map $f(z) = e^z$ sends S conformally to \mathbb{H} , and so

$$\rho = f^*(|dz|/y) = \frac{|e^z| |dz|}{\text{Im } e^z} = \frac{e^x |dz|}{e^x \sin(y)} = \frac{|dz|}{\sin(y)}.$$

The more general version of the Schwarz lemma then reads:

Theorem 2.23 *Let $f : X \rightarrow Y$ be an analytic map between hyperbolic Riemann surfaces. Then f is either a covering map, or it is a contraction for the hyperbolic metric.*

Proof. Pass to the universal covers of the domain and range and apply the usual Schwarz lemma. ■

Proper maps. A *Blaschke product* $B : \Delta \rightarrow \Delta$ is a rational map of the form

$$f(z) = e^{i\theta} \prod_1^d \frac{z - a_i}{1 - \bar{a}_i z}$$

with zeros satisfying $|a_i| < 1$. Using the fact that $1 = z\bar{z}$ on S^1 , it is easy to show that $|B(z)| = 1$ on S^1 , and thus $B : \Delta \rightarrow \Delta$ is proper. Conversely, we have:

Theorem 2.24 *Every proper analytic map $f : \Delta \rightarrow \Delta$ is a Blaschke product.*

Proof. A proper map is surjective, so f has at least one zero, say a . Let $M(z) = (z - a)/(1 - \bar{a}z)$. Then $g = f(z)/M(z) : \Delta \rightarrow \overline{\Delta}$ is analytic, with one fewer zero than f . The proof now follows by induction on the degree of f . ■

Dynamics and the Schwarz Lemma.

Theorem 2.25 *Let $f : \hat{\mathbb{C}} \rightarrow \hat{\mathbb{C}}$ be a rational map of degree $d > 1$. Then the immediate basin of any attracting cycle contains a critical point.*

Corollary 2.26 *The map f has at most $2d - 2$ attracting cycles.*

Theorem 2.27 (Denjoy-Wolff) *Let $f : \Delta \rightarrow \Delta$ be a holomorphic map which is not conjugate to $z \mapsto e^{i\theta}z$. Then there exists a point $p \in \overline{\Delta}$ such that $f^n(z) \rightarrow p$ for all $z \in \Delta$.*

3 Entire and meromorphic functions

This section discusses general constructions of functions on \mathbb{C} with given zeros and poles, and analyzes special cases such as the trigonometric functions and $\Gamma(z)$.

3.1 Infinite products

To start we recall that $\prod(1 + a_n)$ converges (to an element of \mathbb{C}^*) iff $\sum |a_n|$ converges. (Note: we say $\prod(1 - 1/n)$ *diverges to zero*.)

Similarly, $\prod(1 + f_n(z))$ defines an entire function on \mathbb{C} iff $\sum |f_n(z)|$ converges uniformly on compact sets.

Functions with infinitely many zeros. Given $a_n \rightarrow \infty$ in \mathbb{C} , can one find an analytic function $f(z)$ whose zeros are exactly these points? Is this function essentially unique, given e.g. some constraints on its growth?

For example, is $\sin(\pi z)$ simply a ‘polynomial’ with roots at all points of \mathbb{Z} ?

Polynomials work for finitely many zeros. We could try to address the case of infinitely many zeros by defining $f(z) = \prod(z - a_i)$. But this product has no chance of converging. A better choice – assuming $a_i \neq 0$ – is $\prod(1 - z/a_i)$. And this works: we have

Theorem 3.1 *If $\sum 1/|a_n|$ is finite, then $f(z) = \prod(1 - z/a_n)$ defines an entire function with zeros exactly at these points.*

Multiplicities of zeros can also be specified – they just correspond to repetitions of the same number in the sequence a_i .

However this result is still too weak to address the case $a_n = n$ needed for constructing $\sin(\pi z)$.

Weierstrass factors. Here is an elegant expression for an entire function with a zero only at $z = 1$, which is also close to 1 for $|z| < 1$. It is called the *Weierstrass factor* of order p :

$$E_p(z) = (1 - z) \exp \left(z + \frac{z^2}{2} + \cdots + \frac{z^p}{p} \right).$$

By convention, $E_0(z) = (1 - z)$.

The idea behind this expression is that $\log(1/(1 - z)) = z + z^2/2 + z^3/3 + \cdots$, and hence the two terms ‘almost cancel’ to give $(1 - z)/(1 - z) = 1$. (For more insight, compute the logarithmic derivative E'_p/E_p or see the proof below.)

Since we have truncated at the term z^p , it is easy to see that for $|z| < 1/2$ (say) we have

$$|E_p(z) - 1| = O(|z|^{p+1}).$$

For many purposes this bound is sufficient, however it is sometimes useful to have a bound which works for any $z \in \Delta$ and where the implicit constant (which might depend on p) is explicit. Such a bound is provided by:

Theorem 3.2 *For $|z| < 1$, we have $|E_p(z) - 1| \leq |z|^{p+1}$.*

Proof. The negative of the logarithmic derivative of E_p , we obtain:

$$-E'_p(z)/E_p(z) = 1/(1 - z) - 1 - z - \cdots - z^{p-1} = z^p/(1 - z).$$

Hence for all z we have

$$-E'_p(z) = z^p \exp(z + z^2/2 + \cdots + z^p/p) = z^p \sum_{k=0}^{\infty} a_k z^k,$$

with $a_k \geq 0$ for all k . Integrating term by term and using the fact that $E_p(0) = 1$, we find

$$1 - E_p(z) = z^{p+1} \sum_{k=0}^{\infty} b_k z^k$$

with $b_k \geq 0$. We also have $\sum b_k = 1 - E_p(1) = 1$, and hence for $|z| < 1$ we have

$$|1 - E_p(z)| \leq |z|^{p+1} \sum b_k = |z|^{p+1}.$$

■

Theorem 3.3 *For any sequence of nonzero complex numbers $a_n \rightarrow \infty$, the formula $f(z) = \prod_1^\infty E_n(z/a_n)$ converges for all z and defines an entire analytic function with zero set exactly (a_n) .*

Proof. The previous estimate yields convergence of the tail of the series: for all $z \in B(0, R)$, we have:

$$\sum_{|a_n| > 2R} |1 - E_n(z/a_n)| \leq \sum_1^\infty (|z|/|a_n|)^{n+1} < \sum_1^\infty (R/2R)^{n+1} < \infty.$$

■

Canonical products. In general using $E_n(z/a_n)$ is overkill. Then the same argument shows:

Theorem 3.4 *If $\sum 1/|a_n|^{p+1} < \infty$, then $P(z) = \prod E_p(z/a_n)$ defines an entire analytic function.*

Proof. For $|z| < R$ we have

$$\sum_{|a_n| > 2R} |1 - E_p(z/a_n)| \sum_1^\infty (|z|/|a_n|)^{p+1} < \infty. \leq R^{p+1} \sum 1/|a_n|^{p+1} < \infty.$$

■

If $p \geq 0$ is the *least integer* such that $\sum 1/|a_n|^{p+1} < \infty$, then we say $P(z) = \prod E_p(z/a_n)$ is the *canonical product* associated to (a_n) .

The counting function. It is also useful to introduce the *counting function* $N(r) = |\{n : |a_n| < r\}|$. Then $r^{-\beta} N(r)$ is a rough approximation to $\sum_{|a_n| < r} 1/|a_n|^\beta$. Consequently we have:

$$\alpha = \limsup_{r \rightarrow \infty} \frac{\log N(r)}{\log r}.$$

In more detail: suppose $N(r) \leq r^\beta$; then we can collect the points a_n into groups where $2^n < |a_n| \leq 2^{n+1}$; then

$$\sum |a_n|^{-\alpha} \leq \sum (2^n)^{-\alpha} N(2^{n+1}) \leq 2^\beta \sum 2^{n(\beta-\alpha)} < \infty$$

if $\alpha > \beta$. So β is an upper bound for the critical exponent. Similarly, if $N(r) \geq r^\beta$, then β is a lower bound, since $\sum |a_n|^{-\alpha} \geq r^{-\alpha} N(r) \geq r^{\beta-\alpha} \rightarrow \infty$ if $\alpha < \beta$.

Observe that knowledge of $N(r)$ is the same as knowledge of $r_n = |a_n|$ for all n . Thus we can also express functions of r_n in terms of $N(r)$. A typical example is:

$$\sum |a_n|^{-\alpha} = \int_0^\infty N(r) \alpha r^{-\alpha} \frac{dr}{r}.$$

A similar expression will arise in connection with Jensen's formula.

Entire functions of finite order. An entire function $f : \mathbb{C} \rightarrow \mathbb{C}$ is of *finite order* if there is a $\rho > 0$ such that $|f(z)| = O(\exp |z|^\rho)$. The infimum of all such ρ is the *order* $\rho(f)$. It is standard to denote the maximum and minimum of $|f|$ on $|z| = r$ by $M(r)$ and $m(r)$. Thus the order f is given by

$$\rho(f) = \limsup_{r \rightarrow \infty} \frac{\log \log M(r)}{\log r}.$$

Examples: Polynomials have order 0; $\sin(z)$, $\cos(z)$ and $\exp(z)$ have order 1; $\cos(\sqrt{z})$ has order $1/2$; $E_p(z)$ has order p ; $\exp(\exp(z))$ has infinite order.

Hadamard's 3-circles theorem. Here is useful general property of the function $M(r) = \sup_{|z|=r} |f(z)|$. The following result applies not just to entire functions, but also to functions analytic in an annulus of the form $r_1 < |z| < r_2$.

Theorem 3.5 *For any analytic function $f(z)$, the quantity $\log M(r)$ is a convex function of $\log r$.*

Proof. A function $\phi(s)$ of one real variable is convex if and only if $\phi(s) + ar$ satisfies the maximum principle for any constant a . This holds for $\log M(\exp(s))$ by considering $f(z)z^a$ locally. ■

Corollary 3.6 *We have $M(\sqrt{rs}) \leq \sqrt{M(r)M(s)}$.*

The convex functions satisfying $F(\log r) = \log M(r)$ look roughly linear for polynomials, e.g. like $F(x) = \deg(f)x + c$, and look roughly exponential for functions of finite order, e.g. $F(x) = \exp(\rho(f)x)$.

Hadamard's factorization theorem. We can now state a formula that describes every entire function of finite order in terms of its zeros and an additional polynomial.

Theorem 3.7 (Hadamard) *A entire function $f(z) \neq 0$ of finite order ρ can be uniquely expressed in the form:*

$$f(z) = z^m \prod E_p(z/a_n) e^{Q(z)},$$

where (a_n) are the zeros of f , $p \geq 0$ is the least integer such that $\sum 1/|a_n|^{p+1} < \infty$, and $Q(z)$ is a polynomial of degree q . We have $p, q \leq \rho$.

The number p is called the *genus* of f . Note that ordinary polynomials arise as a special case, where $p = q = \rho = 0$.

Remark. This theorem shows that the zeros of f determine f up to finitely many additional constants, namely the coefficients of $Q(z)$. It is tempting to conclude then that if $f(z)$ has no zeros, it is determined by its values at any $\lfloor \rho + 1 \rfloor$ points. This is not quite true, however, since $f(z)$ only determines $Q(z) \bmod 2\pi i\mathbb{Z}$; for example, $\exp(2\pi iz)$ and the constant function 1 agree on the integers.

It is, however, true that $f'/f = Q'$ in this case, so knowing the logarithmic derivative at enough points almost determines Q . This shows:

Corollary 3.8 *Suppose $f(z)$ and $g(z)$ are entire functions of order ρ with the same zeros, and $f'/f = g'/g$ at $\lfloor \rho \rfloor$ distinct points (where neither function vanishes). Then f is a constant multiple of g .*

I. Entire functions without zeros. The simplest case of Hadamard's theorem in which there is no canonical product, is handled by:

Theorem 3.9 *Let $f(z)$ be an entire function of finite order with no zeros. Then $f(z) = e^{Q(z)}$, where $Q(z)$ is a polynomial of degree d , and $\rho(f) = d$.*

To prove it we strengthen our earlier characterization of polynomials by $M(r) = O(r^d)$.

Lemma 3.10 *Let $Q(z)$ be an entire function satisfying $\operatorname{Re} Q(z) \leq A|z|^d + B$ for some $A, B > 0$. Then Q is a polynomial of degree at most d .*

Proof. There is a constant $C > 0$ such that for $R > 1$, Q maps $\Delta(2R)$ into the half-plane $U(R) = \{z : \operatorname{Re} z < CR^d\}$. By the Schwarz Lemma, Q is distance-decreasing from the hyperbolic metric on $\Delta(2R)$ to the hyperbolic metric on $U(R)$. Since $\Delta(R) \subset \Delta(2R)$ has bounded hyperbolic diameter, the same is true for $Q(\Delta(R)) \subset U(R)$. Therefore in the Euclidean metric,

$$\operatorname{diam} Q(\Delta(R)) = O(d(Q(0), \partial U(R))) = O(R^d).$$

This shows $|Q(z)| = O(|z|^d)$ for $|z| > 1$, and hence Q is a polynomial of degree at most d . ■

Proof of Theorem 3.9. Since f has no zeros, $f(z) = e^{Q(z)}$ for some entire function $Q(z)$. Since f has finite order, $|f(z)| = O(e^{|z|^d})$ for some d , and thus $\operatorname{Re} Q(z) \leq |z|^d + O(1)$; now apply the Lemma above. ■

II. Functions with zeros. Next we analyze entire functions with zeros. We will show:

Theorem 3.11 *Let $f(z)$ be an entire function of order ρ with zeros (a_n) . Then $\sum 1/|a_n|^{\rho+\epsilon} < \infty$ for all $\epsilon > 0$.*

In general, a sequence $a_n \rightarrow \infty$ has a *critical exponent* α , the least number such that $\sum 1/|a_n|^{\alpha+\epsilon} < \infty$. The result above states that the critical exponent of the zeros of f satisfies $\alpha \leq \rho(f)$. Thus if $p = \lfloor \rho(f) \rfloor$, then $p+1 > \rho(f) \geq \alpha$, and hence $\sum |a_n|^{p+1} < \infty$. This gives:

Corollary 3.12 *The genus of f satisfies $p \leq \rho(f)$.*

Jensen's formula. Informally, the result above says that if f has many zeros, then $M(r)$ must grow rapidly. The relation between the size of f and its zeros is made precise by the following important result:

Theorem 3.13 (Jensen's formula) *Let $f(z)$ be a holomorphic function on $B(0, R)$ with zeros a_1, \dots, a_n . Then:*

$$\operatorname{avg}_{S^1(R)} \log |f(z)| = \log |f(0)| + \sum \log \frac{R}{|a_i|}.$$

Proof. We first note that if f has no zeros, then $\log |f(z)|$ is harmonic and the formula holds. Moreover, if the formula holds for f and g , then it holds for fg ; and the case of general R follows from the case $R = 1$.

Next we verify that the formula holds when $f(z) = (z - a)/(1 - \bar{a}z)$ on the unit disk, with $|a| < 1$. Indeed, in this case $\log |f(z)| = 0$ on the unit circle, and $\log |f(0)| + \log(1/|a|) = \log |a/a| = 0$ as well.

The general case now follows, since a general function $f(z)$ on the unit disk can be written in the form $f(z) = g(z) \prod (z - a_i)/(1 - \bar{a}_i z)$, where $g(z)$ has no zeros. ■

Remark. The physical interpretation of Jensen's formula is that $\log |f|$ is the potential for a set of unit point charges at the zeros of f .

Counting zeros. Here is another way to write Jensen's formula. Let $N(r)$ be the number of zeros of f inside the circle of radius r . Then:

$$\int_0^R N(r) \frac{dr}{r} = \text{avg}_{S^1(R)} \log |f(z)| - \log |f(0)|.$$

We can now show $\alpha \leq \rho$.

Proof of Theorem 3.11. Since $N(r)$ is an increasing function, by integrating from $r/2$ to r we find

$$N(r/2) \log(r/2) \leq \log M(r) + O(1),$$

and hence $\alpha = \limsup \log N(r)/\log r \leq \limsup \log M(2r)/\log(2r) = \rho$. ■

Of course functions can also grow rapidly without having any zeros. But then Jensen's formula shows the average of $\log |f|$ is *constant* over every circle $|z| = R$; so if f is large over much of the circle, it must also be close to zero somewhere on the same circle. (For example, e^z is very large over half the circle $|z| = R$, and very small over the rest.)

III. Canonical products. Our next task is to determine the order of a canonical product. It will also be useful, to complete the proof of Hadamard's theorem, to obtain *lower bounds* for such a product.

Theorem 3.14 *Let α be the critical exponent of the sequence $a_n \rightarrow \infty$. Then the canonical product $P(z) = \prod E_p(z/a_n)$ has order $\rho(P) = \alpha$.*

Proof. Let $r_n = |a_n|$ and $r = |z|$. Recall that p is the least integer such that $\sum (1/r_n)^{p+1} < \infty$. So we also have $\sum 1/r_n^p = +\infty$. This implies

$p \leq \alpha \leq p+1$. For convenience we will assume $\sum (1/r_n)^\alpha < \infty$. (For the general case, just replace α with $\alpha + \epsilon$.)

As we have seen, the Weierstrass factor

$$E_p(z) = (1-z) \exp \left(z + \frac{z^2}{2} + \cdots + \frac{z^p}{p} \right)$$

satisfies, for ‘small z ’ meaning $|z| < 1/2$, the inequality $|1 - E_p(z)| \leq |z|^{p+1} \leq 1/2$, and hence also the inequality

$$|\log E_p(z)| = O(|z|^{p+1}). \quad (3.1)$$

On the other hand, for ‘large z ’, meaning $|z| \geq 1/2$, we have

$$|\log E_p(z)| = O(|\log(1-z)| + |z|^p). \quad (3.2)$$

(To be precise — since the logarithm is multivalued — these inequalities mean the logarithm can be chosen so the desired bound holds.) The $\log(1-z)$ can term be ignored unless z is very close to 1, in which case $E_p(z)$ is close to zero. So it can also be ignored when bounding $E_p(z)$ from above.

Combining these estimates for $P(z) = \prod E_p(z/a_n)$, we get the upper bound:

$$\log |P(z)| \leq O \left(r^{p+1} \sum_{r_n > 2r} \frac{1}{r_n^{p+1}} + r^p \sum_{r_n \leq 2r} \frac{1}{r_n^p} \right).$$

Now in the second sum, since $p \leq \alpha$, we have

$$\sum_{r_n \leq 2r} \frac{1}{r_n^p} = \sum \frac{r_n^{\alpha-p}}{r_n^\alpha} \leq (2r)^{\alpha-p} \sum \frac{1}{r_n^\alpha} = O(r^{\alpha-p}).$$

Similarly in the first sum, since $\alpha \leq p+1$, we have

$$\sum_{r_n > 2r} \frac{1}{r_n^{p+1}} = \sum_{r_n > 2r} \frac{1}{r_n^{p+1-\alpha}} \frac{1}{r_n^\alpha} \leq (2r)^{\alpha-p-1} \sum \frac{1}{r_n^\alpha} = O(r^{\alpha-p-1}).$$

Altogether this gives

$$\log |P(z)| \leq O(r^\alpha).$$

Thus $\log M(r) = O(r^\alpha)$ and hence $\rho(P) \leq \alpha$. By Jensen’s theorem, the order of $P(z)$ is exactly α . ■

The minimum modulus. To control the result of division by a canonical product, we now estimate its minimum modulus. As an example — for $\sin(z)$, we have $m(r) \asymp 1$ for infinitely many r .

Theorem 3.15 *The minimum modulus of the canonical product $P(z)$ above satisfies $m(r) \geq \exp(-r^{\alpha+\epsilon})$ for arbitrarily large r .*

Proof. We must show $|\log m(r)| = O(r^{\alpha+\epsilon})$. The follows the same lines as the bound $\log M(r) = O(r^\alpha)$ just obtained, since (3.1) and (3.2) give bounds for $|\log E_p(z)|$. The only nuance is that we cannot ignore the logarithmic term in (3.2). We must also decide which values of r to choose, since $m(r) = 0$ whenever $r = |a_n|$.

To this end, we fix $\epsilon > 0$ and exclude from consideration the balls B_n defined by $|z - a_n| < 1/r_n^{\alpha+\epsilon}$. Since the sum of the radii of the excluded balls is finite, there are plenty of large circles $|z| = r$ which avoid $\bigcup B_n$.

To complete the proof, it suffices to show that for z on such circles, we have

$$\sum_{|z-a_n|<r_n} |\log(1 - z/a_n)| = O(r^{\alpha+\epsilon}).$$

Note that the number of terms in the sum above is at most $N(2r) = O(r^\alpha)$. Because we have kept z away from a_n , we have

$$|\log(1 - z/a_n)| = O(\log r).$$

Consequently

$$\sum_{|z-a_n|<r_n} |\log(1 - z/a_n)| = O(N(2r) \log r) = O(r^{\alpha+\epsilon}),$$

as desired. ■

Proof of the Hadamard factorization theorem. Let $f(z)$ be an entire function of order ρ with zeros (a_i) , and let $P(z)$ be the corresponding canonical product. Then P also has order ρ , as we have just seen. Since f and P have the same zeros, the quotient f/P is an entire function with no zeros. The lower bound on $m(r)$ just established implies that f/P also has order ρ , and hence $f/P = \exp Q(z)$ where $\deg Q \leq \rho$. ■

Trigonometric functions. As a first example, we determine the canonical factorization of the sine function:

Theorem 3.16 *We have*

$$\sin(\pi z) = \pi z \prod_{n \neq 0} \left(1 - \frac{z^2}{n^2}\right).$$

Proof. Indeed, the right hand side is a canonical product, and $\sin(\pi z)$ has order one, so the formula is correct up to a factor $\exp Q(z)$ where $Q(z)$ has degree one. But since $\sin(\pi z)$ is odd, we conclude Q has degree zero, and by checking the derivative at $z = 0$ of both sides we get $Q = 0$. ■

Use of the logarithmic derivative. For later application, we record some useful properties of the logarithmic derivative f'/f of an entire function $f(z)$.

1. We have $(fg)' / fg = f'/f + g'/g$.
2. If $f'/f = g'/g$, then $f = Cg$ for some constant $C \neq 0$.
3. We have $f'(az + b)/f(az + b) = a(f'/f)(az + b)$.

Sine, cotangent and zeta. The product formula above gives, under logarithmic differentiation,

$$\frac{(\sin(\pi z))'}{\sin(\pi z)} = \pi \cot(\pi z) = \frac{1}{z} + \sum_1^\infty \frac{1}{z - n} + \frac{1}{z + n}.$$

This formula is useful in its own right: it shows $\pi \cot(\pi z)$ has simple poles at all points of \mathbb{Z} with residue one. This property can be used, for example, to evaluate $\zeta(2k) = \sum_1^\infty 1/n^{2k}$ and other similar sums by the residue calculus — see the examples in Chapter 1.

We note that the product formula for $\sin(z)$ can also be used to prove $\zeta(2) = \pi^2/6$, by looking at the coefficient of z^2 on both sides of the equation. The sine formula also shows $\sum_{a < b} 1/(ab)^2 = \pi^4/5!$, $\sum_{a < b < c} 1/(abc)^2 = \pi^6/7!$, etc., so with some more work it can be used to evaluate $\zeta(2k)$. For example, we have

$$\zeta(4) = \left(\sum \frac{1}{a^2}\right) \left(\sum \frac{1}{b^2}\right) - 2 \left(\sum_{a < b} \frac{1}{(ab)^2}\right) = \frac{\pi^4}{36} - \frac{\pi^4}{60} = \frac{\pi^4}{90}.$$

Translation and duplication formulas. Many of the basic properties of the sine and cosine functions can be derived from the point of view of the uniqueness of an odd entire function with zeros at $\mathbb{Z}\pi$. For example, equations $\sin(z + \pi) = -\sin(z)$ and $\sin(2z) = 2\sin(z)\cos(z)$ hold up to a factor of $\exp(az + b)$ as a consequence of the fact that both sides have the same zero sets.

3.2 The Gamma Function

We now turn to a discussion of the extension of the factorial function $n \mapsto n!$ to the complex numbers.

We begin by studying the canonical product associated to the *negative integers*: namely

$$G(z) = \prod_1^\infty \left(1 + \frac{z}{n}\right) e^{-z/n}.$$

It has half the terms that appear in the factorization of $\sin(\pi z)$; indeed, we have

$$G(z)G(-z) = \frac{\sin(\pi z)}{\pi z}. \quad (3.3)$$

Euler's constant. We have $G(0) = 1$, but what is $G(1)$? To answer this, we define *Euler's constant* by:

$$\gamma = \lim_{n \rightarrow \infty} (1 + 2 + \cdots + 1/n) - \log(n + 1).$$

This expression gives the limiting error obtained when one approximates $\int_1^{n+1} dx/x$ by n unit rectangles lying above the graph of $y = 1/x$. (The sum is finite because these areas can all be slid horizontally to lie disjointly inside a fixed rectangle of base one.)

Then, using the fact that $(1 + 1)(1 + 1/2) \cdots (1 + 1/n) = (n + 1)$, we find $G(1) = \exp(-\gamma)$.

Functional equation. The functions $G(z - 1)$ and $zG(z)$ have the same zeros. How are they related? By Hadamard's theorem, we have $G(z - 1) = z \exp(az + b)G(z)$ for a, b . In fact:

Theorem 3.17 *We have $G(z - 1) = ze^\gamma G(z)$.*

Proof. The logarithmic derivative of $zG(z)$ is given by:

$$\frac{1}{z} + \sum_1^\infty \frac{1}{z + n} - \frac{1}{n},$$

while the logarithmic derivative of $G(z-1)$ is just

$$\sum_1^\infty \frac{1}{z+n-1} - \frac{1}{n} = \frac{1}{z} - 1 + \sum_1^\infty \frac{1}{z+n} - \frac{1}{n+1}.$$

Since $\sum_1^\infty (1/n - 1/(n+1))$ telescopes to 1, these series are equal, and hence $G(z-1)$ and $zG(z)$ agree up to a constant. The value of this constant is determined by setting $z = 1$. ■

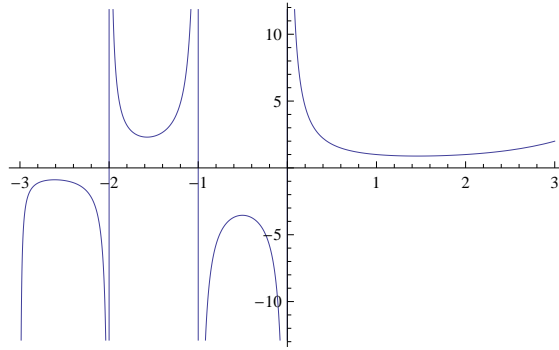


Figure 5. The Γ function.

The Γ function. We now define

$$\Gamma(z) = \frac{1}{ze^{\gamma z}G(z)}.$$

Note especially that we have added a factor of $\exp(\gamma z)$ to compensate for the multiplicative factor of e^γ . This equation is sometimes written more explicitly as

$$z\Gamma(z) = e^{-\gamma z} \prod_{n=1}^{\infty} \left(1 + \frac{z}{n}\right)^{-1} e^{z/n}.$$

From the discussion above we find:

1. $\Gamma(z+1) = z\Gamma(z)$; $\Gamma(1) = 1$; and hence
2. $\Gamma(n+1) = n!$ for $n \geq 0$;
3. $\Gamma(z)$ has no zeros; and
4. $\Gamma(z)$ has simple poles at $0, -1, -2, -3, \dots$

Using the fact that

$$\Gamma(z - n) = \frac{\Gamma(z + 1)}{(z - n) \cdots (z - 1)z},$$

we find

$$\text{Res}_{-n}(\Gamma(z)) = \frac{(-1)^n}{n!}.$$

From (3.3) we obtain *Euler's supplement*

$$\frac{\pi}{\sin(\pi z)} = \Gamma(z)\Gamma(1 - z),$$

which implies, for example:

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

Variants of the sine formula are:

$$\Gamma\left(\frac{1}{2} + z\right)\Gamma\left(\frac{1}{2} - z\right) = \frac{\pi}{\cos \pi z} \quad \text{and} \quad \Gamma(z)\Gamma(-z) = -\frac{\pi}{z \sin(\pi z)}.$$

Using the fact that $\Gamma(\bar{z}) = \overline{\Gamma(z)}$, we find from this last that

$$|\Gamma(iy)|^2 = \frac{\pi}{y \sinh(\pi y)}.$$

We also note that by the functional equation, $\Gamma(z)$ has constant sign on each interval of the form $(n, n + 1)$; in fact the sign is positive for $n > 0$ and $(-1)^n$ for $n < 0$.

Gauss's formula. The formula above for the Γ function would not be easy to discover from scratch. Here is a formula that could also be taken as a simpler definition of the Gamma function.

Theorem 3.18 *We have*

$$\Gamma(z) = \lim_{n \rightarrow \infty} \frac{n!n^z}{z(z+1) \cdots (z+n)}.$$

As motivation, we note that for integers z we have $z! = \lim n^z \binom{z+n}{n}^{-1}$, which is consistent with the formula above if we multiply both sides by z . Indeed, formally we have

$$\binom{z+n}{n} = \frac{(z+1) \cdots (z+n)}{n!} = \frac{(n+1) \cdots (n+z)}{z!} \approx \frac{n^z}{z!},$$

and solving for $z!$ and taking the limit gives the formula above.

Proof. By definition, we have:

$$\frac{1}{\Gamma(z)} = \lim_{n \rightarrow \infty} z e^{\gamma z} \prod_{k=1}^n \left(1 + \frac{z}{k}\right) e^{-z/k}.$$

Now $\gamma - 1 - 1/2 - \cdots - 1/n \approx -\log(n)$, and $(1 + z/k) = (k + z)/k$, so this becomes

$$\frac{1}{\Gamma(z)} = \lim_{n \rightarrow \infty} \frac{z(z+1) \cdots (z+n) n^{-z}}{n!},$$

which gives the formula above. ■

Corollary 3.19 *We have $|\Gamma(z)| \leq |\Gamma(\operatorname{Re} z)|$ for all z .*

Proof. This inequality holds for each term in the product above. ■

This Corollary gives us good estimates on $\Gamma(z)$: it is *uniformly bounded* in any region of the form $0 < a < \operatorname{Re}(z) < b$, and for $-n < \operatorname{Re} z < b$ the function $z(z+1) \cdots (z+n)\Gamma(z)$ is uniformly bounded.

Characterizing $\Gamma(z)$. In fact these properties are enough to characterize the Γ function.

Note that $\exp(2\pi iz)\Gamma(z)$ satisfies the same functional equation as $\Gamma(z)$, and its reciprocal still has order 1; but it violates the growth condition. The boundedness is also needed.

Theorem 3.20 (Wielandt) *If $F(z+1) = zF(z)$ for $\operatorname{Re} z > 0$, $F(1) = 1$ and $F(z)$ is bounded on the strip $\{z : \operatorname{Re} z \in [1, 2]\}$, then $F(z) = \Gamma(z)$.*

Proof. The functional equation allows one to extend $F(z)$ to a meromorphic function on the whole plane, whose poles and their residues agree with those of $\Gamma(z)$. Thus $G(z) = F(z) - \Gamma(z)$ is entire, $G(0) = G(1) = 0$ and $G(z+1) = zG(z)$. Our boundedness assumptions now imply that $G(z) = G(z+1)/z$ is bounded in the strip $S = \{\operatorname{Re} z \in [0, 1]\}$, since it has a removable singularity at $z = 0$. Thus $H(z) = G(z)G(1-z)$ is also bounded in S . The functional equation for G implies $H(z+1) = -H(z)$ (as in the sine formula), and thus H is a constant, which must be zero. ■

A duplication formula. As an application one can prove ‘multiple angle’ formulas for $\Gamma(nz)$. The simplest is:

Corollary 3.21 *We have $2\sqrt{\pi}\Gamma(2z) = 2^{2z}\Gamma(z)\Gamma(z + 1/2)$.*

Proof. Let $F(z) = \Gamma(z/2)\Gamma(z/2 + 1/2)$. Then $F(z + 1) = \Gamma(z/2 + 1/2)\Gamma(z/2 + 1) = zF(z)/2$. So if write instead $F(z) = 2^z\Gamma(z/2)\Gamma(z/2 + 1/2)$, then $F(z + 1) = zF(z)$. Since $F(z)$ is also bounded for $\operatorname{Re} z \in [1, 2]$, we have $F(z) = C\Gamma(z)$ for some C ; and indeed, $C = F(1) = 2\Gamma(1/2) = 2\sqrt{2\pi}$. The result above now follows, upon replacing z with $2z$. ■

The integral representation: the Mellin transform. We now turn to a second motivation for introducing the Γ function.

Theorem 3.22 *For $\operatorname{Re}(z) > 0$, we have*

$$\Gamma(z) = \int_0^\infty e^{-t} t^z \frac{dt}{t}.$$

In other words, $\Gamma(z)$ is the *Mellin transform* of the function e^{-t} on \mathbb{R}^* .

The Mellin transform is an integral against characters $\chi : \mathbb{R}^* \rightarrow \mathbb{C}^*$ (given by $\chi(t) = t^z$), and as such it can be compared to the Fourier transform (for the group \mathbb{R} under addition) and to Gauss sums. Indeed the Gauss sum

$$\sigma(\chi) = \sum_{(n,p)=1} \chi(n) e^{2\pi i n/p}$$

is the analogue of the Gamma function for the group $(\mathbb{Z}/p)^*$.

Proof. Apply the uniqueness theorem above. ■

Stirling’s formula. Here is a very brief introduction to the method of steepest descent, with the aim of explaining Stirling’s formula: as $s \rightarrow \infty$, we have:

$$\Gamma(s) \sim \sqrt{\frac{2\pi}{s}} \left(\frac{s}{e}\right)^s.$$

The idea is to model $\Gamma(s) = \int_0^\infty e^{-t} t^{s-1} dt$ by the more easily understood Gaussian integral:

$$\int_{-\infty}^\infty e^{a-b(t-t_0)^2/2} dt = e^a \sqrt{2\pi/b},$$

which is itself proved using the identity

$$\left(\int_{-\infty}^{\infty} e^{-x^2/2} dx \right)^2 = \int e^{-r^2/2} r dr d\theta = 2\pi.$$

To this end we rewrite $\Gamma(s) = \int_0^\infty e^{\phi(t)} dt$, where $\phi(t) = -t + (s-1) \log t$. Then $\phi'(t) = -1 + (s-1)/t$ vanishes at $t_0 = s-1$, with $\phi''(t_0) = -(s-1)/t_0^2 = -1/(s-1)$. So we have

$$\phi(t) \approx \phi(t_0) + \phi''(t_0)(t-t_0)^2/2 = a - b(t-t_0)^2/2$$

with $a = 1 - s + (s-1) \log(s-1)$ and $b = 1/(s-1)$. This gives

$$\Gamma(s) \sim e^{1-s} (s-1)^{s-1} \sqrt{2\pi(s-1)}.$$

Using the fact that $(s-1)^s \sim s^s/e$ and $\sqrt{s-1} \sim \sqrt{s}$, this gives Stirling's formula.

As a simple check, we note that Stirling's formula implies:

$$\lim (n!/n^n)^{1/n} = 1/e.$$

This can be seen by interpreting $(1/n)(2/n)(3/n) \cdots (n/n)$ as the exponential of an integral, and using the fact that $\int_0^1 \log x dx = -1$.

The name *steepest descent* comes from the fact that the real axis passes through the saddle pass of the absolute value of the integrand at its critical point, and proceeds in the direction where the absolute value decreases fastest. Compare [MH, p. 438].

3.3 Meromorphic functions: The Mittag-Leffler theorem

We now turn to the problem of describing meromorphic functions with *prescribed principal parts*. This means we specify a sequence of *distinct points* $a_n \rightarrow \infty$, and a finite *Laurent tail*

$$p_n(z) = \frac{b_k}{(z-a_n)^k} + \cdots + \frac{b_1}{z-a_n}$$

at each point. (The values of k and b_1, \dots, b_k depend on n .) We say $p_n(z)$ is the principal part of a meromorphic function $f(z)$ if $f(z) - p_n(z)$ is holomorphic at $z = a_n$.

Theorem 3.23 *For any sequence of points $a_n \rightarrow \infty$ and principal parts $p_n(z)$, there exists a meromorphic function $f(z)$ with poles exactly at a_n , and with the prescribed principal parts.*

Proof. Let $q_n(z)$ be the polynomial given by truncating the power series for $p_n(z)$ so that $|p_n(z) - q_n(z)| < 2^{-n}$ when $|z| < |a_n|/2$. Then $f(z) = \sum(p_n(z) - q_n(z))$ is the desired function. ■

Singly-periodic functions. (This is a hint of the theory of elliptic or doubly-periodic functions to come.)

Any meromorphic function satisfying $f(z+1) = f(z)$ has the form $f(z) = F(e^{2\pi iz})$ for some meromorphic function $F : \mathbb{C}^* \rightarrow \mathbb{C}$. This leads to formulas such as:

$$\sum_{-\infty}^{\infty} \frac{1}{(z-n)^2} = \frac{\pi^2}{\sin^2(\pi z)}.$$

Note that the right hand side is a sum of principal parts. Also, compare this to the series for $\pi \cot(\pi z)$.

To verify this equation, first note that $|z-n|^2 \geq |\operatorname{Im} z|^2 + n^2 - O(1)$ for $0 \leq \operatorname{Re} z \leq 1$, and hence the sum on the left converges *uniformly* in this strip so long as $|\operatorname{Im} z| \geq 1$. Thus both sides tends to 0 as $|\operatorname{Im} z| \rightarrow \infty$. But their difference is holomorphic and periodic, and so it must vanish identically.

On the other hand, by evaluating the constant term in the Laurent series on both sides, we obtain the formula $\sum_1^\infty 1/n^2 = \pi^2/6$.

From Mittag-Leffler to Weierstrass. The Weierstrass Theorem — that there exists a function $g(z)$ with prescribed zeros — is a Corollary of the Mittag-Leffler theorem (which historically came later).

Indeed, to construct an entire function with zeros of multiplicities m_n at distinct points $a_n \rightarrow \infty$, we can simply first construct a meromorphic function $f(z)$ with principle part $p_n(z) = m_n/(z - a_n)$ at a_n , and then take $g(z) = \exp(\int f(z) dz)$. Since the residues of $f(z)$ are integers, the integral is only well-defined modulo $2\pi i\mathbb{Z}$, but this is sufficient to make its exponential well-defined.

...this topsy-turvy way of doing things (die Dinge auf dem Kopf zu stellen) should be not sanctioned by anyone who sees mathematics as something other than a disordered heap of mathematical results.

—A. Pringsheim, 1915.

(See [Re, p.131].)

4 Conformal mapping

We now turn to the theory of analytic functions as mappings. Here the dominant operation is composition, rather than addition or multiplication.

4.1 The Riemann mapping theorem

A conformal homeomorphism between plane domains is an isomorphism in the category of Riemann surfaces. The first main result classifies these domains when they are simply-connected.

Theorem 4.1 (Riemann mapping theorem) *Let $U \subset \mathbb{C}$ be a simply connected domain, other than \mathbb{C} itself. Then there exists a conformal homeomorphism $f : U \rightarrow \Delta$.*

The same result holds for any simply-connected region $U \subset \widehat{\mathbb{C}}$ such that $|\widehat{\mathbb{C}} - U| \geq 1$. Note that $\pi_1(U) = (1)$ iff $\widehat{\mathbb{C}} - U$ is connected, so the complement is either empty, one point or a nontrivial continuum.

Since $\text{Aut}(\Delta)$ is large, the Riemann map is not unique. It can be made unique by picking a point $p \in \Delta$ and requiring $f(p) = 0$ and $f'(p) > 0$. The value of $f'(p)$ is an interesting invariant of (U, p) ; its reciprocal is called the *conformal radius* of U with center p .

Examples of Riemann maps.

1. We first note that $\Delta \cong \mathbb{H}$, e.g. by the Möbius transformation $A(z) = -i(z-1)/(z+1)$. We also remark that $A : S^1 \rightarrow \widehat{\mathbb{R}}$ becomes in angular coordinates,

$$A(\theta) = \frac{2}{2i} \frac{e^{i\theta/2} - e^{-i\theta/2}}{e^{i\theta/2} + e^{-i\theta/2}} = \tan(\theta/2).$$

This is the change of variables used in calculus to integrate rational functions of sine and cosine.

2. We also remark that $(\Delta, 0) \cong (\Delta, p)$, and $(\mathbb{H}, i) \cong (\mathbb{H}, p)$, as can be seen using the Möbius transformations $B(z) = (z+p)/(1+\bar{p}z)$ and $C(z) = az+b$ where $p = ai+b$.
3. Of course any region in $\widehat{\mathbb{C}}$ bounded by a circle (or line) is also isomorphic to Δ .
4. Let $L \subset \widehat{\mathbb{C}}$ be a *lune* – the region between two circular arcs. Apply a Möbius transformation so the vertices of L are at 0 and ∞ ; then after a rotation, we can assume $L = \{z : 0 < \arg z < \alpha\}$, and $f(z) = z^{\pi/\alpha}$

send L to \mathbb{H} . (This illustrates the idea that the Riemann map $U \cong \Delta$ is often constructed as a composition of several simple transformations.)

5. The map $f(z) = z + 1/z$ sends the region $\Omega = \mathbb{H} - \overline{\Delta}$ to \mathbb{H} . This is a particular case of a lune.
6. The strip $S = \{z : 0 < \operatorname{Im} z < \pi\}$ is a lune with angle zero; it is sent to \mathbb{H} by $f(z) = e^z$. The region between any two tangent circles is isomorphic to S by a Möbius transformation.

This is related to the problem of extending the harmonic function $u(z) = 1$ for $\operatorname{Im} z > 0$ and 0 for $\operatorname{Im} z < 0$ from S^1 to Δ ; we can take $u = \operatorname{Re} f(z)/\pi$ where $f : \Delta \rightarrow S$ sends ± 1 to the two ends of S .

7. Let us write $S = S_+$ and S_- according to the sign of $\operatorname{Re}(z)$. Note that $f(z) = e^z$ sends the imaginary axis to S^1 . Thus f sends S_- to $\Delta \cap \mathbb{H}$, and sends S_+ to $\mathbb{H} - \overline{\Delta}$. Both of these regions are lunes. The map $g(z) = (z + 1/z)/2$ sends $f(S_-)$ to $-\mathbb{H}$ and $f(S_+)$ to $+\mathbb{H}$.
8. As a variation on the discussion above, we observe that $f(z) = \cos z$ maps the half-strip $T \subset \mathbb{H}$ defined by $0 < \operatorname{Re} z < \pi$ to $-\mathbb{H}$. Indeed, $\cos(z) = (e^{iz} + e^{-iz})/2$ is simply the composition of the map $z \mapsto e^{iz}$, which sends T to $\Delta \cap \mathbb{H}$, with the map $g(z) = (z + 1/z)/2$.

If we reflect T through its sides, we obtain a tiling of \mathbb{C} . The map $\cos(z)$ sends half the tiles to \mathbb{H} and the other half to $-\mathbb{H}$, as can be seen by Schwarz reflection.

9. One last trigonometric example: the map $f(z) = \sin(z)$ sends $T = \{z \in \mathbb{H} : 0 < \operatorname{Re} z < \pi/2\}$ to the first quadrant. As an alternative way to see this, use the fact that $\sin(iy) = \sinh(iy)$ and $\sin(\pi/2 + iy) = \cosh(y)$.

Consequently $\sin^2(z)$ maps T to \mathbb{H} . This is consistent with the previous example, because $\sin^2(z) = (1 - \cos(2z))/2$.

10. We note that Δ^* is isomorphic to $\mathbb{C} - \overline{\Delta}$ by $z \mapsto 1/z$. The Riemann mapping theorem shows there is a conformal map

$$\mathbb{C} - \overline{\Delta} \rightarrow \mathbb{C} - K$$

for any connected set K with $|K| > 1$.

Univalence and compactness. To begin the proof of the Riemann mapping theorem, we recall a few fundamental facts.

Let us say $f : U \rightarrow \mathbb{C}$ is *univalent* if f is injective and analytic. By injectivity, its inverse is analytic, and f provides a homeomorphism between U and $V = f(U)$. (It need *not* provide a homeomorphism between their closures.) We then have:

Lemma 4.2 *Let $f_n : U \rightarrow \mathbb{C}$ be a sequence of univalent maps, converging locally uniformly to $f : U \rightarrow \mathbb{C}$. If f is nonconstant, it too is univalent.*

Proof. Suppose $f(a) = f(b)$ with $a \neq b$, but f is nonconstant. Then $g(z) = f(z) - f(b)$ has a zero at $z = a$ and $g(z) = \lim g_n(z) = f_n(z) - f_n(b)$. It follows by Rouché's theorem that for $n \gg 0$ we have $g_n(a_n) = 0$ with $a_n \rightarrow a$. But then $f_n(a_n) = f_n(b)$, contrary to univalence of f_n . ■

Lemma 4.3 *The space of analytic maps $f : U \rightarrow \overline{\Delta}$ is compact in the topology of locally uniform convergence.*

Proof. This is a consequence of Arzela-Ascoli and the fact that $|f'(z)| \leq 1/d(z, \partial U)$ by Cauchy's integral formula. ■

Lemma 4.4 *If $U \subset \mathbb{C}$ is simply-connected and $0 \notin U$, then there exists a univalent map $f : U \rightarrow \mathbb{C}$ such that $f(z)^2 = z$.*

Proof of the Riemann mapping theorem. Pick $p \in U$ and let \mathcal{F} denote the space of univalent maps $f : (U, p) \rightarrow (\Delta, 0)$. For convenience, we will also impose the condition that $f'(p) > 0$.

We first observe that \mathcal{F} is nonempty. This is clear if U is bounded. It is also true if $U \neq \mathbb{C}$, for in this case we can translate to $0 \notin U$; then the image V of a particular branch of $\sqrt{z} : U \rightarrow \mathbb{C}$ is disjoint from $-V$, so by composing with $1/(z - a)$, $a \in -V$, we obtain a univalent map from U to a bounded region.

Note that if $B(p, r) \subset U$, then by the Schwarz lemma we have $|f'(p)| \leq 1/r$ for all $f \in \mathcal{F}$. Thus $M = \sup_{\mathcal{F}} f'(p)$ is finite. By the preceding lemmas, the bound M is actually realized: there exists an $f \in \mathcal{F}$ such that $f'(p) = M$.

To complete the proof, we need to show $f(U) = \Delta$. If not, there is an $a \in \Delta - f(U)$. Let $A : \Delta \rightarrow \Delta$ be an automorphism such that $A(a) = 0$. Let $s(z) = z^2$, and construct a branch of s^{-1} on the simply-connected region $A(f(U)) \subset \Delta^*$. Let $B : \Delta \rightarrow \Delta$ be an automorphism such that

$$g = B \circ s^{-1} \circ A \circ f : U \rightarrow \Delta$$

satisfies $g(p) = 0$ and $g'(p) > 0$. Then $g \in \mathcal{F}$ as well.

Now notice that $f = (A \circ s \circ B) \circ g = h \circ g$. By construction, $h : \Delta \rightarrow \Delta$ is a proper map of degree two, with $h(0) = 0$. Thus $|h'(0)| < 1$ by the Schwarz lemma. But $g'(0) = h'(0)f'(0)$, so $g'(0) > f'(0) = M$, contrary to the definition of M since $g \in \mathcal{F}$.

Thus f must have been surjective after all, so it provides the desired conformal map between U and Δ . ■

Boundary behavior. It is now convenient to reverse domain and range and consider a Riemann map $f : \Delta \rightarrow U \subset \mathbb{C}$. We will investigate the question of extending f at least to a *continuous* map on ∂U .

Here is one of the main results. Recall that a compact set $J \subset \mathbb{C}$ is a *Jordan curve* if it is homeomorphic to a circle.

Theorem 4.5 *If ∂U is a Jordan curve, then any conformal map $f : \Delta \rightarrow U$ extends to a homeomorphism between $\overline{\Delta}$ and \overline{U} .*

Length–area. Here is the basic argument, exploiting the fact that f stretches all directions by the same factor, $|f'(z)|$.

Let $R(a, b) = [0, a] \times [0, b] \subset \mathbb{C}$. For each $y \in [0, b]$, let $L_x = [0, a] \times \{y\}$. Then we have

Lemma 4.6 *For any conformal map $f : R(a, b) \rightarrow U \subset \mathbb{C}$, there exists a $y \in [0, b]$ such that*

$$b^2 \text{length}(f(L_y))^2 \leq \text{area}(U) \text{area}(R).$$

In other words, the length of the image of some horizontal line is bounded above by $\sqrt{\text{area}(U)a/b}$.

Corollary 4.7 *If $\text{area}(U) = \text{area}(R(a, b))$, then the image of some horizontal line is shorter in U (or at least no longer).*

Proof of Lemma 4.6 The average length of $f(L_y)$, squared, satisfies

$$\begin{aligned} \left(\frac{1}{b} \int_0^b \text{length}(L_y) dy \right)^2 &= \frac{1}{b^2} \left(\int_{R(a,b)} |f'(z)| |dz|^2 \right)^2 \\ &\leq \frac{1}{b^2} \int 1^2 \int |f'|^2 = \frac{\text{area}(U) \text{area}(R(a, b))}{b^2}. \end{aligned}$$

Since the average exceeds the minimum, the result follows. ■

Jordan curves. Here is a basic fact about a Jordan curve $J \subset \mathbb{C}$. Given $a, b \in J$, let $[a, b] \subset J$ be the subarc joining these two points with smallest diameter. Then $\text{diam}[a, b] \rightarrow 0$ if $|a - b| \rightarrow 0$.

Proof of Theorem 4.5. Given a point $z \in \partial\Delta$, map Δ to an infinite strip, sending z to one end. Then there is a sequence of disjoint squares in the strip tending towards that end. The images of these squares have areas tending to zero, so there are cross-cuts $\gamma_n \subset \overline{\Delta}$ enclosing z such that $\text{length}(f(\gamma_n)) \rightarrow 0$. Thus the endpoints of $f(\gamma_n)$ converge to points $a_n, b_n \in J$, with $z \in [a_n, b_n]$ for $n \gg 0$. Since $\text{diam}[a_n, b_n] \rightarrow 0$, the disk bounded by $f(\gamma_n) \cup [a_n, b_n]$ shrinks to z , and this implies $f(z_n) \rightarrow f(z)$ whenever $z_n \rightarrow z$ in Δ .

Consequently f extends to a continuous map $S^1 \rightarrow J$. Since $f|_{\Delta}$ is a homeomorphism, $f|_{S^1}$ is monotone; thus it is injective unless it is constant on an arc. But it cannot be constant on an arc, since f is nonconstant. ■

Local connectivity. The key point in the proof above is the following property of a Jordan domain \overline{U} : if $a, b \in \partial U$ are joined by a short arc $\alpha \subset U$, then the disk cut off by α has small diameter. This principle is called ‘short dam, small lake’.

Recall that a compact set K is *locally connected* if for every open set $U \subset K$ and $x \in U$ there is a connected open set with $x \in V \subset U$.

Exercise. If there exists a continuous surjective map $f : S^1 \rightarrow K$, then K is locally connected.

The argument in the proof of Theorem 4.5 furnishes short cross-cuts for any Riemann map, so it also shows:

Theorem 4.8 *The Riemann map extends continuously to S^1 iff ∂U is locally connected.*

Finally, using Schwarz reflection one has the useful statement:

Theorem 4.9 *The Riemann map extends analytically across an arc $\alpha \subset S^1$ iff $f(\alpha)$ is a real-analytic arc in ∂U .*

Remark. The length–area method also easily shows:

Theorem 4.10 *Any Riemann mapping $f : \Delta \rightarrow \mathbb{C}$ has radial limits almost everywhere; that is, $\lim_{r \rightarrow 1} f(re^{i\theta})$ exists for almost every θ .*

Proof. We may assume the image of f is bounded; then $\int |f'|^2 < \infty$, which implies by Cauchy-Schwarz that $\int |f'| < \infty$ along almost every ray. ■

More generally, it is known that any bounded analytic function has radial limits a.e. By [Ko], not much more can be said — given a $G_{\delta\sigma}$ set $A \subset S^1$ of measure zero, there exists a bounded function that fails to have radial limits exactly on the set A . For *Riemann mappings*, radial limits exist except outside a very small set — the exceptional set A has capacity zero and in particular $\text{H. dim}(A) = 0$.

Harmonic functions and boundary values. Recalling that if u is harmonic and f is analytic then $u \circ f$ is also harmonic, it is now a simple matter to solve the Dirichlet problem (at least implicitly) on any Jordan domain.

Theorem 4.11 *For any Jordan domain $U \subset \mathbb{C}$, there exists a unique map $P : C(\partial U) \rightarrow C(\overline{U})$ such that $Pu|_{\partial U} = u$ and Pu is harmonic in U .*

Proof. Let $f : \overline{\Delta} \rightarrow \overline{U}$ be the Riemann map, and let $Pu = P_0(u \circ f) \circ f^{-1}$, where P_0 is given by the Poisson kernel on the unit disk. ■

Annuli. A domain $U \subset \widehat{\mathbb{C}}$ is an *annulus* if $\widehat{\mathbb{C}} - U = K_1 \sqcup K_2$ has exactly two connected components. Equivalently, $\pi_1(U) \cong \mathbb{Z}$. Note: in general $H_0(S^2 - U) \cong H_1(U)$; this is an example of *Alexander duality*.

The *standard annulus* is $A(R) = \{z : 1 < |z| < R\}$. Other examples of annuli are \mathbb{C}^* and Δ^* . Exercise: none of these annuli are isomorphic.

Theorem 4.12 *The universal cover of any annulus $U \subset \widehat{\mathbb{C}}$ is isomorphic to \mathbb{C} or \mathbb{H} .*

Proof. We may suppose the two complementary components of U contain 0 and ∞ respectively. Then the map $U \hookrightarrow \mathbb{C}^*$ induces an isomorphism $\pi_1(U) \cong \pi_1(\mathbb{C}^*) \cong \mathbb{Z}$. The universal cover of \mathbb{C}^* is given by $\pi : \mathbb{C} \rightarrow \mathbb{C}^*$, $\pi(z) = e^z$, so

$$V = \log U = \pi^{-1}(U) \subset \mathbb{C}$$

gives the universal cover of U . By the Riemann mapping theorem, V itself is isomorphic to \mathbb{H} or \mathbb{C} . ■

Remark. Clearly $U \cong \Delta^*$ or \mathbb{C}^* iff one or both of its complementary components are singletons. This can be proved e.g. by applying the removable singularities theorem to $f : \Delta^* \rightarrow U$.

Theorem 4.13 *Any annulus U is isomorphic to \mathbb{C}^* , Δ^* , or $A(R)$ for a unique $R > 1$.*

Proof. Let $g : V \rightarrow V$ denote a generator for $\pi_1(U) \cong \mathbb{Z}$ acting on its universal cover. If $V \cong \mathbb{C}$ then g is conjugate to $g(z) = z + 1$ and we have $U \cong \mathbb{C}/\langle g \rangle \cong \mathbb{C}^*$ by the map $\pi(z) = \exp(2\pi iz)$. The same reasoning shows $U \cong \Delta^*$ if $V \cong \mathbb{H}$ and g is parabolic.

Otherwise $V \cong \mathbb{H}$ and we can assume $g(z) = \lambda z$, $\lambda > 1$. This means the core curve of U is a geodesic of length $L = \log \lambda$. Then $\pi(z) = z^\alpha$ maps V to $A(R)$ if we choose α correctly. We want π to map $[1, \lambda]$ onto the unit circle, so we want $\pi(\lambda) = \lambda^\alpha = \exp(\alpha L) = 1$; so we take $\alpha = -2\pi i/L$. Note that this is a purely imaginary number, so π sends \mathbb{R}_+ to the unit circle and \mathbb{R}_- to a circle of radius $R = (-1)^\alpha = \exp(\alpha \pi i) = \exp(2\pi^2/L)$. (We put a minus sign in α so that $R > 1$.) ■

Modulus of an annulus. Two natural invariants of an annulus are the number R such that $U \cong A(R)$, and the hyperbolic length L of its core geodesic. These are related by $\log R = 2\pi^2/L$ as we have just seen. For a direct proof, consider the metric $|dz|/|z|$ on $A(R)$. This makes it into a right cylinder with height over circumference given by $h/c = \log R/2\pi$. On the other hand, the same metric makes $\mathbb{H}/(z \mapsto e^L z)$ into a right cylinder with $h/c = \pi/L$. Equating these two expressions gives the desired relation.

The quantity h/c is often called the *modulus* of A , written $\text{mod}(A)$.

The space of all Riemann mappings. It is traditional to denote the set of univalent mappings $f : \Delta \rightarrow \mathbb{C}$ such that $f(0) = 0$ and $f'(0) = 1$ by S , for *schlicht* (plain, simple). So

$$f(z) = z + a_2 z^2 + a_3 z^3 \dots$$

We give S the topology of uniform convergence on compact sets. We will soon show:

Theorem 4.14 *The space S is compact.*

As a consequence, the coefficients a_n are bounded. The deeper Bieberbach Conjecture, now a theorem, asserts that $|a_n| \leq n$. This is also more than sufficient to show that S is compact.

Remark: no nesting. There are no nesting relations among the images of $f \in S$; that is, the Schwarz lemma implies that if $f(\Delta) = g(\Delta)$ then $f = g$. So if $f(\Delta)$ contains some points outside the unit disk, it must also omit some points inside the unit disk.

Maps to the outside. It is somewhat easier to start with the space Σ of all univalent maps $F : \widehat{\mathbb{C}} - \overline{\Delta} \rightarrow \widehat{\mathbb{C}}$ with $F(\infty) = \infty$, normalized so that

their Laurent series has the form:

$$F(z) = z + \sum_{n=1}^{\infty} \frac{b_n}{z^n}.$$

Since F is an open mapping near ∞ , the complement of its image is a compact set $K(F) \subset \mathbb{C}$.

Example. If we set $b_2 = 1$ we get $F(z) = z + 1/z$ which satisfies $K(F) = [-2, 2]$.

Theorem 4.15 *The area of $K(F)$ is given by $A = \pi(1 - \sum n|b_n|^2)$.*

Proof. Integrate $\overline{F}dF$ over the unit circle and observe that, since $|dz|^2 = (-i/2)d\overline{z}dz$, the area A of $K(F)$ is given by:

$$A = -\frac{i}{2} \int_{S^1} \overline{F}dF = -\frac{i}{2} \left(1 - \sum n|b_n|^2\right) \int_{S^1} \frac{dz}{z} = \pi \left(1 - \sum n|b_n|^2\right).$$

■

Corollary 4.16 *We have $\sum n|b_n|^2 \leq 1$. In particular, we have $|b_n| \leq 1/\sqrt{n}$.*

Corollary 4.17 *The space Σ is compact.*

Proof. This follows from the fact that $|b_n| \leq n^{-1/2}$. So if we fix $R > 1$, then for $|z| \geq R$ we have

$$|F(z) - z| \leq \sum n|b_n|/R^{n+1} = C(R) < \infty.$$

This uniform bound implies every sequence has a subsequence converging uniformly on compact sets. As usual, univalence is also preserved in the limit. ■

Remark. Little is known about optimal bounds for $|b_n|$ over Σ . It is conjectured that $|b_n| = O(1/n^{3/4})$; see [CJ].

We can now use the statement $|b_1| \leq 1$ to prove the first case of the Bieberbach conjecture. Note that the case of equality, we must have all other $b_n = 0$, and hence up to a rotation, $F(z) = z + 1/z$ and $K(F) = [-2, 2]$.

Theorem 4.18 *For all $f \in S$ with have $|a_2| \leq 2$.*

Proof. We first note that $F(z) = 1/f(1/z) + a_2$ is in Σ . Indeed, we find:

$$\begin{aligned} 1/f(1/z) &= \frac{z}{1 + a_2/z + a_3/z^2 + \cdots} \\ &= z \left(1 - (a_2/z + a_3/z^2 + \cdots) + (a_2/z + a_3/z^2 + \cdots)^2 - \cdots \right) \\ &= z - a_2 + \frac{a_2^2 - a_3}{z} + \cdots \end{aligned}$$

and thus $F(z)$ has $b_1 = a_2^2 - a_3$. So we find $|a_2^2 - a_3| \leq 1$, which is sort of interesting, but not what we are aiming for yet.

Next we use a nice trick to make $f(\Delta)$ more symmetrical: we consider, instead of $f(z)$, the new map $g(z) = \sqrt{f(z^2)}$. This map is also in S and it is given by

$$g(z) = z\sqrt{1 + a_2z^2 + a_3z^4 + \cdots} = z + (a_2/2)z^3 + \cdots$$

Thus $|a_2/2| \leq 1$ and hence $|a_2| \leq 2$. ■

Corollary 4.19 *S is compact.*

Proof. Suppose $f_i \in S$ and let $F_i(z) = 1/f_i(1/z) + a_2(i) \in \Sigma$. Pass to a subsequence so $F_i(z)$ converges in Σ , and so $a_2(i)$ converges. Then $G_i(z) = F_i(z) - a_2(i)$ converges outside the disk, and hence $f_i(z) = 1/G_i(1/z)$ converges on the unit disk. ■

Here is a useful geometric consequence of the bound on a_2 .

Theorem 4.20 (Koebe 1/4 theorem) *For any $f \in S$ we have $B(0, 1/4) \subset f(\Delta)$.*

Proof. Suppose $p \notin f(\Delta)$. Note that $A(z) = z/(1 - z/p)$ has $A(0) = 0$ and $A'(0) = 1$. Thus

$$g(z) = \frac{f(z)}{1 - f(z)/p} = (z + a_2z^2 + \cdots)(1 + z/p + \cdots) = z + (a_2 + 1/p)z^2 + \cdots$$

belongs to S . Thus $2 \geq |a_2 + 1/p| \geq |1/p| - 2$ and hence $|p| \geq 1/4$. ■

From the Koebe theorem we get an important comparison between the hyperbolic metric ρ and the ‘1/d’ metric $\delta = \delta(z)|dz| = |dz|/d(z, \partial U)$.

Corollary 4.21 *The hyperbolic metric $\rho(z)|dz|$ on a simply-connected region $U \subset \mathbb{C}$ is comparable to the 1/d metric: we have $\rho(z)/\delta(z) \in [1/2, 2]$ for all $z \in U$. Equivalent, $\rho(z)d(z, \partial U) \in [1/2, 2]$.*

Proof. We will check this at a given $p \in U$. Let $f : (\Delta, 0) \rightarrow (U, p)$ be a Riemann mapping. We may suppose $p = 0$ and $f'(0) = 1$; then $f \in S$, and $\rho_U(p) = \rho_\Delta(p) = 2$. By the Schwarz lemma we have $d(p, \partial U) \leq 1$ and by Koebe we have $d(p, \partial U) \geq 1/4$; hence the result. ■

Comparison of S and Σ . The most important mapping in Σ is $F(z) = z + 1/z$; the most important one in S is

$$f(z) = \sum n z^n = z \frac{d}{dz} \frac{1}{1-z} = \frac{z}{(1-z)^2}.$$

These maps are equivalent: $F(z) = 1/f(1/z) + 2$. We have $K(F) = [-2, 2]$ and $K(f) = (-\infty, -1/4]$. The map $f(z)$ shows the Bieberbach conjecture is sharp. We also note the problem with trying to find a ‘Bieberbach conjecture’ for Σ : there is no map which simultaneously maximizes all the b_n ’s. Indeed, by the area theorem, if $b_1 = 1$ then the rest of the b_n ’s are zero.

The distortion theorems. Given $f \in S$, think of $f(\Delta)$ as a splattered egg; then one finds that no matter what, the yolk $f(B(0, r))$, $r < 1$, is still good (not too distorted). For example, the curvature of $f(S^1(r))$ is bounded by a constant K_r independent of f . Also, $f(S^1(r))$ is convex if r is small enough.

Qualitative theorems of this type can be easily deduced from compactness of S . The Koebe distortion theorems make these results more precise. They state:

Theorem 4.22 *For all $f \in S$ and $z \in \Delta$ with $|z| = r$, we have*

$$\frac{(1-r)}{(1+r)^3} \leq |f'(z)| \leq \frac{(1+r)}{(1-r)^3}$$

and

$$\frac{r}{(1+r)^2} \leq |f(z)| \leq \frac{r}{(1-r)^2}. \quad (4.1)$$

The proof can be made rather conceptual. Let $U \subset \mathbb{C}$ be a proper simply-connected region, and let $f : U \rightarrow \mathbb{C}$ be a univalent map. We will use the hyperbolic metric $\rho(z) |dz|$ on U and the Euclidean metric $|dz|$ on \mathbb{C} . It is then natural to study how these metrics compare under f . To this end we define

$$\delta(z) = \log(|f'(z)|/\rho(z))$$

Note that if we replace $f(z)$ with $af(z) + b$, it only changes $\delta(z)$ to $\delta(z) + \log |a|$.

Lemma 4.23 *The gradient of $\delta(z)$ in the hyperbolic metric on U satisfies $|d\delta|/\rho \leq 2$.*

Proof. We can assume $U = \Delta$, $z = 0$ and $f \in S$ by the Riemann mapping theorem and by the observations above. Then $\rho(z) = 2/(1 - |z|^2)$ is stationary at $z = 0$. Thus $|d\delta| = |f''(0)|/|f'(0)| = |2a_2| \leq 4$. Since $\rho(0) = 2$ we get the bound above. ■

Proof of Theorem 4.22. We continue to assume $U = \Delta$ and $f \in S$. Let $D(z) = \exp \delta(z) = |f'(z)|/\rho(z)$. Integrating along the hyperbolic geodesic from 0 to z , we find

$$|\delta(z) - \delta(0)| \leq \int_0^z |d\delta| \leq \int_0^z 2\rho = 2d(0, r),$$

and thus $D(z)/D(0) \leq \exp(2d(0, r))$, $r = |z|$. But $\exp(d(0, r)) = (1 + r)/(1 - r)$, as can be seen by using $i(1 - z)/(1 + z)$ to map to \mathbb{H} . Thus $D(z)/D(0) \leq (1 + r)^2/(1 - r)^2$. It follows that for $f \in S$, since $f'(0) = 1$, we have:

$$|f'(z)| = \frac{|f'(z)|}{|f'(0)|} \leq \frac{(1 + r)^2}{(1 - r)^2} \cdot \frac{\rho(r)}{\rho(0)} = \frac{(1 + r)}{(1 - r)^3}.$$

The reverse inequality is similar: we have

$$|f'(z)| \geq \frac{(1 - r)^2}{(1 + r)^2} \cdot \frac{\rho(r)}{\rho(0)} = \frac{(1 - r)}{(1 + r)^3}.$$

Integrating these bounds gives (4.1). ■

These results also show that S is compact.

Multiply-connected regions. We briefly mention one of the standard forms for a region that has ‘more than two holes’.

Suppose $U = \mathbb{C} - (K_1 \cup \cdots \cup K_n)$, where the K_i are disjoint compact connected sets, none of which is a single point. We then have:

Theorem 4.24 *There exists a unique conformal map $F : U \rightarrow \mathbb{C}$ of the form $F(z) = z + b_1/z + \cdots$ such that $K(F)$ consists of n disjoint horizontal segments.*

Remark: smoothing the boundary. As a preliminary remark, we note that by applying the Riemann mapping to $\widehat{\mathbb{C}} - K_i$ for each i , one can arrange (if desired) that each K_i is a Jordan domain with real-analytic boundary.

An extremal problems for slits. For the proof of Theorem 4.24, it is useful to introduce the family \mathcal{F} of all univalent conformal maps $f : U \rightarrow \mathbb{C}$ of the form above. We then have the following complement:

The map F maximizes $\operatorname{Re} b_1(f)$ over all $f \in \mathcal{F}$.

The proof depends on the following observations. (Cf. [Gol, V.2].)

1. The theorem is true in the case $n = 1$. Indeed, in this case we can assume $U = \mathbb{C} - \overline{\Delta}$, and then $\mathcal{F} = \Sigma$, and $F(z) = z + 1/z$. This is the unique map maximizing $\operatorname{Re} b_1$, by the area theorem.
2. For any pair of maps defined near ∞ by $z + b_1/z + \cdots$, we have $b_1(f \circ g) = b_1(f) + b_1(g)$.
3. Thus the complement is also true in the case $n = 1$. For in this case we can assume (after translation) that U is the image of $g \in \Sigma$. Then $F = f \circ g^{-1}$, and $\operatorname{Re} b_1(F) = \operatorname{Re} b_1(f) - b_1(g) = 1 - \operatorname{Re} b_1(g)$.

But even more is true! Unless U is already a slit domain, $\operatorname{Re} b_1(F) > 0$, by the area theorem applied to g — since $|b_1(g)| \leq 1$.

4. Now return to the case of general U , and suppose $F \in \mathcal{F}$ maximizes $\operatorname{Re} b_1$. We have $K(F) = L_1 \cup \cdots \cup L_n$. Suppose one of these components — say L_1 — is not a horizontal slit. Then we can find a map $G : \mathbb{C} - L_1 \rightarrow \mathbb{C} - [a, b]$ such that $\operatorname{Re} b_1(G) > 0$ by what we have just observed. But then $G \circ F \in \mathcal{F}$ and $b_1(G \circ F) > b_1(F)$, contrary to the extremal property of F .

5. We should also check uniqueness. For this it is useful to think of U as the interior of a compact, smoothly bounded domain in \mathbb{C} , and $f_1, f_2 : U \rightarrow \widehat{\mathbb{C}}$ as normalized maps of the form $f_i(z) = 1/(z-p) + g_i(z)$ whose images are horizontal slit domains. Then the bounded analytic function $h(z) = f_1(z) - f_2(z)$ on U also sends each component of ∂U into a horizontal line. Thus $h(\partial U)$ has winding number zero about points not on these lines, so h must be constant.

Number of moduli. The number of n -slit regions in \mathbb{C} , i.e. those of the form $S = \mathbb{C} - \bigcup_1^n [z_i, z_i + a_i]$ with $a_i \in \mathbb{R}$, has real dimension $3n$. We can normalize by an affine transformation so the first slit is, say, $[-2, 2]$; so up to isomorphism, the number of such slit regions is $3n - 3$.

Now to take an n -connected region U and produce a slit region, we need to choose a point $p \in U$ to send to infinity and we need to choose a ‘horizontal’ direction at p . That gives 3 more real parameters. But of course these choices are only relevant up to the action of the automorphism group of U , which has dimension 3, 1 and 0 for $n = 1, 2$ and $n \geq 3$. Altogether we find:

$$\begin{aligned} d_n &= \dim\{\text{moduli space of } n\text{-connected regions}\} \\ &= 3n - 6 + \dim \text{Aut}(U) \end{aligned}$$

and hence $d_1 = 0$, $d_2 = 1$ (the modulus of an annulus), $d_3 = 3$, $d_4 = 6$, etc.

Rigidity. We remark that although an n -connected region U can have automorphisms, it cannot have a positive-dimensional set thereof, when $n > 2$. To see this, we can assume U is a plane region bounded by real-analytic Jordan curves. Then $\dim \text{Aut}(U) > 0$ would imply there is a holomorphic vector field $v = v(z)(d/dz)$ on U tangent to ∂U . If $v \neq 0$, then the Euler characteristic of U can be expressed as sum of positive contributions, one for each zero of v on ∂U or in U . But $\chi(U) < 0$, so $v = 0$.

A more geometric proof can be given once one knows that $U \cong \Delta/\Gamma$ (a proof will be sketched below). Then any automorphism must be an isometry, and hence send geodesics to geodesics. By looking at intersections of geodesics one can easily conclude that f is the identity.

4.2 Conformal mappings of polygons

To get explicit forms for Riemann mappings, it is useful to initially try to determine, not f itself, but the deviation of f from a simpler class of mappings. This deviation will be a form in general, so we first make some remarks on forms.

Let $f : X \rightarrow \widehat{\mathbb{C}}$ be a meromorphic function on a Riemann surface. Then $\omega = df = f'(z) dz$ is naturally a 1-form. That is, if we change coordinates by setting $z = \phi(w)$, then in these new coordinates we have

$$\phi^*(df) = d(f(\phi(w))) = f'(\phi(w))\phi'(w) dw.$$

We can similarly define quadratic differentials $q(z) dz^2$ which transform by $\phi^*q = q(\phi(z))\phi'(z)^2 dz^2$.

In more intrinsic terms, these forms are sections of the complex cotangent bundle T^*X and its tensor product with itself.

For a meromorphic 1-form, the *residue* $\text{Res}_p(\omega)$ is also coordinate independent, since it can be expressed as $(2\pi i)^{-1} \int_\gamma \omega$ for a small loop around p . Stokes' theorem implies:

Theorem 4.25 (The Residue Theorem) *If ω is a nonzero meromorphic 1-form on a compact Riemann surface X , then $\sum_{p \in X} \text{Res}_p(\omega) = 0$.*

Let's check this in an example: if we take $\omega(z) = dz/P(z)$ where $P(z)$ is a polynomial of degree $d \geq 2$, then it says:

$$\sum_{P(z)=0} \frac{1}{P'(z)} = 0.$$

This can be verified using partial fractions: we have $P(z) = \sum a_i/(z - b_i)$ and $\sum a_i = 0$ because $|P(z)| = O(|z|^{-2})$ for large z . It can also be proved by integrating $dz/P(z)$ around a large circle and taking the limit.

If $\omega = \omega(z) dz$ is a meromorphic 1-form on the sphere, we can set $\pi(z) = 1/z$ and form $\pi^*\omega$ to find:

$$\text{Res}_\infty(\omega) = \text{Res}_0(-\omega(1/z) dz/z^2).$$

In particular $\text{Res}_\infty(dz/z) = -1$.

Theorem 4.26 *Every meromorphic 1-form on the sphere has 2 more poles than zeros.*

Proof. This is true for $\omega = dz/z$, which has a simple poles at 0 and ∞ and no zeros. It then follows for any other 1-form η , since $f = \eta/\omega$ is a meromorphic function, which must have the same number of poles as zeros. ■

Corollary 4.27 *A holomorphic 1-form on $\widehat{\mathbb{C}}$ must be zero. In particular, if ω_1 and ω_2 are 1-forms with the same principal parts, then $\omega_1 = \omega_2$.*

Remark: one forms and vector fields in higher genus. The first assertion can also be seen by integrating ω to obtain a global holomorphic function on $\widehat{\mathbb{C}}$. On a Riemann surface of genus g , a meromorphic 1-form has $2g - 2$ more zeros than poles, and $\dim \Omega(X) = g$. The space of holomorphic vector fields, on the other hand, satisfies $\dim \Theta(\widehat{\mathbb{C}}) = 3$, $\dim \Theta(E) = 1$ on a torus, and $\dim \Theta(X) = 0$ in higher genus. This is because a meromorphic vector field has $2g - 2$ more poles than zeros.

One can use the vanishing of $\Theta(X)$ to see an n -connected plane region U has a zero-dimensional automorphism group for $n \neq 3$; if not, there would be a boundary-parallel holomorphic vector field on U , but then the double X of U would satisfy $\dim \Theta(X) > 0$; while its genus is given by $g = n - 1$.

Measuring distortion: the 3 great cocycles. Now let $C(f)$ be a differential operator that sends meromorphic functions to meromorphic forms. We say C is a *cocycle* if it satisfies:

$$C(f \circ g) = C(g) + g^*C(f).$$

This formula implies the maps satisfying $C(f) = 0$ form a group.

There are 3 important cocycles in complex analysis:

1. The *derivative* $Df(z) = \log f'(z)$.
2. The *nonlinearity* $Nf(z) dz = d Df = (f''/f')(z) dz$.
3. The *Schwarzian derivative*

$$Sf(z) dz^2 = (Nf)' - \frac{1}{2}(Nf)^2 = \left[\left(\frac{f''}{f'} \right)' - \frac{1}{2} \left(\frac{f''}{f'} \right)^2 \right] dz^2.$$

Their values are functions, 1-forms and 2-forms respectively. The groups they annihilate are transformations of the form $f(z) = z + a$, $f(z) = az + b$ and $f(z) = (az + b)/(cz + d)$.

Let us check this for Sf : if $f(z) = (az + b)/(cz + d)$ and $ad - bc = 1$, then $f'(z) = 1/(cz + d)^2$, hence $Nf(z) = -2c/(cz + d)$, which satisfies

$$(Nf)' = \frac{2c^2}{(cz + d)^2} = \frac{1}{2}(Nf)^2.$$

Higher cocycles? There are in fact operators $T_d f$ generalizing Sf with the property that $T_d f = 0$ iff f is a rational map of degree $d > 1$. But these

cannot be cocycles, because the rational maps of degree $d > 1$ do not form a group.

To get some insight into Sf (and T_d), note that if $f(z) = \sum a_i z^i = (az + b)/(cz + d)$, then $(cz + d) \sum a_i z^i = az + b$ and hence there must be a linear relation between most of the coefficients of $zf(z)$ and $f(z)$; in particular, we must have

$$\det \begin{pmatrix} a_1 & a_2 \\ a_2 & a_3 \end{pmatrix} = a_2^2 - a_1 a_3 = 0,$$

which just says that $f'f''' - (3/2)(f'')^2$ vanishes. This expression is also the denominator of Sf . Similarly a rational map of degree two is characterized by the property that

$$\det \begin{pmatrix} a_1 & a_2 & a_3 \\ a_2 & a_3 & a_4 \\ a_3 & a_4 & a_5 \end{pmatrix} = 0.$$

It is natural to divide this expression by a_1^3 just as we have divided Sf by $a_1^2 = (f')^2$, so we have $T_d(af + b) = T_d(f)$.

The Schwarz-Christoffel formula. We can now use the nonlinearity to derive a formula for the Riemann mapping to a polygonal region. It is based on the fact that

$$N(z^\alpha) = (\alpha - 1) \frac{dz}{z}.$$

Theorem 4.28 *Let $f : \mathbb{H} \rightarrow U$ be the Riemann mapping to a polygon with vertices p_i , $i = 1, \dots, n$ and exterior angles $\pi\mu_i$. Then*

$$f(z) = \alpha \int \frac{d\zeta}{\prod_1^n (\zeta - q_i)^{\mu_i}} d\zeta + \beta,$$

where $f(q_i) = p_i$.

Proof. We will show that:

$$Nf = \sum \frac{-\mu_i dz}{(z - q_i)}.$$

We first observe that Nf extends to a holomorphic 1-form on $\mathbb{C} - \{p_1, \dots, p_n\}$. This follows by Schwarz reflection across each complementary interval on the real axis. These reflected maps g_i do not agree, but they differ by linear

maps: $g_j = A_{ij} \circ g_i$, and thus $N(g_i) = N(g_j)$. It remains to show Nf has simple poles with the given residues at the p_i , and at infinity.

Let $\alpha_i = 1 - \mu_i$. The idea of the proof is that $f(z)$ behaves like $(z - p_i)^{\alpha_i}$ near p_i , and thus Nf has the same residue, which is $\alpha_i - 1 = -\mu_i$. To check this, we note that $g(z) = (f(z) - p_i)^{1/\alpha_i}$ extends by Schwarz reflection across q_i , to give a conformal map near q_i . Thus we can locally write

$$g(z)^{\alpha_i} = f(z) - p_i.$$

Taking the nonlinearity of both sides, and using the fact that the residue is preserved under pullback, we find $\text{Res}_{q_i}(Nf) = 1 - \alpha_i = -\mu_i$.

Near infinity, $f(z)$ behaves like $1/z$ which has nonlinearity $-2dz/z$ and hence residue 2 at infinity. This is consistent with the residue theorem, because $\sum \pi\mu_i = 2\pi$. Since a 1-form is determined by its singularities, we have justified the formula for Nf .

Integrating, we find

$$\log f' = \sum -\mu_i \log(z - q_i) + C;$$

by exponentiating and integrating again, we get the desired formula for f . ■

Two Examples. This formula explains the geometric connection between Riemann maps to polygons and transcendental functions we have seen in two examples, namely:

$$\log(z) = \int \frac{dz}{z}$$

maps \mathbb{H} to a bigon with external angles of π , namely the strips $0 < \text{Im } z < \pi$; and

$$\sin^{-1}(z) = \int \frac{dz}{\sqrt{1 - z^2}}$$

maps \mathbb{H} to a triangle with external angles $\pi/2$, $\pi/2$ and π , namely the half-strip defined by $\text{Im } z > 0$, $|\text{Re}(z)| < \pi/2$.

Unit disk version. The key point in the proof above was that the various results of extending f by Schwarz reflection through $\partial\mathbb{H}$ agree up to composition with linear maps. The same is true if \mathbb{H} is replaced by the unit disk. Thus if we choose the domain of f to be Δ , with $q_i \in S^1$, we find exactly the same formula for f .

Regular polygons. As a particular example, we find:

Theorem 4.29 *The map $f(z) = \int (1 - z^n)^{-2/n} dz$ maps the unit disk to a regular n -gon, sending 0 to its center and the n th roots of unity to its vertices.*

Note that $f'(0) = 1$, and the distance from the center of the polygon to one of its vertices is given by

$$R_n = \int_0^1 (1 - z^n)^{-2/n} dz.$$

The Beta function and conformal radius. The quantity R_n can be computed explicitly. To this end, we recall the Euler Beta-function

$$B(\alpha, \beta) = \int_0^1 u^\alpha (1 - u)^\beta \frac{du}{u(1 - u)} = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}.$$

Note that it is elementary to prove the integral above satisfies $B(\alpha, 1) = 1/\alpha$, and integration by parts gives $B(\alpha, \beta) = ((\beta + 1)/\alpha)B(\alpha + 1, \beta - 1)$; this easily implies

$$B(\alpha + 1, \beta + 1) = \frac{\alpha!\beta!}{(\alpha + \beta + 1)!}$$

for integral values of α, β . We also note that $|dx|/x$, which appears in the integral definition of the Γ function, is the hyperbolic metric on \mathbb{R}_+ , and $|dx|/(x(1 - x))$ is the hyperbolic metric on $(0, 1)$. Note that formally these one forms have residue 1 at the endpoints of the interval, just as $2|dx|/(1 - |x|^2)$ does on $(-1, 1)$.

Making the substitution $u = 1 - z^n$, $du = -nz^{n-1}dz$, $dz = -(1/n)u^{1/n-1} du$, we find

$$R_n = \frac{1}{n} \int_0^1 u^{-2/n} (1 - u)^{1/n-1} du = \frac{\Gamma(1 - 2/n)\Gamma(1/n)}{n\Gamma(1 - 1/n)}.$$

Note that $R_n \rightarrow 1$ as $n \rightarrow \infty$, since $\Gamma(1/n) \sim n$, and $R_n < 1$ by the Schwarz lemma.

Now if we let S_n be the side length of the polygon, then $S_n = 2 \sin(\pi/n) R_n$. But $\Gamma(1/n)\Gamma(1 - 1/n) = \pi/\sin(\pi/n)$, so we get

$$S_n = \frac{2\pi\Gamma(1 - 2/n)}{n\Gamma(1 - 1/n)^2}.$$

Note that $nS_n \rightarrow 2\pi$ as $n \rightarrow \infty$, since $\Gamma(1) = 1$. The factor P_n by which the perimeter exceeds the perimeter of the unit circle is given by $(P_3, P_4, P_5, P_6) \approx (1.461, 1.180, 1.098, 1.043)$.

In particular, the conformal radius of the unit square (sides of length 2) is $2/S_4 = 4\Gamma[3/4]^2/\pi^{3/2} = 1.078705\dots$. (It is clear that the conformal radius of the unit square lies between 1 and $\sqrt{2}$.)

Quadrilaterals. A *quadrilateral* Q is a Jordan domain with 4 distinguished points on its boundary. A conformal map between quadrilaterals is required to preserve these points as a set.

Theorem 4.30 *Any quadrilateral is conformally equivalent to a rectangle of the form $R(a) = [0, a] \times [0, 1]$, $a > 0$. This rectangle is unique up to $a \mapsto 1/a$.*

Proof. By the Riemann mapping theorem, any quadrilateral is isomorphic to \mathbb{H} with 4 distinguished points, $q_1, \dots, q_4 \in \mathbb{R}$. By the Schwarz-Christoffel formula, $f(z) = \int \prod (z - q_i)^{-1/2} dz$ maps \mathbb{H} to a Euclidean rectangle, sending q_i to its vertices. ■

Besikovitch lemma. As an exercise, one can now use the method of extremal length to show, for any quadrilateral $Q \subset \mathbb{C}$, we have $\text{area}(Q) \geq AB$ where A and B are the minimum distances between opposite sides of Q .

The Schwarzian derivative and hypergeometric functions. Just as the nonlinearity can be used to find a formula for the Riemann mapping to a polygon with linear sides, the Schwarzian derivative can be used to find the Riemann mapping to a polygon with circular sides.

To describe a result in this direction, we first recall that for $f(z) = z^\alpha$ we have $Nf(z) = (\alpha - 1)/z$, and thus

$$Sf(z) = (Nf)' - \frac{(Nf)^2}{2} = \frac{1 - \alpha^2}{2} \frac{dz^2}{z^2}.$$

The quantity $1 - \alpha^2$ is a version of the residue for the quadratic differential $Sf(z) dz^2$. Now given α, β, γ there is a unique quadratic differential with double poles at $0, 1, \infty$ with the corresponding residues, and no other singularities. It is given by

$$Q(\alpha, \beta, \gamma) = \frac{1}{2} \left[\frac{1 - \alpha^2}{z^2} + \frac{1 - \beta^2}{(z - 1)^2} + \frac{1 - \gamma^2 - (2 - \alpha^2 - \beta^2)}{z(z - 1)} \right] dz^2.$$

The third term is chosen so that $Q \sim (1 - \gamma^2)/z^2$ as $z \rightarrow \infty$.

Theorem 4.31 *Let P be a circular triangle with interior angles $\pi\alpha$, $\pi\beta$ and $\pi\gamma$. Then the Riemann mapping $f : \mathbb{H} \rightarrow P$ sending 0 , 1 and ∞ to these vertices satisfies $Sf = Q(\alpha, \beta, \gamma)$.*

Proof. By Schwarz reflection, Sf extends to a meromorphic quadratic differential on $\widehat{\mathbb{C}}$ whose residues at 0 , 1 and ∞ are determined by the angles of P . ■

Ideal triangles. As an example, if P is an ideal triangle, then we have

$$Q = Q(0, 0, 0) = \frac{z^2 - z + 1}{2z^2(z - 1)^2}.$$

This differential is symmetric under the action of S_3 on \mathbb{H} , and it has zeros at the primitive sixth roots of unity.

Second order differential equations. To find f itself, we remark that any function f satisfying $Sf = Q$ can be expressed as the ratio $f = u_1/u_2$ of two independent solutions to the differential equation

$$u'' + (Q/2)u = 0.$$

In the case at hand, these solutions in turn can be expressed in terms of solutions to the *hypergeometric equation*

$$z(1 - z)u'' + Au' + Bu = 0$$

for suitable constants A and B .

4.3 The Picard theorems

We conclude this section with two further results concerning entire functions that can be related to geometric function theory.

Theorem 4.32 (Little Picard Theorem) *An entire function $f : \mathbb{C} \rightarrow \mathbb{C}$ which omits two values must be constant.*

Corollary 4.33 *A meromorphic function on \mathbb{C} can omit at most two values on $\widehat{\mathbb{C}}$.*

The Little Picard Theorem is equivalent to the assertion that there is no solution to the equation $e^f + e^g = 1$ where f and g are nonconstant entire functions. Similarly, it implies there is no solution to Fermat's equation $f^n + g^n = 1$, $n \geq 3$, unless the entire functions f and g are constant.

Theorem 4.34 (Great Picard Theorem) *An analytic function $f : U \rightarrow \mathbb{C}$ takes on every value in \mathbb{C} , with at most one exception, in every neighborhood of an essential singularity p .*

The first theorem follows from the second by considering the essential singularity at $z = 0$ of $f(1/z)$. These results generalize Liouville's theorem and the Weierstrass-Casorati theorem respectively.

Rescaling arguments. Here is a third result that initially seems unrelated to the first two.

Theorem 4.35 (Bloch's Theorem) *There exists a universal $R > 0$ such that for any $f : \Delta \rightarrow \mathbb{C}$ with $|f'(0)| = 1$, not necessarily univalent, there is an open set $U \subset \Delta$ (perhaps a tiny set near S^1) such that f maps U univalently to a ball of radius R .*

Corollary 4.36 *For any analytic map on Δ , the image $f(\Delta)$ contains a ball of radius $R|f'(0)|$.*

Note that the ball usually *cannot* be centered at $f(0)$; for example, $f(z) = \exp(nz)/n$ satisfies $f'(0) = 1$ but the largest ball about $f(0) = 1/n$ in $f(\Delta) \subset \mathbb{C}^*$ has radius $1/n$.

The optimal value of R is known as *Bloch's constant*. It satisfies $0.433 < \sqrt{3}/4 \leq R < 0.473$. The best-known upper bound comes from the Riemann surface branched with order 2 over the vertices of the hexagonal lattice.

These apparently unrelated theorems can both be proved using the same idea. (Cf. [BD] and references therein; for another way to relate these theorems, see [Re, Ch. 10].)

Proof of Bloch's theorem. Given $f : \Delta \rightarrow \mathbb{C}$, let

$$\|f'(z)\| = \|f'(z)\|_{\Delta, \mathbb{C}} = (1/2)|f'(z)|(1 - |z|^2)$$

denote the norm of the derivative from the hyperbolic metric to the Euclidean metric. By assumption, $\|f'(0)\| = 1/2$. We can assume (using $f(rz)$) that f is smooth on S^1 ; then $\|f'(z)\| \rightarrow 0$ as $|z| \rightarrow 1$, and thus $\sup \|\|f'(z)\|\|$ is achieved at some $p \in \Delta$.

Now replace f with $f \circ r$ where $r \in \text{Aut}(\Delta)$ moves p to zero. Replacing f with $af + b$ with $|a| < 1$, we can also arrange that $f(0) = 0$ and $\|f'(0)\| = 1$; this will only decrease the size of its unramified disk. Then $\|f'(z)\| \leq \|f'(0)\| = 1$, and thus $f|\Delta(1/2)$ ranges in a compact family of nonconstant analytic functions. Thus the new f has an unramified disk of definite radius; but then the old f does as well. ■

Rescaling proof of Picard's theorem. The proof will use the following remarkably general rescaling theorem. This argument is related to Bloch's proof, to Brody's reparameterization theorem and to other results in complex analysis.

Theorem 4.37 *Let $f_n : \mathbb{C} \rightarrow \mathbb{C}$ be a sequence of nonconstant entire functions. There after passing to a subsequence, there is a sequence of Möbius transformations A_n and a nonconstant entire function $g : \mathbb{C} \rightarrow \mathbb{C}$ such that $g = \lim f_n \circ A_n$ uniformly on compact subsets of \mathbb{C} .*

We note that the A_n need not fix infinity, so $f_n \circ A_n$ is undefined at some point $p_n \in \widehat{\mathbb{C}}$, but we will have $p_n \rightarrow \infty$.

Example. For $f_n(z) = z^n$ we can take $f_n \circ A_n(z) = (1 + z/n)^n \rightarrow e^z$.

Metrics. As a preliminary to the proof, for $g : \mathbb{C} \rightarrow \widehat{\mathbb{C}}$ we define

$$\|g'(z)\|_\infty = \frac{|g'(z)|}{(1 + |g(z)|^2)},$$

and $\|g'\|_\infty = \sup \|g'(z)\|$ over all $z \in \mathbb{C}$. This is the norm of the derivative from the scaled Euclidean metric $\rho_\infty = 2|dz|$ to the spherical metric. Note that $g(z) = \exp(z)$ has $\|g'\|_\infty = 1/2$; a function with bounded derivative can be rather wild.

Similarly, for $g : \Delta(R) \rightarrow \widehat{\mathbb{C}}$, we define

$$\|g'(z)\|_R = |g'(z)| \frac{1 - |z/R|^2}{1 + |g(z)|^2}.$$

This is the derivative from a suitably rescaled hyperbolic metric ρ_R on $\Delta(R)$ to $\widehat{\mathbb{C}}$. Clearly $\rho_R \rightarrow \rho_\infty$ uniformly on compact sets. Its key property is that $\|(g \circ A)'\|_R = \|g'\|_R$ for all $A \in \text{Aut}(\widehat{\mathbb{C}})$ stabilizing $\Delta(R)$.

We also note that the set of maps with uniformly bounded derivatives in one of these norms is compact.

Proof of Theorem 4.37. Let us first consider an arbitrary nonconstant analytic function $f(z)$ and a radius $R > 0$. We claim there exists an $S \geq R$ and an $A \in \text{Aut}(\widehat{\mathbb{C}})$ such that $g = f \circ A$ is analytic on $\Delta(S)$, and

$$\|g'(0)\|_S = \|g'\|_S = 1.$$

Indeed, by replacing f with $f(az + b)$, we can assume $\|f'(0)\|_R = 1$. Then $\|f'\|_R \geq 1$. On the other hand, the R -norm of the derivative of f tends to zero at the boundary of $\Delta(R)$. Thus we can choose $B \in \text{Aut}(\Delta(R))$ such

$$M = \|(f \circ B)'(0)\|_R = \|(f \circ B)'\|_R \geq 1.$$

Now just let $g(z) = (f \circ B)(z/M)$, and $S = RM$.

Applying this claim to f_n and a sequence $R_n \rightarrow \infty$, we obtain $S_n \rightarrow \infty$ and maps $g_n = f_n \circ A_n$ with $\|g'_n(0)\|_\infty = 1$ and $\|g'_n\|_{S_n} \leq 1$. Now pass to a convergent subsequence. ■

To complete the proof of Picard's theorem, we observe:

Lemma 4.38 *If $f_n \rightarrow f$ and f is nonconstant, then any value omitted by all f_n is also omitted by f .*

Proof of the Little Picard Theorem. Suppose if $f : \mathbb{C} \rightarrow \mathbb{C}$ is nonconstant and omits 0 and 1. Then $f_n(z) = f_n^{1/n}(z)$ omits more and more points on the unit circle. We can rescale in the domain so the spherical derivative satisfies $\|f'_n(0)\|_\infty \rightarrow \infty$. Passing to a subsequence and reparameterizing, we obtain in the limit a nonconstant entire function that omits the unit circle. This contradicts Liouville's theorem. ■

Classical proof. The classical proof of the Little Picard Theorem is based on the fact that the universal cover of $\mathbb{C} - \{0, 1\}$ can be identified with the upper halfplane.

To see this, it is useful to start by considering the subgroup $\Gamma_0 \subset \text{Isom}(\Delta)$ generated by reflections in the sides of the ideal triangle T with vertices $\{1, i, -1\}$. For example, $z \mapsto \bar{z}$ is one such reflection, sending T to $-T$. By considering billiards in T , one can see that its translates tile the disk and thus T is a fundamental domain for Γ_0 . Thus the quadrilateral $F = T \cup (-T)$ is a fundamental domain for the orientation-preserving subgroup $\Gamma \subset \Gamma_0$, and the edges of $-T$ are glued to the edges of T to give a topological triply-punctured sphere as quotient.

Now let $\pi : T \rightarrow \mathbb{H}$ be the Riemann mapping sending T to \mathbb{H} and its vertices to $\{0, 1, \infty\}$. Developing in both the domain and range by Schwarz reflection, we obtain a covering map $\pi : \Delta \rightarrow \hat{\mathbb{C}} - \{0, 1, \infty\}$.

Given this fact, we lift an entire function $f : \mathbb{C} \rightarrow \mathbb{C} - \{0, 1\}$ to a map $\tilde{f} : \mathbb{C} \rightarrow \mathbb{H}$, which is constant by Liouville's theorem.

Uniformization of planar regions. Once we know that $\hat{\mathbb{C}} - \{0, 1, \infty\}$ is uniformized by the disk, it is straightforward to prove:

Theorem 4.39 *The universal cover of any region $U \subset \hat{\mathbb{C}}$ with $|\hat{\mathbb{C}} - U| \geq 3$ is isomorphic to the unit disk.*

Sketch of the proof. Consider a basepoint p in the abstract universal cover $\pi : \tilde{U} \rightarrow U$, and let \mathcal{F} be the family of all holomorphic maps

$$f : (\tilde{U}, p) \rightarrow (\Delta, 0)$$

that are covering maps to their image. Using the uniformization of the triply-punctured sphere, we have that \mathcal{F} is nonempty. It is also a closed, normal family of functions in $\mathcal{O}(\tilde{U})$; and by the classical square-root trick, it contains a surjective function (which maximizes $|f'(p)|$). By the theory of covering spaces, this extremal map must be bijective. ■

Proof of the Great Picard Theorem. Let $f : \Delta^* \rightarrow \hat{\mathbb{C}} - \{0, 1, \infty\}$ be an analytic function. We will show f does not have an essential singularity at $z = 0$.

Consider a loop γ around the puncture of the disk. If f sends γ to a contractible loop on the triply-punctured sphere, then f lifts to a map into the universal cover \mathbb{H} , which implies by Riemann's removability theorem that f extends holomorphically over the origin.

Otherwise, by the Schwarz lemma, $f(\gamma)$ is a homotopy class that can be represented by an arbitrarily short loop. Thus it corresponds to a puncture, which we can normalize to be $z = 0$ (rather than 1 or ∞). It follows that f is bounded near $z = 0$ so again the singularity is not essential. ■

5 Elliptic functions and elliptic curves

In this section we discuss the field of meromorphic functions on the first compact Riemann surfaces after $\hat{\mathbb{C}}$, namely complex tori of the form $E = \mathbb{C}/\Lambda$.

These functions arise naturally when trying to solve one of the simplest calculus integrals:

$$F(x) = \int \frac{dx}{\sqrt{x^3 + ax + b}}.$$

Just as the inverses of the exponential and trigonometric functions arise from integrating $1/x$ and $1/\sqrt{x^2 + ax + b}$, the integrals of cubic polynomials involve *inverses* of elliptic functions. We will ultimately see this geometrically, using the Schwarz-Christoffel formula.

From the perspective of algebraic geometry, elliptic functions arise if one tries to sweep out or parameterize all solutions of a cubic equation such as

$$y^2 = x^3 + ax + b.$$

A special feature of the case of cubics is that the solutions (in projective space) form a group.

5.1 Doubly-periodic functions

Let $\Lambda \subset \mathbb{C}$ be a lattice – meaning a discrete subgroup such that $E = \mathbb{C}/\Lambda$ is compact. Then $\Lambda \cong \mathbb{Z}^2$ as an abstract group. We can choose a basis such that

$$\Lambda = \mathbb{Z}\alpha \oplus \mathbb{Z}\beta.$$

Every point of E is represented by an essentially unique point in the *period parallelogram* with vertices $0, \alpha, \beta$ and $\alpha + \beta$.

Meromorphic functions $F : E \rightarrow \widehat{\mathbb{C}}$ are the same thing as *doubly periodic functions* $f : \mathbb{C} \rightarrow \widehat{\mathbb{C}}$, i.e. meromorphic functions satisfying

$$f(z + \alpha) = f(z + \beta) = f(z)$$

(and hence $f(z + \lambda) = f(z)$ for all $\lambda \in \Lambda$. For a fixed Λ , these functions form a *field*, denoted $K(E)$.

Note that E forms a group under addition, and hence $E \subset \text{Aut}(E)$. We also have an involution $\eta : E \rightarrow E$ given by $\eta(z) = -z$. This is the full automorphism group of E provided Λ is not a square or hexagonal lattice. In the latter cases, $\text{Aut}(E)$ also contains a cyclic group $\mathbb{Z}/4$ or $\mathbb{Z}/6$.

Because of these automorphisms, if $F \in K(E)$ then so is $F(-z)$ and $F(z + p)$, $p \in E$.

We note that a holomorphic doubly-periodic function must be constant, by Liouville's theorem, because the period parallelogram is compact. So to find other functions we must allow poles. The simplest are constructed for each $k \geq 3$ by setting

$$\zeta_k(z) = \sum_{\Lambda} \frac{1}{(z - \lambda)^k}.$$

This sum does not quite converge for $k = 2$; it can be made to converge by writing

$$\wp(z) = \frac{1}{z^2} + \sum'_{\Lambda} \frac{1}{(z - \lambda)^2} - \frac{1}{\lambda^2}.$$

The critical properties of $\wp(z)$ are that it is *even*, and it has a unique pole of order 2 on E .

Theorem 5.1 *The Weierstrass \wp -function is doubly-periodic.*

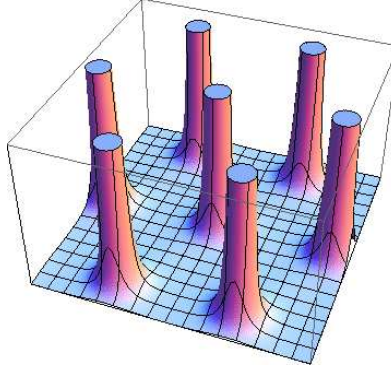


Figure 6. The Weierstrass \wp -function for the hexagonal lattice.

Proof. Note that $\wp'(z) = -2\zeta_3(z)$ is doubly-periodic. Thus $\wp(z+1) = \wp(z) + A$ and $\wp(z+\lambda) = \wp(z) + B$ for some constants A and B . But we also have $\wp(-z) = \wp(z)$; setting $z = -1/2$ and $-\lambda/2$, we find $A = B = 0$. ■

Uniformization of cubic curves. We can now state the main theorems regarding elliptic functions.

Theorem 5.2 *The map $\pi : \mathbb{C} \rightarrow \mathbb{P}^2$ given by*

$$\pi(z) = (\wp(z), \wp'(z))$$

gives an isomorphism between the Riemann surface $E = \mathbb{C}/\Lambda$ and a smooth cubic curve of the form $y^2 = (4x^3 + ax + b)$.

Theorem 5.3 *The field of all doubly-periodic functions for a given lattice Λ satisfies*

$$K(\mathbb{C}/\Lambda) \cong \mathbb{C}[x, y]/(y^2 - (4x^3 + ax + b)),$$

where $(x, y) = (\wp, \wp')$. In particular, every doubly-periodic function is a rational function of \wp and \wp' .

We remark that any smooth cubic curve $C \subset \mathbb{P}^2$ can be put into the form above by applying a change of coordinates (an automorphism of \mathbb{P}^2). Later, we will see that *every* cubic curve occurs for suitable choice of Λ .

Basic properties of elliptic functions. Let $f(z)$ be a nonconstant doubly-periodic function for a lattice $\Lambda \subset \mathbb{C}$ with period parallelogram P . Then f defines a meromorphic function on the compact Riemann surface $E = \mathbb{C}/\Lambda$. We also note that the form dz is invariant under Λ , so we also get a meromorphic 1-form $\omega = f(z) dz$ on E . Here are some basic facts.

1. The sum of the residues of $f(z) dz$ over E , or over points in P , is zero.
2. The function f has the same number of zeros as poles. The number of each is called the *degree* $d = \deg(f)$.
3. If f has poles p_1, \dots, p_d and zeros a_1, \dots, a_d in E , then $\sum p_i = \sum a_i$ in the group law on E .

Proofs. (1) This follows applying Stokes' theorem to the closed form $f(z) dz$ on $E - \{p_1, \dots, p_d\}$, or by integrating $f(z) dz$ around the boundary of P (we may assume f has no poles on ∂P .)

(2) This is a general property of proper maps between Riemann surfaces. For a direct proof, one can also apply the residue theorem to df/f .

(3) This property is special to elliptic curves. Let $P = [0, \alpha] \times [0, \beta]$, and assume f has no zeros or poles in P . Then we have

$$\sum a_i - \sum p_i = (2\pi i)^{-1} \int_{\partial P} \frac{zf'(z) dz}{f(z)}.$$

We wish to show that this quantity lies in Λ . The integrals over opposite edges cancel, up to a terms of the form $\lambda(2\pi i)^{-1} \int_e f'(z)/f(z) dz$ with $\lambda \in \Lambda$. Since the f is periodic, it has an integral winding number $N(e)$ on each edge, and these terms have the form $N(e)\lambda \in \Lambda$. ■

We will later see that we may *construct* an elliptic function with given zeros and poles subject only to constraint (3).

Pushforward. Here is another way to see property (3). Let $f : E \rightarrow \widehat{\mathbb{C}}$ be a nonconstant meromorphic function. Then $f_*(dz) = 0$, since $\Omega(\widehat{\mathbb{C}}) = 0$. Now choose a path $C \subset \widehat{\mathbb{C}}$ running from 0 to ∞ and avoiding the critical values of f . Then $\tilde{C} = f^{-1}(C) \subset E$ gives a collection of arcs connecting (a_i) and (p_i) in pairs, which we can assume have the same indices. We then have

$$0 = \int_C f_* \omega = \int_{\tilde{C}} dz = \sum p_i - a_i \mod \Lambda.$$

The differential equation. We can now give a cubic equation relating $x = \wp(z)$ and $y = \wp'(z)$.

Since $\wp'(z)$ is an odd function of degree three, its zeros coincide with the points $E[2]$ of order two on E . Using our chosen basis $\Lambda = \mathbb{Z}\alpha \oplus \mathbb{Z}\beta$, we can explicitly label representatives of these points:

$$E[2] = \{0, c_1, c_2, c_3\} = \left\{0, \frac{\alpha}{2}, \frac{\beta}{2}, \frac{\alpha + \beta}{2}\right\}.$$

Let $e_i = \wp(c_i)$.

We note that the critical values e_i are distinct; indeed, since $\wp'(z_i) = 0$, the function $\wp(z) - e_i$ has a double zero at z_i , so it cannot vanish anywhere else. This shows:

The zeros of $\wp'(z)$ coincide with the points of order two c_1, c_2, c_3 .

(Morally there is also a critical point at $z = 0$.) Consequently:

The function $\wp(z) - e_i$ has a double zero at c_i and no other zeros.

(Note: it is not at all easy to say where the two zeros of $\wp(z)$ lie, except in the case of symmetric lattices.)

Theorem 5.4 *For all $z \in \mathbb{C}$, we have*

$$\wp'(z)^2 = 4(\wp(z) - e_1)(\wp(z) - e_2)(\wp(z) - e_3).$$

Proof. From what we have seen above, the two sides are doubly-periodic functions with the same zeros and poles. Thus they are multiples of one another. They are equal since near zero they are both asymptotic to $4/z^6$. ■

Corollary 5.5 *The map $\wp : E \rightarrow \widehat{\mathbb{C}}$ presents E as a 2-sheeted covering space of $\widehat{\mathbb{C}}$, branched over e_1, e_2, e_3 and ∞ . It gives the quotient of E by the involution $z \mapsto -z$.*

Proof. By construction \wp is even and 2-to-1, and we have just identified its critical points and values. ■

Corollary 5.6 *The map $\pi(z) = (\wp(z), \wp'(z))$ maps $E = \mathbb{C}/\Lambda$ bijectively to the smooth projective cubic curve E defined by*

$$y^2 = 4(x - e_1)(x - e_2)(x - e_3).$$

Proof. Since E is compact, the function $x = \wp(z)$ maps E onto $\widehat{\mathbb{C}}$; and for a given x , the two possible values of y solving the equation above are given by $y = \wp'(\pm z)$. ■

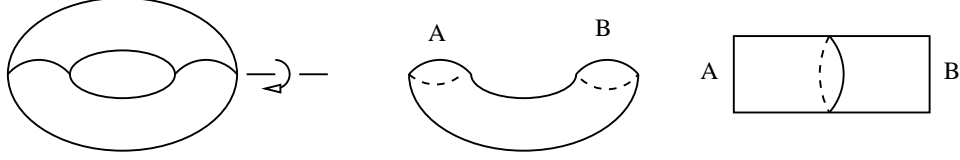


Figure 7. Visualizing the degree two map from E to $E/\eta \cong \widehat{\mathbb{C}}$, where $\eta(z) = -z$.

Power series expansion. Let $\mathcal{O}(d) \subset K(E)$ be the vector space of doubly-periodic functions with a pole of order d at $z = 0$ and no other poles.

Theorem 5.7 *We have $\dim \mathcal{O}(d) = d$ for $d \geq 2$, and $\dim \mathcal{O}(0) = 1$.*

Proof. Clearly a function in $\mathcal{O}(d)$ is uniquely determined by its polar part up to the constant term, $a_d/z^d + \cdots + a_0$. Moreover $a_1 = 0$ by the residue theorem. Thus $\dim \mathcal{O}(d) \leq d$. Using $\wp(z)$ and $\zeta_k(z)$ we obtain the reverse inequality. ■

Which element of $\mathcal{O}(2)$ is \wp ? In fact, it is the unique element with $a_2 = 1$ and $a_0 = 0$. This is evident from its definition, since each term in the series for $\wp(z)$, apart from the initial term $1/z^2$, is normalized to vanish at the origin.

More precisely, using the expansion

$$\frac{1}{(z - \lambda)^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2} \left[\frac{2z}{\lambda} + \frac{3z^2}{\lambda^2} + \cdots \right] = \frac{2z}{\lambda^3} + \frac{3z^2}{\lambda^4} + \frac{4z^3}{\lambda^5} + \cdots,$$

we obtain:

Theorem 5.8 *The Weierstrass \wp -function is given near $z = 0$ by:*

$$\wp(z) = \frac{1}{z^2} + 3G_2z^2 + 5G_3z^4 + \cdots = \frac{1}{z^2} + \sum_{n=1}^{\infty} (2n+1)z^{2n}G_{n+1}$$

where

$$G_n = G_n(\Lambda) = \sum'_{\Lambda} \frac{1}{\lambda^{2n}}.$$

Corollary 5.9 *We have $\wp'(z)^2 = 4\wp(z)^3 - g_2\wp(z) - g_3$, where $g_2 = 60G_2$ and $g_3 = 140G_3$.*

Proof. Neglecting terms of order $O(z^2)$, we have:

$$\begin{aligned}\wp(z)^3 &= \frac{1}{z^6} + \frac{9G_2}{z^2} + 15G_3 \quad \text{and} \\ \wp'(z)^2 &= \left(\frac{-2}{z^3} + 6G_2z + 20G_3z^3 + \cdots \right)^2 = \frac{4}{z^6} - \frac{24G_2}{z^2} - 80G_3.\end{aligned}$$

Thus $4\wp(z)^3 - \wp'(z)^2 = 60G_2/z^2 + 140G_3 + O(z^2) = g_2\wp(z) + g_3$. ■

Corollary 5.10 *We have $\sum e_i = 0$.*

Remark: other constructions of elliptic functions. To construct elliptic functions of degree two: if $\Lambda = \mathbb{Z} \oplus \mathbb{Z}\tau$, one can first sum over \mathbb{Z} to get:

$$f_1(z) = \sum_{n=-\infty}^{\infty} \frac{1}{(z-n)^2} = \frac{\pi^2}{\sin(\pi z)^2};$$

then

$$f(z) = \sum_{n=-\infty}^{\infty} f_1(z - n\tau)$$

converges rapidly, and defines an elliptic function of degree two. Similarly, if $E = \mathbb{C}^*/\langle z \mapsto \alpha z \rangle$, with $|\alpha| \neq 0, 1$, then we

$$F(w) = \sum_{n=-\infty}^{\infty} \frac{\alpha^n w}{(\alpha^n w - 1)^2}$$

defines a function on \mathbb{C}^* with a double pole at $w = 1$ satisfying $F(\alpha w) = F(w)$; thus $f(z) = F(e^z)$ is a degree two doubly-periodic function for the lattice $\Lambda = \mathbb{Z}2\pi i \oplus \mathbb{Z} \log \alpha$. (Note that $z/(z-1)^2$ has simple zeros at $0, \infty$ and a double pole at $z = 1$.)

These functions are not quite canonical; there is a choice of direction in the lattice to sum over first. As a consequence they agree with $\wp(z)$ only up to a constant. This constant is a multiple of the important *quasimodular form*

$$G_1(\tau) = \sum_m \left(\sum_n \frac{1}{(m+n\tau)^2} \right).$$

The real case. Now suppose $\Lambda = \mathbb{Z}\alpha \oplus \mathbb{Z}\beta$ with $\alpha \in \mathbb{R}_+$ and $\beta \in i\mathbb{R}_+$. Then the critical points $(0, c_1, c_2, c_3)$ of $\wp(z)$ bound a rectangle $S \subset \mathbb{C}$ (see Figure 8).

Theorem 5.11 *The value $\wp(z)$ is real if and only if z lies on one of the vertical or horizontal lines through $(1/2)\Lambda$.*

Proof. Since Λ is invariant under both negation and complex conjugation, we have

$$\wp(\bar{z}) = \wp(-\bar{z}) = \overline{\wp(z)}.$$

Thus $\wp(z)$ is real on the locus $R \subset E$ which is fixed under $z \mapsto \bar{z}$ or $z \mapsto -\bar{z}$. This locus is covered in \mathbb{C} by the vertical and horizontal lines through the points of $(1/2)\Lambda$. (E.g. the locus fixed by $z \mapsto \bar{z}$ consists of two horizontal loops on E , because $x + \beta/2 \mapsto x - \beta/2 \sim x + \beta/2$.)

Now as z traverses ∂S , it moves monotonically along $\hat{\mathbb{R}}$ passing through ∞ just once. Thus \wp maps S bijectively to $\hat{\mathbb{R}}$, and the same for $-S$. Since \wp has degree two, there are no other preimages of $\hat{\mathbb{R}}$. ■

Corollary 5.12 *The map $\wp|_S$ gives the unique conformal map from S to $-\mathbb{H}$ such that $\wp(c_i) = e_i$ for $i = 1, 2, 3$ and $\wp(0) = \infty$.*

Proof. The region S is a component of the complement of $\wp^{-1}(\hat{\mathbb{R}})$, containing no critical points of \wp , so it maps homeomorphically to \mathbb{H} or $-\mathbb{H}$. In fact the image is $-\mathbb{H}$ because $\text{Im } \wp(z) \sim \text{Im } 1/z^2 < 0$ for small z in S . ■

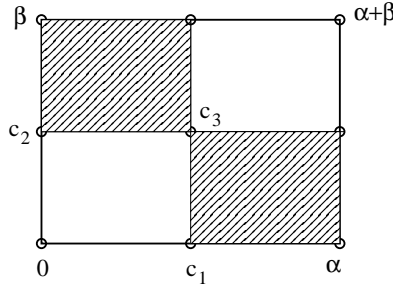


Figure 8. The shaded rectangles map under \wp to \mathbb{H} ; the unshaded ones, to $-\mathbb{H}$.

Theorem 5.13 *Any branch $f : (\pm\mathbb{H}) \rightarrow \mathbb{C}$ of \wp^{-1} satisfies*

$$f(z) = \int \frac{d\zeta}{\sqrt{4\zeta^3 - g_2\zeta - g_3}}.$$

Proof. If $\zeta = \wp(z)$ then $f(\zeta) = z$, and hence $f'(\wp(z))\wp'(z) = 1$. Consequently

$$f'(\zeta) = \frac{1}{\wp'(z)} = \frac{1}{\sqrt{4\wp(z)^3 - g_2\wp(z) - g_3}} = \frac{1}{\sqrt{4\zeta^3 - g_2\zeta - g_3}}.$$

■

Thus the theory of elliptic functions emerges as the special case of the Schwarz-Christoffel formula giving the Riemann map for a rectangle.

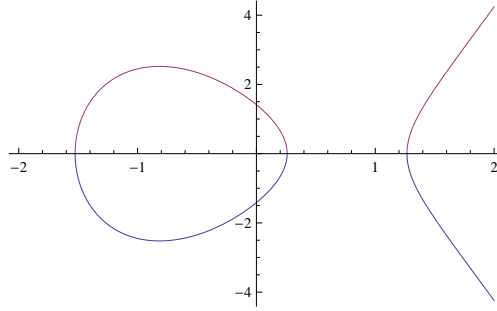


Figure 9. The elliptic curve $y^2 = 4x^3 - 8x + 2$.

Location of roots. Note that the argument shows the roots of $4x^3 - g_2x - g_3 = 0$ are real, and satisfy $e_2 < e_3 < e_1$. We also note that $\wp'(z) \in \mathbb{R}$ along the horizontal lines forming $\wp^{-1}(\mathbb{R})$, and $\wp'(z) \in i\mathbb{R}$ along the vertical lines, since \wp must rotate the latter by 90° to make them real. Thus only the horizontal lines map to the real points of the cubic $y^2 = 4x^3 - g_2x - g_3$.

Periods. This picture makes clear the close relationship between the generators α, β of Λ and the cubic polynomial

$$4x^3 - g_2x - g_3 = 4(x - e_1)(x - e_2)(x - e_3);$$

namely, we have

$$\frac{\alpha}{2} = \int_{e_1}^{\infty} (4x^3 - g_2x - g_3)^{-1/2} dx = \int_{e_2}^{e_3} (4x^3 - g_2x - g_3)^{-1/2} dx,$$

and

$$\frac{\beta}{2i} = \int_{-\infty}^{e_2} (g_3 + g_2x - 4x^3)^{-1/2} dx = \int_{e_3}^{e_1} (g_3 + g_2x - 4x^3)^{-1/2} dx.$$

In all 4 integrals we take the positive square-root; thus, these are simply integrals of $|(\wp')^{-1}(x)|$.

In each case, the fact that the two integrals agree follows from Cauchy's theorem (applied to two homotopic loops in $\mathbb{C} - \{e_1, e_2, e_3\}$), or via a change of variables coming from a Möbius transformation that swaps one interval for the other.

More generally, we have:

Theorem 5.14 *The lattice Λ with invariants g_2 and g_3 is generated by the values of*

$$\pm \int_{\gamma} \frac{dz}{\sqrt{4z^3 - g_2z - g_3}},$$

where γ ranges over all oriented loops in \mathbb{C} which enclose exactly two roots of the cubic in the denominator.

The condition on the roots insures that the integrand can be defined continuously on a neighborhood of γ .

Evaluation of g_2 and g_3 . The preceding formula sometimes permits the evaluation of g_2 and/or g_3 . For example, we have

$$140 \sum'_{\lambda \in \mathbb{Z}[\rho]} \lambda^{-6} = 4 \left(\int_1^{\infty} \frac{dx}{\sqrt{x^3 - 1}} \right)^6 = \frac{256\pi^3 \Gamma[7/6]^6}{\Gamma[2/3]^6}.$$

This follows from the fact that for $\Lambda = \mathbb{Z}[\rho]$, we have $g_2 = 0$ and $g_3 > 0$ satisfies

$$\frac{1}{2} = \int_{e_1}^{\infty} \frac{dx}{\sqrt{4x^3 - g_3}},$$

where $e_1 = (g_3/4)^{1/3}$. Similarly, we have

$$60 \sum'_{\lambda \in \mathbb{Z}[i]} \lambda^{-4} = \frac{64\pi^2 \Gamma[5/4]^4}{\Gamma[3/4]^4}.$$

Function fields. We now return to the case of general elliptic curves.

Theorem 5.15 *The function field of $E = \mathbb{C}/\Lambda$ is generated by $x = \wp$ and $y = \wp'$; more precisely, we have*

$$K(E) = \mathbb{C}(x, y)/(y^2 - 4x^3 + g_2x + g_3).$$

Proof. To see that \wp and \wp' generate $K(E)$ is easy. Any even function $f : E \rightarrow \widehat{\mathbb{C}}$ factors through \wp : $f(z) = F(\wp(z))$, and so lies in $\mathbb{C}(\wp)$. Any odd function becomes even when multiplied by \wp' ; and any function is a sum of one even and one odd. ■

To see that the field is exactly that given is also easy. It amounts to showing that $K(E)$ is of degree exactly two over $\mathbb{C}(\wp)$, and \wp is transcendental over \mathbb{C} . The first assertion is obvious (else \wp would be constant), and if the second fails we would have $K(E) = \mathbb{C}(\wp)$, which is impossible because \wp is even and \wp' is odd. ■

The addition law on an elliptic curve. Consider the curve $E \subset \mathbb{P}^2$ defined by $y^2 = 4x^3 - g_2x - g_3$ and parameterized by the Weierstrass \wp -function via $(x, y) = (\wp(z), \wp'(z))$.

Theorem 5.16 *For any line L , the intersection $L \cap E = \{a, b, c\}$ where $a + b + c = 0$ on $E = \mathbb{C}/\Lambda$.*

Proof. The intersection $L \cap E$ is simply the zero set of $A\wp' + B\wp + C$ for some (A, B, C) . This function has all its poles at $z = 0$. Since the sum of the zeros and poles is zero, its zeros (a, b, c) also sum to zero. ■

Corollary 5.17 *For any $z, w \in \mathbb{C}$ we have*

$$\begin{vmatrix} 1 & 1 & 1 \\ \wp(-z) & \wp(-w) & \wp(z+w) \\ \wp'(-z) & \wp'(-w) & \wp'(z+w) \end{vmatrix} = 0.$$

Corollary 5.18 *The map $p \mapsto -p$ on E is given by $(x, y) \mapsto (x, -y)$.*

Proof. Then the line passes through ∞ which is the origin of E , consistent with the equation $p + (-p) + 0 = 0$. (Alternatively, observe that $x = \wp(z)$ is even and $y = \wp'(z)$ is odd.) ■

Corollary 5.19 *The point $c = a + b$ is constructed geometrically by drawing the line L through (a, b) , finding its third point of intersection $(-c) = (x, -y)$ on E and then negating to get $c = (x, y)$.*

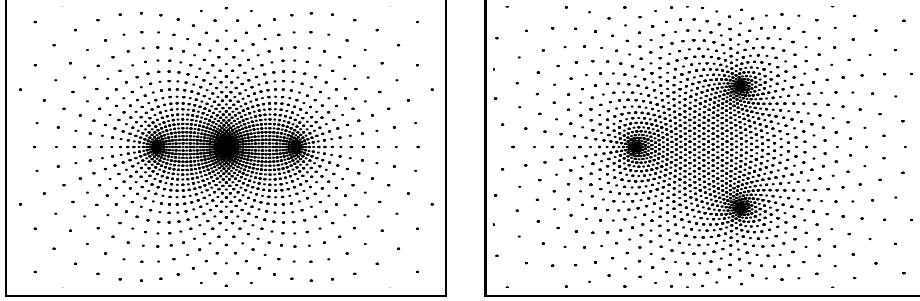


Figure 10. Image of $2^{-5}\Lambda$ under \wp , for $\Lambda = \mathbb{Z}[i]$ and $\mathbb{Z}[\rho]$.

Corollary 5.20 *The point $2a$ can be constructed by taking the line tangent to E at a , finding its other point of intersection with E , and then negating its y coordinate.*

Dynamics of rational maps. There is a rational function $f(z)$ such that

$$\wp(2z) = f(\wp(z)).$$

The preimages of the critical points of f lie along the images of the horizontal and vertical lines in \mathbb{C} under \wp , so they give a way of visualizing the \wp function. For more details, see [Mil].

An elliptic function with given poles and zeros. It is natural to try to construct an elliptic function by forming the Weierstrass product for the lattice Λ :

$$\sigma(z) = z \prod'_{\lambda} \left(1 - \frac{z}{\lambda}\right) \exp\left(\frac{z}{\lambda} + \frac{z^2}{2\lambda^2}\right).$$

Then $(\log \sigma)'' = -\wp(z)$, from which it follows that

$$\sigma(z + \lambda) = \sigma(z) \exp(a_{\lambda} + b_{\lambda}z)$$

for some $a_{\lambda}, b_{\lambda} \in \mathbb{C}$. From this it is easy to see that

$$\frac{\sigma(z - a_1) \dots \sigma(z - a_n)}{\sigma(z - p_1) \dots \sigma(z - p_n)}$$

defines an elliptic function whenever $\sum a_i = \sum p_i$. This demonstrates:

Theorem 5.21 *A divisor $D = \sum a_i - \sum p_i$ on E is principal (it arises from a meromorphic function) iff $\deg D = 0$ and $\sum(a_i - p_i) = 0$ in the group law on E .*

5.2 Aside: Conics and singly-periodic functions

As a point of reference, we describe the parallel theory for conics and rank one lattices in \mathbb{C} .

Let $\Lambda \subset \mathbb{C}$ be a discrete subgroup isomorphic to \mathbb{Z} . Then we can rescale so $\Lambda = \mathbb{Z}$, and form, for $k \geq 2$, the singly-periodic functions

$$Z_k(z) = \sum_{n=-\infty}^{\infty} \frac{1}{(z-n)^k}$$

and the function

$$P(z) = \frac{1}{z} + \sum_{n=-\infty}^{\infty} \left(\frac{1}{(z-n)} - \frac{1}{n} \right).$$

Note that we have the Laurent series (with $\zeta(0) = -1/2$),

$$P(z) = \frac{1}{z} - z \sum_{n=1}^{\infty} \frac{1}{n^2} - z^3 \sum_{n=1}^{\infty} \frac{1}{n^4} - \cdots = \frac{2}{z} \sum_{k=0}^{\infty} \zeta(2k) z^{2k},$$

i.e. the coefficients of $P(z)$ carry interesting invariants of the lattice $\Lambda = \mathbb{Z}$.

We then find:

Theorem 5.22 *The map $P : \mathbb{C}/\mathbb{Z} \rightarrow \widehat{\mathbb{C}} - (\pm\pi i)$ is an isomorphism, sending $[0, 1]$ to $[\infty, -\infty]$.*

Proof. As we have seen,

$$P(z) = \pi \cot \pi z = \pi i \frac{e^{2\pi i z} + 1}{e^{2\pi i z} - 1} = A(e^{2\pi i z}),$$

where $A(z) = \pi i(z+1)/(z-1)$ sends the omitted values 0 and ∞ for $e^{2\pi i z}$ to $\pm i$. ■

Just as we did for elliptic functions, we next note that the power series

$$P(z) = \frac{1}{z} - \frac{\pi^2 z}{3} + \cdots$$

gives

$$-P'(z) = \frac{1}{z^2} + \frac{\pi^2}{3} + \cdots \quad \text{and} \quad P(z)^2 = \frac{1}{z^2} - \frac{2\pi^2}{3} + \cdots$$

yielding the differential equation

$$-P'(z) = P(z)^2 + \pi^2.$$

Here we have used the fact that both $P(z)$ and $P'(z)$ are bounded as $|\operatorname{Im} z| \rightarrow \infty$. Put differently, we have:

Theorem 5.23 *The map $\pi : \mathbb{C} \rightarrow \mathbb{P}^2$ given by*

$$\pi(z) = (x, y) = (P(z), -P'(z))$$

gives an isomorphism between the $\mathbb{C}/\mathbb{Z} \cong \mathbb{C}^$ and the smooth projective conic defined by $y = x^2 + \pi^2$ with two points removed.*

The omitted values on the parabola $y = x^2 + \pi^2$ are $y = 0$, where $x = \pm i\pi$.

We could also verify the final identity using the fact that

$$-P'(z) = \frac{\pi^2}{\sin^2 \pi z}.$$

The fundamental period of Λ can now be expressed as

$$\int_{-\infty}^{\infty} \frac{dx}{y} = \int_{-\infty}^{\infty} \frac{dx}{\pi^2 + x^2} = \int_0^1 dz = 1.$$

I.e. the change of variables $x = P(z)$ transforms the integrand into the standard form dz on \mathbb{C}/\mathbb{Z} .

Remark. Any smooth conic in \mathbb{P}^2 is equivalent to the parabola above, so we have uniformized all conics. The familiar conics $x^2 + y^2 = 1$, $x^2 - y^2 = 1$ and $xy = 1$ are isomorphic to $\mathbb{C}/2\pi\mathbb{Z}$ or $\mathbb{C}/2\pi i\mathbb{Z}$, and are uniformized by $(\cos(t), \sin(t))$, $(\cosh(t), \sinh(t))$ and (e^t, e^{-t}) respectively. Note that all 3 curves have, over \mathbb{C} , two asymptotes, corresponding to the ends of \mathbb{C}^* .

5.3 Moduli spaces and elliptic curves

We now turn to an important complement to the results above.

Theorem 5.24 *Suppose the cubic polynomial $4x^3 + ax + b$ has distinct roots. Then there exists a lattice Λ such that $(x, y) = (\wp(z), \wp'(z))$ satisfies $y^2 = 4x^3 + ax + b$.*

Corollary 5.25 *Any smooth cubic curve in \mathbb{P}^2 is isomorphic to \mathbb{C}/Λ for some $\Lambda \subset \mathbb{C}$.*

Moduli spaces. To approach this ‘metatheory’ of doubly-periodic functions — where the lattice is not fixed but allowed to vary — it is useful to discuss \mathcal{M}_1 , the *moduli space of lattices*, and the *Teichmüller space* $\mathcal{T}_1 \cong \mathbb{H}$ of *marked* lattices.

As a set, we let

$$\mathcal{M}_1 = \{\text{all lattices } \Lambda \subset \mathbb{C}\} / (\Lambda \sim \alpha\Lambda)$$

denote the space of lattices up to similarity. By associating to Λ the complex torus $E = \mathbb{C}/\Lambda$, we find

$$\mathcal{M}_1 = \{\text{all Riemann surfaces of genus one}\} / (\text{isomorphism}).$$

This is because any isomorphism $E_1 \cong E_2$ lifts to an automorphism $f(z) = \alpha z + \beta$ of the universal cover \mathbb{C} .

There is a natural map $\pi : \mathbb{H} \rightarrow \mathcal{M}_1$, given by

$$\pi(\tau) = [\mathbb{Z} \oplus \mathbb{Z}\tau] \in \mathcal{M}_1.$$

This map is surjective, since $\mathbb{Z}\alpha \oplus \mathbb{Z}\beta \sim \mathbb{Z} \oplus \mathbb{Z}(\beta/\alpha)$. We can think of a point $\tau \in \mathbb{H}$ as providing both a lattice $\Lambda = \mathbb{Z} \oplus \mathbb{Z}\tau$ and a marking isomorphism,

$$\phi : \mathbb{Z}^2 \rightarrow \Lambda,$$

coming from the basis $(\tau, 1)$. All other bases for Λ with the same orientation as this one are given by $(a\tau + b, c\tau + d)$, where $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \text{SL}_2(\mathbb{Z})$. The marked lattice with this new basis is equivalent to $(g(\tau), 1)$ where

$$g(\tau) = \frac{a\tau + b}{c\tau + d}.$$

Thus forgetting the marking altogether is the same as taking the quotient of \mathbb{H} by the action of all $g \in \text{SL}_2(\mathbb{Z})$, and hence:

Theorem 5.26 *The moduli space \mathcal{M}_1 is naturally isomorphic to the complex orbifold $\mathbb{H}/\text{SL}_2(\mathbb{Z})$.*

Here $\text{SL}_2(\mathbb{Z})$ acts by Möbius transformations. The quotient is an *orbifold* because the action is not quite free, e.g. $\tau \mapsto -1/\tau$ fixes $\tau = i$.

Lie groups approach. More generally, we can specify a marked lattice $\Lambda \subset \mathbb{R}^n$, normalized so \mathbb{R}^n/Λ has volume one, by a linear map

$$T : \mathbb{Z}^n \rightarrow \Lambda \subset \mathbb{R}^n$$

with $T \in \text{SL}_n(\mathbb{R})$. Two such lattices are similar iff they differ by a rotation, i.e. $T_1 = R \circ T_2$ where $R \in \text{SO}_n(\mathbb{R})$. Thus the Teichmüller space of lattices in \mathbb{R}^n is the homogeneous space

$$\mathcal{H} = \text{SO}_n(\mathbb{R}) \backslash \text{SL}_n(\mathbb{R}).$$

For $n = 2$ we have $\mathcal{H} \cong \mathbb{H}$ because $\mathrm{SO}_2(\mathbb{R})$ is the stabilizer of $\tau = i$ for the usual action of $\mathrm{SL}_2(\mathbb{R})$ on \mathbb{H} .

Finally $T_1(\mathbb{Z}^n) = T_2(\mathbb{Z}^n)$ iff $T_1 \circ T_2^{-1}$ gives an isomorphism of \mathbb{Z}^n to itself, which shows

Theorem 5.27 *The moduli space of lattices in \mathbb{R}^n is isomorphic to*

$$\mathcal{L}(\mathbb{R}^n) = \mathrm{SO}_n(\mathbb{R}) \backslash \mathrm{SL}_n(\mathbb{R}) / \mathrm{SL}_n(\mathbb{Z}).$$

Fundamental domains. We can always normalize a lattice Λ by \mathbb{C}^* so that its shortest nonzero vector is $z = 1$ and the shortest vector in $\Lambda - \mathbb{Z}$ is $\tau \in \mathbb{H}$. Then $|\tau| \geq 1$, and $\mathbb{Z} + \tau \subset \Lambda$ so $|\mathrm{Re} \tau| \leq 1/2$. Moreover $\mathbb{Z} \oplus \mathbb{Z}\tau = \Lambda$; otherwise there would be a vector $v \in \Lambda - \mathbb{R}$ of the form $v = a + b\tau$ with $a, b \in [0, 1/2]$; but then

$$|v| < 1/2 + |\tau|/2 \leq |\tau|,$$

contrary to our assumption that the shortest vector in $\Lambda - \mathbb{Z}$ has length $|\tau|$. The converse holds as well, and we have:

Theorem 5.28 *The region $|\mathrm{Re} \tau| \leq 1/2$, $|\tau| > 1$ in \mathbb{H} is a fundamental domain for the action of $\mathrm{SL}_2(\mathbb{Z})$ on \mathbb{H} .*

The subgroup $\Gamma(2)$. To study \mathcal{M}_1 further, we now introduce the space of cross-ratios

$$\widetilde{\mathcal{M}}_{0,4} = \widehat{\mathbb{C}} - \{0, 1, \infty\}$$

and its quotient orbifold

$$\mathcal{M}_{0,4} = \widetilde{\mathcal{M}}_{0,4} / S_3.$$

Here $\widetilde{\mathcal{M}}_{0,4}$ is the moduli space of ordered quadruples of distinct points on $\widehat{\mathbb{C}}$, up to the action of $\mathrm{Aut}(\widehat{\mathbb{C}})$. Any such quadruple has a unique representative of the form $(\infty, 0, 1, \lambda)$, giving a natural coordinate for this moduli space.

If we reorder the quadruple, the cross-ratio changes, ranging among the six values

$$\lambda, \frac{1}{\lambda}, 1 - \lambda, \frac{1}{1 - \lambda}, \frac{\lambda}{\lambda - 1}, \frac{\lambda - 1}{\lambda}.$$

(There is a natural action of S_4 , but the Klein 4-group $\mathbb{Z}/2 \times \mathbb{Z}/2$ acts trivially.) There are 5 points in $\widetilde{\mathcal{M}}_{0,4}$ with nontrivial stabilizers under the action of S_3 : the points

$$\{-1, 1/2, 2\} = \text{zeros of } 2z^3 - 3z^2 - 3z + 2,$$

which each have stabilizer $\mathbb{Z}/2$ and correspond to the vertices of a square; and the points

$$\{\rho, \bar{\rho}\} = \{1/2 \pm \sqrt{-3}/2\} = \text{zeros of } z^2 - z + 1,$$

which have stabilizer $\mathbb{Z}/3$ and correspond to the vertices of a tetrahedron.

The degree 6 rational map

$$F(z) = \frac{4}{27} \frac{(z^2 - z + 1)^3}{z^2(1 - z)^2}$$

is invariant under S_3 , and gives a natural bijection

$$F : (\widehat{\mathbb{C}} - \{0, 1, \infty\})/S_3 \cong \mathbb{C}$$

satisfying

$$F(0, 1, \infty) = \infty, F(\rho, \bar{\rho}) = 0, \quad \text{and} \quad F(-1, 1/2, 2) = 1.$$

(The last fact explains the 4/27.) We should really think of the image as $\mathcal{M}_{0,4}$ and in particular remember the orbifold structure: $\mathbb{Z}/2$ at $F = 1$ and $\mathbb{Z}/3$ at $F = 0$.

The modular function. We now define a map $J : \mathcal{M}_1 \rightarrow \mathcal{M}_{0,4}$ by associating to any complex torus, the four critical values of the Weierstrass \wp -function. (Note: any degree two map $f : X = \mathbb{C}/\Lambda \rightarrow \widehat{\mathbb{C}}$ is equivalent, up to automorphisms of domain and range, to the Weierstrass \wp function, so the associated point in $\mathcal{M}_{0,4}$ is canonically determined by X .)

More concretely, given $\tau \in \mathbb{H}$ we define the half-integral points $(c_1, c_2, c_3) = (1/2, \tau/2, (1 + \tau)/2)$ and the corresponding critical values by $e_i = \wp(c_i)$. Then their cross-ratio (together with the critical value at infinity) is given by:

$$\lambda(\tau) = \frac{e_3 - e_2}{e_1 - e_2}.$$

For example, we have seen that if $\tau = iy \in i\mathbb{R}_+$ then $e_2 < e_3 < e_1$, so $\lambda(iy) \in (0, 1)$; moreover $e_3 = 0$ and $e_2 = -e_1$ for $\tau = i$, and thus $\lambda(i) = 1/2$.

The value of $\lambda(\tau)$ depends only on the ordering of $E(2)^*$, the three nontrivial points of order two on E . Now $\text{SL}_2(\mathbb{Z}) = \text{Aut}(\Lambda)$ acts on $E(2) \cong (\mathbb{Z}/2)^2$ through the natural quotient

$$0 \rightarrow \Gamma(2) \rightarrow \text{SL}_2(\mathbb{Z}) \rightarrow \text{SL}_2(\mathbb{Z}/2) \rightarrow 0.$$

In particular, λ is invariant under the subgroup $\Gamma(2)$ of matrices equivalent to the identity modulo two.

Now any elliptic element in $\mathrm{SL}_2(\mathbb{Z})$ has trace $-1, 0$ or 1 , while the trace of any element in $\Gamma(2)$ must be even. Moreover, trace zero cannot arise: if $g = \begin{pmatrix} a & b \\ c & -a \end{pmatrix} \in \Gamma(2)$ then $-a^2 - bc = 1$ implies $-a^2 = 1 \pmod{4}$ which is impossible.

By assembling 6 copies of the fundamental domain for $\mathrm{SL}_2(\mathbb{Z})$ (some cut into two pieces), one can then show:

Theorem 5.29 *The group $\Gamma(2)$ is torsion-free, with fundamental domain the ideal quadrilateral with vertices $\{\infty, -1, 0, 1\}$.*

We may now state the main result relating lattices and cross-ratios.

Theorem 5.30 *The natural map*

$$\lambda : \widetilde{\mathcal{M}}_1 = \mathbb{H}/\Gamma(2) \rightarrow \widetilde{\mathcal{M}}_{0,4} = \widehat{\mathbb{C}} - \{0, 1, \infty\}$$

sending a torus to the cross-ratio of the critical values of \wp is a holomorphic bijection, respecting the action of S_3 .

Proof. We first check that λ is injective. If $z = \lambda(\tau_1) = \lambda(\tau_2)$, then the corresponding complex tori E_1, E_2 both admit degree two maps to $\widehat{\mathbb{C}}$ branched over $(0, 1, \infty, z)$. Lifting one map composed with the inverse of the other gives an isomorphism $E_1 \rightarrow E_2$, and hence $\tau_1 \in \mathrm{SL}_2(\mathbb{Z}) \cdot \tau_2$.

We have also seen that there exists a continuous map $\mu : \mathcal{M}_{0,4} \rightarrow \mathcal{M}_1$ which satisfies $\mu(\lambda(\tau)) = [\tau]$; it sends z to the lattice spanned by integrals of the form $dw/(w(w-1)(w-z))$. The existence of this map shows the image of λ is closed. It follows that λ is a bijection. ■

As we have seen, if $\tau \in i\mathbb{R}$ then $e_1, e_2, e_3 \in \mathbb{R}$. This shows:

Theorem 5.31 *The function λ is real on the orbit of the imaginary axis under $\mathrm{SL}_2(\mathbb{Z})$.*

This orbit gives the edges of a tiling of \mathbb{H} by ideal triangles; see Figure 11.

Corollary 5.32 *The map λ gives an explicit Riemann map from the ideal triangle spanned by $0, 1$ and ∞ to \mathbb{H} , fixing these three points.*

Thus $S\lambda^{-1}$ is given by the quadratic differential $Q(0, 0, 0)$ discussed earlier, and λ^{-1} itself can be given as a ratio of solutions to a linear differential equation of order two.

Picard's theorem revisited. This gives the one line proof of the Little Picard Theorem: 'consider $\lambda^{-1} \circ f : \mathbb{C} \rightarrow \mathbb{H}$.'

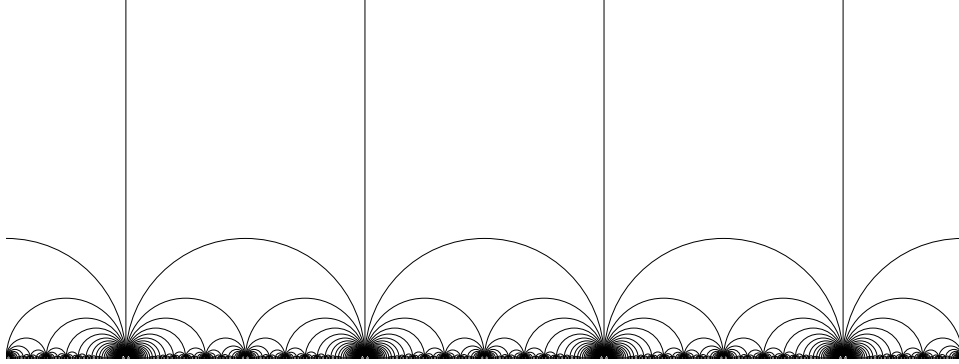


Figure 11. Tiling for $\Gamma(2)$.

The question recently arose in conversation whether a dissertation of 2 lines could deserve and get a Fellowship... in mathematics the answer is yes....

*(Theorem.) An integral function never 0 or 1 is a constant.
(Proof.) $\exp\{i\Omega(f(z))\}$ is a bounded integral function.*

...But I can imagine a referee's report: 'Exceedingly striking and a most original idea. But, brilliant as it undoubtedly is, it seems more odd than important; an isolated result, unrelated to anything else, and not likely to lead anywhere.'

—J. E. Littlewood.

Cf. [Bol, p.39–40]: Here Ω is our λ^{-1} .

The J -function. To remove the ambiguity of ordering, we now define

$$J(\tau) = F(\lambda(\tau)) = \frac{4}{27} \frac{(\lambda^2 - \lambda + 1)^3}{\lambda^2(1 - \lambda)^2}.$$

The S_3 -equivariance of λ then implies:

Theorem 5.33 *The map*

$$J : \mathcal{M}_1 = \mathbb{H}/\mathrm{SL}_2(\mathbb{Z}) \rightarrow \mathcal{M}_{0,4} = (\widehat{\mathbb{C}} - \{0, 1, \infty\})/S_3$$

is a bijection, and a isomorphism of orbifolds.

Forgetting the orbifold structure, we get an isomorphism $J : \mathcal{M}_1 \cong \mathbb{C}$. In other words, $J(\tau)$ is a complex number which depends only on the

isomorphism class of $E = \mathbb{C}/(\mathbb{Z} \oplus \mathbb{Z}\tau)$; we have $J(\tau_1) = J(\tau_2)$ iff $E_1 \cong E_2$; and every complex number arises as $J(\tau)$ for some τ .

Classical proof. Here is the classical argument that J is a bijection. The value of $J(\tau)$ determines a quadruple $B \subset \widehat{\mathbb{C}}$ which in turn determines a unique Riemann surface $X \rightarrow \widehat{\mathbb{C}}$ of degree two, branched over B , with $X \cong \mathbb{C}/\mathbb{Z} \oplus \tau\mathbb{Z}$. Thus $J(\tau_1) = J(\tau_2)$ iff the corresponding complex tori are isomorphic iff $\tau_1 = g(\tau_2)$ for some $g \in \mathrm{SL}_2(\mathbb{Z})$. Thus J is injective.

To see it is surjective, we first observe that $J(\tau + 1) = J(\tau)$. Now if $\mathrm{Im} \tau = y \rightarrow \infty$, then on the region $|\mathrm{Im} z| < y/2$ we have

$$\wp(z) = \frac{1}{z^2} + \sum' \frac{1}{(z-n)^2} - \frac{1}{n^2} + \epsilon(z) = \frac{\pi^2}{\sin^2(\pi z)} + C + \epsilon(z),$$

where $\epsilon(z) \rightarrow 0$ as $y \rightarrow \infty$. (In fact we have

$$\epsilon(z) = \sum'_n \frac{\pi^2}{\sin^2(\pi(z+n\tau))} = O(e^{-\pi y}).$$

We really only need that it tends to zero; and we will not need the exact value of the constant $C = -\pi^2/3$).

Since $\sin(z)$ grows rapidly as $|\mathrm{Im} z|$ grows, we have

$$(e_1, e_2, e_3) = (\pi^2 + C, C, C) + O(\epsilon)$$

and thus $\lambda(\tau) = (e_3 - e_2)/(e_1 - e_2) \rightarrow 0$ and hence $J(\tau) \rightarrow \infty$.

Thus J is a *proper* open map, which implies it is surjective (its image is open and closed). ■

5.4 Modular forms

In this section we will find an explicit formula for $J(\tau)$, and give a complete analysis of the holomorphic forms and functions on \mathbb{H} (with controlled growth) that are invariant under $\mathrm{SL}_2(\mathbb{Z})$.

Modular forms: analytic perspective. A holomorphic function $f : \mathbb{H} \rightarrow \mathbb{C}$ is said to be a *modular form* of weight $2k$ for $\mathrm{SL}_2(\mathbb{Z})$ if it has the following properties:

1. $f(\tau + 1) = f(\tau)$;
2. $f(-1/\tau) = \tau^{2k} f(\tau)$; and
3. $\sup_{\mathrm{Im} \tau > 1} |f(\tau)| < \infty$.

The vector space of all such forms will be denoted by M_k . The product forms of weight $2k$ and 2ℓ has weight $2(k + \ell)$, so $\oplus M_k$ forms a graded ring.

The first two properties imply that

$$f(g(\tau)) = (c\tau + d)^{2k} f(\tau)$$

for all $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$. The first and third properties imply that we can write

$$f(\tau) = \sum_0^\infty a_n q^n,$$

where $q = \exp(2\pi i\tau)$. In other words, $f(\tau)$ descends to a holomorphic function on $\Delta^* = \mathbb{H}/\mathbb{Z}$ with a removable singularity at $q = 0$. In particular, $f(\tau) \rightarrow a_0$ as $\mathrm{Im} \tau \rightarrow \infty$. One sometimes writes $f(i\infty) = a_0$.

If $f(i\infty) = 0$, one says that $f(\tau)$ is *cuspidal form*. The cuspidal forms (*Spitzenformen*) give a natural subspace $S_k \subset M_k$.

Examples.

1. Any holomorphic invariant $F(\Lambda)$ of lattices that is homogeneous of degree $-2k$ and satisfies the right growth conditions defines a modular form of weight $2k$. That is, if $F(t\Lambda) = t^{2k} F(\Lambda)$, then the function

$$f(\tau) = F(\mathbb{Z} \oplus \tau\mathbb{Z})$$

satisfies $f(\tau + 1) = f(\tau)$ and

$$f(-1/\tau) = F(\mathbb{Z} \oplus \tau^{-1}\mathbb{Z}) = F(\tau^{-1}(\mathbb{Z} \oplus \tau\mathbb{Z})) = \tau^{2k} f(\tau).$$

2. In particular, for $k \geq 2$ the *Eisenstein series*

$$G_k(\tau) = \sum' (n + m\tau)^{-2k}$$

are modular forms of weight $2k$. Evidently G_k converges to $2\zeta(2k)$ as $\mathrm{Im} \tau \rightarrow \infty$, so these are holomorphic at infinity but not cuspidal forms.

3. Similarly for the normalized functions $g_2 = 60G_2$, $g_3 = 140G_3$.

Modular forms: algebraic perspective. If we associate to $f(\tau)$ the form

$$\omega = f(\tau) d\tau^k,$$

then the first two conditions above just say that $g^*\omega = \omega$ for all $g \in \mathrm{SL}_2(\mathbb{Z})$. (Recall that $g'(z) = (cz+d)^{-2}$.) Since $dz = (2\pi i)^{-1}dq/q$, the third condition says that ω descends to a form

$$\omega = (2\pi i)^{-k} f(q) \left(\frac{dq}{q} \right)^k$$

on $\Delta^* = \mathbb{H}/\Gamma$ with at worst a pole of order k at $q = 0$.

Thus ω also descends to an S_3 -invariant holomorphic k -form on

$$\widetilde{\mathcal{M}}_{0,4} = \widehat{\mathbb{C}} - \{0, 1, \infty\} = \mathbb{H}/\Gamma(2)$$

with poles of order $\leq k$ at $0, 1$ and ∞ . The converse is also true. This shows:

Theorem 5.34 *The space M_k is naturally isomorphic to the space of S_3 -invariant rational forms $\omega(z) dz^k$ on $\widehat{\mathbb{C}}$ with poles of order $\leq k$ at $0, 1, \infty$ and no other poles.*

This provides a purely algebraic perspective on the initially transcendental-looking theory of automorphic forms.

We now wish to find all the rational holomorphic k -forms $\omega \in M_k$. Here are 2 key properties of any nonzero $\omega \in M_k$:

- (i) The zeros $Z(\omega)$ are invariant under S_3 and satisfy $|Z(\omega)| \leq k$, when counted with multiplicity, and determine ω up to constant multiple.
- (ii) The poles of ω at $0, 1, \infty$ are all of the same order, which can be $k, k-2, k-4$, etc. We have $\omega \in S_k$ iff the order of pole is $k-2$ or less.

Part (i) comes from the fact that ω has at most $3k$ poles, hence at most $3k - 2k = k$ zeros. The parity constraint in (ii) comes from the fact that $g(z) = 1 - z$ leaves ω invariant, and $g'(\infty) = -1$.

Examples.

1. We have $M_0 = \mathbb{C}$; it consists of the constant functions.
2. We have $\dim M_1 = 0$; the group S_3 has no fixed point, so there is no candidate for $Z(\omega)$.

3. We have $\dim M_2 = 1$. The only possibility is $Z(\omega) = \{\rho, \bar{\rho}\}$; thus the quadratic differential

$$F_2 = \frac{(z^2 - z + 1) dz^2}{z^2(z - 1)^2}$$

spans M_2 . (We have met this differential before in the study of Schwarz triangle functions.)

4. Similarly, $\dim M_3 = 1$; it is spanned by the cubic differential

$$F_3 = \frac{(z - 2)(z - 1/2)(z + 1) dz^3}{z^3(z - 1)^3},$$

which has zeros at $-1, 1/2, 2$.

5. The products F_2^2 and F_3F_2 span M_4 and M_5 . This is because S_3 has unique invariant sets with $|Z| = 4$ and 5.
6. We have $\dim M_6 = 2$; it is spanned by F_2^3 and F_3^2 . These two forms are linearly independent because they have different zero sets.

7. *The discriminant*

$$D_6 = \frac{4}{27}(F_2^3 - F_3^2) = \frac{dz^6}{z^4(z - 1)^4}$$

is the first nontrivial cusp form; it has poles of order 4 at $(0, 1, \infty)$ and no zeros. It spans S_6 .

8. Ratios of forms of the same weight give all S_3 -invariant rational functions. Indeed, as we have seen, an isomorphism $\widehat{\mathbb{C}}/S_3 \cong \widehat{\mathbb{C}}$ sending $(\infty, \rho, 2)$ to $(\infty, 0, 1)$, is given by

$$F(z) = \frac{F_2^3(z)}{F_2(z)^3 - F_3(z)^2} = \frac{4}{27} \frac{(z^2 - z + 1)^3}{z^2(z - 1)^2}.$$

More generally, any S_3 -invariant rational function of degree d can be expressed as a ratio of modular forms of degree d .

9. If desired, we can regard F_2 and F_3 as forms on $\widehat{\mathbb{C}}/S_3$ by substituting $w = J(z)$. They then come:

$$F_2 = \frac{dw^2}{4w(w - 1)} \quad \text{and} \quad F_3 = \frac{dw^3}{8w^2(w - 1)}.$$

We also find $D_6 = (1/432)(w^{-4}(w - 1)^{-3}) dw^6$.

Cusp forms. We let $S_k \subset M_k$ denote the space of *cusp forms* with poles of order $\leq k - 2$ at $0, 1, \infty$.

Theorem 5.35 *The map $\omega \mapsto D_6\omega$ gives an isomorphism from M_k to S_{k+6} .*

Proof. Any $F \in M_k$ has poles of order $\leq k$, so FD_6 has poles of order $\leq k + 4$ and hence lies in S_{k+6} . Conversely, if $G \in S_{k+6}$ then $F = G/D_6$ has poles of order at most k at $0, 1, \infty$, and it is otherwise holomorphic since D_6 has no zeros. ■

Every space M_k , $k \geq 2$ contains a form of the type $F_2^i F_3^j$ which is *not* a cusp form; thus $M_k \cong \mathbb{C} \oplus S_k$. By inspection, $\dim S_k = 0$ for $k \leq 5$. On the other hand, any cusp form is divisible by D_6 . This shows:

Corollary 5.36 *We have $\dim M_{6n+1} = n$, and $\dim M_{6n+i} = n + 1$ for $i = 0, 2, 3, 4, 5$.*

Corollary 5.37 *The forms F_2 and F_3 generate the ring $M = \bigoplus M_k$.*

Corollary 5.38 *The forms $F_2^i F_3^j$ with $2i + 3j = k$ form a basis for M_k .*

Proof. By the preceding Corollary these forms span M_k , and the number of them agrees with $\dim M_k$ as computed above. ■

Corollary 5.39 *The ring of modular forms M is isomorphic to the polynomial ring $\mathbb{C}[F_2, F_3]$.*

Proof. The preceding results shows the natural map from the graded ring $\mathbb{C}[F_2, F_3]$ to M is bijective on each graded piece. ■

We have already seen that, under the isomorphism $\mathcal{M}_1 \cong \mathcal{M}_{0,4}$, an analytic modular form is the same as an algebraic modular form. Thus we also have:

Corollary 5.40 *The ring of modular forms is isomorphic to $\mathbb{C}[g_2, g_3]$.*

Corollary 5.41 *For every $k \geq 2$, the quantity $\sum' \lambda^{-2k}$ can be expressed as a polynomial in $\sum' \lambda^{-4}$ and $\sum' \lambda^{-6}$.*

Values of g_2 and g_3 . We now note that for $\tau = i$, the zeros of $4x^3 - g_2x - g_3$ must look like $(-1, 0, 1)$ and thus $g_3(i) = 0$. Similarly for $\tau = \rho$ the zeros are arrayed like the cube roots of unity and hence $g_3(\rho) = 0$.

To determine the values at infinity, we observe that as $\tau \rightarrow \infty$ we have

$$G_2(\tau) \rightarrow 2\zeta(4) = \pi^4/45$$

and

$$G_3(\tau) \rightarrow 2\zeta(6) = 4\pi^6/945.$$

This gives the values $g_2(\infty) = 60G_2(\infty) = (4/3)\pi^4$ and $g_3(\infty) = 140G_3(\infty) = (8/27)\pi^6$, and thus

$$(g_2^3/g_3^2)(i\infty) = 27.$$

The cusp form Δ and the modular function J . By the preceding calculation, the *discriminant*

$$\Delta(\tau) = g_2^3(\tau) - 27g_3(\tau)^2$$

is a cusp form of weight 12. We have seen such a form is unique up to a scalar multiple, and is nonvanishing everywhere except for a simple zero at infinity. This form has a natural meaning: up to a multiple, it is the discriminant of the cubic polynomial $4x^3 - g_2x - g_3$. As we have seen, this polynomial has distinct roots whenever g_2 and g_3 come from a lattice; this explains why $\Delta(\tau)$ has no zeros in \mathbb{H} .

Theorem 5.42 *We have $J(\tau) = g_2^3(\tau)/\Delta(\tau)$.*

Proof. First note that this is a ratio of forms of weight 12 and hence a modular function, i.e. it is invariant under $\text{SL}_2(\mathbb{Z})$. Since $g_2(\infty) \neq 0$, it has a simple pole at infinity, and thus has degree one on \mathcal{M}_1 . We also have $J(\rho) = 0$ since $g_2(\rho) = 0$, and $J(i) = 1$ since $g_3(i) = 0$. ■

Remarks. One can use residues to determine the exact relationship between F_2 and g_2 . Namely, $F_2 \sim (1/4)dw^2/w^2$, while for $q = \exp(2\pi i\tau)$ we have $dq = 2\pi iq d\tau$, and hence

$$g_2 = 60G_2(\tau) d\tau^2 \sim \frac{4\pi^4}{3} \frac{dq^2}{(2\pi iq)^2} = -\frac{\pi^2}{3} \frac{dq^2}{q^2}.$$

Thus $g_2 = -(4\pi^2/3)F_2$. One can similarly calculate g_3/F_3 .

If we let $j(\tau) = 1728J(\tau)$, then

$$j(\tau) = \frac{1}{q} + 744 + \sum_1^{\infty} a_n q^n$$

with $a_n \in \mathbb{Z}$.

Connections with additive and multiplicative number theory. Incredibly, we have

$$\Delta(q) = (2\pi i)^{12} q \prod_1^{\infty} (1 - q^n)^{24}.$$

This gives a close connection between the theory of modular forms and the partition function $p(n)$, since

$$\prod (1 - q^n)^{-1} = \sum p(n) q^n.$$

The coefficients power series for $G_k(\tau)$ involves the function $\sigma_k(n) = \sum_{d|n} d^k$.

The Riemann function $\zeta(s)$ arises as the Mellin transform of a theta function, and then modularity translates into the functional equation.

References

- [BD] F. Berteloot and J. Duval. Sur l'hyperbolicité de certain complémentaires. *Enseign. Math.* **47**(2001), 253–267.
- [Bol] B. Bollabas, editor. *Littlewood's Miscellany*. Cambridge University Press, 1986.
- [CJ] L. Carleson and P. Jones. On coefficient problems for univalent functions. *Duke Math. J.* **66**(1992), 169–206.
- [Gam] T. W. Gamelin. *Uniform Algebras*. Prentice–Hall, 1969.
- [Gol] G. M. Goluzin. *Geometric Theory of Functions of a Complex Variable*. Amer. Math. Soc, 1969.
- [GM] J. D. Gray and S. A. Morris. When is a function that satisfies the Cauchy-Riemann equations analytic? *The American Mathematical Monthly* **85**(1978), 246–256.

- [Ko] S. V. Kolesnikov. On the sets of nonexistence of radial limits of bounded analytic functions. *Russian Acad. Sci. Sb. Math.* **81**(1995), 477–485.
- [MH] J. E. Marsden and M. J. Hoffman. *Basic Complex Analysis*. W. H. Freeman, 1999.
- [Mil] J. Milnor. On Lattès maps. In P. G. Hjorth and C. L. Petersen, editors, *Dynamics on the Riemann sphere*, pages 9–44. European Math. Soc., 2006.
- [Mon] M. G. Monzingo. Why are 8:18 and 10:09 such pleasant times? *Fibonacci. Quart.* **2**(1983), 107–110.
- [Re] R. Remmert. *Classical Topics in Complex Function Theory*. Springer, 1998.
- [Sa] R. Salem. *Algebraic Numbers and Fourier Analysis*. Wadsworth, 1983.