Complex Variable Outline

Richard Koch

January 10, 2022

Contents

List of Figures

Preface

This is a set of lecture notes on complex variable theory. There are no exercises. Please consult other books.

Some authors prefer to write mathematics in a logical order in which results cannot be mentioned until the tools are available to prove them. An advantage of this style is that the readers always know where they stand. But there are disadvantages, best illustrated with an example. All treatments of complex analysis extend $e^x, \sin x, \cos x$, and $\ln x$ to functions of a complex variable $e^z, \sin z, \cos z$, and $\log(z)$. This happens early in the text but the extensions seem arbitrary, made perhaps for historical reasons. Later on it is proved that a differentiable function on R has at most one extension to a differentiable function on C. This is the crucial fact, showing that the complex extensions had been present all along, waiting to be discovered. In these notes, that later result is mentioned immediately, making these complex extensions more exciting.

The extension of calculus to complex numbers is due to Euler. His techniques often did not meet modern standards of rigor, but his results were essentially always correct. The number of crucial discoveries due to Euler is astonishing, and in celebration of his work we describe an example immediately. If $P(z)$ is a polynomial, it can be completely factored over the complex numbers in terms of its roots. Euler applied this to $\sin(z)$, which vanishes at πn, writing

$$\sin(z) = z \prod_{n=1}^{\infty} \left(1 - \frac{z}{\pi n}\right)\left(1 + \frac{z}{\pi n}\right) = z \prod_{n=1}^{\infty} \left(1 - \frac{z^2}{\pi^2 n^2}\right)$$

This generalizes Wallis' product. Who would have believed that Wallis' product is just the fundamental theorem of algebra in disguise? Moreover, Euler used the result to deduce the value of $\sum_{n=1}^{\infty} \frac{1}{n^2}$.

The deeper part of complex analysis depends on integration theory and is due to Cauchy and others. This theory: Cauchy's theorem, the Residue Theorem, the expansion of holomorphic functions as power series, turns out to be a special case of several variable calculus. Many books ignore this fact and develop the theory from scratch, perhaps because the topological portions of several variable calculus, like Green's theorem, aren't proved rigorously. In these lectures,

we develop the deeper theory twice. The first treatment starts with a review of several variable calculus and rapidly obtains the central theorems of complex analysis as a consequence. This is followed by a rigorous treatment using ideas of Emil Artin and Henri Cartan, where modern ideas from topology allow us to integrate over arbitrary continuous paths and prove that integrals over homotopic paths are the same.

The second half of these lectures covers mathematical consequences of complex analysis. In a visit to Oregon, Phillip Griffiths once observed that the nineteenth century centered on complex analysis and the twentieth century centered on algebraic topology; in both cases, almost every mathematician related their work to the central idea of the century in one way or another. Complex analysis contributed to number theory, to topology, to algebraic geometry, to non-Euclidean geometry, to modern algebra in the nineteenth century, just as cohomological ideas have appeared in number theory, algebraic geometry, representation theory, and analysis in the twentieth century. I wanted to push the theory far enough that these connections were clear, rather than just hinting that they were present.

So we discuss the Riemann Zeta Function and actually prove the prime number theory. We discuss Dirichlet series and prove Dirichlet's theorem on primes in an arithmetic progression. We prove the Riemann Mapping Theorem. We discuss elliptic functions, theta functions, and modular functions, and prove Jacobi's formulas for sums of two squares and sums of four squares. We explain how these ideas generalize to arbitrary Riemann surfaces, and then to algebraic curves and algebraic number fields.

Chapter 1

Preliminaries and Examples

1.1 Review of Complex Numbers

A *complex number* is a point (a, b) in the plane R^2. Instead of writing (a, b), we write $a+ib$ and add and multiply as usual, assuming that $i^2 = -1$. If we want to speak of a number in general, we write z instead of $a + ib$.

If $z = (a, b)$, the usual norm $||(a, b)|| = \sqrt{a^2 + b^2}$ is denoted $|z|$ and called the absolute value of z. This absolute value has the important properties $|w+z| \leq |w|+|z|$ and $|wz| = |w|\,|z|$.

The *conjugate* of $z = a + ib$ is $\overline{z} = a - ib$. Notice that $\overline{z}z = |z|^2$.

Nonzero numbers have multiplicative inverses given by

$$\frac{1}{z} = \frac{\overline{z}}{\overline{z}z} = \frac{\overline{z}}{|z|^2}$$

so that dividing by z is the same thing as multiplying by $\overline{z}$ up to a real factor. In particular, the set of complex numbers forms an extension field $R \subset C$. It is easy to see that there are just two automorphisms of C fixing R, the identity map and conjugation.

In polar coordinates we have $(a, b) = (r\cos\theta, r\sin\theta)$ where r is the distance of the point from the origin and θ is the angle of the point from the x-axis. Since $r = |z|$, with complex notation this formula becomes

$$z = |z|\,(\cos\theta + i\sin\theta)$$

No doubt you have seen the formula $e^{i\theta} = \cos\theta + i\sin\theta$. We'll soon deduce it again. Thus the above formula becomes

$$z = |z|e^{i\theta}$$

A consequence of the previous formula is that $|e^{i\theta}| = 1$, a fact we'll use again and again.

If $z_1 = |z_1|e^{i\theta_1}$ and $z_2 = |z_2|e^{i\theta_2}$, we have

$$z_1 z_2 = |z_1 z_2| e^{i\theta_1} e^{i\theta_2} = |z_1 z_2| e^{i(\theta_1+\theta_2)}$$

So multiplying complex numbers is done by multiplying their distances from the origin and adding their angles with the x-axis.

It is a remarkable fact that complex numbers unify the theories of exponential and trigonometric functions; we'll see more of this soon. The formula $e^{i\theta_1}e^{i\theta_2} = e^{i(\theta_1+\theta_2)}$ is a special case of the power law $e^w e^z = e^{w+z}$, but changing the notation slightly makes it a special case of the addition formulas in trigonometry:

$$(\cos\theta_1 + i\sin\theta_1)(\cos\theta_2 + i\sin\theta_2) =$$

$$= (\cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2) + i(\sin\theta_1\cos\theta_2 + \cos\theta_1\sin\theta_2) =$$

$$\cos(\theta_1+\theta_2) + i\sin(\theta_1+\theta_2)$$

1.2 Holomorphic Functions

Complex variable theory is about complex valued functions of a complex variable. Sometimes we denote this $w = f(z)$. There are many examples, like $f(z) = \frac{z^3+iz+3}{z^4+1}$.

Notice that the standard picture of a mountainous surface over the plane only makes sense if the values of f are real. Otherwise, we need four dimensions for the picture. We will soon see that real-valued f are not differentiable in the complex sense and thus not interesting.

On the other hand, old books often contain pictures of the absolute value $|f(z)|$ for famous functions. The next page shows three examples. The first is a plot of the absolute value of $\frac{1}{z(z-1)}$. This function has singularities, known as *poles*, at $z = 0, 1$.

The second is a plot of the gamma function, extended to be defined for all complex numbers. In calculus, $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,dt$ provided $x > 0$. Using integration by parts, we have $\Gamma(x+1) = x\Gamma(x)$; also $\Gamma(1) = 1$. It follows that $\Gamma(n) = (n-1)!$ for all integers n. If we replace x with a complex z, the expression $t^{z-1}e^{-t}$ still makes sense, as we will discover shortly. The integral still converges if $\Re(z) > 0$ and defines a function $\Gamma(z)$. The extension still satisfies the formula $\Gamma(z+1) = z\Gamma(z)$ for all such complex z.

But then we can extend $\Gamma(z)$ to be defined for all z as follows. By algebra, $\Gamma(z) = \frac{\Gamma(z+1)}{z} = \frac{\Gamma(z+2)}{z(z+1)} = ...$ The second equation is defined for $\Re(z+1) > 0$ or $\Re(z) > -1$, the third for $\Re(z) > -2$, etc., and in this way we can extend as far left as we like. Notice that poles then appear at $0, -1, -2, ...$, as shown in the graph.

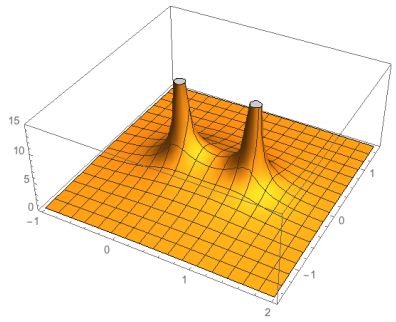


Figure 1.1: $f(z) = \frac{1}{z(z-1)}$

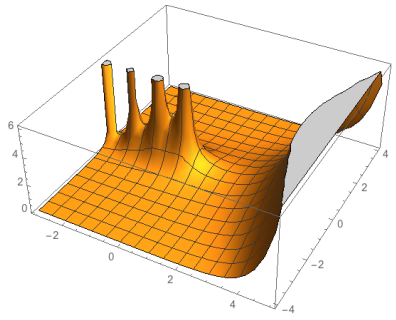


Figure 1.2: $f(z) = \Gamma(z)$

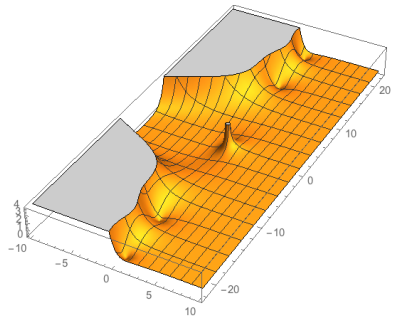


Figure 1.3: $f(z) = \zeta(z)$

The final picture shows $\zeta(z) = \sum_1^\infty \frac{1}{n^z}$, which converges for $\Re(z) > 1$. This is known as the Riemann Zeta function because Riemann proved that it can be extended to be defined everywhere except $z = 1$, by an argument we will give later. Since $\sum_1^\infty \frac{1}{n} = \infty$, the function has a pole as $z = 1$. Riemann proved that this function has infinitely many zeros z_0 satisfying $0 < \Re(z_0) < 1$ and conjectured that these zeros all have real part $\frac{1}{2}$. Four of these zeros as shown in the graph. This conjecture is the most famous (and probably most important) unsolved problem in mathematics.

We define the derivative of a complex valued function by

$$\frac{df}{dz} = \lim_{h\to 0} \frac{f(z+h) - f(z)}{h}$$

Here h is allowed to be complex, so $z + h$ can approach z from all possible directions in the plane; we require that all of these limits be equal. This is a powerful requirement, with beautiful consequences not true in the real version of the theory.

To compute the derivative of f at z, we need to require that f is defined sufficiently near to z in all directions. Thus the domain of f must be an open set $\mathcal{U} \subset C$. We often need to require that $\mathcal{U}$ be connected; for instance, if $\frac{df}{dz}$ is identically zero, we'd like to conclude that f is constant, and that is only true if $\mathcal{U}$ is connected. To save words, we define a *domain* to be a connected open set in C.

Complex variable theory, then, is about a special class of functions called *holomorphic functions*. There are four possible definitions, corresponding to four approaches to the elementary theory. In the end, the four definitions turn out to be equivalent, and after that is proved, the deeper theory opens up in many directions. In these notes, the official definition is the first one.

Definition 1. *A complex-valued function f is* holomorphic *on a domain $\mathcal{U}$ if $\frac{df}{dz}$ exists at each point of $\mathcal{U}$.*

Definition 2. *A complex-valued function f is* holomorphic *on a domain $\mathcal{U}$ if $\frac{df}{dz}$ exists at each point of $\mathcal{U}$, and the resulting function $\frac{df}{dz}$ is continuous on $\mathcal{U}$.*

Definition 3. *A complex-valued function f is* holomorphic *on a domain $\mathcal{U}$ if $\frac{d^k f}{dz^k}$ exists at each point of $\mathcal{U}$ for $k = 1, 2, 3, ...$*

Definition 4. *A complex-valued function f is* holomorphic *on a domain $\mathcal{U}$ if f is given near each $z_0 \in \mathcal{U}$ by a convergent power series*

$$f(z) = c_0 + c_1(z - z_0) + c_2(z - z_0)^2 + ...$$

The first of these definitions is the least restrictive, but special techniques are required at one or two spots because we do not yet know that $\frac{df}{dz}$ is continuous. These notes will follow this approach.

If we assume that $\frac{df}{dz}$ is continuous, then the elementary theory is actually a consequence of several variable calculus and can be developed in an hour or two. We'll also sketch this approach.

The fourth definition is the most restrictive. It is easy to prove that a function given by a convergent power series is differentiable with derivative given by the term-by-term derivative of the power series, so the fourth definition implies all the others. In real variable theory, it is easy to find functions which can be differentiated once, but not twice, and it is also easy to find functions which can be differentiated infinitely often and yet do not have convergent power series. So the fourth definition is only equivalent to the first in complex variable theory.

Weierstrass developed complex variable theory by starting with power series. This is possible, but requires tricky proofs that are avoidable by starting with the first or second definition. For example, if $f(z)$ and $g(z)$ are given by power series, then so is $f+g$, but it is not clear that fg and f/g are given by power series. On the other hand, if f and g have complex derivatives, so do fg and f/g by the ordinary calculus argument, unmodified.

Consider the series

$$1 + z + z^2 + z^3 + ... = \frac{1}{z-1}$$

The series only converges for $|z| < 1$, but the function is defined and differentiable everywhere except $z = 1$. That's why our fourth definition has a complicated condition requiring power series centered at each point of the domain. But then we should really prove a theorem stating that if a power series converges for $|z| < R$, then the expansion about each z_0 in this disk converges to f at least near z_0. This is true, but a direct proof is tricky. The result will become easy by following our approach.

On the other hand, power series are the most natural way to get complex analogues of e^x, $\sin x$, $\cos x$, and $\ln x$, so we state a few elementary facts about them beginning in section 1.4, and then in chapter two. But first

1.3 Differentiation Algorithms

The standard rules for differentiating sums, products, and quotients extend easily to the complex case; we leave these proofs to the reader. The chain rule also extends easily. Therefore, the standard calculus differentiation exercises apply in the complex case. Said another way, you already know how to differentiate.

In the following sections we will define e^z, a^z, $\sin(x)$, $\cos(z)$, and $\log(z)$, but each time we will prove that the derivative is the expected one. So there is nothing to learn even if you find yourself differentiating fancy expressions involving transcendental functions.

On the other hand, the standard pictures do not apply, so it is unclear what $\frac{df}{dz}$ is actually measuring. That will be explained in chapter 3.

1.4 Extending Functions from R to C

We are about to extend e^x from a function defined on the reals to a function defined on the complex numbers. It is reasonable to expect that this is possible in many ways, and that our extensions are chosen for "historical reasons only." Amazingly, this is false. It turns out that there is at most one way to extend a function defined on a subset of R to a holomorphic extension over the complex numbers.

Most standard functions of calculus satisfy identities of various kinds. For example, $\sin^2 x + \cos^2 x = 1$. Amazingly, these identities remain true over the complex numbers, and even more amazingly, there is an abstract theorem which proves this without bothering to check any special case.

Here are the two theorems in question, known as "the identity theorems."

Theorem 1. *Suppose $\mathcal{U}$ is a domain such that $\mathcal{U} \cap \mathcal{R} \neq \emptyset$. If f is a complex valued function defined on $\mathcal{U} \cap \mathcal{R}$, then f can be extended to a holomorphic function on $\mathcal{U}$ in at most one way.*

Theorem 2. *Suppose $\mathcal{U}$ is a domain such that $\mathcal{U} \cap \mathcal{R} \neq \emptyset$. Suppose $f(z)$ and $g(z)$ are holomorphic on $\mathcal{U}$ and $f(x)$ and $g(x)$ satisfy an algebraic identity on $\mathcal{U} \cap \mathcal{R}$. Then $f(z)$ and $g(z)$ satisfy the same identity on all of $\mathcal{U}$.*

Remark: We have to postpone the proofs until later, but a key idea can be given here. We'll sketch the proof of the first theorem from the power series definition of holomorphic functions. Pick a point z_0 in $\mathcal{U} \cap \mathcal{R}$ and expand $f - g$ about z_0. The coefficients of this power series are $\frac{f^k(z_0) - g^k(z_0)}{k!}$. Since $f - g$ is identically zero on $\mathcal{U} \cap \mathcal{R}$, all of these coefficients vanish and the power series is identically zero. So $f - g$ is zero in an open disk about z_0.

Consider the set $\mathcal{V} = \{ z_0 \in \mathcal{U} \mid f - g = 0 \text{ in an open disk about } z_0 \}$. This set is trivially open; by the previous paragraph it is not empty. We prove it closed and consequently all of $\mathcal{U}$ by connectedness of $\mathcal{U}$. Suppose $z_n \in \mathcal{V}$ and $z_n \to z_0 \in \mathcal{U}$. Then $f^k(z_n) - g^k(z_n) = 0$. But $f - g$ can be expanded in a power series near z_0, so all derivatives of $f - g$ exist and are continuous near z_0. It follows that all of these derivatives vanish at z_0, so the coefficients of the series expansion at z_0 are zero and $f - g$ is identically zero near z_0. Thus $\mathcal{V}$ is closed.

1.5 Exponential Functions

We need some fundamental theorems about power series. These will be proved in chapter two, but we state them now because we want to extend known functions on R to holomorphic functions on C using them.

Theorem 3. *Consider an arbitrary power series about the origin:*

$$c_0 + c_1 z + c_2 z^2 + ...$$

There is a constant R, $0 \leq R \leq \infty$, such that the series converges absolutely for $|z| < R$ and diverges for $|z| > R$. This R is called the radius of convergence.

Exercise: Find an example where $R = 0$ and the series converges only at $z = 0$. Find an example where $R = \infty$. Find an example where $R = 1$ and the series diverges at all z with $|z| = 1$. Also find a similar example where the series converges at all z with $|z| = 1$. Finally find a similar example where the series sometimes converges and sometimes diverges at points with absolute value one.

Theorem 4. *If the radius of convergence is R, the series converges absolutely for $|z| < R$, so it can be rearranged at will. Moreover, if $R_0 < R$, the series converges uniformly for $|z| \leq R_0$.*

Theorem 5. *Let the radius of convergence of the following series be R:*

$$f(z) = c_0 + c_1 z + c_2 z^2 + c_3 z^3 + ...$$

Then the radius of convergence of the formally differentiated series is also R:

$$c_1 + 2c_2 z + 3c_3 z^2 + ...$$

Moreover, the function f has a complex derivative at each point z with $|z| < R$, and this derivative is given by the formally differentiated series. Therefore, f has complex derivatives of all orders in the convergence disk, and each is continuous.

Remark: We now use this to extend e^x to e^z:

Definition 5. *The exponential function is the holomorphic function*

$$e^z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + ...$$

Remark: Every calculus course proves that this series converges for all real z. Hence the radius of convergence must be $R = \infty$, so the series always converges.

Theorem 6. *The derivative of e^z is e^z.*

Proof: Power series can be differentiated term by term, so differentiate the series and notice that this reproduces the same series.

Theorem 7.

$$e^{z+w} = e^z e^w$$

Remark: This follows from the identity theorem. The formula $e^{z+w} = e^z e^w$ is known to hold for real z and w. Fix a real w and consider both sides as functions of z. Since the two functions agree on the real numbers, they agree for all z. Now fix a complex z and think of both sides as functions of w. We just proved that they agree for real w, so they must always agree.

Remark: We can also give a direct proof, but it is messier. First we proceed formally.

$$e^z e^w = \sum \frac{z^k}{k!} \sum \frac{w^l}{l!} = \sum_{n=0}^{\infty} \sum_{k+l=n} \frac{z^k w^l}{k!l!} = \sum_{n=0}^{\infty} \frac{1}{n!} \left(\sum_{k+l=n} \frac{n!}{k!l!} z^k w^l \right) = \sum_{0}^{\infty} \frac{(z+w)^n}{n!} = e^{z+w}$$

Thus the equation $e^z e^w = e^{z+w}$ is really all possible binomial theorems wrapped up together.

Exercise 1. *Make this rigorous by recalling that inside the convergence disk, power series converge absolutely and absolutely convergent series can be rearranged.*

Corollary 1. *The function e^z has no zeroes.*

Proof: $e^z e^{-z} = 1$.

Remark: Euler liked to think of power series as "generalized polynomials". If $P(z)$ is a non-constant polynomial, it takes all possible complex values by the fundamental theorem of algebra.

We might try to generalize this result to say that a non-constant holomorphic function defined in the entire complex plane takes all possible complex values, but e^z shows that this is false. However, we will later prove that the values of such a function are dense in the plane.

Around 1861, after everyone thought that basic complex variable theory was complete, Picard surprised everyone by proving that a non-constant holomorphic function on the entire complex plane can omit at most one value.

1.6 Trigonometric Functions

Let x be real and notice that

$$e^{ix} = 1 + ix + \frac{i^2 x^2}{2!} + \frac{i^3 x^3}{3!} + \ldots =$$

$$\left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} + \ldots\right) + i\left(x - \frac{x^3}{3!} + \frac{x^5}{5!} + \ldots\right) =$$

$$\cos x + i \sin x$$

Changing x to $-x$ and adding and subtracting gives

$$\frac{e^{ix} + e^{-ix}}{2} = \cos x$$

$$\frac{e^{ix} - e^{-ix}}{2i} = \sin x$$

The identity theorem says that we must then define

Definition 6.

$$\cos z = \frac{e^{iz} + e^{-iz}}{2} = 1 - \frac{z^2}{2!} + \frac{z^4}{4!} - ...$$

$$\sin z = \frac{e^{iz} - e^{-iz}}{2i} = z - \frac{z^3}{3!} + \frac{z^5}{5!} - ...$$

Theorem 8. *The functions* $\cos z$ *and* $\sin z$ *are holomorphic in the entire complex plane and satisfy the following equations:*

- $\frac{d}{dz} \cos z = -\sin z$; $\frac{d}{dz} \sin z = \cos z$
- $\cos^2 z + \sin^2 z = 1$
- $\cos(z + w) = \cos z \cos w - \sin z \sin w$
- $\sin(z + w) = \sin z \cos w + \cos z \sin w$

1.7 e^z as a Map from C to C

The correct way to think of $f(z)$ geometrically is as a map from a subset of the plane to another subset of the plane. To picture this more accurately, draw a grid (not necessarily rectangular) on the domain and draw the corresponding mapped grid lines on the image. In the next chapter, we will see discover the geometric restrictions that holomorphy places on these transformations.

We want to draw such a picture for e^z. Notice that

$$e^{x+iy} = e^x(\cos y + i \sin y)$$

Consequently, horizontal lines with varying x and fixed y map to radial lines with varying distance to the origin and fixed angle; vertical lines with fixed x and varying y map to circles with fixed radius and varying angle. The resulting picture is the following:

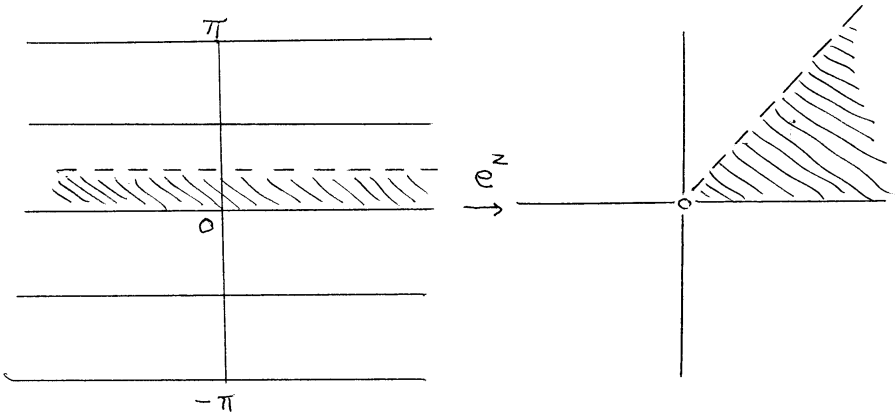


Figure 1.4: e^z as a Mapping $C \to C$

The shaded region on the left maps to the pie shaped shared region on the right; points far to the left map close to the origin, and points far to the right map to points near infinity. The region on the left covers the entire complex plane except the origin, and the map then repeats periodically with period $2\pi i$.

1.8 log(z)

Recall that a complex z can be written in polar coordinates as $z = |z|(\cos\theta + i\sin\theta)$. We call this angle the *argument* of z, $\arg(z)$. Rewriting slightly, $z = |z|e^{i\arg z}$.

We define the complex logarithm, $\log(z)$, to be is inverse to e^z. So

$$e^{\log(z)} = z = e^{\ln|z| + i\arg(z)}$$

We conclude that by definition,

$$\log(z) = ln|z| + i\arg(z)$$

The only problem is that $\arg(z)$ is multiple-valued; we can add any integer multiple of 2π and get another legal value for $\arg(z)$. Consequently, $log(z)$ is multiple valued, and its possible values differ by integer multiples of $2\pi i$. Notice that $\log(z)$ is defined for every complex number except $z = 0$.

The main theorems about holomorphic functions require single-valued functions. So whenever the logarithm occurs, we need to pick a *branch* of the logarithm. That is, we need to

define a continuous single-valued function $\theta(z)$ on the domain $\mathcal{U}$ whose value is one of the legal values of the angle at each point, and define

$$\log(z) = \ln|z| + i\theta(z)$$

for that particular choice of θ.

Often in complex variable theory, we restrict the log to the domain $\mathcal{U}$ formed by removing all non-positive real numbers from the plane, and require that $-\pi < \arg(z) < \pi$. Then $log(z)$ is single-valued on this domain. In a chapter to come, we will prove that $log(z)$ is holomorphic on this domain; the reader can rather easily prove that already. This particular restriction on domain defines what is called *the principal branch of the logarithm*.

Many other domains $\mathcal{U}$ in the plane omit the origin and allow us to define a continuous value of $\arg(z)$ throughout the domain. We can then define a branch of the logarithm on $\mathcal{U}$ by $\log(z) = \ln|z| + i\theta$ and all such functions turn out to be holomorphic. Each such choice is called a *permissible branch of the logarithm*. We will later prove that such a branch exists on $\mathcal{U}$ if and only if $\mathcal{U}$ is simply connected. But often the fancy theory is not needed and it is quite clear how to define $\arg(z)$. The next page shows examples of simply connected domains of the complex plane which omit the origin. Describe one possible choice of $\arg(z)s$ for each such domain:

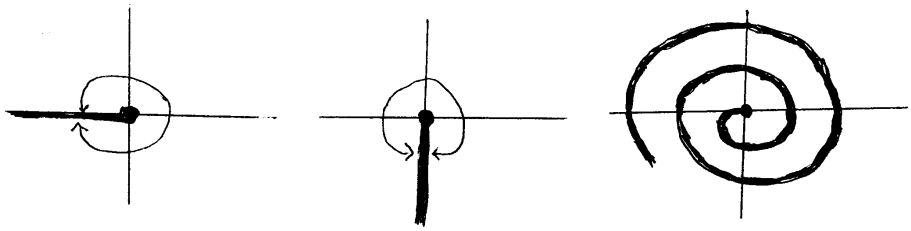


Figure 1.5: Branches of Log(z)

1.9 Complex Exponentials

As in ordinary calculus, we define $a^b = e^{b \ln a}$. Similarly if a and b are complex, we define $a^b = e^{\mathrm{b\ log(a)}}$. This is multiple-valued, and we make it single-valued by choosing an appropriate branch of the logarithm for a. Common sense is often enough to make this choice. We give three examples.

Example 1:

The Riemann Zeta function is defined as $\sum_{n=1}^{\infty} \frac{1}{n^z}$ for $\Re z > 1$. In this case, n is a positive integer, so clearly $n^z = e^{\ln n\ z}$ and we use the ordinary calculus logarithm.

Example 2:

The Gamma function is defined as $\int_0^\infty t^{x-1}e^{-t}\,dt$ for $\Re z > 0$. Again t is a positive real number, so $t^{z-1} = e^{\ln t\ (z-1)}$ and we use the ordinary calculus logarithm.

The rest of this example can be skipped, but shows how we can make use of the multiple values of the logarithm to do interesting things.

We require $\Re z > 0$ in the above formula because otherwise the singularity at the origin is not integrable. On the other hand, e^{-t} vanishes so rapidly at infinity that the integral converges there for all z. For this reason, it is interesting to compute

$$I(z) = \int_\gamma t^{z-1}e^{-t}dt$$

where γ is the curve shown at the top of the next page. Assume the two portions of the path on the right side are actually on the positive x-axis. Of course this uses the integration theory from a later chapter, but let us be courageous and proceed.

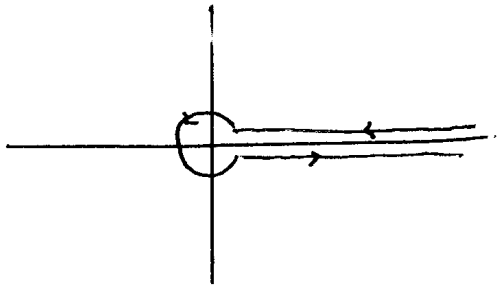


Figure 1.6: γ

The term $t^{z-1} = e^{log(t)\,(z-1)}$ and we assume $\log(t) = \ln t$ as we integrate along the first portion of the path.

Since the limits of integration are at infinity where the integrand goes to zero, this integral exists for all values of z. As we later prove, this and similar integrals define holomorphic functions of z. So $I(z)$ is globally defined.

If $\Re z > 0$, the integrand is small near the origin and we can ignore the integral around the circle there. But the imaginary part of $\log(z)$ will increase from 0 to 2π around this circle. Then as we integrate the last portion in the positive direction along the x-axis, we will be integrating $e^{(\ln t + 2\pi i)(z-1)}\,dt$. The real integrals along the positive x-axis define $\Gamma(z)$, so we conclude that for $\Re z > 0$,

$$I(z) = \left(-1 + e^{2\pi i(z-1)}\right)\Gamma(z)$$

or

$$I(z) = 2ie^{\pi iz}\left(\frac{e^{\pi iz} - e^{-\pi iz}}{2i}\right)\Gamma(z) = 2ie^{\pi iz}\sin(\pi z)\Gamma(z)$$

Thus for $\Re z > 0$, we have

$$\Gamma(z) = \frac{1}{2i}e^{-\pi iz}\frac{1}{\sin(\pi z)}I(z)$$

On the other hand, $I(z)$ is holomorphic for all z, so this formula gives a second holomorphic continuation of the Gamma function to the entire complex plane. We already know this by an easier argument from the identity $\Gamma(z+1) = z\Gamma(z)$.

It turns out that a similar argument analytically continues the zeta function to the entire complex plane, and in that case the second argument is not available to us.

Since $\sin(\pi z)$ vanishes at integers, the Gamma function has poles at the non-positive integers. There are no positive poles because $I(z) = 0$ at positive integers.

Example 3:

The remaining examples are more complicated and can be skipped.

Notice that

$$\sqrt{z} = e^{1/2\ log(z)}$$

Although the logarithm has infinitely many values, it is easy to check that $\sqrt{z}$ defined this way has only two values, which differ only in sign.

If we use the principal branch of the logarithm, we discover that $\sqrt{z}$ is holomorphic, defined everywhere except the negative x-axis. If $\sqrt{x} > 0$ for $x > 0$, then this branch of $\sqrt{z}$ has values which are positive imaginary numbers just above the negative x-axis and negative imaginary numbers just below the negative x-axis.

Another way to think of $\sqrt{z}$ is as a double-valued function which goes through all of its values as we wrap around the origin *twice*. In this case, it is convenient to put the branch line along the *positive* x-axis. If we start at a positive x and wind around the origin counterclockwise once, the values of $\sqrt{z}$ move from $\sqrt{x}$ to $-\sqrt{x}$. As we wind around again, the values move from $-\sqrt{x}$ back to $\sqrt{x}$.

Example 4: The length of an ellipse is given by an integral whose integrand equals the square root of a cubic or quartic, depending on the choice of coordinates. If we want to investigate this integral over the complex numbers, we need to study expressions of the form $\sqrt{z^3 + az^2 + bz + c}$. To be concrete, consider the special case

$$\sqrt{z^3 - z} = \sqrt{(z-1)z(z+1)}$$

The square root is not differentiable at the origin, so we must remove $-1, 0$, and 1 from the plane. Earlier, we asserted that $\log(z)$ has a single-valued branch on any simply connected domain. There is actually a more astonishing theorem: if $f(z)$ is holomorphic and nonzero on a simply connected domain $\mathcal{U}$, then $\log(f(z))$ has a single-valued branch on $\mathcal{U}$. We will prove this later. It is easy to find examples where $\mathcal{U}$ is simply connected, but the range $f(\mathcal{U})$ is not simply connected, so we cannot start by finding a branch of the logarithm on the range of f.

At any rate, let us make a branch cut from $-\infty$ to 1 along the x-axis. The remaining set is simply-connected, and thus we can define $\log(z-1)z(z+1)$ on this set. Then $\sqrt{(z-1)z(z+1)} = e^{\frac{1}{2}log(z-1)z(z+1)}$ is a branch on $\mathcal{U}$. See the left side of the Figure below.

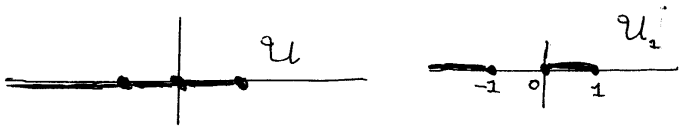


Figure 1.7: $\sqrt{(z-1)z(z+1)}$

Notice that $\sqrt{(z-1)z(z+1)} = \sqrt{z-1}\cdot\sqrt{z}\cdot\sqrt{z+1}$. If we traverse a very small circle centered at 1, z and $z+1$ hardly change and $\sqrt{z}\cdot\sqrt{z+1}$ return to their original value. But $\sqrt{z-1}$ changes sign, and thus $\sqrt{(z-1)z(z+1)}$ changes sign.

A similar argument applies if we traverse a small circle about 0. But z already changes sign on the right side, so it *does not* change sign on the left side. Consequently the entire square root does not change sign on $[-1,0]$ and we can remove that portion of the branch cut. We conclude that our branch of $\sqrt{(z-1)z(z+1)}$ is defined on the entire plane except the intervals $(-\infty,-1]\cup[0,1]$. The function consists of two copies with this domain and differing only in sign. If we cross one of the branch cuts $(-\infty,-1]\cup[0,1]$, we must jump to the other copy of our function, i.e., change the sign of the integrand.

Chapter 2

Series

This short chapter should be skimmed. Its purpose is to show that standard results in infinite series and power series work for complex numbers as well as real numbers. No new exciting ideas appear. The one advantage is that most proofs can be left to the reader, so the central ideas perhaps stand out more clearly than when first taught.

No difficult proofs are skipped. The missing proofs could be supplied, but they would just clutter up the text.

The last section on power series is important for these lectures; read it carefully.

2.1 Infinite Series

Definition 7. *If $\{c_k\}$ is a sequence of complex numbers, we write $L = \sum_{k=0}^{\infty} c_k$ and say the sum* converges *if for each $\epsilon > 0$ there is an integer N such that whenever $n > N$ we have*

$$\left| L - \sum_{k=0}^{n} c_k \right| < \epsilon$$

Theorem 9. *If $\sum c_k$ converges, then $c_k \to 0$.*

Remark: The converse is false, since $\sum \frac{1}{n}$ grows arbitrarily large because

$$\sum \frac{1}{n} = 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) + \left(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}\right) + \ldots \geq 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \ldots$$

Remark: The trouble with the above definition is that it requires knowing L in advance. But we want to use series to define new numbers. The following definition does not mention L.

Definition 8. *A sequence* $\{c_k\}$ *is a* Cauchy sequence *if for each* $\epsilon > 0$ *there is an* N *such that whenever* $N < m < n$ *we have*

$$\left| \sum_{k=m}^{n} c_k \right| < \epsilon$$

Theorem 10. *If a sequence converges, it is a Cauchy sequence.*

Remark: Calculus depends on the existence of the real numbers; it falls apart if we only believe in rationals. The following theorem requires the real numbers and is absolutely crucial for everything that follows.

Theorem 11. *Conversely, every Cauchy sequence has a limit.*

Proof: It suffices to prove this for sequences of real numbers, since

$$\sum (a_k + ib_k) = \left(\sum a_k\right) + i\left(\sum b_k\right)$$

Choose N satisfying the Cauchy condition for .001 and compute $\sum_{k=0}^{N} a_k$. Suppose, for example, we get 13.264. Then the sum of any finite number of additional terms is smaller than .001 in absolute value, and thus an arbitrary finite sum of terms is between 13.263 and 13.265. So every sufficiently large finite sum starts 13.26.

Choose N satisfying the Cauchy condition for .000001 and repeat the calculation. Suppose we get 13.264552. As before, any finite sum will be between 13.264551 and 13.264553 and therefore have the form 13.26455.... Continue in this way.

We conclude that in Platonic heaven there is a unique infinite decimal, perhaps 13.2655178099936000042...., such that any future calculation of a finite number of terms of our sequence gives this Platonic number up to an appropriate number of places. QED.

Remark: Several objections to this proof are possible, and the most important occurs when a term in the expansion is followed by a sequence of zeros or a sequence of nines. For example, suppose our calculation gives 13.2650000000000000000. The next value might be 2 and in that case all of these digits are correct. But if some future value adds something very, very slightly below zero, then almost all the digits were wrong starting with the 5, and 5000000000000 should have been 4999999999999999. The problem, of course, is that some real numbers have two decimal expansions, one ending in all nines and the other ending in all zeros.

If this does not bother you, swell; proceed. But if it does, or if you don't believe in Platonic heaven, keep reading. The standard proof of this theorem begins with an alternate definition of the set of real numbers. The alternate definition makes it clear that Cauchy sequences always converge. The decimal notation of real numbers, including the ambiguity between ending in all nines and ending in all zeros, becomes a theorem in this new approach.

By very good fortune, the convergence of Cauchy sequences is the central idea of the new construction. Indeed, one approach is to start with the rational numbers and then define a real number to be an equivalence class of Cauchy sequences of rational numbers. So all of the magic of real numbers is captured by our theorem, and we need not worry that some other essential point about reals hasn't been made clear.

2.2 A Big Problem

We want to use infinite sums in a relaxed way, using all the tricks we employ with finite sums. But something can go very wrong with these infinite sums.

Recall that $\frac{1}{1+x} = 1-x+x^2-x^3+....$ Integrate both sides to obtain $\log(1+x) = x-\frac{x^2}{2}+\frac{x^3}{3}-\frac{x^4}{4}+....$ In particular

$$\log(2) = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + ...$$

and

$$\frac{1}{2}\log(2) = 0 + \frac{1}{2} + 0 - \frac{1}{4} + ...$$

Adding

$$\frac{3}{2}\log(2) = 1 + \frac{1}{3} - \frac{1}{2} + \frac{1}{5} + \frac{1}{7} - \frac{1}{4} + ...$$

But this is just a rearrangement of the original series, although the sum is different. So the order of the terms matters in infinite series.

This is terrible! When Gauss was six years old, his genius was recognized because he added the numbers from 1 to 100 by rearranging the sum as $(1 + 100) + (2 + 99) + (3 + 98) + ... = 101 + 101 + 101 + ... = 5050$. Tricks like this fail for infinite series.

Definition 9. *A rearrangement of a series $\{c_k\}$ is a one-to-one and onto map $f : N \to N$, where N is the set of non-negative natural numbers. This produces a new series $\{c_{f(k)}\}$.*

Definition 10. *Let a_k be a series of real numbers. We say that $\sum a_k = \infty$ if for each bound B there is an N such that whenever $n > N$ we have $\sum_{k=0}^{n} \geq B$.*

Theorem 12. *Let $\{a_k\}$ be a series of real numbers and suppose $a_k \geq 0$. Either there is a unique L such that $\sum_{k=0}^{\infty} a_k = L$ or else $\sum_{k=0}^{\infty} a_k = \infty$. Any rearrangement of the series has the same sum.*

Remark: This is the first ray of light. Our problem does not occur if every term in the sequence is a positive real number. Incidentally, the proof is very easy and can be left to the reader.

Remark: Now suppose $\{a_k\}$ is an arbitrary sequence of real numbers. We could form the sum of all the positive terms P, change the signs of all the negative terms and form their sum N, and define the sum of all the numbers to be $P - N$. This would work for any sequence whatever,

and give an answer which is either L or ∞ or $-\infty$ (except in one case). The sum would be independent of the arrangement of the terms. Nirvana!

The bad case is when $P = \infty$ and $N = \infty$. Then the subtraction makes no sense.

Definition 11. *Suppose $\{a_k\}$ is a sequence of real numbers. We say this sequence is in L^1 if P and N are both finite. In that case, it is easy to prove that $P-N$ gives $\sum_{k=0}^{\infty} a_k$ according to our original definition. Moreover, any rearrangement of the series gives the same sum. Finally, it is clear that $\{a_k\}$ is in L^1 exactly when $\sum |a_k|$ converges.*

Remark: We say a sequence *converges absolutely* if $\sum |a_k|$ converges. This is the terminology we will use from now on. Our L^1 is commonly used in real analysis, for both series and integrals.

Corollary 2. *Suppose $\{c_k\}$ is a series of complex numbers. We say this series is absolutely convergent if $\sum |c_k|$ converges. Write $c_k = a_k + ib_k$. Since $|a_k| \leq |c_k|$ and $|b_k| \leq |c_k|$, the real and imaginary series converge absolutely to the real and imaginary components of the limit, and a rearrangement of such a series gives the same sum.*

Remark: Finally, suppose $\{a_k\}$ is a series of real numbers which does not converge absolutely. Let us assume that $a_k \to 0$, since otherwise the series certainly does not converge. Then either P is infinite and N is not and the sum is ∞, or P is finite and N is not and the sum is $-\infty$ or $P = N = \infty$. In the first two cases, any rearrangement gives the same sum. So the only unusual case is the third one, and for it we have the following spectacular result:

Theorem 13. *Suppose $\{a_k\}$ is a sequence of real numbers, $a_k \to 0$, and the sum of all positive terms P and the sum of minus all negative terms Q are both infinite. Then for any $L \in R$, some rearrangement of the series converges to L.*

Proof: Start adding positive terms until the sum is larger than L. Then add negative terms until the sum is less than L. Then add more positive terms until the sum is greater than L. Etc. QED.

2.3 An Alternate Approach

There is a research area called "the theory of divergent series" about alternate definitions of sequence convergence. These new methods gives the original result for convergent series, but also give values to certain divergent series. For example, Cesaro's summation method calculates the partial sums $s_n = \sum_{k=0}^{n} a_k$, and then defines the sum of the series to be the limit of $\frac{\sigma_1+\sigma_2+...+\sigma_n}{n}$ as $n \to \infty$. The Cesaro sum of a convergent series is its ordinary sum, and the Cesaro sum of $1 - 1 + 1 - 1 + ...$ is $\frac{1}{2}$. The Fourier series of a continuous function need not converge to the function, but according to a beautiful theorem of Fejer, the Fourier series of a continuous function does converge to f if we use the Cesaro method.

It is possible to imagine an alternate history of mathematics in which the standard definition of summing a series would be to sum the positive elements, sum the negative elements, require that both sums be finite, and subtract the sums. In this alternate world, all infinite sums could be rearranged. Our current definition of convergence would become a "divergent series method," in which a value is squeezed out of a divergent series by arranging the terms in a particular order.

This alternate approach would not, however, avoid using a deep property of real numbers. Suppose a_k are positive numbers. Notice that as we add more and more terms, the partial sums increase. If these partial sums are unbounded, we say the limit is ∞. Otherwise we define the limit of the series to be the least upper bound of the partial sums. Here we use another deep property of real numbers: every set of reals which is bounded above has a least upper bound.

Thus the reals are in some sense defined by two properties; **every Cauchy sequence converges** and **every subset bounded above has a least upper bound**. It is easy to prove that these two properties are equivalent.

There are essentially two approaches to construct the real numbers. Each approach starts with the rational numbers. In one approach, we add additional numbers to make Cauchy sequences converge, and in the other we add additional numbers to create least upper bounds for bounded sets. The two approaches give the same set of real numbers in the end.

2.4 Uniform Convergence of Functions

Definition 12. *Suppose $f_k(p)$ are complex-valued functions defined on some set M. We say $\sum f_k(p)$ converges pointwise to $f(p)$ if for each $p \in M$, $\sum f_k(p)$ converges to $f(p)$.*

Remark: Here M might be an interval $[a, b]$ in R, or an open subset of C, or any other set.

Remark: When we add series, the basic idea is that finite sums $\sum c_k$ look more and more like L. Here the intuition is that finite sums $\sum f_k(p)$ look more and more like $f(p)$. However, this intuition is tricky, as the following example shows.

Fourier and Cauchy were contemporaries. Fourier led an interesting life. He was the secretary of the academic side of the army when France invaded Egypt, so while the army was doing its thing, he and his workers were making careful traces of hieroglyphics. French soldiers also discovered the Rosetta Stone, but that was taken to England by the British when they arrived and defeated the French army. These events eventually led to the decipherment of hieroglyphics, and although partial progress was made by a British amateur named Thomas Young, the main decipherment was done by Jean-Francois Champollion in France. It turned out that hieroglyphics are partly phonetic and similar to the Coptic language, which Champollion learned.

Fourier invented what we now call Fourier series. According to Fourier, any reasonable periodic function on R with period 2π can be written

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx)$$

where

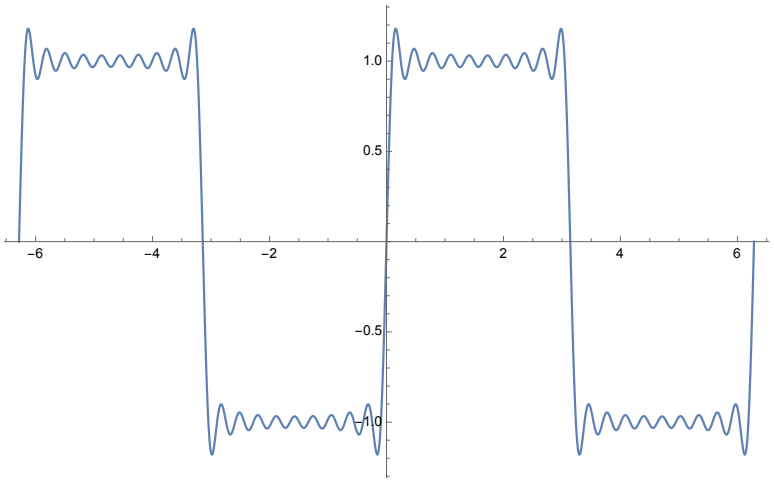
$$a_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos kx \, dx \quad \text{and} \quad b_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin kx \, dx$$

For example, if $f(x)$ is a square wave with $f(x) = -1$ on $(-\pi, 0)$ and $f(x) = 1$ on $(0, \pi)$, then an easy calculation gives $a_k = 0$ always, and $b_k = \frac{4}{\pi k}$ for k odd and zero otherwise. Thus the square wave equals

$$\frac{4}{\pi}\left(\sin x - \frac{\sin 3x}{3} + \frac{\sin 5x}{5} - \frac{\sin 7x}{7} + \ldots\right)$$

When Cauchy heard of this result, he told Fourier not to publish immediately because his manuscript had mistakes. For example, Cauchy had published a book proving that the sum of continuous functions is continuous, and yet the equation just displayed claims to represent a discontinuous function as a sum of continuous functions.

Below is a Mathematica plot of the sum of the first ten terms of this series. It is clear that the sum is converging to the square wave and in the limit will jump from a value of -1 to a value of 1 at multiples of π.



Eventually Cauchy revised his book, and later Dirichlet rigorously proved that this and most Fourier series converge pointwise to $f(x)$. Just before the jump, notice that the sum rises about 10% above the function. This is called the Gibb's phenomenon, happens generally at jumps,

and was discovered when mechanical calculators actually computed Fourier series in the late 1800's.

To rescue his theorem, Cauchy made the following definition:

Definition 13. *Let $f_k(p)$ be complex valued functions on a set M. We say $\sum f_k(x) = f(x)$* uniformly *on this set if for each $\epsilon > 0$, there is an N such that when $n > N$ we have for all $p \in M$ simultaneously*

$$\left| f(x) - \sum_{k=0}^{n} f_k(x) \right| < \epsilon$$

Remark: In short, if convergence is uniform then our intuition that finite sums of the $f_k(x)$ look more and more like $f(x)$ is actually true.

Theorem 14 (The Weierstrass M-Test)**.** *Suppose there are real constants m_k such that for all p*

$$|f_k(p)| \leq m_k$$

and such that $\sum m_k$ converges. Then $\sum f_k(p)$ has a uniform limit $f(p)$.

Theorem 15. *Suppose M is a subset of R or C. If $f_k(p)$ are continuous functions on M and $\sum f_k(p)$ converges uniformly to $f(p)$, then $f(p)$ is continuous.*

Theorem 16. *Suppose $f_k(x)$ are continuous real-valued functions on $[a, b]$ and $\sum f_k(x)$ converges uniformly to $f(x)$. Then*

$$\sum \int_a^b f_k(x)\, dx = \int_a^b f(x)\, dx$$

2.5 Power Series

A power series is a series $\sum_{k=0}^{\infty} c_k z^k$. The lesson of this short final section is that power series are nice in every possible way. They converge uniformly, they converge absolutely, and their convergence domain is easily described. They can be differentiated term by term. See below for the precise results.

Theorem 17. *Given a power series, there is a constant R called the* radius of convergence *such that the series converges for $|z| < R$ and diverges for $|z| > R$. Behavior on the circle $|z| = R$ is varied. The series converges absolutely when $|z| < R$, and it converges uniformly on $|z| \leq R_0$ for any $R_0 < R$.*

Remark: The constant R can be infinite, as for

$$e^z = 1 + z + \frac{z^2}{2!} + ...$$

it can be finite as for

$$\frac{1}{1-z} = 1 + z + z^2 + ...$$

and it can be zero, as for

$$1 + 1!z + 2!z^2 + 3!z^3 + ...$$

When R is finite, the series may diverge everywhere on the boundary, as for $\frac{1}{1-z} = 1+z+z^2+...$, and the series can converge at some boundary points and diverge at others, as for

$$\log(1-z) = -z + \frac{z^2}{2} - \frac{z^3}{3} + ...$$

which converges at $z = 1$ and diverges at $z = -1$. The series can converge on all boundary points, as for example

$$\sum \frac{z^n}{n^2}$$

Proof: Let R be the least upper bound of the set of all r such that $\{|c_k|r^k\}$ is a bounded set. If $|z| > R$, then the $c_k z^k$ are not bounded, so $\sum c_k z^k$ diverges.

If $R_0 < R$, we can find R_1 such that the $|c_k|R_1^k$ are bounded by, say, M and $R_0 < R_1 < R$. Then when $|z| \leq R_0$ we have

$$|c_k z^k| \leq |c_k|R_0^k = |c_k| \left(\frac{R_0}{R_1}\right)^k R_1^k \leq M \left(\frac{R_0}{R_1}\right)^k$$

Since $\left|\frac{R_0}{R_1}\right| < 1$, the series of constants on the right converges, and thus the series on the left converges absolutely and uniformly. QED.

Corollary 3. *The function $f(z) = \sum c_k z^k$ on $|z| < R$ is continuous*

Theorem 18. *Suppose $f(z) = \sum c_k z^k$ on $|z| < R$. Then the radius of convergence of the term by term derivative*

$$\sum k c_k z^{k-1}$$

is also R and $\frac{df}{dz}$ exists and equals this sum.

Corollary 4. *Suppose $f(z)$ is given by a convergent power series on $|z| < R$. Then $f(z)$ has derivatives of all orders, and these derivatives are given by term-by-term differentiation of the power series on $|z| < R$.*

Proof of theorem: If we multiply a convergent series by a constant, the new series converges as well. So we can multiply $\sum k c_k z^{k-1}$ by z and it suffices to study $\sum k c_k z^k$. If the radius of

convergence of the original series is R and $R_0 < R$, choose $R_0 < R_1 < R$ and note that the $|c_k|R_1^k$ are bounded by, say, M, So

$$k|c_k|R_0^k = k|c_k|\left(\frac{R_0}{R_1}\right)^k R_1^k \leq k\left(\frac{R_0}{R_1}\right)^k M$$

We claim that the $k\left(\frac{R_0}{R_1}\right)^k$ are bounded. If so, then the radius of convergence of $\sum kc_k z^k$ is at least R_0, and since R_0 is arbitrarily close to R, the radius of convergence of $\sum kc_k z^k$ must be R. To show our claim, note that $\frac{R_1}{R_0} > 1$ and thus equals $1 + \alpha$ for a positive α. Then $\left(\frac{R_1}{R_0}\right)^k > (1+\alpha)^k = 1 + k\alpha + ... > k\alpha$, and so

$$k\left(\frac{R_0}{R_1}\right)^k < k\frac{1}{k\alpha} = \frac{1}{\alpha}$$

Next we differentiate. If $f(z) = \sum c_k z^k$, then

$$\frac{f(z+h) - f(z)}{h} = \sum c_k \frac{(z+h)^k - z^k}{h} =$$

$$\sum c_k \frac{\Big((z+h) - z\Big)\Big((z+h)^{k-1} + (z+h)^{k-2}z + (z+h)^{k-3}z^2 + ... + z^{k-1}\Big)}{h} =$$

$$\sum c_k \Big((z+h)^{k-1} + (z+h)^{k-2}z + (z+h)^{k-3}z^2 + ... + z^{k-1}\Big)$$

Fix z and think of both sides as functions of h. On the right side we have a sum of continuous functions of h. This sum converges uniformly by the Weierstrass M-Test, because we can restrict to $|z+h| < R_1$ and $|z| < R_1$ and then the kth term is bounded by $|c_k|kR_1^{k-1} < |c_k|kR^{k-1}$ and the initial part of the proof shows that by shrinking R we can make this bounded.

Therefore the left side of the equation is continuous at $h = 0$, or in other words, has a limit as $h \to 0$. Moreover, this limiting value is the value of the sum when $h = 0$ and so $\sum c_k k z^{k-1}$. QED.

Chapter 3

The Cauchy-Riemann Equations and Consequences

The Cauchy-Riemann equations were discovered by Cauchy around 1820. They were later used extensively by Riemann in his thesis and subsequent paper on Riemann surfaces. Hence the equations are named after both men.

3.1 Differentiation Rules

Theorem 19. *The standard differentiation techniques work for* $\frac{df}{dz}$. *These include the rules for sums, products, and quotients, and the chain rule.*

Remark: Most proofs work exactly as before. The chain rule requires slightly more work.

3.1.1 The Product Rule

3.1.2 The Chain Rule

3.1.3 The Inverse Function Rule

3.2 Complex Functions as Maps from the Plane to the Plane

Advanced calculus is partly about maps $R^m \to R^n$. In the special case $m = n = 2$, such a map is often written $(x, y) \to (u, v) = (u(x, y), v(x, y))$. Each function of a complex variable defines such a map. Rather than writing u and v, we almost always call the components f_1 and f_2. So $f(z) : (x, y) \to (u, v) = (f_1(x, y), f_2(x, y))$.

Example 1: Since $e^z = e^{x+iy} = e^x(\cos y + i \sin y)$, we have $f_1(x,y) = e^x \cos y$ and $f_2(x,y) = e^x \sin y$.

Example 2: Since $\text{Log}(z) = |z| + i\theta$, we have $f_1(x,y) = \sqrt{x^2+y^2}$ and $f_2(x,y) = \arctan \frac{y}{x}$.

Example 3: If $f(z) = z^2$, we have $f_1(x,y) = x^2 - y^2$ and $f_2(x,y) = 2xy$.

Example 4: If $f(z) = \sin z = \sin(x,y) = \sin x \cos iy + \cos x \sin iy = \sin x \cosh y - i \cos x \sinh y$. So $f_1(x,y) = \sin x \cosh y$ and $f_2(x,y) = -\cos x \sinh y$.

3.3 The Cauchy-Riemann Equations

Theorem 20. *Suppose $f(z) = f_1(x,y) + i f_2(x,y)$. If $\frac{df}{dz}$ exists at a point, then the four partial derivatives $\frac{\partial f_1}{\partial x}, \frac{\partial f_1}{\partial y}, \frac{\partial f_2}{\partial x}$, and $\frac{\partial f_2}{\partial y}$ all exist at the point and*

$$\frac{\partial f_1}{\partial x} = \frac{\partial f_2}{\partial y}$$

$$\frac{\partial f_1}{\partial y} = -\frac{\partial f_2}{\partial x}$$

Proof: Recall that $\frac{df}{dz}$ requires that limits be equal as we approach the limit from any direction. In particular, the horizontal and vertical limits must be equal. So

$$\frac{df}{dz} = \lim_{h\to 0} \frac{(f_1(x+h,y) + i f_2(x+h,y)) - (f_1(x,y) + i f_2(x,y))}{h}$$

$$= \lim_{h\to)} \frac{f_1(x+h,h) - f_1(x,y)}{h} + i \lim_{h\to 0} \frac{f_2(x+h,y) - f_2(x,y)}{h}$$

$$= \frac{\partial f_1}{\partial x} + i \frac{\partial f_2}{\partial x}$$

The vertical limit is formed by replacing h by ih for a real h:

$$\frac{df}{dz} = \lim_{h\to 0} \frac{(f_1(x,y+h) + i f_2(x,y+h)) - (f_1(x,y) + i f_2(x,y))}{ih}$$

$$= \lim_{h\to)} \frac{f_1(x,h+h) - f_1(x,y)}{ih} + i \lim_{h\to 0} \frac{f_2(x,y+h) - f_2(x,y)}{ih}$$

$$= \frac{\partial f_2}{\partial y} - i \frac{\partial f_1}{\partial y}$$

Example: If $f(z) = e^z = e^x \cos y + ie^x \sin y$ then

$$\frac{\partial f_1}{\partial x} = e^x \cos y = \frac{\partial f_2}{\partial y}$$

$$\frac{\partial f_2}{\partial y} = -e^x \sin y = -\frac{\partial f_2}{\partial x}$$

Exercise: Show that a real-valued holomorphic function $f(z)$ on a domain $\mathcal{U}$ is constant.

3.4 The Converse of the Cauchy-Riemann Equations

Theorem 21. *Let $f(z) = f_1(x,y) + if_2(x,y)$ on a domain $\mathcal{U}$. Suppose $\frac{\partial f_1}{\partial x}, \frac{\partial f_1}{\partial y}, \frac{\partial f_2}{\partial x}, \frac{\partial f_2}{\partial y}$ exist and are continuous on $\mathcal{U}$, and suppose f_1 and f_2 satisfy the Cauchy-Riemann equations on $\mathcal{U}$. Then $\frac{df}{dz}$ exists at each point of $\mathcal{U}$ and is continuous on $\mathcal{U}$.*

Remark: Notice that this is not an exact converse of the previous theorem. We will eventually fill the gap between these theorems by proving that is $\frac{df}{dz}$ exists at each point of $\mathcal{U}$, then this derivative is continuous on $\mathcal{U}$ and thus all four partials are continuous on $\mathcal{U}$.

Sketch of Proof: We want to apply the mean value theorem, which asserts that there is an intermediate value θ between x and $x+h$ with

$$F(x+h) - F(x) = h\frac{dF}{dx}(\theta)$$

We'll use this to approximate

$$\frac{f(x+h, y+k) - f(x,y)}{h+ik}$$

The idea is to move from $(x+h, y+k)$ to (x,y) on variable at a time, and apply the mean value theorem to each move. If we move $f_1(x+h, y+k)$ to $f_1(x+h, y)$ and then move $f_1(x+h, y)$ to $f_1(x,y)$, we obtain

$$k\frac{\partial f_1}{\partial y}(x+h, \theta_1) + h\frac{\partial f_1}{\partial x}(\theta_2, y)$$

If we move $if_2(x+h, y+k)$ to $if_2(x+h, y)$ and then move $if_2(x+h, y)$ to $f_2(x,y)$, we get

$$ik\frac{\partial f_2}{\partial y}(x+h, \theta_3) + ih\frac{\partial f_2}{\partial x}(\theta_4.y)$$

If each of these expressions were evaluated at x, y, we could apply the Cauchy-Riemann equations to obtain

$$\frac{h+ik}{h+ik}\frac{\partial f_1}{\partial x}(x,y) + \frac{ih-k}{h+ik}\frac{\partial f_2}{\partial x}(x,y)$$

which would equal

$$\frac{\partial f_1}{\partial x} + i\frac{\partial f_2}{\partial x} = \frac{df}{dz}$$

So our expression is this last result plus a sum of differences, each difference being a partial evaluated one place minus the same partial evaluated somewhere else. This difference goes to zero in the limit by continuity of the partials, so $\lim \frac{f(x+h,y+k)-f(x,y)}{h+ik}$ exists.

Remark: This theorem gives an alternate way to extend functions to the complex plane and prove the extension holomorphic. For example, $\text{Log}(z) = |z| + i \arg z$ is holomorphic because $f_1(x,y) = \sqrt{x^2+y^2}$ and $f_2(x,y) = \arctan \frac{y}{x}$ satisfy the Cauchy-Riemann equations.

3.5 Linear Approximation of Maps $R^2 \to R^2$

Next we want to discover the geometric meaning of the Cauchy-Riemann equations. This requires a brief revisit to second year calculus.

Theorem 22. *Let $F(x,y) = (u(x,y), v(x,y)) : \mathcal{U} \to R^2$ be a map given by functions with continuous partial derivatives on $\mathcal{U}$. Let $p = (a,b) \in \mathcal{U}$. Then F has a unique linear approximation near p.*

Indeed, each linear transformation $A : R^2 \to R^2$ defines a map

$$\epsilon : \mathcal{U} \to R^2$$

such that

$$F(x,y) = F(a,b) + A\begin{pmatrix} x-a \\ y-b \end{pmatrix} + \epsilon(x,y)$$

There is exactly one linear transformation A such that

$$\lim_{(x,y)\to(a,b)} \frac{\epsilon(x,y)}{||(x-a, y-b)||} = 0$$

Proof: First we prove uniqueness. If two matrices A and B both work, then subtraction gives

$$(A-B)\begin{pmatrix} x-a \\ y-b \end{pmatrix} = \epsilon_2(x,y) - \epsilon_1(x,y)$$

Then

$$\frac{(A-B)\begin{pmatrix} x-a \\ y-b \end{pmatrix}}{||(x-a, y-b)||} \to 0$$

Select a positive real λ and multiply the vector $(x-a, y-b)$ by λ. Notice that λ cancels out of the expression on the left, and yet as $\lambda \to 0, \lambda(x-a, y-b) \to 0$. We conclude that $(A-B)\begin{pmatrix} x-a \\ y-b \end{pmatrix}$ is identically zero, and thus that $A = B$.

Existence is proved using the mean-value theorem techniques introduced earlier. Namely, $u(x,y) - u(a,b) = u(x,y) - u(x,b) + u(x,b) - u(a,b) = (y-b)\frac{\partial u}{\partial y}(x,\theta_1) + (x-a)\frac{\partial u}{\partial x}(\theta_2, b)$. This can be rewritten as the expression we really want, plus ϵ,

$$u(x,y) = u(a,b) + \frac{\partial u}{\partial x}(a,b)(x-a) + \frac{\partial u}{\partial y}(a,b)(y-b) + \epsilon(x,y)$$

where

$$\epsilon(x,y) = \left(\frac{\partial u}{\partial x}(\theta_2, b) - \frac{\partial u}{\partial x}(a,b)\right)(x-a) + \left(\frac{\partial u}{\partial y}(x,\theta_1) - \frac{\partial u}{\partial y}(a,b)\right)(y-b)$$

and a similar result holds for v. Since partial derivatives are continuous, an easy argument concludes the proof.

3.6 The Local Geometry of Holomorphic Maps

Next we apply the previous theorem to a holomorphic map $f(z) = (f_1(x,y), f_2(x,y))$. The theorem states that this map has a unique linear approximation at $z_0 = a + ib$, given by the matrix

$$\begin{pmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} \\ \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} \end{pmatrix}$$

evaluated at (a,b).

By the Cauchy-Riemann equations, this equals

$$\begin{pmatrix} \frac{\partial f_1}{\partial x} & -\frac{\partial f_2}{\partial x} \\ \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_1}{\partial x} \end{pmatrix}$$

Note that

$$\frac{df}{dz} = \frac{\partial f_1}{\partial x} + i\frac{\partial f_2}{\partial x} = \left|\frac{df}{dz}\right|(\cos\theta + i\sin\theta)$$

where $\theta = \arg\left(\frac{df}{dz}\right)$.

The above matrix is then

$$\left|\frac{df}{dz}\right| \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

Theorem 23. *Suppose $f(z)$ is holomorphic and $\frac{df}{dz}$ is continuous. Then near $z_0 = a+ib$, the map $f : R^2 \to R^2$ magnifies by $\left|\frac{df}{dz}\right|$ and rotates by* $\arg\left(\frac{df}{dz}\right)$. *In particular, maps given by holomorphic f are* conformal*; that is, they preserve angles.*

Remark: This theorem is not very informative at zeroes of $\frac{df}{dz}$ because the linear approximation at such points is zero. We later find more precise geometrical information near such points.

Remark: Notice that the determinant of these linear approximations of holomorphic f are greater than or equal to zero. It follows that holomorphic maps preserve orientation and thus cannot be reflections. The reader should directly show, from the Cauchy-Riemann equations and again directly from the definition, that $f(z) = \overline{z}$ is not holomorphic.

Chapter 4

Integration

4.1 Definition of the Complex Integral

We now begin integration theory. It is a remarkable fact that the deep features of complex analysis depend on integration theory. When I was a graduate student, I met a mathematician from Florida who was able to prove these deep results while avoiding integration. Unfortunately, his proofs were much harder than the canonical proofs which follow.

If $f(z)$ is continuous on a domain $\mathcal{U}$, we want to define

$$\int f(z)dz$$

The ordinary calculus integral $\int f(x)dx$ should appear as a special case of this complex integral, so the complex integral must be a line integral over paths in $\mathcal{U}$ rather than a double integral of some sort. Below is a picture of three possible paths.

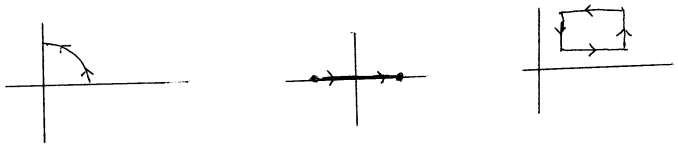


Figure 4.1: Various Integration Paths

We do not require that f be holomorphic. It is complex valued, but in special cases f may be real and form a sort of "curtain" over the path. In these cases, it is easy to guess that $\int f(z)\,dz$ gives the area of this curtain, *but this is false*. Leibniz knew what he was doing when he invented

the modern notation of integration. The "dz" is important. The notation tells us that dz will be real when we integrate horizontally, but imaginary when we integrate vertically.

We could, of course, define integration so that it always give the area of the curtain, but then the fundamental theorem of calculus would be false — and we definitely need that theorem.

Definition 14. *A* parameterized path *is a piecewise continuously differentiable path* $z(t) = x(t) + iy(t)$ *defined on an interval* $a \le t \le b$.

Remark: Instead of $z(t)$, we often write $\gamma(t)$. Too many z's lead to confusion.

Remark: Thus there is a finite decomposition $a = t_0 < t_1 < \ldots < t_n = b$ such that $x(t)$ and $y(t)$ have continuous derivatives on $[t_{i-1}, t_i]$. This includes one-sided derivatives at the t_i. Notice that the derivatives at the t_i from the left and right need not agree, i.e., the path can have corners.

Definition 15. *Let* $f(z)$ *be a continuous complex valued function on a domain* $\mathcal{U}$*, and* $z(t)$ *be a parameterized path in* $\mathcal{U}$*. Then*

$$\int_\gamma f(z)\,dz = \int_a^b f(\gamma(t))\frac{d\gamma}{dt}\,dt$$

Example: Let $f(z) = z^2$ and let γ be the counterclockwise semicircle with radius 1 starting at $z = 1$ and ending at $z = -1$. Thus $z(t) = \cos t + i \sin t, 0 \le t \le \pi$. Then we get the following; all details are given because we want to make a point:

$$\int_\gamma z^2\,dz = \int_0^\pi (\cos(t) + i\sin(t))^2(-\sin t + i\cos t)\,dt =$$

$$\int_0^\pi \Big((\cos^2 - \sin^2 t)(-\sin t) - 2\cos(t)\sin(t)\cos t\Big)\,dt$$

$$+i\int_0^\pi \Big(-2\cos t \sin t \sin t + (\cos^2 t - \sin^2 t)\cos t\Big)\,dt =$$

$$\int_0^\pi \Big(-3\cos^2 t \sin t + \sin^3 t\Big)\,dt + i\int_0^\pi \Big(-3\sin^2 t\cos t + \cos^3 t\Big)\,dt =$$

$$\int_0^\pi \big(-3\cos^2 t \sin t + \sin t(1 - \cos^2 t)\big)\,dt + i\int_0^\pi \Big(-3\sin^2 t\cos t + \cos t(1 - \sin^2 t)\Big)\,dt =$$

$$\int_0^\pi \big(-4\cos^2 t \sin t + \sin t\big)\,dt + i\int_0^\pi \Big(-4\sin^2 t\cos t + \cos t\Big)\,dt =$$

$$\left(4\frac{\cos^3 t}{3} - \cos t\right)\Big|_0^\pi + i\left(-4\frac{\sin^3 t}{3} + \sin t\right)\Big|_0^\pi = \left(-\frac{4}{3} + 1 - \frac{4}{3} + 1\right) + 0 = -\frac{2}{3}$$

Example Continued: Or $\int_\gamma z^2 dz = \frac{z^3}{3}\Big|_1^{-1} = -\frac{1}{3} - \frac{1}{3} = -\frac{2}{3}$.

Theorem 24. *Suppose the continuous $f(z)$ is the complex derivative of a holomorphic function $\frac{dg}{dz}$. Then*

$$\int_\gamma f(z)\, dz = g(\textit{end of}\,\gamma) - g(\textit{beginning of}\,\gamma)$$

Proof:

$$\int_\gamma f(z)dz = \int_\gamma \frac{dg}{dt}dt = \int_a^b \left(\frac{\partial g_1}{\partial x} + i\frac{\partial g_2}{\partial x}\right)\left(\frac{dx}{dt} + i\frac{dy}{dt}\right)dt =$$

$$\int_a^b \left(\frac{\partial g_1}{\partial x}\frac{dx}{dt} - \frac{\partial g_2}{\partial x}\frac{dy}{dt}\right)dt + i\int_a^b \left(\frac{\partial g_2}{\partial x}\frac{dx}{dt} + \frac{\partial g_1}{\partial x}\frac{dy}{dt}\right)dt$$

Apply the Cauchy-Riemann equations to get

$$\int_a^b \left(\frac{\partial g_1}{\partial x}\frac{dx}{dt} + \frac{\partial g_1}{\partial y}\frac{dy}{dt}\right)dt + i\int_a^b \left(\frac{\partial g_2}{\partial x}\frac{dx}{dt} + \frac{\partial g_2}{\partial y}\frac{dy}{dt}\right)dt =$$

$$\int_a^b \frac{d}{dt}g_1(x(t), y(t))dt + i\int_a^b \frac{d}{dt}g_2(x(t), y(t))dt = (g_1 + ig_2)\Big|_a^b = g(\text{end of}\,\gamma) - g(\text{beginning of}\,\gamma)$$

Remark: Notice that the complex integral is just a special case of the line integral studied in advanced calculus. Indeed if (E_x, E_y) is a vector field and γ is a piecewise continuously differentiable path, we defined

$$\int_\gamma E \cdot d\gamma = \int_a^b E(x(t), y(t)) \cdot \left(\frac{dx}{dt}, \frac{dy}{dt}\right)\, dt$$

Thus our integral is given by two line integrals

$$\int_a^b (f_1 + if_2)\left(\frac{dx}{dt} + i\frac{dy}{dt}\right)dt = \int_a^b (f_1, -f_2) \cdot \left(\frac{dx}{dt}, \frac{dy}{dt}\right)dt + i\int_a^b (f_2, f_1) \cdot \left(\frac{dx}{dt}, \frac{dy}{dt}\right)dt$$

Remark: Summarizing, the complex integral of an $f = \frac{dg}{dz}$ is calculated by the fundamental theorem, exactly as in ordinary calculus. The integral can be calculated more generally for

any continuous f, and then the calculation requires the calculation of two line integrals from advanced calculus.

Consequently, it is not necessary to discuss in detail the many special cases and peripheral matters already discussed in advanced calculus. We simply list them so the reader can check them off as known:

- We always assume that the parameter of integration is a single interval $[a, b]$. But in practice we integrate over, say, rectangles by integrating separately over each side, and do not bother to adjust parameterizations to be portions of a single interval. Thus each side could be parameterized by $[0, 1]$
- Suppose $t \in [a, b]$ is actually $t(u)$ for $u \in [c, d]$, where $t(c) = a$ and $t(d) = b$. Then by the chain rule, the integral over $[a, b]$ equals the integral over $[c, d]$. In short, equivalent parameterizations give the same integral.
- The line from (a, b) to (c, d) can be parameterized as $(a + tc, b + td)$ where $0 \leq t \leq 1$.
- The circle of radius r centered at (a, b) is $(a + r\cos t, b + r\sin t)$ where $0 \leq t \leq 2\pi$.
- The graph of the function $y = h(x)$ can be made into a parameterized path by $t \to (t, h(t))$

Important Remark: Let us integrate z^n around a complete circle of any radius about the origin. Here the power n can be any integer, positive or negative. Since z^n is the derivative of $\frac{z^{n+1}}{n+1}$ and since the path begins and ends at the same point, all of these integrals are zero, except one. If $n = -1$, it does not make sense to divide by $n + 1$, so we must integrate $\frac{1}{z}$ directly.

We will do this calculation two different ways. Parameterize the circle as re^{it} for $0 \leq t \leq 2\pi$. Then

$$\int_\gamma \frac{1}{z} dz = \int_0^{2\pi} \frac{1}{re^{it}} rie^{it} dt = i \int_0^{2\pi} dt = 2\pi i$$

Alternately,

$$\int_\gamma \frac{1}{z} dz = \int_\gamma \frac{dLog(z)}{dz} dz = Log(z)\Big|_0^{2\pi}$$

Recall that the logarithm is a multiple valued function. At z moves around the circle counterclockwise, the argument of z increases by 2π. Recall that $Log(z) = ln|z| + i \arg z$. Thus the integral is

$$\int_\gamma \frac{1}{z} dz = 2\pi i$$

Remark: Curiously, the point made in the previous example already occurred in the first public paper Newton wrote about the calculus. Newton integrated by expanding functions in series

and integrating term by term. Some of his expansions had both positive and negative powers of x. In one example, Newton was faced with integrating such a series, and he integrated each term except the $\frac{1}{x}$ term by using the fundamental theorem of calculus. Newton put a box around $\frac{1}{x}$, meaning that he would have to integrate that using numerical methods.

The Newton paper was placed in the Royal Society's offices in London, and there Leibniz read the paper sometime before 1672. Leibniz' notes survive. When he came to the above spot, he wrote $\int \frac{1}{x}$ rather than putting a box around $\frac{1}{x}$.

Remark: At this spot, I recommend taking time to compute several integral, both using the line integral approach and using the fundamental theorem of calculus, just to make sure that the ideas become straightforward and natural.

4.2 A Useful Inequality

Theorem 25. *Let $f(z)$ be a continuous complex valued function and γ a piecewise continuously differentiable path. Then*

$$\left| \int_\gamma f(z)\, dz \right| \leq \left(\max_\gamma |f(z)| \right) (\textit{length of } \gamma)$$

Proof: Ultimately the integral is $\int_a^b f(z(t)) \frac{dz}{dt} dt$. This is an ordinary integral of a (complex-valued) function of a real variable, so

$$\left| \int_\gamma f(z)\, dz \right| \leq \int_a^b \left| f(z(t)) \frac{dz}{dt} \right| dt \leq \max_\gamma |f(z)| \int_a^b \left| \frac{dz}{dt} \right| dt$$

and a brief calculation shows that the remaining integral gives the length of γ.

4.3 Consequences of the Inequality

Corollary 5. *Let $f_n(z)$ and $f(z)$ be continuous complex-valued functions defined on the range of a piece-wise differentiable curve $\gamma(t), a \leq t \leq b$. Suppose $f_n(z) \to f(z)$ uniformly on this image. Then $\int_\gamma f_n(z)\, dz \to \int_\gamma f(z)\, dz$.*

Corollary 6. *Let $f_n(z)$ and $f(z)$ be continuous complex-valued functions defined on the range of a piece-wise differentiable curve $\gamma(t), a \leq t \leq b$. Suppose $\sum_{n=1}^\infty f_n(z) = f(z)$ uniformly on this image. Then $\sum_{n=1}^\infty \int_\gamma f_n(z)\, dz = \int_\gamma f(z)\, dz$.*

Chapter 5

Review of Advanced Calculus

5.1 Differentiation

Two dimensional calculus is about *functions* like $f(x, y)$ and *vector fields* like $\big(E_x(x, y), E_y(x, y)\big)$.

There is a map, grad $: \{functions\} \to \{vector fields\}$, given by

$$\text{grad} f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right)$$

and a second map, curl $: \{vector fields\} \to \{functions\}$, given by

$$\text{curl } E = \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y}$$

We usually think of these maps as forming a sequence

$$\{functions\} \xrightarrow{\text{grad}} \{vector fields\} \xrightarrow{\text{curl}} \{functions\}$$

The composition of these maps is identically zero because $\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}$.

5.2 Geometry of Subsets

In the plane, there are certain special subsets called *signed 0-dimension sets*, *signed 1-dimensional sets*, and *signed 2-dimensional sets*. A signed 0-dimensional set is a finite set of points, each with an attached plus or minus sign. A signed 1-dimensional subset is a finite set of parameterized curves with an arrow on each indicating a direction to trace the curve. A signed 2-dimensional

subset is a region R with the usual orientation, † x-first and y-second. There are maps ∂, called *boundary maps*:

$$\{\text{signed 0-dimensional sets}\} \xleftarrow{\partial} \{\text{signed 1-dimensional sets}\} \xleftarrow{\partial} \{\text{signed 2-dimensional sets}\}$$

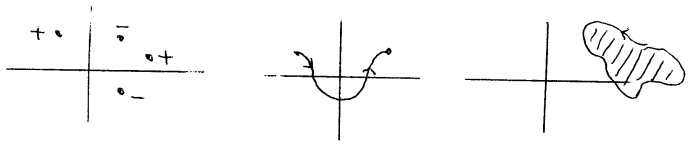


Figure 5.1: k-dimensional Sets

The boundary of a parameterized curve is a set with two points: the end point with a plus sign and the beginning point with a minus sign.

The boundary of a region $\mathcal{R}$ is the boundary curve of this region, traced with a direction so that $\mathcal{R}$ on the left side. This rule for the boundary of a region gives, for instance, the following picture:

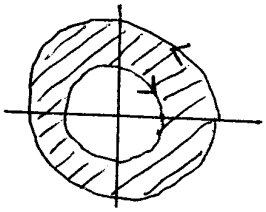


Figure 5.2: k-dimensional Sets

Notice that the composition of these maps is also zero.

5.3 Integration

Consider the sequences introduced in the two previous sections:

$$\{functions\} \xrightarrow{\text{grad}} \{vector fields\} \xrightarrow{\text{curl}} \{functions\}$$

$$\{\text{signed 0-dimensional sets}\} \xleftarrow{\partial} \{\text{signed 1-dimensional sets}\} \xleftarrow{\partial} \{\text{signed 2-dimensional sets}\}$$

Objects on the top line can be integrated over the corresponding sets on the bottom line. For instance, in the first column,

$$\sum \pm f(p_i)$$

where the plus or minus indicates the sign attached to p_i. The integral in the second column is the sum of line integrals over the paths in question:

$$\int_\gamma (E_x, E_y) \cdot d\gamma$$

The integral in the third column is

$$\int \int_{\mathcal{R}} f(x, y) dx dy$$

5.4 Fundamental Theorem of Calculus

At each arrow, there is a theorem relating the integral left of the arrow to the integral right of the arrow. The first of these results is the *generalized fundamental theorem of calculus*:

$$\int_\gamma \text{grad} f \cdot d\gamma = f(\text{end of } \gamma) - f(\text{beginning of } \gamma)$$

The second result is called *Green's theorem*:

$$\int \int_{\mathcal{R}} \left(\frac{\partial E_y}{\partial x} - \frac{\partial E^x}{\partial y} \right) dx dy = \int_{\gamma = \partial \mathcal{R}} E \cdot d\gamma$$

5.5 Differential Forms

This entire theory generalizes to n-dimensions. The main difficulty is defining the objects which play the roles of the functions and vector fields in two dimensions. In the general case, these are called k-forms, where $0 \leq k \leq n$. A single general theorem generalizes the Fundamental Theorem, Green's theorem, and their analogues in three dimensions: the Fundamental Theorem, Stoke's Formula, and the Divergence Theorem. This generalization is known as Stoke's formula.

But for complex variable theory, we only need the two dimensional theory sketched above.

5.6 Solving grad$f = E$ for f

In coordinates, the equation grad$f = E$ becomes

$$\frac{\partial f}{\partial x} = E_x(x, y)$$
$$\frac{\partial f}{\partial y} = E_y(x, y)$$

These equations are the 2-dimensional analogue of

$$\frac{df}{dx} = g$$

in one variable calculus, where a solution f is called *an indefinite integral of* g.

In several-variable calculus, the following results are proved about this equation:

Theorem 26. *If E is a continuous vector field defined on a domain $\mathcal{U}$, and f_1 and f_2 are differentiable functions on $\mathcal{U}$ satisfying gradf_1 = grad$f_2 = g$, then there is a constant c with $f_2 = f_1 + c$.*

Theorem 27. *The equation cannot be solved unless curl $E = 0$.*

Theorem 28. *If curl $E = 0$, the equation can be solved* locally.

That is, every $p \in \mathcal{U}$ has an open neighborhood $\mathcal{V} \subset \mathcal{U}$ on which a continuously differentiable f exists with grad $f = E$.

Remark: This subset can be taken to be an open rectangle. We sketch the proof. Solve $\frac{\partial f}{\partial x} = E_x$, getting

$$f(x, y) = \int_a^x E_x(t, y)\, dt + C(y)$$

Then determine $C(y)$ making $\frac{\partial f}{\partial y} = E_y$. Be sure to understand where the hypothesis that curl $E = 0$ is used in the proof, and how this proof requires a rectangle as domain of E.

Remark: For a general domain $\mathcal{U}$, and an E with curl $E = 0$, there may not exist an f *globally defined* on all of $\mathcal{U}$ with grad $f = E$.

For example, let $\mathcal{U} = R^2 - \{0\}$ and define $E = \left(\frac{-y}{x^2+y^2}, \frac{x}{x^2+y^2}\right)$. Then no f exists on all of $\mathcal{U}$ with $grad\ f = E$. Indeed, the appropriate f is $\arg z$, but this is not globally defined on $\mathcal{U}$.

Theorem 29. *Let E be continuous on the domain $\mathcal{U}$ and fix $p \in \mathcal{U}$. Define a function f on $\mathcal{U}$ as follows: for $q \in \mathcal{U}$, find a piecewise differentiable path γ from p to q and define $f(q) = \int_\gamma E \cdot d\gamma$. Suppose this function is independent of the particular path chosen. Or equivalently, suppose $\int_\gamma E \cdot d\gamma = 0$ for all closed curves in U. Then f is well-defined and grad $f = E$.*

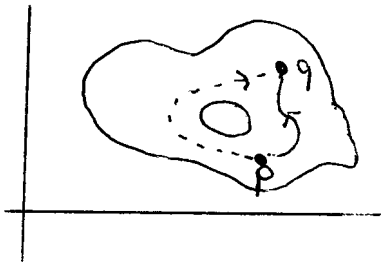


Figure 5.3: Defining f

Remark: It is useful to know the rough idea of the proof. To prove that $\frac{\partial f}{\partial x} = 0$, we find a path from p to the right height for q, and follow that by a horizontal path to q. The initial path is independent of q, so its derivative is zero. An easy calculation shows that the derivative of the integral along the horizontal path is E_x. A similar argument using a different path gives $\frac{\partial f}{\partial y} = E_y$.

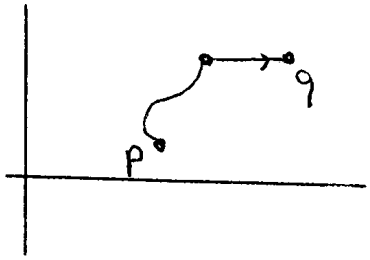


Figure 5.4: Defining f

Theorem 30 (Crucial Result)**.** *If $\mathcal{U}$ has no holes and E is a continuous vector field on $\mathcal{U}$ with curl $E = 0$ there is a* global *f on $\mathcal{U}$ with grad $f = E$.*

Remark: This is proved by showing that the line integral of E does not depend on the path, or equivalently that the line integral of E around closed curves is zero. This follows from Green's theorem since $\int_\gamma E \cdot d\gamma = \int\int$ curl E over the counterclockwise boundary.

Remark: It is at this point that some readers become concerned about rigor. Have we ever defined the "inside" of a curve? If a curve can twist arbitrarily, what does it mean to say it is "counterclockwise?" How precisely was Green's theorem proved?

Readers with these questions should carefully read the more rigorous treatment of these points

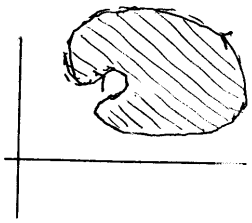


Figure 5.5: Defining f

two chapters ahead. For now, it is helpful to recall the proof of Green's theorem. We want $\int\int \text{curl}\, E$. This requires computing two integrals:

$$\int\int \frac{\partial E_y}{\partial x}\, dx\, dy$$

$$\int\int \frac{\partial E_x}{\partial y}\, dy\, dx$$

In the proof, we integrate these pieces in different orders so we can use the the one-variable fact that the integral of a derivative is the original function. The region of integration must be of a special shape so both of these integrals make sense.

We then extend to the general case by breaking the region into pieces, and noticing that the double integral of the entire region is the sum of the double integrals over pieces, and the line integral of the entire region is the sum of the line integrals over pieces. The second fact requires noticing that internal paths have opposite orientation and cancel out. See the pictures below.

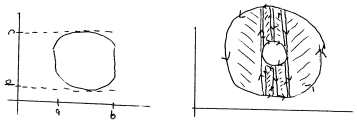


Figure 5.6: Proving Green's Theorem

5.7 Calculus Version of the Residue Theorem

The previous material suffices for the following chapters. But it is fun to prove a more general result, which will reappear in complex analysis as one of our most powerful tools.

Suppose we have a region with finitely many holes, as indicated below. On this region, suppose we have a vector field E with $\text{curl}\, E = 0$. Because of the holes, this E need not equal the gradient of a function.

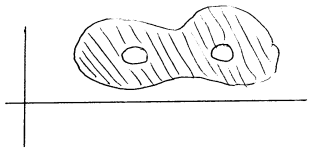


Figure 5.7: Region with Holes

Around each hole, draw a counterclockwise path γ_i. Then a more general form of Green's theorem states

Theorem 31 (Extended Green Theorem)**.** *Suppose γ is a closed path in a region $\mathcal{U}$ with n holes, and let γ_i be counterclockwise paths around each hole. Then*

$$\int_\gamma E \cdot d\gamma = \sum_{holes} (\textit{number of times } \gamma \textit{ goes around ith hole}) \int_{\gamma_i} E \cdot d\gamma_i$$

Proof: See the pictures below. In each picture, the indicated path does not surround any holes, so Green's theorem states the the integral around this path is zero. But the integral around the outside path equals the sum of the integrals around the holes since other parts of the path cancel. Similar pictures show how to handle paths which go around holes a multiple number of times.

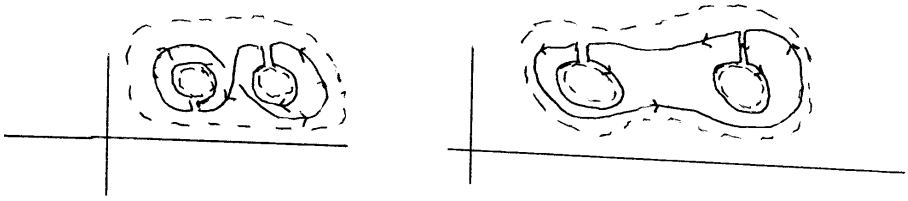


Figure 5.8: Extended Green Theorem

Chapter 6

Complex Analysis as a Chapter in Advanced Calculus

6.1 Cauchy's Theorem

We now begin the deeper theory. In this chapter, a *holomorphic* function on a domain $\mathcal{U}$ is a function with a complex derivative at each point of $\mathcal{U}$, such that the derivative is continuous. We will ultimately prove that the final condition is automatically true if the first condition is true.

Theorem 32 (Cauchy's Theorem)**.** *Suppose f is holomorphic on a domain $\mathcal{U}$ which contains a closed disk D. Then*

$$\int_{\partial D} f(z)\,dz = 0$$

Proof: The integral of f also equals the sum of two line integrals:

$$\int_{\partial D} f(z)\,dz = \int_{\partial D} (f_1, -f_2)\cdot d\gamma + i\int_D (f_2, f_1)\cdot d\gamma$$

The easy way to remember this fact is via Leibniz' notation, since we have $f(z)dz = (f_1 + if_2)(dx + idy) = (f_1\,dx - f_2\,dy) + i(f_2\,dx + f_1\,dy)$. Apply Green's theorem, and notice that $\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y}$ equals one of the following two expressions:

$$-\frac{\partial f_2}{\partial x} - \frac{\partial f_1}{\partial y} \qquad \text{and} \qquad \frac{\partial f_1}{\partial x} - \frac{\partial f_2}{\partial y}$$

Both are zero by the Cauchy-Riemann equations. QED

Remark: Cauchy's theorem is one of our most important results. We later generalize it to state the $\int_\gamma f(z)\,dz = 0$ if f is holomorphic on a domain which includes γ and every point inside γ.

Theorem 33 (Cauchy's Formula)**.** *Suppose f is holomorphic on a domain $\mathcal{U}$ which contains a closed disk D. For every z_0 strictly inside D,*

$$\frac{1}{2\pi i}\int_{\partial D}\frac{f(z)}{z-z_0}\,dz = f(z_0)$$

Remark: This astonishing theorem says that the value of f at every point inside the disk is completely determined by its values on the boundary of the disk.

Proof: Let $g(z) = \frac{f(z)-f(z_0)}{z-z_0}$ when $z \neq z_0$ and $\frac{df}{dz}(z_0$ when $z = z_0$. Since f is holomorphic at z_0, g is continuous at z_0 and holomorphic on the rest of $\mathcal{U}$.

Consider the curve γ shown below; g is holomorphic everywhere inside γ because z_0 is outside. The proof of Cauchy's theorem still works and shows that the integral of g around γ is zero; essential this is Green's theorem applied to the region inside γ.

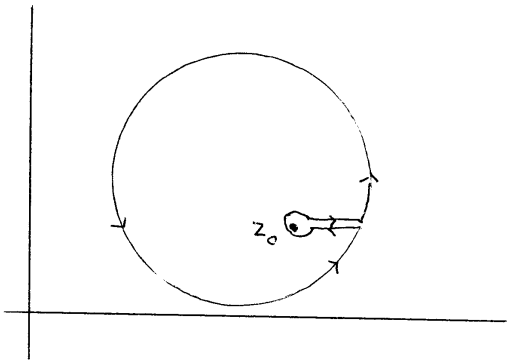


Figure 6.1: Cauchy's Formula

On the other hand, the integral of g over the small circle is bounded by the maximum of $|g|$ near the circle times the length of the circle, and as the circle shrinks to zero, this approaches zero. We conclude that the integral of g around the large outside circle is zero and consequently

$$\frac{1}{2\pi i}\int_{\partial D}\frac{f(z)}{z-z_0}\,dz = \frac{1}{2\pi i}\int_{\partial D}\frac{f(z_0)}{z-z_0}\,dz$$

However, the right side of this expression is $f(z_0)\frac{1}{2\pi i}\int_{\partial D}\frac{1}{z-z_0}\,dz$. Since the derivative of $\mathrm{Log}(z-z_0)$ is $\frac{1}{z-z_0}$, the integral of $\frac{1}{z-z_0}$ is $\mathrm{Log}(z-z_0)$. The real part of this expression is single valued,

namely $\ln|z-z_0|$, so it comes back to itself as we circle z_0. But the imaginary part of $\mathrm{Log}(z-z_0)$ is $\arg(z-z_0)$, and this increases by 2π as we trace the large circle counterclockwise. So the integral on the right side is $\frac{1}{2\pi i} f(z_0)(2\pi i) = f(z_0)$. QED

6.2 Power Series Expansions of Holomorphic Functions

Theorem 34. *Let $f(z)$ be holomorphic on a domain $\mathcal{U}$. Let $z_0 \in \mathcal{U}$. Then f has a power series expansion*

$$f(z) = \sum_{k=0}^{\infty} c_k (z - z_0)^k$$

which converges to f at least in the largest open disk about z_0 inside $\mathcal{U}$.

Corollary 7. *If f is holomorphic on $\mathcal{U}$, it has continuous complex derivatives of all orders.*

Corollary 8. *Suppose a power series $\sum_k^{\infty} c_k (z - z_0)^k$ converges to f on $|z - z_0| < R$. Then at each z_1 in this disk, f is given by a power series $\sum_{k=0}^{\infty} c_k (z - z_1)^k$ centered at z_1 which converges to f near z_1.*

Proof of Theorem: First a notational point. Many authors prefer to reserve z for the arguments of functions and let ζ be the variable of integration. These authors write Cauchy's formula as

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

We won't be dogmatic about the point, but the new notation will be convenient in the following proof because it involves ζ, z, and z_0.

Let D be a closed disk centered at z_0 such that the entire disk is inside $\mathcal{U}$. Call the counterclockwise boundary of this disk γ. By Cauchy's formula, for every z strictly inside the disk,

$$f(z) = \frac{1}{2\pi i} \int_\gamma \frac{f(\zeta)}{\zeta - z} d\zeta = \frac{1}{2\pi i} \int_\gamma \frac{f(\zeta)}{(\zeta - z_0) - (z - z_0)} d\zeta = \frac{1}{2\pi i} \int_\gamma \left(\frac{f(\zeta)}{\zeta - z_0} \right) \left(\frac{1}{1 - \frac{z - z_0}{\zeta - z_0}} \right) d\zeta$$

Since ζ is on the boundary of the disk and z is inside the disk,

$$\left| \frac{z - z_0}{\zeta - z_0} \right| < 1$$

so we can expand the last parenthetical expression as a series converging uniformly in ζ. This gives

$$f(z) = \frac{1}{2\pi i} \int_\gamma \frac{f(\zeta)}{\zeta - z_0} \sum_k \left(\frac{z - z_0}{\zeta - z_0} \right)^k d\zeta$$

We can move the summation sign outside the integral by the last result in chapter 4, so

$$f(z) = \sum_{k=0}^{\infty} \left[\frac{1}{2\pi i} \int_{\gamma} \frac{f(\zeta)}{(\zeta - z_0)^{k+1}} \, d\zeta \right] (z - z_0)^k$$

This is the desired power series about z_0.

If our disk is contained in a larger disk $\tilde{D}$, and if $\tilde{\gamma}$ is the boundary of the larger disk, the same formula holds. But power series expansions are unique, since the coefficients must be

$$\frac{f^{(k)}(z_0)}{k!}$$

by repeated term by term differentiation. So it follows that the original series converges to f in the larger disk, and therefore in the largest open disk about z_0 in $\mathcal{U}$.

6.3 Indefinite Complex Integrals

Theorem 35. *Suppose g, h, and f are holomorphic on a domain $\mathcal{U}$ and $\frac{dg}{dz} = \frac{dh}{dz} = f$ on $\mathcal{U}$. Then $h = g + c$ for some complex constant c.*

Proof: The complex derivative of $g = g_1 + ig_2$ is

$$\frac{\partial g_1}{\partial x} + i\frac{\partial g_2}{\partial x} = \frac{\partial g_2}{\partial y} - i\frac{\partial g_1}{\partial y}$$

by the Cauchy-Riemann equations. So all partials of g agree with corresponding partials of h. It then follows trivially that g and h differ by a constant.

Theorem 36. *Suppose g is holomorphic on a rectangle. Then there is a holomorphic f on this rectangle with $\frac{df}{dz} = g$. Thus indefinite integrals of holomorphic g exist* locally.

Proof: We want to find continuously differentiable real-valued functions f_1 and f_2 with $\operatorname{grad} f_1 = (g_1, -g_2)$ and $\operatorname{grad} f_2 = (g_2, g_1)$, because then $\frac{\partial f_1}{\partial x} = \frac{\partial f_2}{\partial y}$ and $\frac{\partial f_1}{\partial y} = -\frac{\partial f_2}{\partial x}$, so $f = f_1 + if_2$ would satisfy the Cauchy-Riemann equations and by the converse of these equations, f would be holomorphic with derivative g. By several variable calculus, these equations can be solved in a rectangle provided $\operatorname{curl}(g_1, -g_2) = 0$ and $\operatorname{curl}(g_2, g_1) = 0$. Writing these equations out, we need $-\frac{\partial g_2}{\partial x} = \frac{\partial g_1}{\partial y}$ and $\frac{\partial g_1}{\partial x} = \frac{\partial g_2}{\partial y}$, and these equations hold because g is holomorphic.

Remark: Indefinite integrals need not exist *globally*. For example, $\frac{1}{z}$ is holomorphic on $C - \{0\}$, but its indefinite integral would have to be $\text{Log}(z) + c$, and this function is multiple-valued.

Remark: We will later prove that if g is holomorphic on a simply connected $\mathcal{U}$, then it has an indefinite integral.

6.4 Morera's Theorem

Eventually we will prove that if f is holomorphic on a simply-connected $\mathcal{U}$, then the integral of f around any closed curve in $\mathcal{U}$ is zero. At the moment, we know this only in special cases, including integrating around a circle contained with its interior in $\mathcal{U}$.

Theorem 37. *Let f be holomorphic in an open rectangle. Then for any closed curve γ in this rectangle, $\int_\gamma f = 0$.*

Proof: By the previous theorem, there is a g with $f = \frac{dg}{dz}$ (the roles of f and g were reversed in the previous section). So the integral of f over any closed curve is g at the end of the curve minus g at the beginning. Since the beginning and end are the same, this integral is zero.

Remark: The converse of the previous result is true, and thus if $\int f = 0$ for all closed curves, f must be holomorphic. Indeed

Theorem 38 (Morera's Theorem)**.** *Suppose f is continuous on a domain $\mathcal{U}$. If we have*

$$\int_{\mathcal{R}} f(z)\, dz = 0$$

for all rectangles contained, with their interior, in $\mathcal{U}$, then f in holomorphic in $\mathcal{U}$.

Proof: The integral of $(f_1 + if_2)(dx + idy)$ over any rectangle is zero, so the real integrals of $(f_1, -f_2)$ and (f_2, f_1) over any rectangle is zero. The real variable argument on page 35 then shows that if we fix a rectangle contained with its interior in $\mathcal{U}$, then there exist g_1 and g_2 continuously differentiable on this rectangle with $\text{grad} g_1 = (f_1, -f_2)$ and $\text{grad} g_2 = (f_2, f_1)$. It follows that $\frac{\partial g_1}{\partial x} = \frac{\partial g_2}{\partial y}$ and $\frac{\partial g_1}{\partial y} = -\frac{\partial g_2}{\partial x}$. So $g_1 + ig_2$ is holomorphic with derivative f in the fixed rectangle. But we know that holomorphic functions are infinitely differentiable, so f has a continuous derivative at each point in the fixed rectangle. So f is locally holomorphic, hence globally holomorphic.

6.5 Filling the Gap

At last we are able to prove

Theorem 39. *Suppose $f(z)$ is continuous on a domain $\mathcal{U}$ and $\frac{df}{dz}$ exists at each point of $\mathcal{U}$. Then f is holomorphic in $\mathcal{U}$, i.e., this derivative is continuous.*

Remark: This gap opened up when we proved the Cauchy-Riemann equations assuming only existence of derivatives, but required continuity of the derivative to prove the converse. We have used the converse of the Cauchy-Riemann equations at several points in the above exposition.

To fill the gap, we use the previous theorem. Notice that the proof of that theorem didn't use anything about f except that it is continuous and the integral around all rectangles is zero. So it is enough to prove the following, first proved by Goursat:

Theorem 40. *Let $f(z)$ be continuous on a domain $\mathcal{U}$ containing a closed rectangle, and suppose $\frac{df}{dz}$ exists at each point on an inside the rectangle. Then $\int_{\mathcal{R}} f(z)\,dz = 0$.*

Proof: Suppose this is not true, and $\left|\int_{\mathcal{R}} f(z)\,dz\right| = A > 0$. Divide the rectangle into four pieces as shown below. The sum of the integrals around these subrectangles equals the original integral, so at least one subrectangle must exist whose integral is greater in absolute value than $A/4$.

Subdivide this rectangle similarly. We obtain a sequence of rectangles $R_1 \supset R_2 \supset R_3 \supset ...$, such that $\left|\int_{R_n} f\,dz\right| \geq A/4^n$.

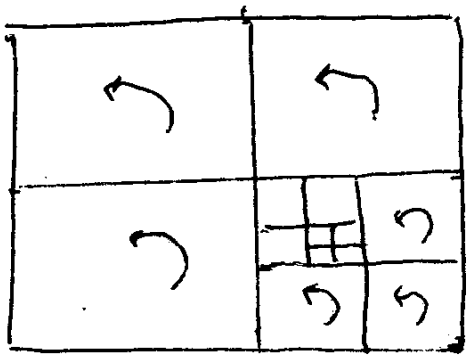


Figure 6.2: Goursat's Argument

By a standard argument, there is a point z_0 in the intersection of these closed rectangles. By assumption, the derivative of f exists at z_0. Thus we can write

$$f(z) = f(z_0) + \frac{df}{dz}\Big|_{z_0}(z - z_0) + \epsilon(z)$$

where

$$\lim_{z \to z_0} \frac{\epsilon(z)}{z - z_0} = 0$$

Therefore, for any positive constant B, we obtain

$$|\epsilon(z)| \leq B|z - z_0|$$

for z sufficiently close to z_0.

But $f(z_0) + \frac{df}{dz}\Big|_{z_0}(z - z_0)$ is trivially holomorphic and its derivative around R_n is zero. Consequently,

$$(\text{maximum of } |\epsilon(z)|)(\text{length of } R_n) \geq \left| \int_{R_n} \epsilon(z) dz \right| > A/4^n$$

Notice that each time we divide into four subrectangles, the length of each of these rectangles is half of the original length. So if the original length is L, then length of $R_n = L/2_n$.

Also, the maximum possible value of $|z - z_0|$ in the original rectangle is the diagonal D of the rectangle. Each time we subdivide into four pieces, this maximum decreases by 2. We conclude that $|\epsilon| \leq BD/2^n$.

Putting this all together,

$$B(D/2^n)(L/2^n) \geq |\epsilon(z)|L/2^n > A/4^n$$

and so $BD > A$. However, A and D are fixed, but B can be chosen as small as we wish by choosen rectangles sufficiently close to z_0, so we have a contradiction.

6.6 Summary

Suppose $f(z)$ is a continuous function on a domain $\mathcal{U}$. We have proved that the following conditions are equivalent:

- f is holomorphic on $\mathcal{U}$
- $\frac{df}{dz}$ exists at each point of $\mathcal{U}$
- $\frac{df}{dz}$ exists at each point of $\mathcal{U}$ and is continuous on $\mathcal{U}$
- $\frac{d^k f}{dz^k}$ exists at each point of $\mathcal{U}$ for all k
- f can be expanded in a power series centered at any $z_0 \in \mathcal{U}$ which converges at least in the largest open disk about z_0 in $\mathcal{U}$
- f is locally $\frac{dg}{dz}$ on $\mathcal{U}$
- $\int_{\mathcal{R}} f(z)dz = 0$ for all rectangles contained, with their interiors, in $\mathcal{U}$
- $f = f_1 + if_2$ where f_1 and f_2 have continuous partial derivatives in $\mathcal{U}$ and satisfy the Cauchy-Riemann equations

Chapter 7

Complex Analysis from a Topological Point of View

This is a chapter for doubters, including those extreme doubters who claim that several variable calculus is hopelessly pictorial and non rigorous. We will reprove everything in the previous chapter from scratch. But don't worry; either the previous proof was sufficient and we'll refer you back to it, or else the previous proof can be dramatically simplified using an powerful version of Cauchy's theorem proved next. In this chapter, a function f is holomorphic on $\mathcal{U}$ if $\frac{df}{dz}$ exists at all points of $\mathcal{U}$; we do not require continuity of this derivative.

The new version of Cauchy's theorem allows us to avoid any mention of Green's theorem. We'll show that this can be done without using fancy results like the Jordan curve theorem, and with complete rigor so every reader will henceforth be willing to deform integration paths with abandon.

The topologists introduce two tools to study domains in the plane: the homotopy group $\pi_1(\mathcal{U})$ and the homology group $H_1(\mathcal{U}, \mathcal{Z})$. Following their lead, we prove a homotopy version of Cauchy's theorem in this chapter. This is followed in the next chapter by a deeper homological version of the theorem.

7.1 Curves and Homotopy

Let $\mathcal{U}$ be a domain in the plane. In this chapter, a *path* is a continuous curve in $\mathcal{U}$. Notice that we do not require differentiability. All of our paths will be parameterized by $[0, 1]$, so a path is a continuous map

$$\gamma : [0, 1] \to \mathcal{U}$$

We say two closed paths are *homotopic* if one can be deformed to the other inside $\mathcal{U}$ through intermediate closed paths. Analogously, two paths which start at p and end at q are *homotopic* if one can be deformed to the other inside $\mathcal{U}$ through intermediate paths which also start at p and end at q. See the pictures below.

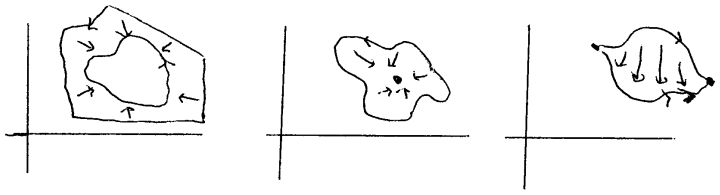


Figure 7.1: Homotopic Paths

Making this rigorous is easy. A *homotopy* between paths $\gamma(t)$ and $\tau(t)$ is a continuous map

$$h(t,u) : [0,1] \times [0,1] \to \mathcal{U}$$

such that $h(t,0) = \gamma(t)$ and $h(t,1) = \tau(t)$. The intermediate paths of the homotopy are obtained by fixing u_0 and writing $h(t,u_0)$.

If γ and τ are closed, we have a *homotopy through closed paths* if $h(0,u) = h(1,u)$ for all u. If γ and τ both start at p and end at q, we have a *homotopy with fixed endpoints* if $h(0,u) = p$ and $h(1,u) = q$ for all u. Usually it is clear which type of homotopy is being used, and we simply say *homotopic*.

It is relatively easy to prove paths homotopic. The following general argument is often enough:

Definition 16. *A domain $\mathcal{U}$ is star-shaped with respect to $p \in \mathcal{U}$ if whenever $q \in \mathcal{U}$, the straight line joining q to p is entirely in $\mathcal{U}$.*

See the pictures which follow.

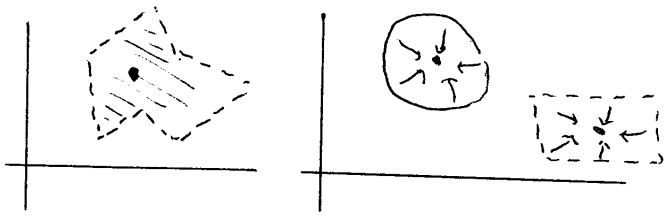


Figure 7.2: Star Shaped Regions

Definition 17. *A domain $\mathcal{U}$ is* simply-connected *if every closed path in $\mathcal{U}$ is homotopic to a constant path, i.e., a point.*

Theorem 41. *A star-shaped region is simply connected, so any closed path is homotopic to a constant.*

Proof: Suppose $\mathcal{U}$ is star-shaped with respect to p. Notice that $h(t,u) = \gamma(t)(1-u) + p(u)$ deforms along lines and thus remains in $\mathcal{U}$. This homotopy starts at γ and ends at p.

Corollary 9. *The following domains are all simply connected*

- *the interior of a rectangle*
- *the interior of a disk*
- *the entire complex plane*

Remark: Simple connectivity is a topological property, so if two sets $\mathcal{U}$ and $\mathcal{V}$ are homeomorphic and one is simply connected, the other is also. This is one way to prove rigorously that a set is simply connected.

For example, $z \to z^6$ maps one-eight of an annulus to three-fourths of an annulus. The first set is star-shaped, hence simply connected. Therefore the second set is simply connected, but not star-shaped.

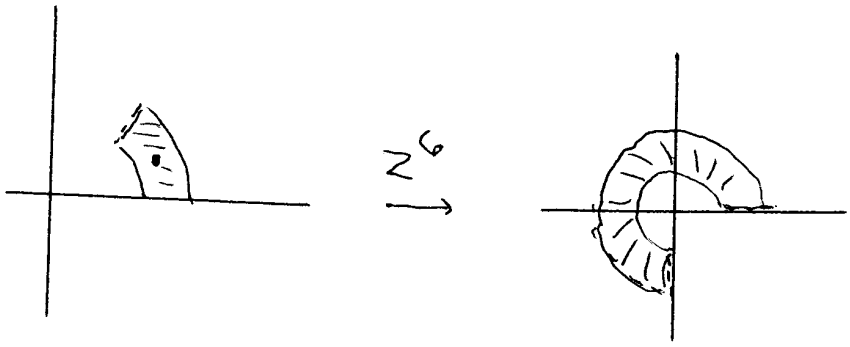


Figure 7.3: Simply Connected Sets

Remark: Finally, here is the big theorem.

Theorem 42 (Homotopy Form of Cauchy's Theorem)**.** *Suppose f is holomorphic on a domain $\mathcal{U}$ and γ and τ are either closed curves in $\mathcal{U}$ or else curves in $\mathcal{U}$ with the same starting and ending points. If γ and τ are homotopic, then*

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

Corollary 10. *If f is holomorphic on a simply connected domain $\mathcal{U}$, then for any closed γ in $\mathcal{U}$,*

$$\int_\gamma f(z)\,dz = 0$$

Remark: It would be unfortunate if we had to restrict attention to piece-wise differentiable homotopies, because we'd constantly need to check differentiability. But remarkably, this isn't necessary. Instead, we'll begin the proof by defining $\int_\gamma f(z)dz$ when γ is merely continuous, and then allow homotopies which are only continuous.

Since continuous curves can have infinite length, and can be space-filling, it seems highly unlikely that we'll be able to define $\int_\gamma f(z)\,dz$ over them. But there is a catch:

IMPORTANT POINT: We previously defined $\int_\gamma f(z)\,dz$ when f is merely continuous on $\mathcal{U}$, but γ is a piecewise-continuously differentiable curve. The new definition will work for all continuous curves, but then $f(z)$ must be holomorphic. The two definitions agree when both are defined.

It will turn out that *defining* the integral over continuous curves is the main difficulty. Once we do that, proving Cauchy's theorem is easy because the same technique is involved. Both tasks ultimately depend on the assertion that if f is holomorphic, then *locally* there is a holomorphic g with $\frac{dg}{dz} = f$, and for completeness we'll reprove that from scratch.

7.2 Holomorphic f Have Local Antiderivatives

We reprove this key assertion from scratch. Suppose f is holomorphic on $\mathcal{U}$ and find an open rectangle $\mathcal{S}$ contained with its interior in $\mathcal{U}$. We only assume that $\frac{df}{dz}$ exists at each point of $\mathcal{S}$.

Turn back to Goursat's theorem in section 6.5, which asserts that the integral of f around any rectangle $\mathcal{R}$ in our $\mathcal{S}$ is zero. This argument doesn't use Green's theorem and remains valid here. So we'll reuse it without repeating the proof.

We can define g on $\mathcal{S}$ by integrating from a fixed point p_0 to z along two sides of a rectangle, either horizontally and then vertically, or else vertically and then horizontally. Goursat's

theorem guarantees that these integrals are equal. See the picture below. We will prove that $\frac{\partial f}{\partial x} = -i\frac{\partial f}{\partial y} = \frac{df}{dz}$. It follows that g has continuous partial derivatives, namely f, and satisfies the Cauchy-Riemann equations, so g is holomorphic with derivative f and we are done.

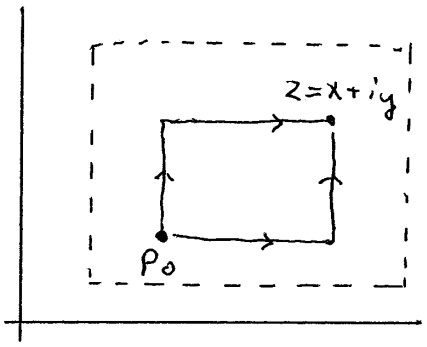


Figure 7.4: $f = \frac{dg}{dz}$ Locally

Take, for instance, the path vertically and then horizontally, illustrated previously. Call the entire path γ and the vertical path γ_1, and parameterize the horizontal path $\tau(t) = t + iy$ for $x_0 \leq t \leq x$. Then

$$g(x+iy) = \int_{\gamma_1} f + \int_{x_0}^{x} f(t,y)dt$$

The first integral is independent of x, so

$$\frac{\partial g}{\partial x} = \frac{d}{dx}\int_{x_0}^{x} f(t,y)\,dt = f(x,y)$$

The equation $\frac{\partial g}{\partial y} = -if(x,y)$ is proved similarly. QED.

7.3 Proof of Homotopy Cauchy Theorem

This proof comes from an elegant short book on complex analysis by Henri Cartan.

We begin by sketching the key idea behind our new definition of $\int_\gamma f(z)\,dz$ for continuous γ and holomorphic f. Consider the special case

$$\int_\gamma \frac{1}{z}\,dz; \qquad \gamma(t) = e^{4\pi i t} : [0,1] \to C - \{0\}$$

It is tempting to integrate using the fundamental theorem of calculus. Thus

$$\frac{d}{dz}\log(z) = \frac{1}{z}; \qquad \log(z) = \ln|z| + i\arg(z)$$

On the unit circle, $\ln|z| = 0$, and $\arg(z)$ increases by 4π as we circle twice. So the integral is 4π.

The trouble with this calculation is that $\log(z)$ is multiply-valued, so it doesn't make sense to treat it as an ordinary holomorphic function on our domain. On the other hand, it does make sense to consider $\log(\gamma(t)) = \varphi(t)$ as a single valued function on $[0,1]$, given by $\varphi(t) = 4\pi t$. This function is continuous, and the key point is that *locally* it equals $g(\gamma(t))$ where g is a branch of the logarithm satisfying $\frac{dg}{dz} = \frac{1}{z}$ and defined on a smaller $\mathcal{V}$ in the domain of $\frac{1}{z}$. So the central idea of the proof is that multiply-valued g become single-valued $\varphi(t) = g(\gamma(t))$.

We'll now sketch the full proof of the homotopy version of Cauchy's theorem. An ambitious reader can fill in the details without further help, but we'll offer a little help at the end. Assume f is holomorphic on $\mathcal{U}$ and γ is a continuous path in $\mathcal{U}$.

A *primitive* for γ is a continuous complex valued function φ on $[0,1]$ which is locally the pull-back of a local g on $\mathcal{V} \subset \mathcal{U}$ with $\frac{dg}{dz} = f$.

Lemma 1. *Such a primitive exists.*

Lemma 2. *A primitive is unique up to a constant.*

Lemma 3. *Consequently, $\varphi(1) - \varphi(0)$ is well defined independent of the chosen primitive. This invariant number is called $\int_\gamma f(z)\,dz$.*

Lemma 4. *If γ is piecewise-continuously differentiable, the old and new definitions of the integral agree.*

Definition 18. *Let $h(t,u) : [0,1]\times[0,1] \to \mathcal{U}$ be a homotopy. A primitive for h is a continuous function φ on $[0,1]\times[0,1]$ which is locally the pullback of a continuous g on $\mathcal{V} \subset \mathcal{U}$ with $\frac{dg}{dz} = f$.*

Lemma 5. *Such a primitive exists.*

Proof of Cauchy's theorem for a homotopy of closed curves: Let $\varphi(t,u)$ be a primitive of h on $[0,1]\times[0,1]$. Then $\varphi(t,0)$ is a primitive for γ and $\varphi(t,1)$ is a primitive for τ. Thus $\int_\gamma f = \varphi(1,0)-\varphi(0,0)$ and $\int_\tau f = \varphi(1,1)-\varphi(0,1)$.

Moreover, $h(0,u)=h(1,u)$ since all curves are closed, so these are the same curve. Therefore $\varphi(0,u)$ and $\varphi(1,u)$ differ by a constant, so $\varphi(1,u)-\varphi(0,u)$ is a constant independent of u. So $\int_\gamma f = \int_\tau f$.

Proof of Cauchy's theorem for a homotopy with fixed endpoints: The proof is the same as the previous proof until the end. All curves start at p and end at q, so $h(0,u)=p$ and $h(1,u)=q$. Since both of these curves are constant, one possible primitive is a constant, so $h\varphi(1,u)=c_1$ and $\varphi(0,u)=c_0$. So $\varphi(1,u)-\varphi(0,u)=c_1-c_0$, and thus $\int_\gamma f = \int_\tau f$.

Remark: PRETTY SIMPLE. But maybe we should give a few more details. Let's start with the definition of a primitive of a curve. So

Definition 19. *A* primitive *for a continuous curve γ in $\mathcal{U}$ and a holomorphic f on $\mathcal{U}$ is a continuous complex-valued φ on $[0,1]$ which is locally the pullback of g such that $\frac{dg}{dz}=f$. This means that for every $t_0\in[0,1]$ there is an open neighborhood $\mathcal{W}$ of $t_0\in[0,1]$ and an open neighborhood $\mathcal{V}$ of $\gamma(t_0)$ in $\mathcal{U}$ and a g on $\mathcal{V}$ with $\frac{dg}{dz}=f$ such that $\varphi(t)=g\circ\gamma(t)$ on $\mathcal{W}$.*

Remark: Please take time to understand this in detail. Everything else will then be easy.

Sketch of Proof of First Lemma: Since g exists locally on $\mathcal{U}$, we can find an open cover $\{\mathcal{W}_\alpha\}$ of $[0,1]$ with a primitive φ_α on each $\mathcal{W}_\alpha$. By compactness, we can find a subdivision $0=t_0<t_1<\ldots<t_n=1$ such that each $[t_{i=1},t_i]$ is in some $\mathcal{W}_\alpha$. So a primitive φ_i exists on $[t_{i-1},t_i]$. Since g is only defined up to a constant, we can add constants to the φ_i at will. Choose these constants to adjust the φ so they match at boundary points t_i. QED.

Sketch of Proof of Second Lemma: Suppose φ_1 and φ_2 are primitives. Locally φ_1 and φ_2 are pullbacks of g_1 and g_2, both with derivative f. By shrinking domains, we can assume that g_1 and g_2 are both defined on the same connected open $\mathcal{V}$. Since their derivatives are equal, they differ by a constant. So $\varphi_2-\varphi_1$ is continuous and locally constant. Since $[0,1]$ is connected, the difference is globally constant. QED.

Sketch of Proof of Fourth Lemma: As earlier, subdivide $[0,1]$ using t_i so φ is the pullback of a single locally defined g on $[t_{i-1},t_i]$. By the fundamental theorem of calculus, the original definition of the integral gives $\int f(z)\,dz = g(\gamma(t_i))-g(\gamma(t_{i-1})=\varphi(t_i)-\varphi(t_{i-1})$ on $[t_{i-1},t_i]$. Both the original integral and the new integral are clearly additive when we subdivide the path, so we are done.

Sketch of the Fifth Lemma: As with an earlier argument, we can subdivide $[0,1]\times[0,1]$ into a grid such that each subrectangle of the grid maps to a $\mathcal{V}$ on which a local g exists. So we obtain primitives φ_{ij} on each subrectangle, which can be modified by adding constants. Starting at top-left, and working across the rows, and then down row by row, adjust these constants so

the φ_{ij} match on boundaries, to obtain a globally defined φ. It is instructive to work through this and convince yourself that these adjustments are possible. Two tricky situations can occur. These are illustrated below. In each case, a new subrectangle is being added to the diagram. The black curve defines a parameterized path with two primitives, namely the primitive defined by the already constructed piece of h and the primitive from the new subrectangle. So these primitive differ by a constant and we can adjust the new primitive for the new subrectangle by adding a constant to make it continuous across the black boundary.

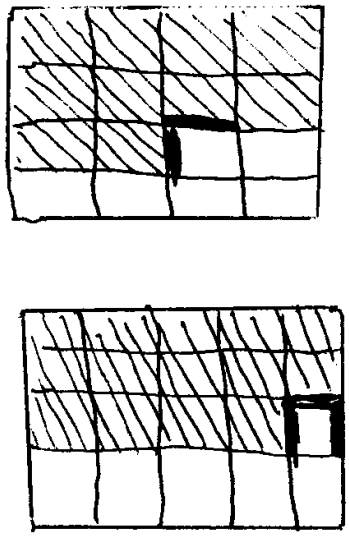


Figure 7.5: A Primitive for the Homotopy

7.4 Chapter 6 Done Rigorously

We now know enough to reprove everything in chapter 6 rigorously. We have to do this in the correct order.

The first theorem in chapter 6 is Cauchy's theorem that $\int_{\partial D} f(z)\,dz = 0$ if f is holomorphic on $\mathcal{U}$ containing the disk D. We already know a very substantial generalization, namely that we can replace ∂D by any closed curve in a simply connected region. Moreover, our original proof required continuity of $\frac{df}{dz}$ and our current proof doesn't need this hypothesis.

The second theorem in chapter 6 is Cauchy's formula $f(z) = \int_{\partial D} \frac{f(\zeta)}{\zeta - z}\,dz$ if f is holomorphic on $\mathcal{U}$ containing the disk D and z is inside this disk. Look back and notice that our original proof used Green's theorem to prove that the integral of $\frac{f(\zeta)-f(z)}{\zeta - z}$ over the disk is equal to the integral over a small circle about z. The quotient expression is continuous at z but may not be holomorphic there, but the boundary of D and the small circle are homotopic by a homotopy which avoids z. So these integrals are equal. The original proof then showed that the integral around the small circle approaches zero. It follows that the integral of $\frac{f(\zeta)-f(z)}{\zeta - z}$ around D is zero. From this point, the original proof still works and easily gives Cauchy's formula.

Next we proved theorem 27 that a holomorphic function can be expanded in a power series. This proof works unchanged, except that the requirement that f have a continuous derivative is no longer needed.

The easy proof of theorem 30 still works. But theorem 31 and theorem 32 can be replaced by a powerful generalization:

Theorem 43. *Let f be holomorphic on a simply connected $\mathcal{U}$. Then there is a global g on $\mathcal{U}$ with $\frac{dg}{dz} = g$.*

Proof: Fix $p \in \mathcal{U}$ and for each $q \in \mathcal{U}$, define $f(z) = \int_\gamma f(z)\,dz$. This f does not depend on the path γ because the integral of f around any continuous closed path is zero by the homotopic Cauchy theorem. The argument used in section 7.2 then shows that $\frac{dg}{dz} = f$.

This completes our redo of chapter 6. The previous theorem has powerful corollaries, as well:

Corollary 11. *If $\mathcal{U}$ is a simply connected domain which does not contain* 0*, it is possible to define a branch of the logarithm on $\mathcal{U}$.*

Proof: Use the theorem to find g on $\mathcal{U}$ with $\frac{dg}{dz} = \frac{1}{z}$. Then

$$\frac{d}{dz}\left(\frac{e^g}{z}\right) = \frac{ze^g\frac{1}{z} - e^g}{z^2} = 0$$

Consequently $e^g(z) = cz$ for some constant $c \neq 0$. Write $c = e^\lambda$. Then $g - \lambda$ is a branch of the logarithm because

$$e^{g-\lambda} = e^g e^{-\lambda} = e^\lambda z e^{-\lambda} = z$$

Corollary 12. *If $\mathcal{U}$ is a simply connected domain which does not contain* 0*, it is possible to choose a single-valued branch of $\sqrt{z}$ on $\mathcal{U}$.*

Proof: Select $\sqrt{z} = e^{\frac{1}{2}\text{Log}(z)}$.

Corollary 13. *If $f(z)$ is holomorphic and never zero on a simply connected domain U, it is possible to define a branch of Log$(f(z))$ on $\mathcal{U}$.*

Proof: It is easy to find examples where $f(\mathcal{U})$ is no longer simply connected, so this does not follow from an earlier corollary.

If such a branch exists, its derivative will be $\frac{f'(z)}{f(z)}$. This function is holomorphic because f has no zeros on $\mathcal{U}$. Find g with $\frac{dg}{dz} = \frac{f'}{f}$. Then

$$\frac{d}{dz}\left(\frac{e^{g(z)}}{f(z)}\right) = \frac{fe^g\left(\frac{f'}{f}\right) - e^g f'}{f^2} = 0$$

So this function is a (nonzero) constant e^λ and $e^{g(z)} = f(z)e^\lambda$. Thus

$$e^{g(z)-\lambda} = f(z)$$

and so $g(z) - \lambda$ is a branch of Log$(f(z))$.

Corollary 14. *If $f(z)$ is holomorphic and never zero on a simply connected domain $\mathcal{U}$, it is possible to define a branch of $\sqrt{f(z)}$ on $\mathcal{U}$.*

Proof:

$$\sqrt{f(z)} = e^{\frac{1}{2}Log(f(z))}$$

Chapter 8

Winding Numbers, Homology, and Cauchy's Theorem

8.1 Winding Numbers

Previously we talked about domains with *holes*, always accompanied with a picture. We'll finally be able to define this rigorously.

The development which follows was introduced into complex variables by Emil Artin, in his lectures on the subject.

Definition 20. *Fix a point z_0. Let γ be a closed curve which does not contain z_0. The* winding number *of γ about z_0 is*

$$W(\gamma, z_0) = \frac{1}{2\pi i} \int_\gamma \frac{1}{\zeta - z_0} d\zeta$$

Intuitively, this expression counts the number of times γ goes around z_0 counterclockwise. Clockwise motions are counted negatively.

Theorem 44.

- *$W(\gamma, z_0)$ is an integer*
- *If γ and τ are homotopic through closed curves which all miss z_0, then $W(\gamma, z_0) = W(\tau, z_0)$*
- *The winding number of the circular path $\gamma(t) = z_0 + Re^{2\pi int}$ is n.*

Proof: The second result follows from Cauchy's theorem, and the third is an easy calculation. We need only prove the first item.

Let φ be a primitive for the curve γ and function $\frac{1}{z-z_0}$. Then φ is locally $g\circ\gamma$ where $\frac{dg}{dz} = \frac{1}{z-z_0}$, and so $\frac{e^{\varphi}}{\gamma-z_0}$ is locally $\frac{e^g}{z-z_0}\circ\gamma$, where $\frac{d}{dz}\left(\frac{e^g}{z-z_0}\right) = 0$. It follows that $\frac{e^{\varphi}}{\gamma-z_0}$ is constant.

Since $\gamma(0) = \gamma(1)$, we conclude that $e^{\varphi(0)} = e^{\varphi(1)}$ and therefore $e^{\varphi(1)-\varphi(0)} = 1$. So $\varphi(1)-\varphi(0)$ is an integer multiple of $2\pi i$ and consequently $\frac{1}{2\pi i}\int_\gamma \frac{1}{z-z_0}\,dz = \frac{1}{2\pi i}(\varphi(1)-\varphi(0))$ is an integer.

Remark: As a corollary, we can prove the fundamental theorem of algebra. There are many proofs of this theorem, but the proof we give is the proof that can be explained to a layman, and thus in a sense the correct proof.

Let $P(z) = z^n + c_1 z^{n-1} + ... + c_n$ be a polynomial over the complex numbers, $n > 0$. We claim this polynomial takes the value 0. If this is false, then for any positive R, $P(Re^{2\pi it})$ is a closed path which avoids 0. Also $H(t,u) = P(Rue^{2\pi it})$ is a homotopy from this closed path to a constant path, and thus $W(P(Re^{2\pi it}), 0) = 0$.

We claim that for R large enough, $P(Re^{2\pi it})$ is homotopic to $Re^{2\pi int}$ by a homotopy avoiding the origin. Since this last path has winding number n, we have a contradiction.

The required homotopy is

$$h(t,u) = \left(Re^{2\pi it}\right)^n + uc_1\left(Re^{2\pi it}\right)^{n-1} + ... + uc_n$$

To complete the proof, we must prove that this express is never zero for R large enough. If it were zero, then

$$\left(Re^{2\pi it}\right)^n = -\left[uc_1\left(Re^{2\pi it}\right)^{n-1} + ... + uc_n\right]$$

for some t. The absolute value of the left side is R^n and the absolute value of the right side is at most $|c_1|R^{n-1} + ... + |c_n| \leq \max\{|c_1|, ..., |c_n|\}R^{n-1}$, so we have a contradiction as soon as $R > \max\{|c_1|, ..., |c_n|\}$.

8.2 The Inside of a Curve

A closed curve $\gamma : [0,1] \to \mathcal{U}$ is said to be simple if γ does not cross itself, that is, $\gamma(t_1) = \gamma(t_2)$ only if $t_1 = t_2$ or one of t_1, t_2 is 0 and the other is 1. According to the Jordan curve theorem, the complement of a simple curve in the plane has exactly two components, one bounded and one unbounded. The bounded component is called the interior of the curve.

It used to be believed that this theorem was necessary for a rigorous treatment of complex analysis. But we can use the winding number to avoid it. Suppose a curve is continuous and closed, but not necessarily simple. We say that a point z_0 is *inside* the curve if it is not on the curve and if $W(\gamma, z_0) \neq 0$.

Theorem 45. *The complement of a closed curve is a union of open connected components. Moreover*

1. *$W(\gamma, z)$ is constant on components of the complement*
2. *Exactly one component is unbounded, and $W(\gamma, z)$ is zero on this component.*
3. *The union of a closed curve and all points inside it is compact.*

Remark: For example, consider the curve below. For each component of the complex, compute the Winding Number of the curve about points in that component. Hint: It suffices to fix a point in the component and then create a homotopy of the curve, which doesn't pass through that point, to a much easier curve. Far away curlicues no longer matter.

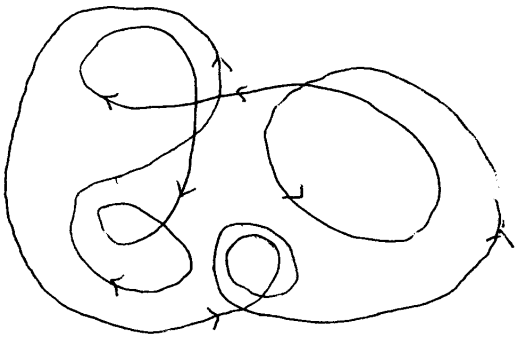


Figure 8.1: A Complicated Curve

Proof: Connected open sets are arcwise connected. So if z_1 and z_2, are points in the same component, we can find a path τ in this component joining z_1 to z_2. We claim that $W(\gamma, \tau(u))$ is constant in u, proving the first assertion.

This follows from the following principle. The map $z \to z + a$ translates the entire plane by a. Equivalently, we can leave the plane fixed and translate functions using the formula $f(z) \to f(z - a)$. Then integrating $f(z)$ around γ gives the same result as integrating $f(z - a)$ around $\gamma + a$, because these equations translate both the function and the path by a.

Sincee $W(\gamma, \tau(u)) = \frac{1}{2\pi i} \int_\gamma \frac{d\zeta}{\zeta - \tau(u)}$, this winding number is equal to $W(\gamma + \tau(u), 0)$. We now have a homotopy of curves, and so W doesn't change by Cauchy's theorem.

Since $\gamma([0, 1])$ is compact, it is bounded. Hence there is a positive real number R such that $\{z \mid |z| > R\}$ does not intersect the curve, and thus must be included in one component of the compliment. That component is the only unbounded component. If z is any point in this infinite ring, $W(\gamma, z) = 0$ because γ is homotopic to a constant curve by a homotopy that avoids z, since $\{z \mid |z| \leq R\}$ is contractible.

The union of the curve and all points inside it is compact because it is closed and bounded. it is closed because its compliment is the union of the remaining open components of $C - \gamma$. It is bounded because the compliment contains the unbounded component of $C - \gamma$.

8.3 The Homological Cauchy Theorem

Theorem 46 (Homological Cauchy Theorem)**.** *Suppose $f(z)$ is holomorphic on a domain $\mathcal{U}$. Let $\gamma_1, \ldots, \gamma_k$ be a finite number of closed curves in $\mathcal{U}$. Suppose $\sum W(\gamma_i, z_0) \neq 0$ implies $z_0 \in \mathcal{U}$. Then $\sum \int_{\gamma_i} f(z)\, dz = 0$.*

Corollary 15. *Let $f(z)$ be holomorphic on a domain $\mathcal{U}$. Let γ be a closed curve in $\mathcal{U}$ such that every point inside γ is in $\mathcal{U}$. Then $\int_\gamma f(z)\, dz = 0$.*

Important Remark: From now on, when we apply this result we will say *γ is a closed curve* instead of *$\gamma_, \ldots, \gamma_i$ are closed curves*, and we will say *every point inside γ is in $\mathcal{U}$* instead of *$\sum W(\gamma_i, z_0) \neq 0$ implies $z_0 \in \mathcal{U}$*.

Example: The picture below shows a domain $\mathcal{U}$ with two holes, and a continuous curve. This curve is not homotopic to a constant in $\mathcal{U}$. However, $W(\gamma, z_0) = 0$ for points in the holes, because γ goes around each hole once clockwise and once counterclockwise. Consequently, $\int_\gamma f(z)\, dz = 0$ if f is holomorphic on $\mathcal{U}$. This example shows that the homological version of Cauchy's theorem is more powerful than the homotopy version.

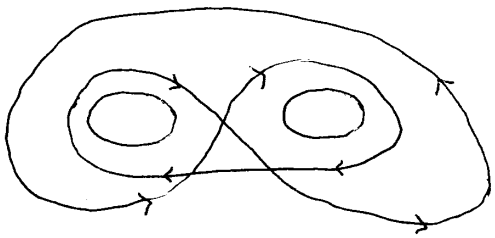


Figure 8.2: A Curve with Zero Integral.

Example: The illustration below shows an application of the Homological Cauchy Theorem to a region with three boundary curves. If f is holomorphic on the shaded region, the sum of the integrals around these three curves is zero by the theorem, even if f is not defined on the holes, because the sum of the Winding Numbers of the three boundary curves is zero inside either hole.

This sort of application will be crucial when we prove the Residue Theorem in a future chapter.

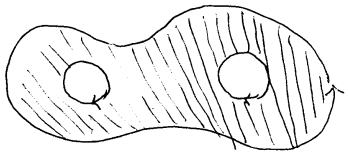


Figure 8.3: A Region with Three Boundary Curves

Remark: We now prove the theorem. This proof was given around 1960 in a complex variable course at Harvard by George Mackey. The details of this proof may be due to Mackey.

Proof: By compactness, we can find ϵ such that whenever $\gamma_i(t)$ is a point on one of the curves, the open disk of radius 5ϵ about this point is in $\mathcal{U}$. (The 5 gives us room to wiggle in a minute.) The set of open disks of radius ϵ forms a cover of the union of the curves, so we can find a common decomposition $0 = t_0 < t_1 < ... < t_n = 1$ such that each $\gamma_i([t_{j-1}, t_j])$ is in one such disk of radius ϵ.

For each i and j, draw this disk and find a homotopy within the larger disk of radius 5ϵ to a curve with one horizontal and one vertical segment, as illustrated below. Replace the original γ_i by the resulting polygonal replacements.

Figure 8.4: Small Path Adjustment

We claim it will be enough to prove the theorem for the new curves. Notice that they still

satisfy the key assumption that $\sum W(\gamma_i, z_0) \neq 0$ implies $z_0 \in \mathcal{U}$, for if such a $z_0 \notin \mathcal{U}$, then the homotopy between the original γ_i and the new polygonal γ_i would avoid z_0 and thus not change the sum of the winding numbers. If we could prove the conclusion of the theorem for the polygonal paths, then any holomorphic function f on $\mathcal{U}$ would satisfy $\sum \int_{\gamma_i} f(z)\, dz = 0$ for the new γ_i. But again, the homotopies were done within $\mathcal{U}$, so this sum of integrals remains unchanged it we returned to the original γ_i.

Think of the endpoints of the polygonal paths as vertices of an irregular lattice, and draw a rectangular piece of that lattice. In some cases, horizontal or vertical segments may need to be subdivided to allow this. We will draw the lattice as almost square, but in practice it may be very irregular, with for instance, long narrow rectangles.

Below is a picture with two polygonal paths indicated. Note that some of the subrectangles of the diagram will be in $\mathcal{U}$ and some not, since $\mathcal{U}$ is not shown.

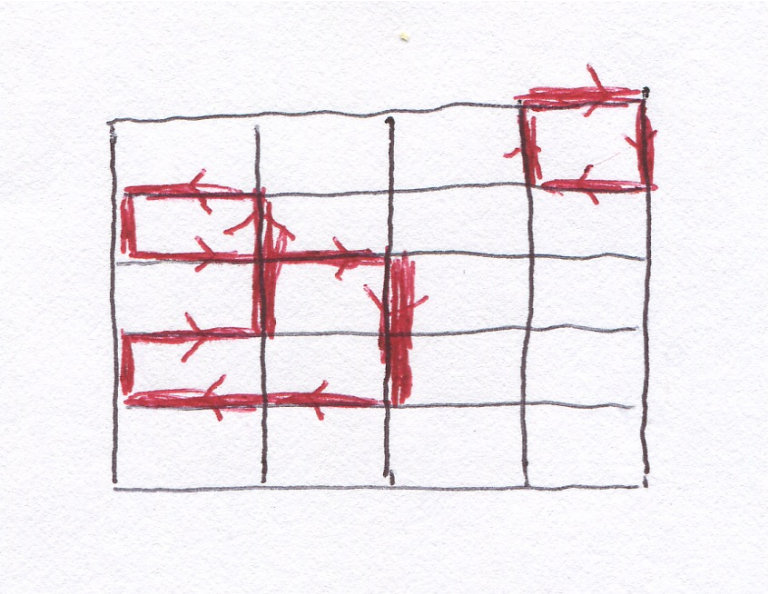


Figure 8.5: Main Diagram of the Proof

In the above special case, the remaining argument is easy. If z_0 is inside the top right rectangle, then $\sum W(\gamma_i, z_0) = -1 \neq 0$, and so $z_0 \in \mathcal{U}$. Thus we can apply Cauchy's theorem to conclude that $\int f(z)\, dz = 0$ around this rectangle.

The lower left of the diagram shows a path surrounding four rectangles. The winding number sum is 1 for one rectangle, and -1 for the remaining three, so the interiors are in $\mathcal{U}$. Break these paths into two paths. By Cauchy's theorem, both integrals vanish. The picture suggests more, namely that we could break the bottom left path into a sum of rectangles. That's shown

in the full proof.

The general case of the proof is really a clever combinatorial argument rather than analysis. Let h_{ij} be formal symbols indicating the horizontal segments with arrows pointing right, and let v_{ij} be formal symbols indicating the vertical segments with arrows pointing up. Form the abelian group $\mathcal{G}$ over the integers generated by these symbols. An element of this group is a formal sum $\sum m_{ij}h_{ij} + \sum n_{ij}v_{ij}$. We can think of elements of this group as general paths, not necessarily connected or closed. Each of our closed polygonal paths σ_i induces elements of this group in an obvious way. Notice that the σ_i could use individual segments more than once, so elements of the form $3h_{11} - 2v_{23}$ are possible. We define integration over an element of the group in the obvious way.

Each subrectangle of the grid induces an obvious element of $\mathcal{G}$ formed by the formal symbols for the four sides of the rectangle with appropriate signs. Denote these elements R_{ij}.

Consider the expression $\sum W_{ij}R_{ij} \in \mathcal{G}$ where $W_{ij} = \sum W(\sigma_i, z_0)$ for some z_0 inside R_{ij}. In other words, we form a formal sum of rectangles, each occurring as many times as the original polygonal paths went around the rectangle. Notice that only rectangles in $\mathcal{U}$ play a role in this expression, by the basic assumption of our theorem.

The combinational lemma we need states that this sum of rectangles equals the sum of the original polygonal σ_i in $\mathcal{G}$. This lemma will clearly prove the theorem.

To prove the result, form

$$\sigma = \sum W_{ij}R_{ij} - \text{formal sum corresponding to polygonal path}$$

This is an element of $\mathcal{G}$ and we would like to prove that it is zero, i.e., that the coefficients of each h_{ij} and each v_{ij} are zero.

We'll give the argument for a horizontal segment h_{ij}. Find a sequence z_n converging to the center of this segment from above and a second sequence w_n converging to the center from below.

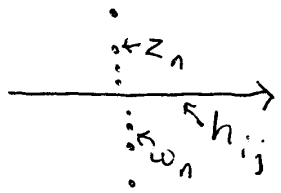


Figure 8.6: A Limit

By definition of σ, the integral of $\frac{1}{z-z_n}$ over σ and the integral of $\frac{1}{z-w_n}$ over σ are both zero. So the integral of

$$\frac{1}{z-z_n} - \frac{1}{z-w_n}$$

over σ is zero. On the other hand, this integrand goes uniformly to zero on all segments except h_{ij}. So if the coefficient of h_{ij} in σ is not zero, we conclude that the limit of

$$\int_{h_{ij}} \left(\frac{1}{z-z_n} - \frac{1}{z-w_n} \right) dz$$

is zero.

On the other hand, consider the the same integrand over the path below.

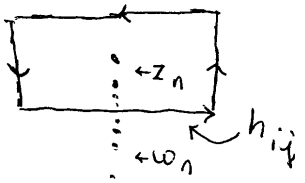


Figure 8.7: Another Limit

By a winding number argument, the value of the integral over this path is $2\pi i$. Again, the integrand converges uniformly to zero on all sides of the rectangle except h_{ij}. So we conclude that the limit of

$$\int_{h_{ij}} \left(\frac{1}{z-z_n} - \frac{1}{z-w_n} \right) dz$$

is $2\pi i$. This contradiction proves the result.

Chapter 9

Isolated Singularities, Laurent Expansions, and the Residue Theorem

9.1 The Identity Theorem

We mentioned the following theorem in chapter two. It is time to give the details.

Theorem 47 (The Identity Theorem)**.** *Suppose f and g are holomorphic on a domain $\mathcal{U}$. If $\mathcal{U}$ intersects R and $f = g$ on this intersection, then $f = g$ always. More generally, if there is a sequence $z_n \in \mathcal{U}$ with a limit $z_0 \in \mathcal{U}$ and $f(z_n) = g(z_n)$, then $f = g$ always*

Proof: It suffices to prove the second result. Let $h(z) = f(z) - g(z)$ and note that $h(z_n) = 0$. Expand h in a power series about z_0

$$h(z) = c_0 + c_1(z - z_0) + c_2(z - z_0)^2 + ...$$

By continuity, $h(z_n) \to h(z_0)$, so $h(z_0) = c_0 = 0$.

Let $k(z) = h(z)/(z - z_0)$ and notice that $k(z_n) = 0$ and k is holomorphic near z_0 because

$$k(z) = c_1 + c_2(z - z_0) + ...$$

By continuity, $k(z_n) \to k(z_0)$, so $c_1 = 0$. Continue. It follows that the power series is identically zero and thus that $h(z)$ is zero on an open neighborhood of z_0.

Consider the set $\mathcal{V}$ of all points $z_0 \in \mathcal{U}$ such that $h = 0$ in an open disk about z_0. This set is trivially open; by the previous paragraph it is not empty. We will prove it closed, so connectedness of $\mathcal{U}$ implies that it is all of $\mathcal{U}$. Suppose $z_n \in \mathcal{V}$ and $z_n \to z_0 \in \mathcal{U}$. Then $h(z_n) = 0$ and the previous argument shows that h is identically zero near z_0.

9.2 Liouville's Theorem

The following beautiful theorem has many consequences. We prove it here because we need it later in the chapter.

Definition 21. *An* entire function *is a function defined and holomorphic on the entire complex plane.*

Theorem 48 (Liouville)**.** *A bounded entire function is constant.*

Proof: Turn back to the proof that a holomorphic function can be expanded in a power series. According to that proof, our entire function has a power series expansion

$$f(z) = \sum_{k=0}^{\infty} \left[\frac{1}{2\pi i} \int_\gamma \frac{f(\zeta)}{\zeta^{k+1}}\, d\zeta \right] z^k$$

valid in the entire plane. Since power series are given by

$$\sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} z^k$$

we conclude that

$$f^{(k)}(0) = \frac{k!}{2\pi i} \int_\gamma \frac{f(\zeta)}{\zeta^{k+1}}\, d\zeta$$

Recall that the absolute value of any complex integral $\int_\gamma f(z)\, dz$ is less than or equal to the maximum of $|f|$ on γ times the length of γ. If f is bounded by B, then the maximum value of $|f|$ on the circle of radius R about the origin is B. The length of this circle is $2\pi R$. We conclude that

$$\left| f^{(k)}(0) \right| \leq \frac{k!}{2\pi} \frac{B}{R^{k+1}} 2\pi R = \frac{k! B}{R^k}$$

As $R \to \infty$, this expression goes to zero for $k \geq 1$, so the power series expansion of f about the origin is just $f(0)$.

Corollary 16 (The Fundamental Theorem of Algebra)**.** *If $P(z)$ is a polynomial of degree at least one, then P has a complex root.*

Proof: We have already proved this, but Liouville's theorem provides a second proof. If P is never zero, then $f(z) = \frac{1}{P(z)}$ is an entire function. But

$$|P(z)| = |z^n + c_1 z^{n-1} + \ldots + c_n| \geq |z|^n \left(1 - \frac{|c_1|}{|z|} - \ldots - \frac{|c_n|}{|z|^{n-1}} \right)$$

and this inequality immediately proves that $\frac{1}{P}$ goes to zero for large z and thus is a bounded entire function, hence constant, a contradiction.

Corollary 17. *Let $f(z)$ be a non-constant entire function. Then the range of f is dense in C.*

Proof: Suppose z_0 is not in the closure of the range. Then $\frac{1}{f(z)-z_0}$ is bounded, hence constant, which implies that f is constant.

Remark: As we remarked earlier, Picard strengthened this by proving that an entire function can omit at most one complex value. We will give his proof later on.

9.3 Isolated Singularities

Suppose f and g are holomorphic on a domain $\mathcal{U}$. If g is not identically zero, it is natural to divide and form $\frac{f}{g}$. This function is defined and holomorphic except at zeros of g, and the identity theorem asserts that these are isolated points. So $\frac{f}{g}$ is a function with isolated singularities.

Definition 22. *Let $\mathcal{U}$ be an open set containing z_0 and let f be holomorphic on $\mathcal{U} - \{z_0\}$. Then we say that z_0 is an* isolated singularity *of f.*

Remark: There is an amazing theory of isolated singularities. One possibility is illustrated by $\frac{\sin z}{z}$. Since z vanishes at the original, the origin is an isolated singularity. But actually,

$$\frac{\sin z}{z} = \frac{z - \frac{z^3}{3!} + \frac{z^5}{5!} - \ldots}{z} = 1 - \frac{z^2}{3!} + \frac{z^4}{5!} - \ldots$$

and the singularity can be eliminating by defining the quotient $\frac{\sin z}{z}$ to be 1 at the origin. We say call such a singularity a *removable singularity*.

Another possibility illustrated by $\frac{1}{z}$. Notice that $\lim_{z\to 0}\left|\frac{1}{z}\right| = \infty$. A singularity with this property is called a *pole*. It turns out that isolated singularities formed as quotients of holomorphic functions like $\frac{f}{g}$ are always removable or poles.

A third possibility is illustrated by $f(z) = e^{1/z}$. If $z_n = -\frac{1}{n}$, the z_n converge to the origin and $f(z) \to 0$. On the other hand, if $z_n = \frac{1}{n}$, the $z_n \to 0$ and yet $f(z_n) \to \infty$. So this singularity is not removable and not a pole. It turns out that for any complex number λ, there is a sequence $z_n \to 0$ with $f(z_n) \to \lambda$. Such a singularity is called an *essential singularity*.

Astonishingly, these are the only possibilities. All sorts of other possibilities suggest themselves. In ordinary calculus, $f(x) = |x|$ is differentiable except at the origin, but only continuous at the origin. This cannot occur in complex analysis, where isolated points of mere continuity are always removable singularities. Similarly, in real variable theory, $f(x) = x/|x|$

is differentiable except at the origin, but not even continuous at the origin. This is also impossible in complex analysis; an isolated singularity such that f is bounded nearby is automatically removable. These results will be proved using the theory of Laurent series, discussed next.

9.4 Newton's Calculus Paper

In 1665, there was a Plague in England and Cambridge University closed for two years. Newton returned to his family's farm, where he had his initial ideas on calculus, gravity, and the theory of light.

Newton continued to work on calculus after he returned to Cambridge. He religiously kept his scratch work, which can be found in volume one of Derrick Whiteside's eight volume edition of Newton's Mathematical Papers. In 1669, Mercator rediscovered one consequence of Newton's work, and Newton's teacher, Isaac Barrow, told him that he would lose priority if he didn't make his work known. So Newton wrote the first public paper on calculus, *De Analysi Per Aequationes Infinitas* or *On analysis by infinite equations*.

The paper was not actually published then, but it was deposited in the archives of the Royal Society in London, where visiting scholars including Leibniz read it. Newton preferred to write a complete book on the calculus, which he finished in 1673. Unfortunately, complications arose and the book was not published until after Newton's death. *De Analysis* was published in 1711; by that time, Newton was a famous man.

The main topic of *De Analysi* is integration theory, and Newton usually integrates by expanding functions in infinite series and integrating term by term. For example, the first substantial example in the paper is the function $y = \frac{a^2}{b+x}$, which Newton expands by straightforward division. Nowadays we would concentrate on the special case

$$\frac{1}{1+x} = 1 - x + x^2 - x^3 + ...$$

Newton integrates term by term to obtain

$$\ln x = x - \frac{x^2}{2} + \frac{x^3}{3} - ...$$

In private papers, Newton used this result to find ln(1.02) to 72 decimal places, and later he wrote to Leibniz "I am ashamed of how much time I wasted as a young man calculating functions to many decimal places."

Immediately after this example, Newton considers $y = \frac{1}{1+x^2}$. This time he notes that there are two ways to divide, suggested by writing the denominator as $1 + x^2$ or as $x^2 + 1$. This gives two

series expansions, and when they are integrating term by term, it gives two results

$$\int \frac{1}{1+x^2} = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + ...$$

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

Newton then writes "Proceed by the former way when x is small enough, by the latter when it is taken large enough."

A few pages later, Newton integrates implicit functions via the same technique. His example is $y^3 + axy + x^2y - a^3 - 2x^3 = 0$. Solving for $y(x)$, he obtains

$$y = x - \frac{1}{4}a + \frac{a^2}{64x} + \frac{131a^3}{512x^2} + \frac{509a^4}{32768x^2} + ...$$

$$\int y(x) = \frac{x^2}{2} - \frac{1}{4}ax + \int \frac{a^2}{64x}\,dx + \frac{131a^3}{512x} + \frac{509a^4}{32768x^2} + ...$$

Incidentally, the integral sign is due to Leibniz and Newton didn't know that symbol when he wrote this paper. Newton just says in words "the area is", until this last formula. Since he cannot integrate $\frac{1}{x}$, he puts a box around the term $\frac{a^2}{64x}$ to indicate that it would have to be evaluated using the numerical techniques from earlier in the paper.

There is much more of interest in this first paper. Newton gives ample examples, which makes for easy reading.

The surprising thing is that Newton introduces series containing powers of both x and $\frac{1}{x}$, which are now called *Laurent series*, and he already knows that the $\frac{1}{x}$ term presents special problems, which will lead us to the notion of the *residue of a singularity*.

9.5 Laurent Series

Definition 23. *A* Laurent Series *is an expansion*

$$\sum_{k=-\infty}^{\infty} c_k(z-z_0)^k = ... + c_{-2}\frac{1}{(z-z_0)^2} + c_{-1}\frac{1}{(z-z_0)} + c_0 + c_1(z-z_0) + c_2(z-z_0)^2 + ...$$

Remark: Given such a series, we let

$$f_1(z) = \sum_{k=0}^{\infty} c_k z^k \text{ for } |z| < R$$

and

$$f_2(z) = \sum_{k=1}^{\infty} c_k z^k \text{ for } |z| < \frac{1}{r}$$

and call the sum of the series the function

$$f(z) = f_1(z - z_0) + f_2\left(\frac{1}{z - z_0}\right) \text{ for } r < |z - z_0| < R$$

Thus a Laurent series converges on a (possibly empty) annulus about z_0. For simplicity of notation, we assume $z_0 = 0$ in the rest of this section, but all of our results obviously apply in the more general case.

Example: It is easy to find such series using essentially Newton's method. Take, for instance, the function $f(z) = \frac{1}{(z-1)(z-2)}$. This can be rewritten using partial fractions as $-\frac{1}{z-1} + \frac{1}{z-2}$. We can divide the plane into three annuli about the origin: $|z| < 1$, $1 < |z| < 2$, and $2 < |z|$. We get three different Laurent expansions as follows. We have

$$\frac{1}{1-z} = 1 + z + z^2 + z^3 + \ldots \qquad |z| < 1$$

$$\frac{1}{1-z} = -\frac{1}{z} - \frac{1}{z^2} - \frac{1}{z^3} - \ldots \qquad |z| > 1$$

$$\frac{1}{z-2} = -\left(\frac{1}{2}\right)\frac{1}{1-\frac{z}{2}} = -\frac{1}{2} - \frac{1}{2^2 z} - \frac{1}{2^3 z^2} - \ldots \qquad |z| < 2$$

$$\frac{1}{z-2} = \left(\frac{1}{z}\right)\frac{1}{1-\frac{2}{z}} = \frac{1}{z} + \frac{2}{z^2} + \frac{2^2}{z^3} + \ldots \qquad |z| > 2$$

and the final series are obtained in the obvious way from these.

Theorem 49.

- *If $f(z)$ is holomorphic in an annulus about the origin, there is a Laurent series which converges to $f(z)$ at least in this annulus.*
- *This series is unique*
- *The series can be differentiated term by term in the annulus*
- *The series can be integrated term by term in the annulus, modulo the standard problems with integrating $\frac{1}{z}$*

Remark: The proof of existence is similar to the proof that holomorphic functions have power series expansions. Fix z_0 in the annulus. Let $g(z)$ be defined on the annulus by $g(z) = \frac{f(z)-f(z_0)}{z-z_0}$. This function is holomorphic on the annulus except possibly at z_0. But expanding f as a power series about z_0 gives $f(z) = \sum_{k\geq 0} c_k(z-z_0)^k$ so that

$$g(z) = \frac{\sum_{k\geq 0} c_k(z-z_0)^k - c_0}{z-z_0} = \sum_{k\geq 1} c_k(z-z_0)^{k-1}$$

Thus g is also holomorphic at z_0.

Draw two curves in the annulus as indicated below, and form the integral

$$\frac{1}{2\pi i}\int \frac{f(\zeta)-f(z)}{\zeta - z}\, d\zeta$$

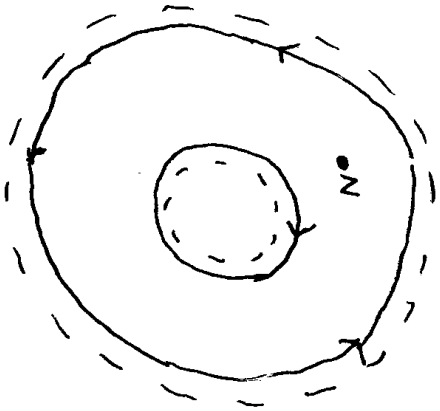


Figure 9.1: Existence of Laurent Series

We assume that z is between the two circular paths. Apply the homological Cauchy theorem, noting that the total winding number is zero on the hole inside the annulus. Consequently this integral is zero and

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

If ζ is on the large circular path γ_1, then $\left|\frac{z}{\zeta}\right| < 1$, so we can write

$$\frac{1}{\zeta - z} = \frac{1}{\zeta}\frac{1}{1-\frac{z}{\zeta}} = \sum_{k=0}^{\infty} \frac{z^k}{\zeta^{k+1}}$$

If ζ is on the small circular path γ_2, then $\left|\frac{\zeta}{z}\right| < 1$ and we can write

$$\frac{1}{\zeta - z} = -\frac{1}{z}\frac{1}{1 - \frac{\zeta}{z}} = -\sum_{k=0}^{\infty} \frac{\zeta^k}{z^{k+1}}$$

We conclude that

$$f(z) = -\sum_{k=0}^{\infty} \left[\frac{1}{2\pi i} \int_{\gamma_2} f(\zeta)\zeta^k d\zeta \right] \frac{1}{z^{k+1}} + \sum_{k=0}^{\infty} \left[\frac{1}{2\pi i} \int_{\gamma_1} \frac{f(\zeta)}{\zeta^{k+1}} d\zeta \right] z^k$$

This gives a Laurent expansion for z between the two circles. But we can replace the inner circle by a smaller homotopic circle, and the outer circle by a larger homotopic circle, to get a formula that is valid for more z. By Cauchy's homotopy theorem, the coefficients do not change. We conclude that the Laurent series converges to f throughout the entire annulus where f is defined.

Proof, continued: To prove the Laurent series unique, suppose $f_1(z) + f_2\left(\frac{1}{z}\right)$ and $g_1(z) + g_2\left(\frac{1}{z}\right)$ are two different Laurent expansions, using the obvious notation for the positive power and negative power terms. Subtracting, $0 = (f_1 - g_1) + (f_2 - g_2)$ and so

$$(f_1 - g_1)(z) = -(f_2 - g_2)\left(\frac{1}{z}\right)$$

In this formula, $f_1 - g_1$ is a power series which converges everywhere in the annulus $r < |z| < R$ and therefore in the disk $|z| < R$. Similarly $f_2 - g_2$ is a series in $\frac{1}{z}$ which converges in the annulus and thus in $r < |z|$. Moreover, $f_2 - g_2$ has no constant term and thus goes to zero at infinity.

Consequently, the formulas defines an entire function, given by $f_1 - g_1$ on $|z| < R$ and by $-(f_2 - g_2)\left(\frac{1}{z}\right)$ on $r < |z|$. This entire function is bounded, and so constant, and so zero. Since a power series expansion is unique, we conclude that the coefficients of the power series for both differences are zero.

Proof, part three: Term by term differentiation and integration of Laurent series are proved the same way. We'll discuss the case of differentiation.

If $f(z) = f_1(z) + f_2\left(\frac{1}{z}\right)$ is a Laurent series, the derivative of f equals

$$\frac{df_1}{dz}(z) + \frac{df_2}{dz}\left(\frac{1}{z}\right)\left(-\frac{1}{z^2}\right)$$

Both f_1 and f_2 are power series which can be differentiated term-by-term. By the displayed formula, the term $c_k z^k$ in f_1 becomes $kc_k z^{k-1}$ and the term $c_{-k} z^k \left(\frac{1}{z}\right)$, i.e., $c_{-k} z^{-k}$ in f_2 becomes $kc_{-k} z^{k-1}\left(\frac{1}{z}\right) \cdot \left(-\frac{1}{z^2}\right)$, i.e., $(-k)c_{-k} z^{-k-1}$.

9.6 Classification of Isolated Singularities

Laurent series occur most often about an isolated singularity z_0, converging in a punctured disk $0 < |z - z_0| < R$. While the case of more general annuli is sometimes useful, readers should think of the punctured disk about a singularity when they hear the term *Laurent expansion*.

Theorem 50. *Let f and g be holomorphic on a domain $\mathcal{U}$ and suppose g is not identically zero. Then all isolated singularities of $\frac{f(z)}{g(z)}$ are removable or poles.*

Proof: If z_0 is a removable singularity of the quotient, expand f and g in power series about z_0:

$$f(z) = (z - z_0)^m f_1(z)$$

$$g(z) = (z - z_0)^n g_1(z)$$

where f_1 and g_1 do not vanish at z_0. Then

$$\frac{f}{g} = (z - z_0)^{m-n} \frac{f_1(z)}{g_1(z)}$$

If $m - n \geq 0$, the singularity is removable. Otherwise, it is clearly a pole.

Theorem 51. *Let $f(z)$ be holomorphic on $\mathcal{U}$ except for an isolated singularity at z_0. Find the Laurent expansion of f in the annulus $0 < |z - z_0| < \epsilon$. Then*

- *If the Laurent expansion has only non-negative terms, the singularity is removable.*
- *If the Laurent expansion has a finite number of negative terms, including at least one such term, the singularity is a pole.*
- *If the Laurent expansion has infinitely many negative terms, the singularity is an essential singularity.*

Remark: This exhausts the possibilities for a Laurent expansion, and thus the possibilities for an isolated singularity. In particular, if f is bounded near an isolated singularity, then the singularity is automatically removable.

Proof: The first result is obvious, since the Laurent expansion is a power series which determines an appropriate value at $z = z_0$.

The second result is also easy. Since the positive terms of the Laurent expansion have a definite value at z_0, we need only consider the negative terms. We want to prove that a non-trivial finite polynomial in $\frac{1}{z-z_0}$ converges to infinity when $z \to z_0$, and this will follow if a non-zero polynomial in z converges to infinity when $z \to \infty$.

Write

$$|P(z)| = |z^k + c_1 z^{k-1} + ... + c_k| \geq |z|^k \left(1 - \frac{|c_1|}{|z|} - \frac{|c_2|}{|z|^2} - ... - \frac{|c_k|}{|z|^k}\right)$$

The required result follows immediately.

Finally consider the case of a Laurent series with infinitely many negative terms. Since the positive terms give a definite value at z_0, it suffices to prove that the negative terms can take any limiting value as $z_n \to z_0$. To simplify notation, we assume $z_0 = 0$. Suppose the result is false, and the limiting value cannot be λ.

Call the sum of the series with negative terms $g(z)$ and let this series converge in $0 < |z| < B_1$; choose $B < B_1$. Consider g on $0 < |z| \leq B$. If $g = \lambda$ infinitely often, a subsequence of points where it happens converges and this limit point must be the origin. By assumption, this cannot happen. By shrinking B, we can assume that g is never equal to λ in $0 < |z| \leq B$. Consider $h(z) = \frac{1}{g(z)-\lambda}$ on $0 < |z| \leq B$. If this function is not bounded near the origin, then there must be a sequence z_n with $|z_n| < \frac{1}{n}$ and $\left|\frac{1}{g(z_n)-\lambda}\right| > n$. So $|g(z_n) - \lambda| < \frac{1}{n}$ and again, by assumption this cannot happen.

Therefore, $h(z) = \frac{1}{g(z)-\lambda}$ is holomorphic in the punctured disk $0 < |z| < B$ and bounded near the origin. Expand this function in a Laurent series. By the proof that this can be done, the negative powers of z in this expansion are

$$-\sum_{k=0}^{\infty} \left[\frac{1}{2\pi i} \int_{\gamma_2} h(\zeta)\zeta^k d\zeta\right] \frac{1}{z^{k+1}}$$

Since h is bounded near the origin, and since this formula holds for γ_2 of arbitrarily small radius, we conclude that all of the negative terms vanish, and consequently h can be extended to be holomorphic at the origin. Solving for $g(z)$, $g(z) = \frac{1}{h(z)} + \lambda$. This is a quotient of holomorphic function and therefore had a removable singularity or pole at the origin, contradicting the assertion that it is given by a Laurent series with infinitely many negative powers.

Remark: Earlier we proved that the range of a non-constant entire function is dense in C. We now know enough to strengthen this result:

Theorem 52. *Let $f(z)$ be a non-constant entire function.*

1. *If f is a polynomial, then it takes each complex value. But if $z_n \to \infty$, then $f(z_n) \to \infty$*
2. *If f is not a polynomial, the values of f are dense in the plane. Indeed, for any fixed λ, there is a sequence $z_n \to \infty$ such that $f(z_n) \to \lambda$*

Proof: If f is not a polynomial, then $f\left(\frac{1}{z}\right)$ has an essential singularity at the origin; the second item is a consequence.

9.7 The Residue at an Isolated Singularity

Let $f(z)$ be holomorphic on a domain $\mathcal{U}$ with an isolated singularity at z_0. The coefficient of $\frac{1}{z-z_0}$ in the Laurent expansion of f about z_0 is called the *residue of f at z_0*, and denoted

$$\text{Res}(f, z_0)$$

Example 1. *The residue of $\frac{\cos z}{z}$ at the origin is 1 because*

$$\frac{\cos z}{z} = \frac{1 - \frac{z^2}{2!} + \frac{z^4}{4!}}{z} = \frac{1}{z} - \frac{z}{2!} + \frac{z^3}{4!} - ...$$

Example 2. *Suppose g and h are holomorphic near z_0 and $f(z) = \frac{g(z}{h(z)}$. Also suppose that $h(z_0) = h'(z_0) = ... = h^{(k-1)}(z_0) = 0$, but $h^{(k)}(z_0) \neq 0$. Then*

$$Res(f, z_0) = k\frac{g^{(k-1)}(z_0)}{h^{(k)}(z_0)}$$

The reader should check that this result is true.

Example 3. *Let $f(z) = \frac{1}{1+z^4}$. This function has isolated singularities at $\frac{\pm 1 \pm i}{\sqrt{2}}$. The residue at z_0 is $\frac{1}{4z_0^3} = \frac{1}{4z_0^3}$. So the residue at $\frac{1+i}{\sqrt{2}} = -\frac{1+i}{4\sqrt{2}}$ and the residue at $\frac{-1+i}{\sqrt{2}} = \frac{1-i}{4\sqrt{2}}$*

9.8 The Residue Theorem

Remark: The following result is the grand climax of our study of Cauchy's theorem. We state it as usually used, and then when multiple closed paths are involved.

Theorem 53 (The Residue Theorem)**.** *Let $f(z)$ be holomorphic on $\mathcal{U}$ except at isolated singularities. Let γ be a continuous closed curve in $\mathcal{U}$ such that every point inside γ is in $\mathcal{U}$. Assume that no isolated singularity is on γ. Then*

$$\int_\gamma f(z)\,dz = (2\pi i) \sum_{singularities} Res(f, z_i) W(\gamma, z_0)$$

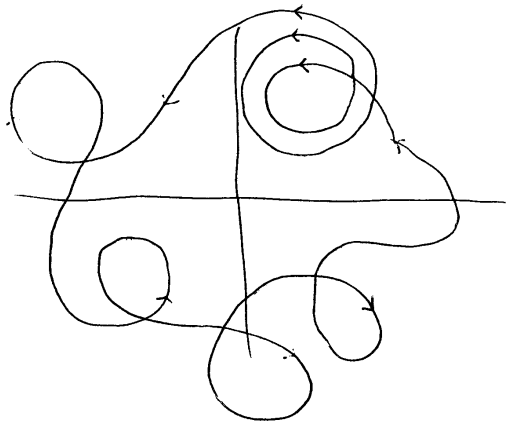


Figure 9.2: The Residue Theorem

Theorem 54 (The Residue Theorem)**.** *Let $f(z)$ be holomorphic on $\mathcal{U}$ except at isolated singularities. Let $\gamma_1, \gamma_2, \ldots, \gamma_k$ be closed continuous curves in $\mathcal{U}$ such that every point inside the γ_i is in $\mathcal{U}$, i.e., whenever $\sum W(\gamma_i, z_0) \neq 0$ then $z_0 \in \mathcal{U}$. Assume no isolated singularity is on a γ_i.*

Then

$$\sum \int_{\gamma_i} f(z)\,dz = (2\pi i) \sum_{singularities, \gamma_i} Res(f, z_0) W(\gamma_i, z_0)$$

Remark: Note that Cauchy's integral formula is a special case of this theorem.

Remark: Here's a small taste of an application. This type of application will be discussed systematically later.

Consider the path γ illustrated below. By the residue theorem,

$$\int_\gamma \frac{dz}{1+z^2} = \frac{1}{2\pi i}\text{Res}(f,i)$$

Since

$$\frac{1}{1+z^2} = \frac{1}{z-i}\left(\frac{1}{z+i}\right) = \frac{1}{z-i}\left(\frac{1}{2i} + \text{higher terms}\right)$$

we have $\text{Res}_i(f,i) = \frac{1}{2i}$, so the complex integral is $(2\pi i)\frac{1}{2i} = \pi$

Figure 9.3: Contour Integral One

This complex integral is

$$\int_{-R}^{R} \frac{dx}{1+x^2} + \int_{\text{semicircle}} \frac{dz}{1+z^2} =$$

On the semicircle, $|z^2| = |1+z^2-1| \leq |1+z^2| - 1$, so $|1+z^2| \geq |z^2| - 1 = R^2 - 1$. So $\left|\frac{1}{1+z^2}\right| \leq \frac{1}{R^2-1}$. Thus the integral over the semicircle is at most this number times the length of the semicircle, or $\frac{\pi R}{R^2-1}$. This expression goes to zero as R goes to infinity. We conclude that

$$\int_{-\infty}^{\infty} \frac{dx}{1+x^2} = \pi$$

In this special case, we can directly verify the result because

$$\int_{-\infty}^{\infty} \frac{dx}{1+x^2} = \arctan x\Big|_{-\infty}^{\infty} = \pi$$

Remark: Here's a second example. We compute $\int_{-\infty}^{\infty} \frac{dx}{1+x^4}$ the same way. The singularities of the integrand inside the semicircle are at $\frac{1+i}{\sqrt{2}}$ and $\frac{-1+i}{\sqrt{2}}$. So the integral is $2\pi i$ times the sum of the residues at these points.

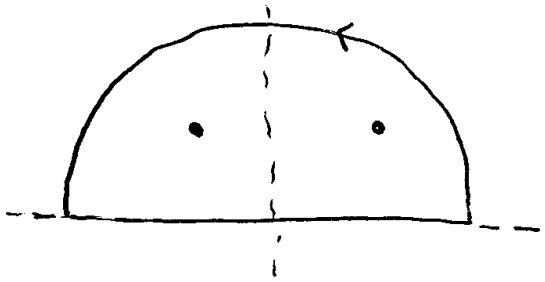


Figure 9.4: Contour Integral Two

By a calculation in the previous section, these residues are $-\frac{1+i}{4\sqrt{2}}$ and $\frac{1-i}{4\sqrt{2}}$. So

$$\int_{-\infty}^{\infty} \frac{dx}{1+x^4} = (2\pi i)\left(\frac{-1-i+1-i}{4\sqrt{2}}\right) = (2\pi i)\left(\frac{-i}{2\sqrt{2}}\right) = \frac{\pi}{\sqrt{2}}$$

Proof of the Residue Theorem: We'll prove the theorem for one γ; the general proof is exactly the same. As in the diagram below, add paths $\tau_1, \ldots, \tau_k$ around the isolated singularities z_i inside γ, going around $-W(\gamma, z_i)$ times. There are only finitely many isolated singularities inside γ since γ together with its inside is compact.

Apply the homological Cauchy theorem. Since γ goes around z_i by $W(\gamma, z_i)$ times and τ_i goes around z_i by $-W(\gamma, z_i)$ times and τ_j doesn't go around z_i at all, the z_i are not inside the path formed by γ and the τ_i. Thus f is holomorphic inside this path, and Cauchy's theorem gives

$$\int_\gamma f(z)\,dz + \sum \int_{\tau_j} f(z)\,dz = 0$$

Therefore

$$\int_\gamma f(z)\,dz = -\sum \int_{\tau_j} f(z)\,dz$$

We can shrink the τ_j as much as we want without changing the integrals. Eventually τ_j will be inside the annulus where the Laurent series for the isolated singularity z_i converges. Separate this Laurent series into two parts: all the terms except the $\frac{1}{z-z_i}$ term and separately this term,

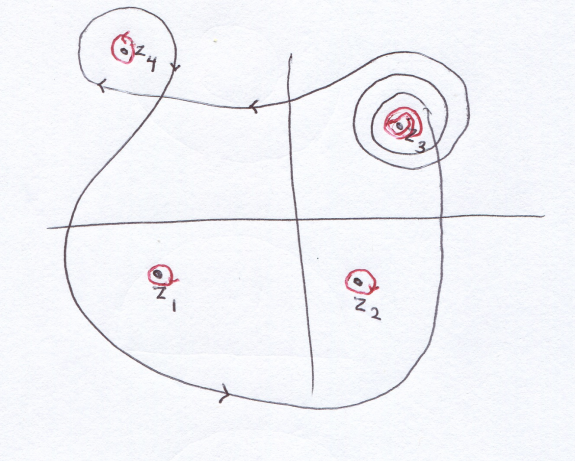


Figure 9.5: Proof of the Residue Theorem

which equals

$$\frac{\text{Res}(f, z_i)}{z - z_i}$$

Since the sum of terms of order not -1 can be integrated term by term, the Laurent series has the form

$$\frac{dh}{dz} + \frac{\text{Res}(f, z_i)}{z - z_i}$$

near z_i, and

$$-\int_{\tau_j} f(z)\, dz = -\int_{\tau_j} \frac{dh}{dz}\, dz - \text{Res}(f, z_i) \int_{\tau_j} \frac{dz}{z - z_i}$$

But $\int_{\tau_j} \frac{dh}{dz}\, dz = h(\text{end of } \tau) - h(\text{start of } \tau) = 0$. Also

$$-\text{Res}(f, z_i) \int_{\tau_j} \frac{dz}{z - z_i} = -(2\pi i)\text{Res}(f, z_i) W(\tau_i, z_i) = (2\pi i)\text{Res}(f, z_i) W(\gamma, z_i)$$

by our choice of τ. In short,

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

Chapter 10

Fundamental Results

We now harvest an array of important results which follow from the Residue theorem and other forms of Cauchy's theorem.

10.1 Meromorphic Functions

Definition 24. *A* meromorphic function *on a domain $\mathcal{U}$ is a holomorphic function with isolated singularities, all removable or poles.*

Remark: We always imagine that all removable singularities are removed, so a meromorphic function has values everywhere. It even has values at poles, if we allow the value to be ∞ there. Note that it would make no sense to assign the value ∞ to an essential singularity, since nearby values can be anything in that case.

Remark: The meromorphic functions on a domain form a *field* we will sometimes denote $\mathcal{M}(\mathcal{U})$. The key point is that $\frac{1}{f}$ has only zeros and poles if f is not identically zero. It has zeros at the poles of f and poles at the zeros of f. All field axioms are easily checked.

Remark: We will focus more and more on meromorphic functions as we proceed; they are the natural objects of study in complex analysis. Notice that from the beginning of these notes on page 8, our examples of holomorphic functions had poles.

10.2 The Order of a Zero or Pole

Definition 25. *Let f be holomophic on an open neighborhood of z_0, except possibly at z_0. Assume f is not identically zero and z_0 is not an essential singularity. Expand f in a Laurent series near*

z_0, and let k be the lowest power of $(z - z_0)^k$ in this expansion. Then the order of f at z_0 *is*

$$Ord_{z_0}(f) = k$$

Remark: If $f(z_0) = 0$, then the order of f at z_0 is the multiplicity of this zero, i.e., the number of zeros at z_0. If z_0 is a pole, the order is negative and counts the multiplicity of infinite values at z_0. If z_0 is neither a zero nor a pole, the order of f at z_0 is zero.

Remark: Thus "order" extends the concept of the multiplicity of a zero of a polynomial introduced in algebra.

Example: $\mathrm{Ord}_0(\sin z - z) = 3$ because $\sin z - z = -\frac{z^3}{3!} + \frac{z^5}{5!} - ...$

Theorem 55. *The order function has the following properties:*

- $Ord_{z_0}(fg) = Ord_{z_0}(f) + Ord_{z_0}(g)$
- $Ord_{z_0}(f + g) \geq \min\{Ord_{z_0}(f), Ord_{z_0}(g)\}$
- $Ord_{z_0}(f)$ *takes all possible values as* f *varies.*
- $Ord_{z_0}(\lambda) = 0$ *if* λ *is a nonzero constant*

Remark: The proof is easy.

10.3 The Riemann Sphere

It is possible to extend complex analysis to arbitrary surfaces embedded in higher dimensions. When given a complex structure, these objects are called *Riemann surfaces*. They play a big role in the theory, first implicitly and later explicitly.

The simplest of these surfaces is the sphere. Imagine this sphere created by adding a point at infinity to the ordinary plane. Stereographic projection sends the sphere minus its north point to the plane, and the added point at infinity replaces this north pole.

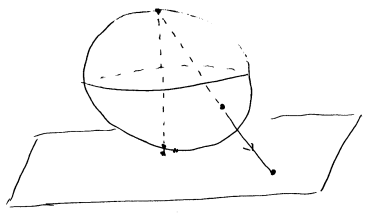


Figure 10.1: Stereographic Projection

We introduce two coordinates on the sphere, called z and w. The z coordinate is the standard coordinate for points on the plane; the added point at infinity has no z coordinate. Non-zero points in the plane also have w coordinates, given by $w = \frac{1}{z}$; we give the point at infinity the w coordinate zero, and the origin in the plane has no w coordinate. The extended plane with these coordinates is known as the *Riemann Sphere*.

We now speak of functions f on the sphere as given by two formulas: $f(z)$ and $g(w) = f\left(\frac{1}{w}\right)$. We can use either of these formulas when speaking of points on the sphere that are not 0 and not ∞. If we are interested in the behavior of a function near the origin, we use $f(z)$. If we are interested in the behavior of a function near ∞, we use $g(w)$ and remove the singularity at $w = 0$.

For example, $f(z) = \frac{z-1}{z-2}$ is meromorphic on the sphere. It has the value $\frac{1}{2}$ at the origin. It has a zero at 1 and a pole at 2, and its value at infinity is 1 because $\lim_{z\to\infty} |f(z)| = 1$.

Surprisingly, there are few meromorphic functions on the Riemann sphere.

Theorem 56. *The field $\mathcal{M}(S^2)$ is exactly the set $C(z)$ of all rational functions $\frac{P(z)}{Q(z)}$ where P and Q are arbitrary polynomials.*

Proof: Suppose $f(z)$ is meromorphic on the sphere. Since the sphere is compact, isolated singularities cannot accumulate and thus there are at most finitely many of them. If there are poles of f at $z_1, \ldots, z_n$ of orders $-k_1, \ldots, -k_n$, we can form

$$h(z) = (z - z_1)^{k_1} \ldots (z - z_n)^{k_n} f(z)$$

and get an entire function with no poles in the finite plane. It suffices to prove that $h(z)$ is a polynomial, because then $f(z)$ will be a rational function.

Since h is meromorphic on the entire Riemann sphere, it has at worst a pole at infinity, and thus $g(z) = h(\left(\frac{1}{z}\right)$ has a Laurent expansion with finitely many negative terms near the origin. It follows that $h(z)$ has a Laurent expansion near infinity with only finitely many positive terms. Since h is holomorphic in the entire plane including the origin, this Laurent expansion converges to h everywhere except possible zero. Since f is holomorphic at the origin, the Laurent expansion must extend to the origin and thus must be a polynomial.

Remark: Turn back to the end of the previous section on the order function. There we listed some simple properties of Ord. All of these properties except the last make sense for an arbitrary field. Even the last makes sense for the field $C(z)$ of rational functions.

It turns out to be relatively easy to prove that the only functions

$$\text{Ord} : C(z) - \{0\} \to Z$$

satisfying these properties are Ord_{z_0} where z_0 is some point on the Riemann surface, including possibly the point at infinity. This fact is exceedingly important in the advanced theory of Riemann surfaces.

If we replace the field $C(z)$ by the field Q of rational numbers, the last statement on the list no longer makes sense. Ignore it. The remaining statements make sense. It is possible to find all such Ord maps $Q - \{0\} \to Z$. There is one such map for each prime number.

Therefore, by some far fetched analogy, the ordinary prime numbers in the case Q correspond to the points on the Riemann sphere in the case $C(z)$. This analogy between number theory and complex analysis continues to hold for deeper and deeper results, and has motivated very important work in the twentieth century.

10.4 The Argument Principle

Theorem 57 (The Argument Principle)**.** *Let γ be a closed continuous path, and let f be meromorphic on an open set including all points inside γ. Suppose f has no poles or zeros on γ. Then*

$$\frac{1}{2\pi i}\int_\gamma \frac{f'(\zeta)}{f(\zeta)}\,d\zeta = W(f\circ\gamma, 0) = \sum_{z_i} W(\gamma, z_i)Ord(f, z_i)$$

where the sum is over zeros and poles of f inside γ.

Proof: If f has a zero or pole at z_0 of order k, then near z_0 we can write $f(z) = (z - z_0)^k h(z)$ where h is nonsingular and not zero near z_0. Then

$$\frac{f'}{f} = \frac{k(z - z_0)^{k-1}h + (z - z_0)^k h'}{(z - z_0)^k h} = \frac{k}{z - z_0} + \frac{h'}{h}$$

near z_0, so the residue at z_0 of $\frac{f'}{f}$ is k. The Residue Theorem then proves that the left side of our equation equals the right side.

We must show that the left and right sides equal the term in the center, which is

$$\mathrm{W}(f\circ\gamma, 0) = \frac{1}{2\pi i}\int_{f\circ\gamma}\frac{1}{\zeta}\,d\zeta$$

and this will follow if

$$\int_\gamma \frac{f'(\zeta)}{f(\zeta)} = \int_{f\circ\gamma}\frac{1}{\zeta}d\zeta$$

If our path were piece-wise differentiable, this would follow from the chain rule. Since our path is continuous, we must prove this by constructing primitives. If $\gamma : [0, 1] \to C$, a primitive for the left integral is a function $\varphi(t)$ on $[0, 1]$ which is locally the pullback via γ of a holomorphic h

with $\frac{dh}{dz} = \frac{f'}{f}$. Locally we can find a branch of the logarithm and form $\text{Log}(f(z))$; the derivative of this function is $\frac{f'}{f}$. So φ is locally the pullback via γ of $\text{Log}(f(z))$, or equivalently the pullback via $f \circ \gamma$ of Log. And this gives the integral on the right. QED.

Remark: Notice that $\text{Log}(f(z)) = \ln|f(z)| + i \arg(f(z))$ and $\frac{d}{dz}\text{Log}(f(z)) = \frac{f'(z)}{f(z)}$. It follows that $\int_\gamma \frac{f'}{f} d\zeta$ measures the increase of the argument of $f(z)$ as we move along γ, hence the name of the theorem.

Theorem 58 (Inverse Function Theorem)**.** *Let f be holomorphic on a domain $\mathcal{U}$ containing z_0 and suppose $\frac{df}{dz}(z_0) \neq 0$. Then there is an open neighborhood $\mathcal{V} \subset \mathcal{U}$ of z_0 such that*

- *$f(\mathcal{V})$ is open*
- *$f : \mathcal{V} \to f(\mathcal{V})$ is a homeomorphism*
- *$f^{-1} : f(\mathcal{V}) \to \mathcal{V}$ is holomorphic*

Proof: Without loss of generality, we can assume that $z_0 = 0$. Choose a small counterclockwise circle γ about the origin which does not go around any other zero of f. The argument principle gives

$$\frac{1}{2\pi i}\int_\gamma \frac{f'}{f} d\zeta = \text{W}(f \circ \gamma, 0) = 1$$

Since the winding number is constant on open connected components of the compliment of paths, we can find a small disk $\mathcal{W}$ about the origin inside $f \circ \gamma$ such that $W(f \circ \gamma, \lambda) = 1$ for all points in this disk.

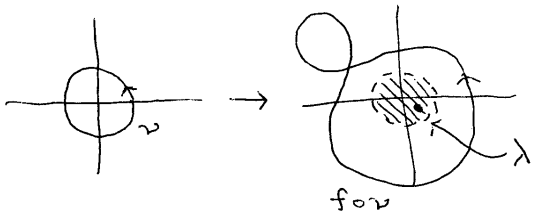


Figure 10.2: Proof of Inverse Function Theorem

By definition, this winding number equals $\frac{1}{2\pi i}\int_{f\circ\gamma} \frac{1}{\zeta - \lambda} d\zeta$, and as in the proof of the argument principle, this equals $\frac{1}{2\pi i}\int_\gamma \frac{f'}{f-\lambda} d\zeta$. By the Residue theorem, this equals the number of times $f = \lambda$ inside γ minus the number of poles inside γ. So $f = \lambda$ exactly once inside γ and f' at the point where this happens is not zero.

Let $\mathcal{V} = f^{-1}(\mathcal{W}) \cap (\text{interior of } \gamma)$. Then $f : \mathcal{V} \to \mathcal{W}$ is one-to-one and onto.

We want to prove that f is a homeomorphism from $\mathcal{V}$ to $\mathcal{W}$ and that f^{-1} has a derivative at each point of $\mathcal{W}$. Formally, the argument for a derivative goes like this:

$$\lim_{w \to w_0} \frac{f^{-1}(w) - f^{-1}(w_0)}{w - w_0} = \lim_{f^{-1}(v) \to f^{-1}(v_0)} \frac{f^{-1}(f(v) = f^{-1}(f(v_0))}{f(v) - f(v_0)} = \lim_{v \to v_0} \frac{1}{\left(\frac{f(v) - f(v_0)}{v - v_0}\right)} = \frac{1}{f'(v_0)}$$

Since each w comes from a unique v and f' exists and is never zero on $\mathcal{V}$, this formal argument will follow if $w \to w_0$ implies $v \to v_0$, and this will follow from the continuity of f^{-1}. So the entire theorem is proved modulo this continuity.

But the continuity of f^{-1} was essentially proved by our earlier argument. Indeed, we know that f' is not zero at each point v of $\mathcal{V}$. The previous argument then showed that we can choose an arbitrarily small open disk about v such that the image of this disk contains an open neighborhood of $f(v)$, and we can then shrink the arbitrarily small open disk to an even smaller open set about v whose image is this open neighborhood of $f(v)$. Since any open set $\mathcal{X}$ in the domain can be covered by open disks inside $\mathcal{X}$, and thus by the even smaller open sets with open image, the image of $\mathcal{X}$ in $\mathcal{W}$ must be open.

10.5 The Open Mapping Theorem

The inverse function theorem together with our study of conformal maps in chapter two gives a complete picture of holomorphic maps near z_0 with $f(z_0) \neq 0$. Namely, we can find an open neighborhood $\mathcal{V}$ of z_0 and an open neighborhood $\mathcal{W}$ of $f(z_0)$ and then $f : \mathcal{V} \to \mathcal{W}$ and $f^{-1} : \mathcal{W} \to \mathcal{V}$ are differentiable homeomorphisms which preserve angle.

We are going to extend this picture to points z_0 where $f'(z_0) = 0$. Suppose the order of the zero of f' at z_0 is $k-1$, where $k - 1 \geq 1$ and thus $k \geq 2$. Then near z_0 we have

$$f(z) = f(z_0) + \frac{f^{(k)}(z_0)}{k!}(z - z_0)^k + \ldots = f(z_0) + (z - z_0)^k \left(c_k + c_{k+1}(z - z_0) + c_{k+2}(z - z_0)^2 + \ldots\right)$$

Note that $h(z_0) \neq 0$. So we can define a branch of the logarithm Log near $h(z_0)$ and write $h(z) = \left(e^{Log(h(z))/k}\right)^k = g(z)^k$ where g is holomorphic near z_0 and $g(z_0) \neq 0$. Therefore $f(z) = f(z_0) + ((z - z_0)g(z))^k$.

Notice that $(z - z_0)g(z)$ is defined near z_0 and has non-zero derivative at z_0. By the inverse function theorem, this is a homeomorphism near z_0 with holomorphic inverse. By shrinking the image set, we can suppose that it is a disk about the origin.

We have proved

Theorem 59. *Suppose f is holomorphic on U and f' has a zero at z_0 of order $k-1>0$. Then f is given by the following composition of maps near z_0. The first map is a translation, the second is a homeomorphism, the third is $z \to z^k$ and the final map is another translation. So up to homeomorphisms, f locally wraps a neighborhood around z_0 k times.*

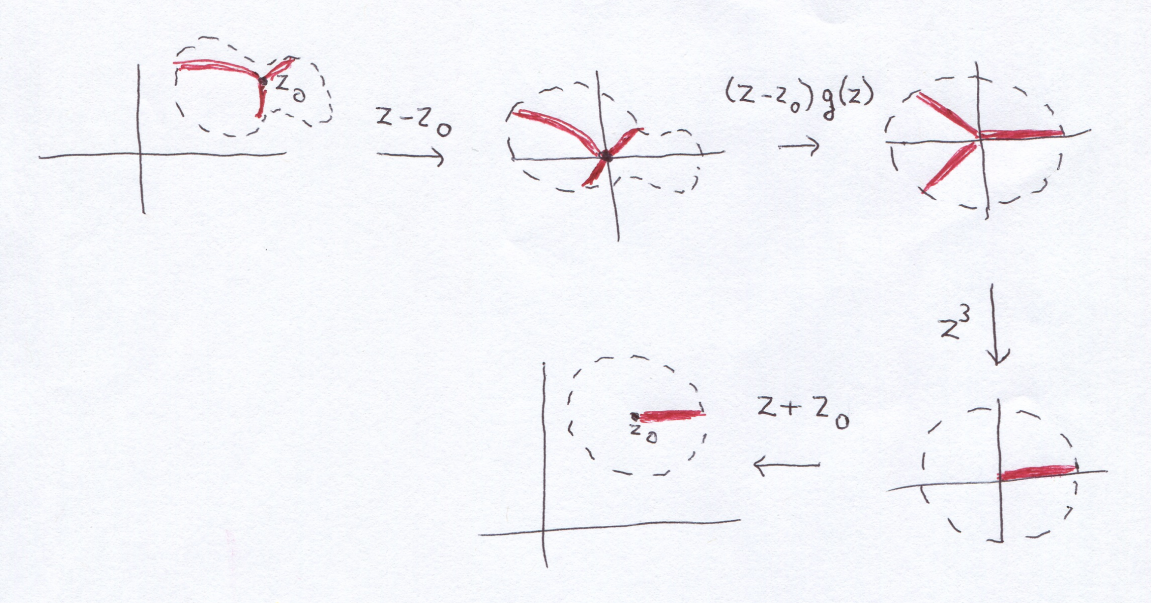


Figure 10.3: Conformal Maps Near a Zero of $\frac{df}{dz}$

Corollary 18 (The Open Mapping Theorem)**.** *Suppose f is a non-constant holomorphic map on a domain $\mathcal{U}$. Then f is an open mapping; it maps open sets to open sets.*

Proof: It suffices to prove this locally, but points where $f'(z_0) \neq 0$ have arbitrarily small open neighborhoods which map to open sets by the inverse function theorem, and this remains true if $f'(z_0) = 0$ by the above sequence of maps.

Corollary 19. *A non-constant holomorphic function on a domain U cannot be real-valued, or take values only on an arbitrary line in the plane, or take values only on the unit circle.*

10.6 The Maximum Principle

Theorem 60 (The Maximum Principle)**.** *The absolute value of a non-constant holomorphic function on a domain U cannot have a local maximum.*

Proof: This is a trivial consequence of the open mapping theorem.

Corollary 20. *Let f and g be continuous on a closed disk and holomorphic on the interior of this disk. If $f = g$ on the boundary of the disk, then $f = g$ everywhere inside.*

Proof: Consider $h = f - g$. Suppose $h = 0$ on the boundary of the disk. Then either $f = g$ everywhere or else $|f(z) - g(z)|$ has a maximum value inside the disk, which cannot happen by the previous result.

10.7 Uniform Limits of Holomorphic Functions

The following amazing theorem is false for real-valued functions, as illustrated by the pictures below.

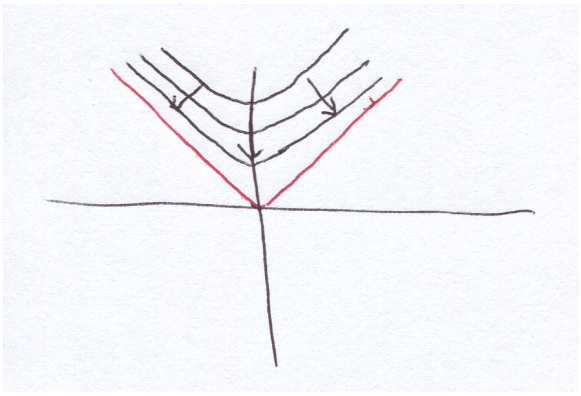


Figure 10.4: Uniform Limits in Real Variables

Theorem 61. *Suppose f_n are holomorphic on a domain U and $f_n(z) \to f(z)$ uniformly on each compact subset of $\mathcal{U}$. Then $f(z)$ is holomorphic and $\frac{df_n}{dz} \to \frac{df}{dz}$ uniformly on each compact subset of $\mathcal{U}$. Consequently all higher derivatives of f_n also converge to the corresponding higher derivative of f uniformly on compact subsets.*

Proof: To prove that f is holomorphic, it suffices by Morera's theorem to prove that $\int_{\mathcal{R}} f(z)\,dz = 0$ for all rectangles contained, together with their interiors, in $\mathcal{U}$. But this holds for f_n since f_n is holomorphic. Moreover, by uniform convergence, $\int_{\mathcal{R}} f_n(z)\,dz \to \int_{\mathcal{R}} f(z)\,dz$.

Next we prove that uniform convergence of functions implies uniform convergence of derivatives. This follows from the following formula for the derivative of f, a generalization of Cauchy's formula which follows from the Residue theorem: if γ is a small circle about z,

$$\frac{df}{dz}(z) = \frac{1}{2\pi i}\int_{\gamma} \frac{f(\zeta)}{(\zeta - z)^2} d\zeta$$

Suppose this circle has radius R. Then for z on or inside the disk of radius $R/2$, we have

$$\left|\frac{df_n}{dz} - \frac{df}{dz}\right| \leq \max_\gamma |f_n(z) - f(z)| \frac{1}{R/2}(2\pi R)$$

Since $f_n \to f$ uniformly on the circle of radius R,

$$\frac{df_n}{dz} \to \frac{df}{dz}$$

uniformly on the interior of a disk of radius $R/2$. Any compact set can be covered by finitely many such disks, so we are done.

Chapter 11

Automorphism Groups of the Complex Plane, the Unit Disk, and the Riemann Sphere

This chapter is the first of three which lead to the Riemann Mapping Theorem, the central result of Riemann's PhD thesis. Results in the preliminary chapters are of independent interest.

If $\mathcal{U}$ is a domain in C or the Riemann Sphere, an *automorphism* of $\mathcal{U}$ is a one-to-one and onto holomorphic map $\varphi : \mathcal{U} \to \mathcal{U}$. It is easy to see that the automorphisms of $\mathcal{U}$ form a group.

We will compute these groups for the open unit disk D, the entire complex plane C, and the Riemann sphere S^2. It turns out that these are the only simply connected Riemann surfaces, although we will not prove that. So our examples were not chosen at random.

11.1 The Automorphism Group of C

Theorem 62. *The automorphisms of C are precisely the maps*

$$\varphi(z) = az + b$$

where a and b are complex and $a \neq 0$.

Proof: Since these are clearly automorphisms, it suffices to prove that all automorphisms have this form. An automorphism is an entire function, so by theorem 54, it is either a polynomial or else for each fixed λ, we can find $z_n \to \infty$ with $\varphi(z_n) \to \lambda$. But φ is onto, so there exists a z_0

with $\varphi(z_0) = \lambda$. Since φ is one-to-one, the derivative of φ at z_0 cannot be zero, so the inverse function theorem states that there bounded open neighborhoods $\mathcal{W}_1$ and $\mathcal{W}_2$ of z_0 and λ such that $\mathcal{W}_1$ is mapped one-to-one and onto to $\mathcal{W}_2$ by φ. The values in $\mathcal{W}_2$ are only taken by φ in $\mathcal{W}_1$, so λ cannot be a limiting value of φ near infinity.

It follows that φ is a polynomial. Since φ is one-to-one, this polynomial cannot have multiple roots or one root of order greater than one. So it is linear. QED.

Remark: The complex plane has a natural metric. Isometries, that is, distance preserving maps, must be conformal by the law of cosines. So our group contains all isometries of the plane. These have the form

$$\varphi(z) = e^{i\theta}z + b$$

If $e^{i\theta} = 1$, these are translations. Otherwise they are rotations about some point. Indeed, any map $\varphi(z) = az + b$ with $a \neq 1$ has a fixed point because $az + b = z$ is solvable, and it is easy to see that when $a = e^{i\theta}$, the resulting map is rotation about this fixed point.

The remaining maps have the form

$$\varphi(z) = me^{i\theta}z + b$$

for $m > 0$; $m \neq 1$. Such maps have fixed points; they magnify about the fixed point by m and then rotate by θ.

The globally conformal maps which are not isometries are therefore, essentially, magnifications.

11.2 The Automorphism Group of the Riemann Sphere

Definition 26. *A* linear fractional transformation *is a map*

$$\varphi(z) = \frac{az + b}{cz + d}$$

with a, b, c, d complex and $ad - bc \neq 0$.

Remark: If we multiply each of a, b, c, d by the same non-zero constant λ, we get the same map. It follows that the set of linear fractional transformations form an object of complex dimension 3 and so real dimension 6. Notice that the condition $ad - bc \neq 0$ does not change the dimension.

Theorem 63. *The automorphisms of the Riemann Sphere are precisely the following linear fractional transformations, with $ad - bc \neq 0$:*

$$\varphi(z) = \frac{az + b}{cz + d}$$

Proof: We leave it to the reader to show that these maps are automorphisms. Notice that $z = -\frac{d}{c}$ maps to ∞, and ∞ maps to $\frac{a}{c}$.

Linear fractional transformations are 3-transitive: any three points can be mapped to any other three points by exactly one φ. For instance, $\varphi(z) = \frac{C-B}{C-A} \cdot \frac{z-A}{z-B}$ maps A to 0, B to ∞, and C to 1.

Now let φ be an arbitrary automorphism of the sphere. Multiplying by a suitable linear fraction transformation, we can suppose that $0 \to 0, 1 \to 1, \infty \to \infty$, and it suffices to show that this composition is the identity.

Since ∞ maps to ∞, no other point maps to infinity and our map is an automorphism of C, and hence has the form $\varphi(z) = az + b$. Since $0 \to 0$, $b = 0$. Since $1 \to 1$, $a = 1$. QED.

Theorem 64. *A linear fractional transformation is either of the form $az + b$ with $a \neq 0$ or of the following form with $c \neq 0, B = \frac{a}{c}, A = -\frac{ad-bc}{c}$:*

$$(Az + B) \circ \frac{1}{z} \circ (cz + d)$$

Proof: A simple calculation.

Remark: It follows that linear fractional transformations are products of translations of the plane, rotations about a point in the plane, magnifications about a point in the plane, and the map $\frac{1}{z}$.

This last map is particularly interesting; it is essentially *inversion in a circle*. Suppose p is a point in the plane and R is a fixed positive number. We define *inversion in the circle* of radius R about p to be the map which interchanges p and ∞, fixes the circle pointwise, and maps the inside to the outside and the outside to the inside by sending $q \neq p$ to the point Q on the ray from p through q, where $|q - p| \cdot |Q - p| = R^2$. So this map turns the circle inside out.

Inversion in a circle reverses orientation, so it is cannot be holomorphic. Actually inversion in a circle of radius 1 about the origin is given by $z \to \frac{z}{|z|^2}$ because the product of the absolute value of z and the absolute value of $\frac{z}{|z|^2}$ is one and $\frac{z}{|z|^2}$ is on the same ray through the origin as z.

Our map $z \to \frac{1}{z}$ equals

$$\varphi(z) = \frac{1}{z} = \frac{\overline{z}}{z\overline{z}} = \frac{\overline{z}}{|z|^2} = \frac{z}{|z|^2} \circ \overline{z}$$

and thus is reflection across the x-axis followed by inversion in the circle. The inclusion of this reflection produces an orientation-preserving map.

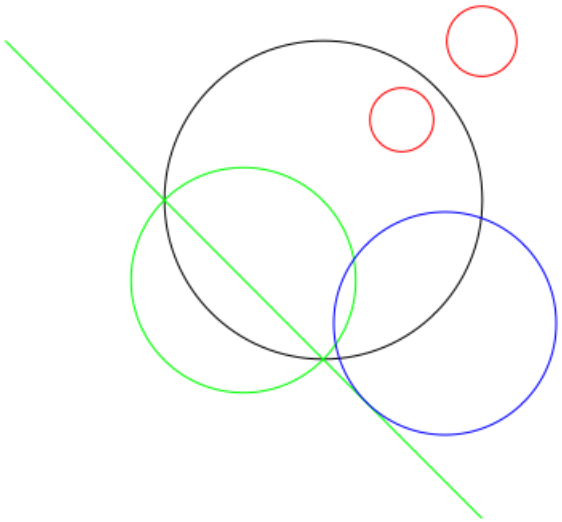


Figure 11.1: Inversion in a Circle

The above picture shows a pure inversion in a circle without the additional reflection. We invert in the black circle. The red circle inside moves to a bigger red circle outside. The blue circle meets the circle of inversion perpendicularly, and a little thought shows that it consequently maps to itself. The green circle goes through the origin, so its image goes through ∞ and becomes a straight line.

Remark: More generally, if we have a sphere S^{n-1} with center p and radius R in R^n, the map *inversion in the sphere* interchanges p and ∞, fixes the sphere pointwise, and sends q on the ray from p to q to the point Q on the same ray, where $|q-p| \cdot |Q-p| = R^2$. It is not difficult to show that this map is conformal.

With somewhat more difficulty, one can prove

Theorem 65. *If $\varphi : \mathcal{U} \to \mathcal{V}$ is one-to-one, onto, and conformal, where $\mathcal{U}$ and $\mathcal{V}$ are domains in R^n and $n \geq 3$, then φ is a composition of translations, rotations, magnifications, reflections, and inversions in spheres.*

Notice that the previous result is a *local result*. In contract, we know that there are enormously many local conformal maps in R^2. So the proper way to generalize complex analysis to higher dimensions is certainly *not* to concentrate on conformal maps as the central idea of the subject.

Theorem 66. *A linear fractional transformation maps circles to either circles or lines, and lines to either circles or lines.*

Proof: This result is obviously true for translations, rotations, and magnifications, since these maps send circles to circles and lines to lines. By theorem 69, it suffices to prove the result for $\frac{1}{z} = \frac{x-iy}{x^2+y^2}$.

The general equation of a circle or straight line in the plane is

$$A(x^2 + y^2) + Bx + Cy + D = 0$$

The image $\left(\frac{x}{x^2+y^2}, \frac{-y}{x^2+y^2}\right)$ satisfies this equation provided

$$A\left(\frac{x^2 + y^2}{(x^2 + y^2)^2}\right) + B\frac{x}{x^2 + y^2} + C\frac{-y}{x^2 + y^2} + D = 0$$

or equivalently

$$A + Bx - Cy + D(x^2 + y^2) = 0$$

Both of these equations have the same form. Indeed, the possible solutions of the first equation are a circle, a line, a point, the empty set, or everything, and similarly for the solutions of the second equations. Since our linear fractional transformation is one-to-one and onto, one equation has solution a circle or line just in case the other equation has tone of these solutions.

Remark: A calculation shows that stereographic projection, illustrated on page 81, is a conformal map from the sphere to the plane. Give the sphere the standard Riemannian metric; then isometries of this sphere are conformal and map to linear fractional transformations.

The site http://users.math.msu.edu/users/shapiro/pubvit/downloads/rs_rotation/rotation.pdf contains a paper by Joel H. Shapiro proving that the linear fractional transformations corresponding to isometries of the sphere are exactly those induced by a unitary matrix

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

Note that the matrix induced by $\frac{az+b}{cz+d}$ is only unique up to multiplying all components by the same nonzero λ. The paper's assertion means that a linear fractional transformation induces an isometry of the sphere if and only if one representative matrix for the transformation is unitary.

Notice that $\frac{1}{z}$ comes from an isometry of the sphere. This is a coincidence, and inversions in other circles do not.

Example: It is traditional to stop here and draw many wonderful pictures of maps obtained from linear fractional transformations. We'll briefly consider one case. Let

$$\varphi(z) = i\,\frac{1-z}{1+z}$$

This maps sends $-1 \to \infty, 1 \to 0$, and $i \to 1$. Thus the maps sends the circle through -1 and 1 and i to the line through ∞ and 0 and 1, i.e., the x-axis. It follows that this map is a conformal isomorphism from the interior of the unit disk to the upper or lower half plane bounded by the x-axis. Since $0 \to i$, it maps the interior of the unit circle to the upper half plane.

Notice that the inverse of this map is

$$\varphi^{-1}(z) = \frac{1+iz}{1-iz}$$

11.3 The Automorphism Group of the Open Unit Disk

We begin with a beautiful, and very important, lemma

Lemma 6 (Schwarz' Lemma)**.** *Let $f(z)$ be a holomorphic map from the open unit disk to itself and suppose that $f(0) = 0$. For each z in the disk,*

$$|f(z)| \leq |z|$$

If this inequality is an equality for some nonzero z, then $f(z)$ is a rotation, $f(z) = e^{i\theta}z$.

Proof: Apply the principle of the maximum to $g(z) = \frac{f(z)}{z}$. On the circle of radius $1-\epsilon$, $|g|$ is at most one and $|z| = 1-\epsilon$, so $|g| \leq \frac{1}{1-\epsilon}$. By the maximum principle, $|g|$ is at most this value everywhere inside this circle of radius $1-\epsilon$. As the circle grows closer to 1, the bound grows tighter, so ultimately we find that $|g| \leq 1$ inside the disk. If $|g| = 1$ at some z inside the disk, then g takes its maximum value inside the disk and thus is constant by the maximum principle.

Corollary 21. *Let f be an automorphism of the unit disk mapping the origin to the origin. Then f is rotation, $f(z) = e^{i\theta}z$.*

Proof: Apply the Schwarz lemma to f and f^{-1}. We conclude that $|f(z) \leq |z|$ and $f(z) \geq |z|$. Hence $|f(z)| = |z|$ and the result follows from the Schwarz lemma.

Theorem 67. *The automorphism group of the open unit disk D is the set of all*

$$f(z) = e^{i\theta}\frac{z - z_0}{1 - z\,\overline{z_0}}$$

as $z_0 \in D$ and $0 \leq \theta < 2\pi$.

Proof: We will prove that $\frac{z-z_0}{1-z\,\overline{z_0}}$ maps the unit circle to itself. Since the origin maps to $-z_0$ and $|z_0| < 1$, this linear fractional transformation must map the interior of the circle to the interior and thus be an automorphism of the unit disk. It is enough to prove that when z is on the unit circle,

$$|z - z_0|^2 = |1 - z\,\overline{z_0}|^2$$

or

$$(z - z_0)(\overline{z} - \overline{z_0}) = (1 - z\,\overline{z_0})(1 - \overline{z}\,z_0)$$

or

$$z\overline{z} - z_0\overline{z} - z\overline{z_0} + z_0\overline{z_0} = 1 - z\overline{z_0} - \overline{z}z_0 + z\overline{z}z_0\overline{z_0}$$

and this holds because the middle terms cancel and $z\overline{z} = 1$ and $z_0\overline{z_0} = z\overline{z}z_0\overline{z_0}$.

If φ is an arbitrary automorphism of the unit disk, it maps 0 to some z_0 in the disk. Then

$$\frac{z - z_0}{1 - z\,\overline{z_0}} \circ \varphi^{-1}(z)$$

is an automorphism which maps 0 to 0, so it equals $e^{-i\theta}z$ for some θ, and thus

$$\frac{z - z_0}{1 - z\,\overline{z_0}} \circ \varphi^{-1}(z) \circ \varphi(z) = e^{-i\theta}z \circ \varphi(z)$$

or equivalently

$$e^{i\theta}\frac{z - z_0}{1 - z\,\overline{z_0}} = \varphi(z)$$

Theorem 68. *The conformal isomorphism $\varphi(z) = i\frac{1-z}{1+z}$ from the unit disk to the upper half plane maps the automorphisms group $e^{i\theta}\frac{z-z_0}{1-z\,z_0}$ to the automorphism group $\frac{az+b}{cz+d}$ of linear fractional transformations with real coefficients a, b, c, d such that $ad - bc > 0$.*

Remark: This group is usually called $SL(2, R)/\pm I$ since multiplying a, b, c, d by ± 1 does not change the linear fractional transformation. Serge Lang wrote an entire book on $SL(2, R)$.

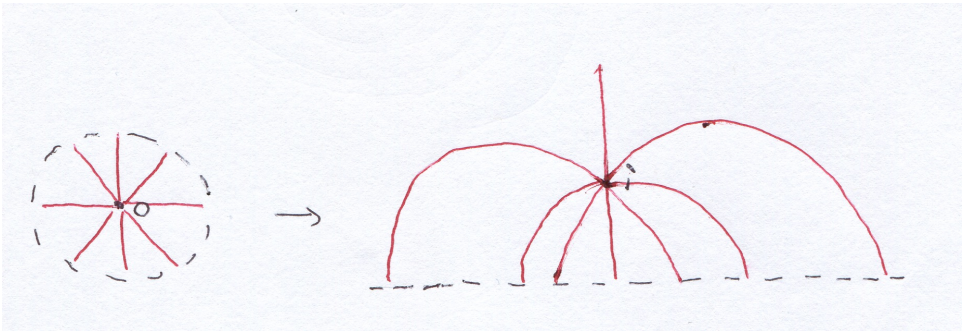


Figure 11.2: Disk is Conformally Equivalent to Upper Half Plane

Proof: Note that the conformal map from the unit disk to the upper half plane is linear fractional, and the automorphism group of the unit disk is linear fractional. It follows that the automorphism group of the upper half plane contains only linear fractional transformations.

Such a transformation preserves the upper half plane if and only if it maps the real line to itself, and does not interchange upper and lower half planes.

Suppose $\frac{az+b}{cz+d}$ is such a transformation At least one of a, b, c, d is non-zero. Say it is a and pick λ such that λa is real. Replace a, b, c, d by $\lambda a, \lambda b, \lambda c, \lambda d$.

If z and $\frac{az+b}{cz+d}$ are both real, then

$$\frac{az+b}{cz+d} = \overline{\left(\frac{az+b}{cz+d}\right)} = \frac{\overline{a}\,\overline{z}+\overline{b}}{\overline{c}\,\overline{z}+\overline{d}} = \frac{\overline{a}\,z+\overline{b}}{\overline{c}\,z+\overline{d}}$$

Thus the right and left sides of this equation are linear fractional transformations which agree on all real numbers, and thus on three numbers, and hence are equal up to a constant multiple. Since a is real and nonzero, $a = \overline{a}$ and the constant multiple is one. Hence a, b, c, and d are all real.

Conversely, if a, b, c, d are real and $ad - bc > 0$, then $\frac{az+b}{cz+d}$ preserves the real line and thus maps the upper half plane conformally to either the upper half of lower half plane. But $\frac{ai+b}{ci+d}$ is

$$\frac{(ai+b)(-ci+d)}{(ci+d)(-ci+d)} = \frac{(ad-bc)i}{c^2+d^2} + \frac{ac+bd}{c^2+d^2}$$

and this is in the upper half plane if and only if $ad - bc > 0$.

Remark: Notice that the automorphism groups of the plane, disk, and sphere are all transitive; every point can be mapped to every other point. This is highly unusual and almost never happens for other Riemann surfaces.

11.4 The Unit Disk and Non-Euclidean Geometry

The results in this section, while important, play no role in the proof of the Riemann mapping theorem.

Remark: The plane and sphere have natural metrics; orientation preserving isometries are conformal. The plane has additional conformal automorphisms (magnifications) and and so does the sphere (inversions in circles). It is a remarkable fact that the disk has a natural metric, and this time *all conformal* maps are orientation preserving isometries. The metric is exactly the non-Euclidean geometry discovered by Lobachevsky, Bolyai, and Gauss, although the connection between geometry and complex analysis was not made until toward the end of the 1800's, by Poincare.

We sketch the resulting theory. According to Riemann, a metric is best given as a quadratic form

$$ds^2 = \sum g_{ij}(x_1, x_2, \ldots, x_n) dx_i dy_j$$

This expression means that we can find lengths of curves $(\gamma_1(t), \dots, \gamma_n(t))$ for $a \leq t \leq b$ by computing

$$\int_a^b \sqrt{\sum g_{ij}(\gamma_1(t), \dots, \gamma_n(t)) \frac{d\gamma_1}{dt} \dots \frac{d\gamma_n}{dt}}\, dt$$

Curves that are locally the shortest between two points are called geodesics and these geodesics replace the *straight lines* of classical geometry. Curves γ have tangent vectors $X = \left(\frac{d\gamma_i}{dt}, \dots, \frac{d\gamma_n}{dt}\right)$ and the inner product between such tangent vectors is given by

$$< X, Y >= \sum g_{ij} X_i Y_j$$

from which angles of intesection of curves can be computed by the standard formulas.

For instance, in R^n we set $g_{ij} = \delta_{ij}$ and recover the usual formulas of advanced calculus.

We now introduce the metric of non-Euclidean geometry. The formula is simplest in the upper half plane, where it is

$$ds^2 = \frac{dx^2 + dy^2}{y^2}$$

In the unit disk, this becomes

$$ds^2 = 4\frac{dx^2 + dy^2}{(1 - (x^2 + y^2))^2}$$

These formulas say that distance over small regions is essentially computed by the standard formula from Pythagoras; however in the case of the upper half plane, distances very close to the x-axis are much larger than they look, and distances high above the x axis are much smaller than they look.

In the case of the disk, the formula says that distances close to the origin are as expected, but distances close to the boundary are much larger than they look. The constant 4 has been added to make the Gaussian curvature equal -1.

The crucial assertion is that our conformal automorphism groups preserve these metrics and become isometries of non-Euclidean geometry. Let's prove that. It is easiest to work in the upper half plane where the group elements are $\frac{az+b}{cz+d}$ for a, b, c, d real, and $ad - bc = 1$. We'll give the essential calculation, letting the reader disentangle that it really works. Let $\mathfrak{I}(z)$ be the imaginary part of z. Then

$$ds^2 = \frac{dx^2 + dy^2}{y^2} = \frac{(dx + idy)(dx - idy)}{\mathfrak{I}(z)^2} = \frac{dz\, \overline{dz}}{\mathfrak{I}(z)^2}$$

Suppose $w = \frac{az+b}{cz+d}$. Then $dw = \frac{(cz+d)a - (az+d)c}{(cz+d)^2}\, dz = \frac{1}{(cz+d)^2}\, dz$, so

$$dw\, \overline{dw} = \frac{dz\, \overline{dz}}{|cz + d|^4}$$

Also

$$\mathfrak{I}(w) = \frac{1}{2}\left(\frac{az+b}{cz+d} - \overline{\frac{az+b}{cz+d}}\right) = \frac{1}{2}\left(\frac{(az+b)(c\overline{z}+d) - (a\overline{z}+b)(cz+d)}{|cz+d|^2}\right)$$

$$= \frac{1}{2}\left(\frac{(ad-bc)(z-\overline{z})}{|cz+d|^2}\right) = \left(\frac{\mathfrak{I}(z)}{|cz+d|^2}\right)$$

Consequently

$$\frac{dw\,\overline{dw}}{\mathfrak{I}(w)^2} = \left(\frac{dz\,\overline{dz}}{|cz+d|^4}\right)\left(\frac{|cz+d|^4}{\mathfrak{I}(z)}\right) = \frac{dz\,\overline{dz}}{\mathfrak{I}(z)^2}$$

Remark: This result will imply that the conformal automorphisms of the unit disk preserve the Poincare metric on that disk if we can prove that the conformal map from the upper half plane to the unit disk, given by $z \to \frac{1+iz}{1-iz}$, maps the Poincare metric on the upper half plane to the Poincare metric on the unit disk. The metric on the upper half plane is $\frac{dz\,\overline{dz}}{\mathfrak{I}(z)^2}$ and the metric on the unit disk is $\frac{4dz\,\overline{dz}}{(1-|z|^2)^2}$ and we want to prove that $w = \frac{1+iz}{1-iz}$ carries the first to the second.

As before,

$$dw = \frac{(1-iz)i - (1+iz)(-i)}{(1-iz)^2} = \frac{2i}{(1-iz)^2}dz$$

and therefore

$$dw\,\overline{dw} = \left(\frac{2i}{(1-iz)^2}\right)\left(\frac{-2i}{1+i\overline{z})^2}\right)dz\,\overline{dz} = \frac{4}{(1+2\mathfrak{I}(z)+|z|^2)^2}dz\,\overline{dz}$$

Now

$$|w|^2 = \left|\frac{1+iz}{1-iz}\right|^2 = \left(\frac{1+iz}{1-iz}\right)\left(\frac{1-i\overline{z}}{1+i\overline{z}}\right) = \frac{1+iz-i\overline{z}+|z|^2}{1-iz+i\overline{z}+|z|^2} = \frac{1-2\mathfrak{I}(z)+|z|^2}{1+2\mathfrak{I}(z)+|z|^2}$$

and

$$(1-|w|^2)^2 = \left(1 - \frac{1-2\mathfrak{I}(z)+|z|^2}{1+2\mathfrak{I}(z)+|z|^2}\right)^2 = \left(\frac{4\mathfrak{I}(z)}{1+2\mathfrak{I}(z)+|z|^2}\right)^2$$

Putting these two results together, we have

$$\frac{4\,dw\,\overline{dw}}{(1-|w|^2)^2} = \frac{16}{(1+2\mathfrak{I}(z)+|z|^2)^2}dz\,\overline{dz}\left(\frac{1+2\mathfrak{I}(z)+|z|^2}{4\mathfrak{I}(z)}\right)^2 = \frac{dz\,\overline{dz}}{\mathfrak{I}(z)^2}$$

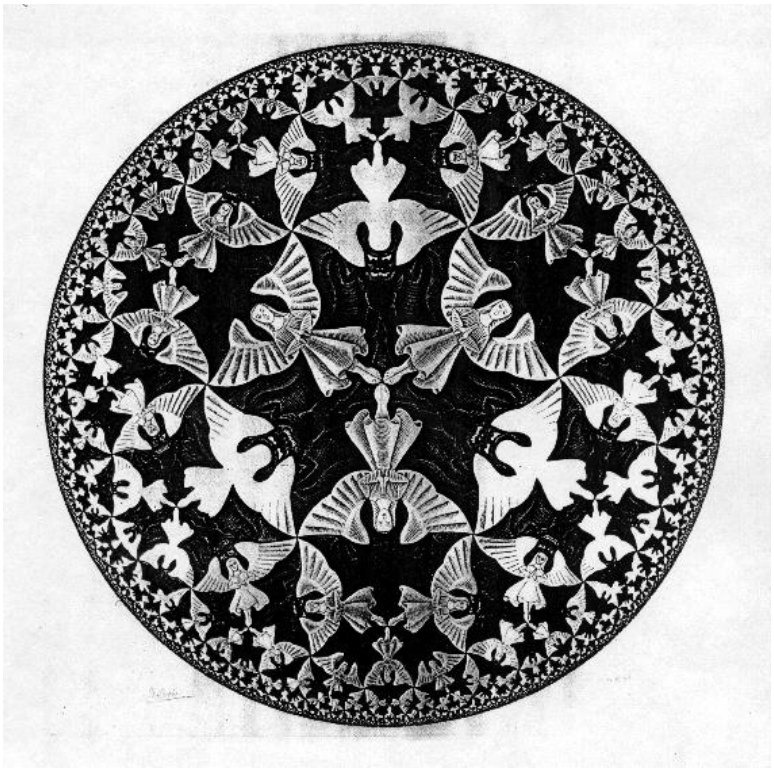


Figure 11.3: Escher's Angels and Devils

Remark: These calculations show that complex analysis on the unit disk is closely related to Non-Euclidean geometry, because the isometries of that geometry are just the complex automorphisms of the disk, and in particular are all linear fractional transformations.

Remark: Escher's wonderful work *Angels and Devils* is a picture of Non-Euclidean geometry. In the picture, each angel and devil has the same non-Euclidean size. As we approach the boundary, the angels and devils only seem to grow smaller; note that there are infinitely many of them. Some people believe that this picture was drawn on a sphere and then projected downward; that is certainly false because a sphere could contain only finitely many angles.

Remark: Let us determine the straight lines in this geometry, that is, the *geodesics*. We first claim that straight lines through the origin are geodesics. This becomes clear when we write the metric in polar coordinates:

$$ds = 2\frac{\sqrt{(dr)^2 + (rd\theta)^2}}{1 - r^2}$$

If we want to go from the origin to the boundary, the r variable must increase from $r = 0$ to r equal to one; any wiggles in θ will only increase the integral for curve length.

Once we know that lines through the origin are geodesics, it follows that images of these lines under automorphisms of the disk are also geodesics. The authomorphism of the disk are all linear fractional, and thus send lines and circles to lines and circles. Since straight lines through the origin meet the boundary circle perpendicularly, this must still hold after we map by an automorphism. We conclude that

Theorem 69. *The geodesics of the Poincare model of Non-Euclidean geometry are straight lines through the origin and circles which meet the boundary at 90 degrees.*

Remark: The previous illustration shows several of these geodesics. They are also visible in Escher's drawing. Notice that lines through the origin alternately bisect angels and devils. Notice that other curves can be seen alternately bisecting angels and devils. They are all circles meeting the boundary at 90 degrees.

We gather below other results from Non-Euclidean geometry.

Start with parallel lines. Select a line L and a point $p \notin L$. Drop a perpendicular from p to $q \in L$. At p, draw the line M perpendicular to this perpendicular. This new line is a parallel to L through p. See the illustration below.

In Non-Euclidean geometry, the sum of the angles of a triangle is always less than π, and consequently the sum of the angles of a quadralateral is always less than 2π. From this, it follows that no other line except the line from p to q is perpendicular to both L and M.

In the diagram, two red lines are shown. These are called *limiting parallels*; they are parallel to L and the lines through p between these two are exactly the lines through p parallel to L.

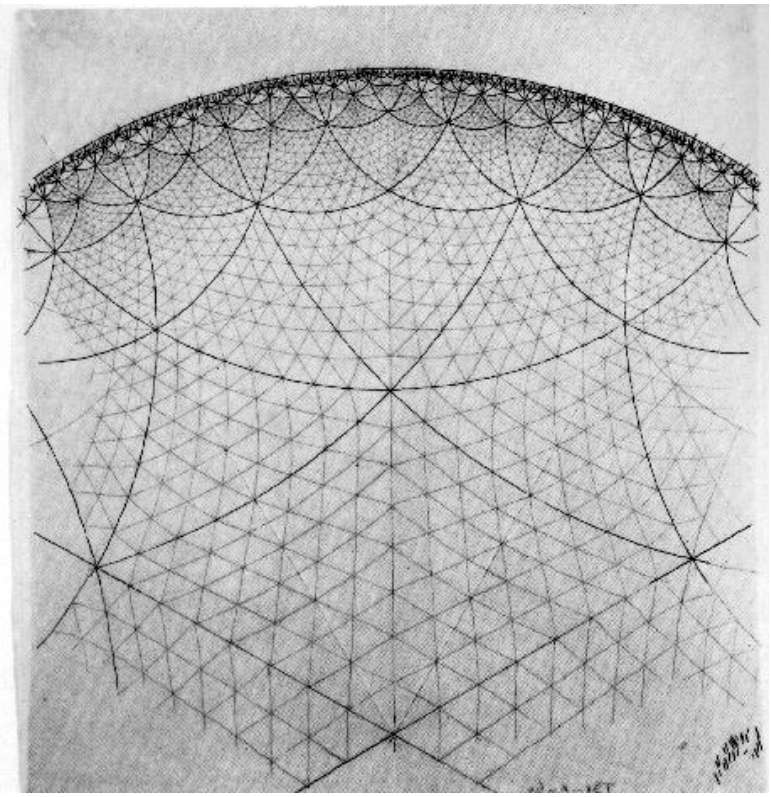


Figure 11.4: Poincare's Model for Non Euclidean Geometry

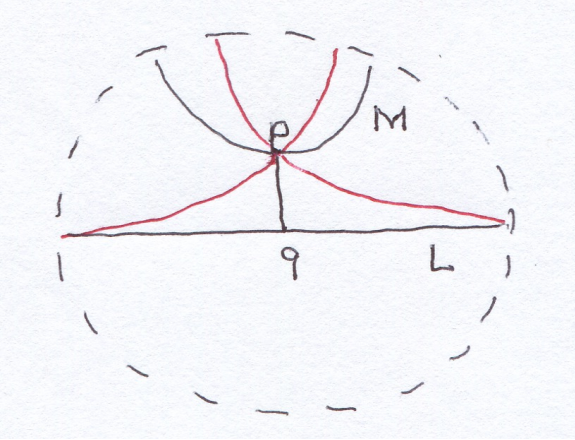


Figure 11.5: Parallel Postulate

In L and M are arbitrary parallel lines in Non-Euclidean geometry, either there is exactly one line segment between them that is perpendicular to both, or else there is no such line segment and the two lines are limiting parallels.

There is a beautiful formula relating the Non-Euclidean length a of the segment pq and the *angle of parallelism* Π, defined as the angle between pq and the first limiting parallel:

$$\cos \Pi = \tanh a = \frac{e^a - e^{-a}}{e^a + e^{-a}}$$

This result was proved in 1840 by Lobachevsky. If a is very small, $\tanh a$ is almost zero and $\cos \Pi$ is almost zero, so Π is almost $\frac{\pi}{2}$. Of course in Euclidian geometry this angle is exactly $\frac{\pi}{2}$. On the other hand, as $a \to \infty$, tanh approaches 1 and so $\cos \Pi \to 1$ and Π is almost zero.

Imagine that the universe is Non-Euclidean, but very large, so for distances a we are likely to encounter, Π is almost $\frac{\pi}{2}$ and the angle between limiting parallels is so small it would be undetectable.

Incidentally, Lobachevsky's formula shows that there is a natural unit of length in Non-Euclidean geometry, since a is measured in this unit and Π is measured in radians, as is usual for angles. There are no *magnifications* in Non-Euclidean geometry. We already know this another way, because the group of conformal automorphisms contains only isometries in the disk case. If the universe were Non-Euclidean, we could determine this natural unit of length by experiment. It would be very, very, very large.

Remark: Next we discuss circles. It is clear that a circle about the origin in the disk is a standard circle. But then a circle about any other origin is a standard circle because the isometry group maps circles to circles or lines. (The center of this circle isn't in the usual position.)

The circumference of a circle with Non-Euclidean radius r is

$$C = 2\pi \sinh r = 2\pi \left(r + \frac{r^3}{3!} + \frac{r^5}{5!} + ... \right)$$

and the area of a circle with Non-Euclidean radius r is

$$A = 2\pi(\cosh r - 1) = 2\pi \left(\frac{r^2}{2!} + \frac{r^4}{4!} + ... \right)$$

Remark: Next a few remarks on Non-Euclidean triangles. The sum of the angles of any triangle is smaller than π. But more is true:

Theorem 70. *The area of a triangle with angles α, β, γ is*

$$\pi - (\alpha + \beta + \gamma)$$

It follows that all triangles have area less than π. The extremal case is illustrated below.

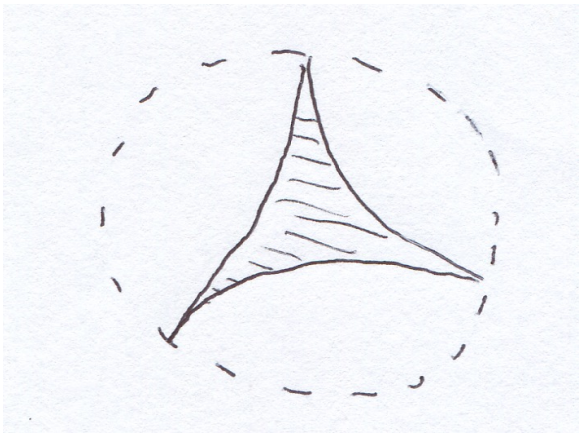


Figure 11.6: Limiting Triangle with Maximum Area

Remark: Similar remarks hold for quadrilaterals, and other polygons.

Remark: In Euclidean geometry, three distinct points in the plane determine a unique line or circle. This theorem is still true for the unit disk in the Euclidean plane, of course. But in terms of Non-Euclidean geometry, there is a puzzle: the curve determined by three points could be a line through the origin, or a circle which meets the boundary at 90 degrees, or a circle entirely inside the disk. But it could also be a line which does not go through the origin, or a circle which meets the boundary at two points but not with angle 90 degrees, or a circle which meets the boundary at just one point. The first three of these curves are Non-Euclidean lines and circles. The last three aren't.

Suppose L is a non-Euclidean line. At each point of this line, erect a perpendicular of length d. Connect the ends of these perpendiculars. This gives a new curve, and in Non-Euclidean geometry it is *not a straight line*. Such curves are called *equidistant curves*. In Poincare's model, these curves are straight lines or circles which meet the boundary at two points, but not at 90 degrees.

Curves which are circles meeting the boundary at exactly one point are called *horocyles*; these can be characterized with an internal property.

So three points in Non-Euclidean geometry determine a unique curve, which is either a straight line, a circle, an equidistance curve, or a horocycle.

Chapter 12

A Topology on the Set of Holomorphic Maps

If a power series $\sum c_n z^n$ has radius of convergence R, then the partial sums $f_n(z)$ converge to the total sum $f(z)$ uniformly on compact subsets of the open disk of radius R, since they converge uniformly on closed disks of radius $R_1 < R$.

At the end of chapter ten, we discovered that the topology of uniform convergence on compact subsets behaves unusually well for holomorphic functions. For instance, if $f_n \to f$ in this sense and each f_n is holomorphic, then the sum is holomorphic and $\frac{df_n}{dz} \to \frac{df}{dz}$, again uniformly on compact subsets.

If $\mathcal{U}$ is a domain, let $\mathcal{H}(\mathcal{U})$ be the set of all holomorphic functions on $\mathcal{U}$ and let $\mathcal{C}(\mathcal{U})$ be the set of all continuous complex-valued functions on $\mathcal{U}$. We want to formalize the previous remarks by making $\mathcal{H}(\mathcal{U})$ and $\mathcal{C}(\mathcal{U})$ into topological spaces with the property that $f_n \to f$ means f_n *converges to* f *uniformly on compact subsets of* $\mathcal{U}$.

Topological spaces induced by metrics have reasonable properties, but more general topologies can behave in unexpected ways. In particular, convergence of sequences determines the topology on metric spaces, but not in general. Therefore, we are going to produce a metric on our function spaces which induces the required meaning of $f_n \to f$. We will never refer to the metric again; it's sole purpose is to allow us to use topological arguments without worrying about subtle points.

12.1 A Metric on $\mathcal{H}(\mathcal{U})$ and $\mathcal{C}(\mathcal{U})$

It is easy to find a countable number of compact subsets $K_i \subset \mathcal{U}$ whose interiors cover $\mathcal{U}$. For instance, we can select closed disks inside $\mathcal{U}$ centered at rational points and with rational radii. If f and g belong to either of our function spaces, define

$$d(f,g) = \sum_i \frac{\min\left[\max_{K_i}(|f-g|), 1\right]}{2^i}$$

This sum converges because $\sum \frac{1}{2^i}$ converges. It is symmetrical in f and g. If $d(f,g) = 0$, then each term must be zero and therefore $f = g$ on each K_i; since the interiors of the K_i cover $\mathcal{U}$, $f = g$ everywhere. Finally, the triangle inequality holds because it holds for the numerators of each term.

Therefore, d makes $\mathcal{H}(\mathcal{U})$ and $\mathcal{C}(\mathcal{U})$ into metric spaces.

Suppose $f_n \to f$ in this metric, and suppose K is compact. Since the interiors of the K_i cover $\mathcal{U}$, we can find a finite number of K_i with $K \subset (K_1 \cup ... \cup K_n)$. Let ϵ be given. Choose ϵ_1 smaller than $\min\left(\frac{1}{2^n}, \frac{\epsilon}{2^n}\right)$. Choose N so that for $n > N$ we have $d(f_n, f) < \epsilon_1$. This inequality will fail if we use 1 in any of the numerators of the first n terms of the distance formula, so for $i \leq n$ we have $\frac{\max_{K_i}(|f_n - f|)}{2^i} < \epsilon_1 < \frac{\epsilon}{2^n}$. Consequently, $|f_n - f| < \epsilon$ on $(K_1 \cup K_2 \cup ... \cup K_n)$ and thus on K. It follows that $f_n \to f$ uniformly on compact subsets of $\mathcal{U}$.

Conversely, suppose $f_n \to f$ uniformly on compact subsets of $\mathcal{U}$. Let $\epsilon > 0$ be given. Choose n such that $\sum_{i>n} \frac{1}{2^i} < \frac{\epsilon}{2}$. Let $K = (K_1 \cup K_2 \cup ... \cup K_n)$. Choose N so large that $n > N$ implies $|f_n - f| < \frac{\epsilon}{2n}$ on each K_i. Then $n > N$ implies $d(f_n, f) < \epsilon$. Therefore $f_n \to f$ in the metric.

Remark: In the next few chapters we will construct important holomorphic functions using the topology of uniform convergence on compact subsets. Our $f_n(z)$ will be more complicated than mere powers of z, and the domain $\mathcal{U}$ will no longer be a disk.

However, we sometimes obtain $f(z)$ in a deeper manner, by using compactness arguments on special subsets of $\mathcal{H}(\mathcal{U})$. These arguments depend on a characterization of compactness we give next. This characterization uses deep properties of holomorphic functions and definitely is false for continuous functions or differentiable functions of a real variable.

12.2 The Key Technical Theorem

Theorem 71 (Montel). *Let D be a closed disk contained in an open domain $\mathcal{V}$. Let f_n be a sequence of holomorphic functions on $\mathcal{V}$. Suppose that this set is uniformly bounded on D, that is, suppose there is a bound B such that for all n we have $|f_n| \leq B$ on D. Then there is a subsequence of the initial sequence which converges uniformly on compact subsets of the interior of D.*

Remark: This theorem is definitely false when we replace $\mathcal{H}$ with $\mathcal{C}$, the space of continuous functions, or $\mathcal{C}^\infty$, the space of C^∞ functions. It doesn't matter if the domain of the functions is a subset of R, or of C, or of a more general space.

For instance, suppose $f_n(x) = 1$ for negative x, $f_n(x)$ is a straight line from $(0,1)$ to $(1/n, 0)$ on $[0, 1/n]$, and then $f_n(x) = 0$ as shown below. Clearly no subsequence converges uniformly. We could round the corners to get the same example for differentiable functions, and round more subtly to get it for C^∞ functions.

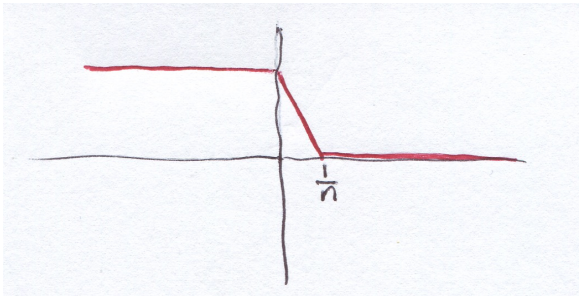


Figure 12.1: Montel's Theorem is False for Real Variables

Proof: Let γ be the counterclockwise boundary of the disk. If f is holomorphic in a neighborhood of the disk, then the derivatives of f at the origin are given by the generalized Cauchy formula

$$f^{(k)}(0) = \frac{k!}{2\pi i} \int_\gamma \frac{f(\zeta)}{\zeta^{k+1}}\, d\zeta$$

Suppose the radius of the disk is R. Apply this result to each f_n in our sequence to conclude that

$$|f_n^{(k)}(0)| \leq \frac{k!}{2\pi} \frac{B}{R^{k+1}} 2\pi R = \frac{Bk!}{R^k}$$

Since this bound does not depend on n, the values of $f_n(0)$ are bounded and we can find a subsequence of the f_n such that these values converge to a value we will call $f(0)$. Since the $\frac{df_n}{dz}(0)$ are bounded, we can find a subsequence of this subsequence so the $\frac{df_n}{dz}(0)$ also converge to a value we will call $\frac{df}{dz}(0)$. Continue. By a standard diagonal argument, we can find

a subsequence of the original sequence which is ultimately a subsequence of each of these subsequences, and therefore such that $f_n^{(k)}(0) \to f^{(k)}(0)$ for all k.

We now claim that the series

$$f(z) = \sum_k \frac{1}{k!} f^{(k)}(0) z^k = \sum c_k z^k$$

has radius of convergence at least R, and thus defines a function f, as implied by our notation. Indeed our inequalities show that $|f^{(k)}(0)| \leq \frac{Bk!}{R^k}$ and consequently that $|c_k| \leq \frac{B}{R^k}$. If $|z| < R$, then $|c_k z^k| \leq B \left|\frac{z}{R}\right|^k$ and this converges by comparison because $\left|\frac{z}{R}\right| < 1$.

To complete the proof, we must show that our subsequence converges uniformly to f on compact subsets of the interior of D. It is enough to show that it converges uniformly to f on $|z| \leq R_1 < R$. Pick a large N, and estimate the difference between f_n and f carefully for the first N terms and then more broadly for the remaining terms, as follows:

$$|f_n(z) - f(z)| \leq \sum_{k=0}^{N} \left| \frac{1}{k!} f_n^{(k)}(0) - f^{(k)}(0) \right| |z|^k + \sum_{k=N+1}^{\infty} \frac{2B}{R^k} |z|^k$$

If $\epsilon > 0$ is given and $|z| \leq R_1 < R$, we can make the last term smaller than $\frac{\epsilon}{2}$ by choosing N large enough. After that we can make the first term smaller than $\frac{\epsilon}{2}$ by choosing n large enough. QED.

12.3 Compact Subsets of $\mathcal{H}(\mathcal{U})$

The main theorem of the previous section contains all the hard work. In this section, we generalize to an arbitrary $\mathcal{U}$ using "abstract nonsense."

Definition 27. *A subset $\mathcal{H}_1$ of $\mathcal{H}(\mathcal{U})$ is* locally uniformly bounded *if for each compact K in $\mathcal{U}$ there is a uniform bound B such that the absolute value of every $f \in \mathcal{H}_1$ on K is at most B.*

Theorem 72 (Main Theorem; Montel)**.** *Let f_n be a sequence of elements of $\mathcal{H}(\mathcal{U})$ and suppose this sequence is locally uniformly bounded. Then there is a subsequence of the original sequence which converges in $\mathcal{H}(\mathcal{U})$.*

Corollary 22. *A subset $\mathcal{H}_1 \subset \mathcal{H}(\mathcal{U})$ is compact in the topology of uniform convergence on compact subsets if and only if it is locally uniformly bounded and closed.*

Proof: We proved the key technical theorem for disks centered at the origin, but of course it remains true for disks with any center.

Choose a countable number of compact disks $D_i \subset \mathcal{U}$ of radius R_i such that the corresponding open disks $\tilde{D}_i$ of radius $\frac{R_i}{2}$ cover $\mathcal{U}$. Our sequence is locally uniformly bounded on D_i. So by the technical lemma, we can find a subsequence of $\{f_n\}$ which converges uniformly on the closure of $\tilde{D}_1$. By the same lemma, we can find a subsequence of this subsequence which converges uniformly on the closure of $\tilde{D}_2$. Etc. By the diagonal argument, we can find a subsequence of the original sequence which is eventually a subsequence of each of these subsequences and so converges uniformly on the closure of each $\tilde{D}_i$. A finite number of D_i cover any compact subset of $\mathcal{U}$, so the final subsequence converges uniformly on each such compact set.

Proof of corollary: Since our $\mathcal{H}$ is a metric space, compactness of $\mathcal{H}_1$ follows from the main theorem. But we need to prove the converse, and for that it is enough to show that a compact subset of $\mathcal{H}$ must be locally uniformly bounded. This is easy. If $\mathcal{H}_1$ is compact but not locally uniformly bounded, then it is not uniformly bounded on some compact K. Find $f_n \in \mathcal{H}_1$ and $z_n \in K$ with $|f_n(z_n)| > n$. Suppose a subsequence $f_{n_i} \to f$ uniformly on K. This f is continuous on the compact K and hence bounded. Since $f_{n_i} \to f$ uniformly on K, the f_{n_i} are bounded on K, a contradiction.

Chapter 13

The Riemann Mapping Theorem

13.1 The Riemann Mapping Theorem

In his PhD thesis, Riemann proved the following amazing theorem:

Theorem 73 (Riemann Mapping Theorem)**.** *Let $\mathcal{U}$ be a simply-connected domain which omits at least one point of the plane. Then there is a one-to-one, onto holomorphic map $f : \mathcal{U} \to D$ to the open unit disk D about the origin.*

If $z_0 \in \mathcal{U}$, we can choose f so $f(z_0) = 0$. If g is a second such map, there is a constant θ such that $g(z) = e^{i\theta} f(z)$.

Remark: The condition $\mathcal{U} \neq C$ is essential, for if $f : C \to D$, then f is a bounded entire function and hence constant.

Remark: Simply connected sets can be very wild and have extremely bad boundaries. See pictures below. So it is amazing that a continuous map exists, let alone a conformal map preserving angles. Even more amazing is uniqueness. A topological proof that a continuous homeomorphism exists would likely involve arbitrary choices, perhaps even uncountably many. This proof involves essentially no choices.

Remark: Riemann's original theorem, from 1851, had different hypotheses and a defective proof. It took over 50 years, until around 1905, to reach the astonishing proof below. See the historical remarks after the proof.

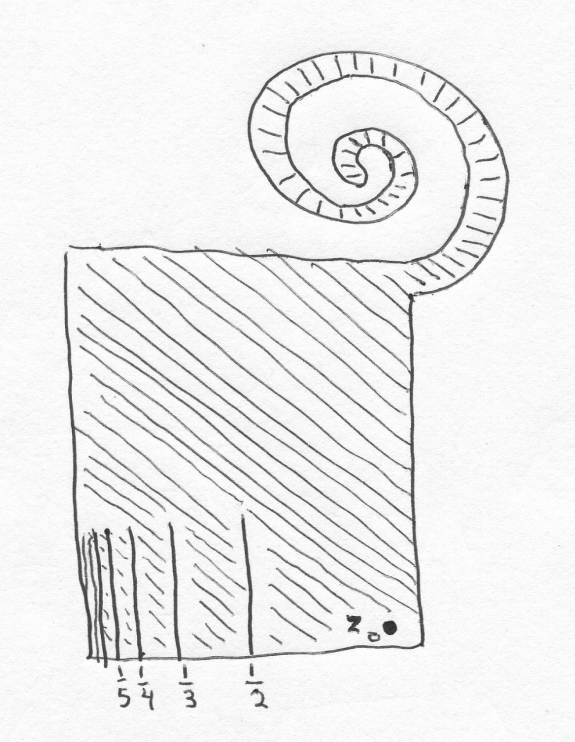


Figure 13.1: Simply Connected Set with Complicated Boundary

Remark: In the above example, an infinite comb of vertical lines has been removed from a square, with x-coordinates $\frac{1}{n}$. The remaining set is simply connected, and can be mapped to the unit disk with z_0 mapping to the origin. It is unclear whether boundary points have a well-defined limit.

Proof, Part 1: We first show that $\mathcal{U}$ can be mapped *into* the unit disk. The main problem is that $\mathcal{U}$ need not be bounded.

Since $\mathcal{U}$ is not everything, we can translate it so $0 \notin \mathcal{U}$. Since $\mathcal{U}$ is simply connected, we can define a branch of $\sqrt{z}$ on the translated $\mathcal{U}$. If λ and $-\lambda$ are non-zero complex numbers, at most one can be in the range of $\sqrt{z}$ on $\mathcal{U}$, because if $\lambda = \sqrt{z_1}$ and $-\lambda = \sqrt{z_2}$, then squaring gives $z_1 = z_2$. . But $\sqrt{z_0}$ is a non-zero number in the range of $\sqrt{z}$; by the inverse function theorem a whole neighborhood $\mathcal{V}$ of $\sqrt{z_0}$ is in the range of $\sqrt{z}$. So no values of $-\mathcal{V}$ are in the range. If $v \in -\mathcal{V}$, the map $f(z) = \frac{1}{z-v}$ is bounded on $\mathcal{U}$ and one-to-one. So the image is conformally equivalent to $\mathcal{U}$ and bounded. It is now easy to map this image into the unit disk by a one-to-one map.

Proof, Part 2: Since the automorphism group of the unit disk is transitive, we can compose

with an automorphism and get a map from $\mathcal{U}$ into the unit disk taking z_0 to zero. From now on, replace $\mathcal{U}$ by its image in the disk and assume $z_0 = 0$. We are going to consider a family of maps $f : \mathcal{U} \to D$ that are one-to-one and fix 0. The first of these maps is the identity, and we will require that all subsequent maps satisfy $|f'(0)| \geq 1$.

Proof, Part 3: We now come to the heart of the proof. If f is a one-to-one map from $\mathcal{U}$ into the disk fixing 0, we will define a positive real number $m(f)$ which measures how close $f(\mathcal{U})$ is to filling the disk. We'll prove that if f is not onto, we can define a new map g with bigger $m(g)$. Then we'll apply a compactness argument to find a map maximizing m.

The obvious definition of m might be the area of $f(\mathcal{U})$, but this definition doesn't work because the $\mathcal{U}$ below would have the maximal possible value and yet not be onto D.

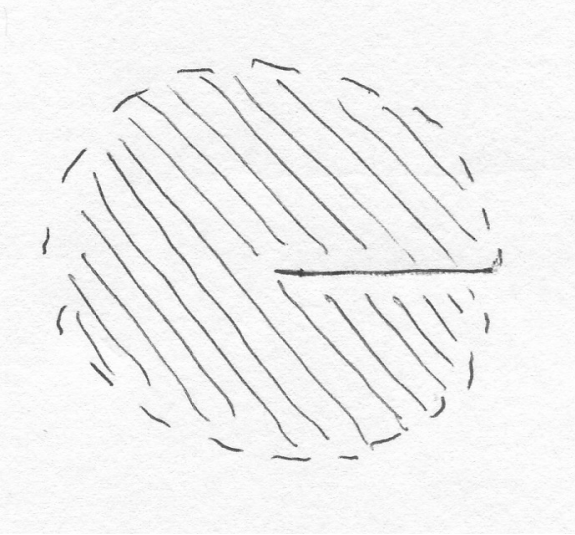


Figure 13.2: $m(f)$ = area fails

The stroke of genius, by I believe Koebe, is to select

$$m(f) = \left|\frac{df}{dz}(0)\right|$$

Proof, Part 4:

Lemma 7. *Suppose f is not onto. Then there is a one-to-one map $g : \mathcal{U} \to D$ such that $g(0) = 0$ and $|g'(0)| \geq 1$ and $m(g) > m(f)$.*

Indeed, suppose $w \notin f(\mathcal{U})$ and define

$$g(z) = \frac{z-\sqrt{-w}}{1-z\overline{\sqrt{-w}}} \circ \sqrt{z} \circ \frac{z-w}{1-z\,\overline{w}} \circ f(z)$$

Reading from right to left, the first map is f, which we want to improve. The second map is an automorphism of the disk which maps w to zero. Since w is not in the image of f, the composition of these two maps is never zero, so we can define a branch of the square root on this composition. The third map is this square root. We have already proved that the square root is one-to-one, so the full composition of the last three maps is a one-to-one map from $\mathcal{U}$ into D. This map sends the origin to $\sqrt{-w}$. The final map is an automorphism of the disk sending $\sqrt{-w}$ back to the origin. So the entire composition is a one-to-one map g from $\mathcal{U}$ into the disk sending the origin to the origin.

We next compute $m(g) = \left|\frac{dg}{dz}(0)\right|$. For this we need

$$\frac{d}{dz}\frac{z-a}{1-z\,\overline{a}} = \frac{(1-z\,\overline{a})-(z-a)(-\overline{a})}{\left(1-z\,\overline{a}\right)^2} = \frac{1-|a|^2}{(1-z\,\overline{a})^2}$$

The required derivative is

$$\left.\frac{1-|\sqrt{-w}|^2}{(1-z\,\overline{\sqrt{-w}})^2}\right|_{\sqrt{-w}} \cdot \left.\frac{1}{2\sqrt{z}}\right|_{-w} \cdot \left.\frac{1-|w|^2}{(1-z\,\overline{w})^2}\right|_0 \cdot f'(0)$$

$$= \frac{1-|\sqrt{-w}|^2}{(1-|\sqrt{-w}|^2)^2} \cdot \frac{1}{2\sqrt{-w}} \cdot (1-|w|^2) \cdot f'(0) = \frac{1+|w|}{2\sqrt{-w}} f'(0)$$

The absolute value of this expression is larger than $|f'(0)|$ because

$$\frac{1+|w|}{2\sqrt{|w|}} > 1$$

Indeed, if $0 < a < 1$, then $\frac{1+a^2}{2a} > 1$.

Proof, Part 5: So much for the concrete part of the proof. Let $\mathcal{H}_1 \subset \mathcal{H}(\mathcal{U})$ be the set of all one-to-one holomorphic maps $\mathcal{U} \to D$ such that $f(0) = 0$ and $|f'(0)| \geq 1$. Since each element of $\mathcal{H}_1$ is bounded by 1, Montel's theorem implies that $\mathcal{H}_1$ is compact provided it is closed. Suppose $f_n \to f$ where $f_n \in \mathcal{H}_1$. Then f is holomorphic on $\mathcal{U}$ and $f(0) = 0$ and $|f'(0)| \geq 1$. We claim that f is one-to-one. This surprising result follows from the following lemma, since $|f'(0)| \geq 1$ implies that f is not contant.

Lemma 8. *Suppose $f_n \to f$ uniformly on compact subsets of a domain $\mathcal{U}$. If each f_n is one-to-one, then either f is one-to-one or else it is constant.*

Proof: Suppose not, so $f(z_1) = f(z_2)$ where $z_1 \neq z_2$. Replace $\mathcal{U}$ by $\mathcal{U} = \{z_2\}$, $f_n(z)$ by $f_n(z) - f_n(z_2)$, and $f(z)$ by $f(z) - f(z_2)$. Then f_n is never zero and $f_n \to f$, but f has a zero. If f is not constant, this zero is isolated, so we can find a small counterclockwise circle γ around the zero with no other zeros on or inside the circle. Then $\frac{f_n'}{f_n} \to \frac{f'}{f}$ uniformly on the circle, and so

$$\frac{1}{2\pi i}\int_\gamma \frac{f_n'(\zeta)}{f_n(\zeta)}\,d\zeta \to \frac{1}{2\pi i}\int_\gamma \frac{f'(\zeta)}{f(\zeta)}\,d\zeta$$

By the argument principle, this integral counts zeros inside γ, so the left side is zero and the right side is at least one, a contradiction.

13.2 History of the Riemann Mapping Theorem

In Riemann's original formulation, the open domain $\mathcal{U}$ has a boundary and the map is to carry $\mathcal{U}$ to the interior of a disk and the boundary of $\mathcal{U}$ to the boundary circle. Riemann claimed the map was unique once we fix the image of one internal point z_0 and one boundary point b_0.

There is no discussion of possible constraints on the boundary. Our earlier illustration shows that the boundary of general $\mathcal{U}$ can be very complicated. Later authors restricted to piecewise differentiable boundaries, or simple closed curves which only cross themselves at the endpoints.

Alfors wrote in 1953 that Riemann wrote "almost cryptic messages to the future" and his statement of the theorem has a form that "would defy any attempt at proof, even with modern methods."

In some sense, Riemann's proof relied on a boundary value problem for harmonic functions on a domain. Here are sketchy details, some in the form of exercises.

Exercise 2. *When discussing the Cauchy-Riemann equations, we wrote $f = f_1 + if_2$. Here we adopt the alternate and common notation $f = u + iv$. Using the Cauchy-Riemann equations, show that if f is holomorphic then u is harmonic, that is*

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

Exercise 3. *Suppose we know u and want to find v so $u + iv$ is holomorphic. By the Cauchy-Riemann equations, we must solve*

$$\frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}$$

$$\frac{\partial v}{\partial y} = \frac{\partial u}{\partial x}$$

By advanced calculus, there is an integrability condition which must hold before these equations can be solved. Show that the integrability condition is exactly the statement that u is harmonic.

Exercise 4. *If u is harmonic on a simply-connected region $\mathcal{U}$, show that v exists on $\mathcal{U}$ making $u + iv$ holomorphic.*

Remark: We want a conformal map $f : \mathcal{U} \to D$ taking z_0 to zero. If this map is one-to-one, then z_0 must be a first order zero and we can form $\frac{f(z)}{z-z_0}$ a nowhere zero function on $\mathcal{U}$. Since $\mathcal{U}$ is simply connected, this function equals e^g for some $g = u + iv$. So we can assume that

$$f = (z - z_0)e^{g(z)} = e^{\ln|z-z_0| + i\arg(z-z_0)}e^{u+iv}$$

On the boundary of our $\mathcal{U}$ we want f to have absolute value one, so we must choose a harmonic u inside $\mathcal{U}$ with boundary values $-\ln|z - z_0|$.

Riemann was interested in physics as well as mathematics, and particularly interested in the theory of electricity and magnetism. In physics, harmonic functions describe steady-state conditions when a system has settled into a state independent of time. For instance, if we have a region with an insulated boundary and spray charges on the boundary, the potential inside is given by a harmonic function with boundary values given by the charges. Thus in physics, it is natural to assume that u exists.

Riemann also knew of the Dirichlet principle. If u represents the potential energy function inside a region, the total energy is given by

$$\int\int_{\mathcal{R}} \left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 \right] dxdy$$

In a steady state, the potential energy should minimize this expression, and the Euler-Lagrange equations for this minimal energy state are exactly the condition that u be harmonic. So to obtain u, we minimize this expression among all possible u satisfying the given boundary conditions.

Let us concentrate on this much of the proof, which produces u. We could then obtain v on the interior, but would still need to show that it also has appropriate boundary conditions.

There is an interesting paper on the history of the Riemann Mapping theorem by Jeremy Grey, Rendiconti Del Circolo Mathematico Di Palermo, Serie II, Supplemento N. 34 (1994). (Also available on the internet.) From this paper, we learn that Riemann's actual proof was more complicated than this sketch. However, Riemann's proof was shown to be invalid in 1871 by his student Prym. Around the same time, Weierstrauss found an example of a calculus of variations problem for which the analogue of the Dirichlet principle failed, and thus managed to

convince mathematicians that an approach via the Dirichlet principle and harmonic functions was suspect.

The next period from 1887 to 1910 was one of extensive work on the mapping theorem by Harnack, Hilbert, Osgood, Cartheodory, and Koebe. Koebe was a particularly important contributor.

In 1887, Harnack published a book which dramatically extended the theory of harmonic functions and solved the Dirichlet problem on many domains. The last pages of this book show how the theory can be used to prove the mapping theorem. Hilbert also worked on justifying the use of the Dirichlet principle for harmonic functions. Building on these results, the American mathematician Osgood may have been the first person to give a rigorous proof of the Riemann mapping theorem, between 1900 and 1902. He cleanly separated the case of mapping the interior from discussions of the boundary, and rigorously proved our form of the mapping theorem. He also proved that if the boundary is a continuously differentiable simple closed curve, then the map of the interior extends continuously to the boundary, and called attention to the possibility that this result might be true for an arbitrary continuous closed boundary curve which does not cross itself.

Around the same time, Jordan proved the Jordan curve theorem: if $\varphi : [0,1] \to C$ is continuous and $\varphi(0) = \varphi(1)$ and otherwise φ is one-to-one, then φ separates the plane into two connected regions, one bounded and one unbounded. His proof was criticized, and then improved by Veblen, Schoenflies, and Brouwer in 1909.

Working with these results, Osgood proved that such a Jordan curve can have positive area. This made clear the difficulty of extending the mapping theorem to Jordan boundaries, let along more general boundaries.

Caratheodory was encouraged by Klein and Hilbert to work on the mapping theorem. In papers published in 1912 and 1913, he replaced Riemann's uniqueness result that f is determined by $f(z_0)$ and $f(b_0)$ where b_0 is a boundary point, with the assertion that f is determined by $f(z_0$ and the argument of $f'(z_0)$ and then very cleanly proved the Riemann mapping theorem on the interior of the domain. According to Gray, this is the first purely function theoretic proof, with no appeal to Harnack's work, replacing that approach by the use of Schwarz's lemma.

In a second paper, Caratheodory proved that the conformal map extends to a homeomorphism of the boundaries if and only if the boundary is a simple Jordan curve in the sense described above.

Gray ends his history of the Riemann mapping theorem by writing: "That the boundary of a domain could be as strange as Osgood and Caratheodory had discovered was a powerful stimulus to precision in these matters. It is reasonable to claim that in proving the Riemann mapping theorem, ... mathematicians came to unite geometric function theory with topology, to the mutual advantage of both."

Indeed, a consequence of Caratheodory's 1913 paper is that if $\varphi : S^1 \to C$ is a simple closed curve, then there is a homeomorphism of the plane mapping the image of φ to the unit circle, and mapping the bounded component of the complement conformally to the disk and the unbounded component conformally to the exterior of the disk. The Jordan curve theorem generalizes this result by proving that if $S^{n-1} \to R^n$ is a one-to-one continuous map, then the complement of the image has two components, one bounded and one unbounded. Around 1920, Alexander proved that in three dimensions, the bounded component is homeomorphic to the open ball $B_3 \subset R^3$. His paper was accepted. As Alexander waited for publication, he came to have more and more doubts, and eventually he found the *Alexander horned sphere*, a counterexample which made him immortal. So the authors of papers on Riemann's result were working on a cliff.

In the midst of the above development, Poincare transformed the subject. In the early part of the 19th century, a central topic was the study of meromorphic functions on a torus, or what is the same thing, doubly periodic meromorphic functions on the plane. We later devote a chapter to that theory. Now Poincare attempted to generalize this subject by replacing the plane by the unit disk. This is a natural object to consider because the disk has a large number of automorphisms, which can be used to map regions to other regions in the same manner that translations map rectangular regions in the plane to other rectangular regions.

Poincare had to find subgroups Γ of the automorphism group which are *discrete*. A *fundamental region* for such a group is a domain $\mathcal{U}$ in the disk such that

- no two elements of $\mathcal{U}$ are equivalent under Γ
- each element of the entire disk is equivalent under Γ to at least one point in the closure of U.

In that case, the elements $\gamma(\overline{\mathcal{U}})$ cover the disk in the same way that rectangles cover the plane. Memomorphic functions on the disk which are invariant under Γ correspond to meromorphic functions on the Riemann surface D/Γ.

Applying the conformal equivalence of the disk and the upper half plane, Poincare's task was equivalent to finding discrete subgroups of $SL(2, R)$. A few such subgroups like $SL(2, Z)$ were known from the earlier theory, and when finding other examples proved difficult, Poincare at first tried to prove that the known examples were the only possibilities.

At this point we can follow Poincare exactly, because he later wrote a book called *The Psychology of Mathematical Invention* about his next steps. Poincare was about to go on vacation, and the night before he drank strong coffee. Then he found that he could not sleep. The next morning, just as he was about to step on the bus, it occurred to him that the unit disk behaves like non-Euclidean geometry. Poincare reports that he then relaxed on vacation, and wrote up the results when he returned to Paris.

The connection to non-Euclidean geometry allowed Poincare to define regions directly using

the kinds of arguments in Euclid, but suitably modified for the non-Euclidean case. This quickly led to discrete groups whose quotient spaces D/Γ were compact surfaces of genus $g \geq 2$. Poincare made the audacious guess that *all such surfaces* arise in this way. The proof turned on the following generalization of the Riemann mapping theorem:

Theorem 74 (The Uniformization Theorem)**.** *Every simply connected Riemann surface is conformally equivalent to S^2, D, or C*

Much of the work on the mapping theory from 1900 on was directed toward proving this generalization. Alfors once wrote that the Riemann mapping theorem is "one of the most important theorems of complex analysis" and the uniformation theorem is "perhaps the single most important theorem in the whole theory of analytic functions of one variable."

13.3 A Short Sketch of Consequences of the Uniformization Theorem

13.3.1 Manifolds

Having talked of Riemann surfaces several times, we finally give a definition. Let us start with ordinary continuous surfaces.

Definition 28. *An n-dimensional manifold M is a topological Hausdorff space which is second countable, together with a covering of M by coordinate neighborhoods $(\mathcal{U}, \varphi, \mathcal{V})$. Here $\mathcal{U} \subset M$ is an open set in M, φ is a homeomorphism from $\mathcal{U}$ to $\mathcal{V}$, and $\mathcal{V} \subset R^n$ is an open set in R^n.*

The essential idea here is that small pieces of M look topologically like small pieces of Euclidean space. The Hausdorff and second countable conditions eliminate bizarre examples like the long line and will not be mentioned again. Note that $\mathcal{U}$ varies over only *some* of the open sets; to give M, it suffices to give a covering of M by enough such coordinate systems.

Remark: If $(\mathcal{U}, \varphi, \mathcal{V})$ is a coordinate neighborhood, then each point of $\mathcal{V}$ has coordinates $(x_1, x_2, \dots, x_n)$ and we can use these to describe points on the manifold via the map φ. Usually, we ignore all the fancy words and just say "suppose we have coordinates $(x_1, \dots, x_n)$ defined on $\mathcal{U}$."

Definition 29. *If $\mathcal{U}_1, \varphi_1, \mathcal{V}_1$ and $\mathcal{U}_2, \varphi_2, \mathcal{V}_2$ are coordinate systems, we get a map from the plane to the plane as in the following picture. It is called the* glueing map*. Such maps explain how the various coordinate systems glue together.*

Remark: This map is only defined over the common intersection. In more detail

$$\varphi_2 \circ (\varphi_1)^{-1} : \varphi_1(\mathcal{U}_1 \cap \mathcal{U}_2) \to \varphi_2(\mathcal{U}_1 \cap \mathcal{U}_2)$$

Example: Recall that we made the Riemann sphere into a manifold by covering it with two coordinates $\mathcal{U}_1 = C \xrightarrow{z} C$ and $\mathcal{U}_2 = C - \{0\} \cup \{\infty\} \xrightarrow{\frac{1}{z}} C$, so the glue map is $1/z$.

Example: Usually we avoid this fancy terminology. The point is that some points of M can belong to intersections with two coordinates: $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$. The glue map explains how to convert from x's to y's or conversely. This is done by giving conversion functions $y_i = y_i(x_1, \ldots, x_n)$.

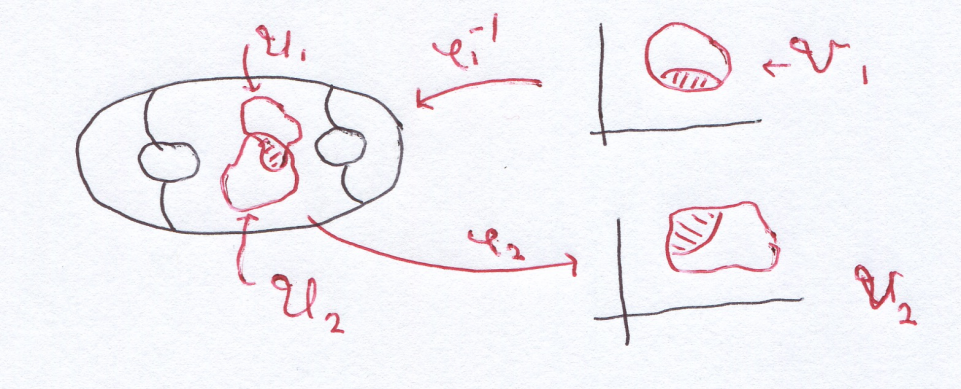


Figure 13.3: Glue map $\varphi_2 \circ (\varphi_1)^{-1}$

If all of the glue maps are C^∞ maps from the plane to the plane, we say that M is a C^∞ *manifold*.

If the dimension of M is two and all of the glue maps are *holomorphic*, we say that M is a *Riemann surface*.

Moise and Munkries proved that any manifold of dimension less than or equal to three has a differentiable structure, which is unique up to differomorphism. (This is dramatically false in dimensions four and above.) So we can assume that surfaces are C^∞.

13.3.2 Riemannian Manifolds

If we have a C^∞ manifolds, we can give it an additional structure allowing us to compute distances and angles; such a structure is known as a *Riemannian metric*. Suppose we have local coordinates near a point p, allowing us to introduce coordinates $(x_1, \ldots, x_n)$. Then if $\gamma_1(t)$ and $\gamma_2(t)$ are differentiable curves through p, we can compute $\frac{d\gamma_1}{dt} = X$ and $\frac{d\gamma_2}{dt} = Y$ and get two tangent vectors at p. A Riemannian structure on M assigns an inner product $< X, Y >$ to any two such vectors defined at p. This makes it possible for us to compute angles between vectors, and lengths of vectors. By linear algebra,

$$< X, Y >= \sum g_{ij} X_i Y_j$$

where g_{ij} is a symmetric positive definite matrix, so equivalently a Riemannian structure is given by functions g_{ij} on each coordinate system.

Riemannian structures were defined on surfaces by Gauss in his great work on differential geometry. The extension to higher dimensional manifolds was sketched by Riemann in his required Habilitationsschrift (inaugural lecture) for the PhD (Gauss was on his committee).

For surfaces, Gauss defined a quantity called the *Gaussian curvature* and explained how to compute this function given the g_{ij}. A necessary and sufficient condition condition that a change of coordinates converts g_{ij} to the Euclidean form δ_{ij} is that $k = 0$ nearby. In other cases, k serves to distinguish various geometries defined by the g_{ij}. Riemann showed how to extend this k to the *Riemannian curvature tensor*, an analogous object in higher dimensions.

The quantity k is usually a function, which makes it difficult to handle. In special cases, however, it is a constant. By magnifying, this constant is either 1, 0, or -1. In these cases, Gauss proved that k completely determines the local geometry. The cases in question are

- $k = 1$: standard sphere
- $k = 0$: standard flat plane
- $k = -1$: non-Euclidean geometry of Bolyai and Lobachevsky

Remark: After Riemann's work, this division into three cases was extended to all dimensions, and the crucial hypotheses became clearer. An *isometry* of a Riemannian manifold is a one-to-one and onto map which preserves inner products, and so distances and angles. A fairly easy theorem asserts that if p and q are points and $v_1, \ldots, v_n$ is an orthonormal basis of tangent vectors at p and $w_1, \ldots, w_n$ is a similar basis of tangent vectors at q, then there is *at most one* isometry mapping p to q and the v_i to the w_i. So an isometry is determined by the image of p and the corresponding rotation about this image. For most Riemannian M, isometries are rare.

Remark: But suppose M is a Riemannian manifold such that all possible isometries occur. The great theorem of the subject asserts that there are only four possibilities up to magnification (provided the dimension of M is at least two). M can be the sphere S^n or the plane R^n, or the generalized non-Euclidean geometry H^n. Or finally, M can be S^n with opposite points identified. This space is called *projective space*. Since it is locally isometric to S^n, many people ignore it and assert that there are only three *very symmetric* Riemannian manifolds.

Every C^∞ manifold can be given a Riemannian metric. The argument employs partitians of unity, that is, functions which are identically 1 near a point p, but then rapidly vanish and are identically zero off the coordinate system. A partition of unity is a collection φ_α of such functions such that $\sum \varphi_\alpha = 1$. To get a global g_{ij}, we define g_{ij} in each coordinate systems, and then form $\sum \varphi_\alpha g_{ij,\alpha}$.

We mention this technique because it would never be used in complex analysis, since the identity theorem would imply that each φ_α is identically zero.

13.3.3 Riemann Surfaces

Given g_{ij} on a surface, coordinates exist such that $g_{ij} = \delta_{ij}$ only when the curvature is zero. However, Gauss and later many others proved that *isothermal coordinates* always exist, where $g_{ij} = \lambda(x, y)\delta_{ij}$. This means that every metric is Euclidean up to magnifications which vary from place to place. In particular, any metric is conformally equivalent to the standard metric of the plane.

Suppose, then, that M is an orientable Riemannian surface. Cover M by isothermal coordinates. The glueing maps are than conformal and hence holomorphic, so M becomes a Riemann surface.

13.3.4 Covering Surface

In topology, there is a beautiful theory of *covering spaces*. It won't be described here, but will be used (only in this chapter). Incidentally, after a twenty year gap, Boubaki has published a new volume, about covering spaces and the fundamental group.

Suppose M is a Riemann surface. Then we can form its universal covering space, $\tilde{M}$. Since M is covered by coordinates with holomorphic glueing maps, this also holds for $\tilde{M}$, which becomes a Riemann surface. Moreover, the deck transformation group maps must be holomorphic.

13.3.5 The Uniformization Theorem

Applying the uniformization theorem, we conclude that $\tilde{M}$ is either the Riemann sphere, the complex plane, or the unit disk. Moreover, the deck transformation group Γ is a discrete subgroup of the automorphism group of $\tilde{M}$ without fixed points, so if $\gamma \neq id$, then $\gamma(z) \neq z$.

A short calculation shows that every linear fractional transformation has fixed points. So if the universal covering space is the sphere, then the original surface is also the sphere. We conclude that

Theorem 75. *The Riemann sphere has a unique Riemann surface structure. Consequently, any Riemannian surface homeomorphic to a sphere has a conformally equivalent second metric equal to the standard metric on a sphere.*

Remark: If the universal cover is the plane, then every deck transformation has the form $\varphi(z) = az+b$. Such maps have fixed points unless $a = 1$. Consequently the deck transformation group must be a discrete subgroup formed only of translations. It is easy to conclude that the only possibilities are $z \to z + l$ where l runs through all elements of a lattice of rank 0, 1, or 2.

Theorem 76. *The only surfaces with universal cover C are C, cylinders, and tori. If a Riemannian manifold is homeomorphic to one of these, it has a conformally equivalent second metric which is flat.*

Remark: Cylinders are conformally equivalent to annuli.

Remark: In all other cases, the universal cover is a disk. Astonishingly, the entire automorphism group of a disk preserves the non-Euclidean metric. Consequently, and astonishingly

Theorem 77. *Every Riemannian surface except the sphere, the complex plane, cylinders, and tori can be given a conformally equivalent metric locally equivalent to non-Euclidean geometry.*

Remark: This result is false in higher dimensions. But Thurston conjectured that it is essentially correct in dimension three. The three geometries of the theorem must be replaced by eight geometries, but all except the non-Euclidean case can essentially be analyzed completely. Moreover, a three-dimensional manifold must be cut into pieces and then the result says that each piece can be given one of these eight geometrical structures.

Remark: This difficult conjecture implies the Poincare conjecture. Like the uniformization theorem, it is proved using difficult theorems in analysis about differential equations. The complete proof has now been given by Hamilton and Perelman, with expositions from a number of Chinese mathematicians.

Chapter 14

The Theorems of Mittag-Leffler and Weierstrass

14.1 The Mittag-Leffler Theorem

Definition 30. *If z_0 is a pole of f, the* principal part *of f at z_0 is the sum of negative terms in the Laurent expansion of f about z_0:*

$$c_{-k}\frac{1}{(z-z_0)^k} + c_{-k+1}\frac{1}{(z-z_0)^{k-1}} + \ldots + c_{-1}\frac{1}{z-z_0}$$

More generally, a principal part *$\mathcal{P}(z)$ at z_0 is a polynomial in $\frac{1}{z-z_0}$ of degree at least one and without constant term.*

Theorem 78 (Mittag-Leffler)**.** *Let $z_1, z_2, \ldots$ be a sequence in the complex plane with no limit points. Let $\mathcal{P}_k(z)$ be a principal part at z_k, for each k. Then there is a meromorphic function $f(z)$ on C with poles precisely at the z_i, and principal part $\mathcal{P}_k$ at z_k.*

Proof: If one of the z_i is zero, set it aside and let $\mathcal{Q}_i = \mathcal{P}_i$. For each remaining k, expand $\mathcal{P}_k$ in a power series $\sum c_i z_i$ about the origin. This series converges for $0 < |z_k|$, and converges uniformly for $0 \leq \frac{|z_k|}{2}$. Let $\mathcal{Q}_k$ be

$$\mathcal{Q}_k = \mathcal{P}_k - \sum_{i=0}^{j} c_i z^i$$

where j has been chosen so $|Q_k| \leq \frac{1}{2^k}$ on $0 \leq \frac{|z_k|}{2}$. Let

$$f(z) = \sum \mathcal{Q}_k(z)$$

There are many ways to see that this converges appropriately and solves the problem. For example, choose a large open disk $\mathcal{U}$ of radius R centered about the origin. The closed disk of radius $2R$ is compact, so only finitely many z_k are inside this disk. Choose M so $k > M$ implies z_k is not in the larger disk. Then $f(z) = \sum_{k=1}^{M} \mathcal{Q}_k + g(z)$ where $g(z)$ is the sum of the remaining terms. Each $\mathcal{Q}_k$ in this remaining sum has absolute value less than $\frac{1}{2^k}$ on $\mathcal{U}$, so $g(z)$ is a uniform limit of holomorphic functions on $\mathcal{U}$ and thus holomorphic. QED.

Theorem 79 (Mittag-Leffler). *Let $\mathcal{U}$ be a domain in the plane. Let $z_1, z_2, \ldots$ be a sequence in $\mathcal{U}$ with no limit points in $\mathcal{U}$. Let $\mathcal{P}_k(z)$ be a principal part at z_k, for each k. Then there is a meromorphic function $f(z)$ on $\mathcal{U}$ with poles precisely at the z_i, and principal part $\mathcal{P}_k$ at z_k.*

Proof: The z_k can now have limit points on the boundary of $\mathcal{U}$; limit points can also occur at infinity.

For each k, let

$$d_k = \inf_{w \notin \mathcal{U}} |z_k - w|$$

Consider two sets of indices

$$\mathcal{A} = \{\, k \mid d_k |z_k| \leq 1 \,\}$$

$$\mathcal{B} = \{\, k \mid d_k |z_k| > 1 \,\}$$

In some sense, A contains indices of z_k which go to the boundary faster than they go to infinity, and B contains indices of z_k which go to infinity faster than they go to the boundary.

We will define

$$f(z) = f_1(z) + f_2(z)$$

where f_1 handles poles for indices in $\mathcal{A}$ and f_2 handles poles for indices in $\mathcal{B}$. Our first claim is that the previous version of the Mittag-Leffler theorem handles $\mathcal{B}$. Indeed, if the corresponding z_k had a finite limit point in the plane, this limit cannot be in $\mathcal{U}$, and so for an appropriate subsequence, $d_k \to 0$. Since $d_k |z_k| > 1$, we conclude that z_k are unbounded, a contradiction.

So it suffices to construct f_1 for $\mathcal{A}$.

If $k \in \mathcal{A}$, a compactness argument finds a b_k in the boundary of $\mathcal{U}$ with $d_k = |z_k - b_k|$. Expand the principal part $\mathcal{P}_k$ about b_k in a Laurent series which converges in the annulus $|z - b_k| > d_k$. See the picture near the top of the next page.

For example, suppose $\mathcal{P} = \frac{1}{z-1}$ and $b_k = 0$. The expansion we need is

$$\mathcal{P} = \frac{1}{z-1} = \frac{1}{z\left(1 - \frac{1}{z}\right)} = \frac{1}{z} \sum_{k \geq 0} \left(\frac{1}{z}\right)^k = \sum_{k \geq 1} \left(\frac{1}{z}\right)^k$$

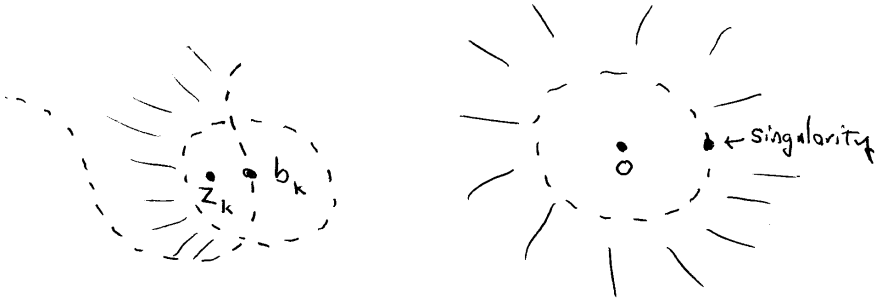


Figure 14.1: Modifying Principal Parts to Force Convergence

In general, this Laurent series will have only terms of negative powers and converge uniformly to $\mathcal{P}$ as long as we stay away from the inner boundary. For each k, choose j such that

$$\mathcal{Q}_k = \mathcal{P}_k - \sum_{i=1}^{j} c_{-i} \frac{1}{(z-b_k)^i}$$

is smaller than $\frac{1}{2^k}$ on $|z-b_k| > 2d_k$, and let

$$f_1(z) = \sum \mathcal{Q}_k(z)$$

This f_1 has the correct principal parts and the modifying terms $\frac{1}{z-b_k}$ have singularities on the boundary but remain holomorphic on $\mathcal{U}$. It suffices to show that on each compact subset of $\mathcal{U}$, $\sum \mathcal{Q}_k$ converges uniformly after we remove the first few terms.

So let $K \subset \mathcal{U}$ be compact. If for all but finitely many k we have $|z-b_k| > 2d_k$ for $z \in K$, then for all but finitely many k we have $|\mathcal{Q}_m| < \frac{1}{2^k}$ and the sum over these remaining terms converges uniformly on K. We are in trouble if for infinitely many k we can find $w_k \in K$ with $|w_k - b_k| \leq 2d_k$. From now on, restrict attention to these k.

In this bad case, one possibility is that $d_k \to 0$. This would imply that K is arbitrarily close to the boundary of $\mathcal{U}$, and this is impossible.

If the d_k to not converge to zero, we can restrict attention to an infinite subset bounded below by, say, B. So $0 < B \leq d_k$ and $B|z_k| \leq d_k|z_k| \leq 1$, and so $|z_k| \leq \frac{1}{B}$. It follows that these z_k have a convergent subsequence. The limit must not belong to $\mathcal{U}$, so $d_k \to 0$, a contradiction.

Remark: Mittag-Leffler was a wonderful Swedish mathematician with many accomplishments, including the above theorems which made him immortal. In the United States, a standard

mathematical rumor asserts that there is no Nobel prize in mathematics because Mittag-Leffler had an affair with Nobel's wife. Curiously, this story is not part of standard lore in Europe. One difficulty with the rumor is that Nobel never married.

Remark: The Mittag-Leffler theorem is the complex analysis version of *partial fraction reduction* as taught in integration theory.

14.2 Infinite Products

We now prepare for Weierstrauss' theorem, an extension of the high school task of factoring polynomials into linear terms.

Remark: We want to define infinite products $g(z) = \Pi g_k(z)$. We'd like various theorems, true for sums, to remain true for products:

- The convergence criterion should be related to uniform convergence on compact subsets.
- It should automatically follow that the product of holomorphic functions is again holomorphic
- It should be possible to rearrange convergent products without changing the result
- There should be a simple formula for the derivative of a product.
- We want to get all of this for free without doing any work!

In addition, we want

- The infinite product is zero if and only if one of the factors is zero.
- The order of $\Pi g_k(z)$ at z_0 should be the sum of the orders of the g_i at z_0.
- In a convergent product, $g_i(z) \to 1$ as $i \to \infty$.

Remark: The naive definition does not have these properties. If the c_k are complex and we define

$$\Pi c_k = \lim_{n\to\infty} \Pi_{k=1}^{n} c_i$$

then it is easy to find examples where the product is zero but none of the terms are zero, and examples where the product converges and yet the terms do not approach one.

Remark: Searching for the correct definition, we recall that logarithms convert products to sums. Perhaps the theory reduces to the known theory of sums using

$$\Pi e^{\mathrm{Log}(g_i)} = e^{\sum \mathrm{Log}(g_i)}$$

The problem is that we cannot take the logarithm of functions with zeros, and the logarithm is multi-valued. On the other hand, when g_i is close to one, the principal branch of the logarithm is well-defined and this formula makes sense.

Remark: This leads to the formal definition:

Definition 31. *If c_i are complex numbers such that $c_i \to 1$, there is an N such that for $i > N$ we have $\Re\, c_i > 0$. Then the standard branch of the logarithm $Log(c_i)$ is defined. If $\sum_{i>N} Log(c_i)$ converges absolutely, we say that the infinite product converges and define to be*

$$(c_1 \cdot c_2 \cdot ... \cdot c_N) \cdot e^{\sum_{i=N}^{\infty} Log(c_i)}$$

Lemma 9. *This product does not depend on the index when we switch to taking logarithms.*

Proof: Notice that $e^{\mathrm{Log}(c_i)} = c_i$. Any remaining words of proof are just fluff.

14.3 A Convergence Criterion for Products

The previous definition requires us to determine if $\sum |\mathrm{Log}(c_i)|$ converges. This can be difficult. Since c_i is close to 1, write $c_i = 1 + d_i$. The crucial lemma of the theory of infinite products asserts that our sum converges absolutely if and only if $\sum |d_i|$ converges absolutely. Testing this condition is *much* easier.

Begin with the following series, which converges if $|z| < 1$:

$$\mathrm{Log}(1+z) = z - \frac{z^2}{2} + \frac{z^3}{3} - ...$$

Therefore if $|z| < \frac{1}{2}$, we have

$$|\mathrm{Log}(1+z)| \leq |z| + |z|^2 + |z|^3 + ... = |z|\frac{1}{1-|z|} < 2|z|$$

and

$$|\mathrm{Log}(1+z)| \geq |z|\left(1 - \frac{|z|}{2} - \frac{|z|^2}{2} - \frac{|z|^3}{2} - ...\right) \leq |z|\left(1 - \frac{1}{4} - \frac{1}{8} - \frac{1}{16} - ...\right) = \frac{1}{2}|z|$$

Notice that if $c_i \to 1$ and $c_i = 1 + d_i$, then eventually $|d_i| < \frac{1}{2}$ and after that, $\sum |\mathrm{Log}(c_i)|$ converges if and only if $\sum |d_i|$ converges.

14.4 Infinite Products of Holomorphic Functions

Definition 32. *Let $f_i(z)$ be holomorphic on a domain $\mathcal{U}$. Write $f_i(z) = 1 + g_i(z)$. We say the infinite product of the f_i converges on $\mathcal{U}$ if for each compact $K \subset \mathcal{U}$, $\sum |g_i|$ converges uniformly on K. In that case, the infinite product is given by*

$$\prod f_i = (f_1(z) \cdot f_2(z) \ldots \cdot f_N(z))\, e^{\sum_{i>N} Log(f_i(z))}$$

In this formula, $i > N$ implies $|f_i(z) - 1| < \frac{1}{2}$.

Remark: The sum of logarithms converges uniformly on compact K, so this function is holomorphic on $\mathcal{U}$. Because the sum converges absolutely, the product can be rearranged at will. It is clear that the order of the product at z_0 is the (finite) sum of orders of factors at z_0.

Remark: Finally, we discuss discuss the derivative of such a product.

Definition 33. *If $f(z)$ is holomorphic, the* logarithmic derivative *of f is*

$$\frac{f'(z)}{f(z)}$$

Remark: Notice that this is the derivative of the logarithm of f, provided this logarithm is defined. Notice also that the log of a product is the sum of the logs, and thus the logarithmic derivative of a product is the sum of the logarithmic derivatives of the terms.

Theorem 80. *If each $f_i(z)$ is holomorphic on $\mathcal{U}$, and $\prod f_(z)$ exists, we have*

$$\frac{\frac{d}{dz}\prod f_i}{\prod f_i} = \sum \frac{f_i'}{f_i}$$

Proof: This is trivially true for finite sums. It also holds for the term

$$e^{\sum_{i>N} \mathrm{Log}(f_i(z))}$$

since the exponent can be differentiated term by term by a previous result.

14.5 The Theorem of Weierstrauss

Weierstrass' theorem does for zeros what Mittag Leffler's theorem does for poles:

Theorem 81 (Weierstrass)**.** *Let $z_1, z_2, \ldots$ be a sequence in the complex plane with no limit points. Let $k_1, k_2, \ldots$ be a related sequence of positive integers. Then there is an entire function $f(z)$ with zeros precisely at the z_i, each of order k_i.*

Remark: The proof is similar to the proof of the Mittag-Leffler theorem. The functions $f_i(z) = (z - z_i)^{k_i}$ have zeros of the required orders at z_i. We modify these functions to obtain convergence, and then take their infinite product.

An infinite product converges if the terms are closer and closer to 1, so the modified functions we seek should have a zero at z_i and yet be very close to 1. Concentrate on the case when $z_i = 1$. We propose the modified function

$$E(z) = (1 - z)e^{-\text{Log}(1-z)}$$

because it has a zero at 1 and yet is identically one. This logarithm does not exist at 1, so we form an approximation.

Since $\frac{1}{1-z} = 1 + z + z^2 + ...$, integration gives

$$-\text{Log}(1 - z) = z + \frac{z^2}{2} + \frac{z^3}{3} + ...$$

Replacing the logarithm with a finite approximation, we are led to consider

Definition 34.

$$E_n(z) = (1 - z)e^{\left(z + \frac{z^2}{2} + ... + \frac{z^n}{n}\right)}$$

Lemma 10. *If* $|z| \leq 1$, *then* $|1 - E_n(z)| \leq |z|^{n+1}$

Proof: The derivative of $E_n(z)$ is

$$(1 - z)e^{\left(z + \frac{z^2}{2} + ... + \frac{z^n}{n}\right)}\left(1 + z + ... + z^{n-1}\right) - e^{\left(z + \frac{z^2}{2} + ... + \frac{z^n}{n}\right)} = -z^n e^{\left(z + \frac{z^2}{2} + ... + \frac{z^n}{n}\right)}$$

It is clear that the exponential term expands in a power series whose coefficients are non-negative real numbers. So

$$\frac{dE_n}{dz} = -z^n - a_1 z^{n+1} - a_2 z^{n+2} - ...$$

and

$$E_n(z) = 1 - \frac{z^{n+1}}{n+1} - a_1 \frac{z^{n+2}}{n+2} - a_2 \frac{z^{n+3}}{n+3} - ...$$

Thus

$$\frac{1 - E_n(z)}{z^{n+1}} = b_0 + b_1 z + b_2 z^2 + ... \quad \text{for} \quad b_i \geq 0.$$

If $|z| \leq 1$,

$$\left|\frac{1 - E_n(z)}{z^{n+1}}\right| \leq \sum b_i |z|^i \leq \sum b_i = \frac{1 - E_n(1)}{1} = 1$$

and the inequality follows.

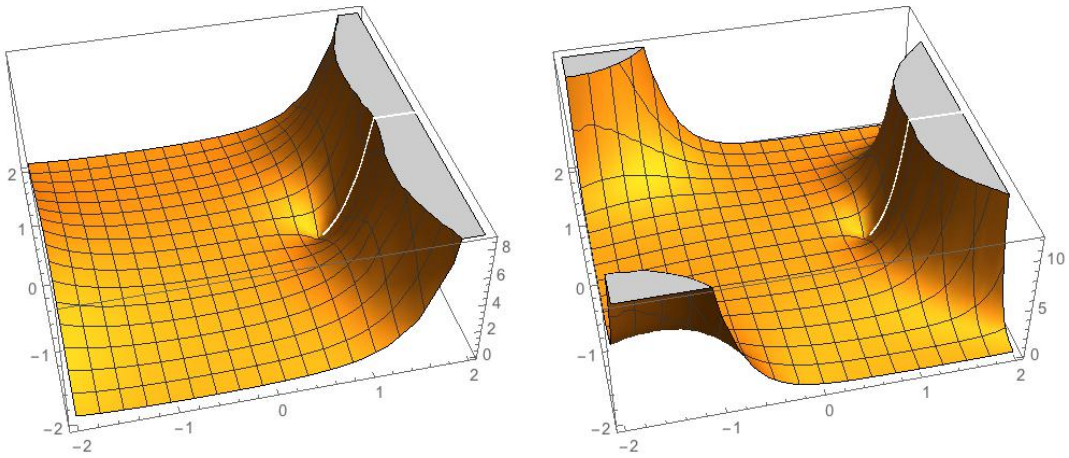


Figure 14.2: $|E_2(z)|$ and $|E_3(z)|$

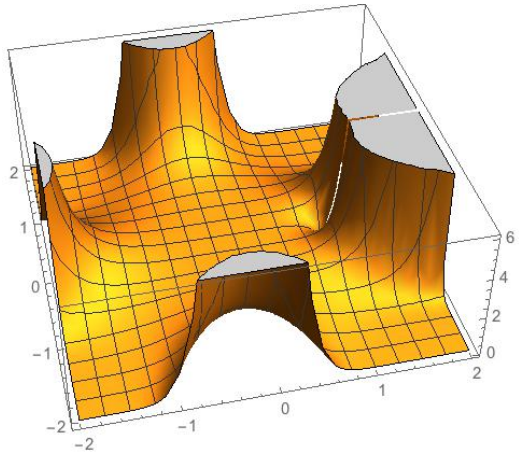


Figure 14.3: $|E_4(z)|$

Proof of Weierstrass Theorem Renumber the given points to $w_1, w_2, w_3, ...$ so each z_i is repeated k_i times, and $0 < |w_1| \leq |w_2| \leq |w_3| \leq ....$ (If $z = 0$ is supposed to be a zero of multiplicity k, replace f by $z^k f(z)$ at the ed of the argument.)

Define

$$f(z) = \prod E_i\left(\frac{z}{w_i}\right)$$

Incidentally, this is a conservative choice making convergence easy to prove; in many situations each term could be E_2. By the previous section, this converges provided

$$\sum \left|1 - E_i\left(\frac{z}{w_i}\right)\right|$$

converges uniformly on compact subsets K of the plane. Since the w_i have no limit points, eventually $\left|\frac{z}{w_i}\right| \leq \frac{1}{2}$ on K. Applying the inequality at the start of this section, eventually

$$\sum \left|1 - E_i\left(\frac{z}{w_i}\right)\right| \leq \sum \left|\frac{z}{w_i}\right|^{i+1} \leq \sum \left|\frac{1}{2}\right|^{i+1}$$

and we have the desired convergence. QED.

Theorem 82 (Weierstrass)**.** *Let $\mathcal{U}$ be a domain in the plane. Let $z_1, z_2, \ldots$ be a sequence in $\mathcal{U}$ with no limit points in $\mathcal{U}$. Let $k_1, k_2, \ldots$ be a related sequence of positive integers. Then there is a holomorphic function $f(z)$ on $\mathcal{U}$ with zeros precisely at the z_i, each of order k_i.*

Proof: The proof follows the outline of the similar Mittag-Leffler proof. For each z_i we find a closest boundary point b_i and distance $d_i = |z_i - b_i|$. Separate indices into $\mathcal{A} = \{\, k \mid b_i|z_i| \leq 1 \,\}$ and $\mathcal{B} = \{\, k \mid b_i|z_i| > 1 \,\}$. Find functions f_1 and f_2 with zeros in $\mathcal{A}$ and $\mathcal{B}$, respectively, and define $f(z) = f_1(z) + f_2(z)$.

The z_k with indices in $\mathcal{B}$ have no finite limit points, so the previous argument gives $f_2(z)$.

Rename the indices in $\mathcal{A}$ so each occurs k_i times. We want to handle these indices as in the Mittag-Leffler proof, perhaps using the same proof pictured below.

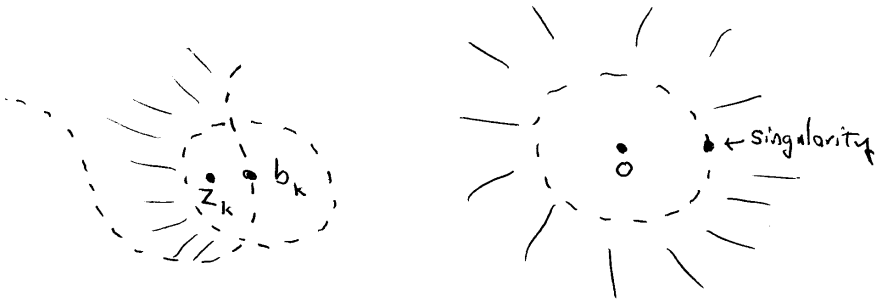


Figure 14.4: Modifying $E_n(z)$ to Force Convergence

The previous proof used expansion of Laurent series in an annulus. Our job is to find an analogue for $E_n(z)$. We have accumulated many tools, and one that comes to mind is linear fractional transformations, which move points on the sphere. Perhaps we could employ

$$E_n\left(\frac{Az + B}{Cz + D}\right)$$

A linear fractional transformation is determined by the images of three points. If our choice mapped b_i to infinity, then the modified E_n would have a singularity at b_i, which is not in $\mathcal{U}$. If

our choice mapped z_i to 1, then the modified E_n would have a zero at z_i. Finally, if our choice mapped ∞ to 0, then points far away, i.e., not close to b_i, would be mapped to near zero where E_n is almost one.

Thus we are led to form

$$E_n\left(\frac{z_i - b_i}{z - b_i}\right)$$

Below, for example, is this function where we require a zero at 2.0, and the closest boundary point is 3.0. This function belows up in a complicated way at 3, which is not in $\mathcal{U}$, but has our required zero at 2. Moreover is close to 1 away from the region containing these two points.

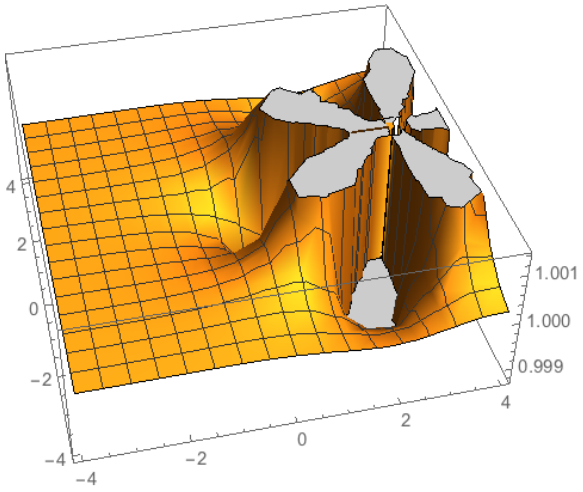


Figure 14.5: $E_4\left(\frac{z_i - b_i}{z - b_i}\right)$; $z_i = 2; b_i = 3$

According to the crucial lemma about E_n, if $\left|\frac{z_i - b_i}{z - b_i}\right| \leq 1$, then

$$\left|E_n\left(\frac{z_i - b_i}{z - b_i}\right) - 1\right| \leq \left|\frac{z_i - b_i}{z - b_i}\right|^{n+1}$$

Apply this when $\left|\frac{z_i - b_i}{z - b_i}\right| \leq \frac{1}{2}$, or equivalently $|z - b_i| \geq 2|z_i - b_i|$, then

$$\left|E_n\left(\frac{z_i - b_i}{z - b_i}\right) - 1\right| \leq \left(\frac{1}{2}\right)^{n+1}$$

So define

$$f_1(z) = \prod E_i\left(\frac{z_i - b_i}{z - b_i}\right)$$

As in the proof of the general Mittag-Leffler theorem, if K is compact in $\mathcal{U}$ and indices belong to $\mathcal{A}$, then for all but finitely many i we have $|z - b_i| > 2d_i$ and so

$$\sum \left| E_n\left(\frac{z_i - b_i}{z - b_i}\right) - 1 \right|$$

converges uniformly.

Corollary 23. *If $\mathcal{U}$ is a domain, there is a holomorphic function $f(z)$ on $\mathcal{U}$ which cannot be extended to be holomorphic on any larger connected open set.*

Proof: Find a sequence $z_n \in \mathcal{U}$ without limit points in $\mathcal{U}$ but such that every boundary point of $\mathcal{U}$ is a limit point. Let $f(z)$ have first order orders at each z_n and no other zeros. Any large open set would contain limit points of the zeros of f, so f would be identically zero on the larger set.

Remark: There is a difficult theory of functions of several complex variables, invented mostly in the 20th century. One reason the theory is difficult is that the previous result is no longer true. Instead, open sets $\mathcal{U} \subset \mathcal{V}$ exist with the property that all functions holomorphic in $\mathcal{U}$ can be extended to be holomorphic in $\mathcal{V}$.

Chapter 15

Classical Sum and Product Formulas

15.1 Beautiful Identities

We get beautiful formulas when we apply the techniques of the Mittag-Leffler and Weierstrass theorems to poles or zeros with a symmetrical pattern, and try to adjust the Principal Part or $E(z)$ as little as possible for convergence. Since the trigonometric functions have periodic zeros, our formulas often sum to trigonometric expressions.

Theorem 83.

$$\frac{\pi}{\sin \pi z} = \frac{1}{z} + \sum_{n \neq 0} (-1)^n \left(\frac{1}{z-n} + \frac{1}{n} \right)$$

$$\frac{\pi^2 \cos \pi z}{\sin^2 \pi z} = \frac{1}{z^2} + \sum_{n \neq 0} (-1)^n \frac{1}{(z-n)^2} = \sum_{n} (-1)^n \frac{1}{(z-n)^2}$$

$$\frac{\pi \cos \pi z}{\sin \pi z} = \frac{1}{z} + \sum_{n \neq 0} \left(\frac{1}{(z-n} + \frac{1}{n} \right)$$

$$\frac{\pi^2}{\sin^2 \pi z} = \frac{1}{z^2} + \sum_{n \neq 0} \frac{1}{(z-n)^2} = \sum_{n} \frac{1}{(z-n)^2}$$

Remark: In the first formulas, the principal parts are $\frac{1}{z-n}$ and $\frac{1}{(z-n)^2}$ up to signs, which alternate. A brief check shows that the second formula is obtained by differentiating the first. The

principal parts of the third and fourth formulas are always $\frac{1}{z-n}$ and $\frac{1}{(z-n)^2}$ without sign alternation, and again the final formula is the derivative of the third one. By differentiating again and again, we obtain similar formulas with principal parts $\frac{1}{(z-n)^k}$, with or without sign alternations. After multiplying by constants and adding, we can get any principal part, repeated with or without sign alternations.

Next we check that the right hand sides converge. It suffices to study the first and third formulas because convergence of the other sums follows by differentiation. In the first and third formulas, we could not simply write $\sum \frac{1}{z-n}$ because the resulting sums would not converge. For instance, if we set $z = 0$ and ignore the singular term $\frac{1}{z}$, we get $\sum \frac{1}{n}$ which diverges. Consequently, we modified the terms using Mittag-Leffler's procedure. If we expand $\frac{1}{z-n}$ about the origin, we get

$$\frac{1}{z-n} = -\left(\frac{1}{n}\right)\frac{1}{1-\frac{z}{n}} = -\frac{1}{n}\sum\left(\frac{z}{n}\right)^k = -\sum_k \frac{z^k}{n^{k+1}}$$

According to Mittag-Leffler, we get convergence by subtracting a finite number of these terms, and our formula suggests that it is enough to subtract $-\frac{1}{n}$. This works because

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

If K is a compact subset of the plane and $z \in K$, then the numerator of this fraction is bounded by, say, B, and for n large enough, we have $|(z-n)n| \geq \left(\frac{n}{2} \cdot n\right)$. So the series converges because $\sum \frac{1}{n^2} < \infty$ and

$$\left|\frac{z}{(z-n)n}\right| \leq \frac{4B}{n^2}$$

Notice that no Mittag-Leffler modification is required for series with higher powers of $\frac{1}{z-n}$.

Next we claim that the left and right sides of the first two equations satisfy $g(z+1) = -g(z)$ and consequently are periodic of period 2; similarly we claim that the left and right sides of the third and fourth equations are periodic of period 1. This is completely trivial for the left sides by ordinary trigonometry. It is also trivial for the right sides of the second and fourth formula because all integers are treated the same up to sign in these formulas. A slight argument is required for the right sides in the first and third formulas, and we'll give it for the first formula and sketch it for the third.

Let $g(z)$ be the sum of the right side of the first formula. Since the second formula is the derivative of the first, and the second formula is periodic up to sign, we have $\frac{d}{dz}g(z+1) = -\frac{d}{dz}g(z)$ and so $\frac{d}{dz}\left(g(z+1)+g(z)\right) = 0$. Consequently, $g(z+1)+g(z)$ is constant. This

constant is $g(1/2)+g(-1/2)$. However, $g(z)$ is an odd function, so the sum of these two terms is zero. Indeed,

$$g(z)=\frac{1}{z}+\sum_{n>0}(-1)^n\left(\frac{1}{z-n}+\frac{1}{n}+\frac{1}{z+n}-\frac{1}{n}\right)=\frac{1}{z}+\sum_{n>0}\frac{2z}{(z^2-n^2)}$$

The argument for the third formula is similar; we want to prove that $g(z+1)=g(z)$, but their derivatives are equal, so $g(z+1)-g(z)$ is constant. In particular, set $z=\frac{1}{2}$. Then $g\left(\frac{1}{2}\right)-g\left(-\frac{1}{2}\right)$ is this constant. As before, g is odd, so the constant is $2g\left(\frac{1}{2}\right)$. However,

$$g(z)=\frac{1}{z}+\sum_{n>0}\left(\frac{1}{z-n}+\frac{1}{z+n}\right)$$

so

$$g\left(\frac{1}{2}\right)=2+\sum_{n>0}\left(\frac{2}{1-2n}+\frac{2}{1+2n}\right)=2\left[1+\sum_{n>0}\left(-\frac{1}{2n-1}+\frac{1}{2n+1}\right)\right]$$

and the value inside the square brackets is

$$1+\left(-1+\frac{1}{3}\right)+\left(-\frac{1}{3}+\frac{1}{5}\right)+\left(-\frac{1}{5}+\frac{1}{7}\right)+\ldots=0$$

Next we claim that the left sides of our four formulas have the same principal parts as the right sides, and consequently the difference between the two sides is an entire function. This is proved using the same sort of calculation in all four cases, so we only consider the first case.

In this first case, $g(z+1)=-g(z)$, so it suffices to check the principal parts at the origin. But $\sin z=z-\frac{z^3}{3!}+\frac{z^5}{5!}+\ldots$ and thus

$$\sin\pi z=\pi z-\frac{\pi^3z^3}{3!}+\ldots=\pi z\left(1-\frac{\pi^2z^2}{3!}+\ldots\right)$$

and so

$$\frac{\pi}{\sin\pi z}=\frac{1}{z}\left(1-\frac{\pi^2z^2}{3!}+\ldots\right)^{-1}=\frac{1}{z}+1+\ldots$$

The final step of the proof of the second and fourth identities proceeds as follows. Let $z=x+iy$ where $0\le x\le 1$ and $|y|\to\infty$. We prove that the left and right sides of our identities go to zero. It follows from this and periodicity that the difference between right and left sides is bounded, hence constant. But since the above limit is zero, this constant must be zero.

We carry this out for the second formula. Notice that

$$\cos \pi z = \frac{e^{i\pi z} + e^{-i\pi z}}{2} = \frac{e^{i\pi x}e^{-\pi y} + e^{-i\pi x}e^{\pi y}}{2}$$

$$\sin \pi z = \frac{e^{i\pi z} - e^{-i\pi z}}{2i} = \frac{e^{i\pi x}e^{-\pi y} - e^{-i\pi x}e^{\pi y}}{2i}$$

We consider the case $y \to \infty$, letting the reader handle the case $-y \to \infty$. Then all terms in these expressions go to zero except the last term, which is in absolute value $\frac{e^{\pi y}}{2}$ in both cases. But in our expression the $\cos \pi z$ term is in the numerator, and the $\sin \pi z$ term is *squared* in the denominator. So the quotient goes to zero.

This argument works for the first, second, and fourth formulas.

Next look at the right side. The second and fourth formulas are the same up to an irrelevant $(-1)^n$. So consider $\sum \frac{1}{(z-n)^2}$. If z is anywhere in the strip $0 \le x \le 1$, then $|z - n| \ge \frac{n}{2}$ for all but finitely many n. For the remaining terms, $\left|\frac{1}{(z-n)^2}\right| \le \frac{4}{n^2}$.

Suppose we want to make the right side less than ϵ by choosing $|y|$ large enough. Choose N so $\sum_{|n|>N} \frac{4}{n^2} < \frac{\epsilon}{2}$. Only finitely many terms remain to be considered, but their sum can also be made less than $\frac{\epsilon}{2}$ because $\lim_{|z|\to\infty} \frac{1}{(z-n)^2} = 0$.

This completes the proof for the second and fourth formulas. However, the argument just given does not work for the first and third formulas. (Try it; the individual terms go to zero, but getting an initial bound on infinitely many terms doesn't work.) So a trick is required in the first and third cases.

But the derivative of the difference between the left and right sides of the first formula is the difference between the left and right sides of the second formula, hence zero. It follows that the difference between the left and right sides of the first formula is constant. Similarly the difference between the left and right sides of the third formula is constant. To show that both of these constants are zero, it suffices to evaluate the first and third formulas at specific points. A good choice is $z = \frac{1}{2}$.

We do both calculations at once. Recall that the two formulas can be rewritten as

$$\frac{\pi}{\sin \pi z} = \frac{1}{z} + \sum_{n\neq 0} (-1)^n \left(\frac{1}{z-n} + \frac{1}{z+n}\right)$$

$$\frac{\pi \cos \pi z}{\sin \pi z} = \frac{1}{z} + \sum_{n\neq 0} \left(\frac{1}{z-n} + \frac{1}{z+n}\right)$$

Substituting $z = \frac{1}{2}$ gives

$$\pi = 2 + \sum_{n \neq 0} (-1)^n \left(\frac{2}{1-2n} + \frac{2}{1+2n} \right)$$

$$0 = 2 + \sum_{n \neq 0} \left(\frac{2}{1-2n} + \frac{2}{1+2n} \right)$$

or

$$\frac{\pi}{2} = 1 + \sum_{n \neq 0} (-1)^n \left(-\frac{1}{2n-1} + \frac{1}{2n+1} \right)$$

$$0 = 1 + \sum_{n \neq 0} \left(-\frac{1}{2n-1} + \frac{1}{2n+1} \right)$$

or

$$\frac{\pi}{2} = 1 - \left(-1 + \frac{1}{3} \right) + \left(-\frac{1}{3} + \frac{1}{5} \right) - \left(-\frac{1}{5} + \frac{1}{7} \right) + \ldots$$

$$0 = 1 + \left(-1 + \frac{1}{3} \right) + \left(-\frac{1}{3} + \frac{1}{5} \right) + \left(-\frac{1}{5} + \frac{1}{7} \right) + \ldots$$

or

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \ldots$$

$$0 = (1-1) + \left(\frac{1}{3} - \frac{1}{3} \right) + \left(\frac{1}{5} - \frac{1}{5} \right) + \left(\frac{1}{7} - \frac{1}{7} \right) + \ldots$$

Both of these formulas are true; the first was one of the first results deduced by Leibniz during the invention of the calculus. This finishes the argument for all four formulas.

Example Substituting $z = \frac{1}{2}$ in the fourth formula gives

$$\pi^2 = \sum \frac{4}{(2n-1)^2}$$

and so

$$\frac{\pi^2}{8} = 1 + \frac{1}{3^2} + \frac{1}{5^2} + \ldots$$

If we set

$$L = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \ldots$$

then

$$L - \frac{\pi^2}{8} = \frac{1}{2^2} + \frac{1}{4^2} + \frac{1}{6^2} + \ldots = \frac{1}{4} \left(1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \ldots \right) = \frac{L}{4}$$

and so

$$\frac{3L}{4} = \frac{\pi^2}{8}$$

and

$$L = \frac{\pi^2}{6} = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + ...$$

Remark: Below is a plot of the third formula. In this case, the limit of the absolute value of the function high above the x-axis is *not zero*; indeed it is π.

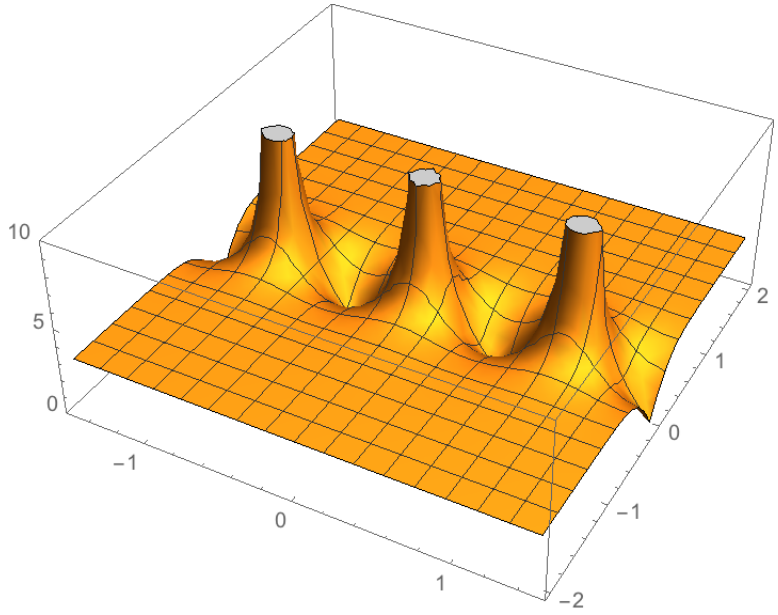


Figure 15.1: $\left|\pi \frac{\cos \pi z}{\sin \pi z}\right|$

15.2 A Product Formula

Theorem 84.

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

Proof: From the proof of Weierstrauss' theorem, recall that

$$E_n(z) = (1 - z) e^{\left(z + \frac{z^2}{2} + ... + \frac{z^n}{n}\right)}$$

If $|z| \leq 1$, then $|1 - E_n(z)| \leq |z|^{n+1}$. If $f(z) = \prod E_{n_i}\left(\frac{z}{w_i}\right)$, then this product converges to a function with zeros w_i provided

$$\sum \left|1 - E_{n_i}\left(\frac{z}{w_i}\right)\right|$$

converges.

If we want a function with simple zeros at the non-zero integers, $f(z) = \prod E_{n_i}\left(\frac{z}{n}\right)$ and we require that $\sum |1 - E_{n_i}\left(\frac{z}{n}\right)|$ converge. As soon as $|\frac{z}{n}| \leq 1$, this becomes the condition $\sum |\left(\frac{z}{n}\right)^{n_i+1}|$ converge, and since $\sum \frac{1}{n^2}$ converges, it suffices to choose $n_i = 1$. Thus the following product, which is essentially the right side of our equation, converges:

$$\prod E_1\left(\frac{z}{n}\right) = \prod \left(1 - \frac{z}{n}\right) e^{\frac{z}{n}}$$

Let us compute the logarithmic derivative of both sides. On the left we get

$$\frac{\pi \cos \pi z}{\sin \pi z}$$

and on the right we get

$$\frac{1}{z} + \sum_{n>0} \frac{\frac{-2z}{n^2}}{(1 - \frac{z^2}{n^2})} = \frac{1}{z} + \sum_{n>0} \frac{-2z}{n^2 - z^2} = \frac{1}{z} + \sum_{n>0} \frac{2z}{z^2 - n^2}$$

These two logarithmic derivatives are equal by the third of our additive formulas. But

$$\frac{f'}{f} = \frac{g'}{g}$$

implies

$$\frac{f'g - fg'}{fg} = 0$$

which implies that $\frac{f}{g}$ has derivative zero, and thus that $\frac{f}{g}$ is a constant. So there is a λ such that

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

Dividing both sides by πz gives

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

The limit of the left side as $z \to 0$ is 1, and the value of the right side at $z = 0$ is λ, so $\lambda = 1$. QED.

Example 1: Set $z = \frac{1}{2}$ to obtain

$$1 = \frac{\pi}{2} \prod_{n>0} \left(1 - \frac{1}{4n^2}\right) = \frac{\pi}{2} \prod \frac{(2n-1)(2n+1)}{(2n)(2n)}$$

This is Wallis' Product:

$$\frac{2}{\pi} = \frac{1}{2} \cdot \frac{3}{2} \cdot \frac{3}{4} \cdot \frac{5}{4} \cdot ...$$

Example 2: We have

$$\frac{\sin z}{z} = 1 - \frac{z^2}{3!} + \frac{z^4}{5!} - \frac{z^6}{7!} + ...$$

and so

$$\frac{\sin \pi z}{\pi z} = 1 - \frac{\pi^2}{3!} z^2 + \frac{\pi^4}{5!} z^4 - \frac{\pi^6}{7!} z^6 + ...$$

On the other hand,

$$\prod \left(1 - \frac{z^2}{n^2}\right) = 1 - \left(\sum \frac{1}{n^2}\right) z^2 + \left(\sum_{m_1<m_2} \frac{1}{m_1^2 m_2^2}\right) z^4 - \left(\sum_{m_1<m_2<m_3} \frac{1}{m_1^2 m_2^2 m_3^2}\right) z^6 + ...$$

Therefore

$$\frac{\pi^2}{6} = \sum \frac{1}{n^2}$$

Moreover,

$$\left(\sum \frac{1}{m_1^2}\right)\left(\sum \frac{1}{m_2^2}\right) = \sum \frac{1}{n^4} + 2 \sum_{m_1<m_2} \frac{1}{m_1^2 m_2^2}$$

or

$$\left(\frac{\pi^2}{6}\right)\left(\frac{\pi^2}{6}\right) = \sum \frac{1}{n^4} + 2\frac{\pi^4}{5!}$$

and so

$$\frac{\pi^4}{90} = \sum \frac{1}{n^4}$$

Finally

$$\left(\sum \frac{1}{m_1^2}\right)\left(\sum \frac{1}{m_2^2}\right)\left(\sum \frac{1}{m_3^2}\right) = \sum \frac{1}{n^6} + 3\left(\sum_{m_1 \neq m_2} \frac{1}{m_1^4 m_2^2}\right) + 3! \sum_{m_1<m_2<m_3} \frac{1}{m_1^2 m_2^2 m_3^2}$$

Moreover,

$$\left(\sum_{m_1 \neq m_2} \frac{1}{m_1^4 m_2}\right) = \left(\sum \frac{1}{n^4}\right)\left(\sum \frac{1}{m^2}\right) - \left(\sum \frac{1}{n^6}\right)$$

Consequently

$$\frac{\pi^6}{6^3} = -2\sum \frac{1}{n^6} + 3\frac{\pi^6}{90 \cdot 6} + 6\frac{\pi^6}{7!}$$

and so

$$\frac{\pi^6}{945} = \sum \frac{1}{n^6}$$

Chapter 16

The Gamma Function

16.1 Three Theorems

Several variable calculus is about functions $f(x,y)$. The central tools of the subject are integration with respect to x or y and differentiation with respect to x or y. There are three main theorems in the elementary theory and they are closely related. The first says that integration with respect to x and then y can be done in either order. The second says that differentiation with respect to x and then y can be done in either order. The third says that integration with respect to x and then differentiation with respect to y can be done in either order.

Theorem 85. *Let $f(x,y)$ be continuous on $[a,b]\times[c,d]$. Then*

$$\int_a^b \int_c^d f(x,y)\,dydx = \int_c^d \int_a^b f(x,y)\,dxdy$$

Proof: Well-known. We use a slight generalization to prove the third main theorem.

Lemma 11 (Fundamental Theorem of Line Integrals)**.** *Let $f(x,y)$ be a continuously differentiable function, and let γ be a path. Then*

$$\int_\gamma grad\ f \cdot d\gamma = f(end) - f(beginning)$$

Lemma 12 (Green's Theorem)**.** *Let $E = \Big(E_x(x,y), F_x(x,y)\Big)$ be a continuously differentiable vector field, and let $\mathcal{R} = [a,b]\times[c,d]$ be a rectangle with counterclockwise boundary γ. Then*

$$\int\int_{\mathcal{R}} \left(\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y}\right) dxdy = \int_\gamma E \cdot d\gamma$$

Proof: Integrate the first term in the order

$$\int_c^d \int_a^b \frac{\partial E_y}{\partial x}\, dx\, dy$$

The inner integral can be done using the fundamental theorem of calculus, and the remaining integral gives the line integral along the left and right sides.

Integrate the second term in the reverse order. Again the inner integral yields to the fundamental theorem of calculus, and the remaining integral gives the line integral along the top and bottom. QED.

Theorem 86. *Let $f(x, y)$ have continuous first and second partials on an open $\mathcal{U}$. Then*

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}$$

Proof: Apply Green's theorem to grad f over a rectangle inside $\mathcal{U}$. Since the boundary is a closed curve, the fundamental theorem of line integrals states that the result is zero. So

$$\int \int_{\mathcal{R}} \left(\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} \right) dxdy = \int \int_{\mathcal{R}} \left(\frac{\partial^2 f}{\partial x \partial y} - \frac{\partial^2 f}{\partial y \partial x} \right) dxdy = 0$$

But if these two expressions are not equal, then their difference are strictly positive or strictly negative and bounded away from zero on *some* small rectangle by continuity, and on that rectangle the integral would be nonzero. QED.

Theorem 87. *Let $f(x, y)$ be continuous on an open set containing $[a, b] \times [c, d]$, and suppose that $\frac{\partial f}{\partial y}(x, y)$ exists and is continuous on this open set. Then on the rectangle*

$$\frac{\partial}{\partial y} \int_a^b f(x, y)\, dx = \int_a^b \frac{\partial f}{\partial y}\, dx$$

Proof: Let $F(y) = \int_a^b f(x, y)\, dx$. Form

$$\int_a^b \int_c^y \frac{\partial f}{\partial y}(x, t)\, dt\, dx = \int_a^b \big(f(x, y) - f(x, c) \big)\, dx = F(y) - F(c)$$

Invert the order of integration to get

$$\int_c^y \int_a^b \frac{\partial f}{\partial y}(x, t)\, dx\, dt = F(y) - F(c)$$

Differentiate both sides with respect to y, and use the fact that for any G we have $\frac{d}{dy}\int_c^y G(t)dt = G(y)$ to get

$$\int_a^b \frac{\partial f}{\partial y}(x,y)\,dx = \frac{\partial F}{\partial y} = \frac{\partial}{\partial y}\int_a^b f(x,y)\,dx$$

16.2 Feynman and Differentiating Under the Integral Sign

In his book *Surely You're Joking, Mr. Feynman*, Feynman writes

> One thing I never did learn was contour integration. I had learned to do integrals by various methods shown in a book that my high school physics teacher Mr. Bader had given me. One day he told me to stay after class. "Feynman," he said, "you talk too much and you make too much noise. I know why. You're bored. So I'm going to give you a book. You go up there in the back, in the corner, and study this book, and when you know everything that's in this book, you can talk again." So every physics class, I paid no attention to what was going on with Pascal's Law, or whatever they were doing. I was up in the back with this book: "Advanced Calculus", by Woods. It had Fourier series, Bessel functions, determinants, elliptic functions, all kinds of wonderful stuff that I didn't know anything about. That book also showed how to differentiate parameters under the integral sign — it's a certain operation. It turns out that its not taught very much in the universities; they don't emphasize it. But I caught on how to use that method, and I used that one damn tool again and again. So because I was self-taught using that book, I had peculiar methods of doing integrals. The result was, when guys at MIT or Princeton had trouble doing a certain integral, it was because they couldn't do it with the standard methods they had learned in school. If it was contour integration, they would have found it; if it was a simple series expansion, they would have found it. Then I come along and try differentiating under the integral sign, and often it worked. So I got a great reputation for doing integrals, only because my box of tools was different from everybody else's, and they had tried all their tools on it before giving the problem to me.

The Wikipedia article on "Differentiation under the integral sign" has many wonderful examples of integrals computed using this method.

In complex analysis, this result gives an extremely useful technique to construct holomorphic functions:

Theorem 88. *Let $f(t,z)$ be continuous on an open set containing $[a,b]\times\mathcal{U}$. Suppose that each*

t, $\frac{df}{dz}(t,z)$ exists, and suppose this function is continuous on $[a,b]\times\mathcal{U}$. Then

$$g(z)=\int_a^b f(t,z)\,dt$$

is holomorphic on $\mathcal{U}$.

Proof: Our previous theorem on differentiating under the integral sign obvious applies to complex-valued functions and to functions with more variables. Since $f(t,z)$ satisfies the Cauchy Riemann equations for each fixed t, the function $g(z)$ satisfies these equations. QED.

Remark: In actual practice, $f(t,z)$ is often also holomorphic in t, and we replace the integral from a to b by a path integral. Once we parameterize the path, we have an expression of the sort covered by the previous theorem. Since the integral of a sum is the sum of the integrals, any piece-wise differentiable path will suffice.

16.3 The Gamma Function

Definition 35. *In the formula below, $t^{z-1}=e^{(z-1)\ln t}$ where $\ln t$ is the ordinary real logarithm. If the real part of z is greater than zero, define*

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

Remark: This is known as the Euler's Gamma Function. As we see in a moment, it extends the factorial function on positive integers to arbitrary complex numbers with positive real part.

We claim that $\Gamma(z)$ is holomorphic. This follows from the previous section, except that we now have an indefinite integral. While we could invent an elaborate theory to handle such integrals, it is easier to be informal. Notice that $|t^{(z-1)}|=t^{x-1}$, so if we restrict z to a compact set, then this term is bounded by t^N for some integer N. But then we can replace e^t by one of the terms in its Taylor expansion to get

$$|t^{z-1}e^{-t}|\le\frac{t^N}{\frac{t^{N+2}}{(N+2)!}}=\frac{(N+2)!}{t^2}$$

This expression is independent of z as long as z is restricted to a compact set, and the integral to infinity of $\frac{1}{t^2}$ converges. It follows that the integral converges as the upper limit goes to infinity, and converges uniformly on compact subsets of the complex plane. So the limit of these holomorphic functions is holomorphic.

We also have to handle a possible singularity at the origin. The term e^{-t} is bounded near the origin and plays no role. The term t^{x-1} is bounded if $x > 1$, so we have convergence if the real part of z is greater than 1. But we can do better. If z is restricted to a compact set in the right half plane, then the real part of $z - 1$ is bounded away from -1, so we can assume that t^{x-1} is bounded by t^a where $-1 < a <= 0$. Notice that $\int_0^1 t^a dt = \left.\frac{t^{a+1}}{a+1}\right|_0^1$ is then finite because t^{a+1} is a positive power of t and thus goes to zero as t goes to zero. Again we obtain uniform convergence on compact subsets as the lower limit goes to zero.

Theorem 89. *We have* $\Gamma(1) = 1$. *If the real part of z is greater than zero, then*

$$\Gamma(z+1) = z\Gamma(z)$$

Corollary 24. *If n is positive integer,* $\Gamma(n) = (n-1)!$

Proof: The first part is a short calculation. The rest follows from integration by parts.

$$\Gamma(z+1) = \int_0^\infty t^z e^{-t}\, dt = \int_0^\infty t^z \frac{d}{dt}\left(-e^{-t}\right)\, dt = -t^z e^{-t}\Big|_0^\infty + z\int_0^\infty t^{z-1}e^{-t}\, dt$$

Remark: It is unfortunate that $\Gamma(n) = (n-1)!$ rather than $n!$, but the notation has been fixed for over 200 years and cannot be changed. However, we are allowed by write $z! = \Gamma(z+1)$ where the real part of z is greater than -1.

Remark: We now provide preliminary evidence that this extension isn't arbitrary. Namely, there is a short formula for the volume of a ball in R^n provided we can take factorials of half-integers.

Theorem 90.

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

Remark: Hence

$$\left(-\frac{1}{2}\right)! = \sqrt{\pi}$$

$$\left(\frac{1}{2}\right)! = \frac{1}{2}\sqrt{\pi}$$

$$\left(\frac{3}{2}\right)! = \frac{3}{2}\cdot\frac{1}{2}\sqrt{\pi}$$

Proof: Notice that $\Gamma\left(\frac{1}{2}\right) = \int_0^\infty t^{-\frac{1}{2}} e^{-t}\, dt$. Substitute $t = u^2$ to convert this to $\int_0^\infty \frac{1}{u} e^{-u^2} 2u\, du$. This equals $2\int_0^\infty e^{-u^2}\, du = \int_{-\infty}^\infty e^{-u^2}\, du$.

Let $A = \int_{-\infty}^{\infty} e^{-u^2}\, du$. Then

$$A^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-u^2-v^2}\, du\, dv$$

Convert to polar coordinates

$$A^2 = \int_0^{2\pi}\int_0^{\infty} e^{-r^2} r\, dr\, d\theta = \pi$$

Hence $A = \sqrt{\pi}$ and $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

Theorem 91. *The area of the sphere $S^{n-1} \subset R^n$, is*

$$\frac{2\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)} = \frac{n\,\pi^{\frac{n}{2}}}{\left(\frac{n}{2}\right)!}$$

Proof: Integrate $e^{-(x_1^2+x_2^2+\ldots+x_n^2)}$ over R^n. In rectangular coordinates this is

$$\int_{\infty}^{\infty}\ldots\int_{\infty}^{\infty} e^{-(x_1^2+x_2^2+\ldots+x_n^2)}\, dx_1 \ldots dx_n = \left(\int_{-\infty}^{\infty} e^{-x^2}\, dx\right)^n = \left(\sqrt{\pi}\right)^n$$

In generalized polar coordinates it is

$$\int_0^{\infty}\int_{S^{n-1}} e^{-\rho^n}\rho^{n-1}\, d\rho\, d\sigma$$

where ρ is the radial direction and σ is integration over the sphere. This expression equals

$$(\text{Area of Sphere})\int_0^{\infty} \rho^{n-1}e^{-\rho^2}\, d\rho$$

Substitute $t = \rho^2$ to obtain

$$(\text{Area of Sphere})\int_0^{\infty} t^{\frac{n-1}{2}}e^{-t}\frac{dt}{2\sqrt{t}}$$

Therefore

$$\left(\sqrt{\pi}\right)^n = \frac{1}{2}(\text{Area of Sphere})\int_0^{\infty} t^{\frac{n}{2}-1}e^{-t}\, dt = \frac{1}{2}(\text{Area of Sphere})\,\Gamma\left(\frac{n}{2}\right)$$

and the result follows.

Theorem 92. *The volume of the unit ball in R^n is*

$$\frac{\pi^{\frac{n}{2}}}{\left(\frac{n}{2}\right)!}$$

Proof: This volume equals $\int_0^1 \rho^{n-1}\, d\rho\, d\sigma = \frac{1}{n}(\text{Area of Sphere}) = \frac{\pi^{\frac{n}{2}}}{\left(\frac{n}{2}\right)!}$

Remark: Although the integral defining $\Gamma(z)$ only converges when the real part of z is positive, the function itself can be extended to the entire complex plane:

Theorem 93. *There is a continuation of $\Gamma(z)$ to a meromorphic function on the entire complex plane. This function has first order poles are $0, -1, -2, -3, \ldots$ and otherwise is holomorphic.*

Proof: The formula $\Gamma(z+1) = z\Gamma(z)$ allows us to write

$$\Gamma(z) = \frac{\Gamma(z+1)}{z}$$

The expression on the right is defined and holomorphic when z with real part greater than -1, except for a first order pole at the origin. Extend further left using

$$\Gamma(z) = \frac{\Gamma(z+1)}{z} = \frac{\Gamma(z+2)}{z(z+1)} = \frac{\Gamma(z+3)}{z(z+1)(z+2)} = \ldots$$

Remark: It is convenient to develop the theory over again starting with an infinite product, because the deeper properties of the gamma function follow easily from this product. In the end, we will want to show that both definitions give the same result. This will follow from the theorem below, which shows that our integral is related to a certain product. For simplicity, we assume z is real, because the identity theorem shows that two holomorphic functions which agree on a non-trivial real interval are equal.

Theorem 94. *Suppose $x > 0$ is real. Then*

$$\Gamma(x) = \lim_{n\to\infty} n^x \frac{n!}{x(x+1)(x+2)\ldots(x+n)}$$

Proof: Recall that $e^{-t} = \lim_{n\to\infty}\left(1 - \frac{t}{n}\right)^n$. Inserting this into the integral, we obtain the following formula; the reader can supply details needed for rigor:

$$\Gamma(x) = \lim_{n\to\infty} \int_0^n t^{x-1}\left(1 - \frac{t}{n}\right)^n dt$$

Substitute $t = nu$ to obtain

$$\Gamma(x) = \lim_{n\to\infty} \int_0^1 n^{x-1}u^{x-1}\left(1-u\right)^n \; n\, du = \lim_{n\to\infty} n^x \int_0^1 u^{x-1}\left(1-u\right)^n \; du$$

For a moment, consider only the integral and integrate by parts starting with $u^{x-1} = \frac{d}{du}\frac{u^x}{x}$.

$$\int_0^1 u^{x-1}\left(1-u\right)^n \; du = \frac{u^x}{x}(1-u)^n\Big|_0^1 + \int_0^1 \frac{u^x}{x} n(1-u)^{n-1}\, du = \frac{n}{x}\int_0^1 u^x\left(1-u\right)^{n-1} \; du$$

Repeating the process gives

$$\frac{n(n-1)}{x(x+1)}\int_0^1 u^{x+1}(1-u)^{n-2}\, du = \frac{n(n-1)(n-2)}{x(x+1)(x+2)}\int_0^1 u^{x+2}(1-u)^{n-3}$$

The next to last result is

$$\frac{n!}{x(x+1)(x+2)\ldots(x+n-1)}\int_0^1 u^{x+n-1}(1-u)^0\, du$$

and the final result is

$$\frac{n!}{x(x+1)(x+2)\ldots(x+n)}\; u^{x+n}\Big|_0^1 = \frac{n!}{x(x+1)(x+2)\ldots(x+n)}$$

The result clearly follows.

16.4 The Gamma Function as an Infinite Product

We are now going to develop the theory of the Gamma function from scratch, without referring to any previous result in the chapter. The new approach leads to a deeper understanding of the central results. At the end, we will prove that both approaches describe the same function.

The Gamma function has simple poles at $0, -1, -2, \ldots$, so $\frac{1}{\Gamma(z)}$ has simple zeros at these points. This suggests that we start with the simplest Weierstrass product which gives these zeros:

Definition 36 (Tentative)**.**

$$G(z) = z\prod_{n>0} E_1\left(-\frac{z}{n}\right) = z\prod_{n>0}\left(1+\frac{z}{n}\right)e^{-\frac{z}{n}}$$

It follows from the standard considerations that this product converges over the entire complex plane to a function with simple zeros at $0, -1, -2, ...$, no other zeros, and no singularities.

Remark Our first result is the analogue of the fundamental equation of the zeta function, namely

$$\Gamma(z+1) = z\Gamma(z)$$

We seek a similar formula for $G(z+1)$. This operation shifts the G-function to the left by one integer, giving a function with simple roots at $-1, -2, -3, ...$ We could get such a function by dividing out the root of G at the origin: $\frac{G(z)}{z}$. Consequently, the geometry of the zeros suggests that

$$G(z+1) = \frac{G(z)}{z}$$

As we will see shortly, this formula is not quite correct, and an attempt to verify it by algebraic manipulation of the product leads to a mess. Instead, we will calculate the logarithmic derivative of both sides and compare the results.

The logarithmic derivative of $G(z)$ is

$$\frac{1}{z} + \sum_{n>0} \frac{\frac{1}{n}e^{-\frac{z}{n}} - \left(1+\frac{z}{n}\right)\frac{1}{n}e^{-\frac{z}{n}}}{\left(1+\frac{z}{n}\right)e^{-\frac{z}{n}}} = \frac{1}{z} + \sum_{n>0} \frac{-\frac{z}{n^2}}{\frac{n+z}{n}} = \frac{1}{z} - \sum_{n>0} \frac{z}{(n+z)n} = \frac{1}{z} + \sum_{n>0}\left(\frac{1}{n+z} - \frac{1}{n}\right)$$

so the logarithmic derivative of $G(z+1)$ is the previous expression with each z replaced by $z+1$, i.e.,

$$\frac{1}{z+1} + \sum_{n>0}\left(\frac{1}{n+z+1} - \frac{1}{n}\right)$$

This equals

$$\frac{1}{z+1} + \left(\frac{1}{z+2} - \frac{1}{1}\right) + \left(\frac{1}{z+3} - \frac{1}{2}\right) + \left(\frac{1}{z+4} - \frac{1}{3}\right) + \left(\frac{1}{z+5} - \frac{1}{4}\right) + ...$$

We can rearrange this slightly

$$\left(\frac{1}{z+1} - \frac{1}{1}\right) + \left(\frac{1}{z+2} - \frac{1}{2}\right) + \left(\frac{1}{z+3} - \frac{1}{3}\right) + \left(\frac{1}{z+4} - \frac{1}{4}\right) + \frac{1}{z+5} + ...$$

A little thought shows that this is

$$\sum_{n>0}\left(\frac{1}{n+z} - \frac{1}{n}\right)$$

The logarithmic derivative of $\frac{G(z)}{z} = z^{-1}G(z)$ is the sum of $\frac{-1}{z}$ and the logarithmic derivative of $G(z)$, and thus equals the same expression

$$\sum_{n>0}\left(\frac{1}{n+z} - \frac{1}{n}\right)$$

So the logarithmic derivatives of $\frac{G(z)}{z}$ and $G(z+1)$ are equal, and therefore there is a constant γ such that

$$\operatorname{Log}\frac{G(z)}{z} = \operatorname{Log} G(z+1) + \gamma$$

It immediately follows that $\operatorname{Log}\frac{G(z)e^{\gamma z}}{z} = \operatorname{Log} G(z+1)e^{\gamma(z+1)}$, and therefore

$$\frac{G(z)e^{\gamma z}}{z} = G(z+1)e^{\gamma(z+1)}$$

Our next task is to determine the constant γ. We do this by setting $z = 0$. Notice that $\frac{G(z)}{z} = \prod_{n>0}\left(1+\frac{z}{n}\right)e^{-\frac{z}{n}}$, so the value of this expression at zero is 1. The value of the right side at $z = 0$ is $G(1)e^{\gamma} = \prod\left(1+\frac{1}{n}\right)e^{-\frac{1}{n}}e^{\gamma}$. So

$$1 = \lim_{N\to\infty}\prod_{n=1}^{N}\left(\frac{n+1}{n}\right)e^{-(1+\frac{1}{2}+\ldots+\frac{1}{N})}e^{\gamma} = \lim_{N\to\infty}(N+1)e^{-(1+\frac{1}{2}+\ldots+\frac{1}{N})}e^{\gamma}$$

This can be rewritten

$$1 = \lim_{N\to\infty} e^{\log(N+1)-(1+\frac{1}{2}+\ldots+\frac{1}{N})}e^{\gamma}$$

and from this it follows that

$$\gamma = \lim_{N\to\infty}\left[(1+\frac{1}{2}+\ldots+\frac{1}{N}) - \log(N+1)\right]$$

This number is called *Euler's constant* or the Euler-Mascheroni constant. It's value is

$$\gamma = 0.57721566490153286060651209008240243104215933593992\ldots$$

It is not known if the constant is rational, algebraic, or transcendental. Note that the constant is the shaded area in the picture below. It is easy to show that the limit defining this constant exists.

Many sources write instead

$$\gamma = \lim_{N\to\infty}\left[(1+\frac{1}{2}+\ldots+\frac{1}{N}) - \log N\right]$$

which is equivalent to our definition in the limit.

Putting all this together, we give a better definition of $G(z)$ and $\Gamma(z)$:

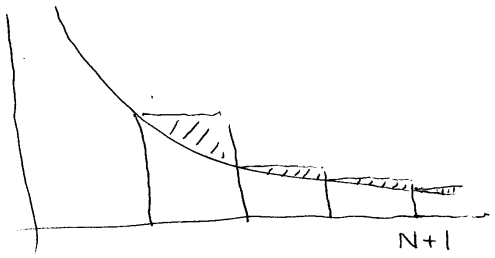


Figure 16.1: Euler-Mascheroni Constant

Definition 37 (Final Definitions)**.**

$$G(z) = z\, e^{\gamma z} \prod_{n>0} \left(1 + \frac{z}{n}\right) e^{-\frac{z}{n}}$$

$$\Gamma(z) = \frac{1}{G(z)}$$

Theorem 95. *These functions have the following properties*

- $G(z)$ *is holomorphic with simple zeros at* $0, -1, -2, \ldots$ *and no other zeros.*
- $G(z+1) = \frac{G(z)}{z}$
- $G(1) = 1$
- $\Gamma(z)$ *is meromorphic with simple poles at* $0, -1, -2, \ldots$*, no other poles, and no zeros.*
- $\Gamma(z+1) = z\,\Gamma(z)$
- $\Gamma(1) = 1$
- *If* $n > 0$ *is an integer,* $\Gamma(n) = (n-1)!$

Theorem 96.

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

Remark: The underlying philosophy of this section is that functions should be almost determined by their zeros, and the basic identities of the subject should be consequences of the geometry of these zeros. In the previous chapter we studied the case when there are zeros at the integers, and reconstructed trigonometric identities. The Gamma function arises from the case when the zeros occur on non-positive integers, forming a *ray* rather than a *lattice line*. There are two obvious geometrical symmetries. We can shift left and get rid of one zero, or we

can reflect and multiply to get zeros at all integers. The first symmetry leads to the fundamental identity $\Gamma(z+1) = z\Gamma(z)$, and the second leads to the current theorem. Notice below that this theorem implies $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

Proof: Since $\Gamma(z+1) = z\Gamma(z)$, we have $\Gamma(-z+1) = (-z)\Gamma(-z)$. So

$$\Gamma(z)\Gamma(1-z) = (-z)\Gamma(z)\Gamma(-z) = \frac{-z}{G(z)G(-z)}$$

But

$$G(z)G(-z) = ze^{\gamma z}\prod_{n>0}\left(1+\frac{z}{n}\right)e^{-\frac{z}{n}} \cdot (-z)e^{-\gamma z}\prod_{n>0}\left(1-\frac{z}{n}\right)e^{\frac{z}{n}} = -z^2\prod_{n>0}\left(1-\frac{z^2}{n^2}\right)$$

Also from chapter 15 we have

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

Consequently

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

and

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

Corollary 25.

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

Proof: $\Gamma^2\left(\frac{1}{2}\right) = \pi$ by the previous theorem.

Theorem 97.

$$\Gamma(z)\,\Gamma\left(z+\frac{1}{2}\right) = 2^{1-2z}\sqrt{\pi}\,\Gamma(2z)$$

Corollary 26.

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

Proof of Corollary: Set $z = 1$ in the previous formula. Then $\Gamma(1)\,\Gamma(\frac{3}{2}) = 2^{-1}\sqrt{\pi}\Gamma(2)$, so $\frac{1}{2}\Gamma(\frac{1}{2}) = \frac{1}{2}\sqrt{\pi}$.

Proof of theorem: The product $G(z)G(z+\frac{1}{2})$ has simple zeros at $0,-1,-2,...$ and $-\frac{1}{2},-\frac{3}{2},...$, and $G(2z)$ has simple zeros at $2z=-n$, i.e., $z=-\frac{n}{2}$. So we expect an identity comparing $G(z)G(z+\frac{1}{2})$ and $G(2z)$. To find the identity, compute the logarithmic derivative of $G(z)$:

$$\frac{1}{z}+\gamma+\sum_n \frac{\frac{1}{n}e^{-\frac{z}{n}}-\frac{1}{n}(1+\frac{z}{n})e^{-\frac{z}{n}}}{(1+\frac{z}{n})e^{-\frac{z}{n}}}=\frac{1}{z}+\gamma-\sum\frac{-\frac{z}{n^2}}{(1+\frac{z}{n})}=\frac{1}{z}+\gamma+\sum\left(\frac{1}{n+z}-\frac{1}{n}\right)$$

It follows that the logarithmic derivative of $G(z)G(z+\frac{1}{2})$ is

$$\frac{1}{z}+\gamma+\sum\left(\frac{1}{n+z}-\frac{1}{n}\right)+\frac{1}{z+\frac{1}{2}}+\gamma+\sum\left(\frac{1}{n+\frac{1}{2}+z}-\frac{1}{n}\right)$$

and the derivative of this expression is

$$\sum_{k\geq 0}\frac{-1}{\left(\frac{k}{2}+z\right)^2}$$

The logarithmic derivative of $G(2z)$ is $\frac{2G'(2z)}{G(2z)}$ and thus twice the logarithmic derivative of $G(z)$ with each z replaced by $2z$. The derivative of this expression produces another 2 and thus is 2^2 times $\frac{d}{dz}\frac{G'(z)}{G(z)}$ with each z replaced by $2z$. Thus it equals

$$2^2\left(-\frac{1}{(2z)^2}-\sum\frac{1}{(n+2z)^2}\right)$$

This is

$$-\sum_{n\geq 0}\frac{1}{(\frac{n}{2}+z)^2}$$

If follows that the logarithmic derivative of $G(z)G(z+\frac{1}{2})$ and $G(2z)$ differ by an additive constant B, so that the logarithmic derivatives of $G(z)G(z+\frac{1}{2})$ and $G(2z)e^{Bz}$ agree. But $\frac{f'}{f}=\frac{g'}{g}$ implies $\frac{f'g-fg'}{fg}=0$ and thus $\frac{d}{dz}\frac{f}{g}=0$. So $\frac{f}{g}$ is a constant e^A and thus $f=e^Ag$. We conclude that

$$G(z)G(z+\frac{1}{2})=e^{A+Bz}G(2z)$$

and therefore that

$$\Gamma(z)\Gamma(z+\frac{1}{2})=e^{-A-Bz}\Gamma(2z)$$

To determine A and B, first set $z=\frac{1}{2}$ and then set $z=1$. We obtain $\sqrt{\pi}=e^{-A-B/2}$, and $\frac{1}{2}\sqrt{\pi}=e^{-A-B}$. Dividing the first by the second gives $2=e^{B/2}$, so $B=2\ln 2$. Thus $\sqrt{\pi}=e^{-A}e^{-(\ln 2)}=e^{-A}\frac{1}{2}$. So $e^{-A}=2\sqrt{\pi}$ and $e^{-A-Bz}=e^{-A}e^{-Bz}=2\sqrt{\pi}e^{-2\ln 2z}=2\sqrt{\pi}2^{-2z}=2^{1-2z}\sqrt{\pi}$.

16.5 The Two Definitions of $\Gamma(z)$ Agree

To prove the result in the section title, it suffices to prove that the two definitions agree on positive real numbers, by the identity theorem. In the section on the integral definition, we proved that

$$\Gamma(x) = \lim_{n\to\infty} n^x \frac{n!}{x(x+1)(x+2)\dots(x+n)}$$

Rewriting this expression

$$\frac{1}{\Gamma(x)} = \lim_{n\to\infty} n^{-x}\, \frac{x(x+1)(x+2)\dots(x+n)}{n!}$$

$$= \lim_{n\to\infty} n^{-x}\, x\,(1+x)\left(1+\frac{x}{2}\right)\left(1+\frac{x}{3}\right)\dots\left(1+\frac{x}{n}\right)$$

$$= \lim_{n\to\infty} x\, e^{-x\log n} e^{\left(1+\frac{1}{2}+\dots+\frac{1}{n}\right)x} \prod_{n>0}\left(1+\frac{x}{n}\right)e^{-\frac{x}{n}}$$

$$= \lim_{n\to\infty} x\, e^{\left(1+\frac{1}{2}+\dots+\frac{1}{n}-\log n\right)x} \prod_{n>0}\left(1+\frac{x}{n}\right)e^{-\frac{x}{n}}$$

$$= xe^{\gamma x} \prod_{n>0}\left(1+\frac{x}{n}\right)e^{-\frac{x}{n}} = G(x)$$

16.6 Stirling's Formula

Recall that $\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$. An easy way to see this is to note that the average of $1, 2, \dots, n$ is $\frac{n+1}{2}$. Instead of adding lots of different numbers, we could just add their average over and over, giving $n\left(\frac{n+1}{2}\right)$.

Stirling's formula is about multiplying rather than adding. Instead of multiplying $1, 2, \dots n$, we could take their *multiplicative average* A and raise it to the nth power. But what is this multiplicative average? Stirling's formula will show that it is $\frac{n}{e}$. So while the additive average of the first n numbers is roughly $\frac{n}{2}$, the multiplicative average is roughly $\frac{n}{2.718}$ and $n! \sim \left(\frac{n}{e}\right)^n$.

However, it is not quite this simple. The expression $\left(\frac{n}{e}\right)^n$ is too small, and as n gets larger, the error grows bigger. We would like the quotient of $n!$ and its approximation to approach one. Stirling provides a very surprising correction to make that happen: $\sqrt{2\pi n}$. Thus the real Stirling formula states that

$$n! \sim \left(\frac{n}{e}\right)^n \sqrt{2\pi n}$$

A rough proof of Stirling's formula without this correction term is easy;

$$\log n! = \sum_{k=1}^{n} \log k \sim \int_1^n \log t \; dt = (t \log t - t)\Big|_1^n = (n \; \log n - n) + 1 \sim n \; \log n - n$$

and so $n! \sim e^{n \; \log n - n} = \left(\frac{n}{e}\right)^n$.

History: Stirling's formula is due to DeMoivre, who provided a complicated formula for the correction term. Stirling observed that that complicated formula gives $\sqrt{2\pi n}$.

Theorem 98 (Stirling). *The quotient of the two terms below converges to* 1 *as* $n \to \infty$.

$$n! \sim \left(\frac{n}{e}\right)^n \sqrt{2\pi n}$$

Proof: There are many proofs of this result. The proof below is motivated by the Euler-Maclaurin formula, discovered around 1735. This formula approximates the difference between $\sum_{k=a}^{b} f(k)$ and $\int_a^b f(x)\,dx$ using expressions given by values of higher derivatives of f, and is important in numerical analysis. Using the formula, Euler could estimate infinite sums with surprising accuracy. In particular he approximated $\sum_{k=1}^{N} \frac{1}{k^2}$ that way, perhaps to check that $\frac{\pi^2}{6}$ is the correct value for the infinite sum.

Below are three examples. The $\mathcal{E}$ are error terms, given by formulas that are easily estimated to show that the remaining error is very small.

$$\begin{aligned}
\sum f - \int f &= \frac{f(a)+f(b)}{2} + \mathcal{E}(f'') \\
&= \frac{f(a)+f(b)}{2} + \frac{1}{6}\frac{f'(b)-f'(a)}{2!} + \mathcal{E}(f''') \\
&= \frac{f(a)+f(b)}{2} + \frac{1}{6}\frac{f'(b)-f'(a)}{2!} - \frac{1}{30}\frac{f'''(b)-f'''(a)}{4!} + \mathcal{E}(f'''''')
\end{aligned}$$

We will use a slightly modified special case of this formula:

$$\sum_{k=1}^{N} f(k) - \int_{\frac{1}{2}}^{N+\frac{1}{2}} f(x)\,dx = -\frac{1}{2}\int_{\frac{1}{2}}^{N+\frac{1}{2}} f''(x)\left[D\left(x+\frac{1}{2}\right)\right]^2 dx$$

where $D(x)$ is the distance from x to the nearest integer. Each integral above is a sum of integrals over smaller intervals, so it is enough to prove a typical case, and the most symmetrical case is

$$f(0) - \int_{-\frac{1}{2}}^{\frac{1}{2}} f(x)\,dx = -\frac{1}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} f''(x)\left[D\left(x+\frac{1}{2}\right)\right]^2 dx$$

We do this integrating over $\left[-\frac{1}{2}, 0\right]$ and $\left[0, \frac{1}{2}\right]$ separately, and integrating both expressions twice by parts. Thus

$$\int_{-\frac{1}{2}}^{0} f(x)\,dx = \int_{-\frac{1}{2}}^{0} f(x)\frac{d}{dx}\left(x+\frac{1}{2}\right)\,dx = f(x)\left(x+\frac{1}{2}\right)\Big|_{-\frac{1}{2}}^{0} - \int_{-\frac{1}{2}}^{0} f'(x)\left(x+\frac{1}{2}\right)\,dx$$

$$= \frac{f(0)}{2} - \int_{-\frac{1}{2}}^{0} f'(x)\frac{d}{dx}\frac{1}{2}\left(x+\frac{1}{2}\right)^2\,dx = \frac{f(0)}{2} - f'(x)\frac{1}{2}\left(x+\frac{1}{2}\right)^2\Big|_{-\frac{1}{2}}^{0} + \frac{1}{2}\int_{-\frac{1}{2}}^{0} f''(x)\left(x+\frac{1}{2}\right)^2\,dx$$

$$= \frac{f(0)}{2} - \frac{f'(0)}{8} + \frac{1}{2}\int_{-\frac{1}{2}}^{0} f''(x)\left(x+\frac{1}{2}\right)^2\,dx$$

and in the same way

$$\int_{0}^{\frac{1}{2}} f(x)\,dx = \int_{0}^{\frac{1}{2}} f(x)\frac{d}{dx}\left(x-\frac{1}{2}\right)\,dx = \frac{f(0)}{2} + \frac{f'(0)}{8} + \frac{1}{2}\int_{0}^{\frac{1}{2}} f''(x)\left(x-\frac{1}{2}\right)^2\,dx$$

and the result follows.

Apply this result to $f(x) = \log x$; it says

$$\sum_{k=1}^{n} \log k - \int_{\frac{1}{2}}^{n+\frac{1}{2}} \log x\,dx = \frac{1}{2}\int_{\frac{1}{2}}^{n+\frac{1}{2}} \frac{1}{x^2}\left[D\left(x+\frac{1}{2}\right)\right]^2\,dx$$

or

$$\log n! = (x\ \log x - x)\Big|_{\frac{1}{2}}^{n+\frac{1}{2}} + \frac{1}{2}\int_{\frac{1}{2}}^{n+\frac{1}{2}} \frac{1}{x^2}\left[D\left(x+\frac{1}{2}\right)\right]^2\,dx$$

From this, we conclude that there is a constant C such that

$$\log n! = \left(n+\frac{1}{2}\right)\log n - n + C + O(1/x)$$

To see this, look first at the integral and observe that the term involving D is bounded by 1. So as $n \to \infty$, the integral converges, and for large n the difference between the finite integral and its limit is $O(1/x)$. In particular, the constant C is mainly determined by the limit of this integral, so when we evaluate the remaining items, there is no need to keep track of constant terms.

When we evaluate $x\log x - x$ at $n+\frac{1}{2}$, we get $\left(n+\frac{1}{2}\right)\log\left(n+\frac{1}{2}\right) - n - \frac{1}{2}$ and when we evaluate it at $\frac{1}{2}$, we get constants. We want that top number to be $\left(n+\frac{1}{2}\right)\log n - n$. The essential difference between what we have and what we want is

$$\left(n+\frac{1}{2}\right)\left(\log\left(n+\frac{1}{2}\right) - \log n\right) = \left(n+\frac{1}{2}\right)\log\left(1+\frac{1}{2n}\right)$$

Since the logarithm of a number is the area under $\frac{1}{x}$ between 1 and that number, $\log\left(1+\frac{1}{2n}\right)$ is at most $\frac{1}{2n}$ and the difference which worries us is at most $\left(n+\frac{1}{2}\right)\left(\frac{1}{2n}\right) = \frac{1}{2}+\frac{1}{4n}$. In fact this distance converges rapidly to $\frac{1}{2}$, finishing our claim.

We have shown that $\log n! \sim \left(n+\frac{1}{2}\right)\log n - n + C$ and thus

$$n! \sim e^{\left(n+\frac{1}{2}\right)\log n - n + C} = n^n\sqrt{n}e^{-n}e^C = \left(\frac{n}{e}\right)^n \sqrt{n}\, e^C$$

From now on, write C rather than e^C in this formula for $n!$ It remains to show that $C = \sqrt{2\pi}$. We can obtain this from Wallis' Product. Recall that

$$\frac{2}{1}\cdot\frac{2}{3}\cdot\frac{4}{3}\cdot\frac{4}{5}\cdot\frac{6}{5}\cdot\frac{6}{7}\cdot\ldots = \frac{1}{2n+1}\prod_{k=1}^{n}\frac{(2k)(2k)}{(2k-1)(2k-1)} \to \frac{\pi}{2}$$

Rewriting this,

$$\frac{1}{2n+1}\prod_{k=1}^{n}\frac{(2k)(2k)(2k)(2k)}{(2k-1)(2k)(2k-1)(2k)} = \frac{1}{2n+1}\frac{2^{4n}\left(n!\right)^4}{\left((2n)!\right)^2} \to \frac{\pi}{2}$$

Replacing each factorial with the Stirling approximation gives

$$\frac{1}{2n+1}\frac{2^{4n}\left(\frac{n}{e}\right)^{4n}\left(\sqrt{n}\right)^4 C^4}{\left(\frac{2n}{e}\right)^{4n}\left(\sqrt{2n}\right)^2 C^2} = \frac{1}{2n+1}\frac{n^2C^4}{2nC^2} = \frac{n}{4n+2}C^2 \to \frac{\pi}{2}$$

Since $\frac{n}{4n+2} \to \frac{1}{4}$, we conclude that $C^2 = 2\pi$ and $C = \sqrt{2\pi}$. QED.

16.7 Extending Stirling's Formula to the Complex Plane

Since $\Gamma(z)$ has no zeros in $\Re(z) > 0$, we can form $\log\Gamma(z)$ on this right half plane, choosing the branch which is real on the positive real axis.

Theorem 99. *If $\Re(z) > 0$,*

$$\log\Gamma(z) = \left(z-\frac{1}{2}\right)\log z - z + \frac{1}{2}\log(2\pi) + O\left(|z|^{-1}\right)$$

The error term means that there is a constant C such that the difference is at most $\frac{C}{|z|}$.

Remark: Thus

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

Recall that $\Gamma(z) = (z-1)!$. This explains why z is in the denominator of the square root, rather than the numerator.

Proof of Theorem: The gamma function has a zero of order 1 at the origin, so $\log \Gamma(z)$ behaves like $\log z$ near the origin. In our approximation, all terms are well behaved at the origin or blow up like $\log z$ there. But $\frac{1}{|z|}$ blows up much faster. So the estimate is true near the origin.

If we remain in the closed right half plane on the set $\epsilon \leq |z| \leq R$, the gamma function and all approximating terms are continuous and thus bounded. Consequently their difference is certainly smaller than $\frac{C}{|z|}$ on this set provided we choose an appropriate C. So it suffices to show that the required estimate is true for large z, say $|z| > 2$.

Recall that

$$G(z) = z\, e^{\gamma z} \prod_{n>0} \left(1 + \frac{z}{n}\right) e^{-\frac{z}{n}}$$

$$\Gamma(z) = \frac{1}{G(z)}$$

$$\gamma = \lim_{N\to\infty} \left[(1 + \frac{1}{2} + ... + \frac{1}{N}) - \log(N+1)\right]$$

Note that $\log(N+1)$ can be replaced by $\log(N)$ in the last formula since

$$\left[(1 + \frac{1}{2} + ... + \frac{1}{N}) - \log(N)\right] = \left[(1 + \frac{1}{2} + ... + \frac{1}{N}) - \log(N+1)\right] + [\log(N+1) - \log N]$$

and $\log(N+1) - \log N = \log(1 + 1/N) \to \log 1 = 0$. It follows that

$$\log \Gamma(z) = -\log G(z) = -\log z - \gamma z + \sum_{n=1}^{\infty} \left(\frac{z}{n} - \log(1 + z/n)\right) =$$

$$\lim_{N\to\infty} \left[-\log z - \left((1 + \frac{1}{2} + ... + \frac{1}{N}) - \log N\right) z + \sum_{n=1}^{N} \left(\frac{z}{n} - \log(n+z) + \log n\right)\right] =$$

$$\lim_{N\to\infty} \left[-\log z + z \log N + \log N! - \sum_{n=1}^{N} \log(n+z)\right] = \lim_{N\to\infty} \left[z \log N + \log N! - \sum_{n=0}^{N} \log(n+z)\right]$$

For a moment we are going to fix z and evaluate this limit. The key step is to obtain a better approximation for $\sum \log(n+z)$. In the proof of Stirling's formula, we obtained the equation

$$\sum_{k=1}^{N} f(k) - \int_{\frac{1}{2}}^{N+\frac{1}{2}} f(x)\, dx = -\frac{1}{2} \int_{\frac{1}{2}}^{N+\frac{1}{2}} f''(x) \left[D\left(x + \frac{1}{2}\right)\right]^2 dx$$

The same result is true if we start the sum at $k = 0$ and integrate from $-1/2$. Let $f(x) = \log(x+z)$ to obtain

$$\sum_{k=0}^{N} \log(k+z) = \int_{-\frac{1}{2}}^{N+\frac{1}{2}} \log(x+z)\,dx + \frac{1}{2}\int_{-\frac{1}{2}}^{N+\frac{1}{2}} \frac{1}{(x+z)^2}\left[D\left(x+\frac{1}{2}\right)\right]^2 dx$$

Consequently

$$\log\Gamma(z) = \lim_{N\to\infty}\left[z\log N + \log N! - \int_{-\frac{1}{2}}^{N+\frac{1}{2}} \log(x+z)\,dx - \frac{1}{2}\int_{-\frac{1}{2}}^{N+\frac{1}{2}} \frac{1}{(x+z)^2}\left[D\left(x+\frac{1}{2}\right)\right]^2 dx\right]$$

Since the integral of $\log(x+z)$ is $(x+z)\log(x+z) - (x+z)$, the first integral in the above formula is

$$\left(N+\frac{1}{2}+z\right)\log\left(N+\frac{1}{2}+z\right) - \left(N+\frac{1}{2}+z\right) - \left(-\frac{1}{2}+z\right)\log\left(-\frac{1}{2}+z\right) + \left(-\frac{1}{2}+z\right) =$$

$$\left(N+\frac{1}{2}+z\right)\log\left(N+\frac{1}{2}+z\right) - \left(-\frac{1}{2}+z\right)\log\left(-\frac{1}{2}+z\right) - N - 1 =$$

$$\left(N+\frac{1}{2}+z\right)\log N + \left(N+\frac{1}{2}+z\right)\log\left(1+\frac{1}{N}(\frac{1}{2}+z)\right) - \left(-\frac{1}{2}+z\right)\log\left(-\frac{1}{2}+z\right) - N - 1$$

It follows that $\log\Gamma(z)$ equals

$$\lim_{N\to\infty}\left[z\log N + \log N! - \left(N+\frac{1}{2}+z\right)\log N - \left(N+\frac{1}{2}+z\right)\log\left(1+\frac{1}{N}(\frac{1}{2}+z)\right) + \left(-\frac{1}{2}+z\right)\log\left(-\frac{1}{2}+z\right)\right.$$

$$\left. +N + 1 - \frac{1}{2}\int_{-\frac{1}{2}}^{N+\frac{1}{2}} \frac{1}{(x+z)^2}\left[D\left(x+\frac{1}{2}\right)\right]^2 dx\right]$$

We want to take the limit of this expression as $N \to \infty$. The term $\left(N+\frac{1}{2}+z\right)\log\left(1+\frac{1}{N}(\frac{1}{2}+z)\right)$ goes to $\frac{1}{2}+z$. Indeed the logarithm in this term goes to zero, so the last two terms in the initial $\left(N+\frac{1}{2}+z\right)$ are sent to zero. Near 1, the log term is about $\frac{1}{N}(\frac{1}{2}+z)$, and when this is multiplied by N, the term is about $\frac{1}{2}+z$.

Consequently $\log\Gamma(z)$ is the limit as $N \to \infty$ of

$$\left[\log N! - N\log N + N - \frac{1}{2}\log N - (\frac{1}{2}+z) + \left(-\frac{1}{2}+z\right)\log\left(-\frac{1}{2}+z\right) + 1 - \frac{1}{2}\int_{-\frac{1}{2}}^{N+\frac{1}{2}} \frac{1}{(x+z)^2}\left[D\left(x+\frac{1}{2}\right)\right]^2 dx\right]$$

But Stirling's formula says that $N! \sim \left(\frac{N}{e}\right)^N \sqrt{2\pi N}$ and so $\log N! - N\log N + N - \frac{1}{2}\log N$ converges to the constant $\sqrt{2\pi}$ as N goes to infinity. We conclude that

$$\log\Gamma(z) = -(\frac{1}{2}+z) + \left(-\frac{1}{2}+z\right)\log\left(-\frac{1}{2}+z\right) + \sqrt{2\pi} + 1 - \frac{1}{2}\int_{-\frac{1}{2}}^{\infty} \frac{1}{(x+z)^2}\left[D\left(x+\frac{1}{2}\right)\right]^2 dx$$

This formula is exactly true. We will use it to estimate $\log\Gamma(z)$ for large z. The key step will show that we can ignore the integral. But let us clear up the initial terms first. Notice that

$$\left(-\frac{1}{2}+z\right)\log\left(-\frac{1}{2}+z\right) = \left(-\frac{1}{2}+z\right)\log z + \left(-\frac{1}{2}+z\right)\left(\log\left(-\frac{1}{2}+z\right) - \log z\right) =$$

$$\left(-\frac{1}{2}+z\right)\log z + \left(-\frac{1}{2}+z\right)\log\left(-\frac{1}{2z}+1\right) =$$

$$\left(-\frac{1}{2}+z\right)\log z - \frac{1}{2}\log\left(-\frac{1}{2z}+1\right) + z\log\left(-\frac{1}{2z}+1\right)$$

If z is large, the first term is large, so we must keep it. For large z, the term $\log\left(-\frac{1}{2z}+1\right)$ goes to zero, so the middle term is irrelevant. In the final term, z is large and the log is small, so we must be more careful. For large z, $\log\left(-\frac{1}{2z}+1\right)$ is about $-\frac{1}{2z}$, so the final term is about $-1/2$

Putting this altogether, we find that for large z, $\log\Gamma(z)$ is an integral term plus about

$$-\left(\frac{1}{2}+z\right) + \left(-\frac{1}{2}+z\right)\log z - \frac{1}{2} + \sqrt{2\pi} + 1 = z\log z - z - \frac{1}{2}\log z + \sqrt{2\pi}$$

and this is exactly the estimate we desire. So it suffices to show that the integral is $O(\frac{1}{|z|})$.

We will estimate the absolute value of the integral using $\left|\int f(z)\,dz\right| \le \int |f(z)|\,dz$. Since $|z| = |(x+z)-x| \le |x+z| + |x|$, we have $|z| - |x| \le |z+x|$. Similarly $|x| - |z| \le |z+x|$. One of the two terms on the left is non-negative, so

$$\frac{1}{(|z|-|x|)^2} \ge \left|\frac{1}{(z+x)^2}\right|$$

The left side will be infinite when $x = |z|$, but if we integrate from $2|z|$ to ∞, our integral is bounded by

$$\int_{2|z|}^{\infty} \frac{1}{(|z|-|x|)^2}\,dx = \int_{2|z|}^{\infty} \frac{1}{(|z|-x)^2}\,dx = \left.\frac{1}{|z|-x}\right|_{2|z|}^{\infty} = \frac{1}{|z|}$$

To finish, we must estimate the integral over $-\frac{1}{2} \leq x \leq 2|z|$. The length of this interval is $2|z| + \frac{1}{2}$. If $z = a + ib$,

$$|(z+x)^2| = (a+x)^2 + b^2 = a^2 + b^2 + 2ax + x^2 \geq |z|^2 + 2ax$$

If $x > 0$ then $2ax > 0$ and this expression is greater than $|z|^2$. Otherwise $-\frac{1}{2} \leq x$ and $2ax$ could be as negative as $-a = -\Re(z)$. Since the real part of z is at most $|z|$, we conclude that $|(z+x)^2| \geq |z|^2 - |z| = |z|(|z| - 1)$.

We only need prove our estimate for large z, so we can assume that $|z| > 2$. Then $|z| - 1 > |z| - |z|/2 = |z|/2$ and $|(z+x)^2| \geq |z|\left(\frac{|z|}{2}\right) = \frac{|z|^2}{2}$. So the key expression in the integrand is

$$\left|\frac{1}{(z+x)^2}\right| \leq \frac{2}{|z|^2}$$

and the integral is at most this times the length of the interval, or

$$\frac{2}{|z|^2}\left(2|z| + \frac{1}{2}\right) \sim \frac{4}{|z|}$$

QED.

Chapter 17

The Zeta Function and the Prime Number Theorem

17.1 Some History

The first calculus results discovered by Leibniz concerned infinite series. In particular, he was asked to evaluate $\sum_{n=1}^{\infty} \frac{1}{n(n+1)}$ and found the ingenious solution

$$\sum_{n=1}^{\infty} \frac{1}{n(n+1)} = \sum_{n=1}^{\infty} \left(\frac{1}{n} - \frac{1}{n+1} \right) = \left(1 - \frac{1}{2} \right) + \left(\frac{1}{2} - \frac{1}{3} \right) + \left(\frac{1}{3} - \frac{1}{4} \right) + \ldots = 1$$

Later by a deeper method he discovered that

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \ldots$$

Both Leibniz and Newton were on the lookout for *general methods* rather than isolated results. Therefore, it was natural for Leibniz to next consider $\sum \frac{1}{n^2}$, but he was unable to calculate this sum. Then Leibniz heard that Newton knew the answer, so he wrote Newton indirectly via the Royal Society, offering to trade his proof of the $\frac{\pi}{4}$ series for the sum $\sum \frac{1}{n^2}$. Newton replied with two long letters revealing large amounts of the calculus. Each ended with an anagram suggesting further results that Newton was keeping secret.

The letters are well-worth reading, but may have disappointed Leibniz. Newton already knew his series for $\frac{\pi}{4}$ and told Leibniz that he knew other series which converged faster, for instance

$$\frac{\pi}{2\sqrt{2}} = 1 + \frac{1}{3} - \frac{1}{5} - \frac{1}{7} + \frac{1}{9} + \frac{1}{11} - \ldots$$

On the other hand, Newton only know approximations for $\sum \frac{1}{n^2}$.

The problem of computing $\sum \frac{1}{n^2}$ passed from Leibniz to his pupils the Bernoulli's, and from them to Euler, who eventually found the value we discussed earlier. But Euler went much further, particularly in his two volume work *Introductio in Analtsin Infinitorum* of 1748, which has been called "the most important mathematics textbook of modern times." In this book Euler wrote of the relation between the zeta function and prime numbers

$$\zeta(x) = \sum_{n=1}^{\infty} \frac{1}{n^x} = \prod_p \frac{1}{\left(1 - \frac{1}{p^z}\right)}$$

and used this result to prove the first real advance in prime number theory since the Greeks, namely

$$\sum_p \frac{1}{p} = \infty$$

This sum for the zeta function only converges for $x > 1$, and it is for these values that the product formula holds.

Since the product formula rests on very easy ideas, we stop to explain. Here is a simple proof that there are infinitely many primes. If not, we could form the *finite* product

$$\left(1 + \frac{1}{2} + \frac{1}{2^2} + ...\right)\left(1 + \frac{1}{3} + \frac{1}{3^2} + ...\right)\left(1 + \frac{1}{5} + \frac{1}{5^2} + ...\right)... = \sum \frac{1}{2^{k_1} 3^{k_2} 5^{k_3} ...} = \sum \frac{1}{n}$$

All terms are positive, so the rearrangement is legal. The final sum equals $\sum \frac{1}{n}$ since every integer can be uniquely factored into primes. On the other hand, $\sum \frac{1}{n}$ diverges, but the individual product terms are all finite because the term for p equals $\frac{1}{1-\frac{1}{p}}$. QED.

This simple argument is also enough to prove Euler's result that $\sum \frac{1}{p}$ diverges. Indeed, the previous argument shows that

$$\lim_{N\to\infty} \prod_{p\le N} \left(1 + \frac{1}{p} + \frac{1}{p^2} + ...\right) = \sum \frac{1}{n} = \infty$$

and so

$$\lim_{N\to\infty} \sum_{p\le N} \log\left(\frac{1}{1 - \frac{1}{p}}\right) = \lim_{N\to\infty} \sum_{p\le N} -\log\left(1 - \frac{1}{p}\right) = \infty$$

If $0 < x < 1$ we have $\frac{1}{1-x} = 1 + x + x^2 + ...$ and so

$$-\log(1-x) = x + \frac{x^2}{2} + \frac{x^3}{3} + ...$$

Therefore

$$\lim_{N\to\infty} \sum_{p\leq N} \left(\frac{1}{p} + \frac{1}{2p^2} + \frac{1}{3p^3} + ...\right) = \infty$$

Rewrite this as

$$\sum_p \frac{1}{p} + \sum \frac{1}{p^2}\left(\frac{1}{2} + \frac{1}{3p} + \frac{1}{4p^2} + ...\right) = \infty$$

But the second term is smaller than $\sum \frac{1}{p^2}\left(1 + \frac{1}{2} + \frac{1}{2^2} + ...\right) \leq 2\sum \frac{1}{n^2}$ and thus finite, and we are done. QED.

Because $\sum \frac{1}{p} = \infty$ and $\sum \frac{1}{n^2}$ converges, there are far more primes than perfect squares.

Euler first proved the divergence of the sum of reciprocals of primes in 1737, and that date is often regarded as the beginning of analytic number theory.

Euler realized that if we replace p by p^x for $x > 1$, then we get a convergent product. We can write

$$\frac{1}{\left(1 - \frac{1}{p^z}\right)} = 1 + \frac{1}{p^x} + \frac{1}{p^{2x}} + \frac{1}{p^{3x}} + ...$$

and our product is an infinite product of such series. Consider the related product using only the values $p = 2, 3, 5$. It would be

$$\left(1 + \frac{1}{2^x} + \frac{1}{2^{2x}} + ...\right)\left(1 + \frac{1}{3^x} + \frac{1}{3^{2x}} + ...\right)\left(1 + \frac{1}{5^x} + \frac{1}{5^{2x}} + ...\right) = \sum \frac{1}{\left(2^{k_1}3^{k_2}5^{k_3}\right)^x}$$

Since every integer can be written uniquely as a product of prime powers, this is $\sum \frac{1}{n^x}$ over integers whose prime factorization contains only 2, 3, 5. In the limit as more and more terms of the product are included, we approach $\sum_n \frac{1}{n^x}$. So Euler's identity is essentially an analytic form of unique factorization into primes.

17.2 More History

One of the frustrations of prime number theory is that it is easy to make conjectures about primes of special forms, but almost impossible to prove such conjectures. Consider, for examples, primes of the form $2^n \pm 1$. These primes occur naturally in various branches of mathematics.

By definition, a Mersenne prime is a prime of the form $2^n - 1$. The importance of these primes was first noticed by Euclid, who proved that if $2^n - 1$ is prime, then $N = 2^n(2^n - 1)$ is perfect, that is, the sum of the divisors of N that are smaller than N is exactly N. Two thousand years later, Euler proved that all even perfect numbers have this form. Nobody knows if there are any odd perfect numbers.

The first few Mersenne primes are $3, 7, 31$, and 127, corresponding to the perfect numbers $6, 28, 496$, and 8128. These are the only perfect numbers known to the Greeks.

The crucial fact about Mersenne primes is that if $2^n - 1$ is prime, then n itself must be prime. This is easily proved. If we restrict attention to prime p, then $2^p - 1$ seems to be prime or not purely by chance. Since the rough distribution of prime numbers is known, a brief calculation gives a conjectured distribution of Mersenne primes, and in particular suggests that there are infinitely many Mersenne primes. But this has never been proved. On the other hand, very extensive computer calculations agree with the conjectured distribution of Mersenne primes. The largest known primes have historically been Mersenne primes.

Prime numbers of the form $2^n + 1$ are called Fermat primes. It is easy to prove that $2^n + 1$ can only be prime if n is a power of two. So we actually write $F_n = 2^{2^n} + 1$. If we restrict attention to powers of two, F_n seems to be prime or not purely by chance.

The first few Fermat primes correspond to $n = 0, 1, 2, 3, 4$ and are $3, 5, 17, 257, 65537$. Fermat primes are important because Gauss proved that an n-sided polygon can be constructed by straightedge and compass if and only if n is a power of 2 times a product of distinct Fermat Primes.

Fermat primes grow very rapidly and thus are difficult to analyze by hand. The number corresponding to $n = 5$ is $F_5 = 4294967297$. This was first studied by Euler, who proved indirectly that it is *not* prime. Aside from the first five primes, no other Fermat primes have been discovered.

Since the rough distribution of primes is known, it is possible to calculate the expected number of Fermat primes, and this calculation suggests that there should only be a small finite number of Fermat primes. But this has never been proved.

17.3 Dirichlet

In the midst of all of these conjectures and questions, it must have been a relief when in 1837 Dirichlet managed to prove something very important along similar lines.

In base ten, every prime digit except 2 and 5 ends in 1, 3, 7, or 9. Since there are infinitely many primes, infinitely many primes must end in at least one of these digits. Dirichlet proved that infinitely many primes end in each digit.

More generally, let $b \geq 2$ and write integers with base b. Let $0 < a < b$. If a and b are not relatively prime and the prime p divides both, then p divides every digit in base b which ends in a, so at most one prime ends in a. Suppose, on the other hand, that a and b are relatively prime. Dirichlet proved that there are infinitely many primes which end in the digit a when written in base b. Obviously each of these primes has the form $a + nb$, and Dirichlet's theorem asserts that this arithmetic progression contains infinitely many primes.

Dirichlet's result seems even greater when his astonishing proof is read. We'll give that proof in a later chapter.

For now, we remark that a small number of special cases of Dirichlet's theorem can be proved along the lines of Euclid's proof that there are infinitely many primes. But most cases are proved as Dirichlet did by showing that

$$\sum_{p \equiv a \pmod{b}} \frac{1}{p} = \infty$$

Dirichlet was born in 1805. After his student days, he traveled to Paris, where he studied mathematics from 1822 to 1826. Among other things, he carefully read Gauss's *Disquisitiones Arithmeticae*. Later reports say that he slept with this book under his pillow. Near the end of his Paris stay, Dirichlet proved Fermat's last theorem for $n = 5$, a result which made him famous.

After a short stay in Breslau, Dirichlet moved to Berlin. At the time he was 23 years old, so his position in Berlin was temporary until 1831. In Berlin, Dirichlet often attended weekly gatherings of Berlin artists in the home of Abraham Mendelssohn Bartholdy, and he eventually married Felix Mendelssohn's sister.

Dirichlet had a reputation as an excellent teacher, and it is in this capacity that he met

17.4 Riemann

Riemann was born in 1826. He attended Gottingen as an undergraduate. Gauss was there, but unapproachable, so Riemann transferred to Berlin where he heard the lectures of Dirichlet. Dirichlet had a profound influence on him — recall the Dirichlet principal used to prove the Riemann mapping theorem — so no doubt Riemann was familiar with Dirichlet's proof of the theorem on primes in an arithmetic progression.

Eventually Riemann returned to Gottingen, and wrote his PhD thesis under Gauss. In 1855, Gauss died and Dirichlet moved to Gottingen as his successor. In this position, he resumed contact with Riemann and used his influence to retain Riemann on the teaching staff.

Dirichlet died in 1859. Riemann was appointed his successor. That year, Riemann was elected a member of the Berlin Academy of Sciences, and required as a new member to send a report

on his recent work. His report was the eleven page paper *On the number of primes less than a given magnitude*, his only paper on number theory and one of the most important number theory papers ever written. The results in this chapter are largely from that paper.

Riemann married in 1862, but developed tuberculosis later that year, and died of it in 1866.

17.5 The Riemann Zeta Function

Definition 38. *The Riemann Zeta Function is the function*

$$\zeta(z) = \sum_{n=1}^{\infty} \frac{1}{n^z}$$

Remark: Here $n^z = e^{z \ln n}$ and $\ln n$ is the real branch of the logarithm.

Theorem 100.

- *This sum converges absolutely for* $\Re(z) > 1$.
- *This sum converges uniformly on compact subsets of* $\Re(z) > 1$.
- *The resulting function is holomorphic on* $\Re(z) > 1$
- *The function is also given by a product*

$$\zeta(z) = \prod_p \frac{1}{\left(1 - \frac{1}{p^z}\right)}$$

- *This sum diverges for* $\Re(z) \leq 1$.

Proof: Notice that $|n^z| = |e^{(x+iy)\ln n}| = n^x$. Since $\sum \frac{1}{n^x}$ converges by the integral test, absolute convergence follows by the comparison test. Uniform convergence on compact subsets also follows by the Weierstrass M-test, since we can bound terms with $\frac{1}{n^{x_0}}$ for a fixed x_0.

The final result is easy for $\Re(z) <= 0$, because then the $|n^{-z}| = n^{-x}$ do not go to zero. If $0 < \Re(z) \leq 1$, the final result is surprisingly tricky. We write

$$\sum_{n=1}^{N} \frac{1}{n^z} = \sum_{n=1}^{N} \left(\frac{1}{n^z} - \int_n^{n+1} \frac{dt}{t^z}\, dt \right) + \int_1^{N+1} \frac{dt}{t^z}\, dt$$

Notice that

$$\int t^{-z}\, dt = \frac{t^{-z+1}}{-z+1}$$

provided $z \neq 1$. It follows that the final integral is the limit of the expression below as $N \to \infty$:

$$\frac{1}{z-1}\left(1 - \frac{1}{(N+1)^{z-1}}\right)$$

If $z = 1+iy$, then the final fraction oscillates, since $(N+1)^{iy} = e^{iy\ln(N+1)}$. Otherwise $\frac{1}{(N+1)^{z-1}} = |(N+1)^{-x-iy+1}| = |(N+1)^{-x+1}|$ goes to infinity because $0 < -x+1 < 1$.

So the final expression does not have a limit, and it suffices to prove that the middle sum does converge.

Note that

$$\left|\left(\frac{1}{n^z} - \int_n^{n+1} \frac{dt}{t^z}\, dt\right)\right| = \left|\int_n^{n+1} \left(\frac{1}{n^z} - \frac{1}{t^z}\right) dt\right| \leq \int_n^{n+1} \left|\frac{1}{n^z} - \frac{1}{t^z}\right| dt$$

Let $f_1(u)$ and $f_2(u)$ be the real and imaginary parts of $\frac{1}{u^z}$ and apply the mean value theorem to these functions. Then

$$f_1(n) - f_1(t) = \frac{df_1}{du}(\xi_1)(n-t)$$
$$f_2(n) - f_2(t) = \frac{df_2}{du}(\xi_2)(n-t)$$

It follows that the above integral is bounded by

$$\left|\frac{df_1}{du}(\xi_1) + i\frac{df_2}{du}(\xi_2)\right| \leq \left|\frac{df_1}{du}(\xi_1)\right| + \left|\frac{df_2}{du}(\xi_2)\right|$$

If we regard u as complex, then the derivative of $f(u) = u^{-z}$ is $-zu^{-z-1}$. Since $\frac{df}{du} = \frac{df_1}{du} + i\frac{df_2}{du}$, the expression above is less than or equal to

$$\left|\frac{df}{du}(\xi_1)\right| + \left|\frac{df}{du}(\xi_2)\right| = |z|\left|(\xi_1)^{-z-1}\right| + |z|\left|(\xi_1)^{-z-1}\right|$$

Note that $n \leq \xi_i \leq (n+1)$ and

$$\left|(\xi_1)^{-z-1}\right| = \left|e^{-(x+1)\ln\xi_i - iy\ln\xi_i}\right| = \left|e^{-(x+1)\ln\xi_i}\right| \leq \left|e^{-(x+1)\ln n}\right| = \frac{1}{n^{\Re z+1}}$$

So

$$\left|\left(\frac{1}{n^z} - \int_n^{n+1} \frac{dt}{t^z}\, dt\right)\right| \leq \frac{2|z|}{n^{\Re(z)+1}}$$

Since $0 < \Re(z) \leq 1$, the sum of these terms converges. QED.

Remark: In a later chapter, we will study *Dirichlet series* in general. These are series of the form

$$f(z) = \sum_{n=1}^{\infty} \frac{c_n}{n^z}$$

where the c_n are arbitrary complex numbers. We will prove that there is an x_0, called the *abscissas of convergence*, such that the series diverges when $\Re z < x_0$ and converges when $\Re(z) > x_0$. We will also prove that there is an x_1, called the *abscissas of absolute convergence*, such that the series does not converge absolutely if $\Re(z) < x_1$, and converges absolutely if $\Re(z) > x_1$. Obviously $x_0 \leq x_1$. We will prove that the distance between them is at most one. In the case of the Riemann zeta function, these numbers are equal.

17.6 Analytic Continuation of the Zeta Function

Riemann's first great result in his eleven page paper is

Theorem 101. *The Riemann Zeta function can be analytically continued to a function meromorphic in the entire complex plane. This function has a pole of order one at $z = 1$ and otherwise is holomorphic.*

Proof: Riemann begins with the integral formula for the Gamma function:

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

Let n be a positive integer and replace t by nt to get

$$\Gamma(z) = \int_0^{\infty} n^{z-1} t^{z-1} e^{-nt} n dt = n^z \int_0^{\infty} t^{z-1} e^{-nt}\, dt$$

So

$$\frac{1}{n^z}\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-nt}\, dt$$

Sum from $n = 0$ to infinity to get

$$\zeta(z)\Gamma(z) = \int_0^{\infty} t^{z-1} \sum_{n=1}^{\infty} e^{-nt}\, dt = \int_0^{\infty} \frac{t^{z-1}}{e^t - 1}\, dt$$

Now look closely at the integral on the right. The denominator is

$$e^t - 1 = t + \frac{t^2}{2!} + \frac{t^3}{3!} + ...$$

When t is large, $e^t - 1$ will swamp any fixed power of t in the numerator, so the integral converges. But when t is close to zero, the denominator is approximately t and the numerator needs to have a positive power of t to get convergence; for real z, that requires $z - 1 > 0$, i.e., $z > 1$.

Riemann proposed replacing the contour along the x-axis from 0 to ∞ by the first contour below, which we will call γ. The new contour avoids the singularity at the origin by circling it and returning to infinity.

The two journeys along the x-axis do not cancel out because the integral contains $t^{z-1} = e^{(z-1)\mathrm{Log}(t)}$. Until this moment, we set $\mathrm{Log}(t)$ to be the real logarithm on the x-axis. Now we set it to the real logarithm for the first portion of the path, and then extend it continuously for the remaining path. In particular, as we trace along the final portion along the x-axis, we will have $\mathrm{Log}(t) = \ln t + 2\pi i$.

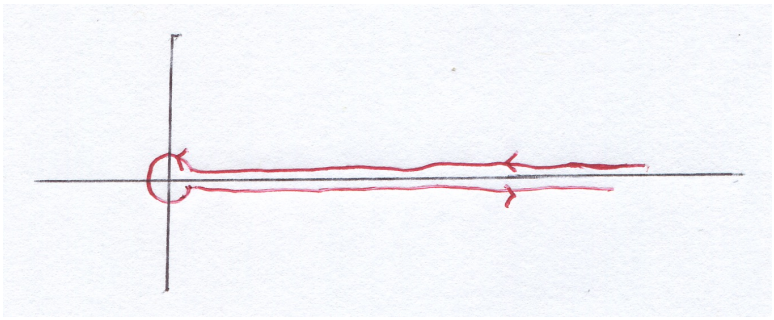


Figure 17.1: New Path for $\zeta(z)$

Figure 17.2: Two Branches of $\mathrm{Log}(z)$

The second illustration above shows how to make this rigorous. The left side shows a branch cut along the negative imaginary axis. Define $\mathrm{Log}(z)$ on the remaining plane so $-\frac{\pi}{2} < \mathrm{Arg}(z) < \frac{3\pi}{2}$; such a branch exists because the remaining plane is simply connected and omits the origin. Similarly define a second branch of the logarithm on the plane minus the positive imaginary axis so $\frac{\pi}{2} < \mathrm{Arg}(z) > \frac{5\pi}{2}$. Notice that these branches agree on the left half plane.

Our actual path lies exactly on the positive x-axis, coming from infinity to close to the origin.

This integral uses the left branch of $\mathrm{Log}(z)$ to form t^{z-1}. Then it begins circling the origin counterclockwise. At some point when this path is in the left half of the plane, we jump to the second branch of $Log(z)$; since the two branches agree on the left half, the exact point we switch makes no difference. We continue using this branch to compute t^{z-1} are we continue circling the origin and then follow the positive x axis from close to the origin on out to infinity.

If we start and end a finite distance from the origin, this integral defines a holomorphic function of z by results of the section on differentiating under the integral sign. Taking the limit as the start and end approach infinity gives an integral with a limit, and the limit is uniform on compact subsets of the z-plane. So in the end, the integral defines a holomorphic function of z.

Let us compute the value of the integral when the real part of z is greater than 1. The integral from infinity to almost zero along the positive x axis is almost

$$-\int_0^\infty \frac{t^{z-1}}{e^t-1}\,dt = -\zeta(z)\Gamma(z)$$

Next we integrate around the small circle. Since $\Re(z)-1>0$, the integrand is bounded near the origin, and the length of the small circle goes to zero, so this part of the integral becomes irrelevant.

Finally we integrate back out along the x-axis. This time $t^{z-1} = e^{(z-1)(\ln t + 2\pi i)}$ and so the integral is

$$\int_0^\infty \frac{t^{z-1}e^{2\pi i(z-1)}}{e^t-1}\,dt = e^{2\pi iz}\zeta(z)\Gamma(z)$$

We conclude that

$$\zeta(z)\Gamma(z)\left(e^{2\pi iz}-1\right) = \int_\gamma \frac{t^{z-1}}{e^t-1}\,dt$$

Every term of this equation is defined and meromorphic on the entire complex plane except $\zeta(z)$, so this equation defines a meromorphic continuation of $\zeta(z)$ to the entire plane.

Notice that $e^{2\pi iz}-1 = e^{\pi iz}2i\left(\frac{e^{\pi iz}-e^{-\pi iz}}{2i}\right) = 2ie^{\pi iz}\sin(\pi z)$. So our equation can be written

$$\zeta(z) = \frac{e^{-\pi iz}}{\Gamma(z)2i\sin(\pi z)}\int_\gamma \frac{t^{z-1}}{e^t-1}\,dt$$

We have proved the important

Theorem 102. *The following formula is universally valid for all arguments of the zeta function, and defines a meromorphic extension to the entire plane:*

$$\zeta(z) = \frac{e^{-\pi iz}}{\Gamma(z)2i\sin(\pi z)}\int_\gamma \frac{t^{z-1}}{e^t-1}\,dt$$

Remark: We now wish to locate the poles of ζ and compute the residue at such poles.

Since $e^{-\pi iz}$ and $\int_\gamma \frac{t^{z-1}}{e^t-1}\,dt$ are holomorphic, and since $\Gamma(z)$ has simple poles at $z = 0, -1, -2, ...$ and no zeros, and since $\sin(\pi z)$ has simple zeros at $z = 0, \pm 1, \pm 2, ...$ and no other zeros, the only singularities are at the integers. However, the zeros and poles cancel to give finite values at $0, -1, -2, ...$

The simple zeros of $\sin(\pi z)$ do not cancel for $n = 1, 2, 3, ...$ unless the integral vanishes. However, $\zeta(z)$ is holomorphic and non-zero for $\Re(z) > 1$ by the product formula, so the integral must indeed vanish at $n = 2, 3, 4, ...$ It follows that the only pole of $\zeta(z)$ is at $z = 1$.

We easily compute the residue at this pole. First consider the integral. Since $z = 1$, this integral is $\int_\gamma \frac{1}{e^t-1}\,dt$. The integrand is not multiple-valued, so the integrals along the x axis cancel and we integrate need only integrate counterclockwise around a circle about the origin. This gives

$$2\pi i \operatorname{Res}_0\left(\frac{1}{e^t-1}\right)$$

But

$$\frac{1}{e^t-1} = \frac{1}{t+\frac{t^2}{2}+...} = \frac{1}{t(1+\frac{t}{2}+...)} = \frac{1}{t}(1+...)$$

So the residue is 1 and the value of the integral is $2\pi i$.

Moreover $\sin(\pi z) = \sin(\pi)+\pi\cos(\pi)(z-1)+... = -\pi(z-1)+....$ So near $z = 1$ we have

$$\zeta(z) = \frac{e^{-\pi iz}}{\Gamma(z)2i\sin(\pi z)}\int_\gamma \frac{t^{z-1}}{e^t-1}\,dt = \frac{e^{\pi i}}{2i(-\pi(z-1)+...}2\pi i = \frac{1}{(z-1)+...} = \frac{1}{z-1}+...$$

We have proved

Theorem 103. *The zeta function is holomorphic on the entire plane except for a simple pole at* $z = 1$. *The residue of this pole is* 1.

17.7 The Functional Equation of the Zeta Function

The second great result in Riemann's paper is :

Theorem 104. *The Riemann Zeta Function satisfies the following equation:*

$$\zeta(z) = 2^z\pi^{z-1}\sin\left(\frac{\pi z}{2}\right)\Gamma(1-z)\zeta(1-z)$$

Remark: This theorem says that if we rotate the plane by 180 degrees about the point $z = \frac{1}{2}$, then ζ is unchanged modulo known factors.

However, ζ has a second symmetry. If f is any entire function, the function $g(z) = \overline{f(\overline{z})}$ is easily proved holomorphic by the Cauchy-Riemann equations. If f is real on a non-trivial interval in the real line, then these functions agree on that interval, and consequently agree everywhere by the identity theorem. So $f(\overline{z}) = \overline{f(z)}$.

This easy result obviously applies to the zeta function. So ζ is invariant modulo known factors under reflection across the x-axis and rotation by 180 degrees about $z = \frac{1}{2}$. These operations generate the dihedral group D_2, containing reflections across the lines $y = 0$ and $x = \frac{1}{2}$, and rotations about $\frac{1}{2}$ by 0 and 180 degrees.

Proof: We are going to deform the contour for the integral which analytically continues the zeta function. We will enlarge the circle about the origin until it becomes a very large square, with "tail pieces" along the x-axis that are so far out that they are irrelevant. We need large squares rather than large circles for reasons that will appear at the end of the proof.

The function $\frac{t^{z-1}}{e^t-1}$ has singularities along the imaginary axis at points $z = 2\pi in$ where $e^z = 1$. Thus we must use the Residue theorem to deform across these singularities. This will give a sum of terms which will eventually give $\zeta(1-z)$. We will choose our squares so that the top and bottom hit the imaginary axis halfway between adjacent singularities, and so at $\pi i(2n+1)$.

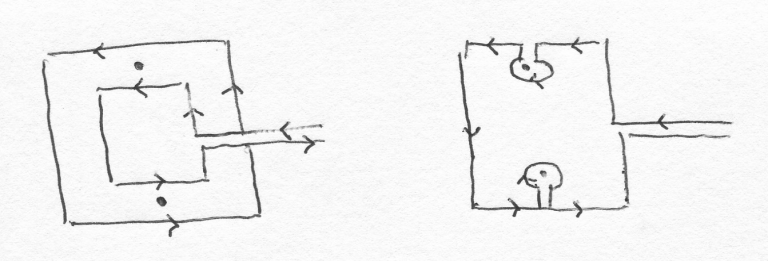


Figure 17.3: Deforming the ζ contour

Here are the details. The poles of $\frac{t^{z-1}}{e^t-1}$ occur at $\pm 2\pi in$ for integers n. Suppose we enlarge a square so it crosses one more singularity at both top and bottom. Let these be at $\pm 2\pi in$. The pictures above show the deformation, and show that we cross the poles by adding a counter-clockwise circle at the top and bottom. So the new integral is the old integral minus

$$2\pi i \left(\text{Res}_{2\pi in} \frac{t^{z-1}}{e^t - 1} + \text{Res}_{-2\pi in} \frac{t^{z-1}}{e^t - 1} \right)$$

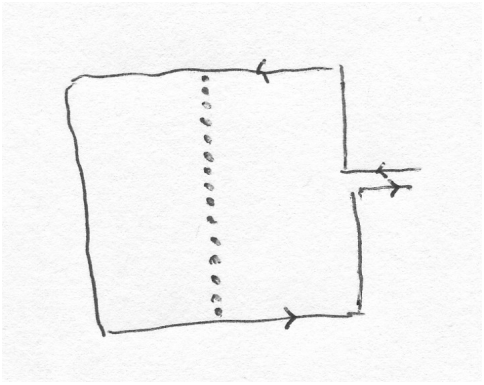


Figure 17.4: Limiting Case of Homotopy

Expanding $f(t) = e^t - 1$ about $2\pi in$ gives

$$f(2\pi in) + \frac{df}{dz}(2\pi in)(t - 2\pi in) + ... = (t - 2\pi in) + ...$$

so

$$\frac{1}{e^t - 1} = \frac{1}{(t - 2\pi in)(1 + ...)} = \frac{1}{t - 2\pi in}(1 + ...)$$

and

$$\frac{t^{z-1}}{e^t - 1} = \frac{e^{(z-1)\mathrm{Log}(2\pi in)}}{t - 2\pi in} + ... = \frac{e^{(z-1)\left(\ln n + \ln 2\pi + \frac{\pi i}{2}\right)}}{t - 2\pi in} + ...$$

This is the expansion at the top. Replace $\frac{\pi i}{2}$ by $\frac{3\pi i}{2}$ to get the corresponding expansion at the bottom.

Putting this together, the addditional factor as we enlarge the square past the two singularities in question is

$$-2\pi i\left(n^{z-1}(2\pi)^{z-1}e^{(z-1)\frac{\pi i}{2}} + n^{z-1}(2\pi)^{z-1}e^{(z-1)\frac{3\pi i}{2}}\right) = -2\pi i\, n^{z-1}(2\pi)^{z-1}\left(-e^{\frac{\pi iz}{2}}i + e^{\frac{3\pi iz}{2}}i\right)$$

This expression equals

$$2\pi\, n^{z-1}(2\pi)^{z-1}e^{\pi iz}\left(e^{\frac{\pi iz}{2}} - e^{\frac{-\pi iz}{2}}\right) = n^{z-1}(2\pi)^z e^{\pi iz}(2i)\sin\frac{\pi z}{2}$$

At the end of the deformation, we have

$$\int_\gamma \frac{t^{z-1}}{e^t - 1}\,dt = (2i)(2\pi)^z e^{\pi iz}\sin\frac{\pi z}{2}\sum_{n=1}^{N}\frac{1}{n^{1-z}} + \int\frac{t^{z-1}}{e^t - 1}\,dt$$

where the integral on the right is over a very large square, together with two tail lines along the x-axis.

Now suppose that $\Re(z) < 0$ so $\Re(1-z) > 1$. Then $\sum_{n=1}^{N} \frac{1}{n^{1-z}}$ converges to $\zeta(1-z)$. We claim that the integral over the large square and the remaining tail pieces also goes to zero as the square gets larger. Since this step is somewhat tricky, we leave it to last.

Then in the limit as the radius of the square goes to infinity we have

$$\zeta(z) = \frac{e^{-\pi i z}}{\Gamma(z) 2i \sin(\pi z)} \int_\gamma \frac{t^{z-1}}{e^t - 1}\, dt = \frac{e^{-\pi i z}}{\Gamma(z) 2i \sin(\pi z)} (2i)\,(2\pi)^z\, e^{\pi i z}\, \sin\frac{\pi z}{2}\, \zeta(1-z)$$

$$= \frac{(2\pi)^z\, \sin\frac{\pi z}{2}}{\Gamma(z)\sin(\pi z)}\, \zeta(1-z)$$

Using the identity

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

or equivalently

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

the previous equation becomes

$$\zeta(z) = \frac{(2\pi)^z \sin\frac{\pi z}{2}}{\pi} \Gamma(1-z)\zeta(1-z) = 2^z \pi^{z-1}\, \sin\left(\frac{\pi z}{2}\right)\, \Gamma(1-z)\zeta(1-z)$$

Therefore, Riemann's functional equation is true for $\Re(z) < 0$. The identity theorem then implies that it is always true.

Since the integral analytically continuing the zeta function converges, the tail pieces along the x-axis go to zero for large squares. So to finish the proof, we must show that when $\Re(z) < 0$, the integral over the large square goes to zero as the sides of the square go to infinity.

If a and b are complex numbers, $|a| = |b + (a-b)| \le |b| + |a-b|$ and so $|a-b| \ge |a| - |b|$. Interchanging a and b, we deduce that

$$|a-b| \ge |\,|a| - |b|\,|$$

Consequently

$$|e^t - 1| \ge |\,|e^t| - 1\,|$$

Moreover,

$$t^{z-1} = e^{(x-1+iy)(\ln|t|+i\theta)} = |t|^{x-1} e^{-y\theta} A$$

where A has absolute value one.

Putting this together,

$$\left|\frac{t^{z-1}}{e^t-1}\right| \leq \frac{|t|^{\Re(z)-1}e^{-y\theta}}{||e^t|-1|}$$

In this formula, y is the imaginary part of z and $0 \leq \theta \leq 2\pi$. Since we are integrating over t, y is constant and θ is bounded, so $e^{-y\theta}$ is bounded by a constant M that does not depend on the square. Hence we can ignore it.

The corners of our square are at $(\pm\pi(2n+1), \pm\pi i(2n+1))$ and the lengths of the sides are $2\pi(2n+1)$.

Look first at the vertical side on the right. Since $\Re(z) < 0$, the term $|t|^{\Re(z)-1}$ is at most $\pi(2n+1)$ to a negative power and so smaller than 1. On the other hand, the denominator is $e^{\pi(n+1}-1$, which goes to infinity very rapidily. Even when we multiply by the length of the side, $2\pi(2n+1)$, this term goes to zero.

Look next at the vertical side on the left. This time, $|e^t| = e^{-\pi(n+1)}$ goes to zero as the square gets large, so the denominator is close to 1. However, the numerator is essentially $|t|^{\Re(z)-1}$. Since $\Re(z) < 0$, the exponent is smaller than -1. So this term is smaller than

$$\frac{1}{(\pi(2n+1))^{1-\Re(z)}}$$

We must multiply by the length of the side, $2\pi(2n+1)$, getting

$$\frac{2}{(\pi(2n+1))^{-\Re(z)}}$$

Since $\Re(z) < 0$, this goes to zero as n approaches infinity.

Finally, we must study the top and bottom of the square. Up to a constant, the integrand is bounded by

$$\frac{|t|^{\Re(z)-1}}{||e^t|-1|}$$

Notice that $|t|$ is at least $|\Im(t)|$; since $\Re(z) - 1 < 0$, the numerator is at most $(\Im(t))^{\Re(z)-1}$. The length of the top and bottom is $2|\Im(t)|$. So if we ignore the denominator, the maximum of the number times the length of the curve is $2|\Im(t)|^{\Re(z)}$ and this goes to zero because $\Re(z) < 0$.

To complete the argument, we now prove that the denominator is bounded below by a positive bound which is independent of the chosen square. Notice that $||e^t|-1|$ is periodic with period $2\pi i$. So this function is the same on the top and bottom of all squares, except that it is cut off on the left and right. Ignoring the cutoff, the function goes to infinity on the right, and goes to 1 on the left. It is never zero because it the squares were chosen to avoid zeros of e^z. Hence it is bounded below, either by 1 or by an interior minimum. QED.

Remark: Riemann later rewrote the functional equation for greater symmetry. The following result defines a function which is essentially the zeta function, but which is unchanged when z is changed to $1-z$:

Theorem 105.

$$\pi^{-z/2}\Gamma\left(\frac{z}{2}\right)\zeta(z) = \pi^{-(1-z)/2}\Gamma\left(\frac{1-z}{2}\right)\zeta(1-z)$$

Proof: Apply the previous functional equation to obtain

$$\pi^{-z/2}\Gamma\left(\frac{z}{2}\right)\zeta(z) = \pi^{-z/2}\Gamma\left(\frac{z}{2}\right)2^z\pi^{z-1}\sin\left(\frac{\pi z}{2}\right)\Gamma(1-z)\zeta(1-z)$$

Substitute $\frac{z}{2}$ for z in the formula $\Gamma(z)\Gamma(1-z) = \frac{\pi}{\sin(\pi z)}$ to obtain

$$\Gamma\left(\frac{z}{2}\right)\sin\left(\frac{\pi z}{2}\right) = \frac{\pi}{\Gamma\left(1-\frac{z}{2}\right)}$$

and substitute this in the right side of the previous formula to obtain

$$\frac{\pi^{z/2}2^z\Gamma(1-z)}{\Gamma\left(1-\frac{z}{2}\right)}\zeta(1-z)$$

Substitute $\frac{1-z}{2}$ for z in the formula $\Gamma(z)\Gamma\left(z+\frac{1}{2}\right) = 2^{1-2z}\sqrt{\pi}\,\Gamma(2z)$ to get

$$\Gamma\left(\frac{1-z}{2}\right)\Gamma\left(1-\frac{z}{2}\right) = 2^z\sqrt{\pi}\,\Gamma(1-z)$$

and substitute this in the previous formula to obtain

$$\frac{\pi^{z/2}2^z\Gamma\left(\frac{1-z}{2}\right)}{2^z\sqrt{\pi}}\zeta(1-z) = \pi^{(z-1)/2}\ \Gamma\left(\frac{1-z}{2}\right)\ \zeta(1-z)$$

Remark: Notice that the poles of $\Gamma\left(\frac{z}{2}\right)$ at $0, -2, -4, ...$ cancel the zeros of $\zeta(z)$ at $-2, -4, ...$ So we end up with a meromorphic function with poles at $z = 0$ and $z = 1$ and zeros exactly at the non-trivial zeros of the zeta function in the critical strip. Riemann then made a slight modification:

Theorem 106. *The following function is entire and unchanged when z is replaced by $1-z$. Its zeros are exactly the non-trivial zeros of the zeta function in the critical strip:*

$$z(1-z)\,\pi^{-z/2}\Gamma\left(\frac{z}{2}\right)\zeta(z)$$

17.8 Bernoulli Numbers

Definition 39. *The* Bernoulli Numbers B_n *are defined by the expansion*

$$\frac{t}{e^t-1}=\sum_{n=0}^{\infty}\frac{B_n}{n!}\,t^n$$

Remark: It is easy to deduce a recursive formula for the B_n. We have

$$t=\left(\sum_{k\geq 0}\frac{B_k}{k!}\,t^k\right)(e^t-1)=\left(\sum_{k\geq 0}\frac{B_k}{k!}\,t^k\right)\left(\sum_{l\geq 1}\frac{1}{l!}\,t^l\right)=t\left(\sum_{k\geq 0}\frac{B_k}{k!}\,t^k\right)\left(\sum_{l\geq 0}\frac{1}{(l+1)!}t^l\right)$$

and therefore

$$1=\sum\left(\frac{B_k}{k!(l+1)!}\right)t^{k+l}$$

So

$$1=\frac{B_0}{0!1!}\qquad\text{or}\qquad B_0=1$$

Also

$$0=\frac{B_0}{0!(n+1)!}+\frac{B_1}{1!n!}+\ldots+\frac{B_n}{n!1!}$$

and so

$$B_n=-\left(\frac{B_0}{n+1}+B_1+\frac{nB_2}{2!}+\frac{n(n-1)B_3}{3!}+\ldots+nB_{n-1}\right)$$

Using this formula and Mathematica, we obtain

$$B_0=1,B_1=-\frac{1}{2},B_2=\frac{1}{6},B_3=0,B_4=-\frac{1}{30},B_5=0,B_6=\frac{1}{42},B_7=0$$

$$B_8=-\frac{1}{30},B_9=0,B_{10}=\frac{5}{66},B_{11}=0,B_{12}=-\frac{691}{2730},B_{13}=0,B_{14}=\frac{7}{6}$$

Theorem 107. *If n is odd and $n>1$, $B_n=0$.*

Proof: We have

$$\frac{t}{e^t-1}+\frac{t}{2}=t\left(\frac{1}{e^t-1}+\frac{1}{2}\right)=t\left(\frac{2+e^t-1}{(e^t-1)2}\right)$$

$$=\frac{t}{2}\left(\frac{e^t+1}{e^t-1}\right)=\frac{t}{2}\left(\frac{e^{\frac{t}{2}}+e^{-\frac{t}{2}}}{e^{\frac{t}{2}}-e^{-\frac{t}{2}}}\right)$$

If we change the sign of t in the last expression, nothing changes, so the expansion involves only even powers of t.

Warning: Since all odd Bernoulli numbers except B_1 are zero and the even Bernoulli numbers alternate sign, it is common to find alternate definitions; a common alternative defines B_n as our $(-1)^{n+1}B_{2n}$.

Remark: Bernoulli numbers occur in surprising places. The series expansion of the tangent has them, numerical approximations for integrals contain them, and we will see shortly that they occur in formulas for the value of the zeta function at integers. The numbers occur in number theory and algebraic topology. For the latter, see Hirzebruch's em New Topological Methods in Algebraic Geometry. Kummer proved that Fermat's last theorem is true for a prime exponent p if p does not divide the numerators of any of $B_2, B_4, \ldots, B_{p-3}$.

Remark: Often Bernoulli numbers are hidden away. One dead giveaway that they are present is the occurrence of the unusual number 691 in B_{12}. If a calculation yields a rational with numerator 691, Bernoulli numbers are there somewhere.

Bernoulli discovered these numbers in an unusual manner. In beginning integration theory, it is useful to have formulas for sums of powers of integers. For instance, $\sum_1^n k = \frac{n(n+1)}{2}$ and $\sum_1^n k^2 = \frac{k(k+1)(2k+1)}{6}$. At the end of his book *Ars Conjectandi*, published in 1713 after his death,

Bernoulli listed the first ten formulas in a table.

$$
\begin{array}{lllllll}
\frac{n^2}{2} & +\frac{n}{2} & & & & & \\
\frac{n^3}{3} & +\frac{n^2}{2} & +\frac{n}{6} & & & & \\
\frac{n^4}{4} & +\frac{n^3}{2} & +\frac{n^2}{4} & & & & \\
\frac{n^5}{5} & +\frac{n^4}{2} & +\frac{n^3}{3} & -\frac{n}{30} & & & \\
\frac{n^6}{6} & +\frac{n^5}{2} & +\frac{5n^4}{12} & -\frac{n^2}{12} & & & \\
\frac{n^7}{7} & +\frac{n^6}{2} & +\frac{n^5}{2} & -\frac{n^3}{6} & +\frac{n}{42} & & \\
\frac{n^8}{8} & +\frac{n^7}{2} & +\frac{7n^6}{12} & -\frac{7n^4}{24} & +\frac{n^2}{12} & & \\
\frac{n^9}{9} & +\frac{n^8}{2} & +\frac{2n^7}{3} & -\frac{7n^5}{15} & +\frac{2n^3}{9} & -\frac{n}{30} & \\
\frac{n^{10}}{10} & +\frac{n^9}{2} & +/\frac{3n^8}{4} & -\frac{7n^6}{10} & +\frac{n^4}{2} & -\frac{3n^2}{20} & \\
\frac{n^{11}}{11} & +\frac{n^{10}}{2} & +\frac{5n^9}{6} & -n^7 & +n^5 & -\frac{n^3}{2} & +\frac{5n}{66}
\end{array}
$$

After displaying this table, Bernoulli wrote "Whoever will examine the series as to their regularity may be able to continue the table." He then writes a general formula without further explanation.

The first two columns have clear patterns. The patterns are more difficult to determine in the remaining columns; can you find them? However, coming down the right diagonal, notice the following multiples of n: $\frac{1}{6}, 0, -\frac{1}{30}, 0, \frac{1}{42}, 0, -\frac{1}{30}, 0, \frac{5}{66}$. These are the Bernoulli numbers, making their first appearance in mathematics.

You might argue that we skipped the first entry, $\frac{1}{2}$, and for the reason that the first Bernoulli number is $-\frac{1}{2}$. This number doesn't quite fit the pattern because all other Bernoulli numbers with odd indices are zero. Indeed, some tables list $B_1 = \pm\frac{1}{2}$. We'll stick with the value $-\frac{1}{2}$ given by our generating function.

The connection between Bernoulli numbers and $\frac{t}{e^t-1}$ was first discovered by Euler.

17.9 The Value of the Zeta Function on Integers

Consider the integral

$$\int_\gamma \frac{t^{z-1}}{e^t-1}\,dt$$

used to analytically continue the zeta functions. If n is an integer, the term t^{z-1} in the formula becomes single-valued, and the integral can be evaluated by the Residue theorem. So we are close to evaluating $\zeta(n)$ for all integers n. But some nasty surprises lie ahead.

We have

$$\int \frac{t^{n-1}}{e^t-1} = \int \sum_{k\geq 0} \frac{B_k}{k!} t^{k+n-2} = 2\pi i \operatorname{Res}\left(\sum_{k\geq 0} \frac{B_k}{k!} t^{k+n-2}\right)$$

By the Residue theorem, the crucial term occurs when $k+n-2=-1$, i.e., $k=1-n$. In that case the integral is $2\pi i \frac{B_{1-n}}{(1-n)!}$.

Since the k in the sum are non-negative, the calculation shows that the integral is zero if $n>1$. This seems impossible since for positive n, $\zeta(n)=\sum \frac{1}{k^n} > 0$. But remember that

$$\zeta(n) = \frac{e^{-\pi i n}}{\Gamma(n) 2i \sin(\pi n)} \int_\gamma \frac{t^{z-1}}{e^t-1}\,dt$$

If $n>1$, both the integral and the term $\sin(\pi n)$ in the denominator vanish, and we have to compute $\zeta(n)$ by taking a limit as $z\to n$. But that requires knowing the contour integral at non-integer values when we can no longer use the Residue theorem. So we are temporarily stuck.

When $n=1$, we can expand $\sin(\pi z) = \sin(\pi) + \pi\cos(\pi)(z-1) + ... = -\pi(z-1) + ...$, so that near $n-1$ we have

$$\zeta(z) = \frac{-1}{2\pi(-\pi(z-1)+...)} 2\pi i \frac{B_0}{0!} = \frac{1}{z-1} + ...$$

This confirms an earlier result that $z=1$ is a pole of order one with residue 1.

From now on, consider the case $n\leq 0$. The formula for ζ has $\Gamma(z)\sin(\pi z)$ in the denominator. Since $\Gamma(z)\Gamma(1-z) = \frac{\pi}{\sin(\pi z)}$, we can write

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

So

$$\zeta(n) = \frac{e^{-\pi i n}\Gamma(1-n)}{2\pi i} 2\pi i \frac{B_{1-n}}{(1-n)!}$$

For psychological reasons only, it is convenient to replace each n with $-n$. Thus for $n \geq 0$ we have

$$\zeta(-n) = \frac{(-1)^n n! B_{n+1}}{(n+1)!} = \frac{(-1)^n B_{n+1}}{n+1}$$

Theorem 108. *For $n \geq 0$ we have*

$$\zeta(-n) = \frac{(-1)^n B_{n+1}}{n+1}$$

In particular, $\zeta(0) = -\frac{1}{2}$, $\zeta(-1) = -\frac{1}{12}$, $\zeta(-3) = \frac{1}{120}$, $\zeta(-5) = -\frac{1}{252}$, $\zeta(-7) = \frac{1}{240}$, ...

Theorem 109. *The zeta function has zeros at* $-2, -4, -6, -8, ...$

Remark: We now use the functional equation of the zeta function:

$$\zeta(z) = 2^z \pi^{z-1} \sin\left(\frac{\pi z}{2}\right) \Gamma(1-z)\zeta(1-z)$$

Apply this when $1 - z = 2n$ and $n \geq 1$. Thus $z = 1 - 2n$ and we get

$$\zeta(1-2n) = \frac{(-1)^{2n-1} B_{2n}}{2n} = 2^{1-2n}\pi^{1-2n-1} \sin\left(\frac{\pi(1-2n)}{2}\right) \Gamma(2n)\zeta(2n)$$

and so

Theorem 110. *If $n \geq 1$,*

$$\zeta(2n) = 2^{2n-1}\pi^{2n} \frac{(-1)^{n+1} B_{2n}}{(2n)!}$$

Examples: $\zeta(2) = \frac{\pi^2}{6}, \zeta(4) = \frac{\pi^4}{90}, \zeta(6) = \frac{\pi^6}{945}, \zeta(8) = \frac{\pi^8}{9450}, \zeta(10) = \frac{\pi^{10}}{93555}, \zeta(12) = \frac{691\pi^{12}}{638512875}$.

Remark: Explicit values for $\zeta(2n+1)$ with $n \geq 1$ are unknown.

Corollary 27. *Starting with $B_2 = \frac{1}{6}$, the signs of Bernoulli numbers with even indices alternate.*

Corollary 28. *For large n,*

$$|B_{2n}| \sim \frac{2(2n)!}{(2\pi)^{2n}}$$

and in particular, $|B_{2n}| \to \infty$.

Proof: Clearly $\zeta(2n) \to 1$

Remark: The Functional Equation trick fails for $\zeta(2n+1)$, but still provides a new piece of information:

Theorem 111. *The zeros of $\zeta(z)$ at $-2, -4, -6, \ldots$ are all simple zeros.*

Proof: By the functional equation, if $n > 0$ then

$$\zeta(-2n) = 2^{-2n}\pi^{-2n-1}\sin\left(\frac{-2\pi n}{2}\right)\Gamma(2n+1)\zeta(2n+1)$$

Every term on the right is nonzero except $\sin(-\pi n)$ and this term has a simple zero. So the zero $\zeta(-2n)$ is simple.

Remark: Finally, we have the important

Theorem 112. *The Riemann zeta function has simple zeros are $-2, -4, -6, \ldots$ All other zeros lie in the strip $0 \leq \Re(z) \leq 1$. The zeros in this strip occur symmetrically. If there is a zero at $x + iy$ with $x = \frac{1}{2}$, there is a zero of the same order at $x - iy$. If there is a zero at $x + iy$ with $x \neq \frac{1}{2}$, there are zeros of the same order at $(1 - x) + iy$, $(1 - x) - iy$, and $x - iy$.*

Proof: There are no zeros with $\Re(z) > 1$ by the product formula, since no term in this formula vanishes.

By the Functional Equation,

$$\zeta(z) = 2^z\pi^{z-1}\sin\left(\frac{\pi z}{2}\right)\Gamma(1-z)\zeta(1-z)$$

The only zeros of $\sin\left(\frac{\pi z}{2}\right)$ occur at integers; the gamma function has no zeros and its poles occur at integers. But we know all possible zeros of the zeta function at integers. Otherwise, $\zeta(z)$ and $\zeta(1-z)$ has the same zeros. Moreover, ζ is real on the x-axis, so $\zeta(\overline{z}) = \overline{\zeta(z)}$.

Remark: Surprisingly, the value $\zeta(-1) = -\frac{1}{12}$ has come up in string theory. There is a UTube video about it at https://www.youtube.com/watch?v=w-I6XTVZXww by the physicist Tony Padillo. His calculation goes like this:

Taking the average of the partial sums, we conclude that $1-1+1-1+\ldots = \frac{1}{2}$. Next define

$$S = 1 - 2 + 3 - 4 + \ldots$$

Shift S right by one and add to deduce that $S = \frac{1}{4}$:

$$2S = 1 + (-2+1) + (3-2) + (-4+3) + \ldots = 1 - 1 + 1 - 1 + \ldots = \frac{1}{2}$$

Finally, let $T = 1 + 2 + 3 + 4 + \ldots$ Then $T = -\frac{1}{12}$ because

$$T - S = 0 + 4 + 0 + 8 + \ldots = 4(1 + 2 + 3 + 4 + \ldots = 4T$$

Padillo points out that the physicists have computed various physical quantities using this result, and experiments show that their results are correct.

Ramanujan also considered this series. In his first letter to Hardy, he sent a list of formulas without proofs. Replying, Hardy no doubt asked for methods of calculation. In his second letter to Hardy, Ramanujan stated that his methods required too much room to state in a letter. He said "I am very much gratified on perusing your letter of the 8th of February. I was expecting a reply from you similar to the one which a Mathematics Professor at London wrote asking me to study carefully Bromwich's Infinite Series and not fall into the pitfalls of divergent series... I told him that the sum of an infinite number of terms of the series $1 + 2 + 3 + 4 + ... =$ - 1/12 under my theory. If I tell you this you will at once point out to me the lunatic asylum as my goal. I dilate on this simply to convince you that you will not be able to follow my methods of proof if I indicate the lines in a single letter."

Hardy, no doubt, knew how to justify the $-\frac{1}{12}$.

17.10 Remarks on the Prime Number Theorem

The prime number theorem can be stated in several equivalent ways, but the most illuminating is

Theorem 113 (Prime Number MetaTheorem)**.** *The probability that a positive integer n is prime is*

$$\frac{1}{\ln n}$$

I once had a student who taught me how to think about logarithms. He said "you mathematics always draw the wrong graph for the logarithm; the correct graph is this one." Then he draw our graph on the top, and the correct graph below that.

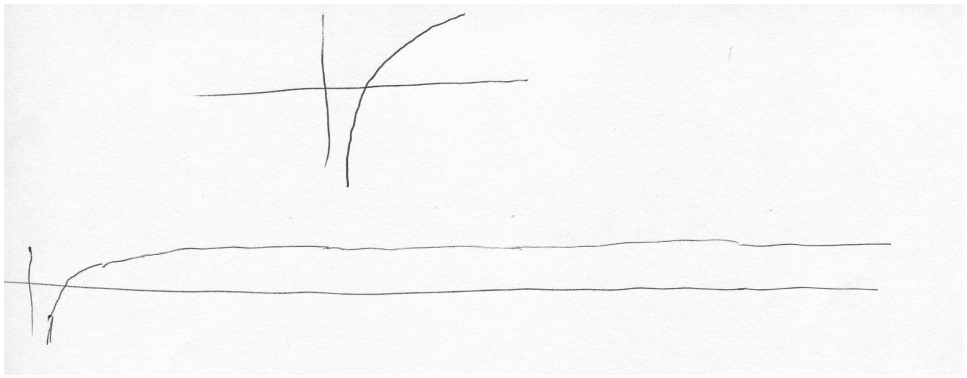


Figure 17.5: Correct Graph of $\ln x$

Indeed, the most crucial fact about the logarithm is that it is flat, and gets *incredibly flat* for large x, and yet manages to grow as large as you wish.

Thus the prime number theorem says that if you look at a table of numbers, the primes will occur randomly at a more or less constant rate. Only very large tables will suggest that their density gradually decreases.

As an example, the RSA encryption scheme used to encode internet transactions requires picking two large prime numbers, advertising their product, but keeping their individual values secret. These primes should have about 100 digits. The recipe for finding such primes is to "guess and check." As it turns out, fast algorithms exist to distinguish primes from nonprimes (for example, most nonprimes do not satisfy Fermat's little theorem). But how many guesses are needed? Well

$$\ln 10^{100} = 100 \ln 10 = 230$$

So a prime should show up after around 230 calculations. However, it is best to avoid guessing even numbers, and then only around 115 guesses are needed.

Compare this with a similar search for perfect squares. A perfect square n^2 with 100 digits occurs when $n \sim 10^{50}$, and the distance between consecutive squares is $(n+1)^2 - n^2 \sim 2n = 2 \cdot 10^{50}$. So it would take around 10^{50} guesses to guess a perfect square. According to Google, there are about 7.5×10^{18} grains of sand on all of the world's beaches, so it is far, far easier to search for one particular grain of sand than to search for a perfect square with 100 digits.

Remark: As stated, the prime number theorem makes no sense. Probability is about repeatable events which can have different outcomes. The probability of getting heads when tossing a quarter is $\frac{1}{2}$ because if you toss a thousand times, about half of the tosses yield heads. But every time you check whether 101 is prime, the answer is "yes."

The real meaning of the prime number theorem is that primes satisfy various statistical tests which would also be satisfied by randomly choosing integers n with probability $\frac{1}{\ln n}$. For example, suppose we count the number of primes less than a million. The number we get is also roughly the number we get by randomly choosing integers according to the $\frac{1}{\ln n}$ rule and counting the number of selected values. This is an experiment we can perform over and over, each time with a different result. But the results cluster around the number of primes less than one million.

It is remarkable that the primes satisfy so many precise rules, and yet behave in this statistical manner. The 20th century mathematician Paul Erdos had a wonderful way of expressing this discovery. He once said "God does not play dice with the universe, but something very strange is happening with the prime numbers."

Notation: We now introduce some notation to deal with counting the primes below a given bound. This is the statistical test we will discuss in this chapter.

Definition 40. *Given a positive real number x, we define*

- $\pi(x) =$ *the number of primes less than or equal to* x
- $li(x) = \int_2^x \frac{1}{\ln t}\, dt$
- $\lambda(x) = \frac{x}{\ln x}$

Here $\pi(x)$ is the counting function we would like to approximate, and $\mathrm{li}(x)$ is the expected value of this function if the probability that n is prime is $\frac{1}{\ln n}$. Actually the expected value is $\sum_{n=2}^{\lfloor x \rfloor} \frac{1}{\ln n}$, but the sum and integral are essentially equal since $\ln x$ is so flat.

Unfortunately, $\int \frac{1}{\ln x}\, dx$ cannot be explicitly integrated. It is easy to get approximate values, as we shall see. The first approximation can be obtained by observing that $\ln t$ is flat and thus essentially equal to its ending value over most of the domain of integration. This gives the approximate value $\lambda(x) = \frac{x}{\ln x}$.

The symbols π and li just introduced are commonly used throughout number theory. The function $\frac{x}{\ln x}$ is seldom given a name.

Measuring approximations: Our goal in this chapter is to prove the prime number theorem, which asserts that $\pi(x)$ is approximately $\psi(x)$. We will easily prove that $\psi(x)$ is approximately $\lambda(x)$, so either approximation can be used.

It is important to define "approximately equal". The prime number theorem is based on

Definition 41. *Let $f(x)$ and $g(x)$ be real-valued functions defined on the positive real numbers. Suppose $\lim_{x\to\infty} f(x) = \infty$ and $\lim_{x\to\infty} g(x) = \infty$. We say f and g are* approximately equal *and write $f(x) \sim g(x)$ if*

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1$$

Remark: Given f and an approximation g, the approximation error is $e(x) = |f(x) - g(x)|$. If $f \sim g$, this error can easily go to infinity; the relation $f \sim g$ is easily shown equivalent to $\frac{e(x)}{f(x)} \to 0$. So $f \sim g$ means that the error is smaller and smaller as a fraction of f.

Imagine that a hand calculator displays answers in exponential notation and has space for only five significant digits. This calculator would show $\pi(10^{10}) = .45505 \times 10^9$ and $\lambda(10^{10}) = .43429 \times 10^9$. According to the prime number theorem, eventually the calculator will show all five digits correctly.

Incidentally, the actual value of $\pi(10^{10})$ is 455, 052, 511. The error made by using $\frac{x}{\ln x}$ is 20, 758, 029. The error made by using li(x) is 3,104. We will have more to say later about the vastly superior approximation afforded by li.

Theorem 114. *We have $lix \sim \frac{x}{\ln x}$. Therefore the following statements are equivalent:*

$$\pi(x) \sim li(x) = \int_2^x \frac{1}{\ln t}\, dt \qquad \text{and} \qquad \pi(x) \sim \frac{x}{\ln x}$$

Proof: Integrate by parts to obtain

$$\int_2^x \frac{1}{\ln t}\, dt = \int_2^x \frac{dt}{dt}\,\frac{1}{\ln t}\, dt = \frac{t}{\ln t}\Big|_2^x - \int_2^x t\frac{-\frac{1}{t}}{(\ln t)^2}\, dt = \frac{x}{\ln x} - \frac{2}{\ln 2} + \int_2^x \frac{1}{(\ln t)^2}\, dt$$

To finish the argument, it is enough to show that $\int_2^x \frac{1}{(\ln t)^2}\, dt$ is small compared to $\frac{x}{\ln x}$. Write

$$\int_2^x \frac{1}{(\ln t)^2}\, dt = \int_2^{\sqrt{x}} \frac{1}{(\ln t)^2}\, dt + \int_{\sqrt{x}}^x \frac{1}{(\ln t)^2}\, dt$$

The first integral is at most $\frac{\sqrt{x}}{(\ln 2)^2}$, which is sufficiently small because $\frac{\sqrt{x}}{x/\ln x} = \frac{\ln x}{\sqrt{x}} \to 0$. The second integral is at most $\frac{x}{(\ln\sqrt{x})^2} = \frac{4x}{(\ln x)^2}$ which is again small because

$$\frac{4x/(\ln x)^2}{x/\ln x} = \frac{4}{\ln x} \to 0$$

17.11 A Crucial Calculation

So far, we have not made use of the product formula for the zeta function. Now we use it. Since $\pi(x)$ is a sum over primes, it is reasonable to convert the product into a sum. This can be done in two ways: by using logarithms, and by using logarithmic derivatives. We'll use the second approach.

The product formula only holds when $\Re(z) > 1$, so assume $\Re(z) > 1$ in the following. Then

$$\frac{\zeta'(z)}{\zeta(z)} = \sum_p \frac{\frac{d}{dz}\left(1-\frac{1}{p^z}\right)^{-1}}{\left(1-\frac{1}{p^z}\right)^{-1}} = -\sum_p \frac{\left(1-\frac{1}{p^z}\right)^{-2}\frac{d}{dz}\left(-e^{-z\ln p}\right)}{\left(1-\frac{1}{p^z}\right)^{-1}} = -\sum_p \frac{e^{-z\ln p}\ln p}{\left(1-\frac{1}{p^z}\right)}$$

Simplifying,

$$\frac{\zeta'(z)}{\zeta(z)} = -\sum_p \frac{\ln p}{p^z\left(1-\frac{1}{p^z}\right)} = -\sum_p \ln p\left(\sum_{n\geq 1}\frac{1}{p^{nz}}\right)$$

If n is a positive integer, define

$$\Lambda(n) = \begin{cases} \ln p & \text{if } n \text{ is a power of } p \\ 0 & \text{otherwise} \end{cases}$$

Theorem 115. *If* $\Re(z) > 1$,

$$-\frac{\zeta'(z)}{\zeta(z)} = \sum_{n\geq 1}\frac{\Lambda(n)}{n^z}$$

Remark: In parallel with this formula and its consequences, we will develop a variant and the variant's consequences. For the variant, we begin as above and modify the algbra slightly.

$$-\frac{\zeta'(z)}{\zeta(z)} = \sum_p \ln p \left(\sum_{n\geq 1} \frac{1}{p^{nz}} \right) = \sum_p \frac{\ln p}{p^z} + \left(\sum_{n\geq 2} \frac{1}{p^{nz}} \right) = \sum \frac{\ln p}{p^z} + \sum_p \frac{\ln p}{p^z(p^z-1)}$$

If n is a positive integer, define

$$\Theta(z) = \begin{cases} \ln p & \text{if } n \text{ is a prime } p \\ \\ 0 & \text{otherwise} \end{cases}$$

Theorem 116. *If* $\Re(z) > 1$,

$$-\frac{\zeta'(z)}{\zeta(z)} = \sum_n \frac{\Theta(n)}{n^z} + \sum_p \frac{\ln p}{p^z(p^z-1)}$$

Theorem 117. *The following series converges absolutely and uniformly on compact sets to a function holomorphic on* $\Re(z) > \frac{1}{2}$*:*

$$\sum_p \frac{\ln p}{p^z(p^z-1)}$$

Proof: We have $|p^z - 1| \geq ||p^z| - |1|| = p^x - 1$. If $\Re(x) > \frac{1}{2}$ and $p \neq 2, 3$ then $p^x - 1 \geq p^x/2$, so

$$\left| \frac{\ln p}{p^z(p^z-1)} \right| \leq \frac{\ln p}{p^x(p^x-1)} \leq \frac{2\ln p}{p^{2x}}$$

and this last sum converges when $2x > 1$. Apply the Weierstrass M-test.

Remark: These equations mark the key moment when the zeta function becomes relevant for the prime number theorem. Notice that the calculations were *very* straightforward.

These results also explain why *zeros* of ζ are important, because they become singularities of $\frac{\zeta'(z)}{\zeta(z)}$.

Our proof of the prime number theorem will involve integrating an expression involving $\frac{\zeta'(z)}{\zeta(z)}$ over a vertical line with $\Re(z) > 1$, and then moving this line to the line $\Re(z) = 1$. In this process, $\sum_p \frac{\ln p}{p^z(p^z-1)}$ is irrelevant because it is holomorphic for $\Re(z) > \frac{1}{2}$. So either formula can be used in the proof.

Remark: The functions $\Theta(n)$ and $\Lambda(n)$ are closely related to extremely natural functions first considered by Tchebyshev.

Definition 42. *If x is a positive real number,*

$$\phi(x) = \sum_{n \le x} \Omega(n) = \sum_{p \le x} \ln p$$

$$\psi(x) = \sum_{n \le x} \Lambda(n) = \sum_{p^k \le x} \ln p$$

Theorem 118. *If $\Re(z) > 1$,*

$$-\frac{\zeta'(z)}{\zeta(z)} = z \int_1^\infty \psi(t) t^{-z-1}\, dt$$

$$-\frac{\zeta'(z)}{\zeta(z)} = z \int_1^\infty \phi(t) t^{-z-1}\, dt + \sum_p \frac{\ln p}{p^z(p^z - 1)}$$

Proof: We have

$$z \int_1^\infty \psi(t) t^{-z-1}\, dt = z \sum_{n=1}^\infty \int_n^{n+1} \psi(t) t^{-z-1}\, dt$$

In the interval $[n, n+1]$, $\psi(t)$ is constantly equal to $\psi(n)$, so this is

$$z \sum_{n=1}^\infty \psi(n) \int_n^{n+1} t^{-z-1}\, dt = z \sum_{n=1}^\infty \psi(n) \left. \frac{t^{-z}}{-z} \right|_n^{n+1} dt = -\sum_{n=1}^\infty \psi(n) \left(\frac{1}{(n+1)^z} - \frac{1}{n^z} \right)$$

This sum is

$$-\psi(2)\left(\frac{1}{3^z} - \frac{1}{2^z}\right) - \psi(3)\left(\frac{1}{4^z} - \frac{1}{3^z}\right) - \psi(4)\left(\frac{1}{5^2} - \frac{1}{4^z}\right) + \ldots$$

$$= -\Lambda(2)\left(\frac{1}{3^z} - \frac{1}{2^z}\right) - (\Lambda(2) + \Lambda(3))\left(\frac{1}{4^z} - \frac{1}{3^z}\right) - (\Lambda(2) + \Lambda(3) + \Lambda(4))\left(\frac{1}{5^z} - \frac{1}{4^z}\right) + \ldots$$

Now notice that $\Lambda(2)$ is multiplied by a collapsing sum starting at $-\frac{1}{2^z}$, so this sum is $\frac{\Lambda(2)}{2^z}$. Similarly $\Lambda(3)$ is multiplied by a collapsing sum starting at $-\frac{1}{3^z}$, so this equals $\frac{\Lambda(3)}{3^z}$. In general, we obtain

$$\sum \frac{\Lambda(n)}{n^z} = -\frac{\zeta'(z)}{\zeta(z)}$$

Exactly the same proof yields the second formula.

17.12 The Intuitive Meaning of $\phi(n)$

For a moment, let us study the intuition encoded in the function:

$$\phi(x) = \sum_{p<x} \ln p$$

The function is easy to understand and we can use the prime number theorem to guess its value. Suppose we divide the interval $0 \leq t \leq x$ into subintervals of varying length so the length of the subinterval containing t is about $\ln t$. Since the probability that a number t is prime is $\frac{1}{\ln t}$, each of our subintervals should contain approximately one prime. If we were to sum $\phi(n)$ over the subinterval, we expect to get one term in the sum, namely $\ln p \sim \ln t$. So the contribution to the sum $\phi(x)$ from primes in the subinterval is the length of the subinterval. Thus when we sum over all subintervals, we should get the total length $[0, x]$. In short, $\phi(x) \sim x$.

We will soon, and easily, convert this idea into a rigorous proof that $\phi(x) \sim x$ is equivalent to the prime number theorem.

Notice next that $\psi(x) = \phi(x) + \phi(\sqrt{x}) + \phi(\sqrt[3]{x}) + ...$ Each of these should be considerably smaller than $\phi(x)$. Indeed, it turns out that the prime number theorem is also equivalent to the statement that $\psi(x) \sim x$.

Theorem 119. *The following are equivalent:*

- *The prime number theorem*
- $\phi(x) \sim x$
- $\psi(x) \sim x$

Proof: Choose $0 < \epsilon < 1$ and notice that

$$\frac{\phi(x)}{\ln x} = \frac{\sum_p \ln p}{\ln x} \leq \frac{\sum_p \ln x}{\ln x} = \pi(x) = \sum_{x^{1-\epsilon} \leq p \leq x} 1 + \sum_{p<x^{1-\epsilon}} 1$$

$$\leq \frac{\sum_{x^{1-\epsilon} \leq p \leq x} \ln p}{\ln x^{1-\epsilon}} + x^{1-\epsilon} \leq \frac{\phi(x)}{(1-\epsilon)\ln x} + x^{1-\epsilon}$$

Taking just the important terms,

$$\frac{\phi(x)}{\ln x} \leq \pi(x) \leq \frac{\phi(x)}{(1-\epsilon)\ln x} + x^{1-\epsilon}$$

Multiply by $\ln x$ and divide by x to get

$$\frac{\phi(x)}{x} \leq \frac{\pi(x)}{x/\ln(x)} \leq \frac{\phi(x)}{(1-\epsilon)x} + \frac{\ln x}{x^\epsilon}$$

Reproduce the starting inequality at the right side to get

$$\frac{\phi(x)}{x} \leq \frac{\pi(x)}{x/\ln(x)} \leq \frac{\phi(x)}{(1-\epsilon)x} + \frac{\ln x}{x^\epsilon} \leq \left(\frac{1}{1-\epsilon}\right)\frac{\pi(x)}{x/\ln(x)} + \frac{\ln x}{x^\epsilon}$$

The same sort of argument finishes the proof that $\psi(x) \sim x$ and $\pi(x) \sim \frac{x}{\ln x}$ are equivalent. We'll just do half. Suppose $\frac{\phi(x)}{x} \to 1$. Then $1 \leq \liminf_{x\to\infty} \frac{\pi(x)}{x/\ln(x)}$ and $\limsup_{x\to\infty} \frac{\pi(x)}{x/\ln(x)} \leq \frac{1}{1-\epsilon}$. The last assertion holds for any $\epsilon > 0$, so $\limsup \frac{\pi(x)}{x/\ln(x)} \leq 1$. Consequently, $\lim \frac{\pi(x)}{x/\ln(x)} = 1$.

We now prove that $\frac{\phi(x)}{x} \sim 1$ and $\frac{\psi(x)}{x} \sim 1$ are equivalent. Notice that

$$\phi(x) \leq \psi(x) = \phi(x) + \phi(\sqrt{x}) + \phi(\sqrt[3]{x}) + ...$$

Choose n so $2^{n-1} \leq x < 2^n$. Then $\sqrt[n]{x} < 2$, so $\phi(\sqrt[n]{x}) = 0$. Therefore

$$\phi(x) \leq \psi(x) = \phi(x) + \phi(\sqrt{x}) + \phi(\sqrt[3]{x}) + ... + \phi(\sqrt[n-1]{x}) \leq \phi(x) + (n-2)\phi(\sqrt{x})$$

and so

$$\phi(x) \leq \psi(x) \leq \phi(x) + (n-2)\sqrt{x}\ln\sqrt{x}$$

But $2^{n-1} \leq x$ and so $(n-1)\ln 2 \leq \ln x$ or $(n-1) \leq \frac{\ln x}{\ln 2}$. We conclude that

$$\phi(x) \leq \psi(x) \leq \phi(x) + \frac{\ln x}{\ln 2}\sqrt{x}\frac{1}{2}\ln x = \phi(x) + \frac{1}{2\ln 2}\sqrt{x}(\ln x)^2$$

Dividing by x, and adding an additional trivial inequality at the end, gives

$$\frac{\phi(x)}{x} \leq \frac{\psi(x)}{x} \leq \frac{\phi(x)}{x} + \frac{1}{2\ln 2}\frac{(\ln x)^2}{\sqrt{x}} \leq \frac{\psi(x)}{x} + \frac{1}{2\ln 2}\frac{(\ln x)^2}{\sqrt{x}}$$

Taking the limit as $x \to \infty$, we conclude that if $\frac{\phi(x)}{x} \to 1$, then $\frac{\psi(x)}{x} \to 1$ and conversely. QED.

17.13 Zeros of $\zeta(z)$ on the Boundary of the Critical Strip

We are very close to proving the prime number theorem. This theorem was proved in 1896, independently by Hadamard and de la Vallee Poussin. A key ingredient of both proofs was the following result, again proved independently by both men. The proof was simplified by Mertens in 1898 to the following argument.

Theorem 120. *The zeta function has no zeros on the boundary of the critical strip.*

Proof: A zero on the boundary of the critical strip would lead to four zeros, two on the left and two on the right. Concentrate on the right side. Zeros there would lead to singularities of $-\frac{\zeta'(z)}{\zeta(z)}$, and consequently by the previous section would produce singularities of $f(z) = \sum_p \frac{\ln p}{p^z}$ as z approaches the singularity from the right.

Suppose $\zeta(z)$ has a zero at $1 + i\alpha$. Call the order of this zero μ. It might or might not also have a zero at $1 + 2i\alpha$. Call the order at this point ν. Then ζ will have zeros at $1 - i\alpha$ and possibly at $1 - 2i\alpha$ of orders μ and ν, and it will have a pole at 1 of order 1.

If g is holomorphic and $g(z) = A(z-z_0)^k + ...$, then $g' = kA(z-z_0)^{k-1} + ...$ and $-\frac{g'}{g} = \frac{-k}{z-z_0} + ...$ We conclude that as $\epsilon \to 0$ through positive values, $f(1+i\alpha+\epsilon) \sim -\mu/\epsilon$, $f(1+2\alpha+\epsilon) \sim -\nu/\epsilon$, $f(1+\epsilon) \sim 1/\epsilon$, with similar results in the lower half plane.

(And now the trick.) Form

$$f(1 + 2i\alpha + \epsilon) + 4f(1 + i\alpha + \epsilon) + 6f(1 + \epsilon) + 4f(1 - i\alpha + \epsilon) + f(1 - 2i\alpha + \epsilon)$$

This sum equals

$$\sum \frac{\ln p}{p^{1+\epsilon}} \left(\frac{1}{p^{2i\alpha}} + 4\frac{1}{p^{i\alpha}} + 6 + 4\frac{1}{p^{-i\alpha}} + \frac{1}{p^{-2i\alpha}} \right) = \sum \frac{\ln p}{p^{1+\epsilon}} \left(\frac{1}{p^{i\alpha}} + \frac{1}{p^{-i\alpha}} \right)^4$$

The crucial point is that this expression is always real and greater than or equal to zero. So if we divide by ϵ and take the limit as $\epsilon \to 0$, we get a non-negative number, specifically $-\nu - 4\mu + 6 - 4\mu - \nu$. Therefore $-2\nu - 8\mu + 6 \geq 0$. Since $\nu \geq 0, \mu \geq 0$, this cannot happen if $\mu \geq 1$. QED.

17.14 Proof of the Prime Number Theorem

Theorem 121. *We have $\phi(x) \sim x$, and therefore*

$$\pi(x) \sim \frac{x}{\ln x} \sim \int_2^x \frac{1}{\ln t}\, dt$$

Remark: The original proofs of 1896 were quite difficult, requiring information on a zero-free region for ζ in the critical strip. Gradually the required extra information was reduced, until Wierner and Ikehara reduced the proof to a Tauberian theorem in 1931. This theorem required no further information about the critical strip, but did require tricky analysis. In 1980, Newman reduced the proof to easy results on contour integration. I find Newman's treatment fairly technical. Dan Zagier wrote several papers simplifying Newton's proof. Our treatment follows Zagier's exposition in the Monthly for 1997.

Step 1:

Theorem 122. *The prime number theorem is a consequence of the statement that the following integral converges:*

$$\int_1^\infty \frac{\phi(t)-t}{t^2}\,dt$$

Proof: Suppose there is a $\lambda > 1$ and a sequence $x_n \to \infty$ with $\phi(x_n) - x_n \geq (1+\lambda)x_n$. Then since ϕ is nondecreasing,

$$\int_{x_n}^{\lambda x_n} \frac{\phi(t)-t}{t^2}\,dt \leq \int_{x_n}^{\lambda x_n} \frac{\lambda x_n - t}{t^2}\,dt = (t = x_n u)\int_1^\lambda \frac{\lambda x_n - u x_n}{x_n^2 u^2} x_n du = \int_1^\lambda \frac{\lambda - u}{u^2}\,du > 0$$

Therefore, out in the tail of the integral where integration over arbitrary intervals goes to zero, there are bumps of the same positive constant size, a contradiction.

The same argument gives the opposite required inequality. Suppose there is a $\lambda < 1$ and a sequence $x_n \to \infty$ with $\phi(x_n) - x_n \leq \lambda x_n$. Then

$$\int_{\lambda x_n}^{x_n} \frac{\phi(t)-t}{t^2}\,dt \leq \int_{\lambda x_n}^{x_n} \frac{\lambda x_n - t}{t^2}\,dt = (t = x_n u)\int_\lambda^1 \frac{\lambda x_n - u x_n}{x_n^2 u^2}\,du = \int_\lambda^1 \frac{\lambda - u}{u^2}\,du < 0$$

Step 2: Consider the equation for $\Re(z) > 1$ developed earlier:

$$-\frac{\zeta'(z)}{\zeta(z)} = z\int_1^\infty \frac{\phi(t)}{t^{z+1}}\,dt + \sum_p \frac{\ln p}{p^z(p^z-1)}$$

Near $z = 1$ we have $\zeta(z) = \frac{1}{z-1} + ...$, so $-\frac{\zeta'(z)}{\zeta(z)} = -\frac{(-1)(z-1)^2+...}{(z-1)^{-1}+...} = \frac{1}{z-1} + ...$ Notice that $z\int_1^\infty \frac{t}{t^{z+1}}\,dt = z\int_1^\infty \frac{1}{t^z}\,dt = z\left.\frac{t^{-z+1}}{-z+1}\right|_1^\infty = -z\frac{1}{-z+1} = \frac{z}{z-1} = \frac{(z-1)+1}{z-1} = 1 + \frac{1}{z-1}$. So

$$-\frac{\zeta'(z)}{\zeta(z)} - \frac{1}{z-1} = z\int_1^\infty \frac{\phi(t)-t}{t^{z+1}}\,dt + 1 + \sum_p \frac{\ln p}{p^z(p^z-1)}$$

In this expression, every term is defined and holomorphic on some open set containing $\Re(z) \geq 1$, so the integral also defines a holomorphic function on $\Re(z) > 1$ which can be extended to be holomorphic on the larger open set. But we have to be careful. This does not mean that the integral exists on this larger open set. It only means that the function given by the integral expression can be analytically continued to the larger open set.

Nevertheless, we are *very* close to the required assertion that $\int_1^\infty \frac{\phi(t)-t}{t^2}$ exists!

Step 3: The final steps require knowing that $\frac{\phi(t)}{t}$ is bounded. We will prove that $\phi(t) < 2t$ following Chebyshev in 1849. (Chevyshev proved a much stronger result). Start with the middle binomial coefficient $\binom{2n}{n} = \frac{(2n)!}{n!n!}$. This number is smaller than 2^{2n} and every prime between n and $2n$ divides it. So $\Pi_{n<p\leq 2n} p < 2^{2n}$. Taking logs, $\sum_{n<p\leq 2n} \ln p = \phi(2n) - \phi(n) \leq 2n \ln 2 < 1.4n$.

We now prove that $\phi(n) \leq 2n$ by induction. This holds for $n = 1$ since $\ln 2 < 2$. The induction step for even numbers follows because

$$\phi(2n) \leq \phi(n) + 1.4n \leq 2n + 1.4n = 3.4n < 4n$$

and the induction step for odd numbers also holds because

$$\phi(2n+1) \leq \phi(2n) + \ln(2n+1) \leq 3.4n + \ln(2n+1) < 2(2n+1)$$

because $\ln(2n+1) < .6n + 2$. A direct check shows that this last inequality holds for the first few n, and after that, the inequality is clear.

Step 4: We next make a convenient change of variables. Let $t = e^u$. Then

$$\int_1^\infty \frac{\phi(t) - t}{t^2}\, dt = \int_0^\infty \frac{\phi(e^u) - e^u}{e^{2u}} e^u\, du = \int_0^\infty \left[e^{-u}\phi(e^u) - 1\right]\, du$$

Let $t = e^u$ and let $z = w + 1$. Then

$$z\int_1^\infty \frac{\phi(t) - t}{t^{z+1}}\, dt = (w+1)\int_1^\infty \frac{\phi(t) - t}{t^{w+2}}\, du = (w+1)\int_0^\infty \frac{\phi(e^u) - e^u}{e^{(w+2)u}}\, e^u\, du$$

$$= (w+1)\int_0^\infty e^{-wu}\left[e^{-u}\phi(e^u) - 1\right]\, du$$

Recall that the second integral is holomorphic in w for $\Re(w) > 0$ and can be extended to be a holomorphic function on an open set containing $\Re(w) = 0$. By step 3, $[e^{-u}\phi(e^u) - 1]$ is bounded.

Step 5: Finally we apply the following theorem, the essential step in Newman's simplified proof, to $f(t) = \left[e^{-t}\phi(e^t) - 1\right]$:

Theorem 123. *Let $f(t)$ be a bounded real-valued function on $[0, \infty)$ which is continuous and locally Riemann integrable. Suppose $g(z) = \int_0^\infty e^{-zt} f(t)\, dt$ exists and is holomorphic on $\Re(z) > 0$ and suppose it can be extended to a holomorphic function $g(z)$ on an open set that contains $\Re(z) = 0$. Then $\int_0^\infty f(t)\, dt$ exists.*

Proof: Let $g_T(z) = \int_0^T e^{-zt} f(t)\, dt$. This function is defined and holomorphic in the entire complex plane. We must prove that $\lim_{T\to\infty} g_T(0)$ exists. We will prove it equals $g(0)$. Caution

is required here because $g(0)$ is not an integral, but instead the extension of a holomorphic function to a larger set.

Select a large real R. By Cauchy's theorem,

$$g(0) - g_T(0) = \frac{1}{2\pi i} \int_\gamma (g(z) - g_T(z))\, e^{zT} \left(1 + \frac{z^2}{R^2}\right) \frac{dz}{z}$$

where γ is the path illustrated below; this path is a circle of radius R, but cut off at horizontal distance δ from the imaginary axis on the left to remain inside the region where g is holomorphic. The δ will depend on R. This holds since the only singularity of the integrand is at zero, where the residue is $g(0) - g_T(0)$. We want this expression to go to zero as $R \to \infty$, so we must estimate the integral along the boundary curve.

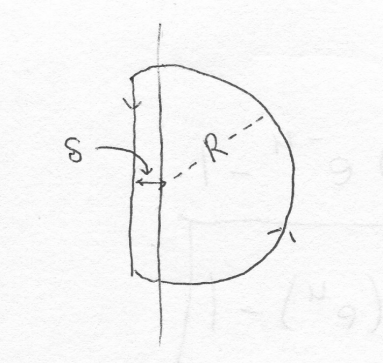


Figure 17.6: Newman's Path

Recall that $f(t)$ is bounded, say by B. Then on the right half of the curve, $|g(z) - g_T(z)| = \left|\int_T^\infty f(t)e^{-zt}\, dt\right| \le B \int_T^\infty \left|e^{-zt}\right| dt = Be^{-\Re(z)T}/\Re(z)$. Also for $\Re x \ge 0$ we have

$$\left|\left(1 + \frac{z^2}{R^2}\right)\frac{1}{z}\right| = \left|\frac{1}{z} + \frac{z}{z\overline{z}}\right| = \left|\frac{1}{z} + \frac{1}{\overline{z}}\right| = \left|\frac{z + \overline{z}}{z\overline{z}}\right| = \frac{2\Re(z)}{R^2}$$

so

$$\left|e^{zT}\left(1 + \frac{z^2}{R^2}\right)\frac{1}{z}\right| \le e^{\Re(z)T} \cdot \frac{2\Re(z)}{R^2}$$

Hence the total contribution from the right side is

$$\frac{Be^{-\Re(z)T}e^{\Re(z)T}2}{R^2} \cdot \pi R = \frac{2\pi B}{R}$$

and this goes to zero as $R \to 0$.

On the left half of the curve, we approximate $g(z)$ and $g_T(z)$ separately. Since $g_T(z)$ is holomorphic everywhere, we can replace the left half by another semicircle without changing the integral. But then exactly the same approximation as before applies. Indeed since $\Re(x) \leq 0$ we have

$$g_T(z)| = \left| \int_0^T f(t)e^{-zt}\, dt \right| \leq B \int_{-\infty}^{T} |e^{-zt}|\, dt = B \int_{-\infty}^{T} e^{-\Re(z)t}\, dt = \left. \frac{Be^{-\Re(x)t}}{-\Re(x)} \right|_{-\infty}^{T} = \frac{Be^{-\Re(x)T}}{|\Re(x)|}$$

The remaining calculation works as before.

Finally we study the term $g(z)$. This time the exact contour matters. The integrand contains the term $g(z)\left(1 + \frac{z^2}{R^2}\right)\frac{1}{z}$ and the term e^{zT}. The first term contains no reference to T. The term e^{zT} goes to 0 in the half plane $\Re(z) < 0$, and uniformly on compact subsets. as $T \to \infty$. Hence we can choose two tails where the curve crosses the $x = 0$ axis, of arbitrarily small length, and off these tails, everything goes to zero as $T \to \infty$. So the left integral can be made arbitrarily small. QED.

17.15 Consequences of the Riemann Hypothesis

We proved that $\pi(x) \sim \frac{x}{\ln x}$ and equivalently $\pi(x) \sim \int_2^x \frac{1}{\ln t}\, dt = \mathrm{li}(x)$. The table on the next page shows some concrete data. Note that the columns for $\frac{x}{\pi x}$ and $\mathrm{li}(x)$ record the *error* of the resulting approximation, for instance $\left|\pi(x) - \frac{x}{\ln x}\right|$:

x	$\pi(x)$	$\pi(x) - \frac{x}{\ln x}$	$\mathrm{li}(x) - \pi(x)$
10	4	0	1
10^2	25	3	5
10^3	168	23	10
10^4	1,119	143	17
10^5	9,592	905	38
10^6	78,498	6,116	130
10^7	664,579	44,158	339
10^8	5,761,455	332,774	754
10^9	50,847,534	2,592,592	1,701
10^{10}	455,052,522	20,758,029	3,104
10^{11}	4,118,054,813	169,923,159	11,588
10^{12}	37,607,012,018	1,416,705,193	38,263
10^{13}	346,065,536,839	11,992,858,452	108,971
10^{14}	3,204,941,750,802	102,838,308,636	314,890
10^{15}	29,844,570,422,669	891,604,962,452	1,052,619
10^{16}	279,238,341,033,925	7,804,289,844,393	3,214,632
10^{17}	2,623,557,157,654,233	68,883,734,693,281	7,956,589
10^{18}	24,739,954,287,740,860	612,483,070,893,536	21,949,555
10^{19}	234,057,667,276,344,607	5,481,624,169,369,960	99,877,775
10^{20}	2,220,819,602,560,918,840	49,347,193,044,659,701	222,744,644
10^{21}	21,127,269,486,018,731,928	446,579,871,578,168,707	597,394,254
10^{22}	201,467,286,639,315,906,290	4,060,704,006,019,620,994	1,932,355,208
10^{23}	1,925,320,391,606,803,968,923	37,083,513,766,578,631,309	7,250,186,216
10^{24}	18,435,599,767,349,200,867,866	339,996,354,713,708,049,069	17,146,907,278
10^{25}	176,846,309,399,143,769,411,680	3,128,516,637,843,038,351,228	55,160,980,939

Remark: This table shows that li is incredibly closer to $\pi(x)$ than $\frac{x}{\ln x}$. The number of primes up to $10^{25} = 10,000,000,000,000,000,000,000,000$ is almost 10^{24}, yet $\frac{x}{\ln x}$ only matches the first two digits of the result. But li matches the first 13 digits, more than half. Indeed, the data suggests that $\mathrm{li}(x)$ consistently gets half of the digits correct. Another way to say this is that

$$|\pi(x) - \mathrm{li}(x)| \sim \sqrt{x}$$

Consider the set of all exponents e such that $|\pi(x) - \mathrm{li}(x)| < Cx^e$ for some constant which may depend on e. Shockingly, the best e currently known is $e = 1$, and of course $|\pi(x) - \mathrm{li}(x)| < x$ is trivial. Results of Riemann imply that e is at least $\frac{1}{2}$.

In 1904, von Koch, who is not related to me, proved that the inf of all such e is also the inf of the e such that the zeta function has no zeros z_0 with $\Re(z_0) > e$. Unfortunately, as far as we know,

the nontrivial zeros of the zeta function can get arbitrarily close to the line $|Re(z) = 1$.

Riemann proved that the Zeta function has infinitely many non-trivial zeros and conjectured that they all have real part $\frac{1}{2}$. This is one of the Field's Problems; a solution earns one million dollars. The Riemann hypothesis is today the most famous unsolved mathematical problem. A solution would imply $|\pi(x) - \text{li}(x)| < Cx^{1/2+\epsilon}$ for every $\epsilon > 0$. Indeed assuming the Riemann hypothesis is true, number theorists have proved that $|\pi(x) - \text{li}(x)| < C\sqrt{x}\ \ln x$.

Many zeros have been computed. They are all on the critical line. Hardy proved that there are infinitely many zeros on the critical line, and Levinson proved that at least 1/3 of the nontrivial zeros are on the criticial line.

17.16 Other Results in Riemann's Paper

The second half of Riemann's paper contained incredible additional results, some with sketchy proofs. All were eventually proved rigorously. Here is a list of some items:

Number of Zeros: Riemann proved that the number of non-trivial zeros in the critical strip of height $0 < h < T$ equals

$$\frac{T}{2\pi} \ln \frac{T}{2\pi} - \frac{T}{2\pi} + O(\ln T)$$

Hence the density of the zeros gradually increases as the height increases.

Product Formula: Earlier we introduced the entire function

$$z(1-z)\pi^{-z/2}\Gamma\left(\frac{z}{2}\right)\zeta(z)$$

This function is unchanged when z is replaced by $1-z$, and its zeros are exactly the non-trivial zeros of zeta in the critical strip.

Riemann proved that the negative of this function is given by a typical Weierstrass product over the roots of ζ:

$$z(z-1)\pi^{-z/2}\Gamma\left(\frac{z}{2}\right)\zeta(z) = e^{Az}\prod_{\rho}\left(1-\frac{z}{\rho}\right)e^{z/\rho}$$

where $A = 1 + \frac{\gamma}{2} - \ln 2\sqrt{\pi}$. In particular, these zeros completely determine the zeta function.

Explicit Formula for Primes: Perhaps the most astonishing result in the paper is an analytic formula for $\pi(x)$. This formula is somewhat complicated; later authors deduced it from a more straightforward formula for Tschebychev's $\psi(x)$:

$$\psi(x) = x - \sum_{\rho}\frac{x^{\rho}}{\rho} - \ln 2\pi - \frac{1}{2}\ln\left(1-\frac{1}{x^2}\right)$$

Notice that the leading term $\phi(x) \sim x$ gives the prime number theorem; this term comes from the pole of zeta at $z = 1$. The sum is over zeros of zeta; more about that term below. The final terms come from the trivial zeros of the zeta function and are very small.

Assuming the Riemann hypothesis,

$$\frac{x^\rho}{\rho} = e^{(1/2+i\Im(\rho))\ln x} = \sqrt{x}\, e^{i\ \Im(\rho)\ \ln x}$$

This means that the correction terms due to the zeros of ζ should have size approximately $\sqrt{x}$, and oscillate. To give a flavor of the oscillation, here is a graph of $\sin(10 \ln x)$ between 0 and 100:

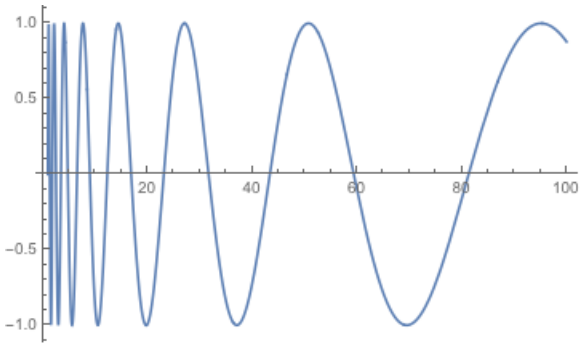


Figure 17.7: $\sin(10 \ln x$

Riemann described the contribution of these oscillatory terms as describing the "alternating thickening and thinning of the primes."

Of course $\psi(x)$ is not continuous, so the sum is not absolutely convergent. This makes it difficult to draw consequences from the sum. Yet it somehow explains the role of the zeros of zeta, which predict the "fine structure of the primes."

Another form of Riemann's formula is obtained by differentiation. Of course $\psi(x)$ is flat except for jumps, so the derivative must be computed in the spirit of distribution theory. We get

$$\psi'(x) = 1 - \sum x^{\rho-1} - \frac{1}{x(x^2-1)}$$

Here the last term is very small and the significant term is the sum in the middle. Zeros on the critical line come in pairs, so we should consider $x^{\rho-1} + x^{\bar{\rho}-1}$. A brief calculation shows that this expression equals $\frac{1}{\sqrt{x}} \cos(\Im(\rho) \ln x)$.

In 2009, David Borthwick gave a talk on prime numbers at Emory University. His slides include the following astonishing picture of the approximation of $\psi'(x)$ using the above formula and

the first 100 pairs of zeros of the zeta function. According to distribution theory, the sum should equal a Dirac delta function at each discontinuity of $\psi(x)$, and so at the powers of the primes. Notice sharp peaks at 2, 3, 4, 5, 7, 8, 9, 11, 13, 16, 17, and 19:

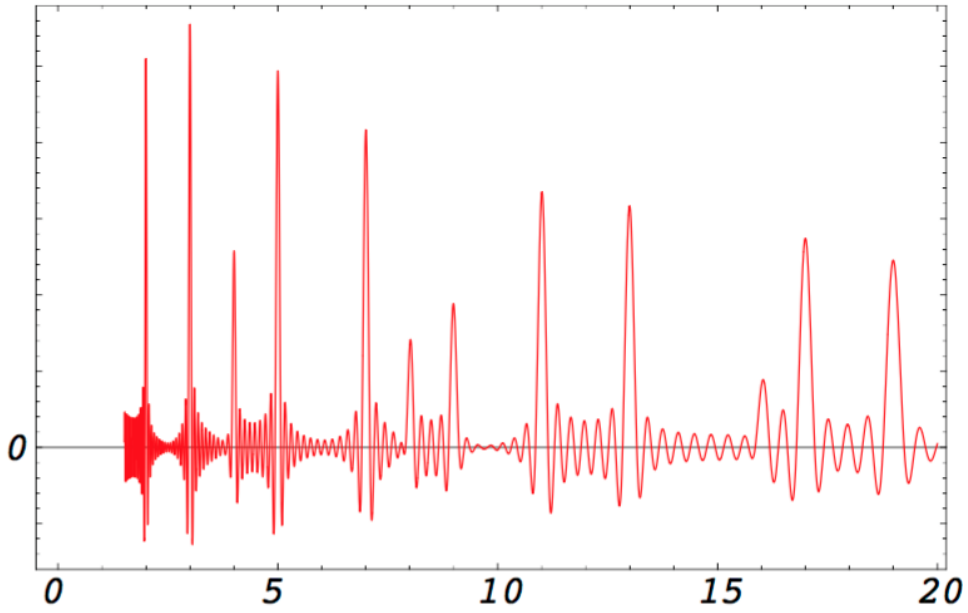


Figure 17.8: Approximation of $\psi'(x)$ with 100 Zeros of Zeta

Chapter 18

Hadamard's Theory of Entire Functions

18.1 Introduction

The fundamental theorem of algebra states that every complex polynomial can be factored as a product of its roots:

$$P(z) = C \prod (z - \lambda_i)$$

Euler had the beautiful idea of generalizing this result to arbitrary entire functions. We cannot write

$$f(z) = C \prod_{\text{zeros}} (z - \lambda_i)$$

because this product does not converge; indeed, when an infinite product converges, the individual terms converge to 1. But if $\lambda_i \neq 0$, we can easily fix this by replacing the term $(z - \lambda_i)$ with the term

$$\left(\frac{\lambda_i - z}{\lambda_i}\right) = \left(1 - \frac{z}{\lambda_i}\right)$$

So we might hope that any $f(z)$ can be factored as

$$f(z) = C\, z^m \prod_{\text{roots}} \left(1 - \frac{z}{\lambda_i}\right)$$

Using this idea, Euler deduced the value of $\sum \frac{1}{n^2} = \frac{\pi^2}{6}$, which made him instantly famous in Europe. To obtain this result, factor $\sin z$ as

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

Then

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

and taking the limit as $z \to 0$ gives $C = 1$.

Now expand the left side as a power series about the origin, multiply out the right side, and equate terms. We obtain

$$1 - \frac{z^2}{3!} + \frac{z^4}{5!} - ... = 1 - \frac{1}{\pi^2} \sum_{n>0} \left(\frac{1}{n^2}\right) z^2 + \frac{1}{\pi^4} \sum_{0<m<n} \left(\frac{1}{m^2 n^2}\right) z^4 + ...$$

Equating the coefficients of z^2 gives

$$\sum_{n>0} \left(\frac{1}{n^2}\right) = \frac{\pi^2}{6}$$

Equating the coefficients of z^4 gives

$$\frac{\pi^4}{5!} = \sum_{0<m<n} \left(\frac{1}{m^2 n^2}\right) = \frac{1}{2}\left[\left(\sum_{m>0} \frac{1}{m^2}\right)^2 - \left(\sum_{m>0} \frac{1}{m^4}\right)\right] = \frac{1}{2}\left[\frac{\pi^4}{36} - \left(\sum_{m>0} \frac{1}{m^4}\right)\right]$$

or

$$\sum_{m>0} \frac{1}{m^4} = \frac{\pi^4}{36} - \frac{2\pi^4}{5!} = \frac{\pi^4}{90}$$

Further work gives the value of the zeta function at every positive even integer. The values of the zeta function at positive odd integers are completely unknown.

Euler worried that the sine function might have additional complex roots. Eventually he showed that all zeros are real using the formula $\sin z = \frac{e^{iz} - e^{-iz}}{2i}$. As a matter of fact, $C\, z \prod_{n \neq 0} \left(1 - \frac{z}{\pi n}\right)$ does not actually converge, but the formula obtained by combining terms with negative and positive n is rigorously true, as we proved earlier in these notes.

18.2 Weierstrass

Let us now return to the general factorization theorem for entire functions. This factorization was made rigorous 100 years later by Weierstrass. We covered his theory in an earlier chapter.

It turns out that the conjectured formula

$$f(z) = C\, z^m \prod_{\text{roots}} \left(1 - \frac{z}{\lambda_i}\right)$$

is too optimistic because the product still need not converge. Weierstrass forced convergence by multiplying $\left(1-\frac{z}{\lambda}\right)$ by an extra term to make it closer to 1. To make his idea easier to follow, consider the special case when $\lambda = 1$. Then

$$(1-z)\,e^{-\log(1-z)}$$

would be identically one if we didn't have to worry about the domain of log. If z is small, $\frac{1}{1-z} = 1 + z + z^2 + ...$ and integration gives

$$log\left(\frac{1}{1-z}\right) = -\left(z + \frac{z^2}{2} + \frac{z^3}{3} + ...\right)$$

Therefore

$$e^{-\log(1-z)} = e^{\log\left(\frac{1}{1-z}\right)} = e^{z+\frac{z^2}{2}+\frac{z^3}{3}+...}$$

We therefore define

$$E_n(z) = (1-z)e^{z+\frac{z^2}{2}+...+\frac{z^n}{n}}$$

and replace the term $1-\frac{z}{\lambda_i}$ in the conjectured factorization by

$$E_n\left(\frac{z}{\lambda_i}\right)$$

for appropriate n. If the λ_i are infinitely many non-zero complex numbers, possibly with repetitions, converging to infinity, we can choose integers n_i so the product

$$\prod E_{n_i}\left(\frac{z}{\lambda_i}\right)$$

converges in the entire complex plane. If $f(z)$ is an entire function with zeros at the λ_i and such that $f(0) \neq 0$, then

$$\frac{f(z)}{\prod E_{n_i}\left(\frac{z}{\lambda_i}\right)} = h(z)$$

is an entire function which is never zero. We earlier proved that we can compute a single-valued logarithm for such a function, so $h(z) = e^{g(z)}$ for some entire function g. Thus we obtain

Theorem 124 (Weierstrass Factorization Theorem)**.** *Suppose $f(z)$ is an entire function with nonzero roots at λ_i and perhaps also a root of order m at the origin. Then there is an entire function $g(z)$ and integers n_i such that*

$$f(z) = e^{g(z)}z^m \prod E_{n_i}\left(\frac{z}{\lambda_i}\right)$$

The trouble with this result is that there is no information about $g(z)$ and no control over the integers n_i, and thus we do not get explicit formulas which we can "mine" as Euler mined his factorization of $\sin z$.

18.3 Riemann

Let us return to Riemann's short paper on the zeta function. This paper starts with Riemann's analytic continuation to the entire complex plane, and with the functional equation of the zeta function. In both of these cases, Riemann gives a completely rigorous proof. After that he writes conjectures without rigorous proofs. Some authors have called these "anticipations." Here is one astonishing case. Riemann introduced the function

$$z(z-1)\pi^{z/2}\Gamma\left(\frac{z}{2}\right)\zeta(z)$$

The terms on the left cancel out the pole of zeta and the trivial zeros at $-2, -4, ...$, and have been chosen so the functional equation becomes the statement that this function is unchanged if each z is replaced by $1-z$. Then Riemann asserts that this function is given by the Weierstrass product

$$e^{-Az}\prod_{\rho}\left(1-\frac{z}{\rho}\right)e^{z/\rho}$$

where $A = 1 + \frac{\gamma}{2} - \ln 2\sqrt{\pi}$ and the ρ_i are the zeros of the zeta function in the critical strip. This is the Weierstrass factorization, but the function $g(z)$ has become a polynomial of degree one and the n_i are all 1 and we obtain a specific formula. In particular, the zeros of the zeta function in the critical strip completely determine the zeta function.

In the years after Riemann, various mathematicians gave rigorous proofs of Riemann's anticipations. This particular result was proved by Hadamard, and in the process he developed a general theory of entire functions which has become part of the canon of complex analysis. We will present Hadamard's theory in this chapter, and use it to obtain this result of Riemann in the following chapter. Hadamard's basic idea is quite straightforward. He showed that we can control $g(z)$ and the integers n_i if we have information about how fast our function grows at infinity. We already proved a very simple case of this theory: if $f(z)$ is entire and $|f(z)| \leq C\,|z|^N$ for some constant C, then f is a polynomial.

18.4 Entire Functions of Finite Order

What sort of growth should we expect if $f(z) = e^{g(z)}z^m\prod E_n\left(\frac{z}{\lambda_i}\right)$ and $g(z)$ is a polynomial? We can get a good idea by looking just at the first term

$$|e^{g(z)}| = |e^{\Re g(z)}e^{i\Im g(z)}| = e^{\Re g(z)} \leq e^{|\Re g(z)|} <= e^{|g(z)|} \leq e^{\sum_{k=0}^{d}|c_k||z|^k|}$$

If $|z| > 1$, this is bounded by

$$e^{\left(\sum_{k=0}^{d} |c_k|\right)|z|^d}$$

So we expect that $|e^{P(z)}|$ is approximately $e^{(|z|^d)}$.

Definition 43. *An entire function has* **finite order** *if there are positive real numbers ρ and R such that whenever $|z| > R$ we have*

$$|f(z)| < e^{(|z|^\rho)}$$

In that case, the greatest lower bound of such ρ is called the **order** *of f.*

Remark: The order of f will be denoted ρ. The inequality $|f(z)| < e^{(|z|^\rho)}$ need not hold, but for every $\epsilon > 0$ there is an R such that $|z| > R$ implies $|f(z)| < e^{\left(|z|^{\rho+\epsilon}\right)}$.

We often get this sort of estimate with an extra constant, C. That is, for every $\epsilon > 0$ there is an R such that $|z| > R$ implies $|f(z)| < e^{C\,(|z|^{\rho+\epsilon})}$. This estimate is good enough, for we can find a larger R_1 such that $|z| > R_1$ implies $C < |z|^\epsilon$ and then

$$|f(z)| < e^{C\,(|z|^{\rho+\epsilon})} < e^{(|z|^\epsilon |z|^{\rho+\epsilon})} = e^{(|z|^{\rho+2\epsilon})}$$

The inequality $|f(z)| < e^{(|z|^{\rho+\epsilon})}$ is equivalent to the inequality $\log|f(z)| < |z|^{(\rho+\epsilon)}$. A reader might worry about zeros of f, but near a zero, $\log|f(z)|$ will be extremely negative and certainly satisfy the inequality, so defining $\log 0 = -\infty$ makes the inequality true also at zeros.

Examples: Polynomials have order 0, e^z and $\sin z$ have order 1. Notice that $\cos(\sqrt{z})$ is well-defined and entire because $\cos w = \cos(-w)$; this function has order $\frac{1}{2}$.

18.5 Rank and Genus

Suppose $f(z)$ is an entire function. We can then list its nonzero roots $\lambda_1, \lambda_2, \lambda_3, \ldots$. We always repeat a root as many times as its order in f, and we list so $|\lambda_1| \leq |\lambda_2| \leq |\lambda_3| \leq \ldots$. We are interested in factorizations of the form

$$f(z) = e^{g(z)} z^m \prod_i E_d\left(\frac{z}{\lambda_i}\right)$$

where he same integer d is used for each of the "Weierstrass factors." We need to find a d making the product converge.

Recall the theory of infinite products developed earlier. It is tricky to define convergence of infinite products, since we want the infinite product to be zero only if one of the terms is zero and we want the individual terms in a convergent product to approach 1. Therefore we defined $\prod_{k=1}^{\infty} a_k = A$ to mean that when we omit a finite number of initial terms, the resulting infinite product converges to a non-zero limit, and A is this limit times the omitted terms.

We want absolute convergence of products so we can rearrange terms, and uniform convergence on compact sets so products of holomorphic functions are holomorphic. Ultimately we discovered that these conditions would hold if we wrote $f_i(z) = 1 + g_i(z)$ and required that $\sum |g_i(z)|$ converge uniformly on compact subsets.

Consequently, $\prod_i E_d\left(\frac{z}{\lambda_i}\right)$ converges in our sense if and only if $\sum |1 - E_d(z/\lambda_i)|$ converges uniformly on compact subsets. We then proved that if $|z| \leq 1$, then $|1 - E_d(z)| \leq |z|^{d+1}$.Since the roots λ_i must go to infinity, when we restrict z to a compact subset we have $|z/\lambda_i| \leq 1$ for all but finitely many i, and thus we need only check that $\sum \left|\frac{z}{\lambda_i}\right|^{d+1}$ converges. Since z is in a compact subset, this will converge provided

$$\sum \left|\frac{1}{\lambda_i}\right|^{d+1}$$

converges.

Definition 44. *If such a d exists for the nonzero roots of an entire function $f(z)$, we say that f has* **finite rank***. In that case, the smallest non-negative integer d satisfying the result is called the* **rank** *of $f(z)$.*

Remark: We will not worry about finding the best possible d. Our theory holds once we find any d that works. We can then write

$$f(z) = e^{g(z)} z^m \prod_i E_d\left(\frac{z}{\lambda_i}\right)$$

We will call an expression of this form a *Hadamard product* if $\sum \left|\frac{1}{\lambda_i}\right|^{d+1}$ converges and g is a polynomial.

We can state the main results of Hadamard Theory. These results will be proved in the following sections.

Theorem 125. *Suppose $f(z)$ is an entire function and d is an integer such that $\sum \left|\frac{1}{\lambda_i}\right|^{d+1}$ converges. If the resulting product for $f(z)$ is a Hadamard product, then f has finite order ρ. Indeed $\rho \leq \max(d+1, \deg g)$.*

Theorem 126 (Hadamard)**.** *Suppose $f(z)$ is an entire function with finite order ρ. Then $f(z)$ has a Hadamard factorization. In fact we can choose a factorization with $\rho \leq \max(d+1, \deg g)$.*

18.6 Functions with Good Factorization Have Finite Order

Here we treat the easier direction. Assume $f(z)$ is an entire function with a Hadamard product. We will prove that f has finite order ρ and $\rho \leq \deg g$ and also $\rho \leq (d+1)$.

Our function f is a product of three terms, $e^{P(z)}$, z^m, and $\prod_i E_d\left(\frac{z}{\lambda_i}\right)$. It suffices to prove the result for each of these terms because then

$$|f(z)| \leq e^{(|z|^{\rho_1+\epsilon})} e^{(|z|^{\rho_2+\epsilon})} e^{(|z|^{\rho_1+\epsilon})} = e^{(|z|^{\rho_1+\epsilon}+|z|^{\rho_2+\epsilon}+|z|^{\rho_3+\epsilon})}$$

and if ρ is the maximum of ρ_1, ρ_2, ρ_3, this is at most

$$e^{(3|z|^{\rho+\epsilon})}$$

and the 3 makes no difference by an earlier remark.

The estimate for $e^{g(z)}$ was given at the start of the section defining order, and the estimate for z^m follows because if $\epsilon > 0$ then $\log |z^m| = m \log |z| \leq |z|^\epsilon$ for all large enough $|z|$.

We now examine the difficult case $\prod_i E_d\left(\frac{z}{\lambda_i}\right)$. We will obtain the desired estimate by carefully estimating $|E_d(z)|$ in three cases: $|z| < \frac{1}{2}$, $\frac{1}{2} \leq |z| \leq 2$, and $2 < |z|$. Since we are actually interesting in $\frac{z}{\lambda_i}$, this corresponds to the cases that $|z|$ is much smaller than the absolute value of the zero, $|z|$ is close to the absolute value of the zero, and $|z|$ is much larger than the absolute value of the zero.

In the calculation below, we use the easy facts $|z_1 z_2| = |z_1|\,|z_2|$ and $|z_1 + z_2| \leq |z_1| + |z_2|$ and $|e^z| \leq e^{|z|}$.

Lemma 13. *Fix a non-negative integer d. There is a constant $M > 0$ such that for all $z \in C$*

$$\log |E_d(z)| \leq M|z|^{d+1}$$

$$\log |E_d(z)| \leq M|z|^d$$

Proof: The calculation is easiest when $2 < |z|$. By definition, $E_d(z) = (1-z)e^{\left(z+\frac{z^2}{2}+\ldots+\frac{z^d}{d}\right)}$. and so

$$\log |E_d(z)| \leq \log |1-z| + |z| + \frac{|z|^2}{2} + \ldots + \frac{|z|^d}{d}$$

This expression is clearly bounded by $C|z|^d$ for some constant C. Since $1 < |z|$, we also have $\log |E_d(z)| \leq C|z|^{d+1}$.

Next we consider $|z| < \frac{1}{2}$. Since $|z| < 1$, $\log(1-z) = -\left(z + \frac{z^2}{2} + \ldots\right)$ and $E_d(z) = e^{-\sum_{k>d} \frac{z^k}{k}}$ and so

$$\log |E_d(z)| = -Re\left(\frac{z^{d+1}}{d+1} + \frac{z^{d+2}}{d+2} + \ldots\right)$$

Since the real part of a complex number is at most its absolute value and the denominators in this series are at least 1, we have

$$\log|E_d(z)| \leq \left(|z|^{d+1} + |z|^{d+2} + ...\right) = |z|^{d+1}\left(1 + |z| + |z|^2 + ...\right)$$

But $|z| < \frac{1}{2}$, so

$$\log|E_d(z)| \leq 2|z|^{d+1}$$

Because $|z| < 1$, this also implies that

$$\log|E_d(z)| \leq 2|z|^d$$

Proof of Lemma, concluded: The final case is $\frac{1}{2} \leq |z| \leq 2$. Notice that $E_d(z)$ is defined everywhere and zero only at $z = 1$. Hence $\log|E_d(z)|$ is the log of a function which is real-valued and positive except at $z = 1$, and in particular is continuous except at $z = 1$. Moreover, as $z \to 1$, the term $|1 - z|$ approaches zero and the exponential term has a finite limit. Thus $\log|E_d(z)|$ approaches minus infinity at this point. Consequently, $\log|E_d(z)|$ is bounded above by some constant D on $\frac{1}{2} \leq |z| \leq 2$. Then trivially on this set

$$\log|E_d(z)| \leq C|z|^d$$

$$\log|E_d(z)| \leq C|z|^{d+1}$$

where, for instance, $C = 2^{d+1}\, D$.

Proof of Theorem, concluded: Finally we must estimate

$$\log\left|\prod E_d\left(\frac{z}{\lambda_i}\right)\right| = \sum \log\left|E_d\left(\frac{z}{\lambda_i}\right)\right| \leq \sum M\left|\frac{z}{\lambda_i}\right|^{d+1} = M\sum\left|\frac{1}{\lambda_i}\right|^{d+1}|z|^{d+1}$$

Recall that we are assuming $\sum\left|\frac{1}{\lambda_i}\right|^{d+1}$ converges. Thus M times this sum is a constant C and $\log\left|\prod E_d\left(\frac{z}{\lambda_i}\right)\right| \leq C|z|^{d+1}$ and thus

$$\left|\prod E_d\left(\frac{z}{\lambda_i}\right)\right| \leq e^{(C|z|^{d+1})}$$

Putting these three results together, we find that our product has order ρ where ρ is at most $\deg P(z)$ and at most $d + 1$.

18.7 Jensen's Formula

Now we want to prove that every entire function with finite order can be written as a Weierstrass product with finite genus. This is Hadamard's central result. The first crucial step is to obtain an integer d such that the sum below converges:

$$\sum \left| \frac{1}{\lambda_i} \right|^{d+1}$$

This step uses Jensen's formula, which we now describe. Suppose $f(z)$ is holomorphic on an open set which contains the closed disk $D(0,r)$ and suppose f has no zeros on this closed disk. Since an open neighborhood of the disk is simply connected, we can define $\log f(z)$ and apply Cauchy's formula to obtain

$$\log f(0) = \frac{1}{2\pi i} \int_{|z|=r} \frac{\log f(\zeta)}{\zeta} \, d\zeta$$

Parameterizing the path by $\gamma(t) = re^{it}$ for $0 \leq t \leq 2\pi$ gives

$$\log f(0) = \frac{1}{2\pi i} \int_0^{2\pi} \frac{\log f(re^{i\theta})}{re^{i\theta}} ire^{i\theta} \, d\theta = \frac{1}{2\pi} \int_0^{2\pi} \log f(re^{i\theta}) \, d\theta$$

Taking the real parts of both sides gives a final result

$$\log |f(0)| = \frac{1}{2\pi} \int_0^{2\pi} \log |f(re^{i\theta})| \, d\theta$$

Notice that $\log |f|$ is the real part of $\log f$ and so harmonic; our formula is then just the mean-value property of harmonic functions.

Jensen's formula generalizes to the case where f has zeros inside the disk, and gives an extremely useful estimate on the number of such zeros in terms of the rate of growth of f.

Theorem 127 (Jensen's Formula)**.** *Suppose f is holomorphic on a neighborhood of the closed disk $D(0,r)$. Suppose f is not zero on the boundary of the disk or its center. If the remaining zeros in the disk are $\lambda_1, \ldots, \lambda_k$, each repeated according to its multiplicity, then*

$$\log |f(0)| + \sum \log \left(\frac{r}{|\lambda_i|} \right) = \frac{1}{2\pi} \int_0^{2\pi} \log |f(re^{i\theta})| \, d\theta$$

Proof: Recall that the automorphisms of the open unit disk have the form

$$z \to e^{i\theta} \frac{z - z_0}{1 - z\overline{z_0}}$$

where z_0 is a point inside the disk and θ is a fixed angle. This map sends boundary points of the disk to other boundary points, and sends z_0 to the origin.

If z_0 is a point in the interior of $D(0,r)$, define a map from the closed $D(0,r)$ to the closed unit disk by mapping z to $\frac{z}{r}$ and mapping z_0 to $\frac{z_0}{r}$ and following this by the automorphism of the unit disk which sends $\frac{z_0}{r}$ to the origin. Thus the map is

$$z \to \frac{\frac{z}{r} - \frac{z_0}{r}}{1 - \frac{z\overline{z_0}}{r^2}} = \frac{r(z-z_0)}{r^2 - z\overline{z_0}}$$

This map sends z_0 to zero, has no other zeros in the closed $D(0,r)$, and has absolute value 1 on the boundary of $D(0,r)$.

Let

$$g(z) = \frac{f(z)}{\prod\left(\frac{r(z-\lambda_i)}{r^2 - z\overline{\lambda_i}}\right)}$$

This map is holomorphic on an open set containing the close $D(0,r)$ and has no zeros on or inside this disk, because the zeros of the denominator exactly cancel out the zeros of the numerator. Moreover, $|g(z)| = |f(z)|$ on the boundary of $D(0,r)$ because the absolute value of the denominator on this boundary is one.

Apply the special case of Jensen's lemma to $\log|g(z)|$. It gives

$$\log|g(0)| = \frac{1}{2\pi}\int_0^{2\pi} \log|f(re^{i\theta})|\, d\theta$$

because the denominator has absolute value one on the boundary. Notice that

$$\log|g(0)| = \log|f(0)| - \sum \log\left|\frac{-r\lambda_i}{r^2}\right| = \log|f(0)| - \sum \log\left|\frac{\lambda_i}{r}\right| = \log|f(0)| + \sum \log\left|\frac{r}{\lambda_i}\right|$$

QED.

Corollary 29. *Suppose f is entire and $f(0) = 1$. Let $n(r)$ be the number of zeros in the disk $D(0,r)$, counting multiplicities. Let $M(r) = max|f(re^{i\theta}|$, that is, the maximum of $|f|$ on the boundary of $D(0,r)$. Suppose f has no zeros on the boundary of $B(0,2r)$. Then*

$$n(r)\log 2 \leq \log M(2r)$$

Proof: By Jensen's formula applied to f and $D(0,2r)$ we obtain

$$\sum \log\left|\frac{2r}{\lambda_i}\right| = \frac{1}{2\pi}\int_0^{2\pi} \log|f(2re^{i\theta}|\, d\theta \leq M(2r)$$

If $|\lambda_i| \leq 1$, then $\left|\frac{2r}{\lambda_i}\right| \geq 2$ and $\log\left|\frac{2r}{\lambda_i}\right| \geq \log 2$. Thus the sum of terms on the left with this property contribute at least $n(r)\log 2$ to the left side. Ignore all other terms, making the inequality even stronger. QED.

18.8 Finite Order Implies Finite Rank

Theorem 128. *Suppose $f(z)$ is an entire function with finite order ρ. Choose $d = \lfloor \rho \rfloor$. In other words, $d \leq \rho < d+1$. If λ_i are the non-zero roots of f, then*

$$\sum \left|\frac{1}{\lambda_i}\right|^{d+1}$$

converges.

Proof: Without loss of generality, we can suppose that $f(0) \neq 0$ and indeed $f(0) = 1$. For if f has a zero of order k at the origin, we can replace f with $\frac{f}{Cz^m}$ for an appropriate nonzero constant C. All the non-zero roots are unchanged and we will show that the order is still at most ρ. Indeed

$$\log\left|\frac{f(z)}{Cz^m}\right| = \log|f(z)| - \log|C| - m\log|z| \leq |z|^{\rho+\epsilon} - \log|C| - m\log|z|$$

and it is enough to show that this is at most $|z|^{\rho+2\epsilon}$ for large z, or

$$|z|^{\rho+\epsilon} - \log|C| - m\log|z| \leq |z|^{\rho+2\epsilon}$$

Dividing by $|z|^{\rho+\epsilon}$, we need to show that

$$1 + \frac{-\log|C| - m\log|z|}{|z|^{\rho+\epsilon}} \leq |z|^{\epsilon}$$

This is clear because the denominator of the quotient is at least $|z|^{\epsilon}$ and this will eventually swamp the numerator, so the left side goes to 1 for large $|z|$ while the right side grows indefinitely.

If $M(r)$ is the maximum value of $|f|$ on the circle of radius r and $n(r)$ is the number of zeros of f in this circle, we proved the $n(r)\log 2 \leq \log M(2r)$ and thus

$$n(r) \leq \frac{\log M(2r)}{\log 2}$$

Since f has order ρ, we have $\log M(2r) < (2r)^{\rho+\epsilon} = 2^{\rho+\epsilon} r^{\rho+\epsilon}$ for large enough r, and so

$$n(r) \leq \frac{2^{\rho+\epsilon}}{\log 2}\, r^{\rho+\epsilon}$$

We are only interested in this inequality for small ϵ, so we can assume that $\epsilon < 1$ and then the first factor on the right is at most $\frac{2^{\rho+1}}{\log 2}$, a number that does not depend on r or ϵ. So for large enough r, this constant is smaller than r^{ϵ} and $n(r) \leq r^{\rho+2\epsilon}$. Since 2ϵ is another arbitrarily small number, we conclude that for any $\epsilon > 0$, we can find R such that $r > R$ implies

$$n(r) \leq r^{\rho+\epsilon}$$

Arrange the roots so $0 < |\lambda_1| \leq |\lambda_2| \leq |\lambda_3| \leq \ldots$ and notice that $k \leq n(|\lambda_k|)$. On the other hand, we have $n(|\lambda_k|) \leq |\lambda_k|^{\rho+\epsilon}$ for all large λ_k by the previous result. So for all large k,

$$k \leq |\lambda_k|^{\rho+\epsilon}$$

Recall that $d \leq \rho < (d+1)$. So when $\rho + \epsilon < (d+1)$ and k is large and $|\lambda_k| > 1$ we will have $k \leq |\lambda_k|^{d+1}$ and so

$$\left|\frac{1}{\lambda_k}\right|^{d+1} \leq \frac{1}{k}$$

We want the sum of the terms on the left to converge, which would be true if the sum of the terms on the right converged. So we are almost, but not quite, in business.

To finish, we must use the inequality $\rho + \epsilon < d + 1$ in a more sophisticated way. Start with the inequality

$$k \leq |\lambda_k|^{\rho+\epsilon}$$

and raise both sides to the power $\frac{d+1}{\rho+\epsilon}$. Raising positive real numbers to any positive power preserves inequalities, so we have

$$k^{\frac{d+1}{\rho+\epsilon}} \leq |\lambda_k|^{d+1}$$

and so for all large k

$$\left|\frac{1}{\lambda_k}\right|^{d+1} \leq \left(\frac{1}{k}\right)^{\frac{d+1}{\rho+\epsilon}}$$

Since $1 < \frac{d+1}{\rho+\epsilon}$, the sum of the terms on the right converges and thus the sum of the terms on the left converges.

18.9 The Hadamard Factorization Theorem

Theorem 129. *Let $f(z)$ be an entire function with finite rank ρ. Let $d = \lfloor \rho \rfloor$, so $d \leq \rho < d+1$. If λ_i are the nonzero roots of f, then $\sum \left|\frac{1}{\lambda_i}\right|^{d+1}$ converges. Therefore $P(z) = \prod E_d\left(\frac{z}{\lambda_i}\right)$ converges to an entire function with exactly the same nonzero roots as f, and we can write*

$$f(z) = e^{g(z)} z^m \prod E_d\left(\frac{z}{\lambda_i}\right)$$

Moreover, $g(z)$ is a polynomial of degree at most ρ.

Proof: It suffices to assume that $f(0) \neq 0$, for otherwise there is an m such that $\frac{f(z)}{z^m}$ is holomorphic and takes a non-zero value at the origin, and easily f and $\frac{f}{z^m}$ have the same order and the same canonical product and $f = e^g z^m \prod E_d$ if and only if $\frac{f}{z^m} = e^g \prod E_d$.

We need only prove that g is a polynomial of degree at most ρ. But unhappily, this is the most difficult step. On the internet many different proofs can be found, some incomplete. It is quite easy to prove that $g(z)$ is a polynomial if we know that $e^{g(z)}$ has finite order. The problem is that we define

$$e^{g(z)} = \frac{f(z)}{z^m \prod E_d\left(\frac{z}{\lambda_i}\right)}$$

so showing that e^g has finite order involves finding a *lower* bound for $\prod E_d$, and yet this product has zeros. These zeros are cancelled by zeros of f, of course, but a mere lower bound on the product will not work.

We'll take a different approach. We will obtain a convenient formula for $g(z)$. Differentiating this formula many times gives similar formulas for the various derivatives of g. If we write the formula for the pth derivative of g where $p > \rho$, we find that this pth derivative is zero, and thus g is a polynomial. This approach ultimately uses the fact that f has finite order, which we know, rather than the fact that e^g has finite order, which we will not know until we prove that $g(z)$ is a polynomial.

Remark: When we are faced with a product and want to differentiate, it is useful to compute the logarithmetic derivative

$$\frac{d}{dz} \log f(z) = \frac{f'(z)}{f(z)}$$

because the derivative of the product is then the sum of the derivatives of the terms. For example, the logarithmetic derivative of fg is

$$\frac{f'g + fg'}{fg} = \frac{f'}{f} + \frac{g'}{g}$$

The logarithmetic derivative of $E_d(z) = (1-z)e^{z+\frac{z^2}{2}+\ldots+\frac{z^d}{d}}$ is therefore

$$-\frac{1}{1-z} + \left(1 + z + z^2 + \ldots + z^{d-1}\right)$$

and the logarithmetic derivative of $E_d\left(\frac{z}{\lambda}\right)$ is

$$-\frac{1}{\lambda - z} + \frac{1}{\lambda} + \frac{z}{\lambda^2} + \frac{z^2}{\lambda^3} + \ldots + \frac{z^{d-1}}{\lambda^d}$$

The logarithmetic derivative of $\prod E_d\left(\frac{z}{\lambda_i}\right)$ is

$$\sum_{i=1}^{\infty}\left(-\frac{1}{\lambda_i - z} + \frac{1}{\lambda_i} + \frac{z}{\lambda_i^2} + \frac{z^2}{\lambda_i^3} + \ldots + \frac{z^{d-1}}{\lambda_i^d}\right)$$

We get a beautiful formula by differentiating this expression d times:

$$\frac{d^d}{dz^d}\left(\frac{\frac{d}{dz}\prod E_d\left(\frac{z}{\lambda_i}\right)}{\prod E_d\left(\frac{z}{\lambda_i}\right)}\right) = -\,d!\sum_{i=1}^{\infty}\frac{1}{(\lambda_i - d)^{d+1}}$$

Since $e^g = f\left(\prod E_d\right)^{-1}$, taking logarithmetic derivaties of both sides gives $g' = \frac{f'}{f} - \sum \frac{E_d'}{E_d}$ and so

$$g^{(d+1)} = \frac{d^d}{d^{d+1}}\left(\frac{f'}{f}\right) + d!\sum_{i=1}^{\infty}\frac{1}{(\lambda_i - d)^{d+1}}$$

This ends the easy algebra. To finish the proof, we need only show that if $f(z)$ is entire of order ρ and $d \leq \rho < (d+1)$, then

$$\frac{d^d}{d^{d+1}}\left(\frac{f'}{f}\right) = -\,d!\sum_{i=1}^{\infty}\frac{1}{(\lambda_i - d)^{d+1}}$$

The demonstration that this is true is ingenious, and starts with Poisson's formula for harmonic functions:

Theorem 130 (Poisson's Formula)**.** *Let $u(z)$ be a real-valued harmonic function on an open set which contains the close disk $D(0,R)$. Then u inside the disk is completely determined by its values on the bounding circle of the disk, by the following formula:*

$$u(re^{i\theta}) = \frac{1}{2\pi}\int_0^{2\pi}\left(\frac{R^2 - r^2}{R^2 - 2Rr\cos(\phi - \theta) + r^2}\right)u(Re^{i\phi})\,d\phi$$

Proof: Find $v(z)$ so $f(z) = u(z) + iv(z)$ is holomorphic on an open set containing $D(0,R)$. To do this, solve the Cauchy-Riemann equations

$$\frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}$$

$$\frac{\partial v}{\partial y} = \frac{\partial u}{\partial x}$$

By standard advanced calculus, these equations can be solved on a simply-connected open set because the integrability conditions below are satisfied, since u is harmonic.

$$\frac{\partial}{\partial y}\left(-\frac{\partial u}{\partial y}\right) = \frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\right)$$

Now apply Cauchy's theorem to f where γ is the counterclockwise circle around the boundary of $B(0,R)$ and z is inside this circle, and thus $\frac{R^2}{\overline{z}}$ is outside the circle:

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

From here, the proof is easy algebra. Let $z = re^{i\theta}$ and $\zeta = Re^{i\phi}$ and note that

$$\left(\frac{1}{\zeta - z} - \frac{1}{\zeta - \frac{R^2}{\overline{z}}}\right) d\zeta = \left(\frac{z - \frac{R^2}{\overline{z}}}{\zeta^2 - \zeta z - \zeta\frac{R^2}{\overline{z}} + \frac{zR^2}{\overline{z}}}\right) d\zeta = \left(\frac{z\overline{z} - R^2}{\zeta^2\overline{z} - \zeta z\overline{z} - \zeta R^2 + zR^2}\right) =$$

$$(\text{since } \zeta\overline{\zeta} = R^2) \quad \left(\frac{|z|^2 - R^2}{\zeta\overline{z} + \overline{\zeta}z - (|z|^2 + R^2)}\right)\frac{d\zeta}{\zeta} = \left(\frac{R^2 - |z|^2}{|z|^2 - 2\Re(\overline{z}\zeta) + R^2}\right)\frac{d\zeta}{\zeta} =$$

$$\left(\frac{R^2 - r^2}{r^2 - 2Rr\cos(\phi - \theta) + R^2}\right)\frac{iRe^{i\phi}d\phi}{Re^{i\phi}} = i\left(\frac{R^2 - r^2}{r^2 - 2Rr\cos(\phi - \theta) + R^2}\right)d\phi$$

So

$$f(re^{i\theta}) = \frac{1}{2\pi}\int_0^{2\pi} f(Re^{i\phi})\left(\frac{R^2 - r^2}{r^2 - 2Rr\cos(\phi - \theta) + R^2}\right)d\phi$$

In this last expression, everything is real except f on the two sides of the equation, so taking the real part of each side gives Poisson's Formula. QED.

Remark: For both holomorphic and harmonic functions, the values inside a disk are determined by the values on the boundary of the disk. But there is a crucial difference. If we start with an arbitrary continuous function on the boundary of a disk, Poisson's formula defines a

function inside this disk. It turns out that this function is C^∞ and harmonic inside the disk, and converges to the given function on the boundary. So Poisson's formula solves the boundary value problem for the Laplace partial differential equation and becomes the first key result in the theory of harmonic functions. Luckily, we do not need to know this result.

Remark: Our proof of Jensen's formula started with the Cauchy Integral Formula. We obtain a different useful result by using the same technique starting with Poisson's formula.

Theorem 131 (Poisson-Jensen)**.** *Suppose $f(z)$ is a holomorphic function on an open neighborhood of the closed disk $D(0,R)$, and suppose $f(0) = 1$ and f has zeros $\lambda_1, \ldots, \lambda_n$ inside the disk and no zeros on the boundary. Then for all $z = re^{i\theta}$ inside the disk,*

$$\log|f(z)| = -\sum_{k=1}^{n} \log\left|\frac{R^2 - \overline{\lambda_k} z}{R(z-\lambda_k)}\right| + \frac{1}{2\pi}\int_0^{2\pi} \frac{R^2 - r^2}{R^2 - 2Rr\cos(\theta-\phi) + r^2} \log|f(Re^{i\phi})| d\phi$$

Remark: Since f has zeros inside the disk, the left side is sometimes undefined. But near these points, $|f(z)|$ is almost zero and $\log|f(z)|$ is almost minus ∞. Moreover, the first term on the side right is also almost minus ∞ close to these points. So in a sense the formula is also true at the zeros of f.

Proof: As in the proof of Jensen's formula, define

$$h(z) = f(z)\prod_{k=1}^{n} \frac{R^2 - \overline{\lambda_k} z}{R(z-\lambda_k)}$$

The extra term cancels the zeros of f, so $h(z)$ has no zeros inside or on the disk. Moreover, as before the extra term has absolute value 1 on the boundary of the disk. Notice that h is holomorphic so we can form a holomorphic function $\log h(z)$. The real part of this function, $\log|h(z)|$, is harmonic. Apply Poisson's formula to $\log|h(z)|$. The integral on the right side of Poisson's formula is exactly the integral in the Poisson-Jensen theorem because the absolute value of our additional factor is one on the boundary of the disk. The rest of the formula follows because

$$\log|h(z)| = \log|f(z)| + \sum \log\left|\frac{R^2 - \overline{\lambda_k} z}{R(z-\lambda_k)}\right|$$

Continuation of the Proof of Hadamard's Theorem: We now construct a formula for $\frac{d^d}{d^{d+1}}\left(\frac{f'}{f}\right)$. Notice first that if $f(z) = u(z) + iv(z)$ is any holomorphic function, we can compute $\frac{df}{dz}$ knowing only u. Indeed

$$\frac{df}{dz} = \frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x} = \frac{\partial u}{\partial x} - i\frac{\partial u}{\partial y}$$

The complex derivative of $\log f(z)$ is $\frac{f'(z)}{f(z)}$ and the real part of $\log f(z)$ is $\log|f(z)|$, so

$$\frac{f'}{f} = \frac{\partial}{\partial x}\log|f(z)| - i\frac{\partial}{\partial y}\log|f(z)|$$

We now compute each of these partials by differentiating the Poisson-Jensen formula for $\log|f(z)|$:

$$\frac{f'}{f} = \left(\frac{\partial}{\partial x} - i\frac{\partial}{\partial y}\right)\left(-\sum_{k=1}^{n}\log\left|\frac{R^2-\overline{\lambda_k}z}{R(z-\lambda_k)}\right| + \frac{1}{2\pi}\int_0^{2\pi}\frac{R^2-r^2}{R^2-2Rr\cos(\theta-\phi)+r^2}\log|f(Re^{i\phi})|d\phi\right)$$

These partials applied to each term of the sum give the logarithmetic derivative of the terms without absolute values by reversing the previous argument, so this term is

$$\sum_{k=1}^{n}\frac{\overline{\lambda_k}}{R^2-\overline{\lambda_k}z} + \sum_{k=1}^{n}\frac{R}{R(z-\lambda_k)}$$

which we prefer to write

$$\sum_{k=1}^{n}\frac{1}{z-\lambda_k} + \sum_{k=1}^{n}\frac{\overline{\lambda_k}}{R^2-\overline{\lambda_k}z}$$

Proof Continued: It is useful to rewrite the Poisson kernel as $\mathfrak{R}\left(\frac{\zeta+z}{\zeta-z}\right)$. Indeed,

$$\mathfrak{R}\left(\frac{\zeta+z}{\zeta-z}\right) = \frac{1}{2}\left(\frac{\zeta+z}{\zeta-z}+\frac{\overline{\zeta}+\overline{z}}{\overline{\zeta}-\overline{z}}\right) = \frac{1}{2}\left(\frac{2\zeta\overline{\zeta}+2z\overline{z}+(z\overline{\zeta}-\zeta\overline{z})+(\zeta\overline{z}-z\overline{\zeta})}{\zeta\overline{\zeta}-(z\overline{\zeta}+\zeta\overline{z})+z\overline{z}}\right)$$

$$= \left(\frac{R^2-r^2}{R^2-2Rr\cos(\theta-\phi)+r^2}\right)$$

This is the only term in the integral which depends on z, and when we compute $\left(\frac{\partial}{\partial x} - i\frac{\partial}{\partial y}\right)$ of it, we just get the derivative of $\frac{\zeta+z}{\zeta-z}$.

Putting everything together, we get

$$\frac{f(z)'}{f(z)} = -\sum_{k=1}^{n}\frac{1}{\lambda_k-z} + \sum_{k=1}^{n}\frac{\overline{\lambda_k}}{R^2-\overline{\lambda_k}z} + \frac{1}{2\pi}\int_0^{2\pi}\frac{2Re^{i\phi}}{\left(Re^{i\phi}-z\right)^2}\log|f(Re^{i\phi})|d\phi$$

Differentiating this d times gives

$$\frac{d^d}{dz^z}\left(\frac{f'(z)}{f(z)}\right) = -\,d!\sum_{k=1}^{n}\frac{1}{(\lambda_k - z)^{d+1}} + d!\sum_{k=1}^{n}\frac{(\overline{\lambda}_k)^{d+1}}{(R^2 - \overline{\lambda}_k z)^{d+1}}$$

$$+(d+1)!\frac{1}{2\pi}\int_0^{2\pi}\frac{2Re^{i\phi}}{\left(Re^{i\phi}-z\right)^{d+2}}\log|f(Re^{i\phi})|d\phi$$

Proof Concluded: Finally, we let $R \to \infty$. The first term of the above sum only depends on the roots of f inside $D(0,R)$, so it changes by summing over more roots. In the limit, this term converges to exactly the result we need to complete the proof. To finish the argument, we must use the hypothesis that f has order ρ to show that the last two terms converge to zero.

Consider the second sum. Suppose $2|z| < R$. Then if $|\lambda_k| < R$ we have $|R^2 - \overline{\lambda_k} z| \geq \frac{R^2}{2}$ and

$$\left|\left(\frac{\overline{\lambda_k}}{R^2 - \overline{\lambda_k} z}\right)\right|^{d+1} \leq \left(\frac{2}{R}\right)^{d+1}$$

So if $n(R)$ is the number of roots inside $D(0,R)$, the second term is bounded by

$$d!\left(\frac{2}{R}\right)^{d+1} n(R)$$

As a corollary to Jensen's formula, we proved that $n(R)\log 2 \leq \log M(2R)$. In turn, this is at most $(2R)^{\rho+\epsilon}$ for large enough R. So the second term is bounded by

$$d!\left(\frac{2}{R}\right)^{d+1}\frac{1}{\log 2}(2R)^{\rho+\epsilon}$$

This is a constant times

$$R^{\rho+\epsilon-d-1}$$

Since $\rho < (d+1)$, the exponent is negative for ϵ small enough, and for such ϵ the term goes to zero as $R \to \infty$.

Finally we handle the integral term. Notice that

$$\int_0^{2\pi}\frac{Re^{i\phi}}{(Re^{i\phi}-z)^{d+2}}\,d\phi = (-i)\int_\gamma \frac{d\zeta}{(\zeta - z)^{d+2}}d\zeta = 0$$

because the integrand is the derivative of another function over the entire path of integration. Consequently we can replace $\log|f|$ by $\log|f| - \log M(R)$ where $M(R)$ is the maximum value of $\log|f|$ on the boundary circle. We want to estimate

$$(d+1)!\frac{1}{2\pi}\int_0^{2\pi}\frac{2Re^{i\phi}}{\left(Re^{i\phi}-z\right)^{d+2}}\left(\log|f(Re^{i\phi})| - \log M(R)\right)d\phi$$

If $|z| < \frac{R}{2}$, we have

$$\left| \frac{2Re^{i\phi}}{(Re^{i\phi} - z)^{d+2}} \right| < \frac{2R}{\left(\frac{R}{2}\right)^{d+2}}$$

Since $M(R) \geq |f(Re^{i\phi}|$, $\log M(R) \geq \log |f|$ and $M(R) - \log |f| \geq 0$, so the integral is at most

$$(d+1)! \frac{1}{2\pi} \int_0^{2\pi} \frac{2^{d+3}}{R^{d+1}} \left(\log M(R) - \log |f(Re^{i\phi})| \right) d\phi =$$

$$(d+1)! \frac{1}{2\pi} \frac{2^{d+3}}{R^{d+1}} \Big(\frac{1}{2\pi} \int_0^{2\pi} \log M(R) d\phi - \frac{1}{2\pi} \int_0^{2\pi} \log |f(Re^{i\phi})| d\phi \Big)$$

By Jensen's formula, and using $f(0) = 1$,

$$\frac{1}{2\pi} \int_0^{2\pi} \log |Re^{i\phi}| d\phi = \sum \log \frac{R}{|\lambda_k|}$$

The term $\frac{R}{|\lambda_k|} > 1$ because $|\lambda_k| < R$, so $\log \frac{R}{|\lambda_k|} > 0$ and

$$\frac{1}{2\pi} \int_0^{2\pi} \log |f(Re^{i\phi})| d\phi > 0$$

Therefore the integral is bounded by

$$(d+1)! \frac{1}{2\pi} \frac{2^{d+3}}{R^{d+1}} \log M(R) \leq (d+1)! \frac{1}{2\pi} \frac{2^{d+3}}{R^{d+1}} R^{\rho+\epsilon}$$

This is a constant times $R^{\rho+\epsilon-d-1}$ and since $\rho - d - 1$ is negative, the entire exponent is negative for small enough ϵ and then the expression goes to zero as $R \to \infty$. QED.

Corollary 30. *Suppose f is an entire function with finite order that is not an integer. Then f has infinitely meny zeros.*

Proof: If f had finitely many zeros λ_i, then $h(z) = \frac{f(z)}{\prod (z - \lambda_i)}$ would be an entire function without zeros. Easy estimates show that h would still have order ρ. By Hadamard's theorem, $g(z)$ is a polynomial, in that case $h(z) = e^{g(z)}$ would have order $\deg g$, which is an integer. This is a contradiction.

Chapter 19

More on the Zeta Function

19.1 Order of the Zeta Function

Recall that the Zeta Function $\zeta(z)$ is meromorphic on the entire plane. It has a pole of order 1 at $z = 1$ and no other poles. It has simple zeros of order 1 at $z = -2, -4, -6, ...$; all other zeros are in the strip $0 < \Re z < 1$. This function satisfies the functional equation

$$\zeta(z) = 2^z \pi^{z-1} \sin\left(\frac{\pi z}{2}\right) \Gamma(1-z)\zeta(1-z)$$

Previously we rewrote this equation using properties of $\Gamma(z)$ and $\sin z$:

$$\pi^{-z/2}\Gamma\left(\frac{z}{2}\right)\zeta(z) = \pi^{-(1-z)/2}\Gamma\left(\frac{1-z}{2}\right)\zeta(1-z)$$

The new function is unchanged when z is replaced by $1 - z$. The function $\Gamma(z/2)$ has simple poles at $0, -2, -4, ...$ because $\Gamma(z)$ has simple poles at the non-positive integers. The poles of Γ at negative even integers cancel the zeros of $\zeta(z)$ there, so the above function only has zeros in the critical strip. It has simple poles at $z = 0, 1$ and no other poles.

This caused Riemann to introduce the function

$$z(1-z)\,\pi^{-z/2}\Gamma\left(\frac{z}{2}\right)\zeta(z)$$

which is entire, is unchanged if z is replaced by $1 - z$, and has zeros only in the critical strip.

Theorem 132. *The function* $f(z) = z(1-z)\,\pi^{-z/2}\Gamma\left(\frac{z}{2}\right)\zeta(z)$ *has order exactly* 1.

Lemma 14. *The order of $f(z)$ is at least* 1.

Proof: We will prove this by looking at the growth of $f(2n)$ as $n \to \infty$. Since

$$\zeta(x) = 1 + \frac{1}{2^x} + \frac{1}{3^x} + ...$$

ζ is a decreasing function of x. Thus for $2 < x$ we have

$$1 \leq \zeta(x) \leq \frac{\pi^2}{6} < 2$$

So we can ignore ζ. The remaining terms at $2n$ can be approximated using Stirling's formula:

$$|2n(1-2n)\pi^{-n}\Gamma(n)| \sim |2n(1-2n)|\,\pi^{-n}\left(\frac{n}{e}\right)^n\sqrt{\frac{2\pi}{n}}$$

This is larger than $\pi^{-n}\left(\frac{n}{e}\right)^n n^{-1/2}$. We want to show that this expression grows faster than $e^{(2n)}$ and thus that its logarithm is eventually larger than $2n$. But this logarithm is

$$-n\log\pi + n\log n - n - \frac{1}{2}\log n = n\Big(\log n - \log\pi - 2\Big) + \Big(n - \frac{1}{2}\log n\Big)$$

and eventually $\log n - \log\pi - 2$ is greater than 2 and $n - \frac{1}{2}\log n$ is greater than zero. QED.

Lemma 15. *The order of $f(z)$ is at most* 1.

Proof: By invariance when z is replaced by $1-z$, we only have to consider z with $\Re(z) \geq 1/2$. It suffices to show that each of the terms in f has order at most 1. This is trivial for $z(1-z)$. Notice, however, that this term cancels out the pole of $\Gamma\left(\frac{z}{2}\right)$ at $z = 0$, so we can ignore the poles of Γ in $\Re(z) \geq 1/2$ and only study growth for large $|z|$.

We have $|\pi^{-z/2}| = |e^{-z/2\log\pi}| \leq e^{|z|\frac{\log\pi}{2}} \leq e^{|z|}$

Thus it suffices to study the growth of the gamma function and the zeta function. We begin with the Gamma function. By Stirling's formula for the complex gamma function, when $\Re(z) > 0$ we have

$$\log\Gamma(z) = \left(z - \frac{1}{2}\right)\log z - z + \frac{1}{2}\log(2\pi) + O\left(|z|^{-1}\right)$$

The singularity near the origin has been cancelled by the term $z(1-z)$, so everything is bounded for small $|z|$. Thus we can assume that $\Re(z) > 0$ and $|z| > 1$. It suffices to show that

$$|\log\Gamma(z)| \leq |z|^{1+\epsilon}$$

and this will follow if for all large $|z|$ we have

$$\left|\left(z - \frac{1}{2}\right)\log z - z\right| < |z|^{1+\epsilon}$$

But applying very gross estimates so we almost give away the farm gives

$$\left|\left(z-\frac{1}{2}\right)\log z - z\right| \leq 2|z|(\log|z|+\pi)+|z| \leq |z|(2\log|z|+2\pi+1)$$

and if ϵ is given in advance, then $(2\log|z|+2\pi+1) < |z|^{\epsilon}$ for sufficiently large $|z|$.

To complete the proof, it suffices to control the growth of $\zeta(z)$. Very early in these notes, we included a picture of the absolute value of the zeta function. Here is that picture again:

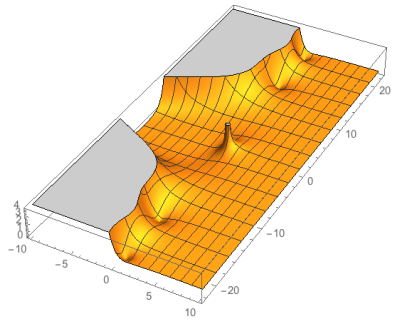


Figure 19.1: $f(z) = \zeta(z)$

Notice that aside from the pole at $z = 1$, the absolute value of the function is actually very small everywhere on $\Re(z) \geq \frac{1}{2}$. We only need to estimate $|\zeta|$ on this set. This suggests that if we play our cards correctly, the proof should be very easy. Keep that in mind while reading the following argument.

The pole of ζ has been cancelled by the term $(1-z)$, so f is bounded on any disc $|z| < B$ and it suffices to assume $\Re(z) > 1/2$ and $|z-1| > 1$. We'll use an argument which analytically continues the zeta function to $\Re(z) > 0$ using an integral formula which can be easily estimated. We start by assuming $\Re(z) > 1$ and writing

$$\zeta(z) = 1+\frac{1}{2^z}+\frac{1}{3^z}+\frac{1}{4^z}+\ldots = \left(\frac{1}{1^z}-\frac{1}{2^z}\right)+2\left(\frac{1}{2^z}-\frac{1}{3^z}\right)+3\left(\frac{1}{3^z}-\frac{1}{4^z}\right)+\ldots =$$

$$z\sum_{n=1}^{\infty} n\int_n^{n+1}\frac{1}{x^{z+1}}\,dx = z\sum_{n=1}^{\infty}\int_n^{n+1}\lfloor x\rfloor\frac{1}{x^{z+1}}\,dx = z\int_1^{\infty}\lfloor x\rfloor\frac{1}{x^{z+1}}\,dx$$

Here $\lfloor x\rfloor$ is the largest integer smaller than x. Notice that

$$z\int_1^{\infty} x\frac{1}{x^{z+1}}\,dx = z\left.\frac{x^{-z+1}}{-z+1}\right|_1^{\infty} = \frac{z}{z-1}$$

So

$$\zeta(z) = \frac{z}{z-1} - z\int_1^{\infty}(x-\lfloor x\rfloor)\frac{1}{x^{z+1}}\,dx$$

Here $(x-\lfloor x \rfloor)$ gives the *fractional part of* x and thus is a sawtooth function which varies between 0 and 1. It immediately follows that the final integral converges. Consequently all previous steps involved convergent series.

The final formula has several significant features. First, the pole at $z = 1$ has been isolated out as the term $\frac{z}{z-1}$ and is no longer part of the integral. Second, the integral actually converges absolutely for $\Re(z) > 0$, because the absolute value of the integrand is bounded by $x^{-\Re(z)-1}$ and the integral of this expression is

$$\frac{1}{x^{\Re(z)}}\bigg|_1^\infty$$

which is finite.

Lemma 16. *The integral converges absolutely and uniformly on* $\Re(z) > \frac{1}{4}$ *and thus defines a holomorphic function on this set. Since the entire expression including* $\frac{z}{z-1}$ *agrees with* $\zeta(z)$ *when* $\Re(z) > 1$*, it is the zeta function on* $\Re(z) > \frac{1}{4}$.

Proof: The integral from 1 to N is a sum of finitely many integrals of the form $\int_n^{n+1}(x-n)\frac{1}{x^{z+1}}\,dx$. Each of these expressions is holomorphic in z, so the integral from 1 to N is holomorphic in z. It suffices to prove that it converges uniformly to the integral from 1 to ∞. The difference between the finite integral and the infinite integral is

$$\int_N^\infty (x - \lfloor x \rfloor)\frac{1}{x^{z+1}}\,dx$$

and the absolute value of this expression is bounded by

$$\int_N^\infty \frac{1}{x^{\Re(z)+1}}\,dx = -\frac{1}{\Re(z)}\,\frac{1}{x^{\Re(z)}}\bigg|_N^\infty = \frac{1}{\Re(z)}\,\frac{1}{N^{\Re(z)}}$$

Since $\Re(z) > \frac{1}{4}$, the first term is bounded by 4 and the second term is bounded by $\frac{1}{\sqrt[4]{N}}$. Therefore this error goes to zero uniformly over the entire domain $\Re(z) > \frac{1}{4}$ and we are done.

Lemma 17. *The expression* $(1-z)\zeta(z)$ *has order of growth at most* 1 *on* $\Re(z) \geq \frac{1}{2}$.

Proof: Since $(1-z)$ is a factor of $f(z)$, this estimate suffices for us. Note that

$$(1-z)\zeta(z) = -z + z(1-z)\int_1^\infty (x - \lfloor x \rfloor)\,\frac{1}{x^{z+1}}\,dx$$

The absolute value of the integral is bounded by

$$\int_1^\infty x^{-\Re(z)-1}\,dx = \frac{x^{-\Re(z)}}{-\Re(z)}\bigg|_1^\infty = \frac{1}{\Re(z)} \leq 2$$

Consequently, the order of growth is far smaller than 1 on this set; indeed the absolute value of our expression is bounded by a constant times $|z|^2$. QED.

Completion of the proof: Finally, we prove that our integral defines a holomorphic function on $\Re(z) > 0$. Notice that we only need this for $\Re(z) > \frac{1}{4}$ because we only care about order on $\Re(z) \geq \frac{1}{2}$. We will restrict to this set because it will make the argument easier. Indeed, now the integral converges uniformly on the entire space.

19.2 Infinitely Many Zeros in the Critical Strip

Theorem 133. *The Riemann Zeta Function has infinitely many zeros in the critical strip.*

Proof: Let $f(z) = z(1-z)\,\pi^{-z/2}\Gamma\left(\frac{z}{2}\right)\zeta(z)$ and let $g(z) = f\left(\frac{1}{2} - z\right)$. Now f is entire and its only zeros are in the critical strip and $f(z) = f(1-z)$. So $g(z)$ is entire and its only zeros are in the strip $-\frac{1}{2} < \Re z < \frac{}{2}$ and $g(-z) = g(z)$. So $g(z)$ can be given by a unique power series about the origin, and yet when we replace z by $-z$, the power series does not change. It follows that this series only has even terms, and thus $g(z) = h(z^2)$ for an entire function h.

Since f has order 1, g also has order 1 and h has order $\frac{1}{2}$. This is not an integer, so h has infinitely many zeros, and therefore g and f have infinitely many zeros.

19.3 The Product Formula

Theorem 134. *The zeta function is given by a product formula*

$$z(z-1)\,\pi^{-z/2}\Gamma\left(\frac{z}{2}\right)\zeta(z) = e^{Az}\prod\left(1-\frac{z}{\rho}\right)e^{\frac{z}{\rho}}$$

over the zeros ρ in the critical strip, where A is some complex constant.

Proof: This is a consequence of the Hadamard Factorization Theorem, since this function is the negative of $z(1-z)\,\pi^{-z/2}\Gamma\left(\frac{z}{2}\right)\zeta(z)$, which has order one. We expect the first term to be e^{B+Az}, and only need to prove that $B = 0$. To do that, we substitute $z = 0$. Then e^B is $-z\Gamma\left(\frac{z}{2}\right)\zeta(0)$. But we proved earlier that $\zeta(0) = -\frac{1}{2}$, so

$$e^B = -\lim_{z\to 0} z\,\Gamma\left(\frac{z}{2}\right)\left(-\frac{1}{2}\right) = \lim_{z\to 0}\frac{z}{2}\,\Gamma\left(\frac{z}{2}\right) = \lim_{z\to 0}\Gamma\left(\frac{z}{2}+1\right) = 1$$

and $B = 0$. We will prove that $A = \log 2\sqrt{\pi} - 1 - \frac{\gamma}{2}$ in the next section.

19.4 A Product Formula for the Zeta Function

Using the previous result, we can find a product formula for the Zeta function. Since

$$z(z-1)\,\pi^{-z/2}\Gamma\left(\frac{z}{2}\right)\zeta(z) = e^{Az}\prod\left(1-\frac{z}{\rho}\right)e^{\frac{z}{\rho}}$$

we have

$$\zeta(z) = \frac{1}{z(z-1)}\pi^{z/2}\frac{1}{\Gamma\left(\frac{z}{2}\right)}e^{Az}\prod\left(1-\frac{z}{\rho}\right)e^{z/\rho}$$

However, we earlier proved that

$$\frac{1}{\Gamma(z)} = G(z) = z\,e^{\gamma z}\prod_{n>0}\left(1+\frac{z}{n}\right)e^{-\frac{z}{n}}$$

So

$$\zeta(z) = \frac{1}{z(z-1)}\pi^{z/2}\frac{z}{2}e^{\gamma z/2}\prod_{n>0}\left(1+\frac{z}{2n}\right)e^{-z/(2n)}e^{Az}\prod\left(1-\frac{z}{\rho}\right)e^{\frac{z}{\rho}}$$

and so

Theorem 135.

$$\zeta(z) = \frac{1}{z-1}\pi^{z/2}\frac{e^{Az+\frac{\gamma z}{2}}}{2}\prod_{n>0}\left(1+\frac{z}{2n}\right)e^{-z/(2n)}\prod\left(1-\frac{z}{\rho}\right)e^{\frac{z}{\rho}}$$

Remark: Notice that this product clearly shows the pole at 1, the trivial zeros at $-2, -4, -6, ...$ and the non-trivial zeros in the critical region.

Taking the logarithmetic derivative, we obtain

Theorem 136.

$$-\frac{\zeta'(z)}{\zeta(z)} = \frac{1}{z-1} - \frac{\ln\pi}{2} - A - \frac{\gamma}{2} - \sum_{n>0}\left(\frac{1}{z+2n}-\frac{1}{2n}\right) - \sum\left(\frac{1}{z-\rho}+\frac{1}{\rho}\right)$$

Remark: We proved that $\zeta(0) = -\frac{1}{2}$. We will prove that $\zeta'(0) = -\frac{1}{2}\log(2\pi)$ at the end of this chapter. Substitution in the above formula shows that $A = \log(2\sqrt{\pi}) - 1 - \frac{\gamma}{2}$.

Remark: On the other hand, if $\Re(z) > 1$, we have $\zeta(z) = \prod\frac{1}{1-\frac{1}{p^z}}$. We computed its logarithmetic derivative just before proving the prime number theorem. We repeat that calculation here.

$$\frac{\zeta'(z)}{\zeta(z)} = -\sum_p \frac{\ln p}{p^z\left(1-\frac{1}{p^z}\right)} = -\sum_p \ln p\left(\sum_{n\geq 1}\frac{1}{p^{nz}}\right)$$

If n is a positive integer, define

$$\Lambda(n) = \begin{cases} \ln p & \text{if } n \text{ is a power of } p \\ 0 & \text{otherwise} \end{cases}$$

Theorem 137. *If* $\Re(z) > 1$,

$$-\frac{\zeta'(z)}{\zeta(z)} = \sum_{n\geq 1}\frac{\Lambda(n)}{n^z}$$

Remark: A glance back to the section on the prime number theorem shows that we introduced the function

$$\Psi(x) = \sum_{n\leq x}\Lambda(n) = \sum_{p^m<x}\log p$$

and proved that the prime number theorem is equivalent to the result $\Psi(x) \sim x$.

Riemann's ultimate goal in his paper was not a proof of the prime number theorem, but rather a formula for the number of primes below a given limit. He found a rather complicated formula, which ultimately depended on the non-trivial zeros of the zeta function. But he only sketched a proof, and a rigorous proof was not found until 1895, in a paper by von Mangoldt. In this paper, von Mangoldt found a much easier formula of the same nature for $\Psi(x)$, and almost all later authors use that formula rather than Riemann's. Here is von Mangoldt's formula:

Theorem 138 (Riemann, von Mangoldt)**.** *The function* $\Psi(x)$ *is given by the formula*

$$\Psi(x) = x - \log(2\pi) - \sum\frac{x^\rho}{\rho} + \sum_{n\geq 1}\frac{x^{-2n}}{2n}$$

Remark: This is an astounding formula, because $\Psi(x)$ is a discontinuous function, Iit rises by a step at each power of a prime. It can only be a sum of continuous functions if the sum does not converge absolutely, and Riemann's sum doesn't. This makes it difficult to draw other conclusions from the sum, since the trick of rearranging the terms isn't available. But the result is so astonishing that we can admire it even when we cannot use it.

Notice that the initial term is x and the prime number theorem is equivalent to the assertion that $\Psi(x) \sim x$. So by that theorem, the first term dominates and the non-trivial zeros of the zeta function cause smaller irregularities in the distribution of the primes.

The last term comes from the trivial zeros of the zeta function. The function $\Psi(x)$ is counting numbers that are prime powers, so we use it when $x > 2$. In that case the last term is at most $\sum_{n\geq 1}\left(\frac{1}{2}\right)^n = 1$. So it is irrelevant, as are small constants. Therefore in essence the theorem says

$$\Psi(x) = x - \sum \frac{x^\rho}{\rho}$$

and implies that the irregularities in the primes are entirely due to the zeros of ζ in the critical strip.

Remark: Actually, the last term can be evaluated explicitly. This term is

$$\sum_{n\geq 1} \frac{x^{-2n}}{2n} = \left(\frac{1}{2x^2} + \frac{1}{4x^4} + \frac{1}{6x^6} + ...\right)$$

and its derivative is

$$-\left(\frac{1}{x^3} + \frac{1}{x^5} + \frac{1}{x^7} + ...\right) = -\frac{1}{x^3}\,\frac{1}{1-\frac{1}{x^2}} = -\,\frac{1}{x(x-1)(x+1)} = \frac{1}{x} - \frac{x}{x^2-1}$$

so up to a constant it equals $\log(x) - \frac{1}{2}\log(x^2-1) = \frac{1}{2}\left(\log x^2 - \log(x^2-1)\right)$ and thus

$$-\frac{1}{2}\log\left(1-\frac{1}{x^2}\right)$$

Once we know this result, we can prove it an easier way. Start with $\frac{1}{1-t} = 1 + t + t^2 + ...$ and integrate both sides from 0 to x to obtain $-\log(1-x) = x + \frac{x^2}{2} + \frac{x^3}{3} + ...$, which is valid when $|x| < 1$. Substitute $\frac{1}{x^2}$ for x to obtain the following when $1 < x^2$:

$$-\log\left(1-\frac{1}{x^2}\right) = \frac{1}{x^2} + \frac{1}{2}\frac{1}{x^4} + \frac{1}{3}\frac{1}{x^6} + ...$$

Divide both sides by 2 to obtain

$$\sum_{n\geq 1} \frac{x^{-2n}}{2n} = \frac{1}{2x^2} + \frac{1}{4x^4} + \frac{1}{6x^6} + ... = -\frac{1}{2}\log\left(1-\frac{1}{x^2}\right)$$

Remark: Recall that the critical zeros are symmetric about the line $x = \frac{1}{2}$ and about the line $y = 0$. We have been a little sloppy about that term, since the formula requires that we compute this sum by taking a limit of finite sums over symmetrical regions $0 < \Re(\rho) < 1, -T < \Im(\rho) < T$.

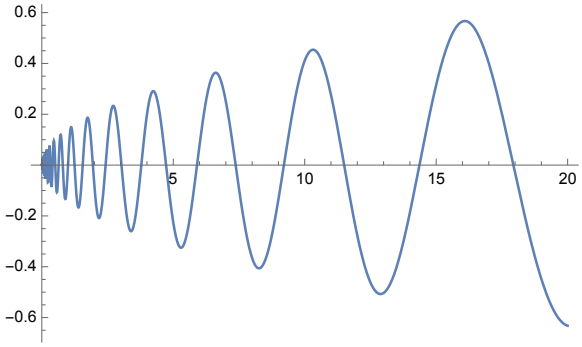
Thus zeros on the critical line count with their conjugate, and zeros off the critical line count with three other similar zeros. Let us determine the contribution of zeros on the critical line. Such a zero has the form $\rho = \frac{1}{2} + iy$ and the contribution is thus

$$\frac{x^\rho}{\rho} = \frac{\sqrt{x}x^{iy}}{\frac{1}{2}+iy} + \frac{\sqrt{x}x^{-iy}}{\frac{1}{2}-iy} = \sqrt{x}\, 2\Re\left(\frac{x^{iy}}{\frac{1}{2}+iy}\right) = \sqrt{x}\, 2\Re\left(\frac{\cos(y\log x)+i\sin(y\log x)}{\frac{1}{2}+iy}\right) =$$

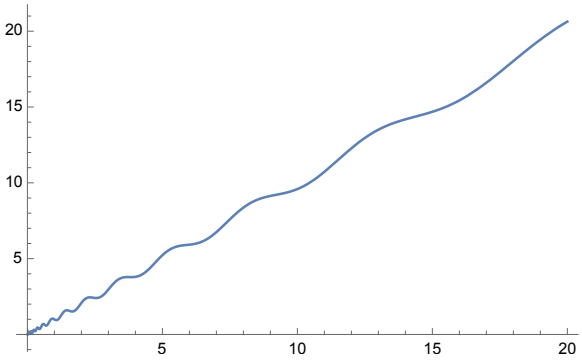
$$\sqrt{x}\,\frac{\cos(y\log x)+2y\sin(y\log x)}{\frac{1}{4}+y^2}$$

Thus the size of the contribution from non-trivial zeros is approximately $\sqrt{x}$ and their contribution oscillates, with an oscillation that is slower and slower as x increases. Riemann had a wonderful expression for this feature. He said the non-trivial zeros caused alternating thickening and thinning of the density of the primes.

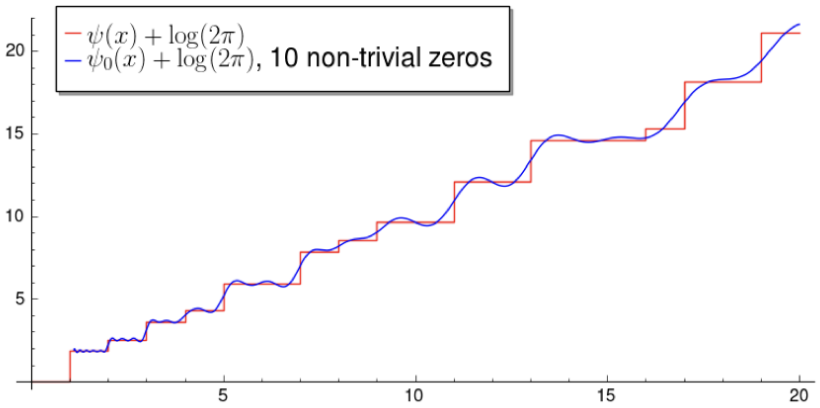
The first non-trivial zero of the zeta function is at $\frac{1}{2} + 14.1347i$. Plugging $y = 14.1347$ into the above formula and plotting gives

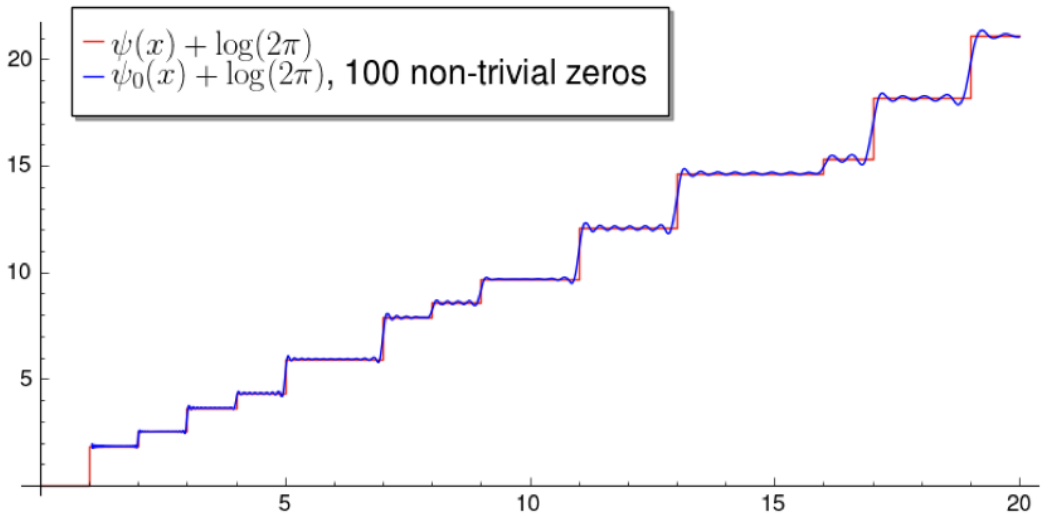


So the first approximation to $\Psi(x)$ looks like



The link http://ism.uqam.ca/~ism/pdf/Hutama-scientificreport.pdf points to a paper by Daniel J. Hutama, in which he calculates this formula using Sage and provides the graph using the first ten non-trivial zeros, and then the first 100 non-trivial zeros.





19.5 Sketch of the Proof

We'll give a very rough sketch of the proof, omitting all details.

In analysis there are three important integral transforms, which convert one function to another function; the Fourier transform, the Laplace transform, and the Mellin transform. We discuss these in a later chapter.

Here is the Mellin transform with its inverse:

$$\hat{f}(y) = \int_0^\infty x^y f(x)\,\frac{dx}{x}$$

$$f(x) = \frac{1}{2\pi i}\lim_{T\to\infty}\int_{c-iT}^{c+iT} x^{-y}\hat{f}(y)\,dy \quad \text{for any large real } c$$

This transform converts a function of a real variable $f(x)$ to a function of a complex variable $\widehat{f}(y)$, and the inverse transformation recovers the original f.

The proof will use the inverse transform without mentioning the direct transform. It is convenient to slightly change notation. So we define an integral transform from a function of a complex variable $f(z)$ to a function of a real variable $\hat{f}(x)$ by

$$\widehat{f}(x) = \frac{1}{2\pi i}\int_{\sigma-\infty}^{\sigma+\infty} f(z)\frac{x^z}{z}\,dz$$

We integrate along a vertical line in the complex plane with real part σ. By Cauchy's theorem, the result should be independent of σ if $f(z)$ has no singularities in the region from one σ to another. We will apply this transform to $-\frac{\zeta'(z)}{\zeta(z)}$, which has singularities whenever ζ has a zero or a pole. Since there are no poles for $\Re(z) > 1$, the transform should be independent of σ provided σ is in this region.

The proof begins with a calculation known as *Perron's Formula*, which almost immediately explains why we find this integral transform useful.

Lemma 18 (Perron's Formula)**.** *If $f(z) = 1$, then $\widehat{f}(x) = 0$ for $x < 1$ and 1 for $x > 1$.*

We will not prove this, but we will use it. Suppose, then, that $f(z) = \frac{1}{n^z}$. Then the transform is

$$\frac{1}{2\pi i}\int_{\sigma-\infty}^{\sigma+\infty}\left(\frac{x}{n}\right)^z\frac{1}{z}\,dz$$

and this is zero for $\frac{x}{n} < 1$ and one for $\frac{x}{n} > 1$. In short, it is zero for $x < n$ and 1 for $x > n$. Thus when we apply the transform to $\sum_n \frac{\Lambda(n)}{n^z}$, integrating term by term, the result is

$$\sum_{n<x}\Lambda(n) = \Psi(x)$$

So far, the integral does not depend on σ, except that we require $\sigma > 1$ so we can apply the formula

$$-\frac{\zeta'(z)}{\zeta(z)} = \sum_n \frac{\Lambda(n)}{n^z}$$

We now want to push σ left. We no longer have the formula $\sum \frac{\Lambda(n)}{n^z}$, but we still have the formula derived from the product formula of the zeta function. So we can use the terms of that formula to determine the effect as we push left.

Notice, however, that the transform of any constant is a function which equals that constant right of 1, but zero left of 1. So we must add any constant in our formula on the left to our formula for $\Psi(x)$ to insure that our calculation still gives the value of $\Psi(x)$.

More generally, any term $g(z)$ on the left contributes $g(0)$ to our formula for $\Psi(x)$.

But in addition, terms with poles or zeros contribute other terms as we move left. Consider, for example, the term $\frac{1}{z-1}$. Let us compute the contribution from moving across this singularity.

Let us write $\int_\sigma$ for $\frac{1}{2\pi i}$ times the integral over the vertical line with real part σ. Then

$$\int_{1-\epsilon} = \int_{1+\epsilon} - \left(\int_{1+\epsilon} - \int_{1-\epsilon} \right) = \Psi(x) - \left(\int_{1+\epsilon} - \int_{1-\epsilon} \right)$$

The second expression is essential a contour integral around the singularity, so

$$\int_{1-\epsilon} = \Psi(x) - \mathrm{Res}_1 \left(\frac{1}{z-1} \frac{x^z}{z} \right) = \Psi(x) - x$$

In short,

$$\Psi(x) = x + \int_{1-\epsilon}$$

In this way, we have unfolded the first term in the explicit formula for $\Psi(x)$.

But notice that the term $\frac{!}{z-1}$ also contributes its value at the origin, which is -1. So in total this term contributes $x - 1$.

Next come several constants, and they contribute themselves to the formula.

Consider next $\left(\frac{1}{z-\rho} + \frac{1}{\rho} \right)$. This term contributes

$$\mathrm{Res}_\rho \left(\frac{1}{z-\rho} \frac{x^z}{z} \right) = \frac{x^\rho}{\rho}$$

But the term $\frac{1}{z-\rho}$ also contributes its value at the origin, $-\frac{1}{\rho}$. The additional term $\frac{1}{\rho}$ contributes $\frac{1}{\rho}$. So the full term contributes $\frac{x^\rho}{\rho}$.

The same story holds at $\left(\frac{1}{z+2n} - \frac{1}{2n}\right)$. This ends our sketch.

A rigorous proof will replace these infinite integrals with integrals from $\sigma - iT$ to $\sigma + iT$ and complete the contour with a line at the top moving left to $x = -T$, and then down, and then right to form a rectangle around finitely many non-trivial zeros and finitely many trivial zeros. This integral can be computed using the residue formula. Then as $T \to \infty$, estimates are required to show that the integral over the three added segments goes to zero. Details are in many books about the zeta function.

19.6 Postlude

Recall that

$$\pi(x) = \sum_{p \le x} 1 \sim \frac{x}{\log x}$$
$$\phi(x) = \sum_{p \le x} \log p \sim x$$
$$\Psi(x) = \sum_{p^m \le x} \log p \sim x$$

The function $\pi(x)$ counts the number of primes below x, $\phi(x)$ counts these primes weighted by $\log p$, and Ψ counts prime *powers* weighted by $\log p$. Each of the three estimates is equivalent to the prime number theorem.

Although the first formula is directly related to numerical data gathered from tables of primes, all three approximations yield the same rule of thumb in the end: the probability that an integer n is prime is about $\frac{1}{\log n}$.

For instance, the statement $\pi(x) \sim \frac{x}{\log x} = x\left(\frac{1}{\log x}\right)$ says that the proportion of primes among the first x integers is $\frac{1}{\log x}$. Since $\log x$ is very flat, the expression $\frac{1}{\log x}$ is constant over much of its range, so this number gives the probability that a number in this range is prime.

The rule of thumb suggests that in an interval of length $\log p$ starting at p there should be about one prime, so the second sum over this interval should give about the length of the interval. Summing over a union of such intervals below x should thus give x. So $\phi(x) \sim x$.

The function $\Psi(x)$ sums over powers of primes. But if $p^2 < x$, then p is considerably smaller than x, so most primes will only contribute $\log p$ to the sum, and only a few small primes will contribute $2 \log p$ or more. In short, the extra terms should not matter.

The graph of the logarithm is very, very flat for large x because $\frac{d}{dx}\log x = \frac{1}{x}$. And yet logarithms converge to infinity in the end. This convergence is very slow, so even enormous numbers have small logarithms. Consequently, our rule of thumb says that primes remain abundant, even for very large numbers. Ultimately primes become scarce, but only for outrageously large numbers. Consider, for instance, numbers with about 100 digits. Since

$$\frac{1}{\log(10^{100})} = \frac{1}{100\log 10} = \frac{1}{100\cdot 2.30} = \frac{1}{203}$$

about one in every 200 numbers is prime. And since half of them are even, the chance of guessing a random prime of this size are about one in a hundred. Contrast that with the chance of randomly guessing a perfect square, which is about one in 10^{50}. If someone used a laser to inscribe your initials on a grain of sand and deposited it on a random sandy beach somewhere in the world, your chance of randomly picking up that particular grain of sand are better than your chance of guessing a perfect square.

Let us stop a moment and recall the long history of $\zeta(z)$ and the distribution of primes. Leibniz began his mathematical career by discovering that

$$\sum \frac{1}{n(n+1)} = \sum\left(\frac{1}{n} - \frac{1}{n+1}\right) = \left(1-\frac{1}{2}\right) + \left(\frac{1}{2}-\frac{1}{3}\right) + \ldots = 1$$

This led him to study $\sum \frac{1}{n^2}$, which he could not evaluate. But rumors suggested that Newton knew the value, and the two men briefly corresponded using as intermediary the secretary of the Royal Society in London. Newton knew many things, but not the sum of this series.

Newton had few students, but Leibniz influenced the Bernoulli's: Johann, Jacob, Daniel, and others. These men came from Basel, Switzerland, and were a contentious bunch. Daniel's father Johann threw him out of the house when Daniel's reputation eclipsed his own, and bore a grudge until his death.

Euler was born in Basel, and his father knew the Bernoulli family. Euler's family moved to Riehen, Switzerland when he was very young, but he was eventually sent to Basel to live with his grandmother, and was tutored by Johann Bernoulli. When Euler was 13, he enrolled in the University of Basel. When he graduated, he applied for a job there, but was passed over. This is the greatest hiring mistake in mathematical history.

Euler ended up in Moscow., and there he eventually solved the *Basel problem* by proving that $\sum \frac{1}{n^2} = \frac{\pi^2}{6}$. His argument gave the values of $\zeta(n)$ for all even integers. This eventually led to the incredibly important formula

$$\zeta(z) = \prod_p \frac{1}{1-\frac{1}{p^z}}$$

relating the zeta function to prime numbers. From this, Euler deduced the first real advancement in the theory of primes since the Greeks: $\sum \frac{1}{p} = \infty$.

After Euler's death, Dirichlet used Euler's techniques to prove that there are infinitely many primes in any arithmetic progression, by showing that the series of their reciprocals diverges. This proof demonstrated the power of Euler's methods.

Twenty years later, Riemann discovered the analytic continuation of the zeta function and its functional equation. The functional equation contained a sin term and a Γ function. The zeros of the sin and the poles of Γ cancelled at odd negative integers to give finite values, and the functional equation then converted these to finite values at the even positive integers, proving Euler's computations in an easier way. But the sin did not cancel poles of Γ at even negative integers, so instead these poles are cancelled by zeros of the zeta function. Applying the functional equation then gives $0 = 0$ rather than a value of zeta at odd positive integers. Thus Riemann explained why Euler could compute values at even positive integers, but not odd positive integers.

The even greater discovery of Riemann is that the zeta function has zeros in the critical strip. The product formula for the zeta function shows that these zeros essentially determine the entire zeta function, so information about prime numbers is hidden somehow in those zeros.

The explicit formula shows in a spectacular way that these zeros understand the prime numbers. Namely, the sum $x - \sum \frac{x^\rho}{\rho}$ converges to a flat function with jumps exactly at the primes and their powers.

Each of these men pushed the theory forward. And yet most were dead before their result was rigorously proved, or its consequences understood.

19.7 The Calculation of $\zeta'(0)$

While not of great importance, the calculation of $\zeta'(0)$ is a nice juggling act using several of our key formulas. According to the functional equation of the zeta function

$$\zeta(z) = 2^z \pi^{z-1} \sin\left(\frac{\pi z}{2}\right) \Gamma(1-z)\zeta(1-z)$$

Change each z to $1-z$ and notice that $\sin\left(\frac{\pi(1-z)}{2}\right) = \cos\left(\frac{\pi z}{2}\right)$ to obtain

$$\zeta(1-z) = 2^{1-z}\pi^{-z} \cos\left(\frac{\pi z}{2}\right) \Gamma(z)\zeta(z) = 2(2\pi)^{-z} \cos\left(\frac{\pi z}{2}\right) \Gamma(z)\zeta(z)$$

Take the logarithmetic derivative to get

$$-\frac{\zeta'(1-z)}{\zeta(1-z)} = -\log(2\pi) - \frac{\pi}{2}\tan\left(\frac{\pi z}{2}\right) + \frac{\Gamma'(z)}{\Gamma(z)} + \frac{\zeta'(z)}{\zeta(z)}$$

To get a value for $\zeta'(0)$, we will set $z = 1$ in this equation, but before doing that we face three problems. First we need to compute $\frac{\Gamma'(1)}{\Gamma(1)}$. Second, the tangent term and the zeta term on the right have poles at $z = 1$, so we need to expand both in Laurent expansions to show that these poles cancel out.

The required Γ calculation is easy using the product rule $\Gamma(z) = G(z)^{-1}$ where

$$G(z) = z\, e^{\gamma z} \prod_{n>0} \left(1 + \frac{z}{n}\right) e^{-\frac{z}{n}}$$

Taking logarithmetic derivatives,

$$\frac{\Gamma'(z)}{\Gamma(z)} = -\frac{1}{z} - \gamma - \sum_{n>0}\left(\frac{\frac{1}{n}}{1+\frac{z}{n}} - \frac{1}{n}\right) = -\frac{1}{z} - \gamma - \sum\left(\frac{1}{n+z} - \frac{1}{n}\right)$$

When $z = 1$, this gives

$$\frac{\Gamma'(1)}{\Gamma(1)} = -1 - \gamma - \left(\frac{1}{2} - 1\right) - \left(\frac{1}{3} - \frac{1}{2}\right) + \ldots = -\gamma$$

Expanding the tangent term about $z = 1$ gives

$$-\frac{\pi}{2}\tan\left(\frac{\pi z}{2}\right) = -\frac{\pi}{2}\frac{\sin\frac{\pi z}{2}}{\cos\frac{\pi z}{2}} =$$

$$-\frac{\pi}{2}\left(\frac{\sin\frac{\pi}{2} + \frac{\pi}{2}\cos\frac{\pi}{2}(z-1) + \ldots}{\cos\frac{\pi}{2} - \frac{\pi}{2}\sin\frac{\pi}{2}(z-1)}\right) = -\frac{\pi}{2}\left(\frac{1+\ldots}{-\frac{\pi}{2}(z-1)+\ldots}\right) = \frac{1}{z-1} + 0 + c_1(z-1) + \ldots$$

At the end of the proof we will show that

$$\frac{\zeta'(z)}{\zeta(z)} = -\frac{1}{z-1} + \gamma + d_1(z-1) + \ldots$$

Using this, we can finally put $z = 1$ in the formula at the bottom of the previous page to get

$$-\frac{\zeta'(0)}{\zeta(0)} = -\log(2\pi) + \left(\frac{1}{z-1+0}\right) - \gamma + \left(-\frac{1}{z-1} + \gamma\right) = -\log(2\pi)$$

We earlier proved that $\zeta(0) = -\frac{1}{2}$, so

$$\zeta'(0) = -\frac{1}{2}\log(2\pi) = -\log\sqrt{2\pi}$$

To complete this argument, we show that the expansion of ζ near $z = 1$ is

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

Then it will follow that

$$\frac{\zeta'(z)}{\zeta(z)} = \frac{\frac{-1}{(z-1)^2} + c + ...}{\frac{1}{z-1} + \gamma + ...} = -\frac{1}{z-1}\left(\frac{1 + c(z-1)^2 + ...}{1 + \gamma(z-1) + ...}\right) = -\frac{1}{z-1}\Big(1+c(z-1)^2+...\Big)\Big(1-\gamma(z-1)+...\Big) =$$

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

We use the formula below, developed when we determined the order of zeta. This formula analytically continues $\zeta(z)$ to $\Re(z) > 0$.

$$\zeta(z) = \frac{z}{z-1} - z\int_1^\infty \frac{x - \lfloor x \rfloor}{x^{z+1}}\, dx = \frac{1}{z-1} + 1 - z\int_1^\infty \frac{x - \lfloor x \rfloor}{x^{z+1}}\, dx$$

To prove that $\zeta(z) = \frac{1}{z-1} + \gamma + ...$, we need only prove

$$1 - \int_1^\infty \frac{x - \lfloor x \rfloor}{x^2}\, dx = \gamma$$

It is useful to rewrite this slightly, using the result $\int_1^\infty \frac{1/2}{x^{z+1}}\, dx = \frac{1}{2}$.

$$\frac{1}{2} - \int_1^\infty \frac{x - \lfloor x \rfloor - 1/2}{x^2}\, dx = \gamma$$

Note that on each interval $[n, n+1]$, the function $x - \lfloor x \rfloor - \frac{1}{2}$ is linear with value $-\frac{1}{2}$ on the left and $\frac{1}{2}$ on the right. So this function is $x - n - \frac{1}{2}$ on $[n, n+1]$ and the integral is the limit as $N \to \infty$ of

$$\frac{1}{2} - \sum_{n=1}^{N} \int_n^{n+1} \frac{x - n - \frac{1}{2}}{x^2}\, dx = \frac{1}{2} - \sum_{n=1}^{N} \int_n^{n+1} \left(\frac{1}{x} - \frac{n + \frac{1}{2}}{x^2}\right) dx =$$

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

In this last sum, the term $\frac{1}{n+1}$ occurs in both the nth and $n+1$th terms. The entry multiplying it, $n + \frac{1}{2}$, is one larger in the second appearance than in the first appearance. So the sum is

actually $\sum_{n=1}^{N} \frac{1}{n+1}$, leaving two stranded terms $-\frac{3}{2}$ and $\left(N+\frac{1}{2}\right)\left(\frac{1}{N+1}\right)$. We conclude that the entire expression is

$$\frac{1}{2} - \log(N+1) + \left(\frac{1}{2} + \frac{1}{3} + ... + \frac{1}{N}\right) + \frac{3}{2} - \frac{N+\frac{1}{2}}{N+1} =$$

$$1 - \log(N+1) + \left(1 + \frac{1}{2} + \frac{1}{3} + ... + \frac{1}{N}\right) - \frac{N+\frac{1}{2}}{N+1}$$

Recall that

$$\gamma = \lim_{N\to\infty}\left[(1 + \frac{1}{2} + ... + \frac{1}{N}) - \log(N+1)\right]$$

The previous expression thus converges to γ as $N \to \infty$ because the last term converges to -1 and cancels the initial 1. QED.

Chapter 20

Dirichlet Series

20.1 Abscissa of Absolute Convergence

Many of the proofs in this chapter are taken from J.P. Serre's book on number theory, *A Course in Arithmetic*.

Definition 45. *A* Dirichlet series *is a series of the following type, where the a_k are complex numbers:*

$$\sum_{k\geq 0} \frac{a_k}{n^z}$$

Theorem 139. *Suppose $\sum \frac{a_k}{n^z}$ is a Dirichlet series.*

- *If the series converges absolutely at z_0, it converges absolutely at z whenever $\Re(z) > \Re(z_0)$.*
- *There is a real number σ_a, called the* abscissa of absolute convergence*, such that the series converges absolutely for $\Re(z) > \sigma_a$ and does not converge absolutely for $\Re(z) < \sigma_a$. This number could be ∞ or $-\infty$.*

Proof: The first item follows from the comparison test, since $|n^z| = n^x > n^{x_0} = |n^{z_0}|$ and therefore $\left|\frac{a_k}{n^z}\right| \leq \left|\frac{a_k}{n^{z_0}}\right|$.

The second item is an immediate consequence of the first.

20.2 Key Results

Theorem 140. *Suppose $\sum \frac{a_k}{n^z}$ is a Dirichlet series.*

- *If the series converges conditionally at z_0, then it converges conditionally on any "wedge" with angle less than π anchored at z_0. Convergence is uniform on the closed wedge. (See the picture at the top of the next page.)*
- *There is a real number σ_c, called the* abscissa of conditional convergence*, such that the series converges conditionally for $\Re(z) > \sigma_c$ and does not converge for $\Re(z) < \sigma_c$.*
- *The sum of the series is holomorphic on the half plane $\Re(z) > \sigma_c$.*
- *We have $\sigma_a - 1 \leq \sigma_c \leq \sigma_a$.*

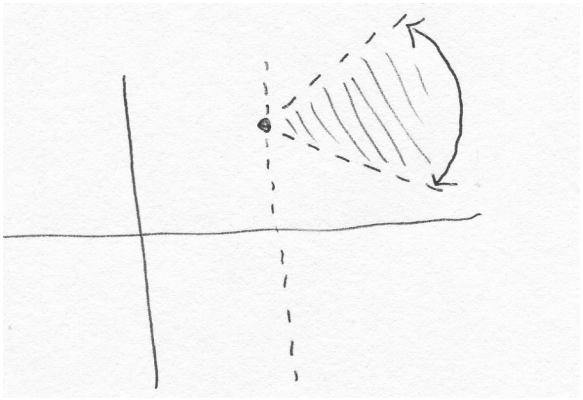


Figure 20.1: Conversion Widge

Proof: The first item clearly implies everything except the last item. Let us clear the decks by proving the last item first. Trivially, $\sigma_c \leq \sigma_a$. Let $\epsilon > 0$ and select z_0 with $\sigma_c < \Re(z_0) < \sigma_c + \frac{\epsilon}{2}$. Then $\sum \frac{a_k}{k^{z_0}}$ converges, and therefore $\left|\frac{a_k}{k^{z_0}}\right| \to 0$. Therefore, these terms are bounded; choose a bound B such that for all k we have $\left|\frac{a_k}{k^{z_0}}\right| \leq B$.

If $\Re(z) > \sigma_c + 1 + \epsilon$, then $\Re(z) > \Re(z_0) + 1 + \frac{\epsilon}{2}$, so

$$\left|\frac{a_k}{k^z}\right| = \left|\frac{a_k}{k^{z_0}}\right| \left|\frac{1}{k^{1+\epsilon}}\right| \leq B \left|\frac{1}{k^{1+\epsilon}}\right|$$

Since the sum of the final terms converges absolutely, $\sum \frac{a_k}{k^z}$ converges absolutely whenever $\Re(z) > \sigma_c + 1 + \epsilon$, for any epsilon, and thus whenever $\Re(z) > \sigma_c + 1$. So $\sigma_a \leq \sigma_c + 1$.

Proof: Finally we prove the "wedge theorem." We can suppose $z_0 = 0$ because

$$\sum \frac{a_k}{n^{z_0+z}} = \sum \frac{\left(\frac{a_k}{n^{z_0}}\right)}{n^z}$$

Therefore assume $\sum a_k$ converges, and thus by the Cauchy condition, for any $\epsilon > 0$ there is an N such that $m, n > N$ implies $\left|\sum_{k=m}^{n} a_k\right| < \epsilon$. We are going to show a similar inequality for $\left|\sum_{k=m}^{n} a_k \frac{1}{n^z}\right|$ for z in the wedge. Indeed, our new inequality will be uniform for z in a compact subset of the wedge, and that will complete the proof.

Lemma 19 (Abel)**.** *Given sequences a_k and b_k, we have*

$$\sum_{k=1}^{n} a_k b_k = a_1(b_1-b_2)+(a_1+a_2)(b_2-b_3)+...+(a_1+a_2+...+a_{n-1})(b_{n-1}-b_n)+(a_1+a_2+...+a_n)b_n$$

Proof of lemma Induction on n. The result is clear for $n = 2$. To get from the nth stage to the $n+1$st stage, we must convert

$$a_{n+1}b_{n+1} + (a_1 + a_2 + ... + a_n)b_n \to (a_1 + a_2 + ... + a_n)(b_n - b_{n+1}) + (a_1 + a_2 + ... + a_{n+1})b_{n+1}$$

and this is clearly true.

Continuation of Wedge Proof: Apply Abel's lemma to a shifted sequence, so that the sum starts with $k = m$ rather than $k = 1$. By the Cauchy condition, all sums $a_m + a_{m+1} + ...$ have absolute value less than ϵ. Consequently,

$$\left|\sum_{k=m}^{n} a_k \frac{1}{k^z}\right| \leq \sum_{k=m}^{n-1} \epsilon \left|\frac{1}{k^z} - \frac{1}{(k+1)^z}\right| + \epsilon \left|\frac{1}{n^z}\right|$$

Note that

$$\left(\frac{1}{k^z} - \frac{1}{(k+1)^z}\right) = z \int_{\ln k}^{\ln(k+1)} e^{-zt}\, dt$$

and the absolute value of this expression is bounded by

$$|z| \int_{\ln k}^{\ln(k+1)} e^{-xt} dt = \frac{|z|}{x}\left(\frac{1}{k^x} - \frac{1}{(k+1)^x}\right)$$

Consequently $\left|\sum_{k=m}^{n} \frac{a_k}{k^x}\right| \leq \epsilon \frac{|z|}{x}\frac{1}{m^x} + \epsilon \frac{1}{n^x}$ We have $\Re(z) = x > 0$ because at the start of the proof we shifted the picture to the right half plane, so $\frac{1}{m^x}$ and $\frac{1}{n^x}$ are at most 1. Therefore

$$\left|\sum_{k=m}^{n} \frac{a_k}{k^x}\right| \leq \epsilon\left(1 + \frac{|z|}{x}\right)$$

The diagram below shows that if θ is half of the total angle of the wedge, then $\frac{|z|}{x} \leq \frac{1}{\cos\theta}$ and thus $1 + \frac{|z|}{x}$ is bounded by a constant on the wedge. This completes the proof.

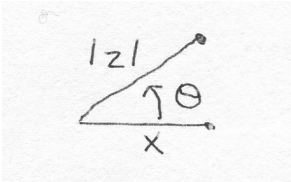


Figure 20.2: $\frac{|z|}{x}$ is bounded

Remark: We end with a very specialized result, because we use it in the crucial step of the next chapter.

Theorem 141. *Suppose $f(z) = \sum \frac{a_k}{k^z}$ is a Dirichlet series with a_k real, $a_k \geq 0$. Suppose this series converges for $\Re(z) > \rho$, and suppose $f(z)$ can be extended to a function holomorphic in a small disk about $z = \rho$. Then $\sigma_c < \rho$ and thus the series actually converges on a larger open half plane.*

Proof: Replacing z by $z - \rho$, we can assume $\rho = 0$.

Since $f(z)$ exists in a small disk about the origin and in the half plane $\Re(z) > 0$, it exists on a disk about $z = 1$ of some positive radius $1 + \epsilon$. Shrink ϵ slightly so f is holomorphic on the *closed* disk of radius $1 + \epsilon$. So the Taylor series about $z = 1$ converges in this closed disk.

The idea of the proof is simple. We will compute this Taylor series. Then we will rearrange it and discover that it equals the Dirichlet series at $z = -\epsilon$. So this Dirichlet series converges to the left of ρ and we are done.

Here are the details. The Taylor series is

$$f(z) = \sum_n \frac{f^{(n)}(1)}{n!}(z-1)^n$$

The derivatives of f at 1 are $\sum a_k k^{-z}$ at 1, which is $\sum \frac{a_k}{k}$, and $\sum a_k(-\ln k)k^{-z}$ at 1, which is $-\sum_k \frac{\ln k a_k}{k}$, and $\sum a_k(-\ln k)^2 k^{-z}$ at 1, which is $\sum \frac{(\ln k)^2 a_k}{k}$, etc. So the Taylor series is

$$f(z) = \sum_n \frac{\sum_k (-\ln k)^n a_k)}{k\ n!}(z-1)^n$$

and therefore

$$f(-\epsilon) = \sum_n \frac{\sum_k (-1)^n (\ln k)^n a_k (-1)^n (\epsilon+1)^n}{k\ n!} = \sum_n \sum_k \frac{((\epsilon+1)\ln k)^n\, a_k}{k\ n!}$$

All the terms in this series are real and positive, so the series converge absolutely and thus can be rearranged to

$$\sum_k \frac{a_k}{k} \sum_n \frac{((\epsilon+1)\ln k)^n}{n!} = \sum_k \frac{a_k}{k} e^{(\epsilon+1)\ln k} = \sum_k \frac{a^k k^{\epsilon+1}}{k} = \sum_k \frac{a_k}{k^{-\epsilon}} = f(-\epsilon)$$

Therefore, the Dirichlet series actually converges at $-\epsilon$, and indeed to $f(-\epsilon)$.

Chapter 21

Dirichlet's Theorem

This optional chapter proves Dirichlet's theorem on primes in an arithmetic progression. The chapter shows that the techniques of the previous chapters can be used to prove other results in number theory.

21.1 Characters of a Finite Abelian Group

Definition 46. *Let G be a finite abelian group. A* character *of G is a group homomorphism $\varphi : G \to S^1$.*

Example: Suppose $G = Z/nZ$. A character φ is determined by $\varphi(1) = e^{i\theta}$ and $\varphi(1)^n = 1$. So $\varphi(1) = e^{2\pi ik/n}$ where $0 \leq k < n$. Therefore G has n characters

$$\varphi_k(t) = e^{2\pi itk/n}$$

Example: Suppose $G = G_1 \times G_2 \times ... \times G_m$. If φ is a character of G, we can restrict φ to G_k and get a character of G_k. Call the resulting characters $\varphi_1, \varphi_2, ..., \varphi_n$. Then clearly

$$\varphi(g_1 \times g_2 \times ... \times g_n) = \varphi_1(g_1) \cdot \varphi_2(g_2) \cdot ... \cdot \varphi_n(g_n)$$

Since every finite abelian group has the form $Z/n_1Z \times Z/n_2Z \times ... \times Z/n_nZ$, these observations quickly give all characters of all finite abelian groups.

Remark: The characters of a finite abelian G themselves form a group $\widehat{G}$, whose group operation is multiplication of values of characters. The previous results easily show that $\widehat{G}$ is isomorphic to G.

Definition 47. *Let $L(G)$ be the vector space of all complex valued functions on G. Define an inner product on $L(G)$ by*

$$< f, g >= \frac{1}{|G|} \int_G \overline{f(x)} g(x)$$

where integrating over G is just summing over G. Thus this formula is actually

$$< f, g >= \frac{1}{|G|} \sum_{x \in G} \overline{f(x)} g(x)$$

Remark: We prefer the integral sign because this entire theory generalizes to situations where the sum becomes an integral sign. But those generalizations do not appear in this chapter.

Theorem 142. *Let G be a finite abelian group. The characters of G are orthonormal in $L(G)$.*

Proof: Let φ and ψ be characters. Fix $y \in G$; then

$$< \varphi, \psi >= \frac{1}{|G|} \sum_{x \in G} \overline{\varphi(x)} \psi(x) = \frac{1}{|G|} \sum_{x \in G} \overline{\varphi(xy)} \psi(xy) = \frac{1}{|G|} < \varphi, \psi > \overline{\varphi(y)} \psi(y)$$

So either $< \varphi, \psi >= 0$ or else $\overline{\varphi(y)} \psi(y) = 1$ for all $y \in G$. Since $\varphi(y) \in S^1$, $\overline{\varphi(y)} \varphi(y) = 1$. So $\overline{\varphi(y)} = \frac{1}{\varphi(y)}$ and $\overline{\varphi(y)} \psi(y) = 1$ if and only if $\psi(y) = \varphi(y)$.

Moreover,

$$< \varphi, \varphi >= \frac{1}{|G|} \int \overline{\varphi(x)} \varphi(x) = \frac{1}{|G|} \int 1 = 1$$

Corollary 31. *Let G be a finite abelian group. The characters of G form an orthonormal basis of $L(G)$. Consequently, for any function f on G, we can find constants c_φ such that*

$$f(x) = \sum_\varphi c_\varphi \varphi(x)$$

Proof: Clearly $L(G)$ has dimension $|G|$; by comments above, this also equals the number of characters of G.

Corollary 32. *In the previous sum $f(x) = \sum c_\varphi \varphi(x)$, we have*

$$c_k =< \varphi, f >$$

Proof: This proof repeats the initial step in Fourier theory. Namely, for a fixed character χ, write

$$< \chi, f >=< \chi, \sum c_\varphi \varphi >= \sum c_\varphi < \chi, \varphi >$$

Since the φ are orthnormal, the only term that matters in the sum is the term when $\varphi = \chi$, and we obtain

$$< \chi, f >= c_\chi$$

21.2 Statement of Dirichlet's Theorem; L Functions

Theorem 143 (Dirichlet's Theorem)**.** *Let a and $b \geq 2$ be positive relatively prime integers. Then there are infinitely many primes in the arithmetic progression $a + bk$. Said another way, in base b there are infinitely many primes whose base-b expansion ends in a.*

Indeed, more is true. Let $\phi(b)$ be the order of $Z_b^\star$. Let $\pi(x, a)$ be the number of primes less than or equal to x and congruent to a modulo b. Then

$$\pi(x, a) \sim \frac{1}{\phi(b)} \frac{x}{\ln x}$$

Hence each residue class contains about the same number of primes.

Example For example, suppose $b = 10$. Then when written in base 10, about $\frac{1}{4}$ of the primes end in each of $1, 3, 7, 9$.

Proof: Dirichlet proved this using Euler's argument and a modified zeta function with a modified product formula. It is tempting to start with $\sum \frac{1}{n^z}$ using only n whose prime factorization contains primes congruent to a modulo b. But this causes many problems, not the least being that we don't know that any such prime exists. Dirichlet's trick is to assign a "weight" $\varphi(p)$ to each prime, multiplicably extend φ to all positive integers, and then study $\sum \frac{\varphi(n)}{n^z}$. Many choices of φ are possible, and by manipulating the choices, Dirichlet eventually proves that $\sum_{p \equiv a \pmod b} \frac{1}{p}$ diverges.

It is crucial that $\sum \frac{\varphi(n)}{n^z}$ satisfy a product formula, and an easy calculation shows that this will happen *if* the weights form a character of the group $Z_b^\star$ of all invertible elements in Z_b. This group is, equivalently, the set of all elements of Z_b whose representatives a are relatively prime to b.

Example The group $Z_6^\star$ has elements represented by $\{1, 5\}$ and the group $Z_{12}^\star$ has elements represented by $\{1, 5, 7, 11\}$. The group $Z_7^\star$ has elements represented by $\{1, 2, 3, 4, 5, 6\}$.

Definition 48. *Let $b \geq 2$ be a positive integer. Let φ be a character of $Z_b^\star$. Then we define the Dirichlet L-function associated with φ by*

$$L(z, \varphi) = \sum_{n \geq 1} \frac{\varphi(n)}{n^z}$$

Here φ has been extended to be zero on integers that do not represent elements of $Z_b^\star$.

Theorem 144. *Each $L(z, \varphi)$ converges absolutely on $\Re(z) > 1$ to a holomorphic function. This*

function has a product representation

$$L(z,\varphi)=\prod_p \frac{1}{\left(1-\frac{\varphi(p)}{p^z}\right)}$$

Remark: Notice that primes p not relatively prime to b contribute 1 to this product.

Proof: The proof is a completely straightforward rewording of the similar proof for the zeta functions.

Remark: The function $L(z,1)$ is

$$L(z,1)=\prod_{p\not| b} \frac{1}{\left(1-\frac{1}{p^z}\right)}=\prod_{p|b}\left(1-\frac{1}{p^z}\right)\zeta(z)$$

The final product has finitely many terms, so it follows that $L(z,1)$ can be analytically continued to a meromorphic function on the entire complex plane with a pole of order 1 at the origin and no other poles. The residue of this pole is $\prod_{p|b}\left(1-\frac{1}{p}\right)$. Notice that this residue equals $\frac{\phi(b)}{b}$, the order of $Z_b^\star$ divided by b. Indeed

$$Z_{p_1^{k_1}p_2^{k_2}\cdots p_l^{k_l}}\cong Z_{p_1^{k_1}}\times\ldots\times Z_{p_l^{k_l}}$$

so $\phi(p_1^{k_1}p_2^{k_2}\ldots p_l^{k_l})=\phi(p_1^{k_1})\phi(p_2^{k_2})\ldots\phi(p_l^{k^l})$. Moreover $\phi(p^k)=p^k-p^{k-1}=p^k\left(1-\frac{1}{p}\right)$.

Remark: The remaining sections of this chapter complete the proof of Dirichlet's theorem.

21.3 Analytic Continuation

Theorem 145. *If $\varphi\neq 1$, $L(z,\varphi)$ can be analytically extended to a holomorphic function on $\Re(z)>0$.*

Proof: Let $\Lambda(x)=\sum_{n\leq x}\varphi(n)$. Notice that $<\varphi,1>=0$ and therefore $\sum\varphi(g)=0$. It follows that Λ is a bounded function.

By a familiar argument, we have

$$z\int_1^{n+1}\frac{\Lambda(t)}{t^{z+1}}\,dt=z\sum_{k=1}^{n}\Lambda(k)\left.\frac{t^{-z}}{-z}\right|_k^{k+1}=-\sum_{k=1}^{n}\Lambda(k)\left(\frac{1}{(k+1)^z}-\frac{1}{k^z}\right)$$

$$=\sum_{k=1}^{n}(\Lambda(k+1)-\Lambda(k))\frac{1}{k^z}-\frac{\Lambda(n)}{(n+1)^z}=\sum_{k=1}^{n}\frac{\varphi(k)}{k^z}-\frac{\Lambda(n)}{(n+1)^z}$$

and so

$$\sum_{k=1}^{n} \frac{\varphi(k)}{k^z} = \frac{\Lambda(n)}{(n+1)^z} + z\int_1^{n+1} \frac{\Lambda(t)}{t^{z+1}}\,dt$$

Let $n \to \infty$ and recall again that Λ is bounded. If $Re(z) > 0$, we obtain

$$\sum \frac{\varphi(k)}{k^z} = z\int_1^{\infty} \frac{\Lambda(t)}{t^{z+1}}\,dt$$

The right hand sum converges to a holomorphic function if $\Re(z) > 0$ and we are done.

21.4 Proof of Theorem Modulo $L(1,\varphi) \neq 0$ for $\varphi \neq 1$

In the previous chapter we proved

$$-\frac{\zeta'(z)}{\zeta(z)} = \sum_p \frac{\ln p}{p^z} + \sum_p \frac{\ln p}{p^z(p^z-1)}$$

and observed that the second term is holomorphic for $\Re(z) > 0$. Since the product formula for $L(z,1)$ excludes primes p which divide b, we obtain

$$-\frac{L'(z,1)}{L(z,1)} = \sum_{p\not|b} \frac{\ln p}{p^z} + \sum_{p\not|b} \frac{\ln p}{p^z(p^z-1)}$$

A similar calculation when $\varphi \neq 1$ gives

$$-\frac{L'(z,\varphi)}{L(z,\varphi)} = -\sum_{p\not|b} \frac{-\left(1-\frac{\varphi(p)}{p^z}\right)^{-2}(\varphi(p)\ \ln p\ \frac{1}{p^z})}{\left(1-\frac{\varphi(p)}{p^z}\right)^{-1}} = \sum_{p\not|b} \frac{\varphi(p)\ \ln p}{p^z\left(1-\frac{\varphi(p)}{p^z}\right)}$$

$$= \sum_{p\not|b} \frac{\varphi(p)\ln p}{p^z} + \sum_{p\not|b} \frac{\varphi(p)^2 \ln p}{p^z(p^z-\varphi(p))}$$

All of this holds for $\Re(z) > 1$.

Notice, however, that the final term of this formula is holomorphic for $\Re(z) > \frac{1}{2}$. The key inequality establishing this is $|p^z - \varphi(p)| \geq ||p^z| - |\varphi(p)|| = |p^x - 1| = p^x - 1$.

Now consider the function $f \in L(Z_b^\star)$ that equals 1 on a and 0 everywhere else. By an earlier result, this function and indeed every element of $L(Z_b^\star)$ can be written as a linear combination of the characters. So there are constants c_φ such that $f(x) = \sum_\varphi c_\varphi \varphi(x)$. Form

$$\sum_\varphi -c_\varphi \frac{L'(z,\varphi)}{L(z,\varphi)} = \sum_\varphi \sum_{p|b} \frac{c_\varphi\ \varphi(p)\ \ln p}{p^z} + \sum_\varphi \sum_{p|b} \frac{c_\varphi\ \varphi(p)^2 \ln p}{p^z(p^z - \varphi(p))}$$

Using the defining property of c_φ, this equals

$$\sum_\varphi -c_\varphi \frac{L'(z,\varphi)}{L(z,\varphi)} = \sum_{p \equiv a \pmod b} \frac{\ln p}{p^z} + \sum_\varphi \sum_{p|b} \frac{c_\varphi\ \varphi(p)^2 \ln p}{p^z(p^z - \varphi(p))}$$

This formula holds if $\Re(z) > 1$. On the other hand, the final term in the formula is holomorphic for $\Re(z) > \frac{1}{2}$ and the initial term on the left is meromorphic on $\Re(z) > 0$ and holomorphic except on singularities of $\frac{L'}{L}$.

Restrict to real x and imagine taking a limit as $x \to 1$ from above. The limit of the final term is finite. The limit of the middle term is

$$\sum_{p \equiv a \pmod b} \frac{\ln p}{p}$$

This term is finite *unless there are infinitely many primes congruent to a modulo b.*

So Dirichlet's theorem will follow if the limit of the left side is infinite. We know that $L(z,1)$ has a pole at $x = 1$, so this term has an infinite limit. We know that $L(z,\varphi)$ for $\varphi \neq 0$ is holomorphic on $\Re(z) > 0$, so it certainly does not have a pole. On the other hand, such an $L(z,\varphi)$ might vanish at 1.

Consequently, Dirichlet's theorem will be true if we can prove

- $c_1 \neq 0$
- if $\varphi \neq 1, L(1,\varphi) \neq 0$

But $c_1 =< 1, f >= \frac{1}{|G|} \sum_{x \in G} \overline{1(x)}\ f(x) = \frac{1}{|G|}$ because every term in the sum vanishes except the term corresponding to $a \in Z_b^\star$.

Remark: We are reduced to proving that is $\varphi \neq 1$, then $L(1,\varphi) \neq 0$. It is at this point that Dirichlet's proof becomes miraculous. Having invented group characters, L-functions, and the apparatus of analytic number theory, there remained this final barrier. But Dirichlet discovered that Gauss had computed the class number of various algebraic number fields, and one of these

formulas expressed the class number as a product of various terms including $L(1,\varphi)$. Since the class number is a positive integer, $L(1,\varphi) \neq 0$.

Luckily, an argument was later found which avoids this step.

21.5 $L(1,\varphi) \neq 0$ if $\varphi \neq 1$

This argument is taken from Serre's book *A Course in Arithmetic*

Consider the product

$$\prod_{\varphi} L(z,\varphi)$$

All terms in this product are holomorphic on $\Re(z) > 0$ except $L(z,1)$, which has a pole at 1. If some $L(1,\varphi) = 0$, the pole will cancel in the product and the entire product will be holomorphic on $\Re(z) > 0$.

Let us sketch the argument which contradicts this possibility. We will reduce the above product to a single Dirichlet series, and this series will turn out to have nonnegative real coefficients. Consequently, the last theorem of the previous chapter says the product has a singularity on the real axis at σ_c. So $\sigma_c \leq 0$. However, we will discover a singularity on the positive x-axis, and this contradiction will establish the theorem.

Now the details. Fix a prime $p \not| b$. Suppose p has order g in $Z_b^\star$. Then $\varphi(p)$ is a gth root of unity. If w varies over all gth roots of unity, we have

$$\prod_{w}(X - w) = X^g - 1$$

We claim that the $\varphi(p)$ are exactly these w, each repeated $\frac{\phi(b)}{g}$ times. If that is true, then

$$\prod_{\varphi}(X - \varphi(p)) = (X^g - 1)^{\phi(b)/g}$$

Indeed, p generates a cyclic subgroup $Z_g \subseteq Z_b^\star$. This subgroup has g distinct characters, each determined by its value on p. Hence the values of these characters on p must run through all of the w. To finish, it suffices to show that each character of Z_p can be extended to a character of $Z_b^\star$ in exactly $|Z_b^\star / Z_p|$ ways.

More generally, suppose G is a finite abelian group and H is a subgroup. We then have an exact sequence

$$0 \to H \to G \to G/H \to 0$$

This sequence induces a similar sequence of character groups

$$0 \leftarrow \widehat{H} \leftarrow \widehat{G} \leftarrow \widehat{G/H} \leftarrow 0$$

and a brief check shows that this sequence is also exact. The tricky part is exactness on the left. This follows from a counting argument: If we replace $\widehat{H}$ in the sequence with the image K of $\widehat{(G)}$, then we certainly get exactness, so

$$|K| = |\widehat{(G)}| - |\widehat{G/H}| = |G| - |G/H| = |H| = |\widehat{H}|$$

Therefore, K is all of $\widehat{H}$. So each character of H comes from at least one character of G, and indeed from exactly $|G|/|G/H|$ such characters.

Continuation of Proof: We just proved

$$\prod_{\varphi}(X - \varphi(p)) = (X^g - 1)^{\phi(b)/g}$$

Each side has $\phi(b)$ copies of X; divide by these to get the rational formula

$$\prod_{\varphi}\left(1 - \frac{\varphi(p)}{X}\right) = \left(1 - \frac{1}{X^g}\right)^{\phi(b)/g}$$

Replace X by p^z, invert, and take the product over p to obtain

$$\prod L(z,\varphi) = \prod_{p,\varphi}\left(\frac{1}{1 - \frac{\varphi(p)}{p^z}}\right) = \prod_{p\not| b}\left(\frac{1}{1 - \frac{1}{p^{gz}}}\right)^{\phi(b)/g}$$

The product on the right equals

$$\prod_{p\not| b}\left(1 + \frac{1}{p^{gz}} + \frac{1}{p^{2gz}} + \ldots\right)^{\phi(b)/g}$$

When expanded, this will be a Dirichlet series with real nonnegative terms, as promised.

To complete the proof, we must show that this series diverges for some positive real z. Note that the pth term is larger than the result by simply raising each term to the power $\phi(b)/g$ and ignoring cross terms:

$$\left(1 + \frac{1}{p^{\phi(b)z}} + \frac{1}{p^{2\phi(b)z}} + \ldots\right)$$

The product over primes $p \not| b$ is larger than

$$\sum_{(n,b)=1} \frac{1}{n^{\phi(b)z}}$$

If $z = \frac{1}{\phi(b)}$, a positive real number, this sum is

$$\sum_{(n,b)=1} \frac{1}{n}$$

This sum diverges because $\sum_p \frac{1}{p}$ diverges and all but finitely many terms of the last sum are terms of $\sum_{(n,b)=1} \frac{1}{n}$.

This completes the proof of the first half of Dirichlet's theorem. We will diverge from Serre's book at this point, and prove the second half using a generalization of Zagier's exposition of Newman's proof of the prime number theorem. The generalization comes from a short paper by Ivan Soprounov at the University of Toronto.

21.6 Proof of $\pi(x, a) \sim \frac{1}{\phi(b)} \frac{x}{\ln x}$

Proof, Step 1: For all φ, $L(z, \varphi)$ has no zeros on the critical line $\Re(z) = 1$.

We already know this is true at $z = 1$ by the crucial lemma used to prove Dirichlet's theorem. It suffices to prove that $A(z) = \prod_\varphi L(z, \varphi)$ has no zeros at complex points of the critical line. Suppose the product has a zero of order μ at $1+i\alpha$. There may or may not be a zero at $1+2i\alpha$; let $\nu \geq 0$ be the order of this zero. Observe that $\overline{L(z, \varphi)} = L(\overline{z}, \overline{\varphi})$ and that $\overline{\varphi}$ is another character. It follows that $\overline{\prod_\varphi L(z, \varphi)} = \prod_\varphi (\overline{z}, \varphi)$. Thus $1 - i\alpha$ and $1 - 2i\alpha$ will have zeros of orders μ and nu.

Taking the logarithmic derivative of $A = \prod L(z, \varphi)$ and using the results at the start of section 19.4, we have

$$-\frac{A'(z}{A(z)} = \sum_{\varphi(p)} \frac{\varphi(p) \ln p}{p^z} + \text{terms holomorphic on } \Re(z) > \frac{1}{2}$$

Let $f(z) = \sum_\varphi \frac{\varphi(p) \ln p}{p^z}$. This function has poles with residue -1 at $1 + i\alpha$, $1 - i\alpha$, and possibly at $1 + 2i\alpha$ and $1 - 2i\alpha$ and a pole with residue 1 at 1. Form

$$f(1 + 2i\alpha + \epsilon) + 4f(1 + i\alpha + \epsilon) + 6f(1 + \epsilon) + 4f(1 - i\alpha + \epsilon) + f(1 - 2i\alpha + \epsilon)$$

This expression equals

$$\sum \frac{\varphi(p)\ln p}{p^{1+\epsilon}}\left(\frac{1}{p^{2i\alpha}}+4\frac{1}{p^{i\alpha}}+6+4\frac{1}{p^{-i\alpha}}+\frac{1}{p^{-2i\alpha}}\right)=\sum \frac{\varphi(p)\ln p}{p^{1+\epsilon}}\left(\frac{1}{p^{i\alpha}}+\frac{1}{p^{-i\alpha}}\right)^4$$

Multiply this expression by ϵ and then take the limit as ϵ approaches zero. The expression goes to $-2\nu-8\mu+6$.

On the other hand, this expression is a non-negative real number, for reasons to be given shortly, so $-2\nu-8\mu+6\geq 0$ and therefore $\mu=0$.

Indeed, everything in sight is positive except possibly $\sum_\varphi \varphi(p)$. If $(p,b)\neq 1$, all φ vanish on p and $\sum_\varphi \varphi(p)=0$. If $p\equiv 1 \pmod b$, all $\varphi(p)=1$ and $\sum_\varphi \varphi(p)=|Z_b^\star|=\phi(b)$. If $p\not\equiv 1 \pmod b$, then there is character τ with $\tau(p)\neq 1$, since otherwise the subgroup of $Z_b^\star$ generated by p, H, is nontrivial and all characters are actually characters of $Z_b^\star/H$, but then there would be fewer than $|Z_b^\star|$ characters. So $\sum\varphi(p)=\sum(\varphi\circ\tau)(p)=\left(\sum\varphi(p)\right)\tau(p)$. Since $\tau(p)\neq 1$, $\sum\varphi(p)=0$.

21.7 Final Steps in Proof

Let $(a,b)=1$. Let $\pi_b(x,a)$ be the number of primes p less than x such that $p\equiv a \pmod b$. Recall that Dirichlet's theorem states that

$$\pi_b(x,a)\sim\frac{1}{\phi(b)}\frac{x}{\ln x}$$

Just as earlier in this chapter, consider the function $f\in L(Z_b^\star)$ that equals 1 on a and 0 everywhere else. This function can be written as a linear combination of the characters. So there are constants c_φ such that $f(x)=\sum_\varphi c_\varphi\varphi(x)$. Form

$$\sum_\varphi -c_\varphi\frac{L'(z,\varphi)}{L(z,\varphi)}=\sum_\varphi\sum_{p\nmid b}\frac{c_\varphi\,\varphi(p)\,\ln p}{p^z}+\sum_\varphi\sum_{p\nmid b}\frac{c_\varphi\,\varphi(p)^2\ln p}{p^z(p^z-\varphi(p))}$$

The expression on the right is holomorphic for $\Re(z)>\frac{1}{2}$, and plays no further role in the proof. The middle term is

$$\sum_{p\equiv a \pmod b}\frac{\ln p}{p^z}$$

Following the previous proof of the prime number theorem, define

$$\phi_b(x,a)=\sum_{p\leq x,p\equiv a \pmod b}\ln p$$

Lemma 20.

$$\sum_{p\equiv a \pmod b} \frac{\ln p}{p^z} = z\int_1^\infty \phi_b(t,a)t^{-z-1}\,dt$$

Proof of lemma: In the interval $[n, n+1]$, $\phi_b(t,a)$ is constantly equal to $\phi_b(n,a)$, so the integral is

$$z\sum_{n=1}^\infty \phi_b(n,a)\int_n^{n+1} t^{-z-1}\,dt = z\sum_{n=1}^\infty \phi_b(n,a)\,\frac{t^{-z}}{-z}\Big|_n^{n+1}\,dt = -\sum_{n=1}^\infty \phi_b(n,a)\left(\frac{1}{(n+1)^z}-\frac{1}{n^z}\right)$$

This sum is

$$\phi_b(1,a)\frac{1}{1^z} + \Big(\phi_b(2,a)-\phi_b(1,a)\Big)\frac{1}{2^z} + +\Big(\phi_b(3,a)-\phi_b(2,a)\Big)\frac{1}{3^z} + \ldots = \sum_{p\equiv a \pmod b}\frac{\ln n}{n^z}$$

Returning to the Proof of Dirichlet's Theorem: Now turn to the expression

$$\sum_\varphi -c_\varphi \frac{L'(z,\varphi)}{L(z,\varphi)}$$

If $\varphi \neq 1$, we have proved that $L(z,\varphi)$ is holomorphic for $\Re(z) > 0$ and has no zeros on the line $\Re(z) = 1$. Consequently, $\frac{L'(z,\varphi)}{L(z,\varphi)}$ is holomorphic on an open set which contains the half plane $\Re(z) \geq 1$.

Recall that

$$L(z,1) = \prod_{p|\ b}\left(1-\frac{1}{p^z}\right)\zeta(z)$$

It follows that $-\frac{L'(z,1)}{L(z,1)}$ has a pole at $z = 1$ with residue 1.

The constant $c_1 = \frac{1}{|G|}\int_G \overline{1}f$ where f equals 1 at one element and zero elsewhere. So $c_1 = \frac{1}{\phi(b)}$. It follows that the function below is holomorphic on an open set containing $\Re(z) \geq 1$:

$$\sum -c_\varphi \frac{L'(z,\varphi)}{L(z,\varphi)} - \frac{1}{\phi(b)}\frac{1}{z-1}$$

Putting this together, we conclude that the following function, defined and holomorphic on $\Re(z) > 1$, can be analytically continued to a function holomorphic on an open set containing $\Re(z) \geq 1$:

$$z\int_1^\infty \phi_b(t,a)t^{-z-1}\,dt - \frac{1}{\phi(b)}\frac{1}{z-1}$$

Note that

$$z\int_1^\infty t^{-z}\,dt = z\,\frac{t^{-z+1}}{-z+1}\bigg|_1^\infty = \frac{z}{z-1} = 1+\frac{1}{z-1}$$

We conclude that the following function, define for $\Re(z)>1$, can be analytically continued to a function defined and holomorphic on an open set containing $\Re(z)\geq 1$:

$$z\int_1^\infty \frac{\phi_b(t,a)-\frac{1}{\phi(b)}t}{t^{z+1}}\,dt$$

Reducing to Convergence of an Integral We cannot automatically conclude from the ability to analytically continue this expression that the integral still converges when $z=1$:

$$\int_1^\infty \frac{\phi_b(t,a)-\frac{1}{\phi(b)}t}{t^{2}}\,dt$$

However, if this integral converges, then we conclude exactly as before that $\phi_b(x,a)\sim\frac{1}{\phi(b)}\,x$. The proof is exactly the same as the proof of theorem 135 and need not be repeated here.

The statement $\phi_b(x,a)\sim\frac{1}{\phi(b)}\,x$ implies the statement $\pi_b(x,a)\sim\frac{1}{\phi(b)}\frac{x}{\ln x}$ by the argument given in the proof of theorem 132. Indeed if $0<\epsilon<1$ then

$$\frac{\phi_b(x,a)}{\ln x} = \sum_{p\leq x, p\equiv a \pmod b}\frac{\ln p}{\ln x} \leq \sum_{p\leq x, p\equiv a \pmod b}\frac{\ln x}{\ln x} = \pi_b(x,a)$$

$$= \sum_{x^{1-\epsilon}\leq p\leq x, p\equiv a \pmod b} 1 + \sum_{p<x^{1-\epsilon}, p\equiv a \pmod b} 1$$

$$= \frac{\sum_{x^{1-\epsilon}\leq p\leq x, p\equiv a \pmod b}\ln p}{\ln x^{1-\epsilon}} + x^{1-\epsilon} \leq \frac{\phi_b(x,a)}{(1-\epsilon)\ln x} + x^{1-\epsilon}$$

Taking the most important terms:

$$\frac{\phi_b(x,a)}{\ln x} \leq \pi_b(x,a) \leq \frac{\phi_b(x,a)}{(1-\epsilon)\ln x} + x^{1-\epsilon}$$

Multiply by $\ln x$ and divide by x to get

$$\frac{\phi_b(x,a)}{x} \leq \frac{\pi_b(x,a)}{x/\ln x} \leq \frac{\phi_b(x,a)}{(1-\epsilon)x} + \frac{1}{x^\epsilon}$$

If $\frac{\phi_b(x,a)}{x}\to\frac{1}{\phi(b)}$, then for a fixed $\delta>0$ and large enough x,

$$\frac{1}{\phi(b)}-\delta \leq \frac{\pi_b(x,a)}{x/lnx} \leq \frac{1}{1-\epsilon}\frac{1}{\phi(b)}+\delta$$

This holds for all $0 < \epsilon$ and δ, so

$$\frac{\pi_b(x,a)}{x/\ln x} \to \frac{1}{\phi(b)}$$

.

Turn back to theorem 136, the essential analytic idea behind Newman's proof of the prime number theorem. This theorem applies to the above case provided the integrand of the first integral is bounded. But $\phi_b(x,a) \leq \phi(x) \leq 2x$ so the bound is trivial. This completes the proof.

21.8 Summary

The Prime Number Theorem and Dirichlet's Theorem on Primes in an Arithmetic Progression are among the greatest theorems in mathematics. The proofs we have given involve many details. It is useful to look back and list the key ideas.

- Both the zeta function and Dirichlet's L functions satisfy product formulas. In our proofs, these products was disguised by taking logarithmic derivatives: $-\frac{\zeta'(z)}{\zeta(z)}$ and $-\frac{L'(z,\varphi)}{L(z,\varphi)}$
- The resulting formulas shift attention from $\pi(x) = \sum_p 1$ to $\phi(x) = \sum_p \ln p$ and from $\pi_b(x,a)$ to $\sum_{p \equiv a \pmod b} \ln p$.
- Both ζ and $L(z,\varphi)$ extend to meromorphic functions left of the critical line $\Re(z) = 1$. In the case of ζ, this required an actual analytic continuation. For $L(z,\varphi)$, it required a very delicate convergence argument. The prime number theorem did not require the full analytic continuation or ζ or L to the entire plane.
- The poles of $\zeta(z)$ and $L(z,1)$ at $z = 1$ play crucial roles, and in some sense give the correct order of magnitude of ϕ.
- The proofs depend critically on proving that there are no zeros of ζ or $L(z,\varphi)$ on the critical line.
- Dirichlet's theorem, both in its naive form and its full form, requires proving that $L(1,\varphi) \neq 0$. This step is not implied by the previous item; instead it requires a separate proof.
- In the end, both proofs requiring moving an integral on $\Re(z) > 1$ to the line $\Re(z) = 1$. The conventional proof of this requires the Ikehara Tauberian Theorem, but here is given by Newman's beautiful argument.

Chapter 22

Elliptic Integrals

The chapter following this one is about doubly periodic functions. Many mathematicians in the 19th century contributed to this theory, polishing and reorganizing the results until now they glisten with a sense of logical inevitability.

But the theory was built on earlier strands of mathematics: isolated but surprising examples, specialized integrals, hints of connections between analysis and number theory. This chapter is about those isolated results, discussed informally and meant to be skimmed for enjoyment. We include a few detailed calculations to keep us honest.

This is a chapter about history rather than mathematics.

22.1 De Analysi per Aequationes Infinitas

We begin with the first public paper on the calculus, written by Newton in 1669 and deposited in the Royal Society in London, where it was read by visiting scholars including Leibniz. This paper was published in 1711.

In the first sentence of the paper, Newton states that his recent discoveries will be "briefly explained rather than narrowly demonstrated" and then he summarizes the calculation of areas by three rules, each followed by explicit examples.

Rule 1: The area under x^n is $\frac{x^{n+1}}{n+1}$.

Rule 2: The area under $f + g$ is the sum of the areas under f and g.

Rule 3: But if f is more complicated than just a polynomial, then expand f in an infinite series and integrate term by term.

Immediately following this rule, Newton computes

$$\frac{1}{1+x} = 1 - x + x^2 - ...$$

by using ordinary grade school division, and then integrates term by term to obtain the area under this curve,

$$\log x = x - \frac{x^2}{2} + \frac{x^3}{3} - ...$$

In a private unpublished paper, he computes ln(1.02) to 79 places using this series.

Since expressions involving square roots come up via conic sections, Newton shows that they can be handled the same way. He considers the function $y = \sqrt{1+x^2}$ and thus the hyperbola $y^2 - x^2 = 1$, and expands in an infinite series using the method of computing square roots by hand which I still learned in grade school:

$$\sqrt{1+x^2} = 1 + \frac{x^2}{2} - \frac{x^4}{8} + ...$$

Integration term by term gives the area:

$$\text{area} = x + \frac{x^3}{6} - \frac{x^5}{40} + ...$$

After writing the first paper, Newton discovered the binomial theorem for non-integer exponents, greatly simplifying this calculation.

More pages on integration follow, and Newton's method of solving equations numerically comes up in the bargain. Newton then writes "Let these observations on investigating the areas of curves suffice. Indeed, since every problem on the length of curves, the quantity and surface of solids and the centre of gravity may ultimately be reduced to an inquiry into the quantity of a plane area bounded by a curved line, there is little necessity to comment further. But I will speak very briefly about these matters."

So in this amazing paragraph, Newton tells us that calculus is a *general method*, not just an algorithm for computing areas. He immediately turns to lengths of curves, essentially obtaining the formula

$$\text{length} = \int \sqrt{(dx)^2 + (dy)^2} = \int \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\, dx$$

One of his initial examples is the circle $y = \sqrt{1-x^2}$, whose length between 0 and x turns out to be $\int \frac{1}{\sqrt{1-x^2}}\, dx$. Expanding and integrating term by term gives

$$\text{length} = x + \frac{1}{6}x^3 + \frac{3}{40}x^5 + \frac{5}{112}x^7 + ...$$

Of course the length of the circle from 0 to x is $\theta = \arcsin x$ where θ is the angle from the y-axis to the point on the circle, measured in radians. Astonishingly, Newton then instructs us to solve this equation for x in terms of θ, getting

$$\sin\theta = \theta - \frac{1}{3!}\theta^3 + \frac{1}{5!}\theta^5 - \frac{1}{7!}\theta^7 + ...$$

Newton writes "let it be noted that when you know 5 or 6 terms of those roots you will be able to prolong them at will by observing analogies," and his modern editor, D. T. Whiteside, writes "The series for the sine and cosine appear here for the first time in a European manuscript."

A footnote in Whiteside's edition tells us that Indian mathematicians had obtained the series one hundred years earlier by a completely different method.

I first read this paper while preparing a talk for high school mathematics teachers on the history of calculus. Then I decided that I ought to know something about Newton the man, so I went to the University of Oregon main library and looked for a biography in the card catalog. To my surprise, the catalog listed some books in Latin by Newton, and after investigating I discovered that the University of Oregon owns an original edition of the first 1711 publication of Newton's paper. We obtained it around 1930 for about $2.

22.2 The Length of an Ellipse

Newton's techniques allow us to compute the length of an ellipse. Recall the standard formula for an ellipse:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

We always assume $a > b$. It is easy to calculate the area of such an ellipse. The area of a circle of radius 1 is π. Magnify in the x direction by a and in the y direction by b; the resulting object is the ellipse. The first magnification changes the area to πa and the second one changes it to πab.

Unfortunately, magnification in only one direction has a complicated effect on lengths, so there is no simple formula for the length of an ellipse. But we can apply Newton's formula to $y = b\sqrt{1 - \frac{x^2}{a^2}}$:

$$\int_0^x \sqrt{1 + \left(\frac{b\left(\frac{-2x}{a^2}\right)}{2\sqrt{1-\frac{x^2}{a^2}}}\right)^2} = \int_0^x \sqrt{1 + \frac{b^2x^2}{a^4 - a^2x^2}} = \int_0^x \sqrt{\frac{a^4 - (a^2-b^2)x^2}{a^4 - a^2x^2}}$$

Assume $a \geq b$ and define

$$k = \sqrt{\frac{a^2 - b^2}{a^2}}$$

This number is called the *eccentricity* of the ellipse. It does not depend on the size of the ellipse, since multiplying both a and b by $\lambda > 0$ leaves it unchanged. Notice that $0 \leq k \leq 1$, where $k = 0$ corresponds to a circle and $k = 1$ corresponds to a line without interior.

The length of the ellipse is then

$$\int \sqrt{\frac{a^2 - k^2x^2}{a^2 - x^2}}\, dx$$

Substitute $x = au$ to obtain the final result

$$a \int \frac{\sqrt{1 - k^2u^2}}{\sqrt{1 - u^2}}\, du = a \int \frac{1 - k^2u^2}{\sqrt{(1 - u^2)(1 - k^2u^2)}}\, du$$

This integral is called an *elliptic integral of the second kind*, because it gives the length of an ellipse and because it is one of three kinds of such integrals classified by Legendre. Giant tables of this integral can be found in the *Handbook of Chemistry and Physics*, and later as one of the functions built into *Mathematica*. Notice that the integral depends on *two* parameters, the top limit of integration u and the eccentricity k.

22.3 Explicit Integration

Recall that by definition a *rational function* is a quotient of polynomials, like $\frac{2x^3+5x+2}{x^2+x+1}$. One of the beautiful theorems of calculus states that every rational function can be explicitly integrated. This isn't obvious because

$$\int \frac{1}{x}\, dx = \log x \qquad \text{and} \qquad \int \frac{1}{1 + x^2}\, dx = \arctan x$$

The theorem is easier to state and prove in the complex case:

Theorem 146. *Any rational function can be integrated explicitly as a sum of rational functions and logarithms.*

Proof: Let $R(z)$ be rational. The denominator $Q(z)$ has only finitely many zeros, so R has only finitely many singularities. Subtract from R the principal part of R at each of these singularities. The result is a rational function with no finite poles, and thus a polynomial. So we can write

$$R(z) = P(z) + \sum_k \sum_{i=1}^{n_k} a_{ki} \frac{1}{(z - \lambda_{ki})^i}$$

This is just the partial fraction decomposition of R used in calculus. Every term in this expression has an explicit rational integral except the terms of the form $\frac{1}{z-\lambda}$, whose integrals are logarithms. QED.

Remark: In calculus we would write $\int \frac{dx}{1+x^2} = \arctan x$. In complex analysis, this is easier:

$$\int \frac{dz}{(z-i)(z+i)} = \frac{1}{2i}\int \left(\frac{1}{z-i} - \frac{1}{z+i}\right) = \frac{1}{2i}\ln(z-i) - \frac{1}{2i}\ln(z+i)$$

The reader can show that this expression can be rewritten to give the calculus expression.

Remark: The previous theorem is usually extended in calculus by adding square roots of quadratics to the mix. Let $y = \sqrt{a+bx+cx^2}$ and let R be a rational function of x and y. Expressions of this type were already integrated in Newton's first paper, and can always be explicitly integrated. We restrict attention to the special case when the radical is $\sqrt{1-x^2}$ because everything else is trivial algebra. Over the reals, $\sqrt{1-x^2}$ only makes sense when $-1 < x < 1$, so we can substitute $x = \sin u$ and then $\sqrt{1-x^2} = \cos u$ and $dx = \cos u\, du$. Thus it suffices to integrate a rational expression in $\sin u$ and $\cos u$.

Now comes what Spivak calls "the world's sneakiest substitution". Substitute

$$\sin u = \frac{2t}{1+t^2} \qquad \cos u = \frac{1-t^2}{1+t^2} \qquad du = \frac{2\,dt}{1+t^2}$$

to convert the integral to the integral of a rational function and reduce to the previous case. Where does this substitute come from? It comes from the following picture.

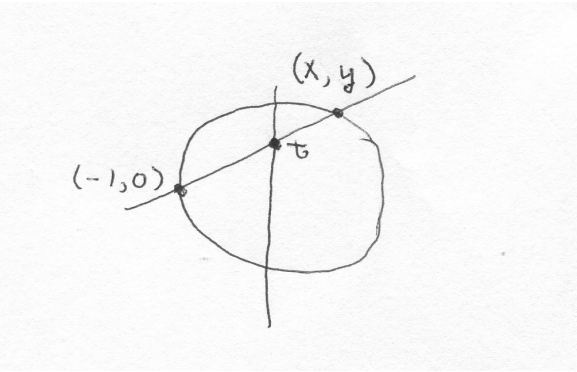


Figure 22.1: Rational Parametrization of Circle

The number t is allowed to be any point on the y-axis, not necessarily inside the circle. As t moves, the point (x, y) traces out each point on the circle except $(-1, 0)$. To find (x, y), it is

only necessary to solve the equations $x^2 + y^2 = 1$ and $y = t(x + 1)$. Substituting the second equation in the first one gives a quadratic equation in x, but we already know that $x = -1$ is one solution, so the other solution can be obtained by factoring.

Notice that if x and y are rational, the slope t will be rational. Conversely if t is rational, we get x by solving a quadratic equation with rational coefficients. But $x = -1$ is one solution, so the other solution is also rational. Thus we obtain a one-to-one correspondence between all rational numbers t and all points (x, y) on the circle with rational coefficients except $(-1, 0)$.

If (x, y) is on the circle and $x = \frac{a}{c}, y = \frac{b}{c}$, then $a^2 + b^2 = c^2$. So our map essentially tells us how to find all right triangles with integer sides: just choose $a = t^2 - 1, b = 2t, c = t^2 + 1$. Precise details can be worked out by the reader or found in any book on number theory. This parametrization is ancient. Large tables of Pythagorian triples have been found in cuneiform tablets, and it seems likely that the authors created them using rules essentially equivalent to our parametrization.

22.4 Legendre

Since integrals of rational expressions in x and the square root of a quadratic can be explicitly found, it was natural to next consider rational expressions in x and the square root of a higher order polynomial. We earlier found that the length of ellipse is given by an integral involving two different square roots of quadrics, or equivalently one square root of a quartic. So mathematicians at first restricted attention to the case when the higher order polynomial has degree 3 or 4. In hindsight that was a lucky choice, because these integrals have special properties not shared when the degree of the polynomial is larger than 4.

Interest in these integrals, called *elliptic integrals*, increased dramatically after a discovery by an Italian mathematician named Fagnano, which caught the attention of Euler. This discovery will be explained in the following sections.

In 1825 through 1832, Legendre published three volumes of a work titled *Traité des fonctions elliptiques et des intégrales eulériennes* in which he showed that elliptic integrals could always be transformed to a combination of three specific forms, known as elliptic integrals of the first kind, of the second kind, and of the third kind. These integrals have a parameter called the *modulus* which corresponds to the eccentricity of the ellipse; the length of a very flat ellipse differs from the length of an almost circular ellipse.

To give a flavor of these results, the elliptic integral of the first kind has the form

$$\int_0^r \frac{dx}{\sqrt{(1-x^2)(1-mx^2)}}$$

where m is the modulus. Notice that $m = 0$ gives $\arcsin(r)$ and thus trigonometric functions.

Care must be taken when consulting tables of these integrals because the notation is sometimes changed in subtle ways. For example, elliptic integrals of the first kind are built into *Mathematica* using the notation

$$\text{EllipticF}[\phi, m] = \int_0^{\sin(\phi)} \frac{dx}{\sqrt{(1-x^2)(1-mx^2)}}$$

However, shortly after these volumes were published, the field was dramatically altered by Jacobi, Abel, and later Weierstrass, who discovered that the inverses of functions given by elliptic integrals are much more tractable and interesting. This will also be explained shortly.

22.5 Gauss and Dividing the Circle

One of the most important books in mathematics is Gauss' book on number theory, *Disquisitiones Arithmeticae.* Written when he was 24 years old in 1801, this book contains the definition of congruence, the first proof of the law of quadratic reciprocity, the introduction of Gauss sums, the theory of binary forms, and much, much more.

About fifteen years ago, an academic named George Van Schach retired to Eugene. He was trained as a mathematician, but spent his career as a librarian in a biology department. Van Schach liked to prowl the UO library, and one day he was allowed to visit the storage area in the basement where books are kept that are no longer in demand but not disposable. There, Van Schach discovered an original edition of Gauss' book on the shelves! It had a UO library card and everything.

The book is in Latin; it is small and unobtrusive because the publisher didn't know that Gauss would become Europe's greatest mathematician. You can find this original edition in the University of Oregon Library's rare book room. There is a modern English translation published by Yale University Press.

Gauss' book has seven chapters and the most unusual is the last, on trigonometry. Gauss writes

> Amont the splendid developments contributed by modern mathematicians, the theory of circular functions holds a most important place. Since the most brilliant modern mathematicians by their industry and shrewdness have formulated for it an extensive discipline, we can hardly expect that any part of the theory, particularly the elements, can be significantly expanded. I will speak of the theory of trigonometric functions as related to arcs which are commensurable with the circumference, or of the theory of regular polygons. Only a small part of this theory has been developed so far, as the present section will make clear. The reader might be surprised to find a discussion of this subject in the present work which deals

> with a discipline apparently so unrelated, but the treatment itself will make abundantly clear that there is an intimate connection between this subject and higher arithmetic.

The second paragraph of chapter seven contains a mysterious statement. Gauss writes

> The principles of the theory which we are going to explain actually extend much farther than we will indicate. For they can be applied not only to circular functions but just as well to other transcendental functions, e.g. to those which depend on the integral $\int \frac{1}{\sqrt{1-x^4}}\,dx$. Since we are preparing a substantial work on transcendental functions, we have decided to consider only circular functions here.

Gauss never finished his work on transcendental functions. But in 1829, Abel published a 200 page paper on elliptic functions, and Gauss acknowledged that his own results could be found in Abel's paper. Sadly, Abel died in 1829.

The main result of Gauss' chapter seven is a determination of all integers n such that an n-sided polygon can be inscribed in a circle using straightedge and compass. Gauss shows that such a construction is possible if

$$n = 2^k p_1 p_2 \ldots p_l$$

where the p_i are distinct odd primes, each of the form $2^{2^m} + 1$. At the end of the chapter he states that no other case is constructible. So the cases n = 7 and 9 are not constructible, but 17 is. The list of such primes begins $3, 5, 17, 257, 65537, \ldots$ and in fact no other primes on this list are known. There are plausibility arguments that the list should be finite and contain roughly four to five elements.

The theory of regular polygons inscribed in a circle is equivalent to the theory of algebraic values of $\cos\left(\frac{2\pi}{n}\right)$. We can inscribe an octogon because $\cos\frac{\pi}{4} = \frac{1}{\sqrt{2}}$; we can inscribe a hexagon because $\cos\frac{\pi}{3} = \frac{1}{2}$, and we can inscribe a pentagon because $\cos\frac{2\pi}{5} = \frac{\sqrt{5}-1}{4}$. Near the end of the chapter, Gauss writes "it is certainly astonishing that although the geometric divisibility of the circle into three and five parts was already known in Euclid's time, nothing was added to this discovery for 2000 years," and he proudly displays the value

$$\cos\frac{\pi}{17} = \frac{\sqrt{17}-1}{16} + \frac{\sqrt{34-2\sqrt{17}}}{16} + \frac{\sqrt{17+3\sqrt{17}-\sqrt{34-2\sqrt{17}}-2\sqrt{34+2\sqrt{17}}}}{8}$$

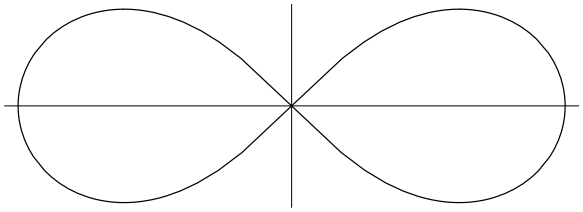
We are about to explain that mysterious elliptic integral $\int \frac{1}{\sqrt{1-x^4}}\,dx$.

22.6 The Lemniscate

Recall that an ellipse with foci p_1 and p_2 is the set of points for which the sum of the distances to the foci is constant. Analogously, we can consider the curve obtained by requiring that the *product* of these distances be constant. If we place the foci at $(\pm a, 0)$ and call the constant c, we obtain the equation $\sqrt{(x+a)^2+y^2}\cdot\sqrt{(x-a)^2+y^2} = c$ or equivalently

$$[(x^2+y^2+a^2)+2ax]\cdot[(x^2+y^2+a^2)-2ax] = c^2.$$

It turns out that the shape of this curve depends on the constant. When c is small, we obtain two ovals surrounding the foci. As the constant increases, these ovals enlarge and merge into one curve. At the exact moment when they merge, we obtain a figure eight shaped curve known as the *lemniscate*. It was introduced by Jacob Bernoulli in 1694.



By symmetry, the ovals merge at the origin and this occurs when $c = a^2$. So the equation of the lemniscate is $[(x^2+y^2+a^2)+2ax]\cdot[(x^2+y^2+a^2)-2ax] = a^4$ or $(x^2+y^2+a^2)^2-4a^2x^2 = a^4$. Expanding, we obtain $\left(x^2+y^2\right)^2 + 2a^2(x^2+y^2) - 4a^2x^2 = 0$ and so

$$\left(x^2+y^2\right)^2 = 2a^2(x^2-y^2)$$

In polar coordinates, this becomes $r^4 = 2a^2r^2\cos 2\theta$. Changing the value of a shrinks or magnifies the lemniscate, so it is convenient to choose the particular value $a = \frac{1}{\sqrt{2}}$. Then the polar form becomes

$$r^2 = \cos 2\theta.$$

We easily deduce

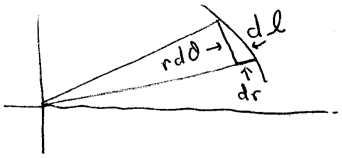
$$(x,y) = \left(\sqrt{\frac{r^2+r^4}{2}}, \sqrt{\frac{r^2-r^4}{2}}\right)$$

Thus if we know a point (x,y) on the lemniscate, we can compute $r = \sqrt{x^2+y^2}$ with straightedge and compass; conversely, knowing r, we can compute x and y. For this reason, we will think of r as the central variable of the theory and gradually eliminate θ.

22.7 The Length of the Lemniscate

Trigonometry is intimately connected to the constant π which gives the length of half of a circle. Similarly the theory of the lemniscate involves a constant ω which gives the length of one lobe of the lemniscate. Using calculus we will calculate this constant.

The picture below shows a typical curve length calculation in polar coordinates.



$$\text{length} = \int \sqrt{(dr)^2 + (rd\theta)^2}$$

Since we want to use r as independent variable, we write this as

$$\text{length} = \int \sqrt{1 + \left(r\frac{d\theta}{dr}\right)^2}\, dr$$

Differentiating $r^2 = \cos 2\theta$ with respect to r gives $2r = -2\sin\theta \frac{d\theta}{dr}$ and thus $r\frac{d\theta}{dr} = \frac{-r^2}{\sin 2\theta}$. So

$$1 + \left(r\frac{d\theta}{dr}\right)^2 = 1 + \frac{r^4}{\sin^2 2\theta} = 1 + \frac{r^4}{1-\cos^2 2\theta} = 1 + \frac{r^4}{1-(r^2)^2} = \frac{1}{1-r^4}$$

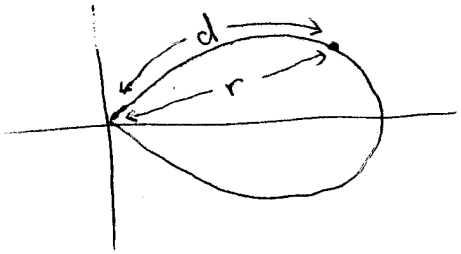
Therefore

$$\text{length} = \int \sqrt{1 + \left(r\frac{d\theta}{dr}\right)^2}\, dr = \int \frac{1}{\sqrt{1-r^4}}\, dr$$

So the integral Gauss mentions in chapter seven gives the length of a lemniscate.

We can use this same integral to find the length of any portion of the lemniscate. We summarize this with a picture. The illustration below shows the radius r and the distance d along the lemniscate to the point determined by r, and these numbers are related by a transcendental equation

$$d = \int_0^r \frac{dx}{\sqrt{1-x^4}}$$



22.8 Fagnano

The theory of the lemniscate was transformed in 1718 when an Italian count named Fagnano made an astonishing discovery. Fagnano tried to compute the integral $\int \frac{dx}{\sqrt{1-x^4}}$. He failed. But he discovered something interesting along the way.

Fagnano's idea was to modify the world's sneakiest transformation slightly for the lemniscatic integral:

$$x^2 = \frac{2t^2}{1+t^4}$$

Then $\sqrt{1-x^4} = \sqrt{1-\frac{4t^4}{(1+t^4)^2}} = \frac{1-t^4}{1+t^4}$ and $2x\frac{dx}{dt} = \frac{4t(1-t^4)}{(1+t^4)^2}$ and so $dx = \frac{1}{x}\frac{2t(1-t^4)}{(1+t^4)^2}dt = \frac{\sqrt{1+t^4}}{\sqrt{2}t}\frac{2t(1-t^4)}{(1+t^4)^2}dt$ or simplifying $dx = \sqrt{2}\frac{1-t^4}{(1+t^4)^{3/2}}dt$. Therefore

$$\int_0^r \frac{1}{\sqrt{1-x^4}}\,dx = \sqrt{2}\int_0^s \frac{1+t^4}{1-t^4}\cdot\frac{1-t^4}{(1+t^4)^{3/2}}dt = \sqrt{2}\int_0^s \frac{1}{\sqrt{1+t^4}}dt$$

provided

$$r^2 = \frac{2s^2}{1+s^4}.$$

Fagnano's substitution has converted the integral into an equally difficult integral. Undeterred, he tried the same trick on the new integral, using the slightly modified substitution

$$t^2 = \frac{2u^2}{1-u^4}$$

Then $\sqrt{1+t^4} = \sqrt{1+\frac{4u^4}{(1-u^4)^2}} = \frac{1+u^4}{1-u^4}$ and $2t\frac{dt}{du} = \frac{4u(1+u^4)}{(1-u^4)^2}$ and so $dt = \frac{1}{t}\frac{2u(1+u^4)}{(1-u^4)^2}du = \frac{\sqrt{1-u^4}}{\sqrt{2}u}\frac{2u(1+u^4)}{(1-u^4)^2}du$ or simplifying $dt = \sqrt{2}\frac{1+u^4}{(1-u^4)^{3/2}}du$. Therefore

$$\int_0^s \frac{1}{\sqrt{1+t^4}}dt = \sqrt{2}\int_0^v \frac{1-u^4}{1+u^4}\cdot\frac{1+u^4}{(1-u^4)^{3/2}}\,du = \sqrt{2}\int_0^v \frac{1}{\sqrt{1-u^4}}du$$

where

$$s^2 = \frac{2v^2}{1-v^4}$$

We seem to be back where we started. But wait. Putting the two substitutions together, we obtain

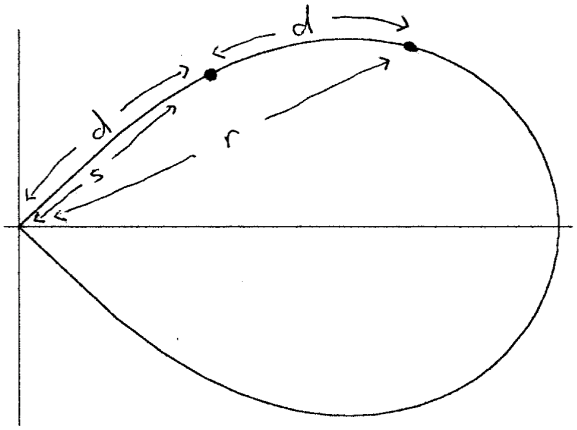
$$\int_0^r \frac{1}{\sqrt{1-r^4}}dr = 2\int_0^v \frac{1}{\sqrt{1-u^4}}du$$

where $r^2 = \frac{2u^2}{1+u^4}$ and $u^2 = \frac{2v^2}{1-v^4}$ and so $r^2 = \frac{4\frac{v^2}{1-v^4}}{1+\frac{4v^4}{(1-v^4)^2}} = \frac{4v^2(1-v^4)}{(1+v^4)^2}$ or $r = \frac{2v\sqrt{1-v^4}}{1+v^4}$. Rewriting so the variable of integration is always x, we obtain

Theorem 147 (Fagnano). *If* $r = \frac{2s\sqrt{1-s^4}}{1+s^4}$, *then*

$$\int_0^r \frac{dx}{\sqrt{1-x^4}} = 2\int_0^s \frac{dx}{\sqrt{1-x^4}}$$

This has an astonishing consequence. Suppose we are given a point on the lemniscate with distance s to the origin. Let d be the length along the lemniscate to this point; we can only approximate d since it is given by a transcendental integral. Continue along the lemniscate an additional distance d, obtaining a new point whose distance along the lemniscate is double the original distance. This new distance can only be approximated. *But the new point can be constructed from the original point using a straightedge and compass!* Indeed r is determined from s by an equation which only involves addition, subtraction, multiplication, division, and square roots. Thus we can "double an arc of the lemniscate" with straightedge and compass.

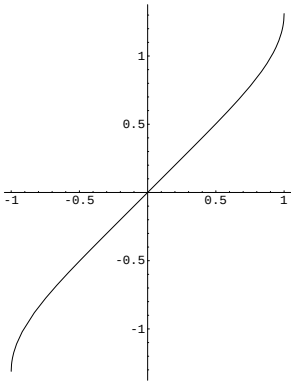


In exactly the same way, we can "bisect an arc of the lemniscate" since the equation $r = \frac{2s\sqrt{1-s^4}}{1+s^4}$ can be solved for s in terms of r and the solution involves only elementary operations and square roots.

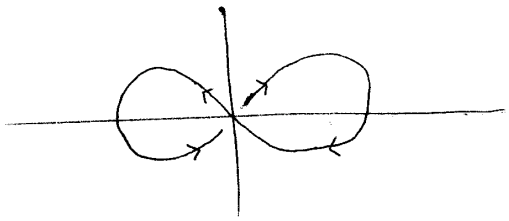
22.9 Inverting the Length Integral

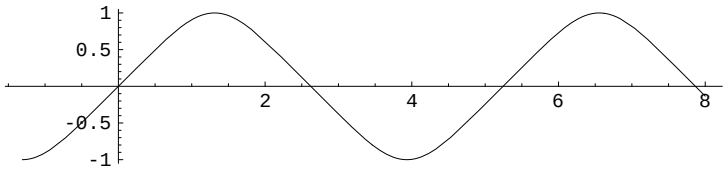
Three people worked on elliptic function theory independently: Gauss, Abel, and Jacobi. Each of them had the same wonderful idea: "elliptic integrals define functions, but nicer functions can be obtained by taking of the inverse of the integral."

Recall the integral $d = \int_0^r \frac{dx}{\sqrt{1-x^4}}$ from section three. This integral defines a function $[-1,1] \to [-\frac{\omega}{2}, \frac{\omega}{2}]$ whose derivative is $\frac{1}{\sqrt{1-x^4}}$ and so positive. Therefore this function is one-to-one and onto.



Call the inverse of the function $\mathrm{sn}(x)$, so $r = \mathrm{sn}(d)$. This is Jacobi's notation. Gauss wrote sinlemn(d) and Abel wrote $\phi(d)$. If we trace the lemniscate as in the first picture below, then d is defined for all real numbers, and the number r depends on d in a periodic way. It is natural to let r be negative for points on the left side of the lemniscate. Thus our inverse can be extended periodically to all d, as shown in the second picture below.





Geometrically, the new function starts with a distance d along the lemniscate, perhaps determined with a tape measure. The function $\mathrm{sn}(d)$ gives the distance r from the endpoint to the origin r. Recall that $(x, y) = \left(\sqrt{\frac{r^2+r^4}{2}}, \sqrt{\frac{r^2-r^4}{2}}\right)$, so in a sense $\mathrm{sn}(d)$ gives the coordinates of the resulting endpoint on the lemniscate. Thus $\mathrm{sn}(d)$ is the analogue $\cos\theta$ and $\sin\theta$ on the unit circle, since θ in radians represents a distance along the circle and $\cos\theta$ and $\sin\theta$ give the endpoint of this arc.

Finally, we want to translate Fagnano's result into an analogous result for $\mathrm{sn}(x)$. According to Fagnano,

$$\int_0^r \frac{dx}{\sqrt{1-x^4}} = 2\int_0^s \frac{dx}{\sqrt{1-x^4}}$$

provided

$$r = \frac{2s\sqrt{1-s^4}}{1+s^4}$$

Apply this equation when $s = \mathrm{sn}(d)$ for some d. Since sn is inverse to the integral, we have $d = \int_0^s \frac{dx}{\sqrt{1-x^4}}$ and so the right side of the first equation is $2d$. Therefore $\int_0^r \frac{dx}{\sqrt{1-x^4}} = 2d$, so $r = \mathrm{sn}(2d)$. Putting these values in the second equation, we have $\mathrm{sn}(2d) = \frac{2\,\mathrm{sn}(d)\sqrt{1-\mathrm{sn}^4(d)}}{1+\mathrm{sn}^4(d)}$.

Definition 49. $cn(x) = \sqrt{1 - sn^4(x)}$

Theorem 148 (Fagnano)**.**

$$sn(2x) = \frac{2sn(x)cn(x)}{1 + sn^4(x)}$$

Remark: Notice that this is analogous to the double angle formula $\sin(2x) = 2\sin(x)\cos(x)$ in trigonometry!

22.10 Inscribing a Square in a Lemniscate

Suppose we wish to inscribe an octagon in a circle with straightedge and compass. We need to construct the point with angle $\theta = \frac{\pi}{4}$; it suffices to construct $\sin\theta$. If we know no geometry, we can still compute this value with the double angle formula. Indeed $\sin 2\theta = 1$, so $2\sin\theta\cos\theta = 1$. Squaring, $4\sin^2\theta(1 - \sin^2\theta) = 1$ and so $\sin\theta$ satisfies the equation

$$\sin^4\theta - \sin^2\theta + \frac{1}{4} = \left(\sin^2\theta - \frac{1}{2}\right)^2 = 0$$

Thus $\sin\theta = \frac{1}{\sqrt{2}}$, a value which can be constructed with straightedge and compass. In chapter 7 of his book, Gauss constructed regular polygons with n sides using a similar analysis of equations satisfied by $\sin\frac{2\pi}{n}$.

We can do the same thing for the lemniscate using Fagnano's formula. An octagon in the full lemniscate is the same thing as an equal division of one lobe of the lemniscate into four equal pieces. If x is the length of one of these pieces, then $\mathrm{sn}(2x) = 1$, so

$$\frac{2\mathrm{sn}(x)\mathrm{cn}(x)}{1 + \mathrm{sn}^4(x)} = 1$$

Squaring and moving the denominator to the other side gives $4\mathrm{sn}^2x(1 - \mathrm{sn}^4x) = (1 + \mathrm{sn}^4x)^2$ or

$$\mathrm{sn}^8x + 4\mathrm{sn}^6x + 2\mathrm{sn}^4x - 4\mathrm{sn}^2x + 1 = 0$$

Miraculously, this polynomial in sn^2x factors:

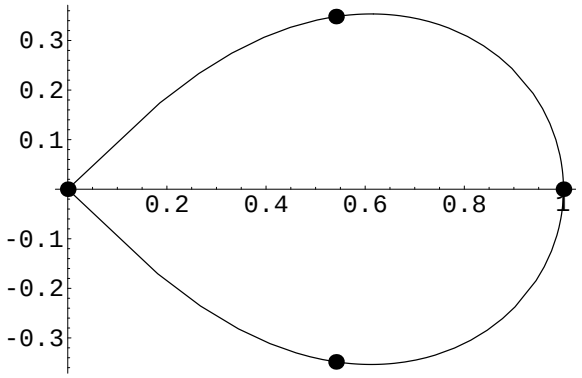
$$(\mathrm{sn}^4x + 2\mathrm{sn}^2x - 1)^2 = 0$$

So $\mathrm{sn}^2x = -1 \pm \sqrt{2}$. Since it is positive, $\mathrm{sn}^2x = \sqrt{2} - 1$ and

$$r = \mathrm{sn}\ x = \sqrt{\sqrt{2} - 1}.$$

Recall that $(x,y) = \left(\sqrt{\frac{r^2+r^4}{2}}, \sqrt{\frac{r^2-r^4}{2}}\right)$. So

$$(x,y) = \left(\frac{(\sqrt{2}-1)+(\sqrt{2}-1)^2}{2}, \frac{(\sqrt{2}-1)-(\sqrt{2}-1)^2}{2}\right) = \left(\frac{2-\sqrt{2}}{2}, \frac{3\sqrt{2}-4}{2}\right)$$



22.11 Euler

We have been describing the work of Fagnano in 1718 as interpreted by Jacobi and Abel in 1829. But something very important happened between those two dates, when the subject was picked up by Euler in 1752.

Fagnano's formula generalizes the double angle formula $\sin(2x) = 2\sin x \cos x$. But in trigonometry, this is just a special case of the addition formula $\sin(x+y) = \sin x \cos y + \cos x \sin y$. Euler discovered that Fagnano's result is similarly a special case of an addition formula:

Theorem 149 (Euler)**.**

$$sn(x+y) = \frac{sn(x)cn(y) + cn(x)sn(y)}{1 + sn^2(x)sn^2(y)}$$

Remark: Here is what Andre Weil says in his wonderful book *Number Theory* about this development:

> Euler had closely followed contemporary work on elliptic integrals. On 23 December 1751 the two volumes of Fagnano's *Produzioni Matematiche*, just published, reached the Berlin Academy and were handed over to Euler; the second volume contained reprints of pieces on elliptic integrals which had appeared between 1714 and 1720 in an obscure Italian journal and had remained totally unknown.

On reading these few pages, Euler caught fire instantly; on 27 January 1752 he was presenting to the Academy a memoir with an exposition of Fagnano's main results.

The most striking of Fagnano's results concerned transformations of the "lemniscatic differential"

$$\omega(z) = \frac{dz}{\sqrt{1-z^4}};$$

how he had reached them was more than even Euler could guess. "Surely his discoveries would shed much light on the theory of transcendental functions", Euler wrote in 1753, "if only his procedure supplied a sure method for pursuing these investigations further; but it rests upon substitutions of a tentative character, almost haphazardly applied ..."

With characteristic generosity Euler never ceased to acknowledge his indebtedness to Fagnano, but surely none but Euler would have seen in Fagnano's isolated results the germ of a new branch of analysis. His first contribution was to extend Fagnano's duplication formula for the lemniscate to a general multplication formula.

Proof: We'd now like to give Euler's proof of the addition formula. Euler did not know about the inverse $\mathrm{sn}(x)$ of the elliptic integral, so his results were stated and proved as equations involving integrals. Recall that Fagnano related the upper limits of two integrals via $r = \frac{2u\sqrt{1-u^4}}{1+u^4}$. Euler guessed that the generalization would relate the upper limits of three integrals via the formula

$$r = \frac{u\sqrt{1-v^4} + v\sqrt{1-u^4}}{1+u^2v^2}.$$

In this equation, think of r as a fixed number, and imagine that the equation determines v as a function of u. Differentiate both sides with respect to u. The equation is symmetrical in u and v and we'd like to preserve this symmetry, so whenever we differentiate u we will write du and whenever we differentiate v we will write dv. In the end we will divide everything by du, so the derivative of u will become $\frac{du}{du} = 1$ and the derivative of v will become $\frac{dv}{du}$ and everything will be fine.

We obtain

$$0 = \frac{(1+u^2v^2)\left(du\sqrt{1-v^4} + u\frac{1}{2\sqrt{1-v^4}}(-4v^3dv) + dv\sqrt{1-u^4} + v\frac{1}{2\sqrt{1-u^4}}(-4u^3du)\right)}{(1+u^2v^2)^2}$$

$$- \frac{(u\sqrt{1-v^4} + v\sqrt{1-u^4})\left(2uv^2du + 2u^2vdv\right)}{(1+u^2v^2)^2}$$

Factoring out terms with du and terms with dv, we obtain

$$0 = \frac{du}{\sqrt{1-u^4}}\left[(1+u^2v^2)\left(\sqrt{1-u^4}\sqrt{1-v^4}-2u^3v\right)-2\left(u\sqrt{1-u^4}\sqrt{1-v^4}+v-vu^4\right)uv^2\right]$$

$$+\ \frac{dv}{\sqrt{1-v^4}}\left[(1+u^2v^2)\left(\sqrt{1-u^4}\sqrt{1-v^4}-2u\,v^3\right)-2\left(v\sqrt{1-u^4}\sqrt{1-v^4}+u-uv^4\right)u^2v\right]$$

We claim the two expressions in square brackets are equal. This can be left to the reader; most terms are symmetrical and so equal, and the others are easily equated. Consequently we can cancel the terms in square brackets and obtain $0 = \frac{du}{\sqrt{1-u^4}} + \frac{dv}{\sqrt{1-v^4}}$. Divide by du to obtain

$$0 = \frac{1}{\sqrt{1-u^4}} + \frac{dv/du}{\sqrt{1-v^4}}$$

Replace u by x and v by $v(x)$ and integrate from $x = 0$ to $x = u$:

$$0 = \int_0^u \frac{dx}{\sqrt{1-x^4}} + \int_0^u \frac{v'(x)\,dx}{\sqrt{1-v^4}}$$

In the second integral, substitute $y = v(x)$:

$$0 = \int_0^u \frac{dx}{\sqrt{1-x^4}} + \int_{v(0)}^{v(u)} \frac{dy}{\sqrt{1-y^4}}$$

Here y is a dummy variable, so it can safely be replaced by x. From the equation $r = \frac{u\sqrt{1-v^4}+v\sqrt{1-u^4}}{1+u^2v^2}$ we discover that $v(0) = r$, so

$$0 = \int_0^u \frac{dx}{\sqrt{1-x^4}} + \int_r^v \frac{dx}{\sqrt{1-x^4}}$$

or equivalently

$$\int_0^r \frac{dx}{\sqrt{1-x^4}} = \int_0^u \frac{dx}{\sqrt{1-x^4}} + \int_0^v \frac{dx}{\sqrt{1-x^4}} \quad \text{if} \quad r = \frac{u\sqrt{1-v^4}+v\sqrt{1-u^4}}{1+u^2v^2}$$

The addition formula we stated follows from this integral equation exactly as the doubling formula for $\mathrm{sn}(x)$ follows from Fagnano's integral equation.

22.12 Inscribing a Triangle in the Lemniscate

Suppose we wish to inscribe a hexagon in a circle. This involves constructing the point $\theta = \frac{2\pi}{6}$ or equivalently $\sin\frac{2\pi}{6}$. If we do not know geometry, we can obtain this value from the addition formula for sin . Indeed $\sin(3\theta) = 0$, so $\sin 2\theta\cos\theta + \cos 2\theta\sin\theta = 0$ and therefore

$$(2\sin\theta\cos\theta)\cos\theta + (1 - 2\sin^2\theta)\sin\theta = 0.$$

Since $\sin\theta \neq 0$, we obtain

$$2\cos^2\theta + (1 - 2\sin^2\theta) = 2(1 - \sin^2\theta) + (1 - 2\sin^2\theta) = 3 - 4\sin^2\theta = 0,$$

so $\sin^2\theta = \frac{3}{4}$ and $\sin\theta = \frac{\sqrt{3}}{2}$, a number which is constructible by straightedge and compass.

Similarly, suppose we want to divide the lemniscate into six equal distances, which is equivalent to dividing a lobe of the lemniscate into three equal distances. If x equals any of these distances, then $\mathrm{sn}(3x) = 0$, so

$$\frac{\mathrm{sn}(2x)\mathrm{cn}(x) + \mathrm{cn}(2x)\mathrm{sn}(x)}{1 + \mathrm{sn}^2(2x)\mathrm{sn}^2(x)} = 0$$

So $\mathrm{sn}(2x)\mathrm{cn}(x) = -\mathrm{cn}(2x)\mathrm{sn}(x)$. Squaring, $\mathrm{sn}^2(2x)\left(1 - \mathrm{sn}^4(x)\right) = \left(1 - \mathrm{sn}^4(2x)\right)\mathrm{sn}^2(x)$. By the duplication formula,

$$\left(\frac{2\mathrm{sn}(x)\mathrm{cn}(x)}{1 + \mathrm{sn}^4(x)}\right)^2 \left(1 - \mathrm{sn}^4(x)\right) = \left(1 - \left[\frac{2\mathrm{sn}(x)\mathrm{cn}(x)}{1 + \mathrm{sn}^4(x)}\right]^4\right)\mathrm{sn}^2(x).$$

Cancelling $\mathrm{sn}^2(x)$, we obtain

$$\frac{4\mathrm{cn}^2(x)}{\left(1 + \mathrm{sn}^4(x)\right)^2}\left(1 - \mathrm{sn}^4(x)\right) = 1 - \frac{16\mathrm{sn}^4(x)\mathrm{cn}^4(x)}{\left(1 + \mathrm{sn}^4(x)\right)^4}$$

and so

$$4\left(1 - \mathrm{sn}^4(x)\right)^2\left(1 + \mathrm{sn}^4(x)\right)^2 = \left(1 + \mathrm{sn}^4(x)\right)^4 - 16\mathrm{sn}^4(x)\left(1 - \mathrm{sn}^4(x)\right)^2$$

Let $X = \mathrm{sn}^4(x)$; we have

$$4(1 - X^2)^2 = (1 + X)^4 - 16X(1 - X)^2$$

and so

$$3X^4 + 12X^3 - 46X^2 + 12X + 3 = 0$$

By a miracle, this factors:

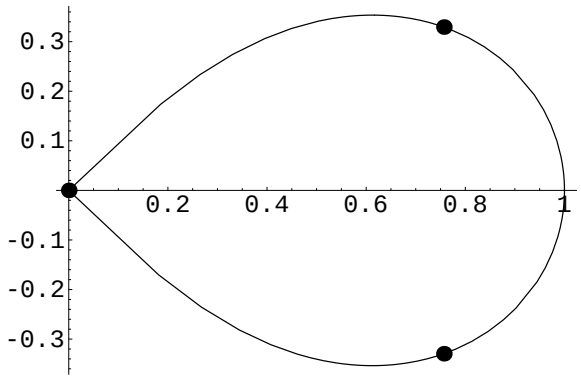
$$(X^2 + 6X - 3)(3X^2 - 6X - 1) = 0$$

Thus $X = \text{sn}^4(x)$ is one of four roots: $2\sqrt{3} - 3, -2\sqrt{3} - 3, \frac{2\sqrt{3}}{3} + 1, -\frac{2\sqrt{3}}{3} + 1$. But only one of these numbers is between 0 and 1, so $\text{sn}^4(x) = 2\sqrt{3} - 3$ and

$$r = \text{sn}(x) = \sqrt[4]{2\sqrt{3} - 3}$$

Therefore

$$(x, y) = \left(\sqrt{\frac{r^2 + r^4}{2}}, \sqrt{\frac{r^2 - r^4}{2}} \right) = \left(\sqrt{\frac{\sqrt{2\sqrt{3} - 3} + 2\sqrt{3} - 3}{2}}, \sqrt{\frac{\sqrt{2\sqrt{3} - 3} - 2\sqrt{3} + 3}{2}} \right)$$



How about dividing a lobe of the lemniscate into five equal parts? Gauss kept a mathematical diary from 1796 to 1814, but it was not found until 1898. It has 146 short entries listing new discoveries. The entry of March 21, 1797 states "divide the lemniscate into five equal parts."

It is clear that the algebra given by the addition formula becomes more and more complicated. But Abel worked his way through this algebra. His 1829 paper contains the following astonishing result:

Theorem 150 (Abel)**.** *The lemniscate can be divided into n equal parts using a straightedge and compass exactly when a regular polygon with n sides can be inscribed in a circle with straightedge and compass.*

For example, the lemniscate can be divided into 17 equal pieces with straightedge and compass, but not into 7 or 9 equal pieces. For a modern proof of this result with many historical references used in this talk, see *Abel's Theorem on the Lemniscate* by Michael Rosen, in the June-July issue of the Mathematical Monthly for 1981.

22.13 *Introductio in Analysin Infinitorum*

In 1748, Euler published a large textbook on analysis, *Introductio in Analysin Infinitorum*. This book was overwhelmingly influential; Euler developed trigonometry using ratios of sides of a right triangle for the first time, based analysis on the notion of a function for the first time, proved that $e^{i\pi} = -1$ and $e^{ix} = \cos x + i \sin x$, showed that that $\sum \frac{1}{n^2} = \frac{\pi^2}{6}$ and $\sum \frac{1}{n^4} = \frac{\pi^4}{90}$, developed the theory of the Riemann zeta function and its connection with number theory, and on and on. We use the letters e and π because Euler used them in this book.

The web page http://math.dartmouth.edu/∼euler/pages/E101.html contains a large pdf zerox copy of the original edition of both volumes of this work, in Latin. Springer published an English translation in 1988. About ten years ago, the author of an article in the Mathematical Monthly wrote an article titled "The Most Influential Textbook of Modern Times" in which he claims that Euler's book is as important for modern mathematics as Euclid was in the ancient world. While writing this report, I typed "most influential textbook of modern" into Google; notice that I didn't even add the word "mathematics." The first entry in Google's return list was Euler's *Introductio*. The second entry was the Communist Manifesto.

Astonishingly, the University of Oregon owns an original edition of Euler's book, housed in the rare book room.

On of the major points Euler makes in this book is that real valued functions should always be extended to the complex plane and studied as functions of a complex variable. A famous result of complex analysis asserts that if a function $f(x)$ can be extended to an open set in the complex plane such that the extension has a complex derivative, then the extension is unique. So it is never a matter of finding a particularly clever extension; instead it is a matter of discovering if God intended an extension.

Therefore it was a foregone conclusion that the new inverses of elliptic integrals would be extended to complex numbers. This step was first taken by Abel and by Jacobi, and both made an astonishing discovery. The functions $sn(x)$ and $cn(x)$ are periodic, so it was expected that their complex extensions would be periodic. But it turned out that these extensions were *doubly periodic*, that is, periodic in two independent directions. We'll use Fagnano's integral $\int \frac{1}{\sqrt{1-z^4}}\, dz$ to show how this conclusion first arose shortly. But first, a word about the work of these two geniuses.

22.14 Abel and Jacobi

Recall that in 1825 through 1832, Legendre published the three volumes of a work titled *Traité des fonctions elliptiques et des intégrales eulériennes* in which he showed that elliptic integrals could always be transformed to a combination of three specific forms, known as elliptic integrals of the first kind, of the second kind, and of the third kind. These integrals depend on an additional parameter called the modulus. To give a flavor of these results, the elliptic integral of the first kind has the form

$$\int_0^r \frac{dx}{\sqrt{(1-x^2)(1-mx^2)}}$$

where m is the modulus. Notice that $m = 0$ gives $\arcsin(r)$ and thus trigonometric functions; $m = -1$ gives the integral $\int \frac{1}{\sqrt{1-x^4}}\, dx$ associated with the lemniscate.

Care must be taken when consulting tables of these integrals because the notation is sometimes changed in subtle ways. For example, elliptic integrals of the first kind are built into *Mathematica* using the notation

$$\text{EllipticF}[\phi, m] = \int_0^{\sin(\phi)} \frac{dx}{\sqrt{(1-x^2)(1-mx^2)}}$$

Shortly after Legendre's book, Abel and Jacobi had this idea of inverting these integrals. Abel's work was published in volumes 2 and 3 of Crelle's journal in 1828 and 1829, in a 197 page article titled *Recherches sur les fonctions elliptiques*. Abel died in 1829; he was 26 years old. Jacobi published his results in 1829 in a book titled *Fundamenta nova theoriae functionum ellipticarum*. This book is still read today; for example, when the largest sporadic simple group, the Monster group, was discovered and its irreducible representations computed, John Conway ran to the library and found the dimensions of these representations in a power series expansion in Jacobi's book. Legendre, incidentally, was very supportive of these developments by Abel and Jacobi.

Jacobi's notation for the inverse of the elliptic integral of the first kind (as defined in the top displayed formula above) is $\text{sn}(x, m)$ as in the talk. This function is build into *Mathematica* as

$$\text{JacobiSN}[x, m]$$

However, in this report I took liberties with $\text{cn}(x) = \sqrt{1 - \text{sn}^4(x)}$ to make the analogy with trigonometric functions clear. Jacobi divides this into two functions, writing

$$\text{cn}(x) = \sqrt{1 - \text{sn}^2(x)}$$

$$\text{dn}(x) = \sqrt{1 - m\,\text{sn}^2(x)}$$

These functions are built into *Mathematica* as JacobiCN$[x, m]$ and JacobiDN$[x, m]$. With these changes, the duplication formula becomes

$$\text{sn}(2x) = \frac{2\text{sn}(x)\text{cn}(x)\text{dn}(x)}{1 - m\ \text{sn}^4(x)}$$

and the addition formula becomes

$$\text{sn}(x + y) = \frac{\text{sn}(x)\text{cn}(y)\text{dn}(y) + \text{sn}(y)\text{cn}(x)\text{dn}(x)}{1 - m\ \text{sn}^2(x)\text{sn}^2(y)}$$

In particular, when $m = -1$ for the lemniscate,

$$\text{cn}(x) = \sqrt{1 - \text{sn}^2(x)}$$

$$\text{dn}(x) = \sqrt{1 + \text{sn}^2(x)}$$

and the duplication and addition formulas become

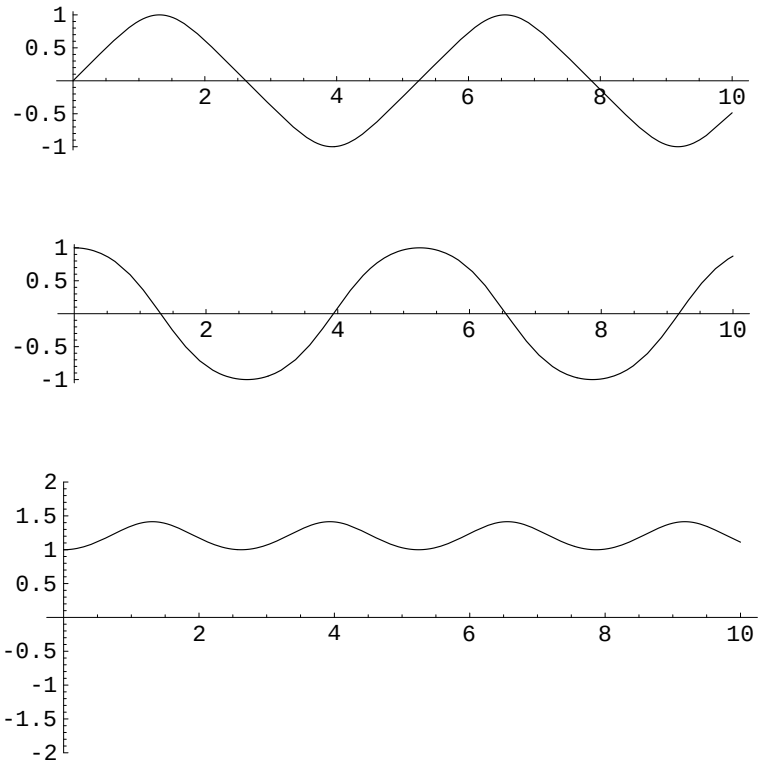
$$\text{sn}(2x) = \frac{2\text{sn}(x)\text{cn}(x)\text{dn}(x)}{1 + \ \text{sn}^4(x)}$$

$$\text{sn}(x + y) = \frac{\text{sn}(x)\text{cn}(y)\text{dn}(y) + \text{sn}(y)\text{cn}(x)\text{dn}(x)}{1 + \ \text{sn}^2(x)\text{sn}^2(y)}$$

Jacobi's change simplifies the parameterization of the lemniscate given in this report as $(x, y) = \left(\sqrt{\frac{r^2+r^4}{2}}, \sqrt{\frac{r^2-r^4}{2}}\right)$. Recall that $r = \text{sn}(x)$ where x is the distance along the lemniscate from the origin to the point. It is easy to deduce that

$$(x, y) = \left(\frac{\text{sn}(x)\text{dn}(x)}{\sqrt{2}}, \frac{\text{sn}(x)\text{cn}(x)}{\sqrt{2}}\right)$$

Below are graphs of $\text{sn}(x)$, $\text{cn}(x)$, and $\text{dn}(x)$ for $m = -1$, that is, for the lemniscate. Notice that the cn graph is much broader at the top than the sn graph, so a periodic displacement of sn no longer gives cn. When $m = 0$, $\text{sn}(x) = \sin(x)$ and $\text{cn}(x) = \cos(x)$ and $\text{dn}(x) = 1$.



22.15 Doubly Periodic Functions

Following the hints from Euler's work, we now study our integral over the complex numbers. Define a map as below, where we integrate over some path from the origin to z.

$$f(z) = \int_0^z \frac{1}{\sqrt{1-t^4}}\, dt$$

We have to worry about points where the integrand blows up; these are the points satisfying $1 - t^4 = 0$; so $\pm 1, \pm i$. Since the square root has multiple complex values, we would like to restrict z to a simply-connected region on which we can choose a unique value for the square root. The easy way to do this is to remove four rays from the plane, as shown in the picture below. The function f is well-defined on the plane minus these four rays.

Next we would like to find the image of our map. We start by finding the image of the first quadrant, and we imagine that the boundary of this quadrant has four pieces, illustrated below. Notice that the forbidden rays are included; this is legal if we push the vertical cut slightly left and the horizontal cut slightly down, so the rays are in our region. We will show that these four paths map to the sides of a rectangle, as on the right below, and conclude that f maps the first quadrant to this rectangle.

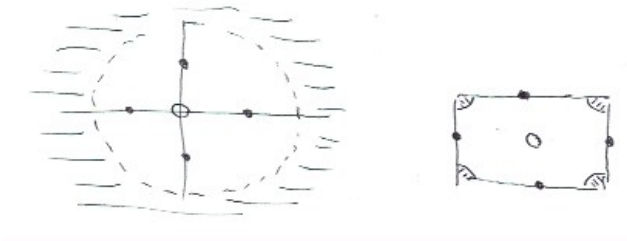
To see this, look first at the horizontal x-axis, which is broken into two paths. On the first path, f is defined by $\int_0^x \frac{1}{\sqrt{1-t^4}}\,dt$ and has the graph shown at the beginning of section 21.9. Thus the image moves along the x-axis until it comes to half the length of a lobe of the lemniscate. On the second path $1 - t^4$ is negative, so $f(x) = i\int_1^x \frac{1}{\sqrt{t^4-1}}\,dt$. This function gradually increases, but it has a finite limit because $\frac{1}{\sqrt{t^4-1}} \sim \frac{1}{t^2}$ and the integral of this to infinity converges. So the image is a finite vertical line.

Next consider the vertical side of the quadrant, which is also broken into two paths. We can parameterize the first path by $\gamma(t) = it$ for $0 \le t \le 1$. Then $\frac{1}{\sqrt{1-\gamma(t)^4}} = \frac{1}{\sqrt{1-t^4}}$, but $d\gamma = i dt$ and so the integral is imaginary and gives a vertical line in the image. The same sort of analysis produces the top segment.

We analyze the remaining quadrants in the same way, but there is a trick to make this easy. Suppose we multiply the first quadrant and its bounding paths by $e^{i(k\pi/4)}$. This leaves the integrand invariant, but the final integral is multiplied by the same amount because $d\gamma$ is multiplied by that amount. So the pictures for the remaining quadrants give the previous picture appropriately rotated. See below.

In the end, then, the entire plane is mapped by our integral to the rectangle shown below. The origin is mapped to the center of the rectangle and the four bad points are mapped to the centers of the four sides. Something strange happens at the four bad points because the angle between the two horizontal paths we used to get the first rectangle is π and yet it becomes $\pi/2$ in the rectangle.

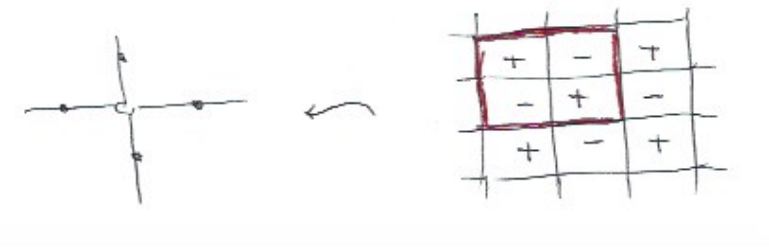
The shading shows that the region close to infinity in the plane is mapped to four quarter-disk regions around the four corners of the rectangle.



We are going to argue that the inverse mapping from the rectangle back to the plane is much nicer than the map given by the integral. The inverse mapping has poles (or rather four pieces of one pole) at the four corners. It has one zero in the center. And at the bad points, the map doubles angles (just like z^2 doubles angles at the origin); an angle of π in the rectangle became an angle of 2π in the plane. This is what we would expect if the derivative of the inverse mapping vanished at the images of the bad points.

But there is more to come. If we rotate the plane on the left about the origin by π, the integral just changes sign. So the values of the inverse mapping on the left and right sides of the

rectangle, and the top and bottom of the rectangle, differ by a sign. It follows that we can glue multiple copies of the rectangle together to get a well-defined inverse mapping on the entire plane, as illustrate below. In this picture, the inverse mapping is given by our original map on rectangles with a plus sign, and by the negative of our map on rectangles with a minus sign.

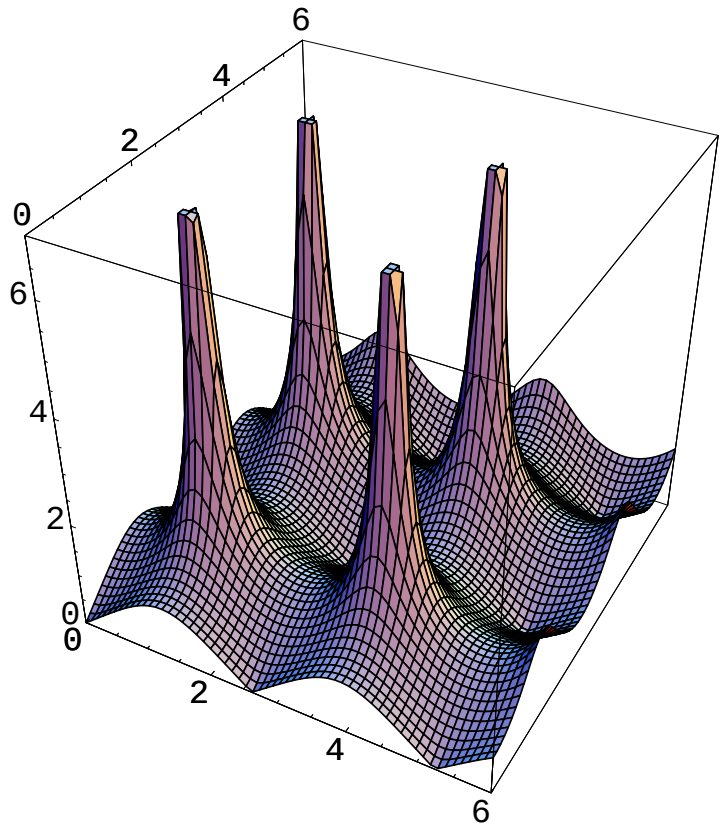


A portion of the picture on the right is shaded in red. Our inverse map is doubly periodic on the entire plane, and this red portion is a fundamental region; its translations left and right and up and down cover the entire plane. The inverse map is completely determined by its values on this fundamental region. Moreover, the values on the left and right sides agree, and the values on the top and bottom sides agree, so we could glue the left and right together and glue the top and bottom together, and argue that the inverse map is really defined on a torus.

As we will see in the next chapter, doubly periodic maps on the plane are either constant or else have poles. Our inverse map is meromorphic with a pole in the center of the red rectangle and a pole shared among the corners. We get a second fundamental region by just pushing the red boundary left and down slightly, and then there are two regular poles inside.

Another result from the next chapter states that a non-constant doubly periodic function takes each complex value the same number of times counting multiplicities. Our map clearly takes most values exactly four times. So it takes every value four times. Thus the two poles must each have order two. Notice that our map has four zeros, one in each of the four pieces. But something interesting happens at the images of the four bad points. There are four of these images inside the red boundary and eight images along the boundary. But opposite boundaries are glued together, so there are really only four images on the boundary. Thus the four bad points correspond to eight unusual points in the red fundamental region. We would expect 16 points in this region if each value in the plane on the left is taken four times. The reason is that the images of the bad points become zeros of the derivative of the inverse mapping, so the inverse mapping at these image points is not conformal, but instead wraps around twice.

Finally, here is a picture of the inverse mapping generated by *Mathematica*. In truth, the picture shows the absolute value of the function. Notice the double periodicity of this function. Notice the zero in the center between the four poles. And notice the flat "passes" half way between two poles where the derivative turns out to vanish.



Remark: Abel and Jacobi showed that elliptic integrals evaluated over the complex numbers have inverse maps which are doubly periodic. It soon became clear that it was much easier to develop the theory in reverse. Rather than starting with an elliptic integral, obtaining an inverse, and showing that it is doubly periodic, it is simpler to develop a general theory of doubly periodic meromorphic functions; then elliptic integral pop out in the end.

The theory of these doubly periodic functions by Abel and Jacobi was later simplified by Weierstrass, and we will follow his approach in the next chapter.

This long history began with Fagnano's formula for doubling the length of the lemniscate, and Euler's generalization to an addition formula for elliptic integrals. We seem to have dropped that topic. But just wait. It will come up naturally in the next chapter.

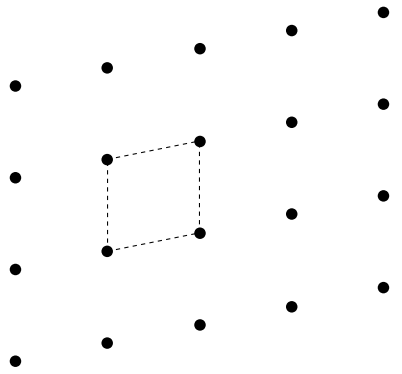
Chapter 23

Doubly Periodic Functions

23.1 Definitions

Definition 50. *A meromorphic function $f(z)$ on C is* doubly periodic *if $f(z+\omega_1) = f(z)$ and $f(z+\omega_2) = f(z)$ for two complex constants ω_1 and ω_2 linearly independent over the reals.*

In that case the constants generate a lattice $Z\omega_1 + Z\omega_2$ and f is completely determined by its values in the parallelogram with sides ω_1 and ω_2. Since f takes the same values on opposite sides of this parallelogram, we can glue these sides together to form a torus. Thus equivalently, a doubly-periodic function is a function on this torus.



Definition 51. *Let $f(z)$ be a meromorphic function on the complex plane. A* period *of f is a number ω such that $f(z+\omega) = f(z)$ for all z. The set of all periods of f is a subgroup of C.*

Theorem 151. *If $f(z)$ is meromorphic and non-constant, the set of all periods is a lattice. So if there are any nonzero periods, then either the periods consist of all integer multiples of a non-zero ω, or else they consist of all integer combinations $m\omega_1 + n\omega_2$ where ω_1 and ω_2 are linearly independent over the reals.*

Proof: The group of periods cannot contain an accumulation point, else there are periods ω_i such that $\omega_i \to \omega$. Thus $f(0) = f(\omega_i) \to f(\omega)$. But then by the identity theorem, f is a constant.

It follows that there are only finitely many periods in any closed disk of radius R. If the set of periods is not just $\{0\}$, we can find a non-zero period ω_1 of smallest positive absolute value. If all periods lie on a line, then any other period p would be between two adjacent multiples of ω_1, say between $k\omega_1$ and $(k+1)\omega_1$, and then $p - k\omega_1$ would be a nonzero period with smaller absolute value.

If the periods are not all on a line, find ω_2 a period of smallest absolute value not on the line. If there are other periods not in the lattice spanned by ω_1 and ω_2, then by subtracting appropriate multiples of ω_1 and ω_2 we can find a period of the form $r\omega_1 + s\omega_2$ with r and s real and between $-\frac{1}{2}$ and $\frac{1}{2}$. So $|r\omega_1 + s\omega_2| \leq \frac{1}{2}|\omega_1| + \frac{1}{2}|\omega_2| \leq |\omega_2|$. If this is a strict inequality, it contradicts the choice of ω_2. So it must be an equality, and $|\omega_1| = |\omega_2|$ and $|r| = |s| = \frac{1}{2}$. But then $\pm\omega_1$ and $\pm\omega_2$ lie on a circle and $r\omega_1 + s\omega_2$ is in the middle of a line joining one of these points to another. However, no point in the middle of such a line can be on the circle.

Remark: If $\omega \neq 0$ is a period of $f(z)$, we can find λ such that $f(\lambda z)$ has period $2\pi i$. Note that e^z has this period.

Theorem 152. *Every meromorphic function with period $2\pi i$ can be written $g(e^z)$ where $g(z)$ is an arbitrary meromorphic function on $C - \{0\}$.*

Proof: The function f is completely determined by it's values in the strip $0 \leq \Im(z) < 2\pi i$. The function e^z maps this strip to $C - \{0\}$ and the map is one-to-one and onto. The result follows immediately,

Remark: Note that $\cos(-iz)$ is periodic with period $2\pi i$. So $\cos(-iz) = g(e^z)$. Indeed

$$\cos(z) = \frac{e^{iz} + e^{-iz}}{2}$$

and

$$g(w) = \frac{w + \frac{1}{w}}{2}$$

23.2 Fundamental Theorems

We prove some easy but beautiful theorems fundamental to the subject.

Theorem 153. *A holomorphic doubly-periodic function is constant.*

Proof: The values of the function are completely determined by values on the closed fundamental parallelogram generated by ω_1 and ω_2. This parallelogram is compact, so the function is bounded. A bounded entire function is constant.

Theorem 154. *A non-constant doubly-periodic function cannot have only one pole, of order one, in the period parallelogram. Indeed the sum of the residues of the poles in this region must be zero.*

Proof: By translating the parallelogram slightly, we can assume that all poles are strictly inside. Let γ be the path obtained by tracing the parallelogram counterclockwise around its boundary. We form

$$\frac{1}{2\pi i}\int_\gamma f(z)\,dz$$

This integral then gives the sum of the residues of the poles. But integration over opposite sides of the parallelogram cancels out because f is periodic, so the integral is zero. QED.

Theorem 155. *A non-constant doubly-periodic function takes each complex value, including ∞, the same number of times in the period parallelogram if we count multiplicities.*

Proof: By the argument principle,

$$\frac{1}{2\pi i}\int \frac{f'(\zeta)}{f(\zeta)}\,d\zeta = \sum_{\text{zeros and poles}} \text{order}(z_i)$$

Our $f(z)$ has only finitely many zeros and poles in the fundamental parallelogram, so we can translate the boundary of this parallelogram by a small amount so no zeros and poles are on the boundary. The integral around this boundary is zero by periodicity, so f has the same number of zeros as poles. Applying the same argument to $g(z) = f(z) - z_0$ shows that f equals z_0 the same number of times. QED.

23.3 The Weierstrass $\wp$-Function

We now set out to construct a doubly-periodic function for the lattice generated by ω_1 and ω_2. By a theorem in the previous section, a non-trivial doubly-periodic function must have at least one pole, and if it only has one pole, the pole must have order at least two. We'll produce such a function. It is easiest to place the poles at the lattice points. This suggests defining

$$f(z) = \sum_{m,n} \frac{1}{(z - m\omega_1 - n\omega_2)^2}$$

Unhappily, this sum does not converge. But the Mittag-Leffler theorem tells us how to fix that. We should expand each term in a Taylor series about the origin and subtract the first few terms of this series to make the remaining portion small near the origin. To save ink, let us write ω for the general term of the form $m\omega_1 + n\omega_2$.

Definition 52. *The Weierstrass $\wp$-function is*

$$\wp(z) = \frac{1}{z^2} + \sum_{\omega \neq 0} \left(\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2} \right)$$

Remark: Suppose we have a compact subset of the complex plane. This subset is contained in a disk of radius R, and on or inside this disk there are only finitely many lattice points. The above formula is interpreted to mean that $\wp(z)$ is the finite sum over these terms, plus the infinite sum of other terms. The first sum is certainly meromorphic; we will prove that the remaining sum converges uniformly on the disk.

Theorem 156. *The infinite sum for $\wp$ converges uniformly on compact subsets of the plane. Consequently, the derivative of the function can be computed term by term, giving*

$$\wp'(z) = \sum_{\omega} \frac{-2}{(z-\omega)^3}$$

The function $\wp(z)$ is even; $\wp(-z) = \wp(z)$. However $\wp'(z)$ is odd; $\wp'(-z) = -\wp'(z)$. Finally, both $\wp(z)$ and $\wp'(z)$ are doubly periodic.

Proof of Convergence: We must estimate

$$\left| \frac{1}{(z-\omega)^2} - \frac{1}{\omega^2} \right| = \left| \frac{-z^2 + 2z\omega}{(z-\omega)^2 \omega^2} \right|$$

We only need estimate for z in a compact set, so we can suppose that $|z| \leq R$. We only need consider large ω, so we can suppose $2R < |\omega|$. Then $|-z^2 + 2z\omega| \leq |z||\omega| + 2|z||\omega| \leq 3R|\omega|$. Also $|z-\omega| \geq |\omega| - |z| \geq |\omega| - |\omega|/2 = |\omega|/2$. Putting this altogether,

$$\left| \frac{1}{(z-\omega)^2} - \frac{1}{\omega^2} \right| \leq \frac{12R}{|\omega|^3}$$

From here we first give the rough idea. There are roughly $2\pi|\omega|$ lattice points near the circle with radius $|\omega|$. So the total contribution from this annulus is

$$\frac{12R}{|\omega|^2} \cdot 2\pi|\omega| = \frac{24R\pi}{|\omega|^2}$$

Therefore, the sum over lattice points in all annuli converges.

In the rigorous proof, we start with the origin and then form the series of larger and larger parallelograms about it. The number of lattice points in the boundaries of these parallelograms is $8, 16, 24, \ldots$, so $8N$ in the Nth ring. Let $\widehat{\omega}$ be the closest lattice point to the origin in the first ring. Then $N\widehat{\omega}$ is the closest lattice point in the Nth ring. For large N, the total contribution to the sum from lattice points in the Nth ring is at most

$$\frac{24R\pi}{|N\widehat{\omega}|^3} \cdot 8N$$

which is a constant times $\frac{1}{N^2}$, and $\sum \frac{1}{N^2}$ converges.

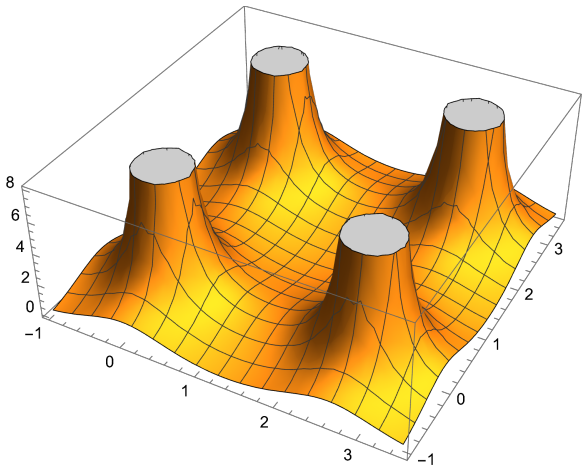
Proof of Remaining Claims: Once we know that our series converges uniformly on compact subsets, we can differentiate term by term, giving the formula for $\wp'(z)$. Notice that this function is automatically doubly periodic because all the roots are treated equally. That isn't true in our formula for $\wp(z)$; the origin has no correction factor. So we must prove that $\wp(z)$ is also periodic.

Notice that $\wp(z)$ is even because whenever ω is a lattice point, so is $-\omega$. Similarly $\wp'(z)$ is odd. By periodicity of $\wp'(z)$, $\wp(z+\omega_1)$ and $\wp(z)$ have the same derivative. So there is a constant C such that $\wp(z+\omega_1) - \wp(z) = C$. Set $z = -\frac{\omega_1}{2}$. Then

$$\wp\left(\frac{\omega_1}{2}\right) = \wp\left(-\frac{\omega_1}{2}\right) + C = \wp\left(\frac{\omega_1}{2}\right) + C$$

So $C = 0$ and $\wp(z+\omega_1) = \wp(z)$. Similarly $\wp(z+\omega_2) = \wp(z)$.

Remark: Here is a picture of $|\wp(z)|$ drawn by *Mathematica*.



23.4 Formula for All Doubly Periodic Functions

Recall that a rational function is just a quotient of two polynomials, like $\frac{x^2+5x+2}{x^2-1}$.

Theorem 157. *Every doubly periodic function has the form*

$$R_1(\wp(z)) + R_2(\wp(z))\,\wp'(z)$$

for rational R_1 and R_2.

Example: $\frac{3\wp^3(z)+2\wp(z)+1}{\wp^4(z)-1} + \frac{3}{\wp(z)}\,\wp'(z)$

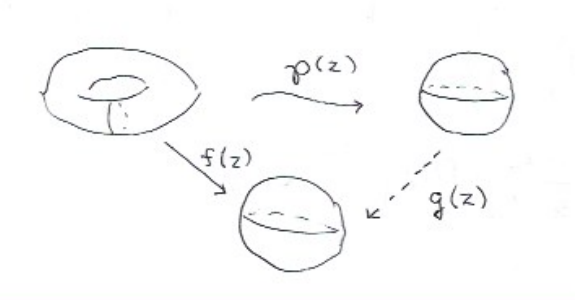
Proof: Let $f(z)$ be doubly periodic. Write

$$f(z) = \frac{f(z)+f(-z)}{2} + \frac{f(z)-f(-z)}{2\wp'(z)}\,\wp'(z)$$

Both fractions are even, so it suffices to prove that every even doubly periodic function is a rational function of $\wp(z)$.

Let $f(z)$ be an even doubly periodic function. We can think of f as a mapping the torus to the Riemann sphere. This sphere is $C \cup \{\infty\}$, so f takes regular points to their values in C and poles to infinity.

Consider the diagram below. It is possible to define a meromorphic $g(z)$ completing the diagram, by an argument given shortly. Since every meromorphic function on the Riemann sphere is rational, $f(z) = R(\wp(z))$ and we are done.



Now we prove the existence of g. The function $\wp(z)$ takes each complex value the same number of times. It has one pole of order 2, so it takes each complex value twice. But $\wp(z) = \wp(-z)$, so the two points on the torus giving a particular value differ by a sign. Since f is an even function, it takes the same value at z and $-z$. So we can define g by using $\wp^{-1}$ to go backward to two points, and then selecting the value of f at these points.

We must pay attention to multiplicities. The value ∞ is only taken at one point, but then there is no ambiguity about the value of $g(\infty)$. Other complex values may come from only one point if taken with multiplicity greater than one, but again there is no ambiguity defining g at these points.

We also have problems if $z = -z$ up to a lattice point, because then we only have one inverse point rather than two. But $\wp'(z) = -\wp'(-z)$ and so $\wp'(z) = 0$ and such points. Therefore they take their values with multiplicity greater than one, so these values are only taken once.

If $\wp(z_0) \neq 0$, then f is a local diffeomorphism near z_0 with holomorphic inverse. So g is holomorphic except at finitely many points where it is merely continuous. But then each of these points is a removable singularity and g is holomorphic everywhere. This concludes the proof.

23.5 The Differential Equation Satisfied by $\wp(z)$

The meromorphic functions on the torus form a field, so they can be added and multiplied. We know that these functions have the form

$$R_1(\wp(z)) + R_2(\wp(z))\,\wp'(z)$$

It is clear how to add two such expressions. To multiply, we need to know $(\wp'(z))^2$.

Theorem 158.

$$\left(\frac{d\wp}{dz}\right)^2 = 4\wp^3 - g_2\wp - g_3$$

where g_2 and g_3 are the following constants:

$$g_2 = 60\sum_{\omega\neq 0}\frac{1}{\omega^4} \qquad g_3 = 140\sum_{\omega\neq 0}\frac{1}{\omega^6}$$

Proof: Both sides of the equation are doubly periodic and have poles only at the lattice points. We will show that that the Laurent expansion of the difference at the origin has only positive powers of z. It follows that the difference is a doubly periodic holomorphic function, so constant, and so zero.

We have the following, where $g(z)$ and the remaining sums are over non-zero lattice points:

$$\wp(z) = \frac{1}{z^2} + \sum\left(\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2}\right) = \frac{1}{z^2} + g(z)$$

$$g'(z) = \sum\frac{-2}{(z-\omega)^3}$$

$$g''(z) = \sum \frac{6}{(z-\omega)^4}$$

$$g'''(z) = \sum \frac{-24}{(z-\omega)^4}$$

$$g''''(z) = \sum \frac{120}{(z-\omega)^6}$$

So $g(0) = 0, g'(0) = 0, g''(0) = \frac{1}{10}g_2, g'''(0) = 0, g''''(0) = \frac{6}{7}g_3$ and

$$\wp(z) = \frac{1}{z^2} + \frac{1}{20}g_2 z^2 + \frac{1}{28}g_3 z^4 + ...$$

$$\wp'(z) = \frac{-2}{z^3} + \frac{1}{10}g_2 z + \frac{1}{7}g_3 z^3 + ...$$

Ignoring positive powers of z, we get

$$\left(\wp'(z)\right)^2 = \frac{4}{z^6} - \frac{2}{5}\frac{g_2}{z^2} - \frac{4}{7}g_3 + ...$$

$$\left(\wp(z)\right)^2 = \frac{1}{z^4} + \frac{1}{20}g_2 + \frac{1}{28}g_3 z^2 + ...$$

$$\left(\wp(z)\right)^3 = \frac{1}{z^6} + \frac{3}{20}\frac{g_2}{z^2} + \frac{3}{28}g_3 + ...$$

$$\left(\wp'(z)\right)^2 - 4\left(\wp(z)\right)^3 + g_2\wp + g_3 = 0 + \text{terms in positive powers of } z$$

QED.

23.6 Elliptic Integrals

Once we know the differential equation satisfied by the $\wp$-function, we can begin making connections of our theory to other fields of mathematics. This section and the next sketch briefly two of these connections. After that, we will return to systematic development of the theory.

Suppose we want to compute the elliptic integral

$$\int \frac{dz}{\sqrt{4z^3 - g_2 z - g_3}}$$

Make the substitution $z = \wp(w)$. Then $dz = \wp'(w)\,dw$ and the integral becomes

$$\int \frac{\wp'(w)\,dw}{\sqrt{4\wp(z)^3 - g_2\wp(z) - g_3}} = \int \frac{\wp'(w)\,dw}{\wp'(w)} = \int dw = w$$

and the integral is given by the inverse of the $\wp$-function. This is the reverse of how the theory was discovered historically. TThe integral came first, and later its inverse was discovered to be doubly-periodic.

There is still work to do. Fairly easy algebra shows that any elliptic integral can be transformed into one involving the Weierstrass cubic, so this substitution can be made. But we need to show that a lattice can be found giving g_2 and g_3. We will prove later on that such a lattice exists, but won't explain how to actually calculate its generators ω_1 and ω_2.

23.7 Field Theory

We can view the field of meromorphic functions on a torus purely algebraically, completely forgetting that the elements of this field are really doubly-periodic functions, or even functions of any sort. Suppose we let X stand for $\wp(z)$ and Y stand for $\wp'(z)$. Then among the elements of the field are all rational functions in X, that is, expressions like $\frac{X^3+7X+4}{X^2-1}$. Algebraists call this the field of rational functions and denote it $C(X)$. It as one of the elementary fields from which more complicated fields are constructed.

A typical way to construct more complicated fields is to start with a ground field, introduce a polynomial $P(Y)$ with coefficients in the ground field, and add a root of this polynomial to the field. Fields constructed in this way with ground field Q are called *algebraic number fields* and are extremely important in number theory. For example, the field $Q(i)$ consisting of all $q_1 + q_2 i$ is formed from Q by adding a root of $Y^2 + 1 = 0$.

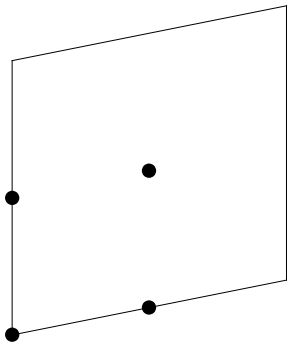
The field of meromorphic functions on a torus is formed in exactly this way. The ground field is all rational expressions in X and we then add a root of $Y^2 - 4X^3 + g_2X + g_3$. Large portions of our theory can be developed purely algebraically using this idea.

Here's an example. There are many tori, depending on the choice of ω_1 and ω_2. Some of these complex tori are conformally equivalent and some are not. We will study this in detail shortly, but we have already seen a hint of it, because the theory of elliptic integrals depends on an extra parameter m. It turns out that two tori are conformally equivalent if and only if their associated fields are isomorphic by an isomorphism which is the identity map on the complex numbers. So conformal equivalence can be determined by g_2 and g_3, and we will show how later on.

23.8 The e_i

We now return to the rigorous theory. We know that $\wp(z)$ takes each value exactly twice, counting multiplicities. We are going to show that there are exactly three complex values taken only once on the torus, because the multiplicity at these points is 2. The value ∞ is also taken only once because our pole has order two.

Recall that $\wp(z) = -\wp(-z)$. This gives two different points unless z and $-z$ are equivalent modulo the lattice. In that case, $z = -z + m\omega_1 + n\omega_2$ and so $z = \frac{m}{2}\omega_1 + \frac{n}{2}\omega_2$. This gives exactly four points in the canonical period parallelogram, as shown below. Notice that opposite boundary points are identified.



There is no need to name the origin. We name the remaining points

$$e_1 = \frac{\omega_1}{2} \quad e_2 = \frac{\omega_2}{2} \quad e_3 = \frac{\omega_1}{2} + \frac{\omega_2}{2}$$

Notice that $\wp'(-e_i) = -\wp'(e_i)$, but also $\wp'(-e_i) = \wp'(e_i)$ by periodicity. So $\wp'(e_i) = 0$. Because $\wp'$ has exactly one pole of order 3, it takes each complex value three times counting multiplicities. So $\wp'$ vanishes exactly at e_1, e_2, e_3. The vanishing of $\wp'$ indicates points of $\wp$ with values of multiplicity greater than one. So the values $\wp(e_i)$ are taken exactly once.

It is easy to calculate values taken exactly once because $\wp'(e_i) = 0$ and so the differential equation gives

$$4\wp(e_i)^3 - g_2\wp(e_i) - g_3 = 0$$

Therefore $\wp(e_i)$ are the three roots of the cubic $4X^3 - g_2X - g_3$.

Since the three values $\wp(e_i)$ are distinct, so are the roots $\lambda_1, \lambda_2, \lambda_3$ of $4X^3 - g_2X - g_3$, and therefore $(\lambda_1 - \lambda_2)^2(\lambda_1 - \lambda_3)^2(\lambda_2 - \lambda_3)^2 \neq 0$. This expression is unchanged under all permutations of the roots, and a famous theorem in algebra then guarantees that it can be written in terms of the coefficients of the polynomial. In the case of the above cubic, this value is $g_2^3 - 27g_3^2$ up to a nonzero factor. We have proved

Theorem 159. *Let* $e_1 = \frac{\omega_1}{2}, e_2 = \frac{\omega_2}{2}, e_3 = \frac{\omega_1}{2} + \frac{\omega_2}{2}$. *Then*

- *The* e_i *are the three zeros of* $\wp'(z)$.
- *The function* $\wp(z)$ *takes the values* $\wp(e_i)$ *exactly once; all other complex values are taken twice.*
- *The exceptional values* $\wp(e_i)$ *are the roots of* $4X^3 - g_2X - g_3$.
- *We have* $g_2^3 - 27g_3^2 \neq 0$.

23.9 Determining an Elliptic Function by its Zeros and Poles

A doubly periodic function is essentially determined by its zeros and poles and their orders. Indeed suppose $f(z)$ and $g(z)$ are doubly periodic with the same zeros and poles and the same orders. Then $\frac{f(z)}{g(z)}$ has no singularities and thus is a non-zero constant, so $f(z) = \lambda g(z)$ for a constant λ.

A finite formal sum $D = \sum d_i z_i$, where the $d_i$are non-zero integers and the z_i are points in the torus, is called a *divisor*. In such a sum, the z_i are mere formal symbols, to keep track of certain points in the torus. Think of a divisor as a request to form a function with zeros of order d_i at the points attached to positive coefficients, and poles of order d_i at points attached to negative coefficients. The *degree* of a divisor is the integer $\sum d_i$. We say D is the *divisor of a function* f if f has zeros and poles at the indicated spots of precisely the indicated orders.

Theorem 160. *If* $D = \sum d_i z_i$ *is the divisor of a doubly-periodic function, then*

- *the degree of* D *is zero*
- *the actual sum of complex numbers* $\sum d_i z_i$ *is a lattice point*

Proof: If D is the divisor of f, then the sum of the positive d_i is the number of times f is zero and the sum of the negative d_i is the negative of the number of times f is infinite, and f takes each value the same number of times.

Form

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

where the integral is taken counterclockwise around the boundary of a fundamental parallelogram. Translate this parallelogram slightly if necessary to insure that f has no zeros or poles on the boundary. Then the integral is the sum of the residues at singularities. There are two kinds of singularities: zeros of f and poles of f. If z_0 is a zero, then near z_0 we can write

$f(z) = (z - z_0)^k g(z)$ and

$$\frac{f'(z)}{f(z)} = \frac{k(z - z_0)^{k-1} g(z) + (z - z_0) g'(z)}{(z - z_0)^k g(z)} = \frac{k}{z - z_0} + \frac{g'(z)}{g(z)}$$

Since $z = (z - z_0) + z_0$ and

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

and the residue is kz_0. At a pole the same calculation applies except that k is negative, and the residue is still kz_0. So the sum of the residues is $\sum d_i z_i$.

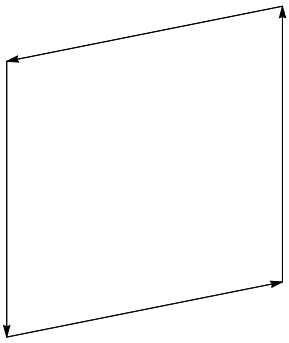
Now we calculate the integral. We'll omit the $2\pi i$ for now. The vertical sides of the fundamental parallogram are parameterized by $\gamma(t) = t\omega_2$ and $t\omega_2 + \omega_1$ for $0 \le t \le 1$. Keeping track of the direction of these paths, the integral is

$$-\int_0^1 \frac{(t\omega_2) f'(t\omega_2)}{f(t\omega_2)} \omega 2 \, dt + \int_0^1 \frac{(t\omega_2 + \omega_1) f'(t\omega_2 + \omega_1)}{f(t\omega_2 + \omega_1)} \omega 2 \, dt$$

Since f is periodic, this difference is

$$\int_0^1 \omega_1 \frac{f'(t\omega_2)}{f(t\omega_2)} \, \omega_2 \, dt = \omega_1 \log f(t\omega_2) \Big|_0^1$$

The values of the function at the two endpoints are the same, so their logarithms differ by a multiple of $2\pi i$. Since our integral has this factor in the denominator, we get an integer multiple of ω_1. Similarly the top and bottom contribution to the integral is an integer multiple of ω_2.



In this proof we used the standard boundary of the parallelogram. If any zeros or poles are on this boundary, just translate the path slightly. QED.

Theorem 161. *Conversely, if $D = \sum d_i z_i$ is a divisor with degree zero, and $\sum d_i z_i$ is a lattice point, there is a doubly-periodic function with divisor D and the chosen zeros and poles. It is unique up to a constant multiple.*

Proof: Suppose we had a doubly periodic function $\sigma(z)$ with simple zeros at the lattice points and no other zeros and poles. Then $\prod \sigma(z-z_i)^{d_i}$ would be the desired function. But of course no doubly periodic function has these properties. We will construct a function which is not doubly periodic, but has these properties. Our function changes in an easy way when we add ω_1 or ω_2 to z, and we will be able to show that the quotient doesn't change at all.

Recall the lemma that $|z| < 1$ implies $|1 - E_n(z)| \leq |z|^{k+1}$. Taking a product over lattice points, this lemma implies that $\sigma(z) = z \prod_\omega E_2\left(\frac{z}{\omega}\right)$ converges uniformly on compact subsets because if z is restrited to a compact subset then for all large ω we have $\left|E_2\left(\frac{z}{\omega}\right)\right| \leq \left|\frac{z}{\omega}\right|^3$ and $\sum \left|\frac{z}{\omega}\right|^3$ converges.

We then get the following formulas; it is traditional to name the second function $\zeta(z)$, but it will only appear in this section.

$$\sigma(z) = z \prod_{\omega \neq 0} \left(1 - \frac{z}{\omega}\right) e^{\left(\frac{z}{\omega}\right) + \frac{1}{2}\left(\frac{z}{\omega}\right)^2}$$

$$\zeta(z) = \frac{\sigma'(z)}{\sigma(z)} = \frac{1}{z} + \sum_{\omega \neq 0} \left(\frac{1}{z-\omega} + \frac{1}{\omega} + \frac{z}{\omega^2}\right)$$

$$\frac{d}{dz}\left(\frac{\sigma'(z)}{\sigma(z)}\right) = -\frac{1}{z^2} - \sum_{\omega \neq 0} \left(\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2}\right) = -\wp(z)$$

Since the derivative of $\zeta(z)$ is periodic, $\zeta(z)$ is periodic up to a constant. It is traditional to name these constants η_1 and η_2. So $\zeta(z+\omega_1) = \zeta(z) + \eta_1$ and $\zeta(z+\omega_2) = \zeta(z) + \eta_2$.

Lemma 21.

$$\eta_1 \omega_2 - \eta_2 \omega_1 = 2\pi i$$

Proof of lemma: Form $\int \zeta(z) dz$ around a fundamental parallelogram. Shift this path down and to the left slightly to avoid the pole of $\zeta(z)$. Since ζ has only one pole with residue one inside this path, the result is $2\pi i$ by the residue theorem. On the vertical portions of the path, $\zeta(z)$ is η_1 larger on the right than on the left, so taking account the direction of integration, the sum of these integrals is $\eta_1 \omega_2$. Similarly the top and bottom integrals give $-\eta_2 \omega_1$.

Continuation of Main Proof: The formula $\zeta(z+\omega_1) = \zeta(z) + \eta_1$ can be rewritten as

$$\frac{\sigma'(z+\omega_1)}{\sigma(z+\omega_1} = \frac{\sigma'(z)}{\sigma(z)} + \eta_1 = \frac{\frac{d}{dz}\left(\sigma(z) e^{\eta_1 z}\right)}{\sigma(z) e^{\eta_1 z}}$$

But if $f(z)$ and $g(z)$ are two functions satisfying $\frac{f'}{f} = \frac{g'}{g}$, then $f'g - fg' = 0$ or $\left(\frac{f}{g}\right)' = 0$, so there is a constant C with $\frac{f}{g} = C$. In short, $f = Cg$. We conclude that

$$\sigma(z + \omega_1) = C\sigma(z)e^{\eta_1 z}$$

To determine C, substitute $z = -\frac{\omega_1}{2}$ and notice that σ is an odd function. Thus

$$\sigma\left(\frac{\omega_1}{2}\right) = C\sigma\left(-\frac{\omega_1}{2}\right)e^{-\frac{1}{2}\eta_1\omega_1}$$

and

$$C = -e^{\frac{1}{2}\eta_1\omega_1}$$

A similar argument works for ω_2 and we conclude that

$$\sigma(z + \omega_1) = -\sigma(z)e^{\eta_1 z + \frac{1}{2}\eta_1\omega_1}$$

$$\sigma(z + \omega_2) = -\sigma(z)e^{\eta_2 z + \frac{1}{2}\eta_2\omega_2}$$

We are finally able to prove our theorem. Recall that we want a periodic function f with divisor $\sum d_i z_i$. Let

$$f(z) = \prod \sigma(z - z_i)^{d_i}$$

We need only prove that this is doubly-periodic. It suffices to study ω_1 and calculate

$$f(z + \omega_1) = \prod \sigma(z - z_i + \omega_1)^{d_i} = \prod \left(-\sigma(z - z_i)e^{\eta_1(z - z_i) + \frac{1}{2}\eta_1\omega_1}\right)^{d_i}$$

Since the sum of the d_i is zero, the initial signs cancel out and our expression equals

$$f(z)e^{\sum d_i(\eta_1(z - z_i) + \frac{1}{2}\eta_1\omega_1)}$$

Because $\sum d_i = 0$, this reduces to

$$f(z)e^{-\eta_1 \sum d_i z_i}$$

We are assuming that $\sum d_i z_i$ is in the period lattice. If it equals zero, we are done.

Otherwise a minor trick fixes the problem. Instead of writing $D = \sum d_i z_i$, simply write the z_i multiple times. Separate the zeros, now called a_i, from the poles, now called b_i. So $D = \sum a_i - \sum b_i$. If this sum equals the lattice point ω, replace one a_i by $a_i - \omega$. Now the sum of the terms in D is zero.

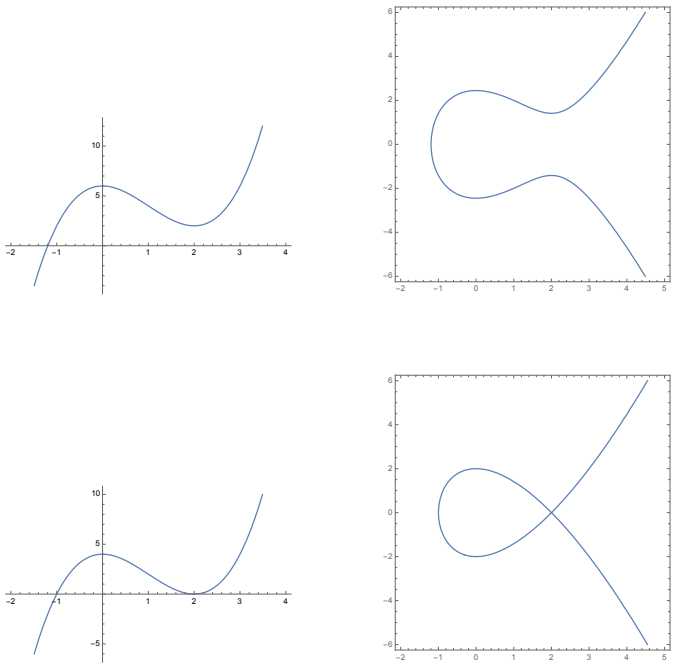
Important Remark: Up until this theorem, the restrictions on zeros and poles have been topological: there must be the same number of zeros and poles, there cannot be a single pole of order 1, and so forth. But according to this theorem, if we are given the points z_i in the torus where f has zeros and poles and we move one of the points very slightly, our multiply periodic function will cease to exist. So the condition in this theorem is *analytical*.

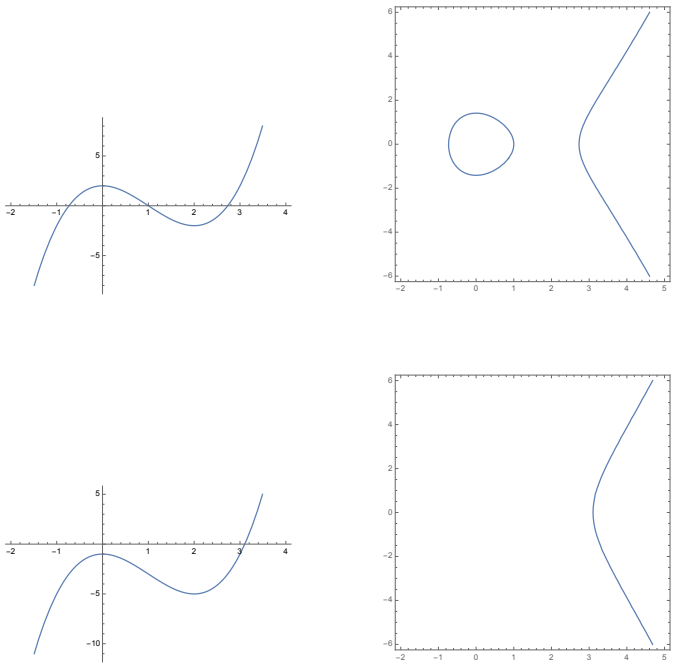
23.10 Algebraic Curves

Rene Descartes is credited with the invention of analytic geometry. He considered polynomials in two variables X and Y and studied curves in the plane formed by their solutions. Thus $X^2 + Y^2 = 1$ gives a circle, $X^2 - Y^2 - 1$ gives a hyperbola, and so forth.

The theory of doubly-periodic functions has led us to the polynomial $Y^2 - 4X^3 + g_2X + g_3$. The constants g_2 and g_3 may be complex, but they certainly can be real and then we get a plane curve from Descartes. It is instructive to graph this curve.

The function $4X^3 - g_2X - g_3$ is a cubic, so it is negative to the left, and positive to the right. When it is positive, there are two possible Y's for each X, differing by a sign. When it is negative, there are no real $Y's$ solving the equation. The picture below shows several cubic functions on the left, and the resulting curve $Y^2 - 4X^3 + g_2X + g_3 = 0$ on the right.





The analytic geometry of Descartes gradually evolved into the modern field of algebraic geometry. One early discovery was that it is better to work over algebraically closed fields, and thus in our case over the complex numbers. Another was that points at infinity should be added to curves by working in projective space. Since we want to do both things, a few more words are appropriate.

When we work over the complex numbers, the enclosing space becomes C^2 rather than R^2, so a 4-dimensional space rather than the plane. Curves like $X^2+Y^2=1$ become two dimensional surfaces in this four-dimensional space. It takes time to develop geometric intuition. One of the trickiest points involves intersecting curves. In real space, we are used to lines and curves intersecting in points. When we switch to the complex case, these lines and curves become planes and surfaces and our intuition tells us that these objects intersect along curved lines. But this intuition is wrong. We are in 4-space rather than 3-space, and surfaces in 4-space usually intersect at isolated points. Think, for example, of the xy and zw planes, which intersect at the origin.

For this reason, it is convenient to draw ordinary pictures in the plane even if we are working over the complex numbers, because these plane pictures correctly show how intersections work.

As for points at infinity, we do not add a single point as we did for the Riemann sphere, but instead add different points in different directions. On the other hand, we do not add a point

infinitely far in front of us and a second one infinitely far behind us; instead these two points are identified. Examples gradually show that this is the correct thing to do.

We'll explain in detail how projective space is constructed, but this section can be skimmed because in the bitter end it will turn out that we add exactly one point at infinity to the curve $Y^2 = 4X^3 - g_2X - g_3$, and this point works over the rationals, or the reals, or the complex numbers.

The construction of the real projective plane is best understood from the following picture. The picture shows R^3 and inside it the plane $z = 1$. We identify $(x, y, 1)$ in this plane with (x, y) in the usual plane, so all of the ordinary points we know and love are in $z = 1$.

A point $(x, y, 1)$ generates a line in R^3 through this point and the origin. We will identify the point $(x, y, 1)$ with this line. Almost all lines through the origin in R^3 are accounted for in this way. The lines that aren't are those in the plane $R^2 \subset R^3$. If we jiggle such a line just a little, then it will intersect the plane $z = 1$ far out toward infinity. So lines through the origin in the plane $z = 0$ behave like points at infinity in our $z = 1$ plane. Notice that there is one such line for every direction.

We now *define* real projective space to be the set of all lines in R^3 through the origin. Most such lines hit the $z = 1$ plane and correspond to ordinary points in $z = 1$. The remaining lines correspond to points at infinity for $z = 1$.

Complex projective space is constructed in the same way. A point in this space is a one-dimensional complex subspace of C^3. The plane at $z = 1$ still consists of all $(c_1, c_2, 1)$ and thus can be identified with the complex plane (c_1, c_2). A complex line in the direction (v_1, v_2, v_3) is formed by the set of all $(\lambda v_1, \lambda v_2, \lambda v_3)$, and intersects our plane in a single point when $\lambda = \frac{1}{v_3}$ unless the line is parallel to this plane because $v_3 = 0$.

Suppose we have a curve, say $Y^2 - X^2 = 1$. There is an easy way to add points at infinity to this curve. Add enough Z's to each monomial to make the polynomial homogeneous. In our case we obtain $Y^2 - X^2 = Z^2$. Notice that if (a, b, c) satisfies this equation, then (ra, rb, rc) satisfies

it for all r. So the solutions of this homogeneous polynomial are lines through the origin, and thus points in projective space. Notice that the line through $(a, b, 1)$ satisfies the equation exactly when (a, b) satisfies the original equation. So the solutions of this new homogeneous equation are just the old solutions together with certain solutions at infinity. In the particular case illustrated, a solution at infinity would have the form $(\lambda a, \lambda b, 0)$ and then $(\lambda b)^2 - (\lambda a)^2 = 0^2$, so $b^2 - a^2 = 0$, so $b = \pm a$. This gives two points at infinity, one in the direction $y = x$ and one in the direction $y = -x$. Since our curve is a hyperbola, this is a reasonable result.

Let us finally come to our equation $Y^2 = 4X^3 - g_2X - g_3$. The homogeneous equation is $Y^2Z = 4X^3 - g_2XZ^2 - g_3Z^3$. Points at infinity are given by setting $Z = 0$ and solving $0 = 4X^3$. So they are lies of the form $(0, Y, 0)$; this corresponds to the vertical direction in the plane. In short, there is only one point at infinity, far up in the picture. Our pictures of the curve make this result reasonable.

Remark: So the solutions of $Y^2 = 4X^4 - g_2X - g_3$ in C^2, together with its point at infinity, correspond to some surface sitting in four dimensional space. What does that surface look like? It seems like a hopeless question, but isn't. The surface is a torus.

Theorem 162. *The map $z \to \big(\wp(z), \wp'(z)\big)$, together with the provision that lattice points map to the point at infinity, is one-to-one and onto from the torus to the projective algebraic curve $Y^2 = 4X^3 - g_2X - g_3$.*

Proof: The image is on the curve because $\wp$ satisfies the appropriate differential equation.

Suppose (X, Y) is a point on the curve. Since $\wp(z)$ takes every complex value, there is a z such that $X = \wp(z)$. Then $\wp(z) = \wp(-z)$ and $\wp'(z) = -\wp'(-z)$. If z is not one of the exceptional points e_1, e_2, e_3, then z and $-z$ are different and $\wp'(z) \neq 0$ and thus $\wp'$ takes different values at z and $-z$. And indeed, there are exactly two points $(X, \pm Y)$ with first coordinate X.

If z is one of e_1, e_2, e_3, then there is only one z with $\wp(z) = X$. But in that case $\wp(z) = X$ is a zero of the function $4X^3 - g_2X - g_3$ and therefore $Y = 0$. So there is only one point $(X, 0)$ on the curve with first coordinate X.

QED.

23.11 The Addition Formula

We now come to the climactic moment of our discussion. The theory of doubly-periodic functions began when Fagnano discovered the doubling formula for the lemniscate. That led Euler to find an addition formula for elliptic integrals, and a corresponding addition formula for their inverses. We are about to formulate this theorem in the language of Weierstrass and give its modern proof.

The theorems we will prove are related to the addition formulas of trigonometry. These results

are mysterious and powerful in high school, but later we rewrite them

$$(\cos\theta_1 + i\sin\theta_1)(\cos\theta_2 + i\sin\theta_2) = \cos(\theta_1+\theta_2) + i\sin(\theta_1+\theta_2)$$

and then

$$e^{i\theta_1}e^{i\theta_2} = e^{i(\theta_1+\theta_2)}$$

and ultimately we are looking at a group homomorphism $R \to S^1$ given by $\theta \to e^{i\theta}$.

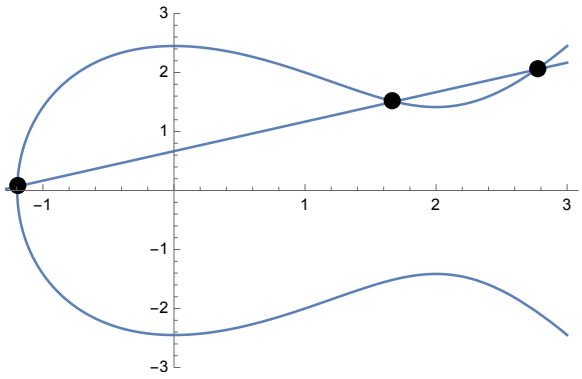
The torus T^2 is a group induced by adding complex numbers modulo a lattice. Let us temporarily introduce $\mathcal{C}$ for the points of the projective curve $Y^2 = 4X^3 - g_2X - g_3$. We have just defined a map

$$T^2 \to \mathcal{C}$$

and our addition theorem is going to claim that this map is a group homomorphism. This will be trivial if we use the map to identify the two objects and define the group structure in $\mathcal{C}$ as the one induced by the structure in T^2. We will not do that.

If we look back at the old result of Fagnano, the remarkable feature is that the limits of the integral giving the doubled length are simple algebraic expressions in the limits of the integral giving the original length. What we would like is a group structure on $\mathcal{C}$ given by simple algebraic expressions.

The key point is that $4X^3 - g_2X - g_3$ is a cubic, and thus this curve intersects lines in three points. So if we are given two points on $\mathcal{C}$, we can draw the line through these points and find the third intersection point. Call these three points p_1, p_2, p_3. The three points play completely symmetric roles, so we cannot expect $p_3 = p_1 + p_2$. Instead we expect $p_1 + p_2 + p_3 = e$, where e is the identity of the group. And indeed, the grand generalization of the research of Fagnano, Euler, and many others, is that this operation makes $\mathcal{C}$ into a group, and our map $T^2 \to \mathcal{C}$ is a group homomorphism.



That is the idea, but there are a wealth of ambiguous details. And maybe the first is: what is the identity of the group $\mathcal{C}$? If our theorem is correct, the map we have defined will take the

identity $0 \in T^2$ to the identity of $\mathcal{C}$. Since $\wp(z)$ has a pole at 0, we must choose $\infty \in \mathcal{C}$ to be the identity of that group.

Recall that ∞ is in the vertical direction. So the straight lines through ∞ will be vertical lines in the plane. Such a line goes through (X, Y) and $(X, -Y)$. So the inverse of (X, Y) must be $(X, -Y)$.

This allows us to precisely define the group operation. Given p_1 and p_2 on our curve, find the straight line through p_1 and p_2 and find the third intersection point $p_3 = (X, Y)$. Then $p_1 + p_2 = (X, -Y)$.

Theorem 163 (The Addition Formula)**.** *The map $z \to (\wp'(z), \wp(z))$ from the torus to the curve $Y^2 = 4X^3 - g_2X - g_3$ is a group isomorphism:*

Proof: Let w_1 and w_2 belong to the torus. Find constants a and b such that

$$\wp'(w_1) - a\wp(w_1) - b = 0$$

$$wp'(w_2) - a\wp(w_2) - b = 0$$

The function $\wp'(z) - a\wp(z) - b$ is doubly periodic with a pole at the origin of order 3 and no other poles. Therefore this function has three zeros, namely w_1, w_2 and one more. By the theorem on zeros and poles of doubly-periodic functions, the sum of these zeros must be 0 in the torus (that is, a lattice point). So the third zero is $-(w_1 + w_2)$. Since $\wp$ is even and $\wp'$ is odd, the corresponding point on the line is $\big(-\wp'(w_1 + w_2), \wp(w_1 + w_2)\big)$. This point is also on the curve, but it is not the sum of the images of w_1 and w_2 because our rule says that we find the third point on the line and then change the sign of its Y component. So the actual sum of the images of w_1 and w_2 is

$$\big(\wp'(w_1 + w_2), \wp(w_1 + w_2)\big)$$

QED.

Remark: We did not actually prove that our addition formula makes the curve into a group. The set certainly has an identity, and each element has an inverse. The missing axiom is associativity. This axiom is true but rather difficult to prove algebraically. But we do not need to give the proof because associativity follows from the addition formula since addition on the torus is associative and our map preserves addition and is one-to-one and onto.

There is still a minor gap. Our proof only works for curves which have an associated lattice and thus associated doubly-periodic functions. We will later prove all curves satisfying $g_2^3 - 27g_3^2 \neq 0$ have such an associated lattice. If this quantity is zero, the curve has a singularity and the group law fails.

23.12 Concrete Addition Formulas

Theorem 164. *The following are concrete addition and duplication formulas for $\wp(z)$:*

$$\wp(u+v) = \frac{1}{4}\left(\frac{\wp'(u)-\wp'(v)}{\wp(u)-\wp(v)}\right)^2 - \wp(u) - \wp(v)$$

$$\wp(2u) = \frac{1}{4}\left(\frac{\wp''(u)}{\wp'(u)}\right)^2 - 2\wp(u) = \frac{1}{4}\left(\frac{\wp(u)\left(6\wp(u)-\frac{1}{2}g_2\right)}{\wp'(u)}\right)^2 - 2\wp(u)$$

$$\wp''(u) = 6\wp^2(u) - \frac{1}{2}g_2\wp(u)$$

Proof: Suppose (X_1, Y_1), (X_2, Y_2), and (X_3, Y_3) all lie in both the line $Y = aX + b$ and the curve $Y^2 = 4X^3 - g_2X - g_3$. Then all three X_i solve $(aX_i + b)^2 = 4X_i^3 - g_2X_i - g_3$. This equation has three roots and thus factors, so

$$4(X - X_1)(X - X_2)(X - X_3) = 4X^3 - g_2X - g_3 - (aX + b)^2$$

Equating coefficients of X:

$$-4(X_1 + X_2 + X_3) = -a^2$$

$$4(X_1X_2 + X_1X_3 + X_2X_3) = -g_2 - 2ab$$

$$-4(X_1X_2X_3) = -g_3 - b^2$$

We only use the first of these results. From it, we can determine X_3 if we know a. It is easy to find a since $Y_1 = aX_1 + b$ and $Y_2 = aX_2 + b$, and thus $a = \left(\frac{Y_1 - Y_2}{X_1 - X_2}\right)$.

Consequently the first equation above gives

$$X_3 = \frac{1}{4}\left(\frac{Y_1 - Y_2}{X_1 - X_2}\right)^2 - X_1 - X_2$$

The sum of (X_1, Y_1) and (X_2, Y_2) is actually $(X_3, -Y_3)$, but no adjustment is needed for the first coordinate. This gives the first formula of the theorem.

The second formula is a consequence of the first if we let $v = u + \Delta u$ and notice that the limit of

$$\frac{1}{4}\left(\frac{\wp'(u)-\wp'(u+\Delta u)}{\wp(u)-\wp(u+\Delta u)}\right)^2 - \wp(u) - \wp(u+\Delta u)$$

equals the limit of

$$\frac{1}{4}\left(\frac{\frac{\wp'(u)-\wp'(u+\Delta u)}{\Delta u}}{\frac{\wp(u)-\wp(u+\Delta u)}{\Delta u}}\right)^2 - \wp(u) - \wp(u+\Delta u)$$

Finally, the third result arises from differentiating the equation

$$\left(\frac{d\wp(z)}{dz}\right)^2 = 4\wp^3(z) - g_2\wp(z) - g_3$$

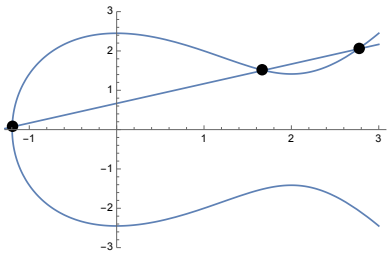
23.13 Full Group Law on the Elliptic Curve

There is a piece of terminology we have not yet used. The curve $Y^2 = 4X^3 - g_2X - g_3$ is called an *elliptic curve*, and sometimes more specifically an *an elliptic curve in Weierstrass form*. History has played a trick on us. The theory started by trying to find the length of an ellipse, and the resulting integrals were called elliptic integrals. Inverting these integrals gave doubly-periodic functions, also called elliptic functions. These solve a differential equation and lead to class of curves called elliptic curves. But these curves are not ellipses.

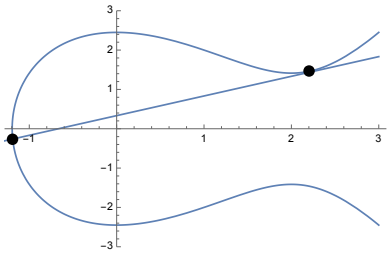
Elliptic curves are studied over many different base fields; in these cases, g_2 and g_3 are elements of the base field. We looked at the fields R and C, but the field Q is intensely studied in number theory. It is also possible to select a finite field and count the number of solutions over that field. In each of these cases we get a group, because the group laws we are obtaining are purely algebraic involving only field operations and not even a square root.

A curve is singular if it does not have a tangent line at some points. Our elliptic curve is singular if and only if $g_2^2 - 27g_3^3 = 0$, a case we have proved does not arise from elliptic function theory. The group law fails if the elliptic curve is singular.

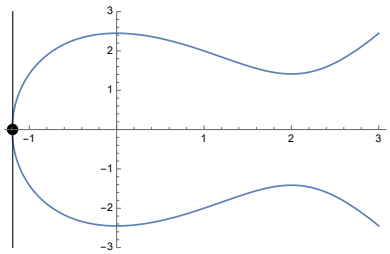
We are going to give the full group law for non-singular elliptic curves. The law falls into two general cases and some exceptions. The first two formulas on the next page cover the so-called *chord case*: we are given two distinct points which are not vertical, draw the chord between them, and find the third intersection point of this chord.



The next two formulas cover the so-called *tangent case*. We select one point whose Y coordinate is not zero and draw the tangent line to the elliptic curve at that point. This line "intersects the curve twice at its tangent." We find the other intersection point.



The final cases cover the cases when the tangent line to our point is vertical. This can only happen when the curve crosses the x-axis.



Theorem 165. *The full group law on the elliptic curve* $Y^2 = 4X^3 - g_2X - g_3$ *is given by the following equations for* $(X_1, Y_1) + (X_2, Y_2) = (X_3, Y_3)$*:*

$$X_3 = \frac{1}{4}\left(\frac{Y_1 - Y_2}{X_1 - X_2}\right)^2 - X_1 - X_2 \quad \text{if } X_1 \neq X_2$$

$$Y_3 = \left(\frac{Y_1 - Y_2}{X_1 - X_2}\right)(X_1 - X_3) - Y_1 \quad \text{if } X_1 \neq X_2$$

$$X_3 = \frac{1}{4}\left(\frac{X_1\left(6X_1 - \frac{1}{2}g_2\right)}{Y_1}\right)^2 - 2X_1 \quad \text{if } X_1 = X_2 \text{ and } Y_1 = Y_2 \text{ and } Y_1 \neq 0$$

$$Y_3 = \left(\frac{X_1\left(6X_1 - \frac{1}{2}g_2\right)}{Y_1}\right)(X_1 - X_3) - Y_1 \quad \text{if } X_1 = X_2 \text{ and } Y_1 = Y_2 \text{ and } Y_1 \neq 0$$

$$(X_1, Y_1) + (X_2, Y_2) = \infty \quad \text{in all remaining cases}$$

Proof: The two formulas for X_3 were obtained in the previous section. The third point is on the line $Y = aX + b$, but the group law changes the sign of Y. So $Y = -aX - b$. We have already determined a formula for a, and since $Y_1 = aX_1 + b$, we have $b = Y_1 - aX_1$. So $Y_3 = -aX_3 - Y_1 + aX_1 = a(X_1 - X_3) - Y_1$.

23.14 Fermat and the Group Law on Elliptic Curves

Surprisingly, these formulas were used by Fermat to solve problems in number theory long before groups were invented. Fermat carefully read a recently printed version of the works of Diophantus of Alexandria, who lived from approximately 200 AD to 290 AD. Diophantus wrote about solving algebraic equations, but in words without our modern algebraic notation, and he was often interested in integer or rational solutions of these equations.

Wikipedia quotes Hermann Hankel about Diophantus as follows: "For our author (Diophantos), not the slightest trace of a general, comprehensive method is discernible; each problem calls for some special method which refuses to work even for the most closely related problems. For this reason it is difficult for the modern scholar to solve the 101st problem even after having studied 100 of Diophantus's solutions."

Many problems considered by Fermat involved polynomials $P(X, Y)$ with rational coefficients, and attempts to find solutions of $P(X, Y) = 0$ with both X and Y rational. Scholars today usually assume that P is irreducible and the resulting curve is non-singular. In that case, the complex solutions in projective space form a compact surface of genus g for some $g \geq 0$. Surprisingly, this g has a strong influence on the rational solutions of the equation. When $g = 0$, we say the curve is *rational* and in that case the method we used to find rational points on the circle $X^2 + Y^2 - 1$ often generalizes. When $g = 1$, we say the curve is *elliptic* and then the techniques of this section come into play. When $g \geq 2$, a celebrated theorem of Gerd Faltings proved in 1983 says that there are only finitely many rational solutions.

Fermat is essentially the first mathematician to seriously investigate the elliptic case. An example shows how very simple problems can lead to such equations. Fermat tried to find a Pythagorian triangle whose area is a perfect square. This is a problem in integers, but it immediately translates into a problem about rational numbers because $a^2 + b^2 = c^2$ if and only if $\left(\frac{a}{c}\right)^2 + \left(\frac{b}{c}\right)^2 = 1$. If the area of the resulting triangle with rational sides is a rational square, then we can go backward and find that the area of the triangle with integer sides is an integral square.

But we know that rational points on the circle are given by

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

and the area of such a triangle is

$$\frac{1}{2}xy = \frac{t(1-t^2)}{(1+t^2)^2}$$

So we are looking for a rational t making this last expression a square, or equivalently a point on the curve $Y^2 = t(1-t^2)$. This curve has rational solutions since $(0,0),(0,1),(0,-1)$ are on the curve, but unhappily each gives a degenerate triangle. Fermat wanted to find other solutions.

Eventually Fermat proved that there are no other rational solutions using a technique he invented called the *method of infinite descent*. He assumed he had a non-trivial solution and then showed that from such a solution he could produce a "smaller" solution. Here the size of a solution is measured by the complexity of the numerators and denominators of the solution. But it is impossible to have infinitely many smaller and smaller solutions, so none can exist.

In many other elliptic cases, however, Fermat found many solutions using what he called the "chord and tangent method." These are exactly our rules for addition on elliptic curves. For instance, given a solution, he found the tangent line to the curve at that point. This line intersected the curve twice at the solution and once somewhere else. That third point must also have rational coordinates since it can be found by elementary algebra. Continue. One reason this method was attractive was that it very rapidly led from simple solutions to solutions given by fractions with incredibly large numerators and denominators. If a reader didn't know about the chord and tangent method, it would seem miraculous that such a solution was ever found.

23.15 Brief Survey of Modern Results

The study of rational solutions of an elliptic curve is an extremely active area of modern research. Perhaps a starting point came around 1901, when Poincare observed that the rational points on an elliptic curve form a group and asked if this group is finitely generated. That is, are there finitely many solutions such that all other solutions are given by the chord and tangent method.

In 1922, Louis Mordell proved that there are. Indeed he proved that the group of rational points is a finitely generated abelian group, and thus has the form

$$Z \times Z \times \ldots Z \times Z_{n_1} \times \ldots \times Z_{n_k}$$

Mordell's proof involved a beautiful generalization of Fermat's method of descent. At one crucial point, the proof is nonconstructive, so for example it does not explain how to find the number of Z factors.

However, using hints from the proof and other special tricks, the number of Z's can often be determined. For the torsion portion of the group, the situation is very much better. For each

special case, there *is* an algorithm to determine the torsion component. Moreover, it is easy to prove that it always has the form Z_n or $Z_2 \times Z_n$. In 1978, Barry Mazur found a list of the exact possible groups. They are Z_n for $n = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$ or 12 and $Z_2 \times Z_n$ for $n = 2, 4, 6, 8$. So there are exactly 15 possible torsion groups.

As for the rank of these groups (the number of Z terms), the highest known rank is 24, found in 2000 by Martin and McMillen. In 2006, Elkies found an example where the rank is unknown, but at least 28. The site https://web.math.pmf.unizg.hr/~duje/tors/rankhist.html lists equations and generators for these examples and others.

It is unknown whether the rank can be arbitrarily large. But gradually a conjecture has developed about average ranks. According to the conjecture, half of elliptic curves have rank zero and half have rank one. That is, if you pick a curve at random, it might have high rank, but chances are that it will have rank either zero or one.

I do not know the current status of these conjectures, but important work has been done by Manul Bhargava and Arul Shankar. As of 2014, they had proved that at least 20.6% of elliptic curves have rank 0 and at least 83.75% have rank either zero or one. See https://mattbaker.blog/2014/03/10/the-bsd-conjecture-is-true-for-most-elliptic-curves/ and search the web for more recent results.

It is also possible to compute the size of elliptic curves over finite fields Z_p. It is easy to estimate the rough size of these groups. The Z_p plane has p^2 points, and evaluating $Y^2 - 4X^3 - g_2X - g_3$ on a point can give p different values. If each of these values is equally likely, then we expect about $p^2/p = p$ to be zero. So when we calculate the actual number of points, we get values which hover around p, sometimes a little larger and sometimes a little smaller.

Now a very rough conjecture. If these numbers are more often a little larger than p, then perhaps the curve has a lot of rational points and so positive rank. But if these numbers are often a little smaller than p, then perhaps the curve has few rational points and so zero rank. In the early 1960's, Peter Swinnerton-Dyer computed the number of points modulo p for a large number of primes. From these calculations, Birch and Swinnerton-Dyer made an initial conjecture making more explicit the idea that many solutions over p implies positive rank. This conjecture was then made far more explicit, eventually leading to a formula for the rank from information about the number of solutions over p. This conjecture of Birch and Swinnerton-Dyer is one of the million dollar prize problems of the Clay institute, and specific details about the conjecture can be found at their web site http://www.claymath.org/millennium-problems/birch-and-swinnerton-dyer-conjecture.

23.16 Appendix on Discriminant

Earlier we introduced the quantity $g_2^3 - 27g_3^2$, which measures whether the polynomial

$$P(X) = 4X^3 - g_2X - g_3 = 4(X - e_1)(X - e_2)(X - e_3)$$

has multiple roots. This expression is called the *discriminant* of the polynomial.

Theorem 166. *If the roots of a polynomial $P(X) = 4X^3 - g_2X - g_3$ are e_1, e_2, e_3, then*

$$g_2^2 - 27g_3^3 = \frac{1}{4}(e_1 - e_2)^2(e_1 - e_3)^2(e_2 - e_3)^2$$

Remark: This expression is symmetrical in the roots; any permutation of the roots leaves it invariant. In turn, the coefficients of the polynomial are given by the elementary symmetric polynomials

$$0 = e_1 + e_2 + e_3$$
$$-\frac{g_2}{4} = e_1e_2 + e_1e_3 + e_2e_3$$
$$-\frac{g_3}{4} = e_1e_2e_3$$

A general theorem in algebra states that any symmetric polynomial $P(X_1, \dots, X_n)$ can be written as a polynomial in the elementary symmetric polynomials $Q(\sigma_1(X_1, \dots, X_n), \dots, \sigma_n(X_1, \dots, X_n))$ and if P has integer coefficients, so does Q. The discriminant $(e_1 - e_2)^2(e_1 - e_3)^2 \dots (e_{n-1} - e_n)^2$ of any polynomial can be written as an integer combination of the roots. This expression is very complicated.

Proof: Differentiating $4X^3 - g_2X - g_3$ gives

$$12X^2 - g_2 = 4\left[(X - e_2)(X - e_3) + (X - e_1)(X - e_3) + (X - e_1)(X - e_2)\right]$$

and so

$$12e_1^2 - g_2 = (e_1 - e_2)(e_1 - e_3)$$
$$12e_2^2 - g_2 = (e_2 - e_1)(e_2 - e_3)$$
$$12e_3^2 - g_2 = (e_3 - e_1)(e_3 - e_2)$$

or

$$(12e_1^2 - g_2)(12e_2^2 - g_2)(12e_3^2 - g_2) = -(e_1 - e_2)^2(e_1 - e_3)^2(e_2 - e_3)^2$$

or

$$(e_1 - e_2)^2(e_1 - e_3)^2(e_2 - e_3)^2 =$$
$$-(12)^3(e_1e_2e_3)^2 + (12)^2g_2(e_1^2e_2^2 + e_1^2e_3^2 + e_2^2e_3^2) - 12g_2^2(e_1^2 + e_2^2 + e_3^2) + g_2^3$$

Notice that

$$0 = (e_1 + e_2 + e_3)^2 = e_1^2 + e_2^2 + e_3^2 + 2(e_1e_2 + e_1e_3 + e_2e_3)$$

and so

$$e_1^2 + e_2^2 + e_3^2 = -2(e_1e_2 + e_1e_3 + e_2e_3) = \frac{g_2}{2}$$

Also $e_1e_2e_3 = -\frac{g_3}{4}$. We only need a formula for one more term in the formula at the bottom of the previous page to complete the argument.

We have

$$\left(\frac{g_2}{2}\right)^2 = (-2(e_1e_2 + e_1e_3 + e_2e_3))^2 = 4(e_1^2e_2^2 + e_1^2e_3^2 + e_2^2e_3^2) + 8(e_1^2e_2e_3 + e_2^2e_1e_3 + e_3^2e_1e_2)$$

The last term on the right is $e_1e_2e_3(e_1 + e_2 + e_3) = 0$, so

$$e_1^2e_2^2 + e_1^2e_3^2 + e_2^2e_3^2 = \frac{g_2^2}{16}$$

This gives us enough information to complete the formula at the bottom of the previous page, and we obtain

$$(e_1 - e_2)^2(e_1 - e_3)^2(e_2 - e_3)^2 =$$

$$-(12)^3\left(-\frac{g_3}{4}\right)^2 + (12)^2g_2\left(\frac{g_2^2}{16}\right) - 12g_2^2\left(\frac{g_2}{2}\right) + g_2^3 =$$

$$-27 \cdot 4g_3^2 + 9g_2^3 - 6g)2^3 + g_2^3 = 4\left(g_2^3 - 27g_3^2\right)$$

QED.

Chapter 24

The Modular Group

24.1 Conformal Equivalence of Tori

We now begin the second half of our investigation, which involves studying when two tori are conformally equivalent. Historically this investigation traces back to the discovery that elliptic integrals contain an extra parameter m related to the eccentricity of the ellipse.

Notice first that each torus has a large number of conformal automorphisms, namely translations. Eventually we will find the full automorphism group, which depends on the particular torus.

Given two tori T_1 and T_2, we want to study holomorphic maps $f : T_1 \to T_2$. Following f with a translation if necessary, we can assume that $f(0) = 0$. Each torus is formed by identifying points of the plane equivalent via a lattice, so we get the diagram below:

$$\begin{array}{ccc} C & & C \\ \downarrow{\scriptstyle \pi_1} & & \downarrow{\scriptstyle \pi_2} \\ T_1 & \xrightarrow{f} & T_2 \end{array}$$

We claim that the map f can be lifted to a map $\widetilde{f}$ making the diagram commute:

$$\begin{array}{ccc} C & \xrightarrow{\widetilde{f}} & C \\ \downarrow{\scriptstyle \pi_1} & & \downarrow{\scriptstyle \pi_2} \\ T_1 & \xrightarrow{f} & T_2 \end{array}$$

Since C is simply connected and the maps $C \to T$ are covering maps, covering space theory shows that the lift of this map to a map $C \to C$ exists. If we include base points and require that all maps take 0 to 0, this map is unique. Because the lift is locally f, it is holomorphic.

Suppose f is a conformal equivalence. Then we can repeat the same argument using f^{-1} and obtain a unique lift of f^{-1} sending 0 to 0.

Composing the lift of f with the lift of f^{-1} gives a lift of $f^{-1} \circ f : T_1 \to T_1$. This is the identity map, and one such lift is $z : C \to C$, so the composition of our lifts must be this map. Similarly we get the identity when we compose in the other order. So each lift is one-to-one and onto and thus an automorphism of C. We already know that every such automorphism has the form $z \to \lambda z + \tau$ for constants λ and τ. Since our lifts map 0 to 0, they have the form $z \to \lambda z$ for a non-zero constant λ. We have proved the following theorem:

Theorem 167. *Every conformal equivalence from T_1 to T_2 sending 0 to 0 is induced by a map from C to C of the form $z \to \lambda z$.*

Note: Readers who have not studied covering space theory can still convince themselves that $\widetilde{f}$ exists. Indeed we essentially gave the proof when we defined complex integration over an arbitrary continuous path. Suppose $\gamma(t)$ is a path in C from the origin to some other point z. Then $\pi(\gamma(t))$ is a similar path in T_1 and $f \circ \pi(\gamma(t))$ is a similar path starting in 0 in T_2. We can lift this path to a path in C starting at 0 by the argument given in that chapter on integration. Define $\widetilde{f}(z)$ to be the end of this lifted path.

If we have two different paths in C starting at 0 and ending at z, then they are homotopic by a homotopy which fixes 0 and z. So the projected paths in T_1 are homotopic, and after applying f, the corresponding paths in T_2 are homotopic. By the exact argument we gave in integration theory, the lifts back up to C are homotopic and thus end at the same point. So our induced map from C to C is well-defined.

It is now a matter of cleaning up the pieces to finish the argument.

24.2 Consequences of the Lift

Suppose a torus T corresponds to a lattice with basis ω_1 and ω_2. Map C to C by $z \to \lambda z$ where $\lambda = \frac{1}{\omega_1}$. We obtain a new lattice whose first generator is 1. If the second generator winds up in the lower half plane, we can replace it by its negative.

Consequently every torus is conformally equivalent to a torus associated with a lattice with basis $\{1, \tau\}$ where τ is in the upper half plane.

You might think this would be the end of the story, but it isn't because the basis for a lattice is not unique. If ω_1 and ω_2 form a lattice basis and a, b, c, d are integers, then $a\omega_2 + b\omega_1$ and

$c\omega_2 + d\omega_1$ are in the lattice. If the determinant of the matrix

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

is one, the inverse matrix also has integer coefficients and thus ω_1 and ω_2 can be written as integer combinations of $a\omega_2 + b\omega_1$ and $c\omega_2 + d\omega_1$, It follows that these two numbers form a new basis for the lattice. Requiring the determinant of the matrix to be 1 rather than -1 makes the map preserve orientations, so if ω_2 is in the plane left of ω_1, the new basis will still have this property. The group of such matrices is called $SL(2, Z)$.

So if we want to determine if two tori T_1 and T_2 are conformally equivalent, we can replace T_1 by a conformally equivalent torus generated by 1 and τ_1 in the upper half plane, and we can replace T_2 by a conformally equivalent torus similarly generated by 1 and τ_2. Then we want to determine when we can find λ so λ times the second lattice is the first lattice. This will happen exactly if the lattice generated by λ and $\lambda\tau_2$ is the lattice generated by 1 and τ_1. So it happens exactly if we can find a matrix in $SL(2, Z)$ such that $\lambda = c\tau_1 + d$ and $\lambda\tau_2 = a\tau_1 + b$. And these equations will be true exactly if $\tau_2 = \frac{a\tau_1+b}{c\tau_1+d}$. We have proved the following theorem:

Theorem 168. *Every torus is conformally equivalent to a torus given by a lattice with basis* $\{1, \tau\}$*, where* τ *is in the upper half plane. Two points* τ_1 *and* t_2 *represent conformally equivalent tori if and only if there is a matrix* $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ *in* $SL(2, Z)$ *with*

$$t_2 = \frac{a\tau_1 + b}{c\tau_1 + d}$$

Definition 53. *The set of all maps from the upper half plane to itself given by actions of* $SL(2, Z)$ *as above is called* the modular group.

Remark: Notice that we can change the sign of all four a, b, c, d without changing the map. The element $\begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$ is in the center of $SL(2, Z)$ and generates a subgroup of order two, and so the modular group is

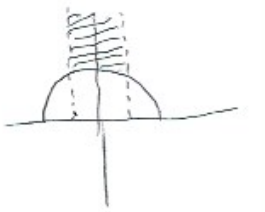
$$SL(2, Z)/\{\pm I\}$$

24.3 The Modular Diagram

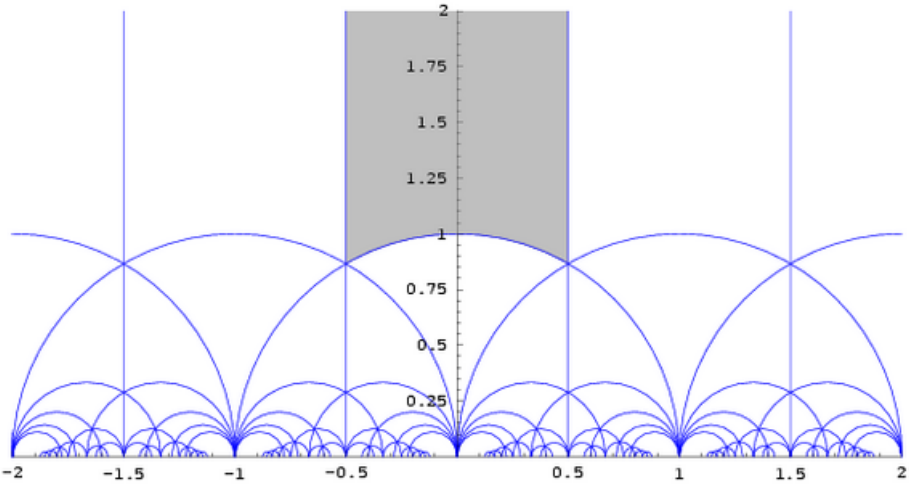
Since $\begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix}$ belongs to $SL(2, R)$, the map $\tau \to \tau + n$ is in the modular group. In particular, the map $\tau \to \tau + 1$ is in the group. Since $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ belongs to $SL(2, R)$, the map $\tau \to -\frac{1}{\tau}$

belongs to the group. It will eventually turn out that these two elements generate the entire group.

Because the first element is in the group, every torus is conformally equivalent to a torus with $-\frac{1}{2} \leq \Im(\tau) < \frac{1}{2}$. Since the second element is in the group, every torus is conformally equivalent to a torus with $|\tau| \geq 1$. It is not clear that we can map any τ to an element satisfying both conditions at once, but we will prove that this can be done. See the picture below.



Indeed, further investigation leads to the picture on the following page, one of the most famous in the entire mathematical literature.



Definition 54. *The fundamental domain of the modular group is the set of all τ satisfying the inequalities $-1/2 \leq Im(\tau) < 1/2$ and $|\tau| \geq 1$. If $|\tau| = 1$, we require $Im(\tau) \leq 0$.*

Remark: In short, the fundamental domain is the shaded region pictured above, but we only take boundary points on the left side of the region.

Members of the modular group map the fundamental domain to the other connected regions in the diagram. Each of these regions comes from a unique element of the modular group, so the modular group is essentially the set of these connected regions. The picture clearly shows

the regions obtained by translating the fundamental region left and right. If we first apply $-\frac{1}{z}$ and then translate, we get the large triangular regions below the fundamental region, and so forth.

Theorem 169. *Any point in the upper half plane can be mapped to a point in the fundamental region by an element of the modular group. This point is unique.*

Proof: We first claim that

$$Im\left(\frac{a\tau+b}{c\tau+d}\right) = \frac{Im(\tau)}{|c\tau+d|^2}$$

This follows from the brief calculation below because $ad-bc=1$:

$$\frac{a\tau+b}{c\tau+d} = \frac{(a\tau+b)(c\overline{\tau}+d)}{(c\tau+d)(c\overline{\tau}+d)} = \frac{(ad+bc)Re(\tau)+(ad-bc)Im(\tau)}{|c\tau+d|^2}$$

Next we notice that for each positive integer k, there are only finitely many integers c and d with $|c\tau+d|^2 \leq k$. Indeed, this expression is $(cRe(\tau)+d)^2+(cIm(\tau))^2$. Only finitely many integers c satisfy $(cIm(\tau))^2 \leq k$, and for each such integer the values of $c\Re(\tau)+d$ form a sequence of numbers along the real line, each of distance one from its neighbors, so only finitely many of them have absolute value bounded by $\sqrt{k}$.

From these facts, it follows that any point in the upper half plane has a representative of maximal height. Select one of these representations and translate it by $z \to z+n$ until it is between $-1/2$ and $1/2$ and on the left boundary if there is a choice. The new representative must have absolute value greater than or equal to one, for otherwise $-\frac{1}{\tau}$ would be higher. If the new representative has absolute value exactly 1 but is on the right side, $-\frac{1}{z}$ will map it to a point with absolute value 1, but on the left side. Hence every point can be mapped to a point in the fundamental region.

We now must consider cases when τ is in the fundamental region, $M \neq I$ is in the modular group, and $M(\tau)$ is also in the fundamental region. In this case, we can suppose that $Im(\tau) \leq Im(M(\tau))$, since otherwise we can work with M^{-1} instead. Thus

$$\frac{Im(\tau)}{|c\tau+d|^2} \geq Im(\tau)$$

and so $|c\tau+d| \leq 1$.

Notice that all points in the fundamental region has imaginary part greater than or equal to $\frac{\sqrt{3}}{2}$. So if τ and $M(\tau)$ are both in the fundamental region, $Im(\tau) \geq \frac{\sqrt{3}}{2}$ and $\frac{Im(\tau)}{|c\tau+d|^2} \geq \frac{\sqrt{3}}{2}$. Write

$\tau = u + iv$, so $v \geq \frac{\sqrt{3}}{2}$. Then $\frac{Im(\tau)}{|c\tau+d|^2} = \frac{v}{(cu+d)^2+(cv)^2} \geq \frac{\sqrt{3}}{2}$. So

$$\frac{\sqrt{3}}{2} \leq \frac{v}{(cu+d)^2+(cv)^2} \leq \frac{v}{(cv)^2} = \frac{1}{c^2 v} \leq \frac{1}{c^2\frac{\sqrt{3}}{2}}$$

and $c^2 \leq \frac{4}{3}$. So $c = -1, 0$, or 1. If $c = 0$, then $ad = 1$, so $a = d = 1$ or $a = d = -1$. Changing the sign of all matrix elements gives the same map, so we can assume $a = d = 1$ and the map is a translation. No non-trivial translation maps a point in the fundamental region back into the fundamental region. So we can assume $c = \pm 1$. If $c = -1$, change the sign of all matrix elements, leaving the map fixed. So from now on, $c = 1$.

Then $ad-b = 1$, so $b = ad-1$. Thus the matrix is $\begin{pmatrix} a & ad-1 \\ 1 & d \end{pmatrix}$ and the map is $\tau \to \frac{a\tau+ad-1}{\tau+d} = a - \frac{1}{\tau+d}$. Recall that $|c\tau + d| \leq 1$, so $|\tau + d| \leq 1$. Thus $|u + iv + d|^2 = (u + d)^2 + v^2 \leq 1$, so $|u+d| \leq 1$. So $|d|-\frac{1}{2} \leq |d|-|u| \leq |d+u| \leq 1$ implies that $d = 0, 1, -1$, and the map is $a-\frac{1}{\tau+1}$ or $a - \frac{1}{\tau-1}$ or $a - \frac{1}{\tau}$. Note that each of these maps is a specific map followed by a translation. We concentrate on the action of the specific map and then consult the modular picture to see if any translation of the image of the map will get us back into the fundamental region.

For example, the map $\tau \to -\frac{1}{2}$ reflects across the circle and sends the fundamental region into the triangular area with vertex on the x -axis. None of these points could translate to belong to the fundamental region, except points on the circle. But on the circle our map is

$$-\frac{1}{\tau} = -\frac{\overline{\tau}}{\tau\overline{\tau}} = -\overline{\tau}$$

so $u + iv \to -u + iv$. The point i is left fixed; other points inside the fundamental region move outside and translations leave them outside. But there is one exception. The bottom left corner moves to the bottom right corner, and if we translate by -1, we are back where we started. So $-\frac{1}{z} - 1$ leaves $-\frac{1}{2} + \frac{\sqrt{3}}{2}$ fixed.

The map $\frac{-1}{\tau+1}$ leaves $-\frac{1}{2} + \frac{\sqrt{3}}{2}$ fixed; this is again the bottom-left corner of the fundamental region. A brief calculation shows that if we compose $-\frac{1}{z} - 1$ with itself, we get $\frac{-1}{z+1}$. If we linearly approximate these maps at the fixed point, our picture shows that $-\frac{1}{z} - 1$ rotates counterclockwise by $\frac{2\pi}{3}$, so the composition $\frac{-1}{z+1}$ rotates around the fixed point by $2\frac{2\pi}{3}$. Thus it maps the fundamental region into the triangle with two curved sides and one straight side, and non-trivial translations of this triangle are never in the fundamental region.

Finally we study the map $a - \frac{1}{z-1}$. When $a = 0$, a brief calculation shows that this map fixes $\frac{1}{2} + \frac{\sqrt{3}}{2}i$. So it must send the fundamental region to a rotated region at this point. We can measure the rotation by calculating the derivative at the fixed point. The derivative is $\frac{1}{(z-1)^2}$ and the value of this at $\frac{1}{2}+\frac{\sqrt{3}}{2}i$ is $\left(\frac{1}{2} + \frac{\sqrt{3}}{2}i\right)^2$, which is a rotation of $2\frac{2\pi}{3}$. The picture then shows that the fundamental region rotates into a triangle with one straight side hanging down from the fixed point, and no translations of this triangle intersect the fundamental region except at images of the fixed point. But we are studying situations where τ is in the fundamental region and $M(\tau)$ is also in the fundamental region. If our M is $a - \frac{1}{1-z}$, then the only possible τ in the fundamental region satisfying this condition is the fixed point, but that is not in the fundamental region.

This finishes the entire proof. The special transformations we discovered in the second half of the proof will be important for us, so we study them next. QED.

24.4 Non-Euclidean Geometry

This is an aside, not used elsewhere in the notes.

In differential geometry, it is common to describe a distance function infinitesimally near each point. The lengths of curves can then be determined by integration. Using the calculus of variations, it is possible to deduce the differential equation satisfied by geodesics, which are the shortest curves between points. Solving this equation then gives the geodesics of the geometry.

For example, to get Euclidean geometry in the plane, we let distance be $ds = \sqrt{dx^2 + dy^2}$. This means that if a curve moves from a point (x, y) to a point $x + dx, y + dy$, the distance it travels is $ds = \sqrt{dx^2 + dy^2}$. The length of a parametric curve $\gamma(t) = (x(t), y(t))$ is then given by

$$\int_{t=a}^{b} \sqrt{dx^2 + dy^2}\, dt = \int_{t=a}^{b} \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{d}{dt}\right)^2}\, dt$$

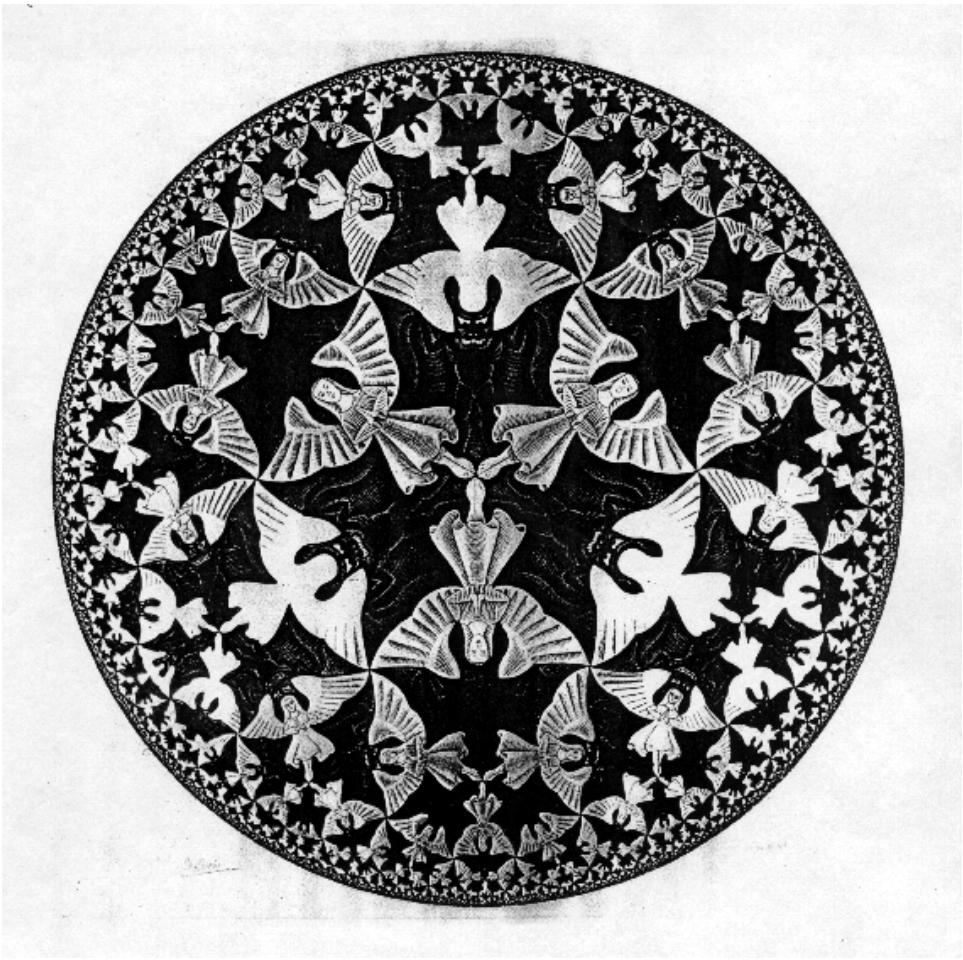
This integral can be taken to be the definition of the formula for infinitesimal distance. The differential equation for geodesics turns out to be $\gamma''(t) = 0$ and thus geodesics are straight lines $\gamma(t) = p + tv$.

A second important example is obtained by providing the unit disk with the distance function

$$ds = \frac{2\sqrt{dx^2 + dy^2}}{1 - (x^2 + y^2)}$$

According to this formula, distance near the origin is what we think it is, but distances near the boundary are much greater than shown in the picture. The 2 is there for historical reasons and can be ignored.

It turns out that with this metric, geodesics are straight lines through the origin and circles which meet the boundary at ninety degrees. One of Escher's most famous pictures was drawn using this geometry. The picture is called *Angels and Devils*.



In the geometry, all of the angels have the same size and all of the devils have the same size. Notice straight lines of symmetry through the origin which bisect both angels and devils. Then notice similar bisecting curves nearer the boundary which form circles perpendicular to the boundary. The picture contains infinitely many angels and devils.

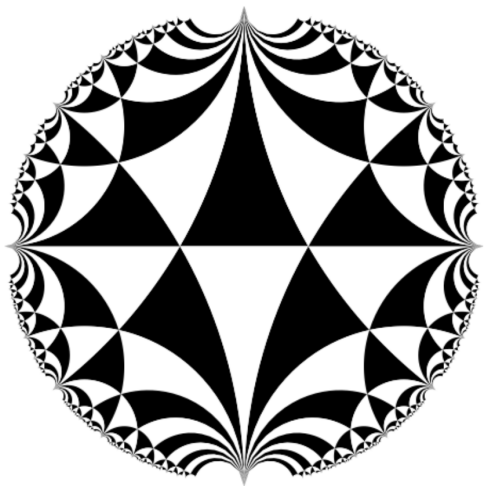
Where does this geometry come from? Recall that Euclid assumed a parallel postulate. The modern version of this postulate states that whenever we have a straight line L and a point P not on L, we can draw a unique line through P parallel to L. It was commonly believed in ancient times that this should have been a theorem rather than a postulate, and many people tried to give a proof. Even Legendre gave such a proof. But around that time, Gauss, Bolya, and Lobachevsky realized that there could be no such theorem because there is a second geometrical object which satisfies all of Euclid's postulates except this one, and in which infinitely many lines through P are parallel to L. Miraculously, it turns out that there is *only one* such geometry, which we now call *Non-Euclidean Geometry*. Much later, Poincare realized that the unit disk, with the above metric, is a model of Non-Euclidean geometry, That is why this particular metric is so important.

Earlier in these notes we proved that the upper half plane and the unit disk are conformally equivalent. It follows that the upper half plane is also a model for Non-Euclidean geometry if we give it the correct metric. It turns out that the appropriate metric is

$$ds = \frac{\sqrt{dx^2 + dy^2}}{y^2}$$

In this metric the geodesics are straight vertical lines and circles which meet the x-axis at ninety degrees.

Notice that the modular picture is constructed using such curves. In Non-Euclidean geometry, triangles can have one vertex at infinity and yet have finite area. The modular diagram is a tiling of the Non-Euclidean plane by such triangles. Each triangle in this tiling has the same size in Non-Euclidean geometry.



It is possible to draw the modular diagram in a disk rather than the upper half plane. We then get a picture similar to the picture of Angles and Devils, but using only lines and circles. Escher's picture is a tiling of the Non-Euclidean plane, but not the modular tiling.

24.5 Conformal Automorphisms of Tori

The covering space argument at the beginning of this chapter shows that every automorphism of one of our tori T is covered by a map of the form $A(z) = az + b$. This map induces a map from T to T exactly when it preserves equivalence classes of points. If L is the lattice, this means that $z_1 - z_2 \in L$ should imply $A(z_1) - A(z_2) \in L$. Simple algebra shows that this condition is equivalent to requiring that if $z \in L$, then $az \in L$.

Automorphisms are maps whose inverses are also well-defined, so $A(z) = az + b$ is an automorphism if $aL \subset L$ and $a^{-1}L \subset L$. It is easy to see that if both conditions hold, a must equal $e^{i\theta}$ for some θ and thus our map is a rotation followed by a translation.

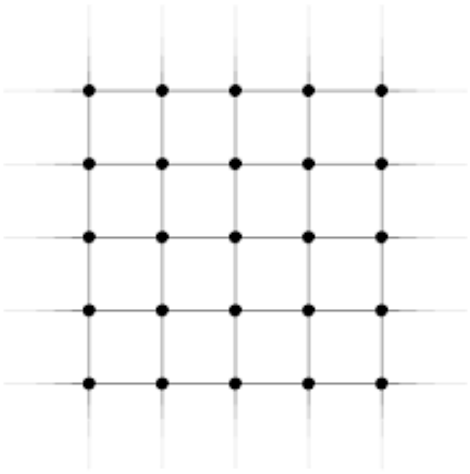
The conditions required of a obviously hold when $a = \pm 1$, so for any lattice or torus, translations induce automorphisms and maps $A(z) = -z + b$ induce automorphisms.

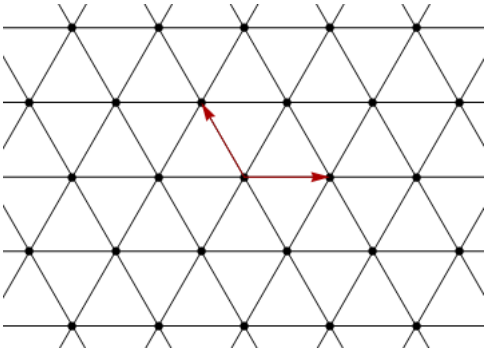
In C, the map $-z + b$ has a unique fixed point at $z = \frac{b}{2}$. So the map is rotation about this point by π.

The action on the torus is a little trickier. Let us first assume $b = 0$ and search for fixed points. A point is fixed if $z = -z + m_1\omega_1 + \omega_2$ or $z = \frac{m_1}{2}\omega_1 + \frac{m_2}{2}\omega_2$. Thus this map has four fixed points in a torus, the origin and e_1, e_2, e_3. Near these points our map looks like a rotation by π.

The same result holds for the map $A(z) = -z + b$, but the fixed points are simply translated slightly. It may be difficult to imagine how four points at once can be fixed by a rotation. But look at the following torus, stick a pin through the four indicated points, and flip the torus up and over while holding the pin fixed. The four pin holes are fixed by this map, which rotates around each fixed point by π.

Translations and these rotations by π are the only automorphisms of most tori. Bu we get special tori when $\tau = i$ or $\tau = -\frac{1}{2} + \frac{\sqrt{3}}{2}i$. These τ give the two famous symmetric lattices in the plane, the square lattice and the hexagonal lattice.





In the square case, multiplication by a preserves the lattice only if $a = \pm 1$ or $\pm i$. These are rotations by $0, \pi/2, \pi$, or $3\pi/2$. In the hexagonal case, multiplication by a preserves the lattice exactly when $a = \pm 1, \pm\left(\frac{1}{2} + \frac{\sqrt{3}}{2}i\right)$, or $\pm\left(-\frac{1}{2} + \frac{\sqrt{3}}{2}i\right)$. These are rotations by $0, \pi/6, 2\pi/6, \pi, 4\pi/6$, or $5\pi/6$. We leave it to the user to determine the fixed points of these rotations in the torus when there is no translation, and then in general.

24.6 The Modular Function j

We now begin a search for a holomorphic function on the upper half plane which is invariant under the action of the modular group. Such a function would attach a number to each torus, and conformally equivalent tori would be assigned the same number. So these numbers could help distinguish tori which belong to different conformal equivalence classes. That is essentially the role of the eccentricity in the theory of elliptic integrals, so we are attempting to find an analogue for doubly-periodic functions.

The differential equation satisfied by an elliptic function depends on two parameters g_2 and g_3 and the formulas for these parameters are

$$g_2 = 60 \sum_{\omega \neq 0} \frac{1}{\omega^4} \qquad g_3 = 140 \sum_{\omega \neq 0} \frac{1}{\omega^6}$$

We can think of these as functions on the upper half plane by writing

$$g_2 = 60 \sum_{\omega \neq 0} \frac{1}{(m_1\tau + m_2)^4} \qquad g_3 = 140 \sum_{\omega \neq 0} \frac{1}{(m_1\tau + m_2)^6}$$

It is natural to generalize and write

$$G_{2k}(\tau) = \sum_{\omega \neq 0} \frac{1}{(m_1\tau + m_2)^{2k}}$$

These series converge for $k > 1$ and are not defined for odd indices because the terms would cancel out in the odd case. The sums are called Eisenstein series after the mathematician who first introduced them. Each is a holomorphic function on the upper half plane.

Eisenstein series are *not* invariant under the modular group, but they do satisfy simple transformation formulas.

Theorem 170. *For each* $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, Z)$,

$$G_{2k}\left(\frac{a\tau + b}{c\tau + d}\right) = (c\tau + d)^{2k} G_{2k}(\tau)$$

Proof: Since $\tau, 1$ and $a\tau + b, c\tau + d$ are different lattice bases for the same lattice, we have

$$G_{2k}(\tau) = \sum \frac{1}{(m_1(a\tau + b) + m_2(c\tau + d))^{2k}} = \frac{1}{(c\tau + d)^{2k}} \sum \frac{1}{\left(m_1\left(\frac{a\tau+b}{c\tau+d}\right) + m_2\right)^{2k}}$$

and the final sum is

$$\frac{1}{(c\tau + d)^{2k}} G_{2k}\left(\frac{a\tau + b}{c\tau + d}\right)$$

Remark: Now we would like to form an expression in these G_{2k} which is invariant under the modular group. This will involve dividing two polynomials in the terms. Recall that $g_2^3 - 27g_3^2$ is never zero. This is a perfect denominator; it changes by $(c\tau + d)^6$ under a modular transformation. The term g_2^3 changes by the same factor, and therefore $j(\tau) = \frac{g_2^3}{g_2^3 - 27g_3^2}$ does not change at all.

Definition 55. *The* modular function *is the following function, holomorphic on the upper half plane and invariant under the modular group:*

$$j(\tau) = \frac{g_2^3}{g_2^3 - 27g_3^2}$$

Warning: Some authors multiply our quotient by 1728. We will see why shortly. Caution is required when reading other sources to understand the conventions being used.

We obtain the following astonishing theorem:

Theorem 171. *The function j is one-to-one and onto from the fundamental region to the complex plane. Consequently, two tori are conformally equivalent if and only if they are associated with the same value of*

$$\frac{g_2^3}{g_2^3 - 27g_3^2}$$

Remark: The ancient theory of ellipses, which began our theory, introduced a single parameter called eccentricity to distinguish ellipses in a fundamental way. And now, at the end of this long development, we have again found a single parameter to distinguish lattices and fields of doubly-periodic functions.

It is particularly fortunate that this parameter is defined using g_2 and g_3 and thus makes sense algebraically when studying abstract fields and algebraic curves associated with doubly-periodic functions.

Proof: The proof of this result is one of the most beautiful in complex analysis. We start by glueing the boundary of the fundamental region together to make a new geometric object. The modular transformation $z \to z+1$ maps the left side to the right side and identifies these sides. The modular transformation $z \to \frac{-1}{z}$ maps the left half of the bottom circle to the right half and identifies these sides. After making the two identifications, the fundamental region becomes homeomorphic to the plane. Moreover, the points in the upper half plane with great height correspond to points far from the origin in this new topological plane.

So we can introduce a point at infinity, exactly as we did when we constructed the Riemann Sphere. With this extra point, the fundamental region together with the point at infinity form a space homeomorphic to a sphere.

Next we give this space a complex structure in the sense that each point has a neighborhood that looks like an open set in the plane and these neighborhoods glue together by holomorphic maps. After this is done, the function j becomes a holomorphic function on the space, except at ∞. We will show that ∞ is a pole of the function, so j is meromorphic on our sphere.

We easily prove that any meromorphic function on our new space must take each value the same number of times. To finish the argument, we need only prove that our function takes ∞ once, including multiplicities. Therefore, all of this abstract nonsense reduces the proof to showing that ∞ is a pole of j of order one.

At the end of the argument, we will avoid the abstract nonsense entirely with a trick. Thus we concentrate on the crucial step. The proof of this step connects our theory to number theory. This connection is just the tip of an iceberg and leads to powerful results in number theory.

Beginning of Formal Proof: Suppose $f(z)$ is holomorphic on the upper half plane and invariant under the modular group. Then it is invariant under $z \to z+1$ and thus periodic with period 1.

So we can write $f(z) = g(e^{2\pi iz})$ where g is holomorphic on $0 < |z| < 1$. It is traditional to write $q = e^{2\pi iz}$. Here q takes values in the unit disk and can be regarded as the independent variable describing these points. Since g is holomorphic on the punctured disk, it can be expanded in a Laurent series $\sum_{k\in Z} c_k q^k$.

Recall that

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

Taking the logarithmetic derivative,

$$\pi \frac{\cos \pi z}{\sin \pi z} = \frac{1}{z} + \sum \frac{\frac{-2z}{n^2}}{\frac{n^2-z^2}{n^2}} = \frac{1}{z} + \sum \frac{2z}{z^2 - n^2} = \frac{1}{z} + \sum \left(\frac{1}{z-n} + \frac{1}{z+n} \right)$$

It is easy to see that the function on the left has period 1, so it has the form $g(e^{2\pi iz})$. Indeed,

$$\pi \frac{\cos \pi z}{\sin \pi z} = \pi \frac{\frac{e^{\pi iz} + e^{-\pi iz}}{2}}{\frac{e^{\pi iz} - e^{-\pi iz}}{2i}} = \pi i \, \frac{e^{2\pi iz} + 1}{e^{2\pi iz} - 1} = \pi i \frac{q+1}{q-1}$$

so

$$g(q) = \pi i \frac{q+1}{q-1} = -\pi i \, (1+q) \frac{1}{1-q} = -\pi i (1+q) \sum_{m=0}^{\infty} q^m = -\pi i - 2\pi i \sum_{m=1}^{\infty} q^m$$

Differentiating $g(e^{2\pi iz})$ with respedt to z gives $g'(e^{2\pi iz})(2\pi i)e^{2\pi iz} = (2\pi i)q \, g'(q)$. The straightforward derivative with respect to z is:

$$-\frac{1}{z^2} - \sum_{n>0} \left(\frac{1}{(z-n)^2} + \frac{1}{(z+n)^2} \right) = -\sum_{n\in Z} \frac{1}{(z-n)^2}$$

The corresponding operation on the q formula is

$$\frac{d}{dz} g(q(z)) = -(2\pi i)^2 \sum_{m=1}^{\infty} m q^m$$

In short

$$-\sum_{n\in Z} \frac{1}{(z-n)^2} = -(2\pi i)^2 \sum_{m>0} m q^m$$

This formula implies that we should substitute $q = e^{2\pi iz}$ in the end. Differentiating after this is easy and we find that

$$(-1)^{k-1}(k-1)! \sum_{n\in Z} \frac{1}{(z-n)^k} = -(2\pi i)^k \sum_{m>0} m^{k-1} q^m$$

We now apply this formula to $G_{2k}(\tau) = \sum_{\omega\neq 0} \frac{1}{(m\tau+n)^{2k}}$. This function is periodic in τ with period 1, so everything we have done applies when we set $z = m\tau$. However, $m = 0$ will be troublesome, so we sum this term first. It will be

$$\sum_{n\neq 0} \frac{1}{n^{2k}} = 2\sum_{n=1}^{\infty} \frac{1}{n^{2k}} = 2\zeta(2k)$$

Therefore

$$G_{2k}(\tau) = 2\zeta(2k) + 2\sum_{m=1}^{\infty}\sum_{n\in Z} \frac{1}{(m\tau - n)^{2k}}$$

At this point we want to evaluate the inner sum by the previous formula, where $z = m\tau$. But that previous formula was a sum over m. To avoid trouble, we change the m in the previous formula to j getting

$$G_{2k}(\tau) = 2\zeta(2k) - 2\sum_{m=1}^{\infty} \frac{(2\pi i)^{2k}}{(-1)^{2k-1}(2k-1)!}\sum_{j>0} j^{2k-1}q^j$$

In this formula, we are supposed to set $q = e^{2\pi i m\tau} = \left(e^{2\pi i\tau}\right)^m$. We want to use $q = e^{2\pi i\tau}$ in the final formula, so we write this as or

$$G_{2k}(\tau) = 2\zeta(2k) + 2\sum_{m=1}^{\infty} \frac{(2\pi i)^{2k}}{(2k-1)!}\sum_{j>0} j^{2k-1}q^{jm} = 2\zeta(2k) + 2\frac{(2\pi i)^{2k}}{(2k-1)!}\sum_{m=1}^{\infty}\sum_{j>0} j^{2k-1}q^{jm}$$

We now define a number theoretic function $\sigma_k(n) = \sum_{d|n} d^k$. So $\sigma_1(6) = 1 + 2 + 3 + 6 = 12$ and $\sigma_2(6) = 1^2 + 2^3 + 3^2 + 6^2 = 50$. A little thought shows that

$$G_{2k}(\tau) = 2\zeta(2k) + 2\frac{(2\pi i)^{2k}}{(2k-1)!}\sum_{m=1}^{\infty} \sigma_{2k-1}(m)q^m$$

where $q = e^{2\pi i\tau}$.

In particular,

$$g_2 = 60G_4(\tau) = 60\left(2\zeta(4) + 2\frac{(2\pi)^4}{3!}\sum_{m=1}^{\infty}\sigma_3(m)q^m\right)$$

$$g_3 = 140G_6(\tau) = 140\left(2\zeta(6) + 2\frac{(-(2\pi)^6)}{5!}\sum_{m=1}^{\infty}\sigma_5(m)q^m\right)$$

Recall that $\zeta(4) = \frac{\pi^4}{90}$, and $\zeta(6) = \frac{\pi^6}{945}$. A brief calculation gives

$$g_2 = \frac{4}{3}\pi^4\left(1 + 240\sum_{m=1}^{\infty}\sigma_3(m)q^m\right)$$

$$g_3 = \frac{8}{27}\pi^6\left(1 - 504\sum_{m=1}^{\infty}\sigma_5(m)q^m\right)$$

Now notice that the outside factor of g_2^3 is

$$\left(\frac{4}{3}\right)^3\pi^{12}$$

and the outside factor of $27g_3^2$ is

$$27\left(\frac{8}{27}\right)^2\pi^{12} = \left(\frac{4}{3}\right)^3\pi^{12}$$

so when we form the modular function, these constants cancel and we get

$$j(\tau) = \frac{\left(1 + 240\sum_{m=1}^{\infty}\sigma_3(m)q^m\right)^3}{\left(1 + 240\sum_{m=1}^{\infty}\sigma_3(m)q^m\right)^3 - \left(1 - 504\sum_{m=1}^{\infty}\sigma_5(m)q^m\right)^2}$$

$$= \frac{(1 + 240q + ...)^3}{(1 + 240q + ...)^3 - (1 - 504q + ...)^2} = \frac{1 + 3\cdot 240q + ...}{(1 + 3\cdot 240q + ...) - (1 - 2\cdot 504q + ...)} =$$

$$\frac{1 + 720q + ...}{1728q + ...} = \frac{1}{1728}\frac{1}{q} + ...$$

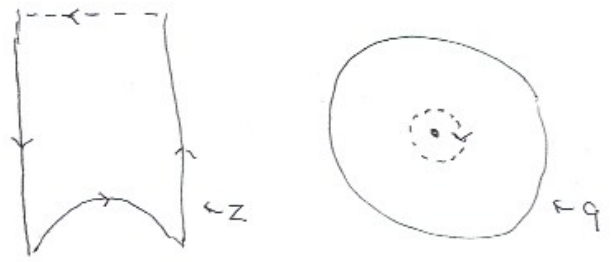
This explains the strange factor 1728 used by some authors.

We claim that this formula shows that the modular function has a pole of order 1 at ∞. The portion of the modular picture higher than 1 consists of equivalent vertical strips; the fundamental region contains one of these strips, with $-\frac{1}{2} \le \Re(z) < \frac{1}{2}$. The periodic function $q = e^{2\pi i\tau}$ maps each of these strips to the punctured disk $0 < |z| < e^{-2\pi}$, where this last bound is provided so we only consider the portion of the strips above $\Im(z) = 1$. It follows that we can think of q as giving coordinates on the fundamental region for all points $\Im(z) > 1$. The missing point $q = 0$ corresponds to the point at ∞. Consequently our expansion of j in terms of q shows that j has a pole of order one at ∞.

End of the Formal Proof: We now supply a trick to complete the proof. Recall the argument principle, which states that the following integral computes the sum of the zeros and poles inside γ, with multiplicities which are positive for zeros and negative for poles.

$$\frac{1}{2\pi i}\int_\gamma \frac{f'(\zeta)}{f(\zeta)}d\zeta$$

Apply this to $j(z)$ and the following path. The integral around the solid lines cancels out and we are left with the integral across the dotted line. Apply the map $\tau \to e^{2\pi i\tau}$ to map the upper portion of the fundamental region to a disk and the dotted path to a dotted path in this disk as pictured below.



The integral in the half plane counts zeros of $j(\tau)$ including multiplicities. There are no poles. If the dotted path is high enough, there will be no zeros above it because according to the second picture these values are very large. The second picture counts poles, but we must consider how the change of variable affects the integral. Our dotted path in the upper half plane is $\gamma(t) = -t + ih$ for $-\frac{1}{2} \le t \le \frac{1}{2}$. The corresponding curve in the disk is $e^{2\pi i(-t+ih)} = e^{-2\pi h}e^{-2\pi it}$. The first term is an irrelevant constant, while the second term is a clockwise path around the origin.

We are integrating $\frac{f(z)}{f'(z)}$. If we express z in terms of q as $z(q)$, then $f(z)$ becomes $f(z(q))$ and the derivative of this with respect to q is $f'(z(q))\frac{dz}{dq}$. However integration requires that we also transform dz to $\frac{dz}{dq}dq$. So the entire integrand changes from

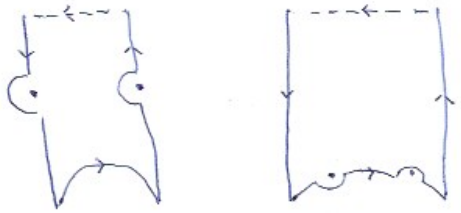
$$\frac{f(z)}{f'(z)}dz$$

to

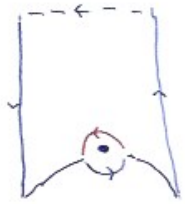
$$\frac{f(z(q))}{f'(z(q))\frac{dz}{dq}}\frac{dz}{dq}\,dq = \frac{f(z(q))}{\frac{d}{dq}f(z(q))}dq$$

This is miraculous and tells us that the integral in the q disk still counts zeros and poles, but with a change of sign because the direction of the path is reversed. Usually zeros are counted positively and poles are counted negatively, but our integral counts zeros negatively and poles positively. There are no zeros in the disk and one pole of order one. So the dotted integral is 1. Therefore the total integral in the half plane is 1. This half plane integral is only counting zeros, and therefore our function has one zero.

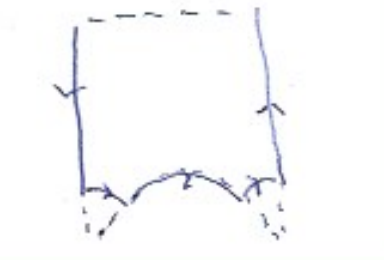
We aren't quite done. There is a problem if $j(\tau)$ is zero on the boundary. We can fix this for most points by modifying the path slightly. The picture on the left shows this modification when a zero is on a vertical boundary, and the picture on the right shows this modification when a zero is on the circle at the bottom.



This trick fails at the points i and $-\frac{1}{2}+\frac{\sqrt{3}}{2}i$. The picture below indicates a different modification at i. We change our original path by using the bottom half of the small circle around i. The main pieces of the integral cancel as before. But the dip below i has nothing to cancel against because the modular transformation which fixes i maps it to the portion above i. Integrals over both of these pieces are equal, and the total integral over the full circle is the order of the zero of j at i. So the integral over just the bottom half of the circle is $\frac{1}{2}$ (order of zero at i). Our modified path counts zeros of j in the fundamental region, but it only counts half of the zero at i.



A slightly more complicated thing happens at $-\frac{1}{2}+\frac{\sqrt{3}}{2}i$. The modified path is shown in the illustration below. We avoid both special points at the left and right corners of the fundamental region. The integrals along this path cancel except the two arcs at the special points and the dotted line at the top.



The picture below shows the special point on the left corner of the fundamental region. The new arc we introduced in this corner is shown, together with an arc to its left, which is the new arc we introduced at the bottom right corner of the fundamental region, but moved by $z \to z-1$ to this spot.

There are six arcs in the six angles around the special point, half equivalent to the arc we added on the left side of the fundamental region and half equivalent to the arc we added on the right side. Therefore integrating around our two arcs equals $\frac{1}{3}$ of the total integral over a small circle surrounding $-\frac{1}{2}+\frac{\sqrt{3}}{2}i$.

But this circle runs clockwise rather than counterclockwise, so the integral around our two additional arcs in the fundamental region equals

$$-\frac{1}{3}\left(\text{order of zero of } j \text{ at } -\frac{1}{2}+\frac{\sqrt{3}}{2}i\right)$$

Our final result is

$$\text{(sum of zeros of } j \text{ strictly inside the fundamental region)} =$$

$$-\frac{1}{3}\left(\text{order of zero of } j \text{ at } -\tfrac{1}{2}+\tfrac{\sqrt{3}}{2}i\right) + \text{(integral along dotted line)}$$

We already know that the integral along the dotted line is one because there is a pole of order one at infinity. So the above result can be rewritten

$$\text{(number of zeros inside)} + \frac{1}{3}\,\text{(zeros at bottom left)} = 1$$

We can make these various path modifications simultaneously to handle several zeros on the boundary. We can also apply the argument to $j(z)-\lambda$ so we are counting points where $f=\lambda$. Call i and $-1/2+i\sqrt{3}/2$ the *special boundary points*. Then for the function $j(z)$ and an arbitrary constant λ we have proved the following:

$$\begin{gathered}\text{(number of times } f=\lambda \text{ at a regular point in the fundamental region)} + \\ \frac{1}{2}\,\text{(number of times } f(i)=\lambda) + \\ \frac{1}{3}\left(\text{number of times } f\left(-\tfrac{1}{2}+\tfrac{\sqrt{3}}{2}i\right)=\lambda\right) \\ = 1\end{gathered}$$

From this it is easy to complete the proof. If $j(z)=\lambda$ for a regular point, then we have already accounted for the final 1, and therefore f can only equal λ at that point and with multiplicity one. Otherwise suppose $f(i)=\lambda$ with multiplicity m, where $0\le m$ is an integer, and let $f(-1/2+i\sqrt{3}/2)=\lambda$ with multiplicity n, where $0\le n$ is an integer. Then $\frac{m}{2}+\frac{n}{3}=1$, and this equation has only two solutions: $m=2$ and $n=0$, or $m=0$ and $n=3$. So the value λ is taken at exactly one point. QED.

Corollary 33. *If g_2 and g_3 are arbitrary complex numbers such that $g_2^3-27g_3^2\neq 0$, then there is a lattice with basis ω_1,ω_2 generating these g_2,g_3.*

Proof: By the theorem can find a lattice with basis $\tau,1$ generating $\widehat{g_2}$ and $\widehat{g_3}$ with

$$\frac{g_2^3}{g_2^3-27g_3^2}=\frac{\widehat{g}_2^3}{\widehat{g}_2^3-27\widehat{g}_3^2}$$

If we replace $\{\tau,1\}$ by $\{\lambda\tau,\lambda\}$, $\widehat{g_2}$ is multiplied by λ^4, $\widehat{g_3}$ is multiplies by λ^6, and the denominator $g_2^3-27g_3^2$ is multiplied by λ^{12}. Select a λ making the denominators of the two invariants equal.

Then the numerators will also be equal and so g_2 and $\widehat{g_2}$ differ by a cube root of unity. Multiply λ by this cube root; this will not change λ^{12} or λ^6, but λ^4 will be multiplied by this cube root. making $g_2 = \widehat{g}_2$. After this change, the denominators are equal and have the same g_2, so their $g_3$can only different by a sign. Replace λ^2 by $-\lambda^2$. This will not change λ^{12} or λ^4, but λ^6 will change sign. QED.

Remark: We earlier discovered that two tori have larger automorphism groups than usual. These tori correspond to the two exceptional points in the modular group's fundamental region. The square lattice is given by $\tau = i$ and the hexagonal lattice is given by $\tau = \omega = -\frac{1}{2} + \frac{\sqrt{3}}{2}i$. We easily compute that values of j at these points. Since $g_3(z)$ is a multiple of $G_6(z)$, and since $i = -\frac{1}{i}$, we have

$$g_3(i) = g_3\left(-\frac{1}{i}\right) = i^6 g_3(i) = -g_3(i)$$

and so $g_3(i) = 0$. Since $g_2^3 - 27g_3^2 \neq 0$, $g_2(i) \neq 0$ and

$$j(i) = \frac{g_2^3}{g_2^3 - 27g_3^2} = 1$$

Similarly the map $-\frac{1}{z} - 1 = \frac{-z-1}{z}$ fixes $\omega = -\frac{1}{2} + \frac{\sqrt{3}}{2}i$, so

$$g_2(\omega) = g_2(-\frac{1}{\omega} - 1) = \omega^4 g_2(\omega)$$

Since $\omega^4 \neq 1$, $g_2(\omega) = 0$ and

$$j(\omega) = \frac{g_2^3}{g_2^3 - 27g_3^2} = 0$$

Remark: We obtained the initial coefficient of the expansion of $j(\tau)$, where $q = e^{2\pi i\tau}$. If we multiply j by 1728, this expansion continues

$$j(\tau) = \frac{1}{q} + 196884q + 21493760q^2 + 864299970q^3 + 20245856256q^4 + ...$$

During the twentieth century, the finite simple groups were completely classified. It turned out that the list contains the alternating groups, a number of infinite families of groups related to Lie groups, and 23 exceptional groups. The largest of these exceptional groups is called the Monster group. In 1978, John McKay discovered that the coefficients of the expansion of j are related to the dimensions of the irreducible representations of the monster group. These dimensions are $1, 196883, 21296876, ...$ which we denote $r_1, r_2, r_3, ...$. The coefficients of the j expansion are $1 = r_1$, $196884 = r_1 + r_2$, $21493760 = r_1 + r_2 + r_3$, $20245856256 = 2r_1 + 2r_2 + r_3 + r_4$ and so forth. This led to a large investigation often called Monster Moonshine, until a precise conjecture was made and finally proved in 1992 by Borcherds.

24.7 Generating the Modular Group

Theorem 172. *The modular group is generated by the maps* $z \to z+1$ *and* $z \to -\frac{1}{z}$.

Proof: We know that the only points in the fundamental region left fixed by a non-identity modular transformation are i and $-\frac{1}{2} + \frac{\sqrt{3}}{2}i$. It follows that all points in the upper half plane except points equivalent to these points can be mapped into the fundamental region by only one modular transformation. Thus to prove the theorem we need only show that every point can be mapped to the fundamental region by applying the two indicated maps several times.

Start with an arbitrary point in the upper half plane. Translate the point until it is in the vertical strip $-\frac{1}{2} \leq \Re(z) < \frac{1}{2}$. If the absolute value of the resulting point is greater than one, we are done. If the absolute value is equal to one, we are done if it is on the left side and otherwise $-\frac{1}{z}$ finishes the job. If the absolute value is less than one, applying $-\frac{1}{z}$ leads to a higher point, and we can start the game all over. But we proved that once a point is at a certain height, there are only a finite number of even larger heights it can reach. QED.

Remark: Because of this result, if $f(z)$ is a periodic function on the upper half plane and we are interested in invariance or almost invariance under the modular group, it suffices to find an identity relating $f\left(-\frac{1}{z}\right)$ to $f(z)$. In mathematics there are many celebrated such identities. Whenever one appears, the modular group is not far away. Note that

$$G_{2k}\left(-\frac{1}{z}\right) = z^{2k} G_{2k}(z)$$

24.8 Classification of Modular Functions

Definition 56. *A* modular function *is a meromorphic function* f *on the upper half plane satisfying the equations* $f(z+1) = f(z)$ *and* $f\left(-\frac{1}{z}\right) = f(z)$ *and such that the point at infinity is not an essential singularity.*

Remark: The function is periodic, so it can be written as a function of $q = e^{2\pi i z}$. This function maps the upper half plane to the disk $0 < |q| < 1$. We are requiring that $q = 0$ be a removable singularity or pole, rather than an essential singularity. Another way to say this is that f must be meromorphic on the fundamental region with ∞ added.

It follows from the previous section that a modular function is invariant under the action of $SL(2, Z)$, and consequently assigns to each torus a complex number which is a conformal invariant.

Theorem 173. *Every modular function is a rational function in* $j(z)$.

Proof: Suppose $f(z)$ is modular and not constant. This f has only finitely many poles in the fundamental region, because otherwise these poles would have to converge to ∞ and so f in the q-disk would have an essential singularity at the origin.

Suppose f has a pole of order k at z_0. Let $j(z_0) = c_0$. Then $j(z) - c_0$ has a simple zero at z_0 and $f(z)\left(j(z) - c_0\right)^k$ has no pole at z_0 and thus one less pole in the fundamental region. Continuing, we eventually find a polynomial P such that $f(z)P(j(z))$ has no poles in the fundamental region.

Consider the Laurent expansion of $f(z)P(j(z))$ at the point at infinity, which has the form

$$\frac{a_k}{q^k} + ... + \frac{a_1}{q} + a_0 + ...$$

Since $j(q) = \frac{1}{q} + ...$, we an clearly find a linear combination of powers of j such that the Laurent expansion of this linear combination has exactly the same negative and constant terms. This linear combination is a polynomial expression $Q(j(z))$. Then $f(z)P(j(z)) - Q(j(z))$ vanishes at ∞ and has no poles. Thus it is constant, and so zero. In short, $f(z) = \frac{Q(j(z))}{P(j(z))}$. QED.

Remark: Suppose $f(z)$ is holomorphic on the upper half plane and satisfies $f(z+1) = f(z)$ and $f\left(-\frac{1}{z}\right) = f(z)$. Then f attaches a complex number to each lattice, and it attaches the same invariant to conformally equivalent tori. If f is one-to-one on the fundamental region, then conversely f distinguishes between tori that are not conformally equivalent.

We claim that such a function cannot have an essential singularity at ∞, and therefore the previous theorem applies. Indeed, pick an arbitrary finite point z_0 where $f'(z_0) \neq 0$ and let $f(z_0) = c_0$. Then f is a local homeomorphism near c_0 and thus there is a small disk about c_0 such that the values in this disk are only taken near z_0, and thus never taken near ∞. So there is a disk about $q = 0$ in the q-disk such that $f(z)$ is never within ϵ of c_0 in this disk. It follows that $\frac{1}{f(z)-c_0}$ is bounded in the disk and thus has a removable singularity at the origin, so $f(z) - c_0$ has at most a pole, and the same is true of $f(z)$.

It is easy to see that the only rational functions of j with no poles in the upper half plane are polynomials in $j(t)$, and the only such polynomials which are one-to-one are $aj(z) + b$ for constants a and b. We have proved

Theorem 174. *Let $f(z)$ be a holomorphic function on the upper half plane such that $f(\tau_1) = f(\tau_2)$ if and only if the torus associated with τ_1 is conformally equivalent to the torus associated with τ_2. Then $f(\tau) = aj(\tau) + b$ for constants a and b.*

Remark: Such a function must satisfy $f(z+1) = f(z)$, and thus must have the form $g(e^{2\pi iz}) = g(q)$. The function g must have a Laurent expansion in q of the form

$$a\Big(\frac{1}{q} + b + 196884q + 21493760q^2 + 864299970q^3 + 20245856256q^4 + ...\Big)$$

The mathematician Richard E. Borcherds won the Fields Medal for proving the Monster Moonshine conjecture about the connection between these numbers and the largest exceptional finite simple group. He teaches at Cambridge and Berkeley and has many UTube lectures on a variety of topics. These videos are highly recommended because Borcherds explains the ideas behind theorems rather than magically obtaining results by calculation.

In a wonderful video on the modular function, Borcherds points out that it's defining equations are very simple: $f(z+1) = f(z)$ and $f\left(-\frac{1}{z}\right) = f(z)$ and yet very mysteriously it leads to the very large integers

$$196884, 21493760, 864299970, 20245856256, ...$$

Where do they come from?

Chapter 25

Consequences of the Modular Theory

The theory of conformal equivalences of tori flows in a straightforward way until the very end. Then suddenly, the crucial theorem that $\frac{g_2^3}{g_2^3 - 27g_3^2}$ completely classifies tori up to equivalence is proved by studying the one point that does *not* correspond to a torus, and expanding the function in a Laurent series about this point. That expansion leads from our earlier product formulas for $\sin z$ to series whose coefficients are number theoretic formulas, and then series associated with the Monster group.

This chapter is a miscellaneous collection of such unexpected connections. The chapter will not be entirely satisfying because we won't have time to develop the full implications of these connections. Each topic will introduce one new idea, and a full theory would then put all of these ideas together to build a more complete edifice. Number theory is lurking in the shadows as coefficients of series expansions, and in the following chapter we will finally reveal some of that theory.

25.1 Picard's Theorem

The modern theory of complex analysis is due to Cauchy and his followers in the first half of the 1800's. By 1850 it was essentially complete. Imagine the surprise, then, when Charles-Emile Picard discovered a previously unknown theorem from this elementary theory in 1879.

The following theorem had been well known:

Theorem 175. *If $f(z)$ is a non-constant entire function, then the range of f is dense in C.*

This has an easy proof. If an open neighborhood of a point z_0 is not in the range, then $\frac{1}{f(z)-z_0}$ is bounded, and thus constant.

Entire functions are natural generalizations of polynomials. According to the fundamental theorem of calculus, every non-constant polynomial has a complex root and thus every non-constant polynomial takes every complex value. But this more powerful theorem does not generalize to entire functions because e^z is never zero.

Theorem 176 (Picard, 1879)**.** *The range of a non-constant entire function can omit at most one point.*

Proof: If $f(z)$ omits two arbitrary values, we can replace f by $af(z) + b$ and make the two omitted values $0 = j\left(-\frac{1}{2} + \frac{\sqrt{3}}{2}i\right)$ and $1 = j(i)$ Let $\mathcal{U}$ be the open set obtained by removing from the upper half plane the points i and $-\frac{1}{2} + \frac{\sqrt{3}}{2}i$ and all points equivalent to them under the modular group. Then a glance at the modular picture shows that $j : \mathcal{U} \to C - \{0, 1\}$ is a covering space. We have the following diagram:

$$\begin{array}{ccc} & & \mathcal{U} \\ & & \big\downarrow j \\ C & \xrightarrow{f} & C - \{0, 1\} \end{array}$$

Since C is simply connected, we can lift f to a map $\tilde{f} : C \to \mathcal{U} \subset H$. Composing with $e^{2\pi iz}$ gives a map from C to the unit disk D:

$$e^{2\pi iz} \circ \tilde{f} : C \to D$$

Since D is bounded, this map is constant and thus $\tilde{f}$ is constant and so f is constant. QED.

25.2 Modular Forms for $SL(2, Z)$

Changing variables have been part of calculus from the very beginning. If $f(w)$ is a function and we later write w in terms of z, then $f(w(z))$ becomes a function of z. But when we differentiate or integrate, a mysterious extra term appears in the transformation formula:

$$\frac{d}{dw}f(w) \to \frac{d}{dz}f(w(z)) = \frac{df}{dw}(w(z))\,\frac{dw}{dz}$$

$$\int f(w)\,dw \to \int f(w(z))\,\frac{dw}{dz}\,dz$$

Leibniz' calculus notation is used because it suggests these extra terms and helps us remember them.

When we encountered the modular group, we often made the transformation $w = \frac{az+b}{cz+d}$. The extra term is then

$$\frac{d}{dz}\left(\frac{az+b}{cz+b}\right) = \frac{(cz+d)a-(az+b)c}{(cz+d)^2} = \frac{1}{(cz+d)^2}$$

This extra term then appeared in the Eisenstein series $G_{2k}(\tau)$, which satisfy the formula

$$G_{2k}\left(\frac{a\tau+b}{c\tau+d}\right) = (c\tau+d)^{2k}G_{2k}(\tau)$$

In this section we will try to understand the significance of these Eisenstein series, paying close attention to that extra term. Since the modular group is generated by $z \to z+1$ and $z \to -\frac{1}{z}$ with extra terms 1 and $\frac{1}{z^2}$, we will pay particular attention to those generating cases.

Definition 57. *A* modular form of level $2k$ *is a holomorphic function on the upper half plane which satisfies the two equations*

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

and

$$f\left(-\frac{1}{z}\right) = z^{2k}f(z)$$

and has a limiting value at infinity. In other words, when written as $g(q) = g(e^{2\pi iz})$, g has a removable singularity at $q = 0$. If this value is zero, we call f a cusp form.

Remark: Since $z \to z+1$ and $z \to -\frac{1}{z}$ generate the modular group, modular forms are almost invariant under all modular transformations. It is easy to see that

$$f\left(\frac{az+b}{cz+d}\right) = (cz+d)^{2k}f(z)$$

In the advanced theory, it is useful to think of modular forms as elements $f(z)(dz)^k$, and then they are completely invariant under the modular group. We never speak of modular forms of odd level.

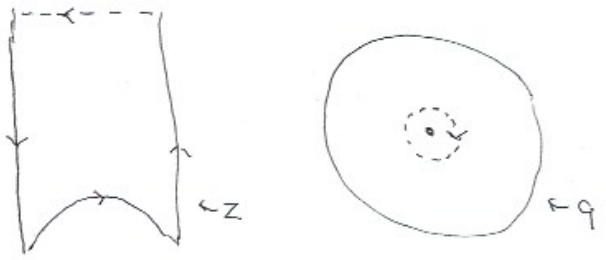
Note that a modular form of level zero is just a modular function. However, every non-trivial modular function has poles, so the only modular forms of level zero are constants.

Theorem 177. *Every modular form of level k has exactly $\frac{k}{12}$ zeros, if zeros are properly counted. The order of a zero is the standard order except at i and $\omega = -\frac{1}{2}+\frac{\sqrt{3}}{2}$. The order at i is half of the standard order, and the order at ω is one third of the standard order.*

Proof: We compute the integral

$$\frac{1}{2\pi i}\int_\gamma \frac{f'(z)}{f(z)}\,dz$$

over the following path, as we did when studying the j function. The argument is basically the same.



Suppose first that none of the zeros is on the boundary of the fundamental region, and that the dotted path is so high that no zeros are above it, except possibly at infinity. Then the integral gives the number of zeros inside the fundamental region. The integral over the dotted line at the top is equivalent to the integral over the dotted circle in the disk and thus counts the zero or pole at infinity, but negatively because the circle is clockwise. For modular forms it gives the negative of the number of zeros at infinity. The two vertical portions of the integral cancel, as they did earlier.

The bottom integral is trickier because of the factor z^k in the modular relation. The bottom can be broken into two paths, the circular path from the lower left to the center point i and the circular path from the lower right corner to i. The map $w = -\frac{1}{z}$ maps the left path to the right path, but the right path is then traversed in the opposite direction.

Recall the change of variable formula. We begin with two open sets $\mathcal{U}$ and $\mathcal{V}$ and a one-to-one and onto holomorphic map $w(z)$ from the first to the second. We think of the independent variable in the first as z and the independent variable in the second as w. Whenever $g(w)$ is a function on $\mathcal{V}$, $g(w(z))$ is a function on $\mathcal{U}$. If γ is a path in $\mathcal{U}$, $w\circ\gamma$ is a path in $\mathcal{V}$. Then

$$\int_\gamma g(w(z))\frac{dw}{dz}\,dz = \int_{w\circ\gamma} g(w)\,dw$$

Therefore when changing coordinates, functions transform as expected via $g(w) \to g(w(z))$, but integrands have a pesky extra term $\frac{dw}{dz}$. Leibniz invented a wonderful notation which helps us remember this term. In his language, the integrand is a differential written $g(w)dw$. When we change coordinates, we have to remember that dw also changes. So the differential becomes $g(w(z))\left(\frac{dw}{dz}\,dz\right)$.

Derivatives also have an extra term. We expect the function $\frac{d}{dw}g(w)$ to become $\frac{d}{dz}g(w(z))$, but in fact $\frac{d}{dz}g(w(z))$ equals $\frac{d}{dw}g(w)\,\frac{dw}{dz}$.

However, the argument principle we are currently using involves an integrand which contains a derivative, and this differential transforms with no extra terms. Indeed

$$\frac{\frac{d}{dw}f(w)}{f(w)}\,dw = \frac{\frac{d}{dw}f(w)}{f(w)}\,\frac{dw}{dz}\,dz = \frac{\frac{d}{dw}f(w(z))\frac{dw}{dz}}{f(w)}\,dz = \frac{\frac{d}{dz}f(w(z))}{f(w(z)}\,dz$$

In our special case, $w(z) = -\frac{1}{z}$ and $f(w(z)) = z^k f(z)$. Therefore

$$\frac{\frac{d}{dz}f(w(z))}{f(w(z)}\,dz = \frac{\frac{d}{dz}z^k f(z)}{z^k f(z)}\,dz = \frac{k}{z} + \frac{f'(z)}{f(z)}\,dz = \frac{f'(w)}{f(w)}\,dw$$

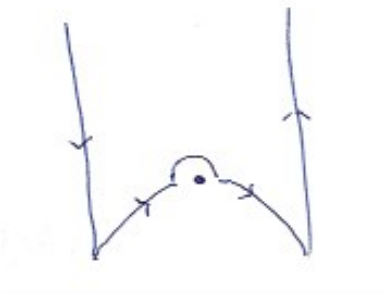
It follows that the integral of $\frac{k}{z} + \frac{f'(z)}{f(z)}$ over the left half of the circle equals the integral of $\frac{f'(w)}{f(w)}$ over the right half of the circle, but traced backward. The integral of $\frac{f'(z)}{f(z)}\,dz$ over the entire circle is thus the negative of the integral of $\frac{k}{z}$ over the left half of this circle. So the total integral of $\frac{1}{2\pi i}\,\frac{f'}{f}$ over the bottom of the fundamental region equals

$$-\frac{1}{2\pi i}\int_\gamma \frac{k}{z}\,dz$$

over the left half of the bottom boundary of the fundamental region. This path is exactly $\frac{1}{12}$ of the complete circle of radius 1 centered at the origin, and the integral moves clockwise rather than counterclockwise, so the value of the integral is $\frac{k}{12}$.

Clearly we have just proved that if f is a non-constant modular form of level k and f has no zeros on the boundary of the fundamental region, then the total number of zeros of f inside the region and at infinity is $\frac{k}{12}$.

We now have to consider zeros on the boundary. Our previous trick of modifying the path to handle zeros on the vertical boundary still works. Handling zeros on the circular boundary at the bottom is tricky, but everything will be clear after we consider a possible zero at i. In that case, we will modify the boundary path as follows:



Ignoring the bump in the center and the factor of $\frac{1}{2\pi i}$ in the integral, on the left we integrate $\frac{k}{z} + \frac{f'(z)}{f(z)}$ and on the right we integrate $-\frac{f'(w)}{f(w)}$; the two major pieces cancel and we are left with the integral of $\frac{k}{z}$. In the limit as the bump goes to zero, this integral becomes $\frac{k}{12}$. The limiting process is a little tricky because if f has a zero at i then $\frac{f'}{f}$ has a pole at i. So the integral over the entire left half of the circle does not converge. But luckily the two pieces cancel out until the bitter end.

Next we must evaluate the small bump integral, and find its limit as the bump gets smaller and smaller. If f has a zero at i, it can be written $(z-i)^m h(z)$ near i, where h is holomorphic and nonzero. So $\frac{f'}{f} = \frac{m}{z-i} + \frac{h'(z)}{h(z)}$. The integral of $\frac{m}{z-i}$ over the bump is almost $\frac{1}{2}$ of the integral around the complete circle, which is $-\frac{m}{2}$ because our curve is clockwise and there is a surpressed $\frac{1}{2\pi i}$. In the limit as the bump decreases to zero, this value becomes exactly $-\frac{m}{2}$. The term $\frac{h'}{h}$ is holomorphic at i and the integral of it goes to zero as the bump shrinks. So integration around our modified path shows that the sum of the zeros inside the fundamental region except for the zero at i is

$$\frac{k}{12} - \frac{m}{2}$$

and thus the sum of all of the zeros in the fundamental region is $\frac{k}{12}$.

When there is a zero on the curved boundary but not at the two exceptional points, then there will be two zeros, one on the left that is in the fundamental region, and one on the right that is just outside. In that case we modify our path to exclude both zeros with two bumps like the single bump pictured at i. The argument is the same as for i and we must take a limit as the bumps decrease to points. The map $z \to -\frac{1}{z}$ moves the bump on the right to the lower half of the bump on the left and we get a single path around our zero. So in the limit, this zero fully counts.

The argument at $\frac{1}{2} - \frac{\sqrt{3}}{2}i$ is left to the reader but sketched briefly here. We make small partial arcs at the two corners to avoid these two zeros. By periodicity, integration over these two arcs

is like one third of the integration clockwise around the exceptional point. In the limit, this becomes one third of the integral of $\frac{m}{z-\omega}$ around this exceptional point. QED.

25.3 Dimensions of Spaces of Modular Forms

Recall that $g_2 = 60G_4$ and $g_3 = 140G_6$. So the fundamental Eisenstein series for elliptic functions are G_4 and G_6.

The set of all modular forms of weight $2k$ clearly forms a complex vector space. It turns out that each of these spaces is finite dimensional. We have

Theorem 178. *The dimension of the space of modular forms for* $2k \leq 12$ *is given by the following table:*

2k	dimension	generators
0	1	1
2	0	
4	1	G_4
6	1	G_6
8	1	G_4^2
10	1	G_4G_6
12	2	G_4^3, G_6^2

Table 25.1: Dimension of Space M_k of Forms of Level k

Proof: A form of level 2 would have $2/12 = 1/6$ zeros, so it only has zeros at special points, but each contributes too much. A form of level 4 would have $1/3$ zeros. It could only have zeros at $\omega = -\frac{1}{2} - \frac{\sqrt{3}}{2}i$. Any two such forms would have zeros of the same order and thus differ by a multiplicative constant. Since G_4 has such a zero, it generates the space. A form of level 6 would have $1/2$ zeros, and such a form could have a zero only at i. Any two would have a zero of the same order and thus differ by a multiplicative constant. Since G_6 has level 6, it generates.

A form of level 8 would have $2/3$ zeros. This can only happen if it has a double zero at ω, as does G_4^2. A form of level 10 would have $10/12 = 5/6$ zeros. This can only happen if it has first order zeros at i and ω, as does G_4G_6.

Finally, a form of level 12 would have 1 zero. This can happen if it has one ordinary zero at any non-special point, or a zero of order 2 at i, or a zero of order 3 at ω. These two special cases give us G_4^3 and G_6^2 and these are certainly linearly independent because their zeros occur at different places. We will prove they generate. Let $\Delta = g_2^3 - 27g_3^2$ and notice that Δ is a linear combination of G_4^3 and G_6^2. It suffices to show that G_4^3 and Δ generate.

Let $f(z)$ be an arbitrary modular form of level 12 and write $g(z) = f(z) - \frac{f(\infty)}{G_4^3(\infty)} G_4^3(z)$. Notice that g has level 12 and has a zero at ∞. If g is identically zero, we are done. Otherwise $\frac{g(z)}{\Delta}$ has level 0 and is holomorphic everywhere, hence constant. So f is a linear expression in Δ and G_4^3 and thus in G_6^2 and G_4^3. QED.

Definition 58. *Let $\mathcal{M}_{2k}$ be the space of all modular forms of level $2k$, and let $\mathcal{M}_0(2k)$ be the subspace of cusp forms of order $2k$, that is, forms which vanish at ∞.*

Theorem 179. *Let $\Delta = g_2^3 - 27g_3^2$ and notice that multiplication by Δ gives a map from $\mathcal{M}(2k)$ to $\mathcal{M}_0(2k+12)$.*

1. *The map Δ is one-to-one and onto.*
2. $\dim \mathcal{M}(2k) = \dim \mathcal{M}_0(2k+12)$
3. $\dim \mathcal{M}(2k) = \dim \mathcal{M}_0(2k) + 1$ *for* $2k > 2$
4. $\dim \mathcal{M}(2k) + 1 = \dim \mathcal{M}(2k+12)$

Proof: Clearly multiplication by Δ is one-to-one, If $f(z)$ is a $2k+12$-level form which vanishes at infinity, then since Δ is level 12 and has a simple zero at infinity and no other zeros, $f(z)/\Delta(z)$ is a level $2k$ form, and it certainly maps to $f(z)$. This proves the first two items.

The third item implies the final one because

$$\dim \mathcal{M}(2k) + 1 = \dim \mathcal{M}_0(2k+12) + 1 = \dim \mathcal{M}(2k+12)$$

To prove the third item, it suffices to find $g(z) \in \mathcal{M}(2k)$ with $g(\infty) = 1$, because then for any f we will have

$$f(z) = [f(z) - f(\infty)g(z)] + f(\infty)g(z)$$

However, functions of the form $G_4(z)^m G_6(z)^n$ suffice for this, because G_4 is zero at one exceptional point and nowhere else, and G_6 is zero at the other exceptional point and nowhere else, and thus these products are not zero at ∞ and can be divided by a constant to make them equal one at infinity. The forms have level $4m+3n$ and these generate all even numbers higher than 2. QED.

Corollary 34. *Every modular form can be written as a polynomial in G_4 and G_6.*

Proof: The $G_4^m G_6^n$ suffice for the third item of the theorem, and $\Delta = g_2^3 - 27g_3^2$ suffices for the first item. QED.

Remark We can consider the algebra of all polynomials $C[G_4(z), G_6(z)]$. This is a graded algebra. The elements in this algebra of level $2k$ give exactly the set $\mathcal{M}_{2k}$. In particular

Theorem 180. *The monomials $G_4^m(z)\, G_6^n(z)$ of level $2k$ form a basis for $\mathcal{M}_{2k}$.*

Proof: It suffices to show that these elements are linearly independent, since we already proved that they generate. A dependence relation would be a polynomial $P(X, Y)$ with constant coefficients such that $P(G_4, G_6)$ has level $2k$ and is identically zero. We claim this would imply the existence of a nontrivial polynomial $Q(X)$ with $Q\left(\frac{G_4^3}{G_3^2}\right)$ identically zero. Assume we had that. Then the value of G_4^3/G_3^2 would always equal some root of Q and by continuity would be constant. But at ω the value of G_4^3/G_3^2 is zero and at i it is infinity, so the expression is certainly not constant.

It remains to show the existence of the polynomial Q. Suppose one monomial in the polynomial is $X^{m_1}Y^{n_1}$. Then $4m_1 + 6n_1 = N$ is the level of the term. If $X^{m_2}Y^{n_2}$ is another monomial, we have $4m^2 + 6n_2 = N$. So $4(m_2 - m_1) + 6(n_2 - n_1) = 0$. It follows that $3|(m_2 - m_1)$ and $2|(n_2 - n_1)$ and we can write $m_2 = m_1 + 3A$ and $n_2 = n_1 - 2A$ for some integer A. So

$$P(X, Y) = X^{m_1}Y^{n_1} \sum_A c_k \left(\frac{X^3}{Y^2}\right)^A$$

The integers A can be both positive and negative, but we can factor out an appropriate negative multiple of $\frac{X^3}{Y^2}$ to make the final sum a polynomial Q in $\frac{X^3}{Y^2}$. QED.

Remark: Since we already know the dimensions of the $\mathcal{M}_{2k}$, an alternate proof would count the number of monomials $G_4^{m_1}G_6^{n_1}$ with $4m_1 + 6m_2 = 2k$ and show that this number agrees with our dimension formula. This reduces the proof to a problem in elementary number theory. The greatest common divisor of 4 and 3 is 2, so there are integers m_1 and n_1 such that

$$4m_1 + 3n_1 = 2k$$

by a lemma in every number theory course ever taught. The next natural question is uniqueness of the pair (m_1, n_1) and this is answered by the argument of the previous paragraph; there are infinitely many other choices, but all have the form $m_1 + 3A, n_1 - 2A$. The final problem is counting solutions with $m_1 \geq 0$ and $n_1 \geq 0$. That is an exercise in some number theory courses, which we avoided. The answer to the exercise is given by the dimension formula.

25.4 A Theta Function

We now switch topics. Our study of the modular function $j(\tau)$ led us to the equation

$$j\left(-\frac{1}{\tau}\right) = j(\tau)$$

and the previous section illustrated the importance of similar formulas.

Another famous example was introduced by Jacobi. He used the convention that $q = e^{\pi i z}$ rather than our $e^{2\pi i z}$, so his functions are periodic with period 2 rather than 1. We will come back to the consequences of this choice in a future section. Then Jacobi defined

$$\theta(q) = \sum_{n=-\infty}^{\infty} q^{n^2} = \sum_{n=-\infty}^{\infty} e^{\pi i n^2 z}$$

Amazingly, this function satisfies a modular identity:

Theorem 181.

$$\theta\left(-\frac{1}{z}\right) = \sqrt{-iz}\,\theta(z)$$

Remark: Since $|q| < 1$, this sum converges very rapidly. Note that $\theta(z)$ is real on the imaginary axis. If $z = ix$ then $\sqrt{-iz} = \sqrt{x}$. We chose the branch of the logarithm making these square roots positive.

History: Jacobi wrote a great book about elliptic functions and modular identities, *Fundamenta nova theoriae funtionum ellipticarum*, published in 1829. This book has been translated into English and is still read today. It contains a wealth of complicated, but amazing results, many closely related to number theory.

Among other things, Jacobi was interested in computing values of elliptic functions. He had his own fundamental functions, the Jacobi elliptic functions, rather than those of Weierstrass, but the principles are the same. Since these functions have poles, series expansions may not converge rapidly near the poles. So Jacobi introduced theta functions, which were not quite doubly periodic but did have rapidly convergent series. He then obtained doubly periodic functions as quotients of these theta functions, where the extra factors replacing periodicity cancelled out.

We'll describe the flavor of these developments without giving full details. Our theta function comes from a more elaborate function introduced by Jacobi:

$$\theta(z,\tau) = \sum_{n=n-\infty}^{\infty} q^{n^2}(w^2)^n \quad \text{where} \quad w = e^{\pi i z}, q = e^{\pi i \tau}$$

Note that this function describes the z behavior of a function for fixed τ, and also the modular behavior of τ. Our fundamental identity for $\theta(z)$ becomes

$$\theta\left(\frac{z}{\tau}, \frac{-1}{\tau}\right) = \sqrt{-i\tau} e^{\left(\frac{\pi}{\tau} i z^2\right)} \theta(z, \tau)$$

If we set $z = 0$ and thus $w = 1$, we obtain the results for $\theta(z)$ described earlier.

Jacobi found an astonishing *triple product identity* for his extended function:

$$\theta(z, \tau) = \sum_{m=-\infty}^{\infty} q^{n^2} (w^2)^n = \prod_{m=1}^{\infty} (1 - q^{2m}) \left(1 + w^2 q^{2m-1}\right) \left(1 + w^{-2} q^{2m-1}\right)$$

This identity can be proved directly without complex analysis, as in Hardy and Wright's book on number theory. It has many wonderful consequences. In the special case when $z = 0$ and $w = 1$, it gives a formula for our θ:

$$\theta(z) = \sum_{m=-\infty}^{\infty} q^{n^2} = \prod_{m=1}^{\infty} (1 - q^{2m}) \left[\left(1 + q^{2m-1}\right)\right]^2$$

Proof of the Modular Identity for $\theta(z)$: We give the proof below by first proving a curious result from Fourier analysis called the Poisson formula. For essentially the first time, we are going to use a result we have not proved. This result is typically proved in undergraduate courses, and states that the Fourier series of a differentiable periodic function converges to that function. Two page proofs of this theorem are available in many books.

We will look at the case of a complex-valued function of a real variable $f(x)$ which is periodic of period 1. Define the Fourier coefficients of f by

$$c_k = \int_0^1 f(t) e^{-2\pi i k t} dt$$

The theorem we assume without proof says that the Fourier series of f converges to f, so

$$f(x) = \sum_{k=-\infty}^{\infty} c_k e^{2\pi i k x}$$

Remark: There is a corresponding theory of Fourier integrals for non-periodic functions, but it is considerably more difficult and we will not use any result from that theory except the definition of the Fourier integral transform. If $f(x)$ is any reasonable function, the Fourier integral transform is by definition

$$\hat{f}(x) = \int_{-\infty}^{\infty} f(t) e^{-2\pi i x t} \, dt$$

Theorem 182 (The Poisson Summation Formula)**.** *Let $f(x)$ be a differentiable function on R with compact support. Let $\hat{f}$ be the Fourier integral transform of f. Then*

$$\sum_{-\infty}^{\infty} f(n) = \sum_{-\infty}^{\infty} \hat{f}(n)$$

More generally,

$$\sum_{-\infty}^{\infty} f(x+n) = \sum_{-\infty}^{\infty} \hat{f}(n) e^{2\pi i n x}$$

*Remark:*We can apply this to a function which is non-zero except on $(0,1)$. In that case the expression $\sum_{-\infty}^{\infty} f(x+n)$ is just the periodic extension of f to the entire line, and the expression $\sum_{-\infty}^{\infty} \hat{f}(n) e^{2\pi i n x}$ is the Fourier series of this function. So Poisson's formula is reasonable.

Remark: When f has compact support, the sum on the left is a finite sum and the Fourier Transform certainly exists. So the proof is straightforward algebra, given below. We will apply the formula in a more general case when f has infinite support but decreases very rapidly at infinity. Technically, this means that for all integers $n > 0$ we have $\lim_{x \to \pm\infty} |x|^n f(x) = 0$. This will cause the finite sum on the left to converge, allow us to interchange sums and integrals in the proof, and cause the Fourier Transform to exist. Details are left to the reader.

Proof: Define $g(x) = \sum_{-\infty}^{\infty} f(x+n)$. Clearly g is C^∞ and periodic with period 1. We will apply the Fourier series result to $g(x)$. The Fourier coefficients for g are

$$c_k = \int_0^1 \sum f(t+n) e^{-2\pi i k t}\, dt = \sum \int_0^1 f(t+n) e^{-2\pi i k t}\, dt =$$

$$\sum \int_n^{n+1} f(u) e^{-2\pi i k (u-n)}\, du = \sum \int_n^{n+1} f(t) e^{-2\pi i k t}\, dt = \int_{-\infty}^{\infty} f(t) e^{-2\pi i k t}\, dt = \hat{f}(k)$$

By convergence of ordinary Fourier series,

$$\sum_{-\infty}^{\infty} f(x+n) = \sum c_k e^{2\pi i k x} = \sum \hat{f}(k) e^{2\pi i k x}$$

The first result follows by setting $x = 0$. QED.

Proof of result for $\theta(z)$: We now come to the main event. We will prove that $\theta\left(-\frac{1}{z}\right) = \sqrt{-iz}\,\theta(z)$ for z on the imaginary axis. The identity theorem then implies that it is true in general. Note that $\theta(ix) = \sum e^{-\pi n^2 x}$ and the identity reads

$$\theta\left(-\frac{1}{ix}\right) = \sqrt{x}\theta(ix)$$

So let us define $f(x) = \sum e^{-\pi n^2 x} = \theta(ix)$. In that case, we have to prove that $f\left(\frac{1}{x}\right) = \sqrt{x} f(x)$. Fix a particular real number t and let $g(x) = e^{-\pi t x^2}$. Then $\sum g(n) = \sum e^{-\pi t n^2} = f(t)$. This is the left side of Poisson's formula for g. We wish to apply the formula, which involves computing the Fourier transform

$$\hat{g}(u) = \int_{-\infty}^{\infty} g(x) e^{-2\pi i u x}\, dx = \int_{-\infty}^{\infty} e^{-\pi t x^2} e^{-2\pi i u x}\, dx = \int_{-\infty}^{\infty} e^{-\pi t (x^2 + \frac{2iux}{t})}\, dx =$$

$$\int_{-\infty}^{\infty} e^{-\pi t \left(x + \frac{iu}{t}\right)^2} e^{\pi t \left(\frac{iu}{t}\right)^2}\, dx = e^{\pi t \left(\frac{iu}{t}\right)^2} \int_{-\infty}^{\infty} e^{-\pi t \left(x + \frac{iu}{t}\right)^2}\, dx$$

Make the substitution $w = x + \frac{iu}{t}$ and $dw = du$ to get

$$\hat{g}(u) = e^{-\pi\, t \left(\frac{u}{t}\right)^2} \int_{-\infty}^{\infty} e^{-\pi t w^2}\, dw$$

The limits of integration should now be $-\infty + \frac{ix}{t}$ and $\infty + \frac{ix}{t}$, but this will make no difference. Make the further substitution $u = \sqrt{t} w$ and $du = \sqrt{t} dw$ to get

$$\hat{g}(u) = e^{-\pi \left(\frac{u^2}{t}\right)} \frac{1}{\sqrt{t}} \int_{-\infty}^{\infty} e^{-\pi u^2}\, du$$

The integral is now a number which we calculate using a trick of Laplace. Call the integral A and notice that

$$A^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\pi(u^2+v^2)} du dv = \int_0^{2\pi} \int_0^{\infty} e^{-\pi r^2} r dr d\theta = 1$$

so $A = 1$. Therefore $\hat{g}(u) = e^{-\pi \left(\frac{u^2}{t}\right)} \frac{1}{\sqrt{t}}$. The right side of Poisson's formula is then

$$\sum e^{-\pi n^2 \left(\frac{1}{t}\right)} \frac{1}{\sqrt{t}} = \frac{1}{\sqrt{t}} f\left(\frac{1}{t}\right)$$

QED.

Remark: Notice that $\theta^2(q) = \sum_{n_1, n_2} q^{n_1^2 + n_2^2}$ with similar formulas for higher powers. Jacobi used this fact and identities which arise in the study of θ to find a formula for the number of ways that an integer can be written as a sum of two squares, and another formula for the number of ways that an integer can be written as a sum of four squares. These formulas generalized great theorems in number theory by Fermat, Euler, and Lagrange. We will describe these proofs later.

There is another surprising use of θ. Riemann analytically continued the zeta function to the entire plane using $\theta(z)$; the functional equation of the zeta function then turned out to be a consequence of our result for $\theta\left(-\frac{1}{z}\right)$! We will explain that next.

25.5 Integral Transforms

There are three important integral transforms, the Fourier transform, the Laplace transform, and the Mellin transform. Each maps one function to another without losing information; indeed each comes with an inversion transform which recovers the original function from the transform.

Here is the Fourier transform with its inverse:

$$\hat{f}(y) = \int_{-\infty}^{\infty} e^{-2\pi i x y} f(x)\, dx$$

$$f(x) = \int_{-\infty}^{\infty} e^{2\pi i x y} \hat{f}(y)\, dy$$

Here is the Laplace transform with its inverse:

$$\hat{f}(y) = \int_{0}^{\infty} e^{-xy} f(x)\, dx$$

$$f(x) = \frac{1}{2\pi i} \lim_{T\to\infty} \int_{c-iT}^{c+iT} e^{xy} \hat{f}(y)\, dy \quad \text{for any large real } c$$

Here is the Mellin transform with its inverse:

$$\hat{f}(y) = \int_{0}^{\infty} x^{y} f(x)\, \frac{dx}{x}$$

$$f(x) = \frac{1}{2\pi i} \lim_{T\to\infty} \int_{c-iT}^{c+iT} x^{-y} \hat{f}(y)\, dy \quad \text{for any large real } c$$

Remark: All three transforms are used in complex analysis, although the Mellin transform is probably the most important for number theory. We will only use some elementary algebraic properties of the transforms. The three transforms are closely related, and in this section we talk informally about the intuition of what they are doing.

We usually apply the Fourier transform to functions defined on R and think of the transform as also defined on R. Although $f(x)$ is allowed to have complex values, we often think of f as real-valued. The inversion formula writes $f(x)$ as a generalized sum of waves $e^{2\pi i x y}$, and the Fourier transform $\hat{f}(y)$ gives the magnitude and phase of each wave. Indeed notice that for real $f(x)$ we have

$$\hat{f}(-y) = \overline{\hat{f}(y)}$$

and so

$$\hat{f}(y)e^{2\pi ixy}+\hat{f}(-y)e^{2\pi ix(-y)} = 2\Re\left(\hat{f}(y)e^{2\pi ixy}\right) = 2\Re\left(|\hat{f}(y)|e^{2\pi i\delta}e^{2\pi ixy}\right) = 2|\hat{f}(y)|\cos(2\pi xy+\delta)$$

where δ is the argument of $\hat{f}(y)$. This is a wave with magnitude $|\hat{f}(y)|$ and phase δ; as a function of x, the wave $\cos 2\pi xy$ has frequency y cycles per second.

Remark: The Laplace transform is also applied to real valued functions on the positive x-axis $f(x)$. But this time, the transformed function is important on the entire right half plane of complex numbers, because the transform formula integrates over vertical lines in this plane. When we apply the Laplace transform to $y = c + it$ to obtain these complex values, we have

$$\hat{f}(c+it) = \int_0^\infty e^{-x(c+it)}f(x)\,dx = \int_0^\infty e^{-ixt}\left(e^{-xc}f(x)\right)\,dx$$

This is essentially the Fourier transform of $f(x)$ with an added exponential decay factor, together with the assumption that $f(x) = 0$ for negative x. So apart from some constant factors, the Laplace transform of f gives the Fourier transform of f on the y-axis, and the Fourier transform of a decaying f on the vertical lines obtained by pushing this axis further right.

To make the analogy even closer, we can use the *bilateral Laplace transform* with the same integrand but integrating from $-\infty$ to ∞. However in physics and electrical engineering, the transform is usually used to analyze a current or wave with a definite beginning, and thus the Laplace transform is the more appropriate tool.

With a slight amount of additional algebra, we can convince ourselves that the inversion formula for the Laplace transform follows from the inversion formula for the Fourier integral. Here is the argument. Let us carefully separate the Laplace and Fourier transforms by reserving the notation $\hat{f}(y)$ for the Fourier transform and writing $\mathcal{L}f(y)$ for the Laplace transform. Then our previous formula says

$$\mathcal{L}f(c+it) = \int_0^\infty e^{-ixt}\left(e^{-xc}f(x)\right)\,dx = \int_0^\infty e^{-2\pi ix\frac{t}{2\pi}}\left(e^{-xc}f(x)\right)\,dx = \widehat{e^{-xc}f(x)}\left(\frac{t}{2\pi}\right)$$

The Fourier inversion formula then gives

$$e^{-cx}f(x) = \int_{-\infty}^\infty e^{2\pi ixt}\widehat{e^{-cx}f(x)}(t)\,dt = \int_{-\infty}^\infty e^{2\pi ixt}\mathcal{L}f(c+2\pi it)\,dt$$

Move e^{-cx} to the other side and substitute Substitute $u = c + 2\pi it$ to obtain

$$f(x) = \frac{1}{2\pi i}\int_{c-\infty}^{c+\infty} e^{xu}\mathcal{L}(f(u))du$$

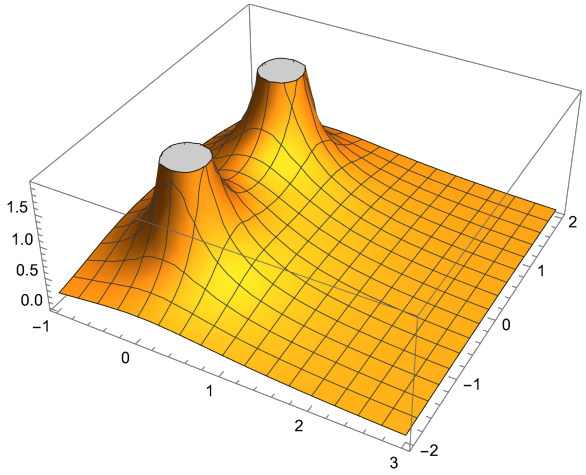
As a concrete example of all of this, let us compute the Laplace transform of $\sin x$. Integration by parts twice gives the indefinite integral of $e^{-xt}\sin x$ as

$$\frac{1}{1+t^2}\left(e^{-xt}\cos x - te^{-xt}\sin x\right)$$

These expressions vanish at infinity and give $\frac{-1}{1+t^2}$ at zero, so the Laplace transform is

$$\frac{1}{1+t^2}$$

This transform has poles at $t = \pm i$ and is well-defined on all vertical lines except the line at $c = 0$.



This function is very well behaved on vertical lines $x = c$ for $c > 0$, where it is gives the Fourier transform of $e^{-c}\sin x$, a rapidly decaying oscillation. As we get closer and closer to $c = 0$, the decay slows down and our $f(x)$ looks more and more like a pure sine wave (for positive x). The Laplace transform, which is giving the Fourier transform on this vertical line, is approaching a delta function with poles at $\pm i$.. Finally when $c = 0$, we have a pure sine wave, or in complex notation $\frac{e^{ix}-e^{-ix}}{2i}$. The Fourier transform can no longer give this wave because it is not a continuous sum of waves, but instead a sum of two isolated waves. Pure real waves correspond to pairs of complex exponentials, hence the two poles in the picture.

Remark: The inversion formulas come with intricate conditions on f and are difficult to prove. In many applications including ours, the inversion formula is not needed. Instead we use the transform to map a known function f to a more difficult function $\hat{f}$. Then easy properties of f become surprising properties of $\hat{f}$.

As an illustration of the complications posed by the inversion formula, consider the Fourier transform. The formula for this transform converges if $|f(x)|$ is integrable. Thus f can have

discontinuous jumps and other bad properties. If we use Lebesgue integration, it can even have certain singularities where it approaches infinity. In the language of Lebesgue, it can be any L^1 function. However, the first significant theorem proved states that $\hat{f}$ is then continuous and $\lim_{x\to\pm\infty} \hat{f}(x) = 0$. So $\hat{f}$ is very nice. But the inversion formula for the Fourier transform is another Fourier transform up to a sign, so the inversion formula could only be true for very nice functions.

However, although $\hat{f}$ goes to zero toward infinity, it need not be integrable. So even further conditions are needed on f for the inversion formula to even make sense. It turns out that if f is continuous and f and $\hat{f}$ are L^1 so both formulas make sense, then the inversion formula is true.

There is one spectacular theorem about the transform. Suppose $f \in L^1$ and also $f \in L^2$. We already know that $\hat{f}$ is continuous, but more is true; $\hat{f} \in L^2$. Moreover the L^2 norms of f and $\hat{f}$ are equal. This is known as Plancherel's theorem. Since $L^1 \cap L^2$ is dense in L^2, this gives a norm preserving map from a dense subset of L^2 to L^2. Easy abstract reasoning then shows that this map can be uniquely extended to give a one-to-one and onto norm preserving isomorphism of L^2. So in a sense, the Fourier transform wants to live in L^2. However, this beautiful result is also very tricky, because the integral defining the Fourier transform will not even exist for some very straightforward elements of L^2, and we must determine a method giving the extension of the map to these elements. Luckily, all of these matters are irrelevant for us.

Remark: The Fourier transform can be generalized to any locally compact abelian topological group, and in that case integration is defined using Haar measure, that is, invariant measure on the group, and the functions $e^{2\pi i x y}$ are replaced with characters of the group. The Fourier transform is then the special case where the group is the additive group R. Notice that the Mellin transform fits this pattern, since $\frac{dx}{x}$ is invariant measure on the multiplicative group of non-zero numbers, and the x^y are group homomorphisms, although to $C^\star$ rather than S^1.

25.6 The Mellin Transform

Remark: We are going to use the Mellin transform. This transformation is also related to taking Fourier transforms on vertical lines. Let us denote the Mellin transform by $\mathcal{M}f(y)$. Then

$$\mathcal{M}f(y) = \int_0^\infty x^y f(x)\,\frac{dx}{x} = \int_0^\infty e^{y\ln x} f(x)\frac{dx}{x} = (\text{substitute } u = \ln x) \int_{-\infty}^\infty e^{yu} f(e^u)\,du$$

and

$$\mathcal{M}f(c+it) = \int_{-\infty}^\infty e^{(c+it)u} f(e^u) du = \int_{-\infty}^\infty e^{itu}\,(e^{cu} f(e^u))\;du$$

The expression on the right is now essentially an inverse Fourier transform, and f is essentially its Fourier transform. So in a sense the Mellin transform reconstructs functions from their

Fourier transforms.

Theorem 183. *The Mellin transform has the following properties:*

- *If the Mellin transform of $f(x)$ is $g(t)$, then the Mellin transform of $f(ax)$ is $a^{-t}g(t)$*
- *The Mellin transform of e^{-x} is $\Gamma(y)$*

Proof: We have

$$\int_0^\infty x^y f(ax)\frac{dx}{x} = (\text{substitute } ax = u)\int_0^\infty \left(\frac{u}{a}\right)^y f(u)\frac{du}{u} = a^{-y}\int_0^\infty u^y f(u)\,\frac{du}{u}$$

and also

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

25.7 $\theta(z)$ and the Zeta Function

Recall that $\theta(q) = \sum q^{n^2}$ where $q = e^{\pi i z}$ and $\theta\left(-\frac{1}{z}\right) = \sqrt{-iz}\theta(z)$. If we restrict attention to the imaginary axis, we have $\theta(ix) = \sum e^{-\pi n^2 x}$ and $\theta\left(\frac{i}{x}\right) = \sqrt{x}\theta(ix)$. In this section only, we follow the convention of redefining θ rather than introducing a new letter. So

$$\theta(x) = \sum_{n=-\infty}^{\infty} e^{-\pi n^2 x}$$

$$\theta\left(\frac{1}{x}\right) = \sqrt{x}\theta(x)$$

It is useful to introduce the closely related function

$$w(x) = \frac{\theta(x)-1}{2} = \sum_{n=1}^{\infty} e^{-\pi n^2 x}$$

We then have the amazing

Theorem 184. *The Mellin transform of $w(x)$ is*

$$\frac{1}{\pi^y}\,\Gamma(y)\,\zeta(2y)$$

Proof: We must form

$$\int_0^\infty \sum_{n=1}^{\infty} x^y e^{-\pi n^2 x}\frac{dx}{x} = (\text{substitute } u = \pi n^2 x)\sum_{n=1}^{\infty}\int_0^\infty \left(\frac{u}{\pi n^2}\right)^{y-1} e^{-u}\frac{du}{\pi n^2} =$$

$$\sum_{n=1}^{\infty} \frac{1}{(\pi n^2)^y} \int_0^{\infty} u^{y-1} e^{-u}\, du = \frac{1}{\pi^y} \sum \frac{1}{n^{2y}}\, \Gamma(y) = \frac{1}{\pi^y}\, \Gamma(y)\, \zeta(2y)$$

QED.

Remark: Next we show that the functional equation of $\theta(x)$ gives the functional equation of the zeta function.

$$w(x) = \frac{\theta(x) - 1}{2} = \frac{\frac{1}{\sqrt{x}}\theta\left(\frac{1}{x}\right) - 1}{2} = \frac{\frac{1}{\sqrt{x}}\left[2w\left(\frac{1}{x}\right) + 1\right] - 1}{2} = \frac{1}{\sqrt{x}} w\left(\frac{1}{x}\right) + \frac{1}{2\sqrt{x}} - \frac{1}{2}$$

We take the Melilin transform of both sides, breaking up the integral into an integral from 0 to 1 plus an integral from 1 to ∞.

$$[\mathcal{M}w](y) = \int_0^1 x^{y-1} w(x)\, dx + \int_1^{\infty} x^{y-1} w(x)\, dx =$$

$$\int_0^1 x^{y-1} \left[\frac{1}{\sqrt{x}} w\left(\frac{1}{x}\right) + \frac{1}{2\sqrt{x}} - \frac{1}{2}\right] dx + \int_1^{\infty} x^{y-1} w(x)\, dx$$

Make the substitution $u = \frac{1}{x}$ in the first integral, getting

$$-\int_1^{\infty} u^{1-y} \left[\sqrt{u} w(u) + \frac{\sqrt{u}}{2} - \frac{1}{2}\right] \frac{-du}{u^2} = \int_1^{\infty} u^{-1-y} \left[\sqrt{u} w(u) + \frac{\sqrt{u}}{2} - \frac{1}{2}\right] du =$$

$$\int_1^{\infty} u^{-1/2-y} w(u) + \frac{1}{2}\left(u^{-1/2-y} - u^{1-y}\right)\, du$$

We conclude that

$$[\mathcal{M}w](y) = \int_1^{\infty} \left(x^{y-1} w(x) + x^{-1/2-y} w(x)\right)\, dx + \int_1^{\infty} \frac{x^{-1/2-y}}{2} dx - \int_1^{\infty} \frac{x^{-1-y}}{2} dx$$

We can compute the last two integrals explicitly. Indeed,

$$\int_1^{\infty} \frac{x^a}{2}\, dx = \left.\frac{x^{a+1}}{2(a+1)}\right|_1^{\infty} = -\frac{1}{2(a+1)} \quad \text{provided } (a+1) < 0$$

The grand conclusion is that

$$[\mathcal{M}w](y) = \pi^{-y}\Gamma(y)\zeta(2y) = \int_1^{\infty} \left(x^{y-1} w(x) + x^{-1/2-y} w(x)\right)\, dx - \frac{1}{2y} - \frac{1}{2(1/2-y)} \quad \text{provided } 0 < y$$

Notice that this last expression is unchanged if we exchange y with $1/2 - y$. So

$$\pi^{-y}\Gamma(y)\zeta(2y) = \pi^{-(1/2-y)}\Gamma(1/2-y)\zeta(2(1/2-y))$$

Replace each y by $y/2$ to obtain the functional equation:

$$\pi^{-y/2}\,\Gamma(y/2)\,\zeta(y) = \pi^{-((1-y)/2)}\,\Gamma((1-y)/2)\,\zeta(1-y)$$

Remark: All this was in Riemann's original paper on the zeta function.

Remark: We avoided issues of convergence in this calculation. The calculation itself was straightforward once two ideas were introduced: replace $\theta(x)$ by $\frac{\theta(x)-1}{2}$ and break up the integral into $\int_0^1$ and $\int_1^\infty$.

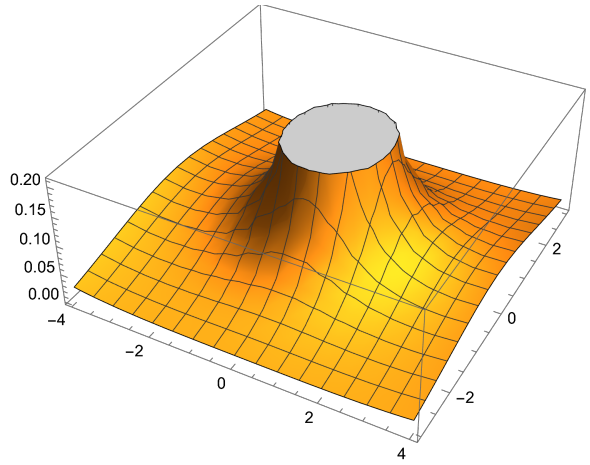
Recall that $\theta = \sum e^{-\pi n^2 x}$. When $x \geq 0$, these terms rapidly decrease to zero *except* the term with $n = 0$, which is identically 1. So θ converges to 1 rather than zero at infinity, and the Mellin integral of θ would not exist. Consequently it was crucial that we use $\theta - 1$, and after that the terms are symmetrical so it was natural to use $\frac{\theta-1}{2}$. Notice that our first calculation can only converge for $2y > 1$ because it reduced by $\sum \frac{1}{n^{2y}}$, which only converges under this condition. So when we changed each $2y$ to y, we needed $y > 1$. The truth is that we could not expect convergence over a larger region because ζ has a pole at one.

Our second calculation requires $y > 0$. This calculation has separated out two poles; after changing $2y$ to y, they are

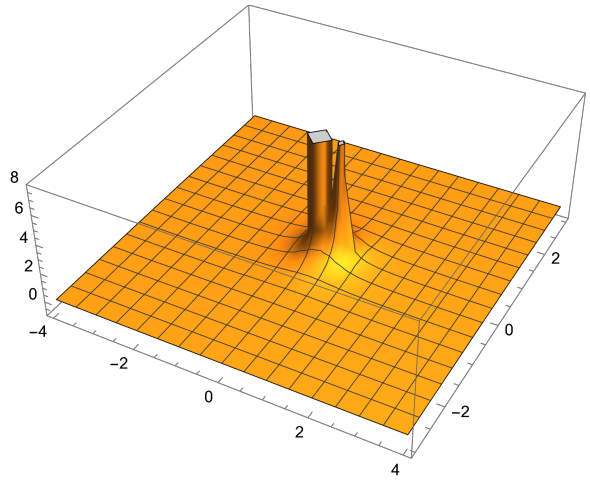
$$-\frac{1}{y} - \frac{1}{1-y}$$

After they are subtracted, the remaining integrals converge. It follows that after the pole at 1 is subtracted away, $\zeta(y)$ can be analytically continued to $\Re(y) > 0$. But then the functional equation allows us to analytically continue it to the rest of the complex plane, and the identity principle implies that the functional equation, originally true in $0 < \Re(y) < 1$, is always true.

Recalll that the Mellin transform is related to the Laplace transform. Below is a picture of the transform of $\frac{\theta(x)-1}{2}$, which can be compared to our picture of the Laplace transform of $e^{-x}\sin x$:



If we increase the resolution, we can see both poles:



25.8 Congruence Subgroups of $SL(2,Z)$

The theta function $\theta(q) = \sum q^{n^2}$ for $q = e^{\pi i \tau}$ satisfies $\theta(\tau+2) = \theta(\tau)$ and $\theta\left(\frac{-1}{\tau}\right) = \sqrt{i\tau}\theta(\tau)$. Consequently, it is associated with a subgroup of the modular group generated by $\tau \to \tau + 2$ and $\tau \to -\frac{1}{\tau}$. In this section we determine that subgroup and its fundamental region. The

subgroup turns out to have finite index in $SL(2,Z)$, so its fundamental region is a union of several fundamental regions for the full modular group.

This is the tip of an iceberg of mathematical research. We could attempt to find all subgroups in $SL(2,Z)$ of finite index. For each we could find a fundamental region and then classify all associated modular forms. This program turns out to be of great interest in number theory, very closely associated with the recent proof of Fermat's last theorem. We will give a small glimpse of the first step of this program.

There is a natural group homomorphism $SL(2,Z) \to SL(2,Z_n)$ obtained by reducing each matrix element modulo n. We will shortly see that this map is onto when $n = 2$, and in fact it is always onto. The kernel of the map is a subgroup of finite index in $SL(n,Z)$ called the *principle congruence subgroup of level n* and denoted $\Gamma(n)$. Concretely, it is the set of all matrices $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with $a,d \equiv 1 \pmod{n}$ and $b,c \equiv 0 \pmod{n}$ and $ad - bc = 1$.

We easily construct other subgroups of finite index by considering groups G with $\Gamma(n) \subset G \subset SL(2,Z)$. These groups are in one-to-one correspondence with subgroups of $SL(2,Z_n)$. Any subgroup of $SL(2,Z)$ obtained in this way is called a *congruence subgroup* of $SL(n,Z)$. When the theory was first studied, it was conjectured that all subgroups of finite index are congruence subgroups, but Felix Klein found a counterexample. Nevertheless, the groups which come up in number theory are often congruence subgroups.

The group $SL(2,Z_n)$ has many subgroups, but an important one contains all upper triangular matrices $\begin{pmatrix} a & b \\ 0 & d \end{pmatrix}$. The associated congruence subgroup of $SL(2,Z)$ is called the *principle congruence subgroup* and denoted $\Gamma_0(N)$. It is these subgroups that are associated with elliptic curves over Q by the modularity theorem at the heart of the proof of Fermat's last theorem.

The matrices $\pm I$ belong to $SL(2,Z)$ and both induce the identity map on the upper half plane. So the modular group of transformations is actually $SL(2,Z)/\pm I$. Thus to obtain subgroups of finite index, we should really consider the map $SL(2,Z)/\pm I \to SL(2,Z_n)/\pm I$. When $n = 2$, $\pm I$ are the same element of $SL(2,Z_n)$ and we can ignore this complication.

Enough of these generalities. Let us consider the subgroup of the modular group generated by $z \to z+2$ and $z \to -\frac{1}{z}$. We call this the *theta group*.

Theorem 185. *The theta group contain* $\Gamma(2)$. *Indeed, it is exactly* $\Gamma(2) \cup \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \Gamma(2)$. *Consequently, it is a congruence subgroup of* $SL(2,Z)$ *associated with the subgroup of order 2 in* $SL(2,Z_2)$ *containing the identity and* $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

Proof: We will prove this in reverse by considering the group $\Gamma(2) \cup \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \Gamma(2)$ and showing that it is generated by the two elements generating the theta group.

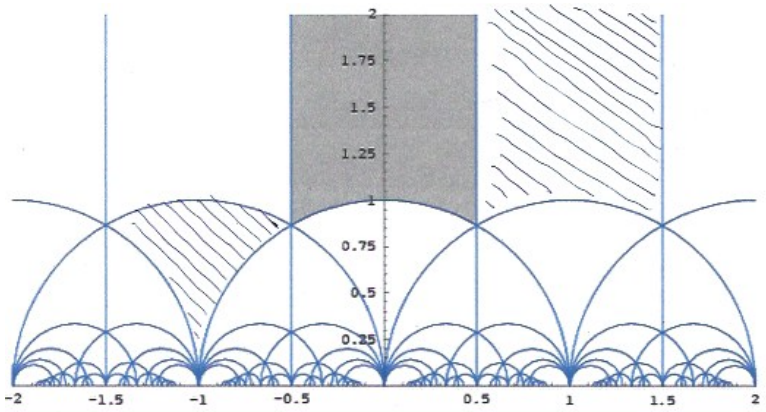
We first claim that the map $SL(2,Z) \to SL(2,Z_2)$ is onto. Indeed, it is easy to see that $SL(2,Z_2)$ has exactly six elements, and they are represented by the following elements of $SL(2,Z)$:

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix},$$

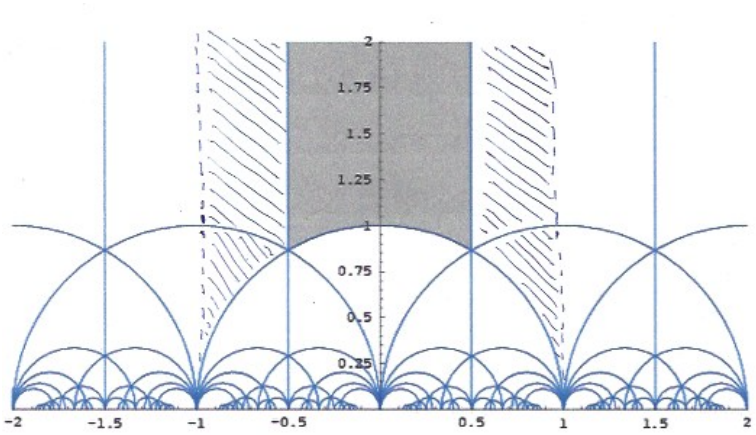
$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix}$$

The group $\Gamma(2)$ is the set of all elements of $SL(2,Z)$ which map to $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, and the set $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \Gamma(2)$ is exactly the set of all elements which map to $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. Consequently the group defined in the theorem has index 3 in $SL(2,Z)$ and the coset representatives are $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix}$, as is easily checked.

From this we easily construct a fundamental region for the group $\Gamma(2) \cup \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \Gamma(2)$. Namely we can take the fundamental region of $SL(2,Z)$, together with its horizontal translate by one, and its image under the map $z \to \frac{1}{z+1}$. This map rotates the fundamental region around the bottom-left corner ω. So the fundamental region is the shaded portion of the following picture:



However, we are free to rearrange this region. We break the right strip into two vertical pieces, and move the right-most piece left two units. Then we break the lower triangle into two pieces and move the left one right by two units. This leaves the following picture.



The great advantage is that the fundamental region can be easily explained It is the set of all z such that $|z| \geq 1$ and $-1 \leq \Re(z) \leq 1$.

This region and its images under $\Gamma(2) \cup \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \Gamma(2)$ cover the plane, so every point in the plane can be moved into this fundamental region by a unique element of the group.

Note that $z \to z+2$ is in $\Gamma(2)$ and thus in our group. Similarly $z \to -\frac{1}{z}$ is in our group. And now finally, we can prove that these two elements generate the group, and therefore the group is exactly the theta group. The proof is exactly like an earlier proof. By a product of translations of the form $z \to z+2$, move a point into the strip $-1 \leq \Re(z) \leq 1$. If the absolute value of the image is greater than or equal to 1, we are done. Otherwise apply the map $z \to -\frac{1}{z}$ and notice that the point is now higher than before. Repeat. But a point can only rise by a finite number of steps, so eventually it will be in the fundamental region. QED.

Remark: We end with another way to think of this fundamental region Notice that the region has three sides, the left vertical side, the semicircle, and the right vertical side. These are all geodesics in non-Euclidean geometry. So our fundamental region is a triangle in non-Euclidean geometry, and images of this triangle under the theta group cover the entire upper half plane.

There is a beautiful theorem about triangles in non-Euclidean geometry. The theorem states that the area of a triangle is

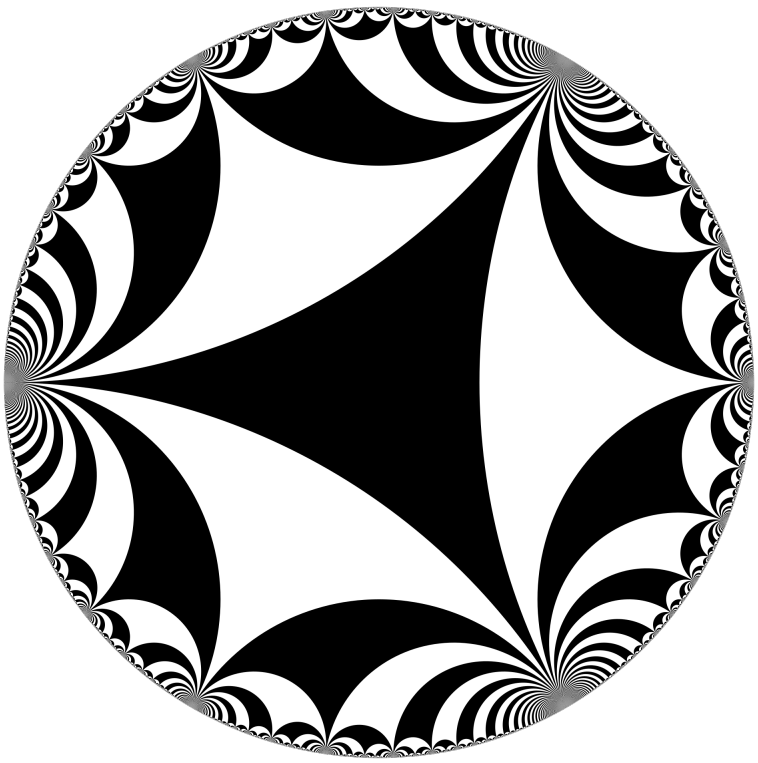
$$\pi - (\alpha + \beta + \gamma)$$

where these are the three angles of the triangle. The sum of the angles of a triangle in non-

Euclidean geometry is smaller than 180 degrees. So if we have a very small triangle, this sum is almost π because in small pieces the geometry looks Euclidean. But big triangles have larger and large areas, so the sum of the angles is smaller. Moreover, there is a limit to the area of a triangle in non-Euclidean geometry: it is never larger than π. This limit is achieved by the largest possible triangles, namely triangles whose vertices are at infinity, where their angle is zero.

The fundamental region of the theta group is such a triangle. Notice that all three vertices are on the boundary of the upper half plane, either on the real axis or way up at infinity.

We can redraw the picture in the unit disk, and then everything becomes clear. Shown below is non-Euclidean space tiled by copies of the fundamental region of the theta group. The theta group is in one-to-one correspondence with these triangles.



Chapter 26

Theta Functions

26.1 Zeros of $\wp(z)$

In our discussion of $\wp(z)$, there is a missing ingredient. We constructed $\wp$ to have a pole of order two at the origin, but we did not determine its zeros. Unfortunately, there is no nice expression for them.

We know that $\wp$ has two zeros or else one zero of order two. For symmetry reasons, we might hope for a single zero of order two. But then $\wp'(z)$ would vanish at the zero, so $\wp(z)$ would satisfy $\wp^3 - g_2\wp - g_3 = 0$ and therefore be one of the roots of $X^3 - g_2X - g_3 = 0$. These roots are e_1, e_2, e_3 and they are taken at $\frac{\omega_1}{2}$, $\frac{\omega_2}{2}$, and $\frac{\omega_1+\omega_2}{2}$. This suggests that better choices for our fundamental function would have been one of

$$\wp(z) - e_1, \wp(z) - e_2, \wp(z) - e_3$$

because these functions have one pole of order two and one zero of order two. However, there is no natural way to pick *one* of the three choices. If we pick $\wp(z) - e_3$ because its zero is in the center of the period parallelogram, a different choice of lattice basis would put it on the outside boundary. Instead it is best to think of the $\wp(z) - e_i$ as three particularly nice functions.

Theorem 186. *Up to translation and multiplication by a constant, there are three doubly-periodic functions with one zero of order two and one pole of order two, namely the $\wp(z) - e_i$.*

Proof: Suppose $f(z)$ is such a function and translate so the pole is at the origin. Since the sum of the residues is zero, the Laurent expansion of f at the pole must be $\frac{a_{-2}}{z^2} + a_0 + a_1 z + \ldots$. So $\frac{1}{a_{-2}} f(z) - \wp(z)$ has no poles and thus is constant, and our function has become $\wp(z) + c$. The reasoning at the beginning of the section shows that c must be $-e_1$, $-e_2$, or $-e_3$.

26.2 The Theta Functions

When we constructed doubly periodic functions with given zeros and poles, we constructed a function $\sigma(z)$ which was almost periodic and had zeros of order one at lattice points. Elliptic functions were then obtained by translating this function to provide zeros at other points, and then forming a quotient to provide given zeros and poles.

The theta functions serve a similar purpose, but they are given by rapidly converging series and thus are easier to use for computations. Jacobi showed that they also have astonishing number theoretic properties. We assume the lattice basis is 1 and τ and define an entire function $\theta_1(z)$ with simple zeros at the lattice points and no other zeros. This function cannot be doubly periodic, but it satisfies known equations when translated by a lattice point. Translations of the function produce zeros at other points, and products and quotients of the translations give all elliptic functions.

The formula for arbitrary translations is messy and not needed. But we obtain nice formulas if we translate by $\frac{1}{2}$, $\frac{\tau}{2}$, or $\frac{1}{2}+\frac{\tau}{2}$, and that is enough to give the $\wp - e_i$ of the previous section. It is very convenient to give special names to these three translations. Let $p = e^{\pi i z}$ and let $q = e^{\pi i \tau}$. Then

$$\theta_1(z,\tau) = i\sum_{n=-\infty}^{\infty}(-1)^n p^{2n-1} q^{\left(n-\frac{1}{2}\right)^2}$$

$$\theta_2(z,\tau) = \sum_{n=-\infty}^{\infty} p^{2n-1} q^{\left(n-\frac{1}{2}\right)^2}$$

$$\theta_3(z,\tau) = \sum_{n=-\infty}^{\infty} p^{2n} q^{n^2}$$

$$\theta_4(z,\tau) = \sum_{n=-\infty}^{\infty}(-1)^n p^{2n} q^{n^2}$$

Remark: When we use these functions to investigate elliptic functions, we fix τ and thus the lattice and just write $\theta_i(z)$. When we are interested in modularity, we usually set $z = 0$ and write $\theta(0)$ as a function of q. Since $z = 0$ gives $p = 1$, we have

$$\theta_1(0) = i\sum(-1)^n q^{\left(n-\frac{1}{2}\right)^2}$$

$$\theta_2(0) = \sum q^{\left(n-\frac{1}{2}\right)^2}$$

$$\theta_3(0) = \sum q^{n^2}$$

$$\theta_4(0) = \sum(-1)^n q^{n^2}$$

Thus $\theta_3(0)$ is the special function studied at the end of the previous chapter.

Remark: These functions are almost periodic. We have

Theorem 187.

$$\theta_1(z+1) = -\theta_1(z)$$
$$\theta_2(z+1) = -\theta_2(z)$$
$$\theta_3(z+1) = \theta_3(z)$$
$$\theta_4(z+1) = \theta_4(z)$$

and, when $b = p^{-2}q^{-1}$

$$\theta_1(z+\tau) = -b\,\theta_1(z)$$
$$\theta_2(z+\tau) = b\,\theta_2(z)$$
$$\theta_3(z+\tau) = b\,\theta_3(z)$$
$$\theta_4(z+\tau) = -b\,\theta_4(z)$$

Proof: Since $e^{\pi i(z+1)} = -e^{\pi iz}$ and $e^{\pi i(z+\tau)} = e^{\pi iz}e^{\pi i\tau}$, adding 1 changes p to $-p$ and adding τ changes p to pq. The theorem easily follows. We'll give the proof in the second group of equations for θ_3 to give the flavor of the argument. There

$$\theta_3(z+\tau) = \sum (pq)^{2n}q^{n^2} = \sum p^{2n}q^{n^2+2n} = \sum p^{2n}q^{n^2+2n+1}q^{-1} = q^{-1}\sum p^{2n}q^{(n+1)^2} =$$
$$p^{-2}q^{-1}\sum p^{2(n+1)}q^{(n+1)^2} = p^{-2}q^{-1}\sum p^{2n}q^{n^2}$$

Theorem 188. *The functions also behave nicely if translated by* $\frac{1}{2}$, $\frac{\tau}{2}$, *or* $\frac{1}{2}+\frac{\tau}{2}$. *Indeed if we set* $a = p^{-1}q^{-1/4}$,

$$\theta_1(z+\tfrac{1}{2}) = \theta_2(z) \qquad \theta_1(z+\tfrac{\tau}{2}) = ia\,\theta_4(z) \qquad \theta_1(z+\tfrac{1}{2}+\tfrac{\tau}{2}) = a\,\theta_3(z)$$

$$\theta_2(z+\tfrac{1}{2}) = -\theta_1(z) \qquad \theta_2(z+\tfrac{\tau}{2}) = a\,\theta_3(z) \qquad \theta_2(z+\tfrac{1}{2}+\tfrac{\tau}{2}) = -ia\,\theta_4(z)$$

$$\theta_3(z+\tfrac{1}{2}) = \theta_4(z) \qquad \theta_3(z+\tfrac{\tau}{2}) = a\,\theta_2(z) \qquad \theta_3(z+\tfrac{1}{2}+\tfrac{\tau}{2}) = ia\,\theta_1(z)$$

$$\theta_4(z+\tfrac{1}{2}) = \theta_3(z) \qquad \theta_4(z+\tfrac{\tau}{2}) = ia\,\theta_1(z) \qquad \theta_4(z+\tfrac{1}{2}+\tfrac{\tau}{2}) = a\,\theta_2(z)$$

Proof: Let $\epsilon = e^{\left(\frac{2\pi i}{4}\right)}$. This time $e^{\pi i(z+\frac{1}{2})} = p\epsilon$, $e^{\pi i(z+\frac{\tau}{2})} = p\sqrt{q}$, and $e^{\pi i(z+\frac{1}{2}+\frac{\tau}{2})} = p\epsilon\sqrt{q}$.

Theorem 189. *The function* $\theta_1(z)$ *has a simple zero at all lattice points L and no other zeros. The function* $\theta_2(z)$ *has a simple zero at all* $L+\frac{1}{2}$ *and no other zeros, the function* $\theta_4(z)$ *has a simple zero at all* $L+\frac{\tau}{2}$ *and no other zeros, and the function* $\theta_3(z)$ *has a simple zero at all* $L+\frac{1}{2}+\frac{\tau}{2}$ *and no other zeros.*

Proof: This follows from previous theorems if we can prove the assertion about θ_1.

Note that $\theta_1(0) = i\sum(-1)^n q^{\left(n-\frac{1}{2}\right)^2}$. The sequence $n - \frac{1}{2}$ for $-1 \le n \le 2$ is $-\frac{3}{2}, -\frac{1}{2}, \frac{1}{2}, \frac{3}{2}, \ldots$ and so each fraction occurs twice, once with an even n and once with an odd n. Consequently $\theta_1(0) = 0$.

Next we let $f(z) = \theta_1(z)$ and compute $\frac{1}{2\pi i}\int \frac{f'(z)}{f(z)}\,dz$ around the boundary of a fundamental parallelogram. This integral counts the zeros and poles inside the boundary by the argument principle. There are no poles, so it counts the zeros. Shift the boundary slightly left and down so the origin is inside. Recall that $\theta_1(z+1) = -\theta_1(z)$, but we need to integrate $\frac{f'}{f}$ and $\frac{\theta'(z+1)}{\theta(z+1)} = \frac{-\theta'(z)}{-\theta(z)} = \frac{\theta'(z)}{\theta(z)}$. So the left and right integrals cancel out. Also $\theta_1(z+\tau) = -b\theta_1(z)$ where $b = p^{-2}q = e^{-2\pi i z}e^{-\pi i \tau}$ and so $\theta_1(z+\tau) = -e^{-2\pi i z}e^{-\pi i\tau}\theta_1(z)$. Consequently using additivity of logarithmetic derivatives,

$$\frac{\theta_1'(z+\tau)}{\theta_1(z+\tau)} = (-2\pi i) + \frac{\theta_1'(z)}{\theta_1(z)}$$

We must add integral over the bottom and the integral over the top, where the bottom path is $\gamma(t) = t$ and the top path is $\gamma(t) = \tau + t$ and $0 \le t \le 1$ and the top counts negatively and the bottom counts positively. (Also we must push both paths left and down slightly, but that makes no difference in the argument we are about to give.) So the top and bottom integrals cancel out except that the top gives an extra term $\frac{1}{2\pi i}\int_0^1(-2\pi i)\,dt = -1$. However, that top integral goes from right to left. So the total integral around the path is 1 and we find that θ_1 has a zero at the origin of order one and no other zeros. QED.

Remark: How could anyone discover these formulas? Since they differ by a translation, it would suffice to find one of them. The easiest to guess is θ_3, which has zeros at $m\frac{1}{2} + n\frac{\tau}{2}$. At each zero, $p = e^{\pi i z} = e^{\pi i \frac{1}{2}m}e^{\pi i \frac{\tau}{2}n} = i^m q^{\frac{n}{2}}$ or equivalently $p^2 = (-1)^m q^n$. We want a sum of terms, each a product of p's and q's, and we want q to be raised to roughly n^2 to get very rapid convergence. To get zeros, some terms need plus signs and some need minus signs. The term involving p needs to depend on n, else we could factor it out and get a p term periodic in only one direction. The most natural thing to try is $\sum p^{2n}q^{n^2}$ because at zeros it would equal $\sum(-1)^n q^{n^2+n}$. Notice that $n = -1$ and $n = 0$ give $-q^0$ and q^0 and cancel, and in general the $-k$ and $k-1$ terms differ by a sign and cancel. Thus this function has the correct zeros and makes a good $\theta_3(z)$.

Important Warning: The constants a and b depend on z, so applying the formulas twice is tricky. For example, $\theta_1(z+\tau) = -b\theta_1(z)$ where b depends on z and τ, so this equation does not imply $\theta_1(z+2\tau) = b^2\theta_1(z)$. Indeed a careful calculation shows that $\theta_1(z+2\tau) = (pq)^{-\frac{1}{4}}\theta_1(z)$.

26.3 Theta Formulas for Weierstrass Functions

Theorem 190.

$$\wp(z) - e_1 = \left(\frac{\theta_1'(0)}{\theta_2(0)}\right)^2 \left(\frac{\theta_2(z)}{\theta_1(z)}\right)^2$$

$$\wp(z) - e_2 = \left(\frac{\theta_1'(0)}{\theta_3(0)}\right)^2 \left(\frac{\theta_3(z)}{\theta_1(z)}\right)^2$$

$$\wp(z) - e_3 = \left(\frac{\theta_1'(0)}{\theta_4(0)}\right)^2 \left(\frac{\theta_4(z)}{\theta_1(z)}\right)^2$$

Proof: Consider the top equation. The expression on the right side has a pole of order two at the origin and a zero of order two at $\frac{1}{2}$. By the first section of this chapter, it equals $\wp(z)+e_1$ up to a constant. To show that this constant is one, we need only check that the Laurent series of both sides at the origin begin $\frac{1}{z^2} + ...$ This is well known for $\wp(z)$. The value $\theta_2(0)$ cancels the value of $\theta_2(z)$ at zero. Since the origin is a zero of θ_1, we have a series expansion $\theta(z) = \theta'(0)z + ...$, so $\frac{1}{\theta_1(z)} = \frac{1}{\theta'(0)}\frac{1}{z} + ...$ and therefore

$$\left(\theta'(0)\frac{1}{\theta_1(z)}\right)^2 = \frac{1}{z^2} + ...$$

The same argument clearly works for the other two terms.

Theorem 191. *We have*

$$e_1 - e_2 = \left(\frac{\theta_1'(0)\theta_4(0)}{\theta_2(0)\theta_3(0)}\right)^2$$

$$e_1 - e_3 = \left(\frac{\theta_1'(0)\theta_3(0)}{\theta_2(0)\theta_4(0)}\right)^2$$

$$e_2 - e_3 = \left(\frac{\theta_1'(0)\theta_2(0)}{\theta_3(0)\theta_4(0)}\right)^2$$

Proof: All three results are proved the same way; we'll show the proof for the first case. Substitute $z = \frac{1}{2}$ in the second formula of the previous theorem and recall that $\wp\left(\frac{1}{2}\right) = e_1$. We obtain

$$e_1 - e_2 = \left(\frac{\theta_1'(0)\theta_3(1/2)}{\theta_3(0)\theta_1(1/2)}\right)^2$$

But a previous result gives $\theta_1(z+1/2) = \theta_2(z)$ and $\theta_3(z+1/2) = \theta_4(z)$. Set $z = 0$ and substitute in the above formula. QED.

Remark: Both previous theorems include $\theta'(0)$, which would be difficult to compute. The following important result of Jacobi eliminates this problem.

Theorem 192. *We have*

$$\theta_1'(0) = \pi\theta_2(0)\theta_3(0)\theta_4(0)$$

Step One of Proof: Several proofs are known, but all are somewhat tricky. This comes from an exercise in Whittaker and Watson. It will be the most difficult proof in this chapter. Consider the function

$$2\frac{\theta_1(z)\theta_2(z)\theta_3(z)\theta_4(z)}{\theta_1(2z)\theta_2(0)\theta_3(0)\theta_4(0)}$$

The numerator has simple zeros at 0, $\frac{1}{2}$, $\frac{\tau}{2}$, and $\frac{1}{2}+\frac{\tau}{2}$ and thus on all points of the lattice $L/2$. So does $\theta_1(2z)$. Therefore the quotient has no zeros. When we replace z by $z+\frac{1}{2}$, the numerator is multiplied by -1 by a previous theorem. When we replace z by $z+\frac{\tau}{2}$, the numerator is multiplied by $-a^4 = -\left(p^{-1}q^{-1/4}\right)^4 = -p^{-4}q^{-1}$ by that same theorem. We will show that $\theta_1(2z)$ has the same properties. If so, then the quotient is doubly-periodic with no zeros or poles, and thus a constant.

The constant is clearly the limit of $2\frac{\theta_1(x)}{\theta_1(2x)}$ as x approaches zero, Since θ_1 has a zero of order one at $x=0$, it has a series expansion $a_1x + a_2x^2 + ... = a_1x\left(1 + b_1x + b_2x^2 + ...\right)$ near the origin, the quotient equals

$$\frac{2a_1x\left(1 + b_1x + ...\right)}{2a_1x\left(1 + 2b_1x + ...\right)}$$

and the limit is clearly 1.

Thus we need to prove the missing formulas for $\theta_1(2z)$. We first prove that $\theta_1\left(2(z+\frac{1}{2})\right) = -\theta_1(2z)$. On the left side we have $2z+1$, which is really $e^{2\pi iz+\pi i} = -p^2$. On the right we have $2z$, which is really p^2. So we must prove that $\theta_1(-p^2,q) = -\theta_1(p^2,q)$ and this is clear because

$$\theta_1(-p^2,q) = i\sum(-1)^n(-p^2)^{2n-1}q^{\left(n-\frac{1}{2}\right)^2} = -i\sum(-1)^n(p^2)^{2n-1}q^{\left(n-\frac{1}{2}\right)^2}$$

Then we must prove that $\theta_1\left(2(z+\frac{\tau}{2})\right) = -p^{-4}q^{-1}\theta_1(2z)$. The term $2(z+\frac{\tau}{2})$ is really p^2q because $e^{\pi i(2z+\tau)} = \left(e^{\pi iz}\right)^2 e^{\pi i\tau}$. The term $2z$ is really p^2. We must prove that

$$\theta_1(p^2q,q) = -p^{-4}q^{-1}\theta_1(p^2,q)$$

But

$$\theta_1(p^2q,q) = i\sum(-1)^n(p^2q)^{2n-1}q^{\left(n-\frac{1}{2}\right)^2} = i\sum(-1)^n(p^2)^{2n-1}q^{\left(n-\frac{1}{2}\right)^2+2n-1} =$$

$$i\sum(-1)^n(p^2)^{2n-1}q^{\left(n+\frac{1}{2}\right)^2-1}$$

Choose m so $m-\frac{1}{2} = n+\frac{1}{2}$ or $m-1=n$. Then the above is

$$i\sum(-1)^{m-1}(p^2)^{2m-3}q^{\left(m-\frac{1}{2}\right)^2}q^{-1} = -ip^{-4}q^{-1}\sum(-1)^m(p^2)^{2m-1}q^{\left(m-\frac{1}{2}\right)} = -p^{-4}q^{-1}\theta_1(p^2,q)$$

We have proves a *doubling formula*:

$$\theta_1(2z) = \frac{2\theta_1(z)\theta_2(z)\theta_3(z)\theta_4(z)}{\theta_2(0)\theta_3(0)\theta_4(0)}$$

Step Two of Proof: We now take the logarithmetic derivative of both sides of this formula.

$$\frac{2\,\theta_1'(2z)}{\theta_1(2z)} = \frac{\theta_1'(z)}{\theta_1(z)} + \frac{\theta_2'(z)}{\theta_2(z)} + \frac{\theta_3'(z)}{\theta_3(z)} + \frac{\theta_4'(z)}{\theta_4(z)}$$

We differentiate once more:

$$4\left[\frac{\theta_1''(2z)}{\theta_1(2z)} - \left(\frac{\theta_1'(2z)}{\theta_1(2z)}\right)^2\right] = \frac{\theta_1''(z)}{\theta_1(z)} - \left(\frac{\theta_1'(z)}{\theta_1(z)}\right)^2 + \frac{\theta_2''(z)}{\theta_2(z)} - \left(\frac{\theta_2'(z)}{\theta_2(z)}\right)^2 + \frac{\theta_3''(z)}{\theta_3(z)} - \left(\frac{\theta_3'(z)}{\theta_3(z)}\right)^2 + \frac{\theta_4''(z)}{\theta_4(z)} - \left(\frac{\theta_4'(z)}{\theta_4(z)}\right)^2$$

and rewrite this as

$$\left[4\frac{\theta_1''(2z)}{\theta_1(2z)} - \frac{\theta_1''(z)}{\theta_1(z)}\right] + \left[\left(\frac{\theta_1'(z)}{\theta_1(z)}\right)^2 - 4\left(\frac{\theta_1'(2z)}{\theta_1(2z)}\right)^2\right] = \frac{\theta_2''(z)}{\theta_2(z)} - \left(\frac{\theta_2'(z)}{\theta_2(z)}\right)^2 + \frac{\theta_3''(z)}{\theta_3(z)} - \left(\frac{\theta_3'(z)}{\theta_3(z)}\right)^2 + \frac{\theta_4''(z)}{\theta_4(z)} - \left(\frac{\theta_4'(z)}{\theta_4(z)}\right)^2$$

Step Three of Proof: We claim that $\theta_1(-z) = -\theta_1(z)$, but $\theta_i(-z) = \theta_i(z)$ for $i = 2,3,4$. Note that $p = e^{\pi iz}$ and $p^{-1} = e^{-\pi iz}$. So this involves proving that the first theta function changes sign if we replace p by p^{-1}, while the other three are unchanged. This is completely clear for θ_3 and θ_4 because the formulas involve p^{2n} and changing p to p^{-1} is the same thing as changing n to $-n$, and the formulas are invariant under this change.

In the formulas for θ_1 and θ_2, replace n by $-n+1$. Then $n-\frac{1}{2}$ is replaced by $-n+\frac{1}{2}$ and both of these expressions have the same square. Also p^{2n-1} is replaced by $p^{-2n+1} = \left(p^{-1}\right)^{2n-1}$. Finally, the $(-1)^n$ in θ_1 is changed to $(-1)^{-n+1} = -(-1)^n$. So the total effect for θ_1 is to change p to p^{-1} and change the sign of the function, and the total effect for θ_2 is to just change p to p^{-1}. Said another way, if we insert p^{-1} for p at the start, θ_1 changes sign and θ_2 remains unchanged.

Step Four of Proof: We now set $z = 0$ in the formula obtained in step 2. This is tricky for terms involving θ_1 because $\theta_1(0) = 0$, but for the remaining terms we have $\theta_2'(0) = \theta_3'(0) = \theta_4'(0) = 0$ because $\theta_i(-z) = \theta_i(z)$. So the right side becomes

$$\frac{\theta_2''(0)}{\theta_2(0)} + \frac{\theta_3''(0)}{\theta_3(0)} + \frac{\theta_4''(0)}{\theta_4(0)}$$

We now evaluate the right side. Since $\theta_1(z)$ is odd and vanishes at $z = 0$, we have

$$\theta_1(z) = a_1 z + a_3 z^3 + ...$$
$$\theta_1'(z) = a_1 + 3a_3 z^2 + ...$$
$$\theta_1''(z) = 6a_3 z + ...$$

Let $\lambda = \frac{a_3}{a_1} = \frac{\theta_1'''(0)/3!}{\theta_1'(0)} = \frac{1}{6}\frac{\theta_1'''(0)}{\theta_1'(0)}$. Then

$$\frac{\theta_1'(z)}{\theta(z)} = \frac{a_1 + 3a_3 z^2 + ...}{a_1 z + a_3 z^3 + ...} = \frac{(1 + 3\lambda z^2 + ...)}{z(1 + \lambda z^2 + ...)} = \frac{(1 + 3\lambda z^2 + ...)(1 - \lambda z^2)}{z(1 + \lambda z^2 + ...)(1 - \lambda z^2)} =$$

$$\frac{1}{z}\left(1 + 2\lambda z^2 + ...\right) = \frac{1}{z} + 2\lambda z + ...$$

So

$$\left(\frac{\theta_1'(z)}{\theta_1(z)}\right)^2 = \frac{1}{z^2} + 4\lambda + ...$$

and

$$\left(\frac{\theta_1'(z)}{\theta_1(z)}\right)^2 - 4\left(\frac{\theta_1'(2z)}{\theta_1(2z)}\right)^2 = \left(\frac{1}{z^2} + 4\lambda + ...\right) - 4\left(\frac{1}{4z^2} + 4\lambda + ...\right) = -12\lambda + ...$$

Similarly,

$$\frac{\theta_1''(z)}{\theta_1(z)} = \frac{6a_3 z + ...}{a_1 z + a_3 z^3} = 6\lambda + ...$$

and

$$4\frac{\theta_1''(2z)}{\theta_1(2z)} - \frac{\theta_1''(z)}{\theta_1(z)} = 4(6\lambda) - 6\lambda = 18\lambda$$

The complete expression on the left side of our big equation is thus

$$18\lambda - 12\lambda = 6\lambda = \frac{\theta_1'''(0)}{\theta_1'(0)}$$

In short

$$\frac{\theta_1'''(0)}{\theta_1'(0)} = \frac{\theta_2''(0)}{\theta_2(0)} + \frac{\theta_3''(0)}{\theta_3(0)} + \frac{\theta_4''(0)}{\theta_4(0)}$$

Step Five of the Proof: The four theta functions satisfy the heat equation:

$$\frac{\partial^2\theta}{\partial z^2} = 4\pi i \frac{\partial\theta}{\partial\tau}$$

Indeed $\frac{\partial}{\partial z}\theta(e^{\pi iz}) = \frac{\partial\theta}{\partial p}\pi i e^{\pi iz} = \pi i p \frac{\partial\theta}{\partial p}$, so $\frac{\partial^2}{\partial^2 z}\theta(e^{\pi iz}) = (\pi i p)^2 \frac{\partial^2\theta}{\partial p^2}$. Similarly, $\frac{\partial\theta}{\partial\tau} = \pi i q \frac{\partial\theta}{\partial q}$.

Consider then θ_1. Differentiating with respect to p changes p^{2n-1} to $(2n-1)p^{2n-1}$ and differentiating twice changes it to $(2n-1)^2 p^{2n-1}$. On the other hand, differentiating $q^{\left(n-\frac{1}{2}\right)^2}$ with respect to q changes the term to $\left(n-\frac{1}{2}\right)^2 q^{\left(n-\frac{1}{2}\right)^2}$. Note that $(2n-1)^2 = 4\left(n-\frac{1}{2}\right)^2$. So the derivative with respect to p is four times the derivative with respect to q. We get an extra πi comparing $(\pi i)^2$ to πi, so all told we get an extra factor of $4\pi i$. Exactly the same calculation works in the other cases.

Step Six of the Proof: Consider the expression

$$\frac{\theta_2(0)\theta_3(0)\theta_4(0)}{\theta_1'(0)}$$

This is still a function of τ, and we can form its logarithmetic derivative with respect to τ. The previous steps show that this logarithmetic derivative is zero, and consequently the above function is a constant independent of τ. This is the result we have been aiming for all this time! To finish the proof, we need only show that the constant is $\frac{1}{\pi}$.

We are going to let $q \to 0$ by letting $\tau \to i\,\infty$. Direct inspection shows that $\theta_3(0)$ and $\theta_4(0)$ approach 1. The terms of $\theta_2(0)$ approaching zero the slowest are $q^{\left(\frac{1}{2}\right)^2}$ and $q^{\left(-\frac{1}{2}\right)^2}$, so this sum behaves like $2q^{\frac{1}{4}}$. Finally

$$\frac{\partial\theta_1}{\partial z}(e^{\pi iz}) = \frac{\partial\theta_1}{\partial p}(\pi i p) = (\pi i) i \sum (-1)^n (2n-1) p^{2n-1} q^{\left(n-\frac{1}{2}\right)^2}$$

and

$$\theta_1'(0) = -\pi \sum (-1)^n (2n-1) q^{\left(n-\frac{1}{2}\right)^2}$$

The significant terms in this sum occur when $n = 0$ and $n-1$ making the exponent of q equal $\frac{1}{4}$. These two terms give $\pi q^{\frac{1}{4}}$ and $\pi q^{\frac{1}{4}}$ for a total contribution of $2\pi q^{\frac{1}{4}}$. So $\frac{\theta_2(0)}{\theta_1'(0)}$ contributes $\frac{1}{\pi}$ in the limit.

Whew. QED.

Remark: We can now rewrite a previous theorem in nicer form:

Theorem 193. *We have*

$$e_1 - e_2 = \pi^2\, \theta_4^4(0)$$

$$e_1 - e_3 = \pi^2\, \theta_3^4(0)$$

$$e_2 - e_3 = \pi^2\, \theta_2^4(0)$$

$$e_1 = \frac{1}{3}\,\pi^2\,\left[\theta_3^4(0) + \theta_4^4(0)\right]$$

$$e_2 = \frac{1}{3}\,\pi^2\,\left[\theta_2^4(0) - \theta_4^4(0)\right]$$

$$e_3 = -\frac{1}{3}\,\pi^2\,\left[\theta_2^4(0) + \theta_3^4(0)\right]$$

$$g_2 = \frac{2}{3}\pi^4\left[\theta_2^8(0) + \theta_3^8(0) + \theta_4^8(0)\right]$$

$$\Delta = g_2^3 - 27g_3^2 = \frac{1}{4}(e_1 - e_2)^2(e_1 - e_3)^2(e_2 - e_3)^2 = \frac{1}{4}\pi^{12}\left[\theta_2(0)\theta_3(0)\theta_4(0)\right]^8$$

$$j = \frac{g_2^3}{\Delta} = \frac{64}{54}\,\frac{\left[\theta_2^8(0) + \theta_3^8(0) + \theta_4^8(0)\right]^3}{\left[\theta_2(0)\theta_3(0)\theta_4(0)\right]^8}$$

Proof: The first three are immediate. Since e_1, e_2, e_3 are the roots of $4X^3 - g_2X - g_2$ and this polynomial has no X^2 term, $e_1 + e_2 + e_3 = 0$. So $(e_1 - e_2) + (e_1 - e_3) = 2e_1 - (e_2 + e_3) = 2e_1 + e_1 = 3e_1$. The second three formulas follow from this.

Write $4X^3 - g_2X - g_3 = 4(X - e_1)(X - e_2)(X - e_3) = 4X^3 + 4(e_1e_2 + e_1e_3 + e_2e_3)X - 4e_1e_2e_3$. Thus $g_2 = -4(e_1e_2 + e_1e_3 + e_2e_3)$ and $g_3 = 4e_1e_2e_3$. We have

$$(e_1 - e_2)^2 + (e_1 - e_3)^2 + (e_2 - e_3)^2 = 2(e_1^2 + e_2^2 + e_3^2) - 2(e_1e_2 + e_1e_3 + e_2e_3)$$

$$0 = (e_1 + e_2 + e_3)^2 = (e_1^2 + e_2^2 + e_3^2) + 2(e_1e_2 + e_1e_3 + e_2e_3)$$

and subtracting twice the second from the first gives

$$(e_1 - e_2)^2 + (e_1 - e_3)^2 + (e_2 - e_3)^2 = -6(e_1e_2 + e_1e_3 + e_2e_3) = -6\left(-\frac{g_2}{4}\right) = \frac{3}{2}g_2$$

This is the third formula from the bottom.

Differentiating $4X^3 - g_2X - g_3$ gives

$$12X^2 - g_2 = 4\left[(X-e_2)(X-e_3) + (X-e_1)(X-e_3) + (X-e_1)(X-e_2)\right]$$

and so

$$12e_1^2 - g_2 = (e_1-e_2)(e_1-e_3)$$
$$12e_2^2 - g_2 = (e_2-e_1)(e_2-e_3)$$
$$12e_3^2 - g_2 = (e_3-e_1)(e_3-e_2)$$

or

$$(12e_1^2 - g_2)(12e_2^2 - g_2)(12e_3^2 - g_2) = -(e_1-e_2)^2(e_1-e_3)^2(e_2-e_3)^2$$

or

$$(e_1-e_2)^2(e_1-e_3)^2(e_2-e_3)^2 =$$
$$-(12)^3(e_1e_2e_3)^2 + (12)^2 g_2(e_1^2e_2^2 + e_1^2e_3^2 + e_2^2e_3^2) - 12g_2^2(e_1^2+e_2^2+e_3^2) + g_2^3$$

In this formula, $e_1e_2e_3 = \frac{g_3}{4}$ and $e_1^2+e_2^2+e_3^2 = -2(e_1e_2+e_1e_3+e_2e_3) = -2\left(\frac{g_2}{4}\right) = -\frac{g_2}{2}$.

Notice that $\frac{g_2^2}{4} = (e_1^2+e_2^2+e_3^2)^2 = \left[(e_1+e_2+e_3)^2 - 2(e_1e_2+e_1e_3+e_2e_3)\right]^2$. But $e_1+e_2+e_3 = 0$, so

$$\frac{g_2^2}{4} = 4(e_1e_2+e_1e_3+e_2e_3)^2 = 4(e_1^2e_2^2+e_1^2e_3^2+e_2^2e_3^2) + 8(e_1^2e_2e_3 + e_2^2e_1e_3 + e_3^2e_1e_2)$$

The last term is $8(e_1+e_2+e_3)(e_1e_2e_3) = 0$, so

$$\frac{g_2^2}{4} = 4(e_1^2e_2^3 + e_1^2e_3^2 + e_2^2e_3^2)$$

The grand conclusion is that

$$(e_1-e_2)^2(e_1-e_3)^2(e_2-e_3)^2 =$$
$$-(12)^3\left(\frac{g_3}{4}\right)^2 + (12)^2\frac{g_2^3}{16} - 12\frac{g_2^3}{2} + g_2^3 = 4(g_2^3 - 27g_3^2)$$

Warning: Some authors define the discriminant of $4X^3 - g_2X - g_3$ to be

$$\Delta = 16(e_1-e_2)^2(e_1-e_3)^2(e_2-e_3)^2$$

and thus 64 times as large as $g_2^3 - 27g_3^2$. When this convention is used, the next to last formula must be multiplied by 64 and thus the $\frac{1}{4}$ on the right side becomes instead 16. The definition of j has an extra factor, and the last formula changes by omitting the 64 on the right. Therefore when consulting other sources, caution is recommended.

26.4 Product Formulas

Since the θ_i have simple zeros at known spots, the methods of Weierstrass can be used to decompose them as infinite products. The convergence of each theta is rapid, so the most naive version of Weierstrass' results suffices. Nevertheless, these products turn out to be beautiful and very powerful.

Theorem 194 (Jacobi). *Let* $C = \prod_{n=1}^{\infty}(1-q^{n^2})$. *Then*

$$i\theta_1(z) = Cq^{\frac{1}{4}}(p-p^{-1})\prod_{n=1}^{\infty}(1-q^{2n}p^2)(1-q^{2n}p^{-2})$$

$$\theta_2(z) = Cq^{\frac{1}{4}}(p+p^{-1})\prod_{n=1}^{\infty}(1+q^{2n}p^2)(1+q^{2n}p^{-2})$$

$$\theta_3(z) = C\prod_{n=1}^{\infty}(1+q^{2n-1}p^2)(1+q^{2n-1}p^{-2})$$

$$\theta_4(z) = C\prod_{n=1}^{\infty}(1-q^{2n-1}p^2)(1-q^{2n-1}p^{-2})$$

These lead to beautiful product formulas for null values obtained by setting $z = 0$, so $p = 1$:

Theorem 195 (Jacobi). *Let* $C = \prod_{n=1}^{\infty}(1-q^{2n})$. *Then*

$$\pi^{-1}\theta_1'(0) = 2Cq^{\frac{1}{4}}\prod_{n=1}^{\infty}(1-q^{2n})^2$$

$$\theta_2(0) = 2Cq^{\frac{1}{4}}\prod_{n=1}^{\infty}(1+q^{2n})(1+q^{2n})$$

$$\theta_3(0) = C\prod_{n=1}^{\infty}(1+q^{2n-1})(1+q^{2n-1})$$

$$\theta_4(0) = C\prod_{n=1}^{\infty}(1-q^{2n-1})(1-q^{2n-1})$$

Step One of Proof: We first prove the formula for θ_3. The zeros of this function are at spots where $z = \frac{1}{2} + \frac{\tau}{2} + m + n\tau$ and so

$$p = e^{\pi i z} = e^{\frac{\pi i}{2}} e^{\frac{\pi i \tau}{2}} e^{\pi i m} e^{\pi i n \tau} = i q^{\frac{1}{2}} (-1)^m q^n$$

If this holds, then

$$p^2 = -q^{2n+1}$$

Conversely, if this condition holds, then $p = \pm i q^{\frac{1}{2}} q^n$ and the $\pm$ is given by $(-1)^m$, which can indeed be both 1 and -1. Note that $2n+1$ takes values $\ldots, -5, -3, -1, 1, 3, 5, \ldots$ and so

$$p^2 = -q^{(2n+1)} \text{ or } p^2 = -\left(\frac{1}{q}\right)^{(2n+1)}$$

for $n \geq 1$. Since p and q are never zero, the two possibilities are equivalent to $1 = -q^{2n+1} p^{-2}$ or $1 = -q^{2n+1} p^2$. The function$\left(1 + q^{2n+1} p^{-2}\right)\ \left(1 + q^{2n+1} p^2\right)$ satisfies all of these conditions. Thus to get all zeros, we form the product $\prod \left(1 + q^{2n+1} p^{-2}\right)\ \left(1 + q^{2n+1} p^2\right)$. However, the index n would then run from 0 to ∞. We prefer starting this index at 1, so we form the product

$$\prod_{n=1}^{\infty} \left(1 + q^{2n-1} p^{-2}\right)\ \left(1 + q^{2n-1} p^2\right)$$

This product converges since $|q| < 1$. Consequently, the quotient

$$\frac{\theta_3(z, \tau)}{\prod_{n=1}^{\infty} (1 + q^{2n-1} p^{-2})\ (1 + q^{2n-1} p^2)}$$

is entire and never zero. We will prove it constant in z by showing that the product satisfies the same periodicity conditions as $\theta_3(z)$. We know that $\theta_3(z+1) = \theta_3(z)$ and $\theta_3(z+\tau) = b\theta_3(z)$ where $b = p^{-2} q^{-1}$. We will show that the same formulas hold for our product. The map $z \to z + 1$ sends $e^{\pi i z}$ to $e^{\pi i z} e^{\pi i} = -e^{\pi i z}$, so it sends $p \to -p$. Our product is clearly invariant under this map. The map $z \to z + \tau$ sends $e^{\pi i z} \to e^{\pi i z} e^{\pi i \tau}$ and so sends $p \to pq$. Thus it modifies the product to become

$$\prod_{n=1}^{\infty} \left(1 + q^{2n-3} p^{-2}\right)\ \left(1 + q^{2n+1} p^2\right)$$

Looking at the first of the two terms, we find that the new product has an extra term $(1 + q^{-1} p^{-2})$. Looking at the second of the two terms, we find that the new product has a missing term $(1 + qp^2)$. Hence to get the new product from the original product, we must multiply by

$$\frac{(1 + q^{-1} p^{-2})}{(1 + qp^2)} = \frac{(1 + b)}{(1 + b^{-1})} = b$$

We conclude that

$$\theta_3(z,\tau) = C(\tau)\prod_{n=1}^{\infty}\left(1+q^{2n-1}p^{-2}\right)\left(1+q^{2n-1}p^{2}\right)$$

where C is constant in z and depends only on τ.

Step Two of Proof: We get the remaining series using a trick, which has the advantage of showing that we get the same $C(\tau)$ in all cases. Recall that $\theta_3(z+\frac{1}{2}) = \theta_4(z)$, $\theta_3(z+\frac{\tau}{2}) = a\theta_2(z)$, and $\theta_3(z+\frac{1}{2}+\frac{\tau}{2}) = ia\theta_1(z)$. If we can prove that the products transform in the same way, the remaining formulas immediately follow.

Consider the first case, $\theta_3(z+\frac{1}{2}) = \theta_4(z)$. Then $e^{\pi iz} \to e^{\pi iz}e^{\frac{\pi i}{2}}$ and so $p \to ip$. The product for $\theta_3(z)$ is $C\prod_{n=1}^{\infty}(1+q^{2n-1}p^2)(1+q^{2n-1}p^{-2})$ and when we replace p by ip, this changes to $C\prod_{n=1}^{\infty}(1-q^{2n-1}p^2)(1-q^{2n-1}p^{-2})$, which is the product for $\theta_4(z)$.

Consider the second case, $\theta_3(z+\frac{\tau}{2}) = a\theta_2(z)$, where $a = p^{-1}q^{-\frac{1}{4}}$. Then $e^{\pi iz} \to e^{\pi iz}e^{\pi i\frac{\tau}{2}}$ and so $p \to pq^{\frac{1}{2}}$. The product for $\theta_3(z)$ is $C\prod_{n=1}^{\infty}(1+q^{2n-1}p^2)(1+q^{2n-1}p^{-2})$ and when we replace p by $pq^{\frac{1}{2}}$, this changes to $C\prod_{n=1}^{\infty}(1+q^{2n-1}p^2q)(1+q^{2n-1}p^{-2}q^{-1})$ which equals

$$C\prod_{n=1}^{\infty}(1+q^{2n}p^2)(1+q^{2n-2}p^{-2})$$

Compare this to the product for $\theta_2(z)$,

$$Cq^{\frac{1}{4}}(p+p^{-1})\prod_{n=1}^{\infty}(1+q^{2n}p^2)(1+q^{2n}p^{-2})$$

Our product has one extra term, $(1+q^0p^{-2}) = (1+p^{-2})$. To get the second product from ours, we must divide out this term and then multiply by $q^{\frac{1}{4}}(p+p^{-1})$, or equivalently multiply by

$$\frac{q^{\frac{1}{4}}(p+p^{-1})}{(1+p^{-2})} = \frac{q^{\frac{1}{4}}p(1+p^{-2})}{(1+p^{-2})} = q^{\frac{1}{4}}p = a^{-1}$$

So $a^{-1}\theta_3(z+\frac{\tau}{2}) = \theta_2(z)$, which is the desired $\theta_3(z+\frac{\tau}{2}) = a\theta_2(z)$.

The final formula is $\theta_3(z+\frac{1}{2}+\frac{\tau}{2}) = ia\theta_1(z)$. Then $e^{\pi iz} \to e^{\pi iz}e^{\pi i\frac{1}{2}}e^{\pi i\frac{\tau}{2}}$ or $p \to piq^{\frac{1}{2}}$. Inserting this in the product for θ_3 gives

$$C\prod(1-q^{2n}p^2)(1-q^{2n-2}p^{-2})$$

Compare this to the product

$$i\theta_1(z) = Cq^{\frac{1}{4}}(p-p^{-1})\prod_{n=1}^{\infty}(1-q^{2n}p^2)(1-q^{2n}p^{-2})$$

To get from the top product to the bottom product, we must divide out the term $(1-q^0p^{-2})$ on top, and multiply by the extra $q^{\frac{1}{4}}(p-p^{-1})$ on the bottom, which involves multiplying by

$$\frac{q^{\frac{1}{4}}(p-p^{-1})}{1-p^{-2}} = \frac{q^{\frac{1}{4}}p(1-p^{-2})}{1-p^{-2}} = q^{\frac{1}{4}}p = a^{-1}$$

So $a^{-1}\theta_3(z+\frac{1}{2}+n\frac{\tau}{2}) = i\theta_1(z)$, or

$$\theta_3(z+\frac{1}{2}+n\frac{\tau}{2}) = ia\theta_1(z)$$

Step Three of Proof: Finally we evaluate $C(\tau)$. In the four products, take $z=0$ and $p=1$ to produce the following results:

$$\pi^{-1}\theta_1'(0) = 2Cq^{\frac{1}{4}}\prod(1-q^{2n})^2$$

$$\theta_2(0) = 2Cq^{\frac{1}{4}}\prod(1+q^{2n})(1+q^{2n})$$
$$\theta_3(0) = C\prod(1+q^{2n-1})(1+q^{2n-1})$$
$$\theta_4(0) = C\prod(1-q^{2n-1})(1-q^{2n-1})$$

In the first formula, $i\theta_1(z) = Cq^{\frac{1}{4}}(p-p^{-1})\prod_{n=1}^{\infty}(1-q^{2n}p^2)(1-q^{2n}p^{-2})$, when we set $z=0$ and $p=1$ we get $0=0$ because $\theta_1(0)=0$ and $p-p^{-1}=0$. We must move $p-p^{-1}$ to the other side and take the limit as $z\to 0$ to get something interesting. Write $\theta_1(z) = c_1z+c_2z^2+...$ and $p-p^{-1} = e^{\pi iz}-e^{-\pi iz} = (1+\pi iz+...)-(1-\pi iz+...) = 2\pi iz+...$ to get the quotient

$$\frac{i\theta_1(z)}{p-p^{-1}} = \frac{i(ic_1z+...)}{2\pi iz+...} \to \frac{ic_1}{2\pi i} = \frac{\theta_1'(0)}{2\pi}$$

We now apply the result $\theta_1'(0) = \pi\theta_2(0)\theta_3(0)\theta_4(0)$. This gives

$$2Cq^{\frac{1}{4}}\prod(1-q^{2n})^2 = 2Cq^{\frac{1}{4}}C^2\prod(1+q^{2n})^2\prod(1+q^{2n-1})^2\prod(1-q^{2n-1})^2$$

or

$$\prod(1-q^{2n})^2 = C^2\prod(1+q^{2n})^2\prod(1-q^{4n-2})^2$$

Step Four of Proof: Let $f(q) = \prod(1+q^n)^2\prod(1-q^{2n-1})^2$ and notice that the right side of the above equation is $C^2 f(q^2)$. We claim that $f(q) = f(q^2)$. Indeed

$$f(q^2) = \prod(1+q^{2n})^2\prod(1-q^{4n-2})^2 = \prod(1+q^{2n})^2\prod(1-q^{2n-1})^2\prod(1+q^{2n-1})^2 =$$

$$\left[\prod(1+q^{2n})^2\prod(1+q^{2n-1})^2\right]\prod(1-q^{2n-1})^2 = \left[\prod(1-q^n)^2\right]\prod(1-q^{2n-1})^2 = f(q)$$

But then $f(q) = f(q^2) = f(q^4) = ...$ Note that $q^n \to 1$ as $n \to \infty$, so all of these values including $f(q^2)$ are 1. Consequently $\prod(1-q^{2n})^2 = C^2$ and $C = \pm\prod(1-q^{2n})$.

Let $\tau \to i\infty$ so $q \to 0$. Then $C(\tau)$ approaches 1 or -1 depending on the sign of the above equation. Moreover, $\theta_3(\tau) = C\prod(1+q^{2n-1})(1+q^{2n-1})$ approaches 1 or -1 as well. On the other hand, $\theta_3(0) = sum_{n=-\infty}^{\infty} q^{n^2}$ approaches 1. So the sign is positive and $C = \prod(1-q^{2n})$. QED.

26.5 Jacobi's Triple Product

The product formula for $\theta_3(z)$ is particularly simple and has many applications in number theory. In Hardy and Wright's book on number theory, an elementary proof of the theorem is given.

Theorem 196 (Jacobi's Triple Product)**.**

$$\theta_3(z) = \sum_{n=-\infty}^{\infty} p^{2n}q^{n^2} = \prod_{n=1}^{\infty}(1-q^{2n})^2(1+q^{2n-1}p^2)(1+q^{2n-1}p^{-2})$$

26.6 Euler's Pentagonal Number Theorem

We end with three applications of the previous results to number theory.

26.6.1 Partitions

A *partition* of a positive integer n is a sum of positive integers with sum n. Order does not matter, and repetitions are allowed. Let $p(n)$ be the number of partitions of n. By convention we set $p(0) = 1$. Then $p(1) = 1, p(2) = 2, p(3) = 3, p(4) = 5, p(5) = 7$. For instance, the partitions of 5 are

$$5, 4+1, 3+2, 3+1+1, 2+2+1, 2+1+1+1, 1+1+1+1+1$$

It is not a good idea to compute $p(n)$ for larger n by writing down all the partitions because $p(100) = 190,569,292$.

26.6.2 Euler's First Formula

While there is no formula for $p(n)$, there is a recursive way to compute its values. This recursive procedure was summarized by Euler in the following beautiful formula, which can be taken as a formal product so we can ignore convergence matters:

$$\sum_{n=0}^{\infty} p(n)X^n = (1 + X + X^2 + ...)(1 + X^2 + X^4 + ...)(1 + X^3 + X^6 + ...)(...)$$

The proof is easy. To compute such an expression, we take one term from the first sum, one term from the second, one term from the third, etc., and multiply them. Then we sum over all such possibilities. So we have

$$\sum X^{n_1+2n_2+3n_3+...} = \sum (\text{number of such representations of } n)\, X^n$$

However, each term $n = n_1 + 2n_2 + 3n_3 + ...$ describes a partition of n with n_1 ones and n_2 twos and n_3 threes and so forth.

26.6.3 Another Interpretation

Euler's first formula describes a way to organize a computation of p_n. This method can also be described in a manner that doesn't use algebra. Notice that the term $1 + x + x^2 + ...$ in Euler's product counts partitions containing only 1; each integer can be written as such a sum in only one way. The product

$$(1 + x + x^2 + ...)(1 + x^2 + x^4 + ...)$$

counts partitions containing only 1's and 2's and thus equals

$$1 + x + 2x^2 + 2x^3 + 3x^4 + 3x^5 + 4x^6 + ...$$

because 0 and 1 have no extra partitions with 2's, 2 and 3 have one additional partition with 2's, 4 and 5 have two additional partitions with 2's, etc.

In a similar way we can count partitions using 1's, 2's, and 3's, and then partitions using 1's, 2's, 3's, and 4's, and so forth. Let's consider one of these cases in detail. It turns out that

$$(1 + x + x^2 + ...)(1 + x^2 + x^4 + ...)(1 + x^3 + x^6 + ...)(1 + x^4 + x^8 + ...) =$$

$$1 + x + 2x^3 + 3x^3 + 5x^4 + 6x^5 + 9x^6 + 11x^7 + 15x^8 + 18x^9 + 23x^{10} + 27x^{11} + 34x^{12} + 39x^{13} + ...$$

where the coefficient of x^n counts partitions of n containing only 1's, 2's, 3's, and 4's. Suppose we now want to count partitions of n containing only 1's, 2's, 3's, 4's, and 5's. Let's concentrate on the case $n = 13$. From the above product we see that there are 39 such partitions with no 5's. A partition containing exactly one 5 will contain 1's, 2's, 3's and 4's adding up to $13 - 5 = 8$, and the above product shows that there are 18 such partitions. A partition of 13 containing exactly two 5's will contain 1's, 2's, 3's, and 4's adding up to $13 - 5 - 5 = 3$ and there are 3 such partitions. There are no partitions of 13 with four or more 5's. So the total number of partitions of 13 with 1's, 2's, 3's, 4's, and 5's is $39 + 15 + 3 = 57$.

Euler's first formula is just a fancy way to summarize this technique. To count all partitions of n, first count all partitions of all $k <= n$ containing only 1's, and then count all partitions of all $k <= n$ containing only 1's and 2's, and then count all partitions of all $k <= n$ containing only 1's, 2's, and 3's, and continue in this way up to partitions containing only 1's, 2's, 3's, ..., n's. If $a_0, a_1, a_2, ...$ are counts of partitions with summands less than t, then the $b_0, b_1, b_2, ...$ counting partitions with summands less than or equal to t are given by the formula

$$b_k = a_k + a_{k-t} + a_{k-2t} + ...$$

26.6.4 A Program

If you are a programmer, you'll understand this much better by writing a program to do the calculation. Almost any language will do, but you have to remember that the numbers will get large and might overflow.. I'll write a program in *Mathematica* because I have it handy and it can deal with arbitrarily large integers.

```
F[limit_] :=
Block[{N, f, i, j, k, list1, list2},(* local variables *)
N = limit + 1;(* number of series coefficients *)
f[s_] := 1;
list1 = Array[f, N];(* series coefficients *)
list2 = Array[f, N]; (* fill initial list with 1's *)

k = 2;
While[k <= limit,
```

```
        Print[" "];   Print["Partitians using 1 through ", k];
        For[i = 1, i <= N, i++,
             sum = list1[[i]];
             For[j = i - k, j > 0, j = j - k,
                  sum = sum + list1[[j]];];
                  list2[[i]] = sum;
                  ];
        For[i = 1, i <= N, i++, list1[[i]] = list2[[i]];];
        For[i = 1, i <= N, i++, Print[i  - 1, ":  ", list2[[i]]]];
        k++;
        ];
    ];
```

This program takes 50 seconds to compute the first fifty values of p_n. In particular, $p_{50} = 204,226$.

26.6.5 Euler's Second Formula

Recall that $\frac{1}{1-x} = 1 + x + x^2 + x^3 + ...$ This is even true "formally" since

$$(1-x)(1+x+x^2+x^3+...) = (1-x) + (x-x^2) + (x^2-x^3) + ... = 1$$

Theorem 197.

$$\frac{1}{(1-x)(1-x^2)(1-x^3)...} = \sum p_n x^n$$

Proof: This follows immediately from Euler's first formula by taking inverses of the series making up that formula.

26.6.6 The Pentagonal Number Theorem

After writing this formula, Euler multiplied out the denominator by hand, hoping to find a pattern. One of my sources says he multiplied the first fifty terms, while another says he multiplied as many as one hundred terms.. Amazingly, he found that

Theorem 198 (The Pentagonal Number Theorem)**.**

$$(1-x)(1-x^2)(1-x^3)... = 1 - x - x^2 + x^5 + x^7 - x^{12} - x^{15} + x^{22} + x^{26} - x^{35} - x^{40} + ...$$

Notice that pairs of terms with minus signs alternate with pairs with positive signs. Notice that the intervals between the exponents of pairs with the same sign increase by one, then two, then

three, etc. Notice that the intervals between exponents of pairs with opposite signs increase by three, then five, then seven, etc.

The Pentagonal Number Theorem leads to a rapid method of computing the partition numbers. Indeed rewriting theorem 4 using theorem 3 gives

$$(1 - x - x^2 + x^5 + x^7 - x^{12} - x^{15} + ...)(1 + p_1 x + p_2 x^2 + p_3 x^3 + ...) = 1$$

Consequently the coefficient of x^n in the product is zero, and so

$$p_n - p_{n-1} - p_{n-2} + p_{n-5} + p_{n-7} + p_{n-12} + p_{n-15} + ... = 0$$

Each of these expressions is a finite sum because $p_0 = 1$ and $p_k = 0$ for negative k by definition. These formulas then allow us to compute the p_n inductively starting with the value $p_0 = 1$. Thus $p_1 - p_0 = 0$, so $p_1 = 1$. Then $p_2 - p_1 - p_0 = 0$, so $p_2 = 2$. Etc.

Using this revised formula, it takes my computer only a second to find the first 50 values of p_n and only 50 seconds to find the first 1000 values. And indeed, the computer gives $p_{1000} = 24,061467,864,032,622,473,692,149,727,991$.

For the record, I'll show the Mathematica program I used to do this computation. You can easily rewrite this program in your favorite language, but keep in mind that the integers computed by the program will be very large.

```
FPentagonal[limit_] :=
Block[{N, f, P, Pinverse, k, index, n,  i},(* local variables *)
N = limit + 1; (* number of series coefficients *)
f[s_] := 0;
P = Array[f, N];(* the p(n), initially filled with zeros *)
P[[1]] = 1; (* p(0) = p(1) = 1; arrays in Mathematica are one-based *)
P[[2]] = 1;
Pinverse = Array[f, N];(* inverse of p(n) series *)
k = 1;(* now fill in Pinverse using the Pentagonal Number Theorem *)
index = k (3 k - 1) / 2;
While[index <= N,
     Pinverse[[index]] = (-1)^k;
     index = k (3 k + 1) / 2;
     If[ index <= N,  Pinverse[[index]] = (-1)^k];
     k = k + 1;
     index = k (3 k - 1) / 2;
     ];
For[n = 2, n < N, n = n + 1,(* compute p(n) as inverse of Pinverse *)
     P[[n ]] = 0;
     For[i = 1, i < n, i = i + 1,
```

```
            P[[n]] = P[[n]] -  P[[n - i]] Pinverse[[i]]
            ];
        P[[n]] = P[[n]] - Pinverse[[n]];
        Print[n , ":  ", P[[n]]];
        ]
    ]
```

I don't know how many values of p_n were computed by Euler. In 1918, MacMahon in England computed the first 200 values of p_n. This table was extended to 600 by Gupta in 1935, and to 1000 by Gupta, Gwyther and Miller in 1958. I don't know if a computer was used for this final table.

Ramanujan was in England when MacMahon computed these values, and Ramanujan noticed that $p_4 = 5$ and every fifth number after that is divisible by 5. Also $p_5 = 7$ and every seventh entry after that is divisible by 7. Also $p_6 = 11$ very eleventh entry after that was also divisible by 11. But the pattern stops there. Ramanujan eventually proved these results, but the proof is not easy.

The interaction between MacMahon and Ramanujan is shown in the movie *The Man Who Knew Infinity* starring Dev Patel and Jeremy Irons.

Why is the theorem called "the pentagonal number theorem"? The reason isn't very interesting mathematically, but here it is. Exponents in the pentagonal series with nonzero coefficients have the form $n = \frac{k(3k-1)}{2}$ and $n = \frac{k(3k+1)}{2}$. The numbers $\frac{k(3k-1)}{2}$ are called "pentagonal numbers" because they count the number of dots in a pentagonal pattern, just as the numbers $n = \frac{k(k+1)}{2}$ and $n = k^2$ are "triangular numbers" and "square numbers" because they count dots in triangular and square patterns.

26.6.7 Proving the Pentagonal Number Theorem

Jordan Bell wrote an interesting paper on the history of the Pentagonal Number Theorem. It can be found at http://arxiv.org/pdf/math/0510054v2.

The first mention of the theorem is in a letter from Daniel Bernoulli to Euler on January 28, 1741. Bernoulli is replying to a (lost) letter from Euler about the expansion, and he writes "The other problem, to transform $(1-x)(1-x^2)(...)$ into $1-x-x^2+x^5+...$, follows easily by induction, if one multiplied many factors. The remainder of the series I do not see. This can be shown in a most pleasant investigation, together with tranquil pastime and the endurance of pertinacious labor, all three of which I lack."

Euler mentions the theorem many more times over the next few years, in letters we do possess to Niklaus Bernoulli, Christian Goldbach, d'Alembert, and others, and in the first publication

of 1751. (This paper was written on April 6, 1741 and had no proof. Euler wrote so many papers that the publishers fell dramatically behind; they were publishing new papers many years after his death.) A typical entry, from a letter to Goldbach, reads "If these factors $(1-n)(1-n^2)(1-n^3)$ etc. are multiplied out onto infinity, the following series $1-n-n^2+n^5+n^7-$ etc is produced. I have however not yet found a method by which I could prove the identity of these two expressions. The Hr. Prof. Niklaus Bernoulli has also been able to prove nothing beyond induction." Here the word "induction" means "by experiment" rather than "a proof by induction".

Euler is not above a little trickery. Learning that d'Alembert wanted to leave mathematical research to regain his health, he wrote him "If in your spare time you should wish to do some research which does not require much effort, I will take the liberty to propose the expression $(1-x)(1-x^2)(1-x^3)(1-x^4)$ etc., which upon expansion by multiplication gives the series $1-x-x^2+x^5+x^7-$ etc., which would seem very remarkable to me because of the law which we easily discover within it, but I do not see how his law may be deduced without induction of the proposed expression." Eventually d'Alembert wrote back "regarding the series of which you have spoken, it is very peculiar, but I only see induction to show it. But no one is deeper and better versed on such matters than you."

Euler finally was able to prove the theorem on June 9, 1750, in a letter to Goldbach. His proof is algebraic. The proof was first published in 1760, and Euler gives more details about points which were vague in his letter to Goldbach. You can consult Bell's paper if you want to follow this original Euler proof.

26.6.8 Franklin's Proof

In 1881, the American mathematician Franklin gave a proof which involves no algebra at all. Hans Rademacher called this proof "the first major achievement of American mathematics." Franklin's proof is described in Hardy and Wright's book on number theory; it is short and beautiful.

26.6.9 Jacobi and the Pentagonal Number Theorem

Surprisingly, the Pentagonal Number theorem is a special case of Jacobi's Triple Product. Indeed consider the product $\prod(1-q^{2n})(1+p^2q^{2n-1})(1+p^{-2}q^{2n-1})$. Replace q by $q^{\frac{3}{2}}$ to get $\prod(1-q^{3n})(1+p^2q^{3n-\frac{3}{2}})(1+p^{-2}q^{3n-\frac{3}{2}})$. Now replace p by $\frac{i}{q^{\frac{1}{4}}}$ to get

$$\prod(1-q^{3n})(1-q^{3n-2})(1-q^{3n-1}) = \prod(1-q^n)$$

These changes map $\sum p^{2n}q^{n^2}$ to $\sum p^{2n}q^{\frac{3n^2}{2}}$ and on to $\sum(-1)^n q^{\frac{n}{2}}q^{\frac{3n^2}{2}} = \sum(-1)^n q^{\frac{1}{2}n(3n-1)}$.

These are sums from $-\infty$ to ∞. When $n = 0$, we get one. The remaining terms are paired. We get

$$1 + \sum_{n=1}^{\infty} (-1)^n \left(q^{\frac{(-n)(-3n-1)}{2}} + q^{\frac{(n)(3n-1)}{2}} \right) = 1 + \sum (-1)^n \left(q^{\frac{n(3n-1)}{2}} + q^{\frac{n(3n+1)}{2}} \right)$$

It is easy to check that this final sum is $1 - q - q^2 + q^5 + q^7 - q^{12} - q^{15} + q^{22} + ...$

Remark: It is possible to go backward and prove Jacobi's triple product identity from the Pentagonal Number theorem. That was Jacobi's original approach.

26.7 Sums of Two Squares

26.7.1 Introduction

Since $\theta_3(0) = \sum q^{n^2}$, $\theta_3(0)^k = \sum q^{n_1^2+n_2^2+...+n_k^2} = \sum r_k(n) q^n$ where $r_k(n)$ is the number of ways that n can be written as a sum of k squares. It is sometimes possible to obtain formulas for $r_k(n)$ from the theory of theta functions. Such formulas were first obtained by Jacobi and astonished the mathematical world.

26.7.2 Elementary Results

Writing integers as a sum of two squares has an ancient history. The formula $(a^2+b^2)(c^2+d^2) = (ac - bd)^2 + (ad + bc)^2$, which is a consequence of $(a + ib)(c + id) = (ac - bd) + i(ad + bc)$ but was known long before the invention of complex numbers, shows that if two numbers are sums of squares, so is their product. Consequently, it is important to determine when a prime can be written as a sum of squares. Of course $2 = 1^2 + 1^2$.

Theorem 199. *If $p \equiv 3 \pmod 4$, then p is not a sum of two squares.*

Proof: Modulo four, the squares are $0^2 = 0, 1^2 = 1, 2^2 = 0, 3^2 = 1$, and the sum of two of these is $0, 1, 2$ but never 3. QED.

Theorem 200. *If $p \equiv 1 \pmod 4$, then p is a sum of two squares.*

Proof: This is the central theorem of the subject, and several different proofs will be given in the following pages.

Theorem 201. *A positive integer n can be written as a sum of two squares if and only if in the prime factorization of n, each $p \equiv 3 \pmod 4$ occurs an even number of times.*

Proof: The previous results show that this condition is sufficient; in particular $p^2 = p^2 + 0^2$ is a sum of squares. The proof that the condition is necessary requires two easy lemmas.

Lemma 22. *Suppose p is an odd prime. If $p \equiv 3 \pmod 4$, the number -1 is not a perfect square in Z_p.*

Since $Z_p^\star$ is a group of order $p-1$, every a in this group satisfies $a^{p-1} = 1$. If $a = b^2$, then $a^{\left(\frac{p-1}{2}\right)} = b^{p-1} = 1$. When $p = 4k+3$, $\frac{p-1}{2} = 2k+1$, so $(-1)^{\left(\frac{p-1}{2}\right)} = (-1)^{2k+1} = -1$. This is not 1, so -1 is not a square in Z_p.

Lemma 23. *Suppose n is a sum of squares and the prime p divides n. If $p \equiv 3 \pmod 4$, then p occurs in the factorization of n an even number of times.*

Assume $n = a^2 + b^2$ and write this equation in Z_p. Since p divides n, we get $a^2 + b^2 = 0$ or $a^2 = -b^2$. If p does not divide b, then in Z_p we get $\left(\frac{a}{b}\right)^2 = -1$, which is impossible by the previous lemma. Similarly p divides a. So $a = Ap$ and $b = Bp$ and $n = (A^2 + B^2)p^2$. Thus p^2 is a factor of n and n divided by p^2 is still a sum of squares. If p is still a factor of $\frac{n}{p^2}$, we can continue the argument. Eventually p is no longer a factor, so it occurred evenly often in the original n.

Remark: That finishes the proof of the theorem. For later purposes, it is useful to expand the first lemma as follows:

Lemma 24. *Suppose p is an odd prime. For any nonzero element in Z_p, $a^{\frac{p-1}{2}} = \pm 1$*

Proof: Since the non-zero elements of Z_p form a multiplicative group of order $p-1$, any nonzero a satisfies $a^{p-1} = 1$. So the square of $a^{\frac{p-1}{2}}$ is 1, and the polynomial $X^2 - 1 = (X-1)(X+1)$ has only two roots, ± 1.

Lemma 25. *Suppose p is an odd prime. A nonzero element $a \in Z_p$ is a square if and only if $a^{\frac{p-1}{2}} = 1$*

Proof: If a is a square, $a = b^2$ and $a^{\frac{p-1}{2}} = b^{p-1} = 1$. Exactly half of the nonzero elements of Z_p are squares because the map $a \to a^2 : Z_p^\star \to Z_p^\star$ has kernel ± 1, so at least half of the elements satisfy $X^{\frac{p-1}{2}} - 1 = 0$. But this polynomial has at most $\frac{p-1}{2}$ roots. So the nonsquares do not satisfy this equation, and thus must all satisfy $a^{\frac{p-1}{2}} = -1$.

Lemma 26. *The element $-1 \in Z_p$ is a square if and only if $p = 2$ or $p \equiv 1 \pmod 4$.*

Proof: $(-1)^{\frac{p-1}{2}} = 1$ when $p \equiv 1 \pmod 4$ and -1 when $p \equiv 3 \pmod 4$.

26.7.3 Alternate Proofs

Fermat wrote an elaborate description of the theorem that every $p \equiv 1 \pmod 4$ can be written as a sum of two squares in a letter to Mersenne from December 25, 1640. He claims to have

a proof, but it is lost. Euler reconstructed Fermat's work, and after much work announced a proof of the two squares theorem on May 6, 1747. This proof worked by descent; first he proved that some multiple mp, for $m < p$, is a sum of two squares and then he showed that if $m \neq 1$, a smaller m can be found.

Many other proofs are known. In this section we sketch one based on the Gaussian integers, because it gives new insight into the significance of the result.

A Gaussian integer is an expression $a + bi$ with $a, b \in Z$. Our immediate goal is to show that number theory can be done in this new ring, with units, primes, unique factorization, and all the rest.

If A and B are Gaussian integers, we say A divides B and write $A|B$ if $\frac{B}{A}$ is itself a Gaussian integer, or equivalently if there is a Gaussian integer C with $AC = B$. A non-zero Gaussian integer is called a *unit* if it divides all Gaussian integers. In the ordinary integers, the units are ± 1. We have

Theorem 202. *The units in the Gaussian integers are* ± 1 *and* $\pm i$.

Proof: If $A = a + ib$ is a Gaussian integer, define its *norm* to be $N(A) = a^2 + b^2$. Note that $N(AB) = N(A)N(B)$. If A is a unit, then for any B there is a C with $AC = B$ and so $N(A)N(C) = N(B)$. Thus $N(A)$ divides all $N(B)$ and so all $c^2 + d^2$. The only ordinary positive integer with this property is 1, so $N(A) = 1$, and $a^2 + b^2 = 1$. The integer solutions of this are ± 1 and $\pm i$.

Remark: We call two Gaussian integers *associates* if they differ by a multiplicative unit. The same definition makes sense in ordinary integers, where n and $-n$ are associates. Ambiguities caused by this equivalence are often resolved in the integers by dealing only with positive integers. In the Gaussian integers, instead of making a choice we gives results "up to associates."

Definition 59. *A non-zero Gaussian integer is a* prime *if it is not a unit and only divisible by its associates and by units.*

Theorem 203. *Every nonzero Gaussian integer can be written as a product of primes. This product is unique up to order and associates.*

Step One of Proof: In grade school, we learn to divide whole numbers with remainder. Given a and b positive integers, we write $\frac{a}{b} = q + \frac{r}{b}$ where q is the quotient and r is the remainder and $r < b$. So $a = bq + r$ where $0 \leq r < b$.

Lemma 27. *If A and $B \neq 0$ are Gaussian integers, we can find Gaussian integers Q and R with*

$$A = BQ + R$$

and $N(R) < N(B)$

Proof: In the field $Q(i)$, write $\frac{A}{B} = q_1 + q_2 i$ where the $q_i \in Q$. Each q_i is within $\frac{1}{2}$ of an integer, so find integers n_1 and n_2 with $|n_1 - q_1| \leq \frac{1}{2}$ and $|n_2 - q_2| \leq \frac{1}{2}$. Then

$$\frac{A}{B} = (n_1 + n_2 i) + ((q_1 - n_1) + (q_2 - n_2)i) = Q + S$$

where everything in sight is a Gaussian integer except S. Let $R = BS$ so we have

$$A = BQ + BS = BQ + R$$

Since A and BQ are Gaussian integers, so is R. Moreover, $N(R) = N(B)N(S)$ and

$$N(S) \leq \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 < 1$$

QED.

Lemma 28. *Suppose P is a Gaussian prime and $P|AB$. Then $P|A$ or $P|B$.*

Proof: Suppose P does not divide A and consider the ideal generated by P and A. That is, consider the set of all $M_1 P + M_2 A$. Choose a non-zero element G in this ideal with smallest norm. By the previous lemma, $P = GQ + R$ with $N(R) < N(G)$. Both P and GQ are in the ideal, so R is in the ideal, and consequently $R = 0$, so $G|P$. By the same reasoning, $G|A$.

Since P is a prime, G is either a unit or else an associate of P. Since P does not divide A, G cannot be an associate of P, so G is a unit. It follows that the ideal contains all of $Z[i]$ and in particular contains 1.

So $1 = M_1 P + M_2 G$; multiplying by B gives $B = M_1 BP + M_2 BG$. Since $P|BP$ and $P|AB$, $P|B$. QED.

Step Two of Proof of Main Theorem: Let A be a nonzero Gaussian integer. If we can factor $A = A_1 A_2$ with neither a unit, then $N(A) = N(A_1)N(A_2)$ and $1 < N(A_i) < N(A)$. Continue factoring as long as possible. Since the norms of the terms are decreasing, the process must end.

The deeper step is uniqueness. Suppose $A = P_1 P_2 \ldots P_k = Q_1 Q_2 \ldots Q_l$ where the P_i and Q_j are prime. Since $P_1|A$, P_1 divides the product of the Q's and the previous lemma ultimately shows that P_1 divides some Q_i. Renumber so this is Q_1. Since P_1 and Q_1 are both prime, they are associates. Multiplying some other Q_i by a unit makes $P_1 = Q_1$, so they can be cancelled. Continue the argument. QED.

26.7.4 Determination of Gaussian Primes

The previous section was grunge work to show that ordinary number theory works in $Z[i]$. Now the fun begins.

Theorem 204. *An ordinary prime $p \in Z$ belongs to $Z[i]$, but may factor in $Z[i]$.*

- *Every prime in $Z[i]$ can be obtained by factoring some ordinary prime p.*
- *No prime in $Z[i]$ is a factor of two different ordinary primes.*

Proof: If A is a Gaussian integer which is not a unit,

$$A\overline{A} = N(A) = p_1 ... p_k$$

and so A divides this product of ordinary primes and thus some particular prime p. So there is a Gaussian integer B with $AB = p$. Taking norms, $N(A)N(B) = p^2$. Because A is not a unit, $N(A) \neq 1$. There are thus two possibilities. Either $N(A) = p^2$ and $N(B) = 1$ or else $N(A) = N(B) = p$. In either case, A is a factor of p, and since $N(A) = p$ or p^2, p is unique. QED.

Remark: Consider the two cases. If $N(A) = p^2$ and $N(B) = 1$, then B is a unit and the equation $AB = p$ says that A is an associate of p. Since A is a Gaussian prime, so is p.

If $N(A) = N(B) = p$, then $A = a + ib$ and $N(A) = a^2 + b^2 = p = (a + ib)(a - ib) = A\overline{A}$, so p factors into two terms and both are Gaussian primes.

This argument started by selecting a Gaussian prime A. But we could have started by factoring an ordinary prime p and letting A be one of its factors. Consequently, we have proved

Theorem 205. *Let p be an ordinary prime in Z. Exactly one of the following is true, and every Gaussian prime arises in this way, up to associates:*

- *p is a sum of two squares $a^2 + b^2$ and p factors as $(a + ib)(a - ib)$ and both are Gaussian primes.*
- *p is not a sum of two squares and p is itself a Gaussian prime.*

Example 1: Every $p \equiv 3 \pmod 4$ is a Gaussian prime.

Example 2: We have $2 = (1 + i)(1 - i)$. Both are Gaussian primes and they are associates. Indeed

$$1 + i = (i)(1 - i)$$

26.7.5 Gaussian Proof of the Two Squares Theorem

The previous section does not prove the two squares theorem, but it shows that the theorem is is part of a larger picture, that is, determining the primes and ultimately prime ideals in algebraic number fields and obtaining them by factoring principle ideals generated by ordinary primes.

However, we can give a very slick proof of Fermat's theorem once we know about the Gaussian integers. Here is the proof.

Theorem 206. *If $p \equiv 1 \pmod 4$, then p is a sum of two squares, and p factors in the Gaussian integers.*

Proof: Since $(-1)^{\frac{p-1}{2}} = 1$ in Z_p, the number $-1 \in Z_p$ is a square. So there is an integer n with $p|(n^2+1)$. Consequently in $Z[i]$, p divides $(n+i)(n-i)$. If p were itself prime, then it would divide one of these, say $n+i$. But $\frac{n}{p} + \frac{1}{p}i$ is not a Gaussian integer. So p is not a Gaussian prime, and the previous theorem asserts that it is a sum of two squares. QED.

26.7.6 The One Sentence Proof:

In the 1990 American Mathematical Monthly, Don Zagier published a one sentence proof of Fermat's theorem. Here is that sentence:

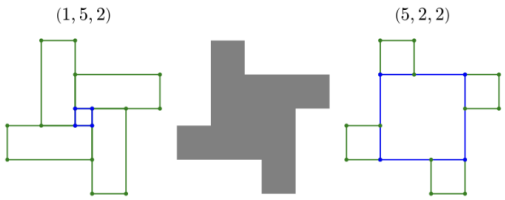
> The involution on the finite set $S = \{(x,y,z) \in N^3 : x^2 + 4yz = p\}$ defined by
>
> $$(x,y,z) \to \begin{cases} (x-2z, z, y-x-z) & \text{if } x < y-z \\ (2y-x, y, x-y+z) & \text{if } y-z < x < 2y \\ (x-2y, x-y+z, y) & \text{if } x > 2y \end{cases}$$
>
> has exactly one fixed point, so $|S|$ is odd and the involution defined by $(x,y,z) \to (x,z,y)$ also has a fixed point. □

After this paper was published, Christian Elsholtz wrote a paper explaining the motivation for Zagier's proof, and still later this motivation was explained using "windmill theory" by A. Spivak in mathematical lecture notes from the mathematical circle in Moscow State University in 2007. There are now several UTube videos giving this proof. Here is a brief sketch.

If an odd prime $p = a^2 + b^2$, then one of a and b is even and one is odd. So we can write $p = a^2 + 4bc$. We will study solutions of this more general statement in positive integers (a,b,c), and the order of these integers is important. Notice that many solutions come in pairs, since (a,b,c) and (a,b,c) are both solutions. If these are the same, then our solution is (a,b,b) and $p - a^2 + 4b^2 = a^2 + (2b)^2$, which is what we want.

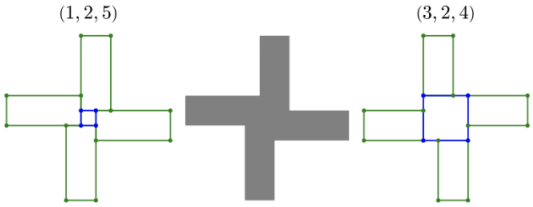
We now replace p by an arbitrary integer n and study triples (a,b,c) with $n = a^2 + 4bc$ from an entirely different point of view. It will turn out that when n is prime, the number of solutions is odd. This finishes the proof by the previous paragraph.

The data gives us a square with side a, and four rectangles with dimensions $b \times c$. We use these objects to create a windmill as follows. Without turning it, glue a $b \times c$ rectangle to the top of the square so the left sides line up. Then symmetrically glue the other three rectangles around the remaining sides. See the picture below.



In the particular case shown above, two different triples give the same windmill, one with a small square and large wings, and one with a large square and small wings. It turns out that this is usually true and Zagier's map explains how to go from one triple to the other. We call this the "two interior squares" theorem for windmills.

There are a few complications, and one is that the map sometimes sends a windmill to its mirror image, formed by gluing the first wing so the right sides line up rather than the left. Both windmills correspond to the same triple (a, b, c) so we agree to call these windmills equivalent. See the example below.



There are exceptional cases where the two interior squares theorem fails. It can happen that the two squares merge together and there is only one way to get the windmill. Or the large square can grow so large that one windmill is a perfect square with no wings. Or the interior square can shrink to a single point. But many of these exceptions are impossible when n is a prime. To complete the proof, it is necessary to show that only one exception survives in this case.

With these hints, you can probably construct the complete proof yourself, perhaps reducing the pictures to algebra, or perhaps staying with pictures for the entire journey.

26.7.7 Jacobi's Proof

In 1829, Jacobi proved the following beautiful result:

Theorem 207. *Suppose n is a positive integer. Let d_1 be the number of divisors $d|n$ such that $d \equiv 1 \pmod 4$ and let d_3 be the number of $d|n$ such that $d \equiv 3 \pmod 4$. Then the number of ways that n can be written as a sum of two squares is*

$$4(d_1 - d_3)$$

Remark: This is a count of ordered pairs (a, b) where $a, b \in Z$ and $a^2 + b^2 = n$. So for example there are eight possible (a, b) giving 5, namely $(\pm 1, \pm 2)$ and $(\pm 2, \pm 1)$.

Example 1: Suppose p is prime and $p \equiv 3 \pmod 4$. The divisors of p are 1, p. One is congruent to 1 mod 4 and the other isn't, so Jacobi's formula gives 0; p is not a sum of two squares.

Example 2: Suppose p is prime and $p \equiv 1 \pmod 4$. Both divisors 1, p are congruent to 1 mod 4, so Jacobi's formula gives 8. So $p = a^2 + b^2$ and the eight possibilities are $(\pm a, \pm b)$ and $(\pm b, \pm a)$. Thus p is a sum of two squares and the representation is essentially unique.

Example 3: Suppose $n = p^2$ where $p \equiv 1 \pmod 4$ and p is prime. The divisors are 1, p, p^2 and each is congruent to 1 mod 4. So Jacobi's formula gives 12. One possibility is $p^2 = p^2 + 0^2$, and this accounts for 4 of the 12 since we can replace p by $-p$ or write $0^2 + p^2$. The other possibility is obtained by writing $p = a^2 + b^2 = (a + ib)(a - ib)$ and so $p^2 = (a + ib)^2(a - ib)^2 = \big((a^2 - b^2) + 2abi\big)\big((a^2 - b^2) - 2abi\big) = (a^2 - b^2)^2 + (2ab)^2$. This accounts for 8 of the 12. So there are essentially two representations. For example, $13 = 2^4 + 3^2$, and therefore either $13^2 = 13^2 + 0^2$ or $13^2 = (3^2 - 2^3)^2 + (2 \cdot 2 \cdot 3)^2 = 5^2 + 12^2$.

Proof: The proof follows from previous formulas about theta functions. First we sketch the argument. We proved

$$\sqrt{\wp(z) - e_3} = \frac{\theta'(0)\theta_4(z)}{\theta_4(0)\theta_1(z)}$$

Substituting $z = \frac{1}{2}$ in this formula gives $\sqrt{e_1 - e_3}$ and another of our formulas gives

$$\sqrt{e_1 - e_3} = \pi\theta_3^2(0)$$

Since $\theta_3(0) = \sum_{-\infty}^{\infty} q^{n^2}$,

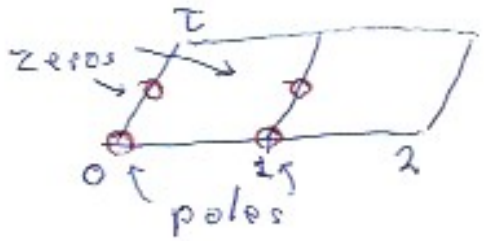
$$\theta_3(0)^2 = \sum q^{n_1^2 + n_2^2} = \sum_0^\infty \tau_2(n) q^n$$

We conclude that

$$\left.\frac{\theta'(0)\theta_4(z)}{\theta_4(0)\theta_1(z)}\right|_{z=\frac{1}{2}} = \pi \sum_{-\infty}^{\infty} \tau_2(n) q^n$$

We find another formula for the left side by studying $f(z) = \frac{\theta'(0)\theta_4(z)}{\theta_4(0)\theta_1(z)}$. Since $\theta_1(z+\tau) = -b\theta_1(z)$ and $\theta_4(z+\tau) = -b\theta_4(z)$, our function is periodic with period τ. Since $\theta_1(z+1) = -\theta_1(z)$ and $\theta_4(z+1) = \theta_4(z)$, our function is periodic with period 2. So it is doubly-periodic with periods $2, \tau$.

Recall that θ_1 has simple zeros at lattice points, so our function has simple poles at 0 and 1. Since θ_4 has simple zeros at $L + \frac{\tau}{2}$, our function has simple zeros at $\frac{\tau}{2}$ and $1 + \frac{\tau}{2}$. See the fundamental parallelogram below.



Let us compute the residue at the origin. Near the origin we have

$$f(z) = \frac{\theta'(0)\theta_4(0)}{\theta_4(0)\Big(0 + \theta'(0)z + \frac{\theta''(0)}{2}z^2 + \ldots\Big)} = \frac{1}{z} + \ldots$$

The residue at $z = 1$ must then be -1 because the sum of the residues of a doubly-periodic function is zero.

Notice that $f(z)$ satisfies all of our general theorems about doubly-periodic functions. It has two poles of order 1, and thus two zeros of order 1. The sum of the points where there are poles or zeros is a lattice point. The sum of the residues is zero. Finally, $f(z+1) = -f(z)$, so the left half of the period parallelogram gives one value of $\sqrt{\wp(z) - e_3}$ and the right half gives the other value and the square root is thus well defined and doubly -periodic on the entire complex plane in this ingenious way.

We are going to decompose $f(z)$ using partial fractions. Since half of its poles have residue 1 and half have residue -1, we naively expect that

$$f(z) = \sum_{m,n} \frac{1}{z - (2n + m\tau)} - \sum_{m,n} \frac{1}{z - ((2n-1) + m\tau}$$

We worry about convergence of this sum, and we also worry that we are equating it to $\sum \tau_2(n)q^n$, so a sum in p, q would be preferable to a sum in z, τ. Both of these worries can be handled at once by converting everything to the p, q language.

The poles with residue one occur when $z = 2m + n\tau$ and so $p = e^{\pi i z} = e^{2\pi i m} e^{\pi i \tau n} = q^n$. The poles with residue minus one occur when $z = (2m+1)+n\tau$ and so $p = e^{\pi i z} = e^{\pi i (2m+1)} e^{\pi i \tau n} = -q^n$. Since $p = q^n$ exactly when $1 = \frac{q^n}{p}$ and $p = -q^n$ exactly when $1 = -\frac{qp^n}{p}$, we are tempted to write

$$f(z) = \sum_{-\infty}^{\infty} \frac{1}{1 - \frac{q^n}{p}} - \sum_{-\infty}^{\infty} \frac{1}{1 + \frac{q^n}{p}}$$

We need to convince ourselves that this new formula still produces simple poles with the correct residues. Note that $\frac{q^n}{p} = e^{\pi i(-z+\tau n)}$ and so

$$\frac{1}{1 - \frac{q^n}{p}} = \frac{1}{1 - (1 + \pi i(-z + 2m + \tau n) + ...)} = \frac{1}{(\pi i(z - 2m - \tau n) + ...)} = \frac{1}{\pi i}\left(\frac{1}{z - 2m - n\tau} + ...\right)$$

The same calculation for the other term gives

$$\frac{1}{1 + \frac{q^n}{p}} = \frac{1}{1 + e^{\pi i(-z+n\tau)}} = \frac{1}{1 - e^{\pi i(-z+(2m+1)+n\tau)}} = \frac{1}{1 - (1 - \pi i(-z + (2m+1) + \tau n) + ...)} =$$

$$\frac{1}{-\pi i(z - (2m+1) - \tau n) + ...} = \frac{-1}{\pi i}\left(\frac{1}{z - (2m+1) - n\tau} + ...\right)$$

Notice that $p = e^{\pi i z} = e^{\pi i (z+2m)}$ and thus any expression involving p and q will be periodic in z with period 2. Moreover, the above expression is also periodic in z with period τ. Indeed $e^{\pi i(z+\tau)} = e^{\pi i z} e^{\pi i \tau} = pq$, so it suffices to show that the expression is unchanged if p is replaced by pq. But this replacement sends

$$\sum_{-\infty}^{\infty} \frac{1}{1 - \frac{q^n}{p}} \to \sum_{-\infty}^{\infty} \frac{1}{1 - \frac{q^{n-1}}{p}}$$

and since we are summing over all n, the sum is unchanged. A similar argument makes the second term of $f(z)$ invariant.

So both $f(z)$ and

$$\pi i \left(\sum_{-\infty}^{\infty} \frac{1}{1 - \frac{q^n}{p}} - \sum_{-\infty}^{\infty} \frac{1}{1 + \frac{q^n}{p}} \right)$$

are doubly periodic with periods 2, τ and have the same residues at all poles, and the difference between $f(z)$ and this expression is a constant. We will show that this constant is zero, so $f(z)$ is given by our partial fraction expansion.

First we simplify our sum using the formula $\frac{1}{1-x} - \frac{1}{1+x} = \frac{2x}{1-x^2}$. so

$$= \pi i \left(\sum_{-\infty}^{\infty} \frac{1}{1 - \frac{q^n}{p}} - \sum_{-\infty}^{\infty} \frac{1}{1 + \frac{q^n}{p}} \right) = 2\pi i \left(\sum_{-\infty}^{\infty} \frac{q^n p^{-1}}{1 - q^{2n} p^{-2}} \right)$$

Let us evaluate $f(z)$ and this expression at $\frac{\tau}{2}$. Since $\theta_4\left(\frac{\tau}{2}\right) = 0$, f vanishes at this point. Also $p = e^{\pi i \frac{\tau}{2}} = q^{\frac{1}{2}}$, so the right side becomes

$$2\pi i \left(\sum_{-\infty}^{\infty} \frac{q^{n-\frac{1}{2}}}{1 - q^{2n-1}} \right)$$

In this sum, we can pair the $n = 0$ and $n = 1$ terms, air the $n = -1$ and $n = 2$ terms, and so forth. For instance, the first pair contains $\frac{q^{-\frac{1}{2}}}{1-q^{-1}}$ and $\frac{q^{\frac{1}{2}}}{1-q^1}$. Writing $X = q^{-\frac{1}{2}}$, the first term has $\frac{X}{1-X^2}$ and the second term has $\frac{X^{-1}}{1-X^{-2}} = -\frac{X}{1-X^2}$ and the two terms cancel out. Indeed all terms cancel out and the expression is zero. So both $f(z)$ and this expression vanish at $\frac{\tau}{2}$ and thus they are equal everywhere.

We now have

$$f(z) = 2\pi i \left(\sum_{-\infty}^{\infty} \frac{q^n p^{-1}}{1 - q^{2n} p^{-2}} \right) = 2\pi i \left(\sum_{n=0}^{\infty} \frac{q^n p^{-1}}{1 - q^{2n} p^{-2}} + \sum_{n=1}^{\infty} \frac{q^{-n} p^{-1}}{1 - q^{-2n} p^{-2}} \right) =$$

$$2\pi i \left(\sum_{n=0}^{\infty} \frac{q^n p^{-1}}{1 - q^{2n} p^{-2}} - \sum_{n=1}^{\infty} \frac{q^n p}{1 - q^{2n} p^2} \right)$$

where the second half is corect because $\frac{X}{1-X^2} = -\frac{\frac{1}{X}}{1-\left(\frac{1}{X}\right)^2}$.

Following our plan at the start of the proof, we now set $z = \frac{1}{2}$ and so $p = e^{\pi i \frac{1}{2}} = i$. This gives

$$f(\frac{1}{2}) = \pi \sum_{n=0}^{\infty} \tau_2(n) q^n = 2\pi \left(\sum_{n=0}^{\infty} \frac{q^n}{1 + q^{2n}} + \sum_{n=1}^{\infty} \frac{q^n}{1 + q^{2n}} \right)$$

or

$$\sum_{n=0}^{\infty} \tau_2(n) q^n = 1 + 4 \sum_{n=1}^{\infty} \frac{q^n}{1 + q^{2n}} = 1 + 4 \sum_{n=1}^{\infty} q^n (1 - q^{2n} + q^{4n} - q^{6n} + ...)$$

and so

$$\sum_{n=1}^{\infty} \tau_2(n) q^n = 1 + 4 \sum_{n=1}^{\infty} (q^n - q^{3n} + q^{5n} - q^{7n} + ...)$$

On the right, how can we get a constant times q^N? It could come from q^n, and of course 1 divides n and $n = N$. It could come from q^{3n} and then 3 divides N and $n = \frac{N}{3}$. It could come from q^{5n} but then 5 divides N and $n = \frac{N}{5}$. And so forth. Adding up all of these possibilities, we get something when d is a divisor of N and $d \equiv 1 \pmod 4$ or d is a divisor of N and $d \equiv 3 \pmod 4$. All divisors of the first type come with a plus sign, and all those of the second type come with a minus sign. So our sum is

$$4 \sum_{N=1}^{\infty} (d_1(N) - d_3(N))\, q^N$$

26.8 Sums of Four Squares

In 1834, Jacobi proved

Theorem 208. *Suppose n is a positive integer. Then $\tau_4(n)$ is* 8 *times the sum of all divisors of n if n is odd, and* 24 *times the sum of all odd divisors of n if n is even. Said another way, $\tau_4(n)$ is the* 8 *times the sum of all divisors of n that are not divisible by* 4.

Since this number is always positive, a corollary is that every positive integer can be written as a sum of four squares. This was first proved by Lagrange in 1770.

Remark: Many different proofs of this theorem are available on the internet, some short and mysterious. Many use modular forms introduced in the previous chapter. In the end, the proof below seemed most illuminating to me. It comes from Elliptic Curves by Henry McKean an Victor Moll, which has been my primary source for all of the material on theta functions. The entire book is highly recommended.

Proof: Clearly $\theta_3^4(0) = \sum_{n=0}^{\infty} \tau_4(n) q^n$. We earlier proved that $\theta_3^4(0) = \frac{e_1 - e_3}{\pi^2}$. We will obtain the formula by expanding the right side.

Start with the formula $\sqrt{\wp(z) - e_3} = \frac{\theta_1'(0)}{\theta_4(0)} \frac{\theta_4(z)}{\theta_1(z)}$. Since

$$\wp(z) = \frac{1}{z^2} + \sum \left(\frac{1}{(z - m_1\omega_1 + m_2\omega_2)^2} - \frac{1}{(m_1\omega_1 + m_2\omega_2)^2} \right)$$

the Laurent expansion is

$$\wp(z) = \frac{1}{z^2} + a_1 z + ...$$

and

$$\wp(z) - e_3 = \frac{1}{z^2} - e_3 + ... = \frac{1}{z^2}\left(1 - e_3 z^2 + ...\right) = \frac{1}{z^2}\left(1 - \frac{e_3}{2}z^2 + ...\right)^2$$

so

$$\sqrt{\wp(z) - e_3} = \frac{1}{z} - \frac{e_3}{2}z + ...$$

Expanding the other side,

$$\frac{\theta_1'(0)}{\theta_4(0)}\frac{\theta_4(z)}{\theta_1(z)} = \frac{\theta_4(0) + \frac{\theta_4''(0)}{2}z^2 + ...}{\theta_4(0)} \cdot \frac{\theta_1'(0)}{\theta_1'(0)z + \frac{\theta_1'''(0)}{3!}z^3 + ...}$$

Here we used that $\theta_4(z)$ is an even function and $\theta_1(0)$ is an odd function. The above formula simplifies to the following, where $a_3 = \frac{\theta_1'''(0)}{3!\theta_1'(0)}$ is a constant that ultimately will not matter:

$$\left(1 + \frac{\theta_4''(0)}{2\theta_4(0)}z^2 + ...\right)\left(\frac{1}{z + a_3 z^3 + ...}\right) = \left(\frac{1}{z} + \frac{\theta_4''(0)}{2\theta_4(0)}z + ...\right)\left(1 - a_3 z^2 + ...\right) = \frac{1}{z} + \left(\frac{\theta_4''(0)}{2\theta_4(0)} - a_3\right)z + ...$$

Comparing the two expansions, we obtain

$$e_3 = 2a_3 - \frac{\theta_4''(0)}{\theta_4(0)}$$

Exactly the same calculation can be done starting with $\sqrt{\wp(z) - e_1} = \left(\frac{\theta_1'(0)}{\theta_2(0)}\right)\left(\frac{\theta_2(z)}{\theta_1(z)}\right)$ and gives

$$e_1 = 2a_3 - \frac{\theta_2''(0)}{\theta_2(0)}$$

Subtracting

$$\theta_3(0)^4 = \frac{e_1 - e_3}{\pi^2} = \frac{1}{\pi^2}\left(\frac{\theta_4''(0)}{\theta_4(0)} - \frac{\theta_2''(0)}{\theta_2(0)}\right)$$

Recall next that theta functions satisfy the heat equation: $\theta''(0) = 4\pi i\frac{\partial}{\partial\tau}\theta$. Applying this result gives

$$\theta_3(0)^4 = \frac{4i}{\pi}\left(\frac{\frac{\partial\theta_4}{\partial\tau}}{\theta_4(0)} - \frac{\frac{\partial\theta_2}{\partial\tau}}{\theta_2(0)}\right)$$

Notice that

$$\frac{\partial}{\partial\tau}\theta(q(\tau)) = \frac{\partial\theta}{\partial q}\pi i e^{pii\tau} = \pi i q\frac{\partial\theta}{\partial q}$$

So now letting a prime indicate differentiation with respect to q, we have

$$\theta_3(0)^4 = 4q\left(\frac{\theta_2'(0)}{\theta_2(0)} - \frac{\theta_4'(0)}{\theta_4(0)}\right)$$

Note that the right side is the logarithmetic derivative of $\frac{\theta_2(0)}{\theta_4(0)}$.

We now use the infinite product representations for the first time. We have

$$\frac{\theta_2(0)}{\theta_4(0)} = 2q^{\frac{1}{4}}\frac{\prod(1+q^{2n})^2}{\prod(1-q^{2n-1})^2} = 2q^{\frac{1}{4}}\frac{\prod(1+q^{2n})^2(1-q^{2n})^2}{\prod(1-q^{2n-1})^2(1-q^{2n})^2} = 2q^{\frac{1}{4}}\frac{\prod(1-q^{4n})^2}{\prod(1-q^n)^2}$$

Taking the logarithmetic derivative of this product, we obtain

$$\theta_3(0)^4 = 4q\left(\frac{\frac{1}{4}q^{-\frac{3}{4}}}{q^{\frac{1}{4}}} + \sum\frac{2(1-q^{4n})(-4n)q^{4n-1}}{(1-q^{4n})^2} - \sum\frac{2(1-q^n)(-nq^{n-1})}{(1-q^n)^2}\right)$$

$$= 1 + \left(-\sum\frac{32nq^{4n}}{(1-q^{4n})} + \sum\frac{8(nq^n)}{(1-q^n)}\right)$$

$$= 1 + \left(-\sum_{n=1}^{\infty} n\frac{32q^{4n}}{(1-q^{4n})} + \sum_{n=1}^{\infty} n\frac{8q^n}{(1-q^n)}\right)$$

We conclude that

$$1 + \sum_{n=1}^{\infty}\tau_4(n)q^n = 1 - \sum_{n=1}^{\infty}32n\left(q^{4n}+q^{8n}+q^{12n}+...\right) + \sum_{n=1}^{\infty}8n\left(q^n+q^{2n}+q^{3n}+...\right)$$

Consider the sum on the right, and compute the contribution to the coefficient of q^N. The kth term of the series associated with n contributes something only if $kn = N$, and then it contributes $8n$. So the series associated with n contributes something only when n divides N, and then it contributes 8 times this divisor. Thus the coefficient of q^N from the last sum is 8 times the sum of all divisors of N.

We get the same story from the first sum, but this time the nth term in the sum contributes something only if $k(4n) = N$, and then it contributes $32n = 8(4n)$. So we get a contribution only from divisors $4n$ of N, that is, divisors which are divisible by 4, and the total contribution is 8 times the sum of all such divisors.

QED.

Remark: In particular, the number of ways a prime can be written as a sum of four squares is $8(p+1)$. This is a very nice result, but it counts representations in an awkward way. It would be more interesting to write representations in the form $n = a^2 + b^2 + c^2 + d^2$ where all are non-negative and $a \geq b \geq c \geq d$ and count these expressions. However, the resulting counts do not have a regular pattern; Jacobi gets a concise result because he counts some patterns more than others. If all the numbers are non-zero and different, there are 4! ways to order the numbers, and 2^4 ways to select signs, so Jacobi counts this 384 times. If all the numbers are positive, but one is repeated three times, there are 4 ways to arrange the terms and 2^4 to select signs, and Jacobi counts these cases 48 times.

Here is data using the new count rather than Jacobi's count, for primes less than 100:

3 :	1	43 :	3
5 :	1	47 :	2
7 :	1	53 :	3
11 :	1	59 :	3
13 :	2	61 :	4
17 :	2	67 :	4
19 :	2	71 :	2
23 :	1	73 :	5
29 :	2	79 :	3
31 :	2	83 :	4
37 :	3	89 :	5
41 :	3	97 :	6

Remark: It is easy to analyze Jacobi's counting method. We get a "symbolic" formula:

$$\sum_{d|n, 4\nmid d} d = 1(a^2) +$$

$$3(2a^2) + 6(a^2 + b^2) +$$

$$4(3a^2) + 12(2a^2 + b^2) + 24(a^2 + b^2 + c^2) +$$

$$2(4a^2) + 8(3a^2 + b^2) + 12(2a^2 + 2b^2) + 24(2a^2 + b^2 + c^2) + 48(a^2 + b^2 + c^2 + d^2)$$

Here a, b, c, d are distinct positive integers. The top line describes the case where three of the squares are zero, the second line gives the case where two of the squares are zero, and so forth. The second line says that each $(a^2 + a^2 + 0^2 + 0^2)$ contributes 3 to the sum and each $(a^2 + b^2 + 0^2 + 0^2)$ contributes 6 to the sum.

If p is prime, the full sum is $p + 1$. In this case the term $1(a^2)$ never occurs, the term $3(2a^2)$ only occurs for $p = 2$, the term $4(3a^2)$ only occurs when $p = 3$, and the terms $2(4a^2)$ and $12(2a^2 + 2b^2)$ never occur. Note the $6(a^2 + b^2)$ only occurs if $p \equiv 1 \pmod 4$. It is fun to try out this formula for other primes.

I cannot resist trying $p = 97$. Then $(p + 1) = 98$. Leaving out impossible terms and dividing both sides by 2 gives

$$49 = 3(a^2+b^2)+6(2a^2+b^2)+12(a^2+b^2+c^2)+4(3a^2+b^2)+12(2a^2+b^2+c^2)+24(a^2+b^2+c^2+d^2)$$

Since 49 is odd and all coefficients are divisible by 2 except $3(a^2 + b^2)$, we conclude that it is possible to write $97 = a^2 + b^2$ and indeed $97 = 9^2 + 4^2 + 0^2 + 0^2$. Since this representation is unique, we conclude that

$$49-3 = 46 = 6(2a^2+b^2)+12(a^2+b^2+c^2)+4(3a^2+b^2)+12(2a^2+b^2+c^2)+24(a^2+b^2+c^2+d^2)$$

Dividing both sides by 2 gives

$$23 = 3(2a^2 + b^2) + 6(a^2 + b^2 + c^2) + 2(3a^2 + b^2) + 6(2a^2 + b^2 + c^2) + 12(a^2 + b^2 + c^2 + d^2)$$

Since 23 is odd, we conclude that it must be possible to write $97 = 2a^2 + b^2$ and indeed we discover that $97 = 2 \cdot 6^2 + 5^2$. This is again unique, so

$$20 = 6(a^2 + b^2 + c^2) + 2(3a^2 + b^2) + 6(2a^2 + b^2 + c^2) + 12(a^2 + b^2 + c^2 + d^2)$$

Since 20 is not divisible by 6, it must be possible to write $97 = 3a^2 + b^2$ and indeed $97 = 3(4)^2 + 7^2$. This is unique, so

$$18 = 6(a^2 + b^2 + c^2) + 6(2a^2 + b^2 + c^2) + 12(a^2 + b^2 + c^2 + d^2)$$

Dividing by 6 gives

$$3 = (a^2 + b^2 + c^2) + (2a^2 + b^2 + c^2) + 2(a^2 + b^2 + c^2 + d^2)$$

If $a > b > c > d > 0$ and $a^2 + b^2 + c^2 + d^2 = 97$, there are few possibilities and we soon rule them all out. So

$$3 = (a^2 + b^2 + c^2) + (2a^2 + b^2 + c^2)$$

Indeed 97 can be written as $2a^2 + b^2 + c^2$ in three different ways:

$$2(6)^2 + 4^3 + 2^2, \;\; 8^2 + 2(4^2) + 1^2, \;\; 8^2 + 5^2 + 2(2)^2$$

In the end, we have written 97 as a sum of four squares in 6 ways, as predicted by our table.

Remark: In particular, Jacobi's four square theorem implies Fermat's theorem that every $p \equiv 1 \pmod 4$ can be written as a sum of two squares. Indeed, $\frac{(p+1)}{2}$ is then odd, and equal to

$$3(a^2+b^2)+6(2a^2+b^2)+12(a^2+b^2+c^2)+4(3a^2+b^2)+12(2a^2+b^2+c^2)+24(a^2+b^2+c^2+d^2)$$

so the first term cannot be zero and representations of $p = a^2 + b^2$ must exist.

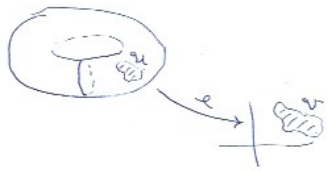
Chapter 27

Riemann Surfaces

27.1 Compact Riemann Surfaces

The compact 2-dimensional surfaces have been classified topologically, and turn out to be tori with g holes, and various un-oriented surfaces like the projective plane and the Klein bottle. Since small pieces of a surface look like open sets in the complex plane, it is natural to try to extend complex analysis to surfaces. Because holomorphic maps preserve orientation, this turns out to be impossible on the non-oriented surfaces, so we restrict attention to tori with g holes. The Riemann sphere is a torus with no holes, and elliptic function theory is complex analysis on tori with one hole. We need only extend to tori with more holes.

By definition, a coordinate system on a surface M is a connected open set $\mathcal{U} \subset M$, a connected open set $\mathcal{V}$ in the plane, and a homeomorphism $\varphi : \mathcal{U} \to \mathcal{V}$. See the picture below for the intuition behind this definition.



Because φ is a homeomorphism, continuous functions on $\mathcal{U}$ are the "same thing" as continuous functions on $\mathcal{V}$. This allows us to completely ignore $\mathcal{U}$ and work on $\mathcal{V}$. To describe a function f on the surface, we just give a formula for the resulting $f(z)$ on $\mathcal{V}$. In particular, we say our function is holomorphic or meromorphic on $\mathcal{U}$ if $f(z)$ is holomorphic or meromorphic on $\mathcal{V}$.

However, we cannot cover the entire surface by one coordinate system, so we have to allow several of them. A compact surface can be covered by a finite number of such coordinate systems: $(\mathcal{U}_i, \varphi_i, \mathcal{V}_i)$. We say these coordinates make M a *Riemann Surface* if M is connected and Hausdorf and the compositions $\varphi_j \circ \varphi_i^{-1} : \mathcal{V}_i \to \mathcal{V}_j$ are all holomorphic.

This condition looks abstract but isn't. Suppose we have two coordinate systems and a function which is $f(z)$ in $\mathcal{V}_1$ and $g(w)$ in $\mathcal{V}_2$. The map $\varphi_2 \circ \varphi_1^{-1}$ just gives w in terms of z. Moreover $g(w(z)) = f(z)$. We want f to be holomorphic exactly when g is holomorphic, and this will hold if $w(z)$ is holomorphic.

Our discussion of the Riemann sphere is easily cast in this language. The Riemann sphere is the complex plane with an additional point at ∞. We get one coordinate system by taking $\mathcal{U}$ to be everything except ∞, $\mathcal{V} = C$. and $\varphi(z) = z$. We get a second coordinate system by taking $\mathcal{U}$ to be all points in the Riemann sphere except 0, $\mathcal{V}$ to again be C, and $\varphi(z) = \frac{1}{z}$.

Our discussion of a fundamental region for a lattice also fits this language, giving a Riemann surface structure on a torus. We take $\mathcal{U} = \mathcal{V}$ to be the interior of the fundamental region, with $\varphi(z) = z$. We can cover the boundary as well using a small number of slightly shifted fundamental regions. ion with identifications on the boundary, then the $\mathcal{U}$ obtained above are more complicated, but easily explained by picture:

We have been cavalier about a few points in our definition of a Riemann surface. First, $\varphi_2 \circ \varphi_1^{-1}$ is only defined on the image of $\mathcal{U}_i \cap \mathcal{U}_j$ rather than on the entire $\mathcal{V}_j$. In practice this is obvious and not worth comment. Second, we are not restricted to using the coordinate systems listed when defining the Riemann surface structure, but can introduce any other coordinate system compatible with these. Our discussion of the torus shows this point in action. We provided a specific coordinate system for points inside the fundamental region, but many alternatives for boundary points. It is not worth worrying about this.

On the other hand, it is not enough to select a surface up to homeomorphism and then say "pick a coordinate system." We saw this with tori because there are uncountably many inequivalent complex structures on a torus. Thus the particular coordinate system chosen is important.

We will mainly deal with compact Riemann surfaces, but the definition makes sense for noncompact surfaces as well. In that case, infinitely many coordinate patches may be needed to cover the surface.

27.2 Functions and Differentials

With the definition out of the way, we can get down to work. We generalize complex variable theory to Riemann surfaces by breaking problems into pieces and using local coordinates to calculate the pieces. For example, we say that a function is holomorphic if $f(z)$ is holomorphic

in each coordinate piece, and we say that a function is meromorphic if $f(z)$ is meromorphic in each coordinate piece. We say f has a zero at p if $f(z)$ is zero at the appropriate spot in one, and hence all, coordinate systems, and we say that the zero has order k if $f(z)$ has a zero of order k in those coordinate systems. We differentiate functions by differentiating in a coordinate system. Finally we integrate over a contour γ by breaking γ into pieces, each entirely in some coordinate system, integrating the pieces in the appropriate coordinate system, and adding the results.

Sadly, there is a very significant complication. If a small piece of a calculation is in two different coordinate systems, we must check that both systems give the same answer. Both differentiation and integration fail this test! Consider, for example, the Riemann sphere. Let $f(z) = z$ and notice that $\frac{df}{dz} = 1$ is constant. In the second coordinate system, $z = \frac{1}{w}$. But $\frac{d}{dw}\frac{1}{w} = -\frac{1}{w^2}$ and this is decidedly not constant.

Integration also fails on the sphere. Let $\gamma(t) = e^{it}$ for $0 \leq t \leq 2\pi$. This is the standard counterclockwise circle about the origin, and $\int_\gamma z\,dz = 0$. But if $w = \frac{1}{z}$, then the path becomes e^{-it} in the w coordinates and the function becomes $\frac{1}{w}$ and a counterclockwise integral of $\frac{1}{w}$ around a counterclockwise circle is $-2\pi i$.

Luckily, Leibniz rescues us from both problems with his wonderful notation. According to Leibniz, we do not integrate functions $f(x)$ but instead integrate differentials $f(x)\,dx$. Did your calculus teacher always insist on including the dx?

We say a *function* on a Riemann surface is an assignment to each local coordinate of a function $f(z)$, such that on the intersection of two coordinates with assignments $f(z)$ and $g(w)$ we have $f(z) = g(w(z))$. We say a *differential* on a Riemann surface is an assignment to each local coordinate of a symbol $f(z)\,dz$ where $f(z)$ is a holomorphic or meromorphic function, such that on the intersection of two coordinates with assignments $f(z)\,dz$ and $g(w)\,dw$ we have

$$f(z)\,dz = g(w(z))\frac{dw}{dz}(z)\,dz$$

This neatly solves our problem: the derivative of a function is a differential and we call this differential df, so

$$df = \frac{df}{dz}dz$$

Similarly, differentials can be integrated over contours using the formula

$$\int_\gamma f(z)\,dz = \int_a^b f(\gamma(t))\frac{d\gamma}{dt}\,dt$$

The reader can easily check that both formulas have the correct invariance related to coordinate changes.

It is fun to check these ideas on the Riemann sphere. We will integrate three differentials over the counterclockwise unit circle in C. In the $w = \frac{1}{z}$ coordinates this becomes the clockwise unit circle. The differentials we integrate are

$$z\,dz, dz, \frac{1}{z}\,dz, \frac{1}{z^2}dz$$

Note that

$$dz = \frac{dz}{dw}\,dw = \frac{d}{dw}\left(\frac{1}{w}\right)\,dw = -\frac{1}{w^2}\,dw$$

So the differentials become

$$-\frac{1}{w^3}\,dw, -\frac{1}{w^2}\,dw, -\frac{1}{w}\,dw, -dw$$

In both z and w coordinates, only the third differential has non-zero integral, and the integrals are equal because the w circle travels in the opposite direction as the z circle.

Remark: We say a differential $f(z)\,dz$ is *holomorphic* if $f(z)$ is holomorphic in all coordinate systems; otherwise it is a meromorphic differential. The above examples show two differentials, $z\,dz$ and dz, which appeared to be holomorphic in z coordinates but turned out to be meromorphic in w coordinates. But that is because both had one bad point, at ∞, and that bad point was not visible in the z coordinates. If a differential is not singular at p in one coordinate system, it is clearly not singular at p in all other coordinate systems.

Remark: In higher dimensions, there are differentials of degree 1, of degree 2, etc, and these objects are called *differential forms*. For surfaces, we only consider one type of differential, but we will often still call it a differential form.

Remark: On the sphere, the only globally holomorphic functions are constant. There are no globally holomorphic differentials on the sphere because if $f(z)dz$ were such a form, then $g(w)dw = -f\left(\frac{1}{w}\right)\frac{1}{w^2}\,dw$ is holomorphic at the origin and thus $f\left(\frac{1}{w}\right)$ must vanish at the origin, so the original $f(z)$ is bounded, hence constant, hence zero.

On the torus, dz is a well-defined holomorphic differential form. In this special case there is a one-to-one correspondence between functions and one-forms, sending $f(z)$ to $f(z)\,dz$. In particular, all holomorphic functions are constant and all holomorphic differential forms are a constant times dz.

A beautiful theorem of Riemann states that on a compact surface of genus g, there are exactly g linearly independent holomorphic differential forms. We will keep running into this result in this last chapter. It is not easy to prove.

Remark: We now know enough to prove three general theorems for compact Riemann Surfaces.

Theorem 209. *The order of a zero or pole is well-defined, independent of coordinate system, for both functions and differential forms.*

Proof: Replacing $f(z)$ by $\frac{1}{f(z)}$ if necessary, it suffices to handle zeros. We may translate both z and w coordinates so the point in question is the origin in both coordinates. Expand $f(z)$ and $g(w(z))$ in power series about the origin. Then

$$f(z) = a_k z^k + ...$$

$$g(w) = b_m w^m + ...$$

$$w(z) = c_1 z + ...$$

where $a_k \neq 0, b_m \neq 0, c_1 \neq 0$. So $f(z) = g(w(z))$ implies

$$a^k z^k + ... = b_m(c_1 z + ...)^m + ... = b_m c_1^m z^m + ...$$

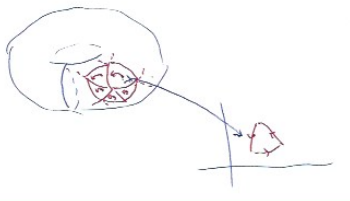
and so $k = m$. Differentials are handled the same way, since $\frac{dw}{dz}$ is never zero because this map is one-to-one. QED.

Theorem 210. *On a compact Riemann surface, all globally holomorphic functions are constant.*

Proof: The function $|f(z)|$ has a maximum because M is compact. By the maximum principle, $f(z)$ is constant in any coordinate system containing this maximum point. By the identity theorem, it is constant in any open set which intersects such a set. Since M is connected, it is globally constant.

Theorem 211. *On a compact Riemann surface, each non-constant meromorphic function takes all values the same number of times, counting multiplicity.*

Proof: If f is not constant, then $\frac{1}{f}$ is a well-defined meromorphic function on the surface and df is a well-defined meromorphic differential form on the surface, and so $\frac{df}{f}$ is also a well-defined meromorphic differential form on the surface. Thus integration of this differential form is well-defined.



Triangulate the surface as in the picture above so each triangle is in a single coordinate $\mathcal{U}_i$. Select the triangulation so no zero or pole is on the boundary of a triangle. For each triangle,

form the path γ_i which the triangulation traces around the triangle, picking the directions of these paths consistently over the surface; this is possible because M is oriented. Then the following sum of integrals cancels out and yet gives the expression on the right.

$$\frac{1}{2\pi i}\sum \int_{\gamma_i} \frac{df}{f} = \sum_{\text{triangles}} (\text{order of zeros in triangle}) - (\text{order of poles in triangle})$$

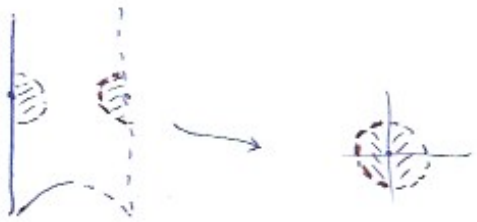
So the total number of zeros equals the total number of poles. Replace f by $f - \lambda$ to show that the total number of times f equals λ is also the total number of poles. QED.

27.3 Return to the Modular Function

In an earlier chapter we proved that the modular function $j(\tau)$ takes each complex value exactly once in the fundamental region of the modular group. The proof contained two steps. The first was an intricate calculation to show that $j(\tau)$ has a pole of order one at infinity. The second was an integral around the boundary of the fundamental region to show that j must take all other values exactly once. We can replace this second step with the theorem that a meromorphic function on a compact Riemann surface takes each value the same number of times, provided we make the compactified fundamental region into a Riemann surface. We'll do that now.

We only need provide local coordinates near each point of the surface. This is trivial for interior points of the region. The modular function is periodic with period one, and thus can be written as a function of $q = e^{2\pi i \tau}$; the map $q(\tau)$ sends the intersection of the upper half plane and the region $-\frac{1}{2} \le \Re(z) < \frac{1}{2}$ to the punctured disk $0 < |q| < 1$. We send ∞ to $q = 0$, and then q provides local coordinates for the region near ∞.

It is easy to obtain local coordinates for points on the vertical boundary, as illustrated in the picture below.

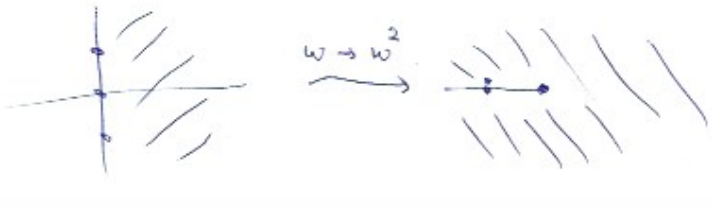


For the bottom boundary, consider the map $z \to w = \frac{z-i}{z+i}$. This is linear-fractional and thus an automorphism of the sphere. It maps $i \to 0$ and $-i \to \infty$ and $-1 \to i$ and $1 \to -i$. In addition,

it maps $0 \to -1$. Consequently we have the following picture, where the interior of the circle goes to the left half plane, the exterior of the circle goes to the right half plane, and the circle including the portion bounding the fundamental region goes to the imaginary axis.

In the fundamental region, the map $z \to -\frac{1}{z}$ identifies sides of the circular boundary and fixes i. This map induces a corresponding map in the w-coordinates. To see what it is, we compute $w \to z \to -\frac{1}{z} \to w$. A brief calculation shows that this map is $w \to -w$.

We follow our map for w with the map $w \to q = w^2$. The picture is shown below. The q's then form the coordinate neighborhood in the plane for points on the bottom boundary and in particular for i. The most interesting feature is that the point i in the fundamental region gets a full neighborhood in the q coordinates, even though it seems to get only half a neighborhood in the fundamental region. We call i a *branch point*, and will say more about them in following sections; we always handle such points using a map $z \to z^k$ to convert a partial neighborhood into a complete neighborhood.

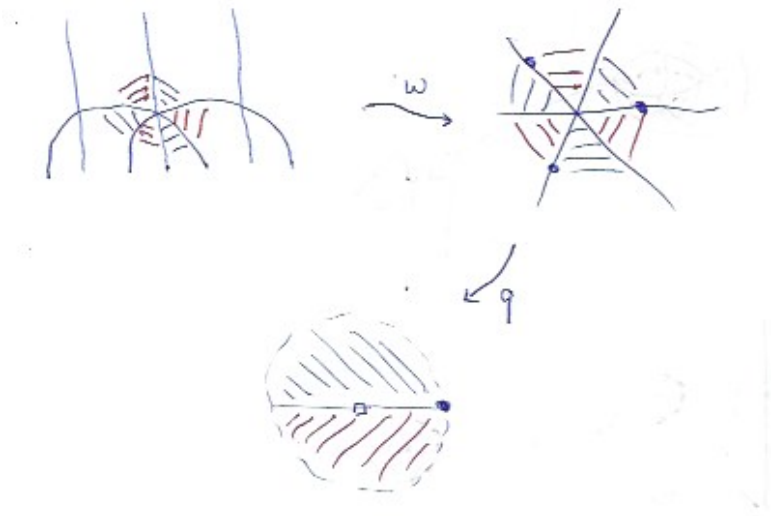


Finally, we need coordinates near $\omega = -\frac{1}{2} + i\frac{\sqrt{3}}{2}$. Consider the map $z \to w = \frac{z-\omega}{z-\omega^2}$. Like all linear fractional transformations, this maps lines and circles to lines and circles. It maps $\omega \to 0$ and $\omega^2 \to \infty$, so it maps the unit circle to some straight line through the origin. Moreover, the vertical line through $x = -\frac{1}{2}$ goes through both ω and ω^2, so it also maps to another straight line through the origin. Since holomorphic maps preserve angles and the angle between the vertical left boundary and the circular boundary is $\frac{2\pi}{6}$, the two lines in w coordinates must meet at angle $\frac{2\pi}{6}$.

We are too lazy to determine the precise lines produced in the q coordinates, and it doesn't

matter. For simplicity we assume the x-axis is one of these lines. We then obtain the following pictures:

These pictures make it clear that we can account for both corners in the fundamental region using a wedge of angle $\frac{2\pi}{3}$ in the w coordinates, and we can identifies the corresponding boundaries using the map $w \to q = w^3$

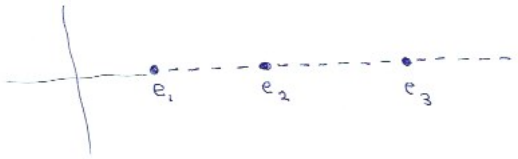


In this way, the fundamental region of the modular group becomes a Riemann surface, and indeed a surface equivalent to the Riemann sphere, since $j(\tau)$ is a one-to-one and onto map from the compactified modular surface to the sphere.

27.4 Riemann's Construction

Consider the function $f(z) = \sqrt{(z-e_1)(z-e_2)\dots(z-e_k)}$ where the e_i are distinct numbers in the complex plane. We are going to show that f can be made into a single-valued holomorphic function on a compact surface, which will turn out to have g holes where $g = \frac{k-2}{2}$ if k is even, and $g = \frac{k-1}{2}$ if k is odd. The a_i can be anywhere, but it is easier to draw pictures if they are on the real axis, and we will do so.

Begin by connecting all the points with straight line segments as shown below. Connect the final point out to infinity by such a line. See the picture below. Remove these line segments. The remaining set is simply connected, so we can define a branch of $\sqrt{(z-e_1)(z-e_2)\dots(z-e_k)}$ on this set.



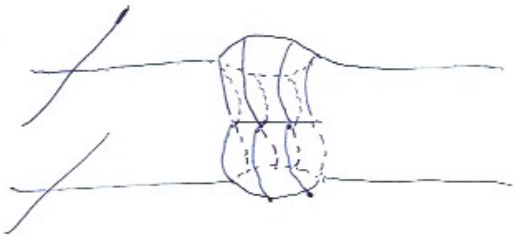
Notice that all terms inside the radical except $z-e_1$ are well-behaved at e_1, so the square root has the form $f(z) = g(z)\sqrt{z-e_1}$. If we draw a small circle around e_1, the value of $g(z)$ will return to its original value as we traverse the circle, but $\sqrt{z-e_1}$ will change sign as we traverse. So all along the interval $[e_1, e_2]$ the values of the radical above and below the cut will differ by a sign.

The same analysis holds at e_2, but this time the cut at $[e_1, e_2]$ accounts for the sign change and therefore the values of the radical agree on $[e_2, e_3]$. So remove this cut. Proceed for further e_i. If the number k of e_i is even, there will be $\frac{k}{2}$ cuts and the cut from the last one to ∞ will be gone. If k is odd, there will be $\frac{k-1}{2}$ cuts and the final cut from e_k to ∞ will remain. See the picture below. The remaining cuts are called *branch cuts* and $f(z) = \sqrt{(z-e_1)(z-e_2)\dots(z-e_k)}$ is well-defined and holomorphic on the plane with these cuts removed.

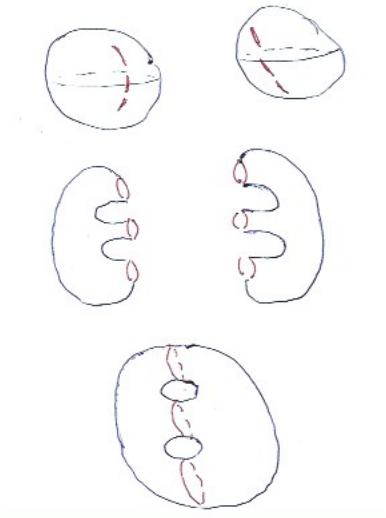
The square root has two values with differ by a sign, but only one of these values is taken on

our copy of the complex plane with cuts. Form a second copy of the cut plane and define $f(z)$ on this copy to be the negative of $f(z)$ on the original copy.

Now the idea is to glue these two planes together and obtain a single more complicated object on which $f(z)$ lives. The glueing takes place along the branch cuts. The value of f just above this cut on the first plane agrees with the value of f just below this cut on the second plane. So we must glue the upper portion on the first plane to the lower portion on the second plane, and the lower portion on the first plane to the upper portion on the second plane. In some sense, the two planes should cross each other at the cut line without touching. Note that the cut line itself becomes two lines on the new object. See the attempted picture below.



While this new object seems difficult to visualize, it is easy to understand if we draw another way. The complex plane is the Riemann sphere with the point at infinity removed. So our two planes are two spheres with cutout lines. We glue the two spheres together alone these lines to obtain a torus with g holes. See the picture below.



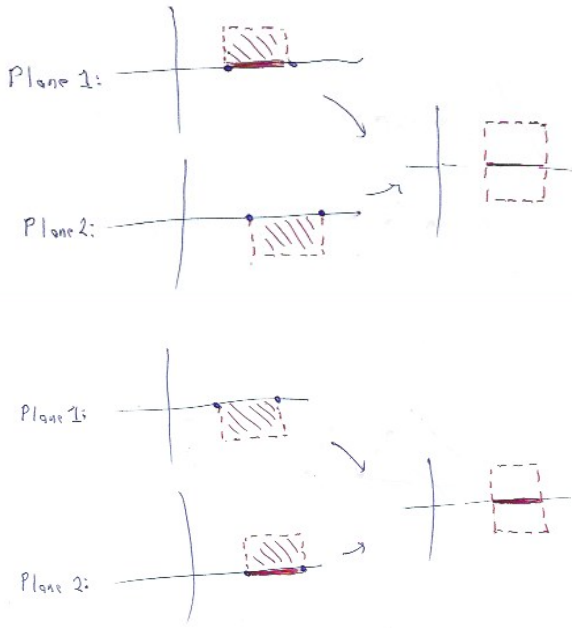
The exact points on our new surface will become clear when we define a coordinate system for the surface. It will turn out that the object contains all points on the two planes after the

branch cuts are removed. It contains two copies of points on the branch cuts. It contains one point at each e_i. If there is no branch cut to infinity, it contains two points at infinity, one on the Riemann sphere for the first plane and one on the Riemann sphere for the second plane. But if a branch cut extends to infinity, these points become a single point at infinity.

From now on, let us denote the surface we have just constructed $\mathcal{S}$. Of course we have constructed many surfaces, depending on the a_i and their number k.

We end this discussion by producing coordinate patches. Almost all points are on one of the cut planes, and each is a trivial coordinate system.

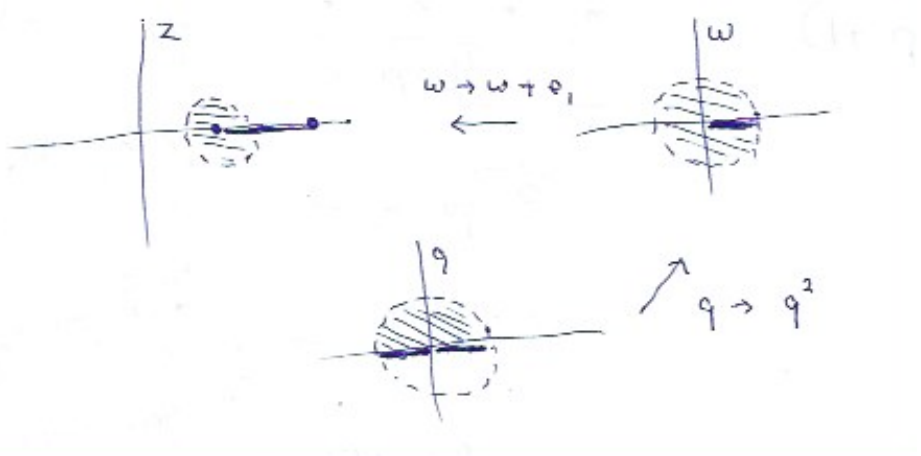
We produce two coordinate patches for each branch cut as explained in the following picture:



It remains to find coordinate patches at the e_i and the one or two ∞. Let us handle infinity first. If no branch cut extends to infinity, then there is a constant B such that each plane contains all z with $|z| > B$. For each plane, introduce $w = \frac{1}{z}$. This is a coordinate system mapping $|z| > B$ to $0 < |w| < \frac{1}{B}$, and we map the point at infinity on the plane to $w = 0$ in these coordinates.

Finally we deal with the e_i and possibly ∞ if a branch cut extends that far. The treatment of ∞ will be exactly like the treatment of the a_i and can be left to the reader. Each of these points is called a *branch point*. As before, note that $f(z) = g(z)\sqrt{z - e_i}$ at branch points. The map $w \to z = w + e_i$ sends a small disk about the origin to a disk about the branch point. The map

$q \to w = q^2$ sends the upper half of a disk about the origin to the full disk. The branch cut has become two cuts in the q disk, on the left and right of the origin. We can map the bottom half of the q-disk to a small disk about e_i in the second copy of the plane. Consequently, the full disk about q produces coordinates for the two disks about e_1, where one of these is the full disk in the first plane and the other is the full disk in the second plane. Moreover, $f(z)$ is holomorphic in these new coordinates because $f(z)$ on the first plane equals $-f(z)$ on the second, so these match along branch cuts.



When working on this surface $\mathcal{S}$ in actual practice, the branch cuts and branch points more or less take care of themselves. We only need caution when computing the orders of meromorphic functions at branch points.

In this section, we have produced compact Riemann surface structures on all compact oriented surfaces. It will soon become clear that the structures we produced when $g = 0$ or $g = 1$ are conformally equivalent to the Riemann sphere, and one of our lattice torus structures. In general, these structures are called *hyperelliptic Riemann surfaces*, and it turns out that every Riemann surface of genus $g \leq 2$ is conformally equivalent to such a surface. But for $g \geq 3$, there are other structures that we leave to deeper treatments of Riemann surface theory.

27.5 The Field of Meromorphic Functions

Because Riemann surfaces are connected, the identity theorem holds on them. It follows that the set of meromorphic functions forms a field, because if f is not identically zero, then its zeros are isolated, and isolated zeros of f become poles of $\frac{1}{f}$.

We found all doubly periodic functions using analysis to construct $\wp(z)$ and then showing that $\wp$ and $\frac{d\wp(z)}{dz}$ generate the full field of meromorphic functions. It may then be surprising that

we can find the complete field of meromorphic functions on our new hyperelliptic Riemann surfaces without any analysis at all.

We know two functions, $f(z) = z$ and $g(z) = \sqrt{(z-e_1)(z-e_2)\dots(z-e_k)}$. The first just sends each point z in either plane to the complex number z. The second was our motivation for creating the surface; $g(z)$ is single-valued and holomorphic on this surface.

However, there will be questions about both functions at the branch points e_i and at the point or the points at infinity. We could calculate the functions in local coordinates to show that they behave well at these points as well, but that isn't necessary. The e_i and the points at infinity are isolated singularities. Both f and g are bounded near the e_i, so these singularities are removable. Both f and g only take values larger than some bound near points at infinity, so these isolated singularities are poles. Done.

Notice that the value z is taken by f exactly once in either place, except for values at branch points and infinity. So f takes almost all values twice, and therefore takes every value twice by a general theorem we proved. In particular, it takes infinity twice. So if k is even and there are two points at infinity, then each is a pole of order one. And if k is odd and there is only one point at infinity, then it is a pole of order two.

The function f has two distinct zeros of order one, unless one of the e_i is zero. In that case, f takes the value zero at only one point, so it is a zero of order 2. Note that all of these remarks apply to $f(z) + c$ for constants c.

Theorem 212. *Every meromorphic function on a hyperelliptic surface $\mathcal{S}$ has the form*

$$R_1\Big(f(z)\Big) + R_2\Big(f(z)\Big)\, g(z)$$

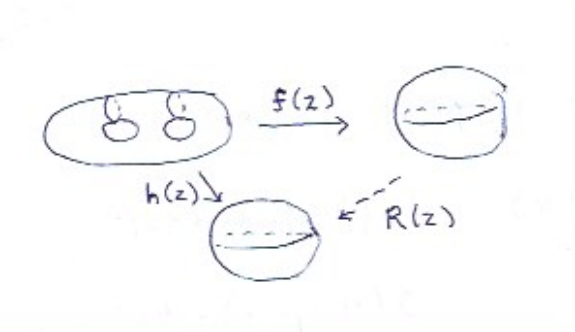
where R_1 and R_2 are rational functions, that is quotients of polynomials.

Proof: Notice that $f(z)$ is an even function, in the sense that it takes the same value on equivalent points in both planes, and notice that $g(z)$ is an odd function, in the sense that the values it takes on equivalent points in the two planes differ in sign. If $h(z)$ is an arbitrary meromorphic function on our surface, we can write

$$h(z) = \frac{h(z) + h(z_1)}{2} + \frac{h(z) - h(z_1)}{2g(z)} g(z)$$

where z is a point in either plane, and z_1 is the corresponding point in the other plane. These functions are defined except possibly at the e_i or ∞, but these are isolated singularities and thus removable. Each coefficient is an even function, and therefore it suffices to prove that every even meromorphic function is a rational function in $f(z)$.

Consider the diagram below. It is possible to define a meromorphic $R(z)$ completing the diagram, by an argument given shortly. Since every meromorphic function on the Riemann sphere is rational, $h(z) = R(f(z))$ and we are done.



Now we prove the existence of R. The function $f(z)$ takes each complex value twice, usually at z and z_1. But $h(z) = h(z_1)$. So we can define R by using f^{-1} to go backward to two points, and then selecting the value of h at these points. This leaves a finite number is isolated points where R is not defined, but all of these points are removable singularities. QED.

Remark: When there are one or two branch points, the functions we have just constructed seem to behave just like $\wp(z)$ and $\wp'(z)$, and indeed we are just reproducing that theory, modulo a 4 which is easily explained away. We found that the field of meromorphic functions on a torus is the field $C(X)$ of all rational functions in an indeterminant X, extended by adding a root of

$$Y^2 - 4X^3 - g_2X - g_3 = Y^2 - 4(X - e_1)(X - e_2)(X - e_2)$$

. In the general case of a hyperelliptic $\mathcal{S}$, we have an easy generalization. This time the field of meromorphic functions is the field $C(X)$ extended by adding a root Y of

$$Y^2 - (X - e_1)(X - e_2)...(X - e_k)$$

27.6 Differentials and Integration

Recall that differentials are objects which look like $g(z)dz$ in local coordinates. If $f(z)$ is a meromorphic function on $\mathcal{S}$, the expression $\frac{df}{dz}$ is not well-defined because it does not transform correctly under coordinate changes, but $df = \frac{df}{dz}dz$ is well-defined. Moreover, we cannot integrate $f(z)$ because it does not transform correctly, but we can integrate $f(z)dz$.

Suppose ω is a differential form, equal to $\omega(z)dz$ in local coordinates. If $f(z)$ is a meromorphic function, then $f(z)\omega = f(z)\omega(z)dz$ is again a differential form, because f transforms in a trivial way and $\omega(z)dz$ transforms as differentials do. Conversely, if both $f(z)dz$ and $g(z)dz$ are differentials, then $\frac{f(z)}{g(z)}$ is a well-defined function. It follows that if ω is a non-zero differential, then the most general differential is $f\omega$ for f an arbitrary meromorphic function on $\mathcal{S}$.

Consider the case of a hyperelliptic surface $\mathcal{S}$. One of the functions on this surface is $f(z)$. In either of the main coordinate systems which cover all points except the branch points and infinite points, $f(z) = z$ and $df = dz$. It follows that the most general differential on $\mathcal{S}$ has the form

$$R_1(z)\,dz + R_2(z)\sqrt{(z-e_1)(z-e_2)\dots(z-\bar{e}_k)}\,dz$$

These are the objects that can be integrated over the surface:

$$\int_\gamma R_1(z)\,dz + \int_\gamma R_2(z)\sqrt{(z-e_1)(z-e_2)\dots(z-e_k)}\,dz$$

If $k = 3$ or 4, this is an elliptic integral; indeed the formula for the length of an ellipse has this form. So we started with such integrals from 1670 or so, and we have returned to our starting point with these hyperelliptic surfaces!

The most significant feature of complex integration is that it does not depend on the fine structure of the path γ. We are free to deform γ without changing the integral, as long as the deformation does not cross a pole of the form. The first part of these lectures was all about the intricate details when a singularity is crossed.

But with Riemann surfaces, there is another possible reason that two paths could produce a different result: one of the paths might go around a hole. We would like to understand the influence of the holes on integrals without having to separate out this influence from the influence of singularities of $f(z)$. The easiest way to make this separation is to integrate differentials ω which are holomorphic everywhere. Then the coefficients will have no singularities, and the only way that integrals over different paths from one point to another can differ is because one of the paths went around a hole. So we now search for holomorphic differentials, which are also called *differentials of the first kind*.

On the Riemann sphere, such a differential would equal $P(z)\,dz$ where P is a polynomial. At infinity we introduce $z = \frac{1}{w}$ so that $dz = -\frac{1}{w^2}\,dw$. Then

$$P(z) = P\left(\frac{1}{w}\right)\frac{-1}{w^2}\,dw$$

This has a singularity at $w = 0$ even if P is just a constant, and the singularity is worse if P has higher degree. So there are no holomorphic differentials on the sphere.

On a torus, dz is a global differential and the most general differential is $f(z)\,dz$ for meromorphic f. If f is not constant, then it has at least one pole and the differential is not holomorphic. So there is exactly one holomorphic differential up to multiplication by a constant, and it is dz.

Notice that the horizontal and vertical sides of a fundamental paralellogram produce closed paths around the hole of the doughnut. See the picture below. The integral of ω around one of

these paths is 1 and the integral around the other is τ. So when we investigate the effect of the holes on integration, we find the generators of the period lattice. This is a hint that studying holomorphic differentials is crucial to understanding compact Riemann surfaces.

Let us turn to the general hyperelliptic case. The full set of meromorphic functions is generated by $f(z) = z$ and $g(z) = \sqrt{(z-e_1)(z-e_2)\dots(z-e_k)}$, but we could just have well have used $f(z)$ and $\frac{1}{g(z)}$ and we will see shortly that this is a useful change. Thus we are looking for holomorphic differentials of the form

$$R_1(z)\,dz + \frac{R_2(z)}{\sqrt{(z-e_1)(z-e_2)\dots(z-e_k)}}\,dz$$

Differentials from the left half of this formula cannot be holomorphic at infinity, for reasons sketched earlier. If ∞ is a branch point, the coordinate system introduces a square root and the transformation for dw involves $\frac{1}{w}$ rather than $\frac{1}{w^2}$, but that doesn't help enough.

So we must study terms of the second type. At e_i we introduced local coordinates q with $z = w + e_i$ and $w = q^2$. Thus $dz = dw$ and $dw = \frac{dw}{dq}\,dq = 2q\,dq$. Also $\sqrt{z-e_i} = \sqrt{w} = q$, so

$$\frac{1}{\sqrt{z-e_i}}dz = \frac{1}{q}\,2q\,dq = 2\,dq$$

The remaining portion of the square root is holomorphic and non-zero at e_i. It follows that when $R_1(z) = 1$ these differentials are holomorphic, except perhaps at infinity. But at infinity the terms vanish, and certainly for large k there are enough terms to cancel out the pole of dz there. Details follow.

Consider the entire term

$$\frac{R(z)}{\sqrt{(z-e_1)(z-e_2)\dots(z-e_k)}}\,dz$$

The singularities at e_i are cancelled by the q coordinates, but just barely. Any singularity of $R(z)$ at such points will not cancel. So if the above is a holomorphic differential, then $R(z)$ cannot have any poles in the plane, and thus must be a polynomial.

Suppose now that k is even. Then there are two poles at infinity, one for each copy of the plane. Local coordinates for one of these poles are given by $z = \frac{1}{w}$ and $dz = -\frac{1}{w^2}\,dw$. So the differential becomes

$$\frac{P\left(\frac{1}{w}\right)}{\sqrt{1-e_1w)(1-e_2w)\dots(1-e^kw)}w^{-k/2}}\,\frac{-1}{w^2}\,dw = \frac{P\left(\frac{1}{w}\right)}{\sqrt{1-e_1w)(1-e_2w)\dots(1-e^kw)}}\,w^{\frac{k}{2}-2}\,dw$$

If $k = 4$, the power of w is zero and this is holomorphic at the origin provided P is just a constant. So the space of holomorphic differentials is one dimensional. If $k = 6$, the power of w in the final term is one and thus the entire differential is holomorphic if the degree of P is at most one. So the space of holomorphic differentials has dimension 2. In general, for even $k = 2K$, the power of w in the final term is $K - 2$ and the degree of P can be at most $K - 2$, so the space of holomorphic differentials has dimension $K - 1$. Notice that when $k = 2K$, the genus of the surface is $\frac{2k-2}{2} = K - 1$. So the space of holomorphic differentials has dimension g in this case.

Suppose finally that k is odd, say $k = 2K-1$. The genus of the surface is then $g = \frac{k-1}{2} = K-1$. This time there is only one point at infinity and it is a branch point. Local coordinates are given by $z = \frac{1}{w}$ and $w = q^2$ where half of the q-disk maps to points near infinity in the first plane, and the other half of the disk maps to points near infinity in the second plane.

We first work in w coordinates on each plane. Then as before we get

$$\frac{P\left(\frac{1}{w}\right)}{\sqrt{1-e_1w)(1-e_2w)\dots(1-e^kw)}}\, w^{\frac{k}{2}-2}\, dw$$

Note that $w = q^2$ and so $dw = 2qdq$. Note also that $k = 2K - 1$. Then the above formula becomes

$$\frac{P\left(\frac{1}{q^2}\right)}{\sqrt{1-e_1q^2)(1-e_2q^2)\dots(1-e^kq^2)}}\, q^{k-4}2q\, dq = \frac{P\left(\frac{1}{q^2}\right)}{\sqrt{1-e_1q^2)(1-e_2q^2)\dots(1-e^kq^2)}}\, 2q^{2K-4}\, dq$$

Notice first that this is an expression in q^2, so what happens on the upper part of the q disk is repeated on the lower half, as predicted. The expression will be holomorphic at $q = 0$ just in case $-2\deg(P) + 2K - 4 \geq 0$ or $\deg(P) \leq K - 2$ and the dimension of the space of holomorphic differentials is thus $\deg(P) + 1 = K - 1 = g$.

We have proved

Theorem 213. *On a hyperelliptic Riemann surface, the dimension of the space of holomorphic differentials is g. These differentials have the form*

$$\frac{P(z)}{\sqrt{(z-e_1)(z-e_2)\dots(z-e_k)}}\, dz$$

where $\deg(P) < g$.

It turns out that the space of holomorphic differentials has dimension g on any compact Riemann surface with g holes, even if it is not hyperelliptic.

27.7 Integrating Holomorphic Differentials

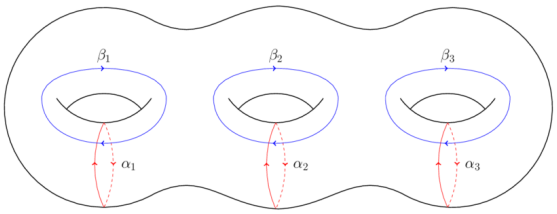
If ω is a differential on a surface $\mathcal{S}$ and γ is a path on the surface which does not go through a singularity of ω, we can compute $\int_\gamma \omega$. The study of such integrals played a significant role in our earlier theory, so it is not surprising that it continues to play an important role in Riemann surface theory. In our earlier theory, we discovered that

- it is possible to define $\int_\gamma \omega$ even if γ is merely a continuous path, and
- if two paths are homotopic with fixed end points, or homotopic through close curves, they have the same integral

The homotopy must avoid singularities of the differential. The central idea behind both arguments was that when the differential is $f(z)\,dz$, we can find a function $g(z)$ locally which satisfies with $dg = f(z)\,dz$. Since this result is purely local, it also holds on Riemann surfaces, and the entire theorem just quoted then holds on an arbitrary Riemann surface with the same proof.

A very interesting special case occurs when ω is a holomorphic differential. In that case there are no singularities to worry about and paths can be deformed arbitrarily. But we cannot conclude that $\int_P^Q \omega$ is independent of the path from P to Q because there can still be distinct paths from P to Q which are not homotopic, due to the topology of $\mathcal{S}$. The study of distinct paths from P to Q is equivalent to the study of closed paths from P to itself obtained by following one path from P to Q, and then following the other path back to P. Thus for further investigation, we want to study integration of ω around closed paths on the surface.

A picture of the surface with three holes was created by Christian Neurohr and is shown below. In the picture, six oriented closed paths are shown, labeled α_i and β_i. We will instead use the letters a_1, a_2, a_3 and b_1, b_2, b_3. It can be shown that any oriented closed path on this surface is equivalent to $\sum_i m_i a_i + \sum n_j b_j$ where the m_i and n_j are integers, in a certain sense described below.



Fix a point P on the surface and consider the set of closed paths starting and ending at P. Call two of these paths *equivalent* if they are homotopic. Define multiplication of paths $\gamma_1 \circ \gamma_2$ to be γ_1 followed by γ_2. Then the set of equivalence classes of paths forms a group, denoted $\pi_1(\mathcal{S})$

and called the *fundamental group* of the surface.

We want to think of the a_i and b_j as such paths starting and ending at P. We do this as follows, using a_1 as an example. Pick a point P_1 on a_1 and a path γ from P to P_1. Replace a_1 with the path γ, followed by a_1, followed by γ^{-1}.

It then turns out that the fundamental group of the torus is generated by the a_i and b_j. This group is non-abelian, so we are claiming that any closed path from P to P is homotopic to product of $a_i, a_i^{-1}, b_i, b_i^{-1}$, where repetitions are allowed and order matters. The group has relations, so this product representation is not unique, but we will ignore the relations for now.

Suppose that ω is a holomorphic differential form. By Cauchy's theorem, if γ_1 and γ_2 are closed paths which are homotopic through closed paths, then $\int_{\gamma_1} \omega = \int_{\gamma_2} \omega$. Moreover, if γ_1 and γ_2 are closed paths through P, then $\int_{\gamma_1 \circ \gamma_2} \omega = \int_{\gamma_1} \omega + \int_{\omega_2} \omega = \int_{\gamma_2 \circ \gamma_1} \omega$. Finally, when we added tail pieces to a_i and b_j so they could start and end at P, the integrals over the tail pieces cancel out. We conclude that the integral of ω over a path γ starting and ending at p depends only on the total number of a_i and b_j in its expression in the fundamental group and not on their order. In short,

$$\int_\gamma \omega = \sum_{m_i} m_i \int_{a_i} \omega + \sum_{n_j} n_j \int_{b_j} \omega$$

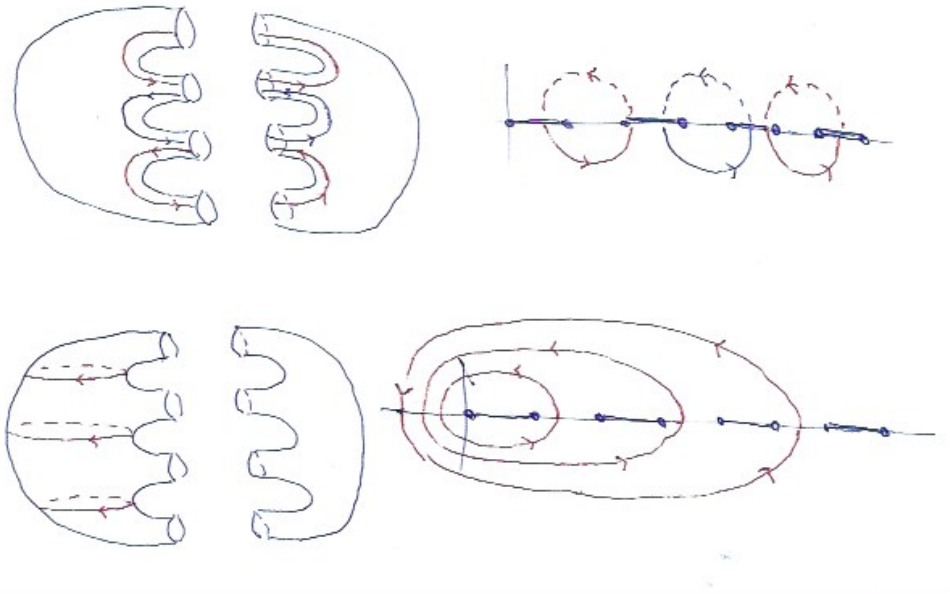
We define the *first homology group*

$$H_1(\mathcal{S}) = \frac{\pi_1(\mathcal{S})}{[\pi_1(\mathcal{S}), \pi_1(\mathcal{S})]}$$

to be the fundamental group of the surface divided by the normal subgroup generated by all commutators. Said another way, we add additional relations to create an abelian group from the fundamental group. According to the above discussion, this group is generated by the images of the a_i and b_j. We will later prove that there are no additional relations, so the group is isomorphic to $Z \oplus ... \oplus Z$, and thus has rank $2g$. Since the group is a topological invariant, this will prove that surfaces with different numbers of holes are not homeomorphic.

To summarize, each holomorphic differential form defines $2g$ numbers, namely $\int_{a_i} \omega$ and $\int_{b_j} \omega$. These numbers are called the *periods* of ω, and they determine the extent to which integration of ω from one point to another depends on the particular path chosen.

In the case of hyperelliptic surfaces, the periods are obtained as by integrating around the following paths:



Chapter 28

Brief Course in Surface Topology

Algebraic topology is a great invention of the twentieth century: homology groups, cohomology groups, homotopy groups, exact sequences, triangulation, and all that. The start of this subject is often set at 1895, when Henri Poincare published the paper *Analysis situs* in the Journal de l'Ecole Polytechnique, and followed it with five supplements published between 1899 and 1905.

Poincare's work generalized and explained results from Euler, Riemann, Betti, and others. In particular, significant hints came from Riemann surface theory. We want to use some of these results, and rather than giving isolated references to topology, we prefer to develop a small amount of theory from scratch. Details are left to the reader with the important caveat that nothing is hidden under the rug. All assertions made without proof are easy exercises.

28.1 The Fundamental Group

We briefly discussed the fundamental group at the end of the last chapter. Here is a precise formulation. Consider an arbitrary topological space X. A *continuous parameterized path* in X is a continuous map $\gamma : [0,1] \to X$, which we interpret as a path $\gamma(t)$ traced in time. Fix a point $x_0 \in X$ once and for all and restrict attention to paths γ such that $\gamma(0) = \gamma(1) = x_0$. If γ_1 and γ_2 are two such paths, we define $\gamma_1 \circ \gamma_2$ by

$$(\gamma_1 \circ \gamma_2)(t) = \begin{cases} \gamma_1(2t) & \text{if } 0 \le t \le \frac{1}{2} \\ \gamma_2(2t-1) & \text{if } \frac{1}{2} \le t \le 1 \end{cases}$$

We say two closed paths γ_1 and γ_2 are *homotopic* if there is a continuous

$$h(t,u) : [0,1]\times[0,1] \to X$$

such that $h(t,0) = \gamma_1(t)$ and $h(t,1) = \gamma_2(t)$ and $h(0,u) = x_0$ and $h(1,u) = x_0$. Intuitively, for each fixed u we have a closed path $h(t,u)$. When $u = 0$ we obtain γ_1 and as u increases we obtain deformations until at $u = 1$ we obtain γ_2.

Theorem 214. *Homotopy is an equivalence relation.*

Proof: Easy.

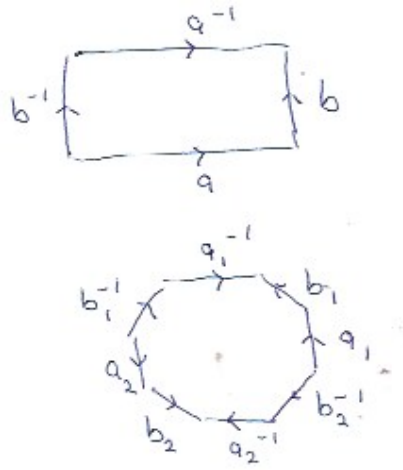
Definition 60. *The set of homotopy equivalence classes of paths starting and ending at x_0 is called the* fundamental group *of X and denoted $\pi_1(X, x_0)$.*

Theorem 215. *The operation $\gamma_1 \circ \gamma_2$ preserves equivalence classes and thus defines a multiplication on $\pi_1(X, x_0)$. Under this operation, π_1 is a group.*

Proof: This is very straightforward.

Remark: This set $\pi_1(X)$ depends on the base point x_0. If X is arcwise connected and $\tau(t)$ is a path from x_0 to x_1, then τ induces a one-to-one and onto group isomorphism between the fundamental group with base point x_0 and the group with base point x_1. We will often omit the base point from our notation.

Remark: We are ready to rigorously prove our previous assertions about π_1 for surfaces, namely that the fundamental group is generated by a_i and b_j where these are the paths previously discussed, modified to start and end at the base point x_0 as before. We will refer to our previous pictures, but in addition the following pictures will be important.



The first shows our standard construction of a torus. Opposite sides are glued together so the arrows match. All four corners become one point on the torus, and the sides become two closed curves a and b, which are indeed the generators of the fundamental group discussed earlier. We typically read off the boundary counterclockwise, writing "a" for a side with arrow pointing in our direction and "a^{-1}" for arrows in the opposite direction. Then a torus becomes $aba^{-1}b^{-1}$.

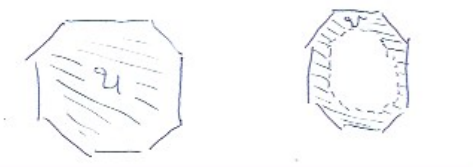
On a surface of higher genus g, the curves a_i and b_j can be drawn meeting at exactly one point. If we cut the surface along these curves, we get a disk or polygon, with two boundary segments for each curve. Reading counterclockwise, we obtain $a_1 b_1 a_1^{-1} b_1^{-1} \ldots a_g b_g a_g^{-1} b_g^{-1}$. All corners of the polygon glue to a single point on the surface, which we take to be the base point of the fundamental group. The reader can draw pictures necessary to show that this process reproduces our original pictures of surfaces with g holes.

Theorem 216. *The curves a_i and b_j generate $\pi_1(\mathcal{S})$, so every closed curve starting and ending at the base point is homotopic to a curve that is a product of $a_i^{m_i}$ and $b_j^{n_j}$.*

Remark: Such a symbol may appear multiple times, since the group is not commutative. We do not claim that such a product representation is unique.

The idea of the proof is very easy. Think of the surface as a polygon with boundary, as in the previous picture. Given a closed curve in the surface, we pick a point inside this polygon not on the curve, and define a homotopy by pushing the inside of the disk out to the boundary. In the end, we get a curve which lies directly on the a_i and b_j.

The problem is that space-filling curves exist, so our curve might go through every point of the disk. To get around this problem, let $\mathcal{U}$ be the interior of the polygon, and let $\mathcal{V}$ be a connected neighborhood of the boundary. See the picture below.



These open sets cover the surface. If $\gamma : [0,1] \to \mathcal{S}$ represents an element of the fundamental group, we can find a decomposition $0 = t_0 < t_1 < \ldots < t_n = 1$ such that each $\gamma([t_{i-1}, t_i])$ is inside $\mathcal{U}$ or $\mathcal{V}$. If the image of a segment is in $\mathcal{U}$, the segment is homotopic to a segment in $\mathcal{U}$ whose image is a straight line from $\gamma(t_i)$ to $\gamma(t_{i+1})$. After this homotopy, our curve may do awful things in $\mathcal{V}$, but the portion of the curve that is not in $\mathcal{V}$ lies on a finite number of line segments. Thus there is a point in $\mathcal{U}$ not on the image of the new curve. Apply a homotopy which pushes the entire polygon to the boundary, leaving the boundary fixed.

The boundary is a little tricky because glueing occurs there. On the other hand, each a_i or b_j on the boundary becomes a circle after glueing, and these circles only meet at the common base point. So we can ignore the remaining surface and suppose that our space X is a bouquet of circles, pictured below.

Again cover this with open sets. Let $\mathcal{U}$ be a contractable neighborhood of the base point, containing a small portion of each circle, as shown above. For each circle let $\mathcal{V}_i$ be the entire circle together with the set $\mathcal{U}$. So the $\mathcal{V}_i$ cover X and each contains one entire circle and contractable pieces of the remaining circles. As before, find $0 = t_0 < t_1 < ... < t_n = 1$ such that each $\gamma([t_{i-1}, t_i])$ is contained in some $\mathcal{V}_i$. If two adjacent segments have images in the same $\mathcal{V}_i$, merge them to form just one segment.

Consider γ restricted to $[0, t_1]$. This path starts at the base point and is always in some particular $\mathcal{V}_i$. Similarly γ restricted to $[t_1, t_2]$ is always in some $\mathcal{V}_j$ for $i \neq j$. So $\gamma(t_1)$ is in both $\mathcal{V}_i$ and $\mathcal{V}_j$ and thus in $\mathcal{U}$. We can thus find a $c_1 \leq t_1 \leq c_2$ such that $\gamma(t) \in \mathcal{U}$ for $c_1 < t < c_2$. Replace the path γ between time c_1 and time c_2 by a path which moves from $\gamma(c_1)$ at time c_1 to the base point at time t_1, and then from the base point to $\gamma(c_2)$ at time c_2. Since the space $\mathcal{U}$ is contractable, any two paths with the same end points are homotopic, so the new path is homotopic to the original but equal to the base point at t_1.

Repeat the same argument for all other division points t_i. We now have a path which is a sum of closed paths on each $[t_i, t_{i+1}]$ and on this interval has image in a particular circle and in $\mathcal{U}$. Since $\mathcal{U}$ is contractible, we can apply a homotopy to this segment which does not change the endpoints but leaves the entire image of the segment in a particular circle. Since this holds for each segment, our path is almost what we want. It stays in one circle from time 0 to time t_1 when it is back at the base point. Then it moves in some other circle from time t_1 to time t_2. Etc.

The only trouble is that our path may move erratically on the circles, perhaps backtracking and the like. Our final argument will show that we can replace each such path with a paths which move regularly: either sitting still, or moving counterclockwise m times at a steady pace, or moving clockwise in a similar way.

Lemma 29. *The fundamental group of a circle is Z, represented by steadily moving paths that travel around the circle clockwise or counterclockwise m times.*

Proof: Consider the map $R \to S^1$ by $r \to e^{2\pi ir}$. This is a covering map and the inverse image of $1 \in S^1$ is Z. Suppose $\gamma : [0,1] \to S^1$ is a path which begins and ends at 1. By covering space theory, this map has a unique lift to a map from $[0,1] \to R$ which starts at 0. This map must end at an integer n. It is easy to show that two maps are homotopy if and only if their lifts end at the same point, and the rest of the theorem then follows.

Remark: This completes the proof of the main theorem. We have not completely determined] the fundamental group of a surface, because the generators we obtained may satisfy additional relations. Indeed it is easy to find such a relation. Consider the picture of the surface as a polygon with sides identified. The path around the boundary of this polygon generates the element $a_1 b_1 a_1^{-1} b_1^{-1}$ ... in $\pi_1(\mathcal{S})$. But this element represents the identity because the path can be shrunk to the base point through the polygon.

Definition 61. *If X is an arcwise connected topological space, define the first homology group of X as*

$$H_1(\mathcal{S}) = \frac{\pi_1(\mathcal{S})}{[\pi_1(\mathcal{S}), \pi_1(\mathcal{S})]}$$

That is, we quotient out the commutator of the fundamental group, obtaining an abelian group.

Remark: The fundamental group depends on the base point, but since X is arcwise connected, the homology group is independent of this point up to isomorphism. By the main theorem of this section, the first homology group of a surface is a finitely generated abelian group with generators a_i and b_j. We will ultimately prove that these generators are independent and therefore the group is isomorphic to $Z \oplus Z \oplus ... \oplus Z$, where there are $2g$ copies of Z. It follows that g is a topological invariant of a surface, so two surfaces with different g cannot be homeomorphic.

28.2 Several-Variable Calculus and Riemann Surfaces

Long ago, we argued that complex variable theory is a special case of several-variable calculus. We are going to pick up that thread again.

Suppose $\mathcal{S}$ is a Riemann surface. This surface is covered by coordinate systems, and we have transition functions from one coordinate to another. Each coordinate system consists of an open set $\mathcal{U} \subset \mathcal{S}$, an open set $\mathcal{V} \subset C$, and a homeomorphism $\varphi : \mathcal{U} \to \mathcal{V}$. Since $\mathcal{V}$ is an open subset of the complex plane, it is also an open subset of R^2. The local coordinate system allows us to describe points in $\mathcal{U}$ using a complex coordinate z, but we could also use a pair of real coordinates x, y.

If we have two complex coordinates z and w, then in the overlapping region of these systems, w becomes a function of z and we require that $w(z)$ be holomorphic. If we have two real coordinates x, y and u, v, then in the overlapping region of these systems, u and v become functions of x and y and we require that $u(x, y), v(x, y)$ be C^∞. We could only require them to be differentiable, but it is convenient to go all the way and require derivatives of all orders.

When we think of a surface as having complex coordinates, we call $\mathcal{S}$ a *Riemann surface*. When we think of it as having real coordinates, we call it a *differentiable manifold*. We have more freedom if we are working with a differentiable manifold because we can choose new coordinates that are merely C^∞ compatible with the existing coordinate cover, rather than holomorphically compatible. It is very useful to think of a Riemann surface as a special kind of differentiable manifold. In the next chapter, we will pin down the extra things true of Riemann surfaces, but in this chapter we only assume that $\mathcal{S}$ is a differentiable manifold.

28.3 Differential Forms

On Riemann surfaces, we talk about functions and differentials. In coordinates they look similar: $f(z)$ for functions and $g(z)\,dz$ for differentials. But they transform differently. In exactly the same way, we have objects on differentiable manifolds which mimic the key objects of several variable calculus. These objects are called 0-forms, 1-forms, and 2-forms. In coordinates they look like

$$f(x, y)$$

$$\omega_1(x, y)dx + \omega_2(x, y)dy$$

$$\Omega(x, y)dx \wedge dy$$

A zero form is just a function and a one-form is similar to a vector field.

The transition rules when coordinates change are easy to remember because Leibniz' notation tells us what to do. If we have coordinates u, v and x, y and $u(x, y), v(x, y)$ are the transition equations, then

$$f(u, v) \to f(u(x, y), v(x, y))$$

$$\omega_1(u, v)du + \omega_2(u, v)dv \to \omega_1\left(\frac{\partial u}{\partial x}dx + \frac{\partial u}{\partial y}dy\right) + \omega_2\left(\frac{\partial v}{\partial x}dx + \frac{\partial v}{\partial y}dy\right)$$

$$\Omega(u, v)du \wedge dv \to \Omega\left(\frac{\partial u}{\partial x}dx + \frac{\partial u}{\partial y}dy\right) \wedge \left(\frac{\partial v}{\partial x}dx + \frac{\partial v}{\partial y}dy\right) = \Omega\left(\frac{\partial u}{\partial x}\frac{\partial v}{\partial y} - \frac{\partial u}{\partial y}\frac{\partial v}{\partial x}\right)dx \wedge dy$$

On Riemann surfaces, the derivative maps functions to differentials: $f(z) \to \frac{df}{dz}\, dz$. Similarly on two-dimensional manifolds, there are maps from functions to one-forms and from one-forms to two-forms:

$$f(x,y) \stackrel{d}{\longrightarrow} \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy$$

$$\omega_1(x,y)dx + \omega_2(x,y)dy \stackrel{d}{\longrightarrow} \left(\frac{\partial \omega_2}{\partial x} - \frac{\partial \omega_1}{\partial y}\right) dx \wedge dy$$

In graduate courses both maps are called "d", but in several variable calculus the first map is called the gradient and the second map is called the curl. Note that $d^2 = 0$.

The coefficients of forms are usually real-valued, but the theory works exactly the same if we allow complex-valued coefficients. That is because the real and imaginary parts are treated separately and never combined by our d-maps. While first reading this section, think of the coefficients are real-valued functions. We will apply the theory when they are complex valued.

Definition 62. *Let $\Lambda^k(\mathcal{S})$ be the set of all global k-forms on the surface. Thus Λ^0 is the set of C^∞ functions, Λ^1 is the set of global 1-forms $\omega_1 dx + \omega_2 dy$, and Λ^2 is the set of global 2-forms $f(x,y)dx \wedge dy$. Recall that we have maps named d with square zero:*

$$0 \to \Lambda^0 \stackrel{d}{\longrightarrow} \Lambda^1 \stackrel{d}{\longrightarrow} \Lambda^2 \to 0$$

Let the coefficients be complex-valued. Then for $k = 0, , 1, 2$ define a complex vector space

$$H^k(\mathcal{S}) = Ker(\Lambda^k \stackrel{d}{\longrightarrow} \Lambda^{k+1})/Im(\Lambda^{k-1} \stackrel{d}{\longrightarrow} \Lambda^k)$$

The $H^k(\mathcal{S})$ are called the deRham cohomology groups of the surface. They are only non-zero for H^0, H^1, and H^2.

Remark: We will soon see that these groups contain all of the crucial topological information about the surface.

Example: Note that $H^0(\mathcal{S})$ is the set of all functions $f(x,y)$ with $df = 0$. Thus

$$\frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy = 0$$

so both partials vanish and f is locally constant. Since our surfaces are connected, f is globally constant. Therefore $H^0(\mathcal{S}) = C$, the set of all constant functions on the surface.

28.4 Calculating Cohomology Groups

There is a central theme of this section. Suppose M is a manifold, and suppose K is a submanifold of dimension k. Then we can integrate k forms on M over K. If two k forms define the same element of $H^k(M)$, they have the same integral by a generalized Stokes theorem. So integration over K defines a map $H^k(M) \to C$. If two embeddings of K are homotopic, integration using either embedding gives the same map from cohomology, again by Stokes' theorem. So the map depends only on rough topological properties of K, not on analytical details of the object. Therefore, geometric objects inside M can detect non-trivial elements of cohomology.

In the case of surfaces, integration over points can detect elements of H^0, integration over closed curves can detect elements of H^1, and integration over the whole surface can detect elements of H^2. This becomes clear as soon as we give examples.

Example 1: To start off, a trivial example. Suppose K is an isolated point p of M. These points are "geometrical objects of dimension zero." A 0-form on M is just a function, and integration of this function over a point is evaluation at p. So we get a map $H^0(M) \to C$ given by $f \to f(p)$. Since elements of $H^0(M)$ are constant functions, this map is an isomorphism. So integration over one point completely determines an element of H^0.

Example 2: The group $H^2(\mathcal{S})$ is the set of all two forms $f(x,y)\,dx\wedge dy$, modulo 2-forms equal to $d\omega$. Two forms on a 2-dimensional surface can be integrated over the surface. Indeed suppose $\Omega = f(x,y)\,dx\wedge dy$ where f is non-zero only on a compact subset of a coordinate neighborhood $\mathcal{V}$. We can compute

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dxdy$$

If u, v is a second coordinate system with transition functions $u(x,y), v(x,y)$, we get the same result because

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(u,v)\,dudv = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(u(x,y),v(x,y))\,\det\begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{pmatrix} dxdy$$

and this Jacobian determinant is provided by the transformation rule for $du \wedge dv$. However, there is a slight complication. When we change variables in integrals, we must also change the limits of integration. If we integrate from $-\infty$ to ∞, this isn't possible and instead the absolute value of the determinant should appear in the integral. On a Riemann surface, the Jacobian determinant is positive because holomorphic maps preserve orientation. On a surface with g holes, we can choose coordinates so the transition functions always have positive Jacobian, and we assume this done. There would be trouble applying this theory to Klein bottles and the like, but we will never do that.

If the 2-form is not supported by a single coordinate system, the surface can be cut into pieces, each in a single coordinate open set, and the integral can be calculated as a sum of integrals

over these pieces. From now on, we denote this integral

$$\int_S \Omega$$

Lemma 30. *If ω is a 1-form,*

$$\int_S d\omega = 0$$

Proof: This is true by Green's theorem. According to that theorem,

$$\int\int_{\mathcal{R}} \left(\frac{\partial \omega_2}{\partial x} - \frac{\partial \omega_1}{\partial y} \right) dxdy = \int_{\partial \mathcal{R}} \omega \cdot d\gamma$$

where $\mathcal{R}$ is a region in the plane and $\partial\mathcal{R}$ is its counterclockwise boundary. Cut up the surface into regions, each in a coordinate system, and apply the theorem. The double integral of $d\omega$ becomes a sum of integrals around the boundaries of these regions, but this sum cancels because adjacent boundary components are integrated in opposite directions.

It follows that integration of 2-forms induces a well-defined map

$$H^2(S) \to C$$

If a 2-form has non-zero integral, then it represents a non-zero element of $H^2(M)$. In fact it can be shown that $H^2(S) = C$ and the integration map is an isomorphism, but we will not need that result.

Example 3: Finally, suppose $\gamma : [a, b] \to S$ is a parameterized path. If $\omega = \omega_1 dx + \omega_2 dy$ is any 1-form, we can define the line integral of ω over γ by the well-known formula

$$\int_\gamma \omega = \int_a^b \left(\omega_1, \omega_2 \right) \cdot \left(\frac{d\gamma_1}{dt}, \frac{d\gamma_2}{dt} \right) dt$$

This is independent of coordinates and induces the contour integral of complex analysis which we have used since the beginning of these notes.

If there is a function f with $df = \omega$, and if γ is a closed curve, then $\int_\gamma df = 0$ because

$$\int_a^b df = \int_a^b \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right) \cdot \left(\frac{d\gamma_1}{dt}, \frac{d\gamma_2}{dt} \right) dt = \int_a^b f(\gamma(t))\, dt = f(\gamma(b)) = f(\gamma(a)) = 0$$

It follows that line integration over γ gives a well-defined map

$$H^1(S) \to C$$

Remark: Integration of one-forms over closed curves is the most important case for us, and we can improve the above treatment. As explained in our summary of several variable calculus in chapter five, if ω is a one-form and $d\omega = 0$, then locally there is a function f with $df = \omega$. In other words, if we are given $\omega_1(x,y)$ and $\omega_2(x,y)$ and $\frac{\partial \omega_2}{\partial x} - \frac{\partial \omega_1}{\partial y} = 0$, then we can find $f(x,y)$ with $\frac{\partial f}{\partial x} = \omega_1$ and $\frac{\partial f}{\partial y} = \omega_2$. In that case, we can compute line integrals locally using

$$\int_\gamma \omega_1 dx + \omega_2 dy = \int_\gamma df = f(\text{end}) - f(\text{start})$$

In chapter 7, we made contour integration, Cauchy's theorem, etc., rigorous by using this local information. What was done there works in exactly the same way when integrating one-forms ω satisfying $d\omega = 0$ on an arbitrary differentiable manifold, and we obtain the following theorem:

Theorem 217. *Let ω be a one-form on a differentiable surface and suppose $d\omega = 0$. Then*

- *It is possible to define $\int_\gamma \omega$ for any merely continuous path γ on the surface. The definition agrees with the definition given above if the path is differentiable.*
- *Suppose γ_1 and γ_2 are continuous closed paths on the surface and they are homotopic by a homotopy through closed paths. Then*

$$\int_{\gamma_1} \omega = \int_{\gamma_2} \omega$$

- *Suppose γ_1 and γ_2 are continuous closed paths which start at P and end at Q and suppose these paths are homotopic by a homotopy through paths which all start at P and end at Q. Then*

$$\int_{\gamma_1} \omega = \int_{\gamma_2} \omega$$

Remark: A dubious reader who look back at chapter 7 and notice that the proofs work in this case with essentially no change.

Remark: We defined the homology group $H_1(S)$ using homotopy classes of merely continuous closed paths on the surface, and we defined the deRham cohomology group $H^1(S)$ to contain equivalence classes of one-forms ω satisfying $d\omega = 0$. Using the previous theorem, we can find a beautiful connection between these two objects. Suppose γ is a closed continuous path representing an element of H_1, and suppose ω is a closed one-form representing an element of H^1. We can then form the complex number

$$\int_\gamma \omega$$

Theorem 218. *The number*

$$\int_\gamma \omega$$

depends only on the equivalence class of γ in H_1 and only on the equivalence class of ω in H^1. In particular, if this integral is not zero, then γ and ω represent non-zero elements of H_1 and H^1.

Proof: If γ_1 and γ_2 represent the same element of H_1, then they are homotopic through closed paths and consequently give the same integral of ω.

If ω_1 and ω_2 represent the same element of H^1, then $\omega_1 - \omega_2 = df$ for some function f. Since γ is a closed curve, it follows that

$$\int_\gamma \omega_1 - \int_\gamma \omega_2 = \int_\gamma df = f(\text{end of } \gamma) - f(\text{beginning of } \gamma) = 0$$

28.5 Generators of $H^1(\mathcal{S})$

We are going to complete the ideas of the previous section by proving that $H_1(\mathcal{S}) = Z \oplus Z \oplus ... \oplus Z$ where the various Z are generated by the a_i and b_j. We will simultaneously prove that $H^1(\mathcal{S}) = C^{2g}$.

Let a_i and b_j be the elements of H_1 induces by these closed paths. Define a map $H^1(\mathcal{S}) \to C^{2g}$ given by

$$\omega \to \left(\int_{a_1} \omega, \dots, \int_{a_g} \omega, \int_{b_1} \omega, \dots, \int_{b_g} \omega \right)$$

Lemma 31. *The above map is is one-to-one.*

Proof: Suppose ω is a closed one-form, and all $2g$ integrals above are zero. Let $p_0 \in \mathcal{S}$ be the base point for the fundamental group. Define a function f on $\mathcal{S}$ by

$$f(p) = \int_\gamma \omega$$

where γ is a curve from p_0 to p. We must show that f is well-defined, or equivalently that this integral does not depend on the particular path. If γ_1 and γ_2 are two possible paths, then γ_1 followed by the inverse of γ_2 is a closed path starting and ending at p_0 and it suffices to show that the integral around this path is zero. This follows because any closed path is homotopic to a product of a_i and b_j.

To finish the proof, it suffices to show that $df = \omega$, because then ω represents 0 in H^1. But we already gave that argument in section 5.6 when we reviewed advanced calculus. QED.

Theorem 219. *The map $H^1 \to C^{2g}$ given by*

$$\omega \to \left(\int_{a_1} \omega, \dots, \int_{a_g} \omega, \int_{b_1} \omega, \dots, \int_{b_g} \omega \right)$$

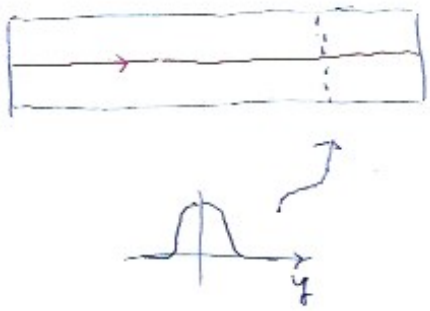
is onto. Consequently, this map is an isomorphism.

Proof: The proof uses a very beautiful construction. From the picture below, we see that each of the paths a_i and b_j is in a ribbon in the surface. A coordinate system can be created for each ribbon making it homeomorphic to the strip $[0, L] \times (-1, 1)$ in the plane.

In this coordinate system, define a form

$$\omega = 0\, dx + f(y)\, dy$$

where $f(y)$ is a "bump function" which is only non-zero very close to $y = 0$ and yet has integral over y equal to 1. See the pictures below. Notice that $d\omega = 0$ because f is independent of x. A careful reader might wonder if such a C^∞ function even exists. We will construct one at the end of the proof. Notice that the path is in the center of this ribbon, moving horizontally.



We have reverted to thinking of the a_i and b_j as originally presented, where base points are ignored. Notice that a_i and b_i intersect, but these paths do not intersect any other a_j or b_j. If the strips are narrow enough, this also holds for them.

Let us label the resulting 1-forms ω_{a_i} and ω_{b_j}. The only paths intersecting the a_i and b_i strips are a_i and b_i, so all other integrals of ω_{a_i} and ω_{b_i} are zero. Surprisingly,

$$\int_{a_i} \omega_{a_i} = 0 \quad \text{and} \quad \int_{b_i} \omega_{b_i} = 0$$

The reason is that these integrals only depend on the path up to homotopy and we can slide a_i up or down the strip until it no longer intersects the region where $\omega_{a_i} \neq 0$. However

$$\int_{a_i} \omega_{b_i} \neq 0 \quad \text{and} \quad \int_{b_i} \omega_{a_i} \neq 0$$

The reason is that b_i intersects the a_i strip by moving from top to bottom, and sliding it in its strip just slides left or right in the a_i strip, so it still intersects. In local coordinates where it matters, the b_i path will be $\gamma(t) = (0, t)$ and the integral of ω_{a_i} will become

$$\pm \int_{-1}^{1} f(y)\, dy = 1$$

This calculation shows that ω_{a_i} maps to an element of C^{2g} with only one nonzero component, and ω_{b_i} maps to another element of C^{2g} with only one nonzero components, and as we vary ω we obtain a complete basis of C^{2g} as image elements, so our map is onto. QED.

Corollary 35. *The elements a_i and b_j in $H_1(\mathcal{S})$ are independent, so this homology group is the full $Z \oplus ... Z$ using $2g$ copies of Z.*

Proof: If not, then some $\sum(m_i a_i + \sum n_j b_j)$ in H_1 is zero in the group. Therefore if we integrate any $\omega \in H^1$ over this path, we must get zero. But when we integrate, say, ω_{b_1} over this path, we get the coefficient of a_1. QED.

Remark: At this point, we know that $H^0(\mathcal{S}) = C, H^1(\mathcal{S}) = C^{2g}$, and $H^2(\mathcal{S}) = C$ or larger.

28.6 The Cohomology Product

Remark: There is a product structure on $\oplus H^k(X)$ induced by the wedge product. Indeed, if λ is a k form and ω is an l form, then $\lambda \wedge \omega$ is a $k + l$ form. Moreover,

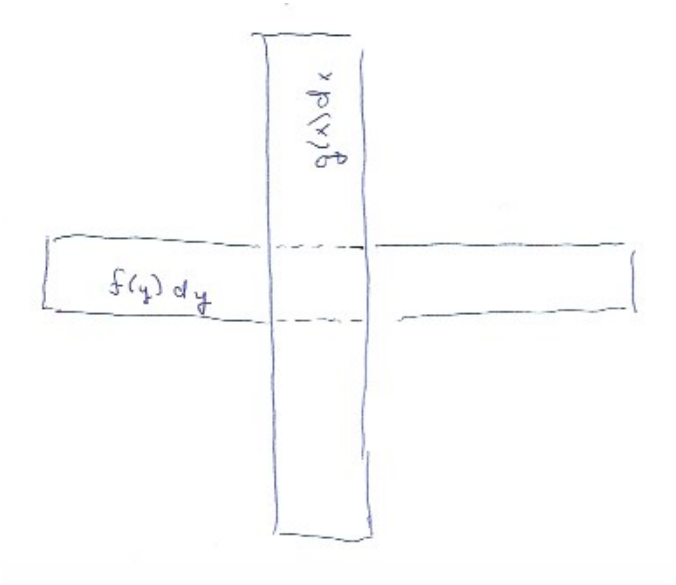
$$d(\lambda \wedge \omega) = (d\lambda) \wedge \omega + (-1)^k \lambda \wedge (d\omega)$$

$$\lambda \wedge \omega = (-1)^{kl} \omega \wedge \lambda$$

A tedious but elementary calculation then shows that there is an induced product

$$H^k \times H^l \to H^{k+l}$$

For surfaces, the only important product is $H^1 \times H^1 \to H^2 = C$. We now compute this product for a surface of genus g. Note first that $\lambda \wedge \omega = -\lambda \wedge \omega$ for these particular dimensions, so the product of an element with itself is zero. Since the supports of ω_{a_i} and ω_{a_j} do not intersect if $i \neq j$, with similar statements for ω_{a_i} and ω_{b_j} and ω_{b_i} and ω_{b_j}, the only interesting products are $\omega_{a_i} \wedge \omega_{b_i}$. These live on two strips which intersect in a square, as seen below:



The product is zero except on the intersection square, where it is $f(x)f(y)\ dx \wedge dy$. This element represents the desired product, which lives in $H^2(\mathcal{S})$ and is mapped to C by

$$\int_{-1}^{1}\int_{-1}^{1} f(x)f(y)\ dxdy = 1$$

The result, however, depends on the orientation of the curves a_i and b_j, so the result is ± 1.

We conclude that in cohomology, $\omega_{a_i} \wedge \omega_{a_j} = 0, \omega_{b_i} \wedge \omega_{b_j} = 0$, if $i \neq j$ then $\omega_{a_i} \wedge \omega_{b_j} = 0$, and $\omega_{a_i} \wedge \omega_{b_i} = -\omega_{b_i} \wedge \omega_{a_i} = \pm 1$. Note that we are free to change the orientations of a_i and b_j, since nothing we have done so far depends on these orientations. The convention is to select orientations so $\omega_{a_i} \wedge \omega_{b_i} = 1$.

Definition 63. *The map $H^1(\mathcal{S}) \times H^1(\mathcal{S}) \to C$ defined by*

$$\omega_1 \times \omega_2 \to \int\int \omega_1 \wedge \omega_2$$

is called the intersection pairing.

Remark: This is well-defined and independent of particular representatives because it is given by the cohomological product $H^1 \times H^1 \to H^2$ followed by the previously defined integration map $H^2 \to C$.

Theorem 220. *The intersection pairing is given by the following formula:*

$$\int\int \omega_1 \wedge \omega_2 = \sum_i \left(\int_{a_i} \omega_1 \int_{b_i} \omega_2 - \int_{b_i} \omega_1 \int_{a_i} \omega_2 \right)$$

Remark: This is the first moment in our development when the orientation of the paths a_i and b_j matters. In all previous results, we could change the direction of paths arbitrarily without changing the final result. We still allow the orientation of the a_i to be chosen arbitrarily, but b_i intersects a_i exactly once and at that spot, we require that the angle from a_i to b_i be negative, where angle is measured between $-\pi$ and π. Said another way, if we select coordinates in which the intersection point is the origin and a_i moves along the positive x-axis, then b_i must cross into the lower half plane at the origin. Once we have orientations of the a_i, this fixes the orientations of all b_i.

Proof: The formula depends on the choice of the paths a_i and b_i, but otherwise both sides are well-defined invariant expressions in cohomology. Both sides are linear in ω_1 if ω_2 is fixed, and linear in ω_2 if ω_1 is fixed. Finally, both sides are skew-symmetric in ω_1 and ω_2.

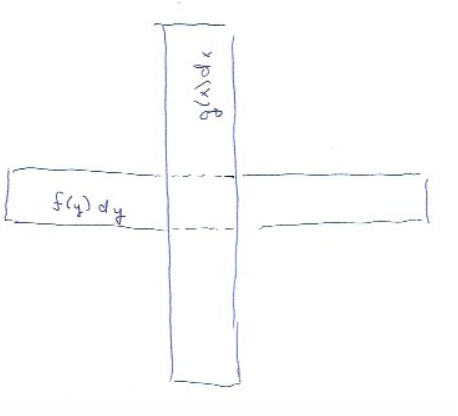
Thus it suffices to check the formula using particular representatives of the generators of the cohomology group H^1. We know that the ω_{a_i} and ω_{b_j} generate H^1. Using bilinearity and skew-symmetry, it suffices to check the following three cases:

- when $\omega_1 = \omega_{a_i}, \omega_2 = \omega_{a_j}$
- when $\omega_1 = \omega_{a_i}, \omega_2 = \omega_{b_j}$
- when $\omega_1 = \omega_{b_i}, \omega_2 = \omega_{b_j}$

Both sides vanish in the first and third cases, and in the second case unless $i = j$. When $i = j$, the formula reads

$$\int\int \omega_{a_i} \wedge \omega_{b_i} = -\int_{b_i} \omega_{a_i} \int_{a_i} \omega_{b_i}$$

This expression involves three integrals, and our previous calculations show that all three integrals are ± 1. Thus the required result is true modulo a sign. To determine the sign, we need to revisit the following picture:



In this picture, the horizontal rectangle represents a strip with the path a_i in the center, moving to the right. This determines the x-axis, and we then determine the direction of the y-axis so the standard orientation in the plane is the orientation on the surface. Similarly the vertical strip represents a strip with the path b_i in the center, moving down. We can use the same x and y coordinates in both pictures and let the path b_i move down along the y-axis, so the narrow coordinate of the vertical strip is x. Notice that the orientation on this vertical strip is correct because of our choice of orientation for b_i. In this joint coordinate system, $\omega_{a_i} = f(y)dy$ and $\omega_{b_i} = f(x)dx$. Thus

$$\int\int \omega_{a_i} \wedge \omega_{b_i} = \int_{-1}^{1}\int_{-1}^{1} f(x)f(y)\, dxdy = 1$$

Similarly,

$$\int_{b_i} \omega_{a_i} = \int_{1}^{-1} f(y)\, dy = -1$$

and

$$\int_{a_i} \omega_{b_i} = \int_{-1}^{1} f(x)dx = 1$$

and the formula

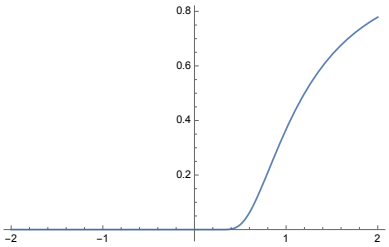
$$\int\int \omega_{a_i} \wedge \omega_{b_i} = -\int_{b_i} \omega_{a_i} \int_{a_i} \omega_{b_i}$$

is valid. QED.

28.7 Appendix

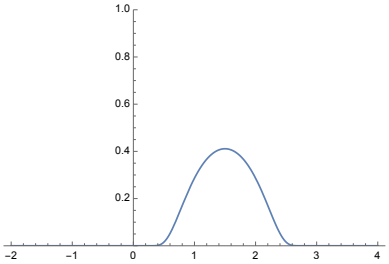
As promised, we construct a C^∞ bump function. Define

$$h_1(x) = \begin{cases} e^{-\frac{1}{x^2}} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$



This function is obviously C^∞ off the origin; the reader can show that it is also C^∞ at the origin. The trick is to notice that all of its derivatives to the right have the form $P\left(\frac{1}{x}\right)e^{-\frac{1}{x^2}}$ where P is a polynomial.

Let $h_2(x) = h_1(x)h_1(3-x)$.



This function is ≥ 0 and zero except on $[0, 3]$. It can be translated to be centered at the origin, and scaled to be as narrow as desired and then to have area one.

Remark: It follows that C^∞ functions do not satisfy the identity theorem. This is the major difference between Riemann surface theory and the theory of differentiable manifolds. The identity theorem severely constrains complex analysis, but a consequence is that the theory produces surprising results. The constraints imposed on C^∞ functions are mild, so differentiable manifolds are an ideal place to study topology and its methods of deforming, pushing here, and pulling there.

Chapter 29

Cohomology and Riemann Surfaces

The Riemann surface is thus a two-dimensional manifold, but with something extra: we only allow coordinate systems compatible with the defining coordinates using holomorphic transition functions, rather than arbitrary C^∞ maps. There is a very beautiful way to capture this additional information. Differentials have basis dx, dy, but when we use complex coefficients, we can use a different basis $dz = dx + idy$ and $d\overline{z} = dx - idy$. Some authors says that forms $f(x,y)dz$ have type Λ_{10} and forms $f(x,y)d\overline{z}$ have type Λ_{01}. The key observation is

Theorem 221. *If the transition functions from one coordinate systems to another are holomorphic, then forms of type Λ_{10} in one coordinate system are forms of type Λ_{10} in the second coordinate system. Conversely if this is true for a coordinate change, then the coordinate transition functions are holomorphic.*

Proof: This is just the Cauchy-Riemann equations in disguise. Suppose one coordinate system is (x,y) and a second is (u,v) and these are related by $u(x,y), v(x,y)$. Then in one system the Λ_{10} forms are

$$du + idv = \left(\frac{\partial u}{\partial x}dx + \frac{\partial u}{\partial y}dy\right) + i\left(\frac{\partial v}{\partial x}dx + \frac{\partial v}{\partial y}dy\right) = \left(\frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x}\right)dx + \left(\frac{\partial u}{\partial y} + i\frac{\partial v}{\partial y}\right)dy$$

We want the right side to be a function $f(x,y)$ times $dz = dx + idy$, and such an f exists exactly when

$$i\left(\frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x}\right)dx = \left(\frac{\partial u}{\partial y} + i\frac{\partial v}{\partial y}\right)$$

This equation is the Cauchy-Riemann equations in disguise.

Remark: On an arbitrary two-dimensional manifold, it is silly to talk about the basis dz and $d\overline{z}$ because this decomposition is not coordinate invariant. But on a Riemann surface, it is invariant and thus important.

Definition 64. *Suppose $f(x,y)$ is a complex-valued C^∞ function on a Riemann surface, not necessarily holomorphic. Then we define $\frac{\partial f}{\partial z}$ and $\frac{\partial f}{\partial \overline{z}}$ by the formula*

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

Remark: It is easy to deduce the formulas for these operators:

$$\frac{\partial f}{\partial z} = \frac{1}{2}\left(\frac{\partial}{\partial x} - i\frac{\partial}{\partial y}\right)(f_1 + if_2)$$

$$\frac{\partial f}{\partial \overline{z}} = \frac{1}{2}\left(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right)(f_1 + if_2)$$

Theorem 222. *A function $f(x,y)$ is holomorphic if and only if $\frac{\partial f}{\partial \overline{z}} = 0$ and in that case*

$$\frac{df}{dz} = \frac{\partial f}{\partial z}$$

Proof: Cauchy-Riemann.

Remark: We are now ready to apply the previous deRham cohomology theory to Riemann surfaces. The crucial additional information in the case of a Riemann surface is that differential forms

$$\omega_1(x,y)dx + \omega_2(x,y)dy$$

can also be written

$$f_1(x,y)dz + f_2(x,y)d\overline{z}$$

and this new decomposition is invariant under holomorphic coordinate changes. The natural question to ask is whether this decomposition drops down to the cohomological level, so we can write

$$H^1(\mathcal{S}) = H^{10}(\mathcal{S}) \oplus H^{01}(\mathcal{S})$$

where elements in the first space are represented by forms $f(x,y)dz$ and those in the second space are represented by $f(x,y)d\overline{z}$. The surprising answer is "yes", but this is a deep theorem.

Let us concentrate for a moment on elements of $H^1(\mathcal{S})$ which can be represented by a form $\omega = f_1(x,y)dz$. Such a form represents a cohomology element if $d\omega = 0$, and by our earlier results this is equivalent to

$$\left(\frac{\partial f}{\partial z}dz + \frac{\partial f}{\partial \overline{z}}d\overline{z}\right) \wedge dz = 0$$

Note that $dz \wedge dz = 0$ and $d\overline{z} \wedge dz = 2idx \wedge dy$, so this equation is equivalent to the equation

$$\frac{\partial f}{\partial \overline{z}} = 0$$

Functions which satisfy this equation are holomorphic. So surprisingly, $H^{10}(\mathcal{S})$ would contain elements of the form $f(z)\,dz$ for holomorphic f, that is, holomorphic differentials.

Two such forms would be equivalent exactly when their difference had the form

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

But then $\frac{\partial f}{\partial \overline{z}} = 0$, so f would be a global holomorphic function, so constant, and $\frac{\partial f}{\partial z} = \frac{df}{dz} = 0$.

We have proved

Theorem 223. *Every holomorphic differential induces an element of $H^1(\mathcal{S})$. The map from the vector space of holomorphic differentials to the first cohomology group is one-to-one.*

Remark: This is a truly astonishing theorem, and it illustrates the difference between Riemann surface theory and the theory of two-dimensional manifolds. Topology is a rough theory which allows us to deform objects and identifies elements that differ by appropriate deformations. Thus elements in the deRham cohomology group has many representatives. Even the elements ω_a which we have defined as non-zero only in a small strip about the path a have many representatives, since the strip can be made arbitrarily narrow.

But complex variable theory is a very rigid theory, and on compact Riemann surfaces its objects do not admit deformations. And indeed, each cohomology class of H^1 admits *at most one* representative that is a holomorphic differential.

Not all classes in H^1 can be represented by holomorphic differentials. Indeed

Theorem 224. *If ω and η in $H^1(\mathcal{S})$ are represented by holomorphic differentials, then*

$$\omega \wedge \eta \in H^2(\mathcal{S})$$

is zero.

Proof: This is trivial because $dz \wedge dz = 0$.

Definition 65. *Let*

$$H^{10}(\mathcal{S}) \subset H^1(\mathcal{S})$$

be the set of all elements of H^1 with representatives of the form $f(x,y)dz$. Define $H^{01}(\mathcal{S})$ similarly with representatives $f(x,y)d\overline{z}$.

Theorem 225. *The intersection of these two subspaces is* $\{0\}$*. The space* H^{10} *is the space of all holomorphic differentials, and the space* H^{01} *is the space of all conjugates* $\overline{f(z)}d\overline{z}$ *of holomorphic differentials. Hence the dimension of the space of holomorphic differentials is at most g, and it is exactly g precisely when* $H^{10} \oplus H^{01} = H^1$*.*

Proof: This can safely be left to the reader.

29.1 Riemann's Bilinear Relations

There is a natural Hermitian inner product on the space of holomorphic differentials. Indeed suppose ω_1 and ω_2 are holomorphic differentials. Then $\overline{\omega_1} \in H^{01} \subset H^1$ and $\omega_2 \in H^{10} \subset H^1$ and so $\overline{\omega_1} \wedge \omega_2 \in H^2$ and we can integrate.

Definition 66. *If* ω_1 *and* ω_2 *are holomorphic differentials,*

$$< \omega_1, \omega_2 > = \; -i \int \int \overline{\omega_1} \wedge \omega_2$$

Remark: Note that $d\overline{z} \wedge dz = (dx - idy) \wedge (dx + idy) = 2idx \wedge dy$, so this inner product in coordinates equals

$$-i \int \int \overline{f_1(z)} f_2(z)\, d\overline{z} \wedge dz = 2 \int \int \overline{f_1(z)} f_2(z)\, dxdy$$

Clearly this expression is antilinear in the first variable, linear in the second variable. Clearly $\overline{< \omega_1, \omega_2 >} =< \omega_2, \omega_1 >$. And finally, if $\omega \neq 0$ then $< \omega, \omega >$ is positive.

We can rewrite the formula for this inner product using the intersection formula

$$\int \int \omega_1 \wedge \omega_2 = \sum_i \left(\int_{a_i} \omega_1 \int_{b_i} \omega_2 - \int_{b_i} \omega_1 \int_{a_i} \omega_2 \right)$$

We obtain

$$< \omega_1, \omega_2 > = \; -i \sum_i \left(\int_{a_i} \overline{\omega_1} \int_{b_i} \omega_2 - \int_{b_i} \overline{\omega_1} \int_{a_i} \omega_2 \right)$$

and in particular

$$< \omega, \omega > = \; 2 \sum_i \Im \left(\overline{\int_{a_i} \omega} \int_{b_i} \omega \right)$$

Theorem 226. *Let $\mathcal{H}$ be the vector space of holomorphic differentials. If the Riemann surface is hyperelliptic, we proved that the complex dimension of $\mathcal{H}$ is g. In fact, it is always g, but we will not prove that. But if it is true, then it is possible to find a basis $\omega_1, \dots, \omega_g$ for $\mathcal{H}$ such that*

$$\int_{a_i} \omega_j = \delta_{ij}$$

Proof: Consider the map $\mathcal{H} \to C^g$ given by

$$\omega \to \left(\int_{a_1} \omega, \dots, \int_{a_g} \omega \right)$$

If this map is an isomorphism, the theorem immediately follows. Otherwise, since $\mathcal{H}$ and C^g have the same dimension, the map must have a non-trivial kernel. Thus there is a holomorphic differential $\omega \neq 0$ such that $\int_{a_i} \omega = 0$ for all i. By then by the previous theorem, $< \omega, \omega > = 0$, which cannot happen because $\omega \neq 0$. QED.

Remark: Let $\omega_1, \dots, \omega_g$ be such a basis for the holomorphic differentials. We obtain a $g \times g$ complex matrix by writing $B_{ij} = \int_{b_i} \omega_j$. A knowledge of this matrix gives complete information on the periods of our basis elements, since

$$\int_{a_i} \omega_j = \delta_{ij} \quad \text{and} \quad \int_{b_i} \omega_j = B_{ij}$$

Theorem 227. *The complex matrix B has the following properties:*

- *B is symmetric*
- *the imaginary part of B is positive-definite*

Proof: Since ω_i and ω_j are holomorphic 1-forms, $\omega_i \wedge \omega_j = 0$ and so the integral of their wedge product is zero. But for closed 1-forms, the intersection formula gives

$$\int \int \omega_1 \wedge \omega_2 = \sum_i \left(\int_{a_i} \omega_1 \int_{b_i} \omega_2 - \int_{b_i} \omega_1 \int_{a_i} \omega_2 \right)$$

In our case this formula says that

$$0 = \sum_k \left(\delta_{ki} B_{kj} - B_{ki} \delta_{kj} \right) = B_{ij} - B_{ji}$$

On the other hand, we have

$$< \omega_i, \omega_j > = -i \sum_k \left(\int_{a_k} \overline{\omega_i} \int_{b_k} \omega_j - \int_{b_k} \overline{\omega_i} \int_{a_k} \omega_j \right) = -i \sum_k (\delta_{ki} B_{kj} - \overline{B_{ki}} \delta_{kj})$$

So

$$< \omega_i, \omega_j >= -i\,(B_{ij} - \overline{B_{ji}}) = -i(B_{ij} - \overline{B_{ij}}) = -i(2i\mathfrak{I}B_{ij}) = 2\mathfrak{I}B_{ij}$$

By definition, a symmetric real matrix is positive definite if for any non-zero real vector $(r_1, \dots, r_g)$ we have $\sum r_i B_{ij} r_j > 0$, and that holds for our matrix because $< \sum r_i\omega_i, \sum r_i\omega_i >$ is positive because our inner product is Hermitian. QED.

Remark: Suppose we a dealing with a torus based on the lattice with basis 1 and τ. We can choose a to be the horizontal path from the origin to 1, and we can choose b to be the path from the origin to τ. A basis for the holomorphic differentials is dz. Then the periods of this differential are $\int_a dz = 1$ and $\int_b dz = \tau$. So these periods recover the lattice basis.

We can generalize this to an arbitrary compact Riemann surface. We replace dz by $\omega_1, \dots, \omega_g$. For each i, we form the periods

$$P_i = \left(\int_{a_1} \omega_i, \int_{a_2} \omega_i, \dots, \int_{a_g} \omega_i \right)$$

and

$$Q_i = \left(\int_{b_1} \omega_i, \int_{b_2} \omega_i, \dots, \int_{b_g} \omega_i \right)$$

Now each period is a g-dimensional vector in C^g. We have $2g$ of these vectors, two for each ω_i. In fact, the initial vectors in the pairs are the standard basis vectors $e_1, \dots, e_g$ for C^g. These vectors are the analogs of $1 \in C$. The second vectors are more general elements of C^g, the analogs of τ.

Theorem 228. *The $2g$ period vectors just defined in C^g are linearly independent over the reals, and thus form a lattice basis for a lattice $L \subset C^g$ of maximal rank.*

Proof: Suppose r_i and s_i are real numbers and $\sum r_i P_i + \sum s_i Q_i = 0$. Notice that $P_i = e_i$ are the standard basis vectors of C^g and $Q_i = (B_{1i}, B_{2i}, \dots, B_{gi})$. Each B_{ij} has a real part and an imaginary part. Since we are looking at a linear combination with real coefficients, The sum represents a vector in C^g and the imaginary parts of the coefficients are $\sum s_i \mathfrak{I}(Q_i)$ where $\mathfrak{I}Q_i$ are vectors with coefficients $(\mathfrak{I}B_{1i}, \mathfrak{I}B_{2i}, \dots, \mathfrak{I}B_{gi})$. But the matrix $\mathfrak{I}B$ is positive-definite, so this sum can only vanish if all the $s_i = 0$. Then since the P_i are the standard basis, the $r_i = 0$ as well.

Definition 67. *The* Jacobian *of a compact Riemann surface $\mathcal{S}$ is the complex torus C^g/L.*

Remark: If $g = 1$, this reconstructs the original Riemann surface, showing that all surfaces with $g = 1$ are conformally equivalent to period parallelograms of lattices in the complex plane. If $g > 1$, it associates a higher dimensional torus with the surface.

Remark: Suppose p and q are points on $\mathcal{S}$. Find a path γ from p to q and compute

$$\left(\int_\gamma \omega_1, \int_\gamma \omega_2, \dots, \int_\gamma \omega_g\right) \in C^g$$

This point determines the integral over γ of every holomorphic differential because such a differential is a linear combination of the ω_i.

However, the resulting point depends on the particular path from p to q. If γ_2 is a second path, we can form a closed path based at p by following γ to q and then following γ_2^{-1} back to p. This loop defines an element of the fundamental group with base point p, but for purposes of integration, only the induced element in H_1 matters. This element has the form $\sum (m_i a_i + n_i b_i)$ where the m_i, n_i are integers. So the difference between the elements in C^g given by the two paths is a sum of $m_i P_i + n_i Q_i$ where P_i and Q_i are the lattice generators. Thus the difference is a lattice point.

It follows that the following element in C^g/L is independent of the path from p to q:

$$\left(\int_p^q \omega_1, \int_p^q \omega_2, \dots, \int_p^q \omega_g\right) \in C^g/L$$

This result is the final outcome of our attempt to understand the effect of topology on integrating holomorphic differentials.

29.2 Other Results

We will not have time to prove more about the Jacobian, but we list here three crucial results.

Theorem 229. *Fix p in the map just discussed. Then $q \to C^g/L$ defines a map from the Riemann surface to its Jacobian. This map is one-to-one, and embeds the surface as an algebraic curve in its Jacobian.*

Theorem 230 (Torelli)**.** *Two compact Riemann surfaces of genus g are conformally equivalent if and only if their Jacobians are equivalent as holomorphic manifolds.*

Remark: Long ago we proved Weierstrass' theorem asserting that we can find a function on a domain $\mathcal{U}$ with predetermined zeros and poles of predetermined orders, provided these points have no limit point in $\mathcal{U}$. Further conditions must be met on a compact Riemann surface. It is useful to invent a terminology for this study as follows:

Definition 68. *A* divisor *on a Riemann surface is a finite formal sum $D = \sum d_p p$ where the d_p are integers and the p are points. Thus a divisor is just an assignment of positive and negative integers to finitely many points on the surface. The* degree *of a divisor is the integer $deg(D) =$*

$\sum d_p$. *If f is a meromorphic function on the surface, the divisor of f is defined to be the assignment of the order of the pole to each pole p, and the order of the zero to each zero p. Poles have negative order and zeros have positive order.*

Remark: Since a meromorphic function takes each value the same number of times, the degree of the divisor of a meromorphic function must be zero.

Remark: On the Riemann sphere, the converse is easily proved true. If a divisor has degree zero, we can find a meromorphic function with that divisor.

Remark: However, on a torus this condition is not sufficient. For example, we cannot find a doubly periodic function with only one pole of order one, so the divisor $D = p - q$ cannot come from a meromorphic function. We proved that a divisor with degree zero comes from a meromorphic function on the torus if and only if the actual sum $\sum d_p p$ is a lattice point.

Theorem 231 (Abel-Jacobi)**.** *Let D be a divisor on a compact Riemann surface and suppose $deg(D) = 0$. Then a series of paths can be found joining poles with zeros so that $|d_p|$ paths start at each pole p and $|d_p|$ paths end at each zero p. Let γ denote the union of these paths. Each path induces an element of C^g by integration, so summing these results over all paths gives an element of C^g. Then D is the divisor of a meromorphic function if and only if this sum is in the lattice L.*

Chapter 30

The Riemann-Roch Theorem

30.1 Introduction to Riemann-Roch

The central theorem in the theory of compact Riemann surfaces is due to Bernhard Riemann. Riemann stated and proved the result in a paper published in 1857, and Gustav Roch added a crucial detail in 1865. We will not prove this result in general, but we will discuss it in several of the following sections.

Suppose D assigns positive integers to a finite number of points on a Riemann Surface. Let $L(D)$ be the collection of meromorphic functions whose only poles are at these points; in addition, we require that the order of the pole at p is at most d_p. For example, if D assigns 3 to just one point p, then $L(D)$ contains constants and functions with a pole of order $1, 2$, or 3 at the point.

It is convenient to think of D as a finite formal sum $\sum d_p p$ where the d_p are positive integers. We called such sums *divisors* earlier, but then we allowed the d_p to be integers of either sign. Moreover, earlier $d_p > 0$ denoted a zero and $d_p < 0$ denoted a pole. We'll reconcile these notations in a moment. Recall that $\deg(D)$ is the sum of the d_p.

Theorem 232. *The set $L(D)$ is a finite dimensional complex vector space and*

$$\dim L(D) \leq \deg(D) + 1$$

Proof: It is obvious that $L(D)$ is closed under addition and scalar multiplication of functions.

Choose local coordinates near each point of the divisor, and expand elements of $L(D)$ in Laurent series about p. This series has d_p negative coefficients. Map $f \in L(D)$ to these negative terms at p for each p. There are $\deg(D)$ such terms, so we have a sequence

$$0 \to \mathcal{K} \to L(D) \to C^{\deg(D)}$$

where $\mathcal{K}$ is the kernel. Functions in $\mathcal{K}$ have no poles and thus are constant. The final map need not be onto, but the sequence shows that the dimension of $L(D)$ is at most $\dim \mathcal{K} + \dim C^{\deg(D)} = 1 + \deg(D)$.

Theorem 233 (Riemann's Inequality)**.**

$$\dim L(D) \geq \deg(D) + 1 - g$$

Remark: This special case of the Riemann-Roch theorem is highly non-trivial because it is an existence theorem. It asserts that there are many non-trivial meromorphic functions on a surface. It is very difficult to prove that there is even one non-trivial meromorphic function on a compact Riemann surface, and indeed in higher dimensions that are compact holomorphic manifolds with no non-constant meromorphic functions. Yet the inequality guarantees that for any point p, there are non-trivial functions with a pole only at p.

Suppose we fix p and look at the divisor $d_p p$ for $d_p = 0, 1, 2, \ldots$. When $d_p = 0$, we only get constants and $\dim L(D) = 1$. When $d_p = 1$, the dimension of $L(D)$ can only be 1 or 2. It will have dimension 2 only if there is a function with a a single pole of order 1 at p. But such a function would take each value exactly once and thus produce a one-to-one and onto map from the surface to the Riemann sphere, so $g = 0$. Otherwise the dimension of $L(D) = 1$. In that case we say that $d_p = 1$ is a *gap value* because there is no function with a pole of that order. Looking at Riemann's inequality, we notice that the left side did not increase, but $\deg(D)$ increased by 1. In a sense, this gap "used up" one of the values of g.

Proceeding in this way, Riemann's inequality guarantees that there are at most g gap values. We will later discover that at each point there are exactly g gap values. We will also discover that at all but finitely many points, these gap values are $k = 1, 2, \ldots, g$. On the other hand, this is not always true. For example, in the hyperelliptic case we constructed a function with a single pole of order 2 on surfaces of arbitrary genus.

Roch's contribution to the theorem was to discover another vector space $I(D)$ making up the difference. This space also has finite dimension, and the full theorem becomes

Theorem 234 (Riemann-Roch)**.**

$$\dim L(D) - \dim I(D) = \deg(D) + 1 - g$$

Remark: The space $I(D)$ contains holomorphic differentials with the additional property that at each p where $d_p > 0$, the differential has a zero of order at least d_p.

Consider the special case $D = 0$. Then $L(D)$ contains only constants, and $I(D)$ contains all holomorphic differentials. So the Riemann-Roch theorem states that

$$1 - \dim(\text{holomorphic differentials}) = 0 + 1 - g$$

and thus that the dimension of the space of holomorphic differentials is g. We verified this for hyperelliptic surfaces and claimed it true in general in earlier sections.

Example 1: Consider the Riemann sphere. Recall that there are no holomorphic differentials on the sphere. So the Riemann-Roch theorem asserts that

$$\dim L(D) = \deg(D) + 1$$

Looking back at the start of this section, we conclude that if we are given a finite number of points and principle parts of a Laurent expansion at each point, there is a meromorphic function with these principle parts, and it is unique up to a constant. This is exactly the partial fraction decomposition of rational functions. Recall that any rational function can be expanded into partial fractions

$$\sum_{z_i} \left[\sum_k \frac{c_k}{(z - z_i)^k} \right] + P(z)$$

where P is a polynomial. The partial fraction terms are Laurent expansions at finite points, and P is the Laurent expansion at infinity.

Example 2: Consider a torus. There is only one holomorphic differential, dz, and it has no zeros. So as soon as D is nonzero, $I(D) = 0$ and the Riemann-Roch theorem states that

$$\dim L(D) = \deg(D) + 1 - 1 = \deg(D)$$

When $D = d_p p$ and $d_p = 1$, the dimension of $L(D)$ is one, so this space only has constants and there is no doubly-periodic function with a single first order pole. But after that, each time we increase d_p by one, the dimension of $L(D)$ increases, so there are doubly-periodic functions with poles at a single point of order $2, 3, 4, \ldots$.

30.2 The Theorem of Riemann-Roch

From now on, a *divisor* is any finite formal sum $D = \sum d_p p$ where the d_p are integers, positive or negative. If f is a meromorphic function, we let (f) be the divisor of f, that is, $\sum d_p p$ where d_p is the order of a zero, or the negative of the order of a pole. Similarly if ω is a meromorphic differential, we let (ω) be its divisor, the sum $\sum d_p p$ where d_p is the order of a zero or the negative of the order of a pole.

Definition 69. *The space $L(D)$ is the set of all meromorphic functions f on the surface such that $(f) + D \geq 0$.*

Remark: This generalizes our definition in the previous section. If $d_p > 0$, then f can have a pole at p, but the order of the pole must be less than or equal to d_p. If $d_p < 0$, then f must have a zero at p of order at least $|d_p|$. As before, $L(D)$ is a complex vector space and clearly is finite dimensional.

Definition 70. *The space $I(D)$ is the set of all meromorphic differentials ω with $(\omega) - D \geq 0$.*

Remark: Thus if $d_p > 0$, then ω must have a zero at p of degree at least d_p. If $d_p < 0$, then ω can have a pole at p, but the order of this pole must be at most $|d_p|$.

Remark: Below is the complete Riemann-Roch theorem. Note that $\dim I(D) \geq 0$, so this result implies Riemann's inequality.

Theorem 235 (Riemann-Roch)**.**

$$\dim L(D) - \dim I(D) = \deg(D) + 1 - g$$

30.3 Linear Equivalence

Definition 71. *We say two divisors D_1 and D_2 are* linearly equivalent *if there is a meromorphic function f such that $D_1 - D_2 = (f)$. Since $(f_1 f_2) = (f_1)(f_2)$, this is an equivalence relation.*

Theorem 236. *Linear equivalence has the following properties:*

- *$D_1 \sim D_2$ implies $\deg(D_1) = \deg(D_2)$*
- *$D_1 \sim D_2$ implies $\dim L(D_1) = \dim L(D_2)$*
- *$D_1 \sim D_2$ implies $\dim I(D_1) = \dim I(D_2)$*
- *Consequently, the Riemann-Roch theorem is true for D if and only if it is true for all divisors equivalent to D.*

Proof: The first item holds because the degree of (f) is always zero. The equation $(f) + D \geq 0$ is equivalent to the equation $\Big((f) + (g)\Big) + \Big(D - (g)\Big) \geq 0$, or $(fg) + (D - (g)) \geq 0$, so $f \in L(D)$ if and only if $fg \in L(D - (g))$. It immediately follows that these spaces are isomorphic, and thus have the same dimension. The remaining results are proved in the same way.

Remark: This result is useful when we want to verify the Riemann-Roch theorem in a case when we already know the meromorphic functions.

Remark: All meromorphic differentials are linearly equivalent, since if $f(z)dz$ and $g(z)dz$ are differentials, then $\frac{g(z)}{f(z)}$ is a well-defined meromorphic function invariant under coordinate changes, and

$$\frac{g(z)}{f(z)} f(z)dz = g(z)dz$$

We could then arbitrarily pick a meromorphic divisor τ and its divisor $K = (\tau)$. This is often called the *canonical divisor*; it is only defined up to linear equivalence. Any other divisor has the form $\omega = f\tau$, and so $I(D)$ is the set of all $f\tau$ such that $(f\tau) - D \geq 0$, that is, $(f) + (\tau) - D \geq 0$, or $(f) + K - D \geq 0$. But this is the definition of $L(K - D)$.

In summary, once we find a convenient meromorphic differential and compute it's divisor K, the Riemann-Roch theorem takes the form

$$\dim L(D) - \dim L(K-D) = \deg(D) + 1 - g$$

This alternate way of writing the Riemann-Roch theorem is often given as the main theorem.

Example: On a torus, one differential is dz, which has no zeros or poles. Consequently we can take $K = 0$ and rewrite the theorem

$$\dim L(D) - \dim L(-D) = \deg(D)$$

Theorem 237. *The degree of K is $2g-2$*

Proof: Write the Riemann-Roch theorem for D and $K-D$:

$$\dim L(D) - \dim L(K-D) = \deg D + 1 - g$$

$$\dim L(K-D) - \dim(D) = \deg(K-D) + 1 - g$$

Add to get

$$0 = \deg(D) + \deg(K-D) + (2-2g)$$

Set $D = 0$ to obtain $0 = \deg(K) + (2-2g)$.

Example 1: On the Riemann sphere, the differential dz has no finite poles or zetos, but has a pole of degree 2 at infinity. Hence its divisor has degree -2.

Example 2: On a torus, the differential dz has no zeros or poles, and $2g-2=0$.

Example 3: On the hypergeometric surface with $2g$ branch points and thus g branch cuts, one of the differentials is

$$\frac{1}{\sqrt{(z-e_1)(z-e_2)\ldots(z-e_{2g})}}\,dz$$

This differential has $2g$ zeros, each of order 1, at the e_i, and one pole of order 2 at infinity. Hence the degree of K is $2g-2$.

30.4 Riemann-Roch and the Riemann Sphere

Theorem 238. *The Riemann-Roch theorem is true on the Riemann sphere.*

Proof: In this case we know the meromorphic functions and differentials precisely, so we can give a direct proof.

Let us begin by noting that any two divisors with the same degree are linearly equivalent. To prove this, we need only show that a divisor with degree zero is the divisor of a meromorphic function. Let D be such a divisor. It suffices to find a meromorphic function with the prescribed zeros and poles, with order, ignoring infinity, for both the given divisor and the divisor of our function will have degree zero and thus our function will also match D at infinity.

So suppose the prescribed zeros in the plane are at $z_1, \ldots, z_m$ with orders $d_1, \ldots, d_m$ and the prescribed poles in the plane are at $w_1, \ldots, w_n$ with orders $e_1, \ldots, e_n$, and define

$$f(z) = \frac{\prod (z - z_i)^{d_i}}{\prod (z - w_j)^{e_j}}$$

Having proved this result, it suffices to prove the Riemann-Roch theorem for one divisor in each linear equivalence class, and thus for one divisor of each possible degree. Let D_n be the divisor which is n at infinity and zero everywhere else. The degree of this divisor is n, and we must prove that

$$\dim L(D_n) - \dim L(K - D_n) = n + 1$$

Note that $L(D_n)$ contains meromorphic functions with no poles in the plane, and so polynomials. For $n \geq 0$, the point at infinity must be a pole of order at most n, so the degree of the polynomials in $L(D_n)$ must be at most n, and thus $\dim L(D_n) = n + 1$. On the other hand, if $n < 0$, then we require that our polynomial vanish at infinity, so $L(D)$ only contains the 0 element and $\dim L(D_n) = 0$.

To complete the argument, we compute $L(K - D_n)$. Note that $-D_n = D_{-n}$. Using dz as a meromorphic differential, there are no zeros or poles in the plane, but a pole of order 2 at infinity. So $K = D_{-2}$ and $K - D_n = D_{-2-n}$. The elements of this space are meromorphic functions on the Riemann sphere with no finite poles, so they are polynomials. When $-2 - n$ is negative, these polynomials vanish at infinity, so $L(K - D)$ is the zero space. That happens when $-2 - n < 0$ or $-2 < n$ or $n = -1, 0, 1, 2, \ldots$. But for $n = 0, 1, 2, \ldots$, $L(D)$ already has dimension $n+1$ and the Riemann-Rock theorem is true. For $n = -1$, both of $L(D)$ and $L(K-D)$ are zero, but then $n + 1 = 0$.

Finally, for $-2 - n \geq 0$, $L(K - D)$ contains polynomials of degree at most $-2 - n$, so this space has dimension $(-2 - n) + 1 = -1 - n$. In this case we already know that $L(D_n)$ contains only the zero function, so $\dim L(D_n) - \dim L(K - D_n) = -(-1 - n) = n + 1$, as desired. QED.

Remark: This proves the Riemann-Roch theorem for the sphere, assuming that we give a sphere the standard conformal structure. But we can also work backwards. Suppose the Riemann-Roch theorem holds on all compact Riemann surfaces, and in particular all surfaces with $g = 0$. Then

Theorem 239. *The Riemann-Roch theorem for $g = 0$ implies that all Riemann surface structures on a sphere are conformally equivalent to the Riemann Sphere.*

Proof: Select a point p in the surface and let $D = p$. Then

$$\dim L(D) \geq \dim L(D) - \dim L(K - D) = \deg D + 1 - g = 2$$

Therefore $L(D)$ contains a non-constant meromorphic function. This function has at most one pole, at p, and this pole has order at most one. If there is no pole, then the function is constant, so it has one pole of order one and thus takes each complex value exactly once. It follows that f maps our surface to the Riemann sphere in a one-to-one, onto, holomorphic manner, proving that the two objects are conformally equivalent. QED.

30.5 Riemann-Roch and Doubly-Periodic Functions

We can also carry out the program of the previous section on a torus. In this case there is a holomorphic differential dz with no zeros and no poles, so $K = 0$. Thus the Riemann-Roch theorem says

$$\dim L(D) - \dim L(-D) = \deg D$$

Theorem 240. *The Riemann-Roch theorem is true on a torus given by a lattice in C.*

Proof: We prove this using a series of lemmas.

Lemma 32. *If $L(D) \neq 0$, then* $\deg(D) \geq 0$.

Proof: Suppose $f \in L(D)$. Then $(f) + D \geq 0$, so $\deg(f) + \deg(D) \geq 0$. But the degree of (f) is zero because f takes each value the same number of times, so $\deg(D) \geq 0$.

Lemma 33. *If* $\deg(D) = 0$*, then either D is the divisor of a function and* $\dim L(D) = 1$*, or else D is not the divisor of a function and* $\dim L(D) = 0$.

Proof: Assume that $L(D)$ is not zero. Then there is a non-zero meromorphic function f with $(f) + D \geq 0$. Either $(f) + D$ is identically zero or else $\deg((f) + D) > 0$. But $\deg((f) + D) = \deg(f) + \deg(D) = \deg(D) = 0$. So $(f) + D = 0$ and $D = -(f) = (1/f)$.

Remark: The Riemann Roch theorem on a torus then follows for $\deg(D) = 0$. Namely, either D is the divisor of a function f, and $-D$ is the divisor of $\frac{1}{f}$ and $\dim L(D) - \dim L(-D) = 1 - 1 = 0 = \deg(D)$, or else D is not the divisor of a function and $\dim L(D) - \dim L(-D) = 0 - 0 = 0 = \deg(D)$.

Remark: In all other cases, $\deg(D) > 0$ or $\deg(D) < 0$. In these cases on a torus, only one of $L(D)$ and $L(-D)$ is non-zero. Clearly it suffices to prove that when $\deg(D) > 0$, we have $\dim L(D) = \deg(D)$.

Lemma 34. *If* $\deg(D) > 0$*, then*

$$\dim L(D) \leq \deg(D)$$

Proof: It suffices to prove this when $\deg(D) = 1$, because any divisor of higher degree can be obtained from such a D by adding one point at a time, and $\dim L(D+P) \leq \dim L(D)+1$.

If $\deg(D) = 1$, then $D = E+P$ where $\deg(E) = 0$. Then $\dim L(D) = \dim L(E+P) \leq \dim L(E)+1$. But earlier we showed that either $\dim L(E) = 0$ or else $\dim L(E) = 1$ and E is the divisor of a function f. If $\dim L(E) = 0$, then $\dim L(E+P) \leq 0+1 = 1$, as desired. If E is the divisor of f then $D = (f)+P$, so D and P are linearly equivalent and $\dim L(D) = \dim L(P)$. But an element of $L(P)$ would have at most one pole of order one, which is impossible on a torus, and so $L(P)$ only contains constants and $\dim L(P) = 1 = \deg(D)$.

Remark: To finish the proof of Riemann-Roch on a torus, we must prove that if $\deg(D) > 0$ then $\dim L(D) \geq \deg(D)$.

We first show this when $\deg(D) = 1$. The divisor D is a formal sum $\sum d_i P_i$. But we can interpret this as an actual sum of points in the torus, and let $P = \sum d_i P_i$ be the resulting point. Form the new divisor $P - D$, which has degree zero. Moreover, as an actual sum of points in the torus, this new sum gives zero. So it satisfies the Abel-condition we proved in the chapter on elliptic functions, and there is a meromorphic function f so $(f) = P - D$ and so $(f) + (D - P) = 0$. It follows that $f \in L(D-P)$ and so $\dim L(D-P) \geq 1$. Therefore $\dim L(D) \geq \dim L(D-P) \geq 1$, as desired.

We prove the inequality in remaining cases by induction on the degree. To prove the induction step, suppose we have a divisor of the form $D+P$ where the result is true for D. It is enough to prove that $\dim L(D+P) > \dim L(D)$, and thus enough to find a meromorphic function that is in $L(D+P)$ but not in $L(D)$. Write $D = \sum n_i P_i$, a formal sum. Select $\deg(D+P)$ points Q_i in the torus such that none are P and the actual sum

$$\sum n_i P_i + P - \sum_i Q_i$$

is zero modulo the lattice. It is easy to find such points, of course. Notice that if we interpret this as a formal sum, its degree is zero. As an actual sum in the torus, it is also zero modulo the lattice. Hence there is a doubly-periodic function f with the negative of this expression as divisor. So

$$0 = (f) + \sum n_i P_i + P - \sum Q_i \leq (f) + \sum n_i P_i + P = (f) + D + P$$

and $f \in L(D+P)$. However, none of the Q_i is P, the order of f at P is one smaller than the coefficient of P in D, and so $f \notin L(D)$. QED.

Remark: Thus the Riemann-Roch theorem is true on tori whose complex structure is given by a lattice in C. But conversely, when theorem is proved abstractly for all compact Riemann surfaces, it implies that every complex structure on a torus is conformally equivalent to a structure given by a lattice. Indeed

Theorem 241. *The Riemann-Roch theorem for surfaces with $g = 1$ implies that every such surface is conformally equivalent to a surface associated to a lattice and period parallelogram.*

Proof: We are assuming that $\dim L(D) - \dim L(K - D) = \deg D$ and therefore that

$$\dim L(D) \geq \deg D$$

Fix a point P in the surface and apply this inequality to $D = nP$ for $n = 1, 2, 3, ...$ The space $L(P)$ contains meromorphic functions with at most one pole at P and no other poles. Constants satisfy this property. If f is not constant, then it has a first order pole at P and thus takes every complex value exactly once. So f is a one-to-one and onto continuous map from the surface to the Riemann sphere, and the genus of the surface is zero rather than one. Consequently $L(P)$ contains only constants and has dimension one.

By the Riemann-Roch theorem, $L(2P)$ must have dimension at least $\deg(2P) = 2$, and yet $\dim L(2P) \leq \dim L(P) + 1$. So $L(2P)$ has dimension two, and by the same reasoning $L(nP)$ has dimension n whenever $n \geq 1$. It follows that there is a function on the surface with only one pole, at P and of order n, for each $n \geq 2$.

Pick f such a function with a pole of order 2 at P. Pick g a function with a pole of order 3 at P. Then f^2 has a pole of order 4 at P and fg has a pole of order 5 at P. Finally, both f^3 and g^2 have poles of order 6 at P. So $L(6P)$ has dimension 6 and contains $1, f, g, f^2, fg, f^3$ and g^2. These seven elements must be linearly dependent, so there are constants such that

$$g^2 = a_1 f^3 + a_2 fg + a_3 f^2 + a_4 g + a_5 f + a_6$$

Our next goal is to revise our choices of f and g to make this equation take the form

$$g^2 = 4f^3 - g_2 f - g_3$$

for new constants g_2 and g_3. When we picked g, we could have picked $\lambda_1 g + \lambda_2 f + \lambda_3$ if $\lambda_1 \neq 0$. If we replace g by $g + \lambda_2 f$, we can choose λ_2 to cancel the fg term. After that, we can pick λ_3 to cancel the g term. Finally we can pick λ_1 to change the coefficient of the f^3 term. So

$$g^2 = 4f^3 + a_1 f^2 + a_2 f + a_3$$

When we picked f, we could have picked $f + \lambda$. By choosing λ correctly, we can cancel the f^2 term. So now

$$g^2 = 4f^3 + a_1 f + a_2$$

Rename a_1 and a_2.

Next form the Riemann surface for $\sqrt{4z^3 - g_2 z - g_3}$ exactly as we did when we constructed hyperelliptic Riemann surfaces. When we factor $4z^3 - g_2 z - g_3$, we obtain three roots, but two might agree, or sll three might be equal. We can still construct a Riemann surface before. If all three roots are distinct, this surface is a torus. If there are only two roots, e_1 and e_2, and e_1 is a double root, then the square root does not change sign as we circle e_1, so it need not be cut out. But we must still make one branch cut from e_2 to ∞. If all three roots are equal, the square root changes sign as we circle it, and we must make one branch cut from e_1 to ∞. So in all cases we obtain a Riemann surface. The surface is a torus if the three roots are distinct, and a sphere otherwise.

The expression $\psi : p \to (f(p), g(p))$ is a map from our unknown surface $\mathcal{S}$ to the Riemann surface we have just constructed. The usual arguments show that it is a well-defined holomorphic map.

We now apply the following result, whose proof can safely be left to the reader:

Theorem 242. *If $\psi : \mathcal{S}_1 \to \mathcal{S}_2$ is a holomorphic map from one compact Riemann surface to another, then either ψ is constant or else it is onto.*

Therefore our map from the unknown surface to the hyperelliptic surface is onto.

Since f has exactly one pole, of order two, f takes each complex value exactly twice, counting multiplicities. So most of the time there will be two points p_1, p_2 in our unknown surface so $f(p_1) = f(p_2)$. These values are either on the same sheet or on separate sheets of the hyperellipic surface. If $g(p_1) = g(p_2)$, then one of the two points in the hyperelliptic surface corresponding to $f(p_1)$ is not covered at all. Since our map is onto, $g(p_1) \neq g(p_2)$. It follows that the map ψ is one-to-one. So it is a conformal equivalence and our unknown surface is the hyperelliptic surface we just constructed.

Either $g(p_1) = g(p_2)$ or else these two points map to different points on our hyperelliptic surface. If $g(p_1) = g(p_2)$, then

We now apply the following result, whose proof can safely be left to the reader:

Theorem 243. *If $f : \mathcal{S}_1 \to \mathcal{S}_2$ is a holomorphic map from one compact Riemann surface to another, then either f is constant or else it is onto.*

Therefore out map from the unknown surface to the hyperelliptic surface is onto.

If the roots of the cubic are different, the new surface is holomorphic to a torus given by a lattice using the inverse of the map $z \to (\wp(z), \wp'(z))$. Composing ψ with this map, we have a conformal equivalence between our surface and a torus associated with a lattice.

If some roots of the polynomial $4z^3 - g_2 z - g_3$ are equal, then the hyperelliptic surface we constructed is a sphere, so our unknown surface is conformally a sphere. But that is impossible

since we are assuming $g = 1$. QED.

30.6 Riemann-Roch on Surfaces with $g = 2$

Theorem 244. *The Riemann-Roch theorem for surfaces with $g = 2$ implies that every such surface is hyperelliptic.*

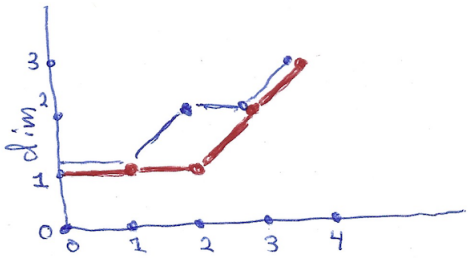
Warning: This result is false for every $g \geq 3$.

Proof: The argument is similar to the argument given in the previous section. According to the Riemann-Roch theorem,

$$\dim L(D) - \dim L(K - D) = \deg D + 1 - 2 = \deg(D) - 1$$

We apply this to $D = nP$ for $n \geq 0$. Since $L(0)$ contains only functions with no poles, it contains constants only and has dimension one. Then we gradually add more copies of P to D. When we add one more P, $\dim L(D)$ can either stay the same or increase by one, and $\dim L(K-D)$ can either stay the same or decrease by one. Since the right side increases by one, either $\dim L(D)$ increases by one and $\dim L(K - D)$ stays the same, or the reverse.

However, there is another constraint. When we add an additional P, the degree of $K - D$ decreases by one. As soon as this degree is negative, $L(K - D)$ is just the zero function. We proved that the degree of $K - D$ is $2g - 2$, so when $g = 2$ it is 2. Starting with $D = 0$, when we add one P, this degree is one, and when we add two P's, this degree is zero, and when we add three P's, the space $L(K - P)$ vanishes, that is, has dimension zero. This dramatically reduces the possibilities for $\dim L(nP)$ as we increase n, and we can display these possibilities in a little chart:



We explain. Two routes as possible for the sequences of $\dim L(nP)$, shown in red and black. A flat spot in a route occurs when $\dim L(D)$ does not change because $\dim L(K - D)$ changes, and we know that $\dim L(K - D)$ changes twice. So each route shown has two flat sections. Moreover, a flat section can only start when $\deg(K - D) \geq 0$, so in our case when $n \leq 2$.

Notice that one other route satisfies these conditions; that route increases between 0 and 1 and then is flat until $n = 3$. But that route would give a meromorphic function with a pole at P of order one and no other poles. Such a function would take each complex value exactly once, and thus make our surface conformally equivalent to the Riemann sphere, contradicting the assumption that $g = 2$.

Similar diagrams can be drawn for all values of g, but the initial complications before the diagram stabilizes increase. The red path shows Riemann's original inequality. When $g = 0$, the red path is the only possibility by Riemann's inequality. When $g = 1$, one other path is possible, but it would give a meromorphic function with only one pole of order one, which is impossible. So only the red path is realized when $g = 0$ or $g = 1$. Alternate paths appear for $g \geq 2$.

As we will see in the next section, the red path is taken for all but finitely many points P on the surface. But points do exist when the black path is taken. Let us leave that assertion for the next section and choose such a P now. Then the dimensions of $L(nP)$ for $n = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$ are $1, 1, 2, 2, 3, 4, 5, 6, 7, 8, 9$. Each time the dimension increases, there is a new meromorphic function with a pole only at P of prescribed order. We can list these functions as follows:

nP	dimension	function
0	1	1
1	1	
2	2	f
3	2	
4	3	f^2
5	4	g
6	5	f^3
7	6	fg
8	7	f^4
9	8	f^2g
10	9	g^2, f^5

We stop here because we have ten functions in a nine dimensional space. So these functions are linearly dependent, and we get an equation of the form

$$g^2 + a_1 f^2 g + a_2 fg + a_3 g = a_4 + a_5 f + a_6 f^2 + a_7 f^3 + a_8 f^4 + a_9 f^5$$

If we replace g by $g - \frac{a_1}{2} f^2$, we get rid of the $f^2 g$ term. If we then replace g by $g - \frac{a_2}{2} f$, we get rid of the fg term. If we replace this g with $g - \frac{a_3}{2}$, we get rid of the g term and have a new equation; renaming the remaining constants we have

$$g^2 = a_5 f^5 + a_4 f^4 + a_3 f^3 + a_2 f^2 + a_1 f + a_0$$

Replacing f appropriately, we can make the coefficient of f^5 equal one, and then factor to obtain

$$g^2 = (f - e_1)(f - e_2)(f - e_3)(f - e_4)(f - e_5)$$

Incidentally, we have ignored a small number of special cases. The original coefficient of g^2 might be zero, or the coefficient of f^5 might be zero. Usually these special cases lead to surfaces with $g < 2$ and can be ignored. Details are left to the reader.

We recognize the basic equation of a hyperelliptic Riemann surface with $g = 2$ holes. If the e_i are not all distinct, then some of the branch points are not needed and we get a surface of smaller genus. The arguments given in the previous section apply without change, proving that our surface is conformally equivalent to the resulting hyperelliptic surface. It follows that if e_i are not distinct, then our surface is conformally equivalent to a surface with smaller g. Since we assume $g = 2$, the e_i must be distinct. QED.

30.7 Riemann Surfaces and Algebraic Curves

In the previous two sections, we proved that our surface is conformally equivalent to the Riemann surface of a function $g(z)$ satisfying $g^2 = P(z)$. We used branch points and branch cuts to create this surface as a domain for $g(z) = \sqrt{P(z)}$. But some readers may have suspected that we could bypass that surface entirely and map directly to the curve $Y^2 - P(X) = 0$. Since we allow X and Y to be complex, this curve is actually a two dimensional surface in C^2, a four dimensional space.

We can draw pictures of the solutions of $Y^2 = P(X)$ over the reals as in high school analytic geometry, so we call it a curve. These pictures turn out to be more helpful than one might expect. Nevertheless, the equation is giving us a surface, so why bother constructing an entirely different surface using branch cuts and all that?

There is a reason, but we'll hold off explaining for a moment.

One problem is that we allow X and or Y to be infinite. However, the geometers have already provided a beautiful solution for this problem. They think of our curves, which are really surfaces, as living in projective space. This projective space consists of equivalence classes of complex triples (X, Y, Z), not all zero, where (X, Y, Z) is equivalent to $\lambda(X, Y, Z)$ when $\lambda \neq 0$. If $Z \neq 0$, then the equivalence class has a unique point $(X, Y, 1)$ and such points form the ordinary plane. Points with $Z = 0$ correspond to points at infinity. If $Y \neq 0$, such a point is equivalent to $(X, 1, 0)$, giving infinitely many points at infinity. If $Y = 0$, the point is equivalent to $(1, 0, 0)$, giving one final point at infinity.

If we are given a polynomial $P(X, Y)$, we add an appropriate power of Z to each monomial to obtain a *homogeneous* polynomial $P(X, Y, Z)$. If (X, Y, Z) solves this polynomial, so does $\lambda(X, Y, Z)$, so we can think of the solutions as certain points in projective space. When we set

$Z = 1$, we obtain the original $P(X, Y)$, so this curve has the original points in the plane. But it now has additional points at infinity.

We did exactly this when studying elliptic functions. We discovered that $(\wp')^2 = 4\wp^3 - g_2\wp - g_3$ and then introduced the curve $Y^2 = 4X^3 - g_2X - g_3$, with one additional point at infinity. We obtained a one-to-one and onto map from the period parallelogram to this curve.

So with this slight modification, curves $P(X, Y) = 0$ from high school analytic geometry become surfaces in projective space, and it begins to look like these surfaces are exactly the compact Riemann surfaces.

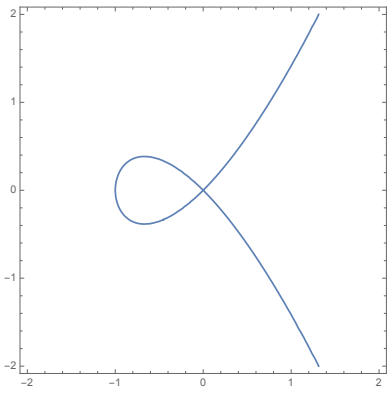
In addition, there is another astonishing reason to think of Riemann surfaces as curves. The meromorphic functions on surfaces, which we carefully constructed using analysis, become rational functions on these curves. That is, they have the form

$$\frac{p(X, Y)}{q(X, Y)}$$

for polynomials p and q. We already caught hints of this fact. On the Riemann sphere, the meromorphic functions are ordinary rational functions of z. On a torus, the meromorphic functions are rational functions of $\wp$ together with other rational functions of $\wp$ multiplied by $\wp'$. So curves would allow us to avoid analysis entirely and just deal with algebra.

Followers of Riemann took exactly this route and developed an elaborate theory of algebraic curves. Lo and behold, the theorem of Riemann-Roch is still the central theorem of this theory, but reinterpreted as an entirely geometric result and given a completely algebraic proof.

However, there is one problem, which perhaps is the penalty assessed by God for avoiding analysis entirely. These curves can have singularities where the surface is not locally homeomorphic to an open set in the plane, so they aren't really full-fledged surfaces in the topological sense. An example is the curve $Y^2 = X^2(X + 1)$. A graph is shown below, and something fishy is going on at the origin.



We can use the implicit function theorem to find these singularities. If (X_0, Y_0) is a zero of $P(X, Y)$ and $\frac{\partial P}{\partial Y} \neq 0$ at (X_0, Y_0), then we can solve for Y in terms of X near this point, obtaining $Y = \varphi(X)$ nearby. So the solutions form a surface near this point. If $\frac{\partial P}{\partial X} \neq 0$, we can solve for X in terms of Y. We only have problems is both of these partials are zero, and those are the singularities of the curve.

In the case $Y^2 = X^2(X+1)$, the two partials are $3X^2 + 2X$ and $2Y$, so the singularities can only occur at $(0,0)$ and $(-2/3, 0)$ and only the first of these is on the curve.

Curves have tangent lines, and that continues to be true when we allow X and Y to be complex, except that now these lines look like two-dimensional planes in four-space. There is an easy way to find the equation of the tangent line at the origin, if the origin is on the curve. Write down the linear terms of the polynomial and set this expression to zero. To find the equation somewhere else, translate the curve so the interesting point is the origin, calculate $P(X, Y)$ for the translation, find the linear equation of the tangent line, and translate back.

This technique still works at singularities, except that now there may be several tangent lines. If the singularity is at the origin, write down the sum of the terms of $P(X, Y)$ of lowest degree, and set this sum to zero. Since the sum of these terms is a homogeneous polynomial, whenever a point p is a solution, so is λp for constant λ. So the solutions are a series of lines through the origin.

Let us apply this to $Y^2 - X^2(X+1) = Y^2 - X^3 - X^2$. The terms of lowest degree are $Y^2 - X^2 = (Y-X)(Y+1)$, so the curve has two tangent lines at the origin, $Y = \pm X$. We can easily visualize these lines in the picture.

Suppose $P(X, Y)$ is an irreducible polynomial and suppose the curve $P(X, Y) = 0$ has no singularities in the finite plane and no singularities at infinity. If the degree of P is d, geometers have proved that the genus of the resulting surface is $g = \frac{d(d+1)}{2}$. These numbers are $1, 3, 6, 10, 15, \ldots$. So our hyperelliptic curve giving $g = 2$ cannot possibly be non-singular everywhere. Let us study its singularities.

The hyperelliptic curve is $Y^2 - (X - e_1)(X - e_2) \ldots (X - e_n) = 0$. The partial with respect to Y is $2Y$ and this is only zero if $Y = 0$ and thus at the points $(e_1, 0), (e_2, 0), \ldots, (e_n, 0)$. The partial with respect to X is $\sum_i (X-e_1)(X-e_2) \ldots \widehat{(X-e_i)} \ldots (X-e_n)$ where the hat means that that particular term should be omitted. If all of the e_i are distinct, then the value of this expression at $(e_i, 0)$ is not zero. Therefore this hyperelliptic curve has no singularities in the finite plane.

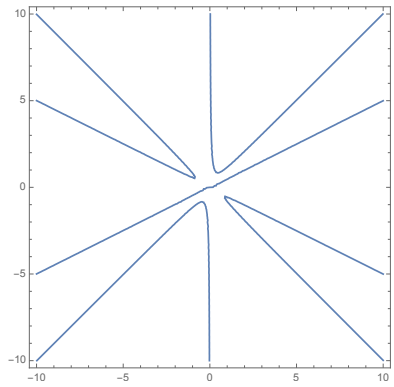
To find points at infinity, we add an appropriate number of Z's to monomials, obtaining

$$Y^2 Z^{n-2} - (X - e_1 Z)(X - e_2 Z) \ldots (X - e_n Z) = 0$$

Setting $Z = 0$ gives $X^n = 0$, so there is one point at infinity at $(0, 1, 0)$. Then the set of all $(X, 1, Z) = (X, Z)$ is another coordinate plane in projective space, and the point at infinity of our

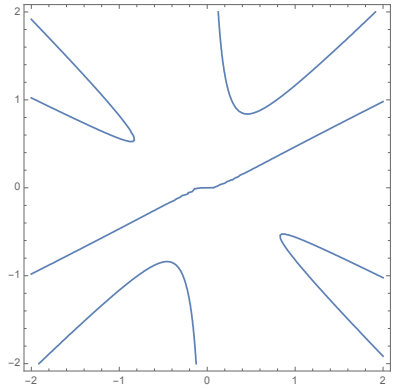
curve is $(0,0)$ and the equation of the curve is $Z^{n-2} - (X - e_1 Z)(X - e_2 Z) \ldots (X - e_n Z) = 0$.

Let us consider the case when there are five e_i with values $-2, -1, 0, 1, 2$. Then we get $Z^3 = X(X^2 - Z^2)(X^2 - 4Z^2)$. The terms of lowest degree are Z^3, so there is only one tangent line at $(X, Z) = (0,0)$ and it is $Z = 0$. On the other hand, both partials vanish at $(0,0)$, so the point at infinity is singular. But how bad can it be? Let's draw some pictures over the reals. The program *Mathematica* gives:



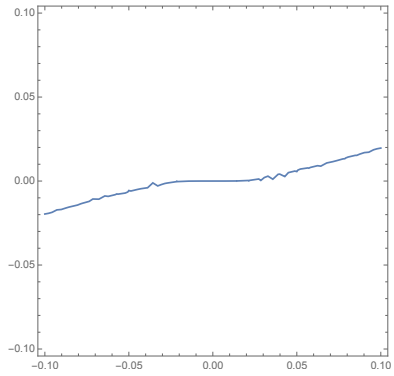
At first we suspect computational error, since the picture seems to show lines $X = 0, X = \pm Z$, and $X = \pm 2Z$. However, substitution back into the full equation shows that these lines are not exact solutions. Moreover, the tangent line at the origin is $Z = 0$, which isn't even visible in the picture.

Let us zoom closer to the origin. We obtain



Recall that the line at infinity is given by $Z = 0$. None of the points of the curve except the origin is on this line. So the crucial piece of the curve is the diagonal line that goes through the

origin. But this line seems to be a little wobbly. Let us magnify once more.



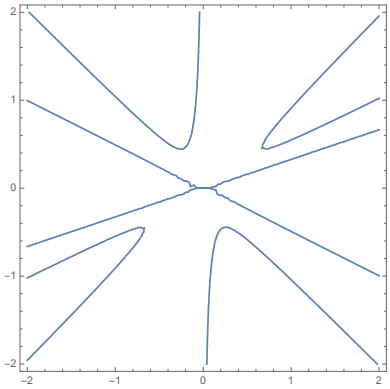
The line is still wobbly, but we finally can see that it is tangent to $Z = 0$ at the origin. But this is bad news because we would hope that our curve crossed the line at infinity decisively at an angle, rather than tangentially.

These pictures illustrate that the structure of curves at singularities can be complicated. These complications must be mastered if Riemann surface theory is translated to a theory of curves, because singularities are inevitable.

A better way to think of our story is that every curve can be *parameterized* by a compact Riemann surface. Given a curve, we can find a Riemann surface and a mapping from the surface to the curve given by holomorphic functions. This map will be onto, and it will be one-to-one to non-singular points. But singular points of the curve can be images of more than one point on the surface, and the map may not have a good inverse near them.

We present one final numerical experiment. If there are six e_i, we still get a surface with genus $g = 2$. This time there are three branch cuts, none extending to infinity. Therefore, the Riemann surface has two points at infinity.

Let us examine pictures for the special case when the e_i are $0, \pm 1, \pm 2, 3$. We have $Y^2 = X(X^2 - 1)(X^2 - 4)(X - 3)$. We can find points at infinity on this curve by writing $Y^2Z^4 = X(X^2 - Z^2)(X^2 - 4Z^2)(X - 3Z)$ and then setting $Z = 0$ to obtain $0 = X^6$. So we only get one point $(0, 1, 0)$ at infinity. Near this point we have coordinates $(X, 1, Z) = (X, Z)$ satisfying $Z^4 = X(X^2 - Z^2)(X^2 - 4Z^2)(X - 3Z)$. The resulting picture is on the next page. This time, two different pieces of the Riemann surface join at the point at infinity on the curve. So both points at infinity on the Riemann surface map to the same point on the curve. Notice that both are tangent at the singularity.



30.8 Gap Sequences

According to the Riemann-Roch theorem,

$$\dim L(nP) - \dim L(K - nP) = n + 1 - g$$

When $n = 0$, this reads $1 - \dim L(K) = 1 - g$, or

$$\dim L(K) = g$$

.

Recall also that $\deg(K) = 2g - 2$.

When we increase n by 1, one of two things happen. Either $\dim L(nP)$ increases by one, so there is a meromorphic function on the surface with a pole of order $n + 1$ and no other poles, and $\dim L(K-nP)$ does not change, or else $\dim L(nP)$ does not change, so there is no meromorphic function on the surface with a pole of order $n + 1$ and no other poles, and $\dim L(K - nP)$ decreases by one. Moreover, in all cases the degree of $K - nP$ goes down by one, and when this degree becomes negative, then $L(K - nP)$ only contains the zero function.

Definition 72. *Fix a point P on a compact Riemann surface S. A positive integer n is called a* gap value *for P if there is no meromorphic function on S with a pole of order exactly n at P and no other poles.*

Theorem 245. *For a fixed point P,*

- *there are precisely g gap values*
- *if* $g \geq 1$*, then* 1 *is a gap value*
- *gap values are at most* $2g - 1$
- *if* $g = 0$*, there are no gap values*
- *if* $g = 1$*, the gap values are 1*
- *if* $g = 2$*, the gap values are 1 and 2, or 1 and 3*

Remark: We are going to show that the gap values are usually $1, 2, \ldots, g$. Suppose the actual gap values, from small to large, are $n_1, n_2, \ldots, n_g$. Define the value of the gap sequence to be $\sum(n_i - i)$. This number will vanish precisely when the gap sequence is $1, 2, \ldots, g$, and otherwise be positive.

Definition 73. *A point P on a Riemann surface is a* Weierstrass point *if its gap sequence is not* $1, 2, \ldots, g$

30.9 The Weierstrass Gap Theorem

Theorem 246 (Weierstrass Gap Theorem)**.** *The sum of the values of gap sequences over all points of a surface is*

$$(g-1)g(g+1)$$

Corollary 36. *There are only finitely many Weierstrass points on any compact Riemann surface.*

Corollary 37. *If* $g \geq 2$*, Weierstrass points definitely exist.*

Corollary 38. *There are exactly* 6 *Weierstrass points on a Riemann surface of genus* $g = 2$*. These points are the six* e_i*, or the five* e_i *and* ∞*.*

Proof: Leaving these corollaries to the reader, we prove the theorem. First a summary. Choose a basis $\omega_1, \ldots, \omega_g$ for the space of holomorphic differentials. In each coordinate system we can compute the Wronskian of these functions. We will discover that when two coordinate systems z and w are related by $w(z)$, the Wronskian transforms by $\left(\frac{dw}{dz}\right)^g$. Consequently, the Wronskians are sections of a line bundle given by the canonical bundle to the power g. This will allow us to determine the degree of their associated divisor.

We will then compute this divisor directly. It turns out to depend on the location and orders of zeros of the holomorphic differentials, and this in turn is related to the gap sequences at these zeros.

Now the details. In local coordinates, $\omega_i(z) = f_i(z)dz$ for holomorphic functions f_i. The Wronskian is the determinant of

$$\begin{pmatrix} f_1^{(g-1)} & f_2^{(g-1)} & \dots & f_g^{(g-1)} \\ f_1^{(g-2)} & f_2^{(g-2)} & \dots & f_g^{(g-2)} \\ \dots & \dots & \ddots & \dots \\ f_1 & f_2 & \dots & f_g \end{pmatrix}$$

If this is the expression in w coordinates, we obtain the corresponding expression in z coordinates by replacing each $f_i(w)$ by $f_i(w(z))\frac{dw}{dz}$. Then $\frac{df_i}{dw}$ is replaced by $\frac{df_i}{dz}\left(\frac{dw}{dz}\right)^2 + f_i\frac{d^2w}{dz^2}$. And so on for higher derivatives.

If $g = 2$, each term on the top line would be a sum of two expressions, so the entire matrix could be written as a sum of two matrices, and the determinant would be the sum of two determinants. But after factoring out $\frac{dw}{dz}$ from the bottom line, and factoring out $\left(\frac{dw}{dz}\right)^2$ from the top line of one matrix, and $\frac{dw}{dz}$ from the top line of the other matrix, we would find that the second matrix has equal rows and thus its determinant is zero. The determinant of the first matrix would be the determinant in w coordinates times $\frac{dw}{dz}$ to the power $1 + 2 = 3$.

This process happens in general. Each term on a line will become the sum of several terms, but most of these terms will lead to matrices with repeated lines. So the only determinant which survives will contain the original determinant times $\frac{dw}{dz}$ to the power

$$1 + 2 + \dots + g = \frac{g(g+1)}{2}$$

We defined a differential to be an expression of the form $\lambda(z)\ dz$ in each coordinate system, such that when z and w are two coordinates related by $w(z)$, then $\lambda(w)dw$ in the w coordinates equals $\lambda(w(z))\ \frac{dw}{dz}\ dz$ in the z coordinates. Such a differential has a divisor D, given by assigning to each point p the degree of $\lambda(z)$. The transformation rule shows that this divisor is well-defined independent of coordinates.

We could have also defined an m-differential to be an expression of the form $\lambda(z)\,(dz)^m$ in each coordinate system, such that the transformation rule replaces $\lambda(w)\,(dw)^m$ by $\lambda(w(z))\left(\frac{dw}{dz}\right)^m (dz)^m$. We assign no meaning to $(dz)^m$. Think of it as a crutch invented by Leibniz to help us remember the transformation rule. So in reality an m-differential assigns a meromorphic function λ to each coordinate system, such that these functions transform according to the stated rule.

Just as before, we can define the divisor D of an m-differential, and this expression is well-defined independent of coordinate system.

Suppose we have an m-differential λ and we also have a global meromorphic function f. We can then form a new m-differential $f\lambda$ by multiplying each coordinate expression for λ by f. This new object is also an m-differential because it obeys the same transformation rules as λ. The divisor associated to the new m-differential has changed, and indeed $(f\omega) = (f) + (\omega)$. So the new object has divisor $(f) + D$.

Once we have one m-differential, all others can be obtained in this way. Indeed if one differential equals $\lambda_1 (dz)^m$ and a second equals $\lambda_2 (dz)^m$, then dividing gives a meromorphic function $\frac{\lambda_2}{\lambda_1}$ and the transformation rule says that this function is invariant under coordinate changes and thus is a global meromorphic f. Clearly $f\lambda_1 = \lambda_2$.

In particular, if $(f) + (\lambda_1) = (\lambda_2)$. Since the divisors of meromorphic functions have degree zero, we discover that

Theorem 247. *The divisors of all m-differentials have the same degree.*

Recall that K denotes the canonical divisor, i.e., the equivalence class of the divisors of differentials. We proved that the degree of K is $2g - 2$. If ω is any differential, then ω^m is clearly an m-differential; if the divisor of ω is D, then the divisor of ω^m is mD. We conclude that the degree of mD is $m(2g - 2)$ and so =the degree of the divisor of an m-differential is $m(2g - 2)$. In particular, the degree of the Wronskian is

$$\frac{g(g+1)}{2}(2g-2) = (g-1)g(g+1)$$

This is the half way point in the proof.

We now want to find the degree of the Wronskian at a point p. This degree is independent of the basis chosen for the space of holomorphic differentials. Indeed if $g_1, \ldots, g_g$ is a second basis, then $g_i = \sum a_{ij} f_j$ where A is a matrix of constant complex numbers with non-zero determinant. Then $W(g)^T = AW(f)^T$ where W denotes the matrix giving the Wronskian, so the determinants of both sides agree up to a non-zero constant.

Find a coordinate neighborhood of p in which p is the origin. Let V be the space of holomorphic differentials. Find a non-zero holomorphic differential $f_1 dz$ with smallest possible degree d_1 at p. If f_2 is a differential with the same degree, then $f_1 = c_1 z^k + \ldots$ and $f_2 = c_2 z^k + \ldots$, then $f_2 - \frac{c_2}{c_1} f_1$ has higher degree. Therefore we can find a subspace of differentials V_2 such that $V = C f_1 \oplus V_2$ and every differential in V_2 has higher degree than d_1 at p. Continue. Eventually we find a basis $f_1, f_2, \ldots, f_g$ with degrees $d_1 < d_2 < \ldots < d_g$ at p.

Theorem 248. *The degree of W at p is*

$$\sum_{i=1}^{g}(d_i - i + 1)$$

Proof: This does not depend on functions coming from holomorphic differentials, so we prove it for any g holomorphic functions near the origin with degrees $d_1 < d_2 < ... < d_g$ at the origin. We prove this by induction. It is trivial if $g = 1$. Suppose the result is true for $g - 1$. To prove it for g, we expand by minors along the bottom row and find that up to sign,

$$W = \sum_i (-1)^i f_i W(f_1', f_2', ..., \widehat{f_i'}, ..., f_g')$$

By induction, the degree of the first summand is

$$d_1 + \sum_{i=2}^{g}((d_i - 1) - (i - 1) + 1) = d_1 + \sum_{i=2}^{g}(d_i - i + 1) = \sum_{i=1}^{g}(d_i - i + 1)$$

This calculation assumed that if f_i has degree d_i, then $\frac{df_i}{dz}$ has degree $d_i - 1$. This is true unless $d_i = 0$. If $f(z) = c_0 + c_1 z + c_2 z^2 + ...$, then $f' = c_1 + 2c_2 z + ...$ and the degree of the derivative is not -1 but instead 0 or higher. In the previous calculation, all f_i for $i \geq 2$ have degrees at least one and the problem does not arise.

The previous calculation works exactly the same for any of the expansion terms for the Wronskian as long as $0 < d_1$. But if $d_1 = 0$, the expansion term we calculated gives the result we desire, and every other expansion term has higher order. That is what we want. So to finish the argument, it suffices to prove that $d_1 = 0$.

If $d_1 > 0$, then all f_i have a zero at p, and therefore $I(0)$ and $I(p)$ have the same dimension. But in that case $L(0)$ and $L(p)$ would have different dimensions, and so there would be a meromorphic function with a pole of order one at p and no other poles. This cannot happen if $g > 0$. We can assume $g > 0$ since the gap theorem is trivially true for $g = 0$.

Remark: Putting these results together, we find that

$$\sum_p \sum_i \Big(d_i - (i - 1)\Big) = (g - 1)g(g + 1)$$

To finish the proof, we must show that the d_i are precisely the gap values at p. By definition, d is a gap value if no meromorphic function with poles only at p has a pole of order d. Note that $L(dP)$ contains functions with poles at p of order at most d, and $L((d-1)p)$ contains functions with poles of order at most $d-1$. So d is a gap value precisely when $L((d-1)p)$ and $L(dp)$ have

the same dimension. By Riemann-Roch, this happens exactly when $I((d-1)p)$ and $I(dp)$ have different dimension. The first contains holomorphic differentials with zeros of order $(d-1)$ or higher at p, and the second contains holomorphic differentials with zeros of order d or higher at p. These spaces have different dimensions exactly when there is a holomorphic differential with order exactly d at p. QED.

30.10 Generalizing the Riemann-Roch Theorem

After the Riemann-Roch theorem was decisively proved, attempts were made to generalize it to higher dimensions, either for compact complex manifolds or for algebraic varieties. Generalizations were discovered, but none seemed decisive.

Then, around 1955, the dam broke and it became clear how to proceed. Progress depended on new ideas from the twentieth century, both in topology and in several complex variable theory.

One of these new ideas was the notion of a vector bundle on a space X. If M is a surface embedded in R^3, we can assign a tangent plane to each point of M. As the point varies, this plane also varies. The resulting space of tangent vectors is called $T(M)$ and each element has the form (p, v) where $p \in M$ and v is a tangent vector starting at p.

Since the plane is 2-dimensional, it has a basis e_1, e_2. These basis vectors must move as p changes, and a key discovery was that it is often impossible to define $e_1(p), e_2(p)$ continuous tangents over the entire M so these vectors are linearly independent at each point of M. If $M = S^2$ is a sphere, it is even impossible to choose a single continuous vector $e_1(p)$ which is non-zero everywhere. So we cannot just say "pick a basis and write tangent vectors as $(r_1(p), r_2(p))$.

Situations soon arose when a different kind of vector space was attached to each p. So instead of just the tangent space, mathematicians discussed *vector bundles*. Here is the definition:

Definition 74. *Let X be a topological space. A k-dimensional* vector bundle *on X is an assignment to each $p \in X$ of a k-dimensional vector space E_p, and an assignment of a topology to $E = \cup E_p$, such that the natural map $\pi : E \to X$ is continuous, and such that every point $p \in X$ has an open neighborhood $\mathcal{U}$ such that $\pi^{-1}\mathcal{U}$ is homeomorphic to $\mathcal{U} \times C^k$ by a homeomorphism which maps each E_q to $q \times C^k$ by a linear isomorphism.*

Remark: This defines a complex topological vector bundle. It is obvious how to modify the definition if real vector bundles are desired, or if we want C^∞ vector bundles over a C^∞ manifold, or holomorphic vector bundles over a Riemann surface or complex holomorphic manifold of higher dimension.

Remark: Vector bundles are often obtained by glueing pieces together in a more down-to-earth, coordinate dependent manner. In this approach we choose a coordinate cover $\mathcal{U}_\alpha$ and say that

elements of E look like $\mathcal{U}_\alpha \times C^k$ over $\mathcal{U}_\alpha$, but for each α and β we have a transition matrix $\rho_{\alpha\beta,ij}$ where ρ is a $k \times k$ matrix whose entries are continuous functions on $\mathcal{U}_\alpha \cap \mathcal{U}_\beta$ and $(p, c_1, \ldots, c_k)$ over $\mathcal{U}_\alpha$ equals $(p, \sum \rho_{1j} c_j, \ldots \sum \rho_{kj} c_j)$ over $\mathcal{U}_\beta$.

Thus we can create a vector bundle by choosing a cover $\mathcal{U}_\alpha$ and then providing transition matrices $g_{\alpha\beta}$ defined on $\mathcal{U}_\alpha \cap \mathcal{U}_\beta$ for all pairs α, β. These transition matrices must satisfy certain compatibility relations to get a vector bundle. $\rho_{\alpha\alpha}$ must be the identity, and $\rho_{\alpha\beta}\rho_{\beta\gamma} = \rho_{\alpha\gamma}$.

Definition 75. *Suppose E is a vector bundle over X. A* section *of E is a continuous map $s : X \to E$ such that $\pi \circ s$ is the identity. In other words, s is a continuous assignment to each $p \in X$ of a vector in E_p.*

Remark: And now the point of all this. Divisors D define complex line bundles over the surface, and $L(D)$ is the space of holomorphic sections of these line bundles. Here's how this works. Suppose D is a divisor. Select a coordinate cover of the surface. In each open $\mathcal{U}$, find a meromorphic function h_α such that $(h_\alpha) = D$. This h_α is only defined locally on $\mathcal{U}$. Define transition functions by $g_{\alpha\beta} = h_\alpha / h_\beta$. Note that the compatibility conditions hold, so these transition functions define a line bundle. A cross section of this bundle will be a holomorphic function s_α on each $\mathcal{U}_\alpha$, such that $s_\beta = g_{\alpha\beta} s_\alpha = h_\alpha / h_\beta \ s_\alpha$, or $s_\beta h_\beta = s_\alpha h_\alpha$. These pieces thus define a global meromorphic function and on each $\mathcal{U}_\alpha$ we have $(s_\alpha h_\alpha) = (s_\alpha) + D \geq D$. Hence the global function belongs to $L(D)$.

Remark: Jean Leray was a French mathematician born in 1906 who became an expert on partial differential equations. During World War II he was interned in a prisoner of war camp in Austria. Leray was afraid that he would be asked to do war work by the Germans if they learned of his expertise in differential equations, so he temporarily switched fields to topology. In the camp, he invented two of the key ideas from topology of the twentieth century: the notion of sheaves, and the notion of spectral sequences. After the war, he returned to partial differential equations.

A sheaf is a family of functions on a space defined by purely local properties. Thus we can consider the sheaf of differentiable functions, or holomorphic sections of a vector bundle because each of these properties can be verified locally. But the sheaf of bounded functions makes no sense because a continuous function is locally bounded, but not necessarily globally bounded.

For any sheaf $\mathcal{F}$, Leray defined a series of cohomology groups $H^k(\mathcal{F})$ for $k \geq 0$. The group $H^0(\mathcal{F})$ is just the set of all globally defined elements of the sheaf, but elements of H^k for $k \geq 1$ are more obscure.

The Riemann Roch theorem describes $\dim L(D) - \dim L(K - D)$. Here D defines a holomorphic line bundle, whose sections form a sheaf $\mathcal{L}(D)$, and $H^0(\mathcal{L}(D)) = L(D)$. It turns out that $H^1(\mathcal{L}(D)) = L(K - D)$. Moreover, the sheaf cohomology groups $H^k(\mathcal{L}) = 0$ for $k \geq 2$. So the

Riemann-Roch theorem is about

$$\dim H^0(\mathcal{L}) - \dim H^1(\mathcal{L})$$

It became clear around 1950 that the proper generalization of Riemann-Roch involves a holomorphic vector bundle E on a compact complex manifold, its sheaf $\mathcal{E}$ of holomorphic cross sections, and the expression

$$\sum_{k=0}^{n} (-1)^k \dim H^k(\mathcal{E})$$

These cohomology groups depend on intricate properties of the vector bundle E, so slight deformations of this bundle can change all of the dimensions. However, the alternating sum turns out to depend only on topological properties of E, and not on precise holomorphic conditions.

These topological properties were analyzed by Hirzebruch around 1955. They involve the characteristic classes of E and of the tangent bundle T of M. Using these ideas, Hirzebruch proved the general Riemann-Roch theorem in all dimensions. At first, Hirzebruch's proof worked only for projective algebraic varieties. Later Atiyah and Singer proved the Atiyah-Singer index theorem about elliptic partial differential operators, and a special case gave Hirzebruch's Riemann-Roch theorem for all compact complex manifolds.

A second ingredient called the Serre Duality Theorem and added in the same time frame. On a manifold of complex dimension n, this theorem shows that H^k of the sheaf of holomorphic sections of E is isomorphic to H^{n-k} of the sheaf of holomorphic sections of a related vector bundle E_1. In the case of Riemann surfaces, $n = 1$ and this theorem makes H^1 of the sheaf for E isomorphic to H^0 of a the sheaf for a related vector bundle. Since H^0 of a sheaf is just the set of global sections of a bundle, this duality theorem replaces an obscure cohomology group with a very concrete group of sections. Indeed when the vector bundle E in question is the line bundle associated with D, the associated bundle is associated with $K - D$.

30.11 Proving the Riemann Roch Theorem

There are many proofs of the Riemann-Roch theorem, some algebraic and some analytical. We'll sketch one step of the modern proof, the argument that involves sheaf theory.

A *sheaf* $\mathcal{F}$ on a topological space X is an assignment to each open set $\mathcal{U} \subset X$ of a group of functions $\mathcal{F}(\mathcal{U})$ on $\mathcal{U}$, such that the following properties are true:

- if $f \in \mathcal{F}(\mathcal{U})$ and $\mathcal{V} \subset \mathcal{U}$, then f restricted to $\mathcal{V}$ is in $\mathcal{F}(\mathcal{V})$
- if $\mathcal{U}_\alpha$ is an open cover of $\mathcal{U}$ and $f_\alpha \in \mathcal{F}(\mathcal{U}_\alpha)$ for all α and the f_α agree on intersections in the sense that $f_\alpha = f_\beta$ on $\mathcal{U}_\alpha \cap \mathcal{V}_\beta$, then there is an $f \in \mathcal{F}(\mathcal{U})$ whose restriction to any $\mathcal{U}_\alpha$ is f_α.

These conditions just say that belonging to $\mathcal{F}$ is a *local property*; if a candidate is locally in the sheaf, then it is in the sheaf. Thus we can form the sheaf of continuous functions, or the sheaf of differentiable functions, or the sheaf of C^∞ functions, because all of these are local properties. But we cannot form the sheaf of bounded functions, because being bounded is not local.

A morphism of sheaves $\varphi : \mathcal{F} \to \mathcal{G}$ if by definition a series of maps $\varphi : \mathcal{F}(\mathcal{U}) \to \mathcal{G}(\mathcal{U})$ for each open $\mathcal{U}$, compatible with restrictions from $\mathcal{U}$ to $\mathcal{V}$ when $\mathcal{V} \subset \mathcal{U}$. This is just a fancy way to say that φ only depends on local properties of the functions. Thus differentiation induces a map of appropriate sheaves.

Suppose

$$\mathcal{F} \xrightarrow{\varphi} \mathcal{G} \xrightarrow{\psi} \mathcal{H}$$

is a sequence of sheaf morphisms. We say the sequence is *exact at* $\mathcal{G}$ if the composition of the two maps is zero, and whenever $g \in \mathcal{G}(\mathcal{U})$ and $\psi(g) = 0$ in $\mathcal{H}(\mathcal{U})$ and $p \in \mathcal{U}$, there is an open $\mathcal{V}$ such that $p \in \mathcal{V} \subset \mathcal{U}$ and an $f \in \mathcal{F}(\mathcal{V})$ such that $\varphi(f) = g|_{\mathcal{V}}$.

Thus if g goes to zero, it does not necessarily come from $\mathcal{F}$ globally, but in small pieces it comes from $\mathcal{F}$. This agrees with the central idea that sheaf theory is about *local* behavior.

As an example, suppose M is a manifold; we define the sheaves Λ^k of k-forms in an obvious way, and obtain maps of sheaves

$$\Lambda^1 \xrightarrow{d} \Lambda^2 \xrightarrow{d} \Lambda^3 \xrightarrow{d} \dots$$

This is an exact sequence of sheaves because if $d\omega = 0$, then $\omega = d\lambda$ locally, although not necessarily globally.

Notice that we can form the sequence

$$\Lambda^1(M) \xrightarrow{d} \Lambda^2(M) \xrightarrow{d} \Lambda^3(M) \xrightarrow{d} \dots$$

It is still true that $d^2 = 0$, but this sequence is no longer necessarily exact. Indeed the deRham cohomology groups exactly measure the deviation from global exactness, since

$$H^k(M) = \frac{Ker(\Lambda^k(M) \xrightarrow{d} \Lambda^{k+1}(M))}{Im(\Lambda^{k-1}(M) \xrightarrow{d} \Lambda^k(M))}$$

Up to this point, sheaf theory is just a language to speak about the difference between local and global properties. Now we come to the profound part of the theory. If $\mathcal{F}$ is a sheaf on a space X, it is possible to define cohomology groups $H^(\mathcal{F})$ for $k \geq 0$. The starting cohomology group

is easy: $H^0(\mathcal{F}) = \mathcal{F}(X)$. There are many ways to define the higher groups, but all are rather abstract.

Here are the two crucial theorems about this cohomology theory:

Theorem 249. *Suppose*

$$0 \to \mathcal{F} \xrightarrow{\varphi} \mathcal{G} \xrightarrow{\psi} \mathcal{H} \to 0$$

is an exact sequence of sheaves. Then there are maps $H^k(\mathcal{G}) \to H^{k+1}(\mathcal{F})$ making the following sequence exact:

$$0 \to H^0(\mathcal{F}) \to H^0(\mathcal{G}) \to H^0(\mathcal{H}) \to H^1(\mathcal{F}) \to H^1(\mathcal{G}) \to H^1(\mathcal{H}) \to H^2(\mathcal{F}) \to H^2(\mathcal{G}) \to ...$$

Theorem 250. *Suppose we have an exact sequence of sheaves*

$$0 \to \mathcal{F} \xrightarrow{d} \mathcal{L}_0 \xrightarrow{d} \mathcal{L}_1 \xrightarrow{d} \mathcal{L}_2 \xrightarrow{d} ...$$

where the d maps are sheaf morphisms. Suppose also that $H^k(\mathcal{L}_i) = 0$ for $k \geq 1$. Then

$$H^k(\mathcal{F}) = \frac{Ker(\mathcal{L}_k(X) \xrightarrow{d} \mathcal{L}_{k+1}(X))}{Im(\mathcal{L}_{k-1}(X) \xrightarrow{d} \mathcal{L}_k(X))}$$

Example: Let R denote the sheaf of locally constant real-valued functions on a manifold M. An element of $R(\mathcal{U})$ is constant on each connected component of $\mathcal{U}$, but possibly different on different components. We have an exact sequence of sheaves

$$0 \to R \to \Lambda^1 \xrightarrow{d} \Lambda^2 \xrightarrow{d} \Lambda^3 \xrightarrow{d} ...$$

It turns out that $H^k(\lambda^i) = 0$ for $k \geq 1$, essentially because partitions of unity can be applied to elements of these sheaves. The previous theorem then states that the cohomology groups $H^k(M, R)$ are given by the deRham cohomology groups.

Remark: We are now ready to explain the connection of this theory with the Riemann-Roch theorem. Suppose D is a divisor. We then easily define a sheaf $\mathcal{L}(D)$; namely $\mathcal{L}(D)(\mathcal{U})$ is the set of meromorphic functions on $\mathcal{U}$ such that $(f) + D \geq 0$ for all points in $\mathcal{U}$. Suppose P is a point on the surface. Then we have an exact sequence of sheaves

$$0 \to \mathcal{L}(D) \to \mathcal{L}(D + P) \to \mathcal{L}(D + P)/\mathcal{L}(D) \to 0$$

Here the last sheaf consists of functions over $\mathcal{U}$ which belong to $L(D + P)$ divided out by functions which belong to $L(D)$. If $P \notin \mathcal{U}$, this is zero because P imposes no extra condition in $\mathcal{U}$. But if $P \in \mathcal{U}$, then there is one extra degree of freedom in the numerator. These functions are

only defined locally, so we don't have to worry about extending to create global functions on the entire surface, and thus the quotient is C. So intuitively this quotient is a *tower sheaf* which is C at P and 0 everywhere else. Let us denote this quotient sheaf $\mathcal{Q}$.

The previous paragraph is somewhat vague becauseour definition of a sheaf was not general enough. In a course on sheaf theory, more care is taken in the definition; as a result, when $\mathcal{F} \subset \mathcal{G}$ are sheaves, it is easy to define the quotient sheaf.

The exact sequence obtained above induces an exact sequence in cohomology. Here it is:

$$0 \to L(D) \to L(D+P) \to \mathcal{Q}(\mathcal{S}) \to H^1(\mathcal{L}(D)) \to H^1(\mathcal{L}(D+P)) \to H^1(\mathcal{Q}(\mathcal{S}) \to$$

At this point, three technical results are required.

Theorem 251. *We have*

- $H^0(\mathcal{Q}) = C$
- $H^1(\mathcal{Q}) = 0$
- $H^1(\mathcal{L}(D))$ *is finite dimensional for any* D

Thus the previous exact sequence becomes

$$0 \to L(D) \to L(D+P) \to C \to H^1(\mathcal{L}(D+P)) \to H^1(\mathcal{L}(D+P)) \to 0$$

Now we apply an easy algebraic result. If we have an exact sequence of finite dimensional vector spaces which starts and ends with zero spaces, then $\sum(-1)^i \dim V_i = 0$. So from the above sequence we conclude that

$$\dim L(D) - \dim L(D+P) + 1 - \dim H^1(\mathcal{L}(D)) + \dim H^1(\mathcal{L}(D)) = 0$$

or

$$\dim L(D) - \dim H^1(\mathcal{L}(D)) - \deg(D) = \dim L(D+P) - \dim H^1(\mathcal{L}(D+P)) - \deg(D+P)$$

It follows from this result that the number attached to a divisor D by the three-term formula on the left is a constant independent of D, and thus equals the number attached to $D = 0$. Indeed we can obtain any D from $D = 0$ by adding or subtracting points one at a time.

When $D = 0$, the number in question is

$$\dim L(0) - \dim H^1(\mathcal{L}(0)) - 0 = 1 - \dim H^1(\mathcal{L}(0))$$

The Serre Duality theorem shows that $\dim H^1(\mathcal{L}(0)) = g$. Thus our result gives

$$\dim L(D) - \dim H^1(\mathcal{L}(D)) - \deg(D) = 1 - g$$

or

$$\dim L(D) - \dim H^1(\mathcal{L}(D)) = \deg(D) + 1 - g$$

Finally, the Serre Duality Theorem gives

$$\dim H^1(\mathcal{L}(D)) = \dim L(K - D)$$

30.12 Field Theory

The set of meromorphic functions $\mathcal{M}$ on a compact Riemann surface is a field. Among these functions are constants, so $C \subset \mathcal{M}$. Since C is algebraically closed, any additional element $f \in \mathcal{M}$ satisfies no algebraic relations, so as a field $C(f)$ is the set of all $\frac{P(f)}{Q(f)}$ where P and Q are polynomials.

It can be shown that whenever f is not constant, the field extension $C(f) \subset \mathcal{M}$ is finite algebraic. Field theory then tells us that there is an element $g \in \mathcal{M}$ such that

$$\mathcal{M} = C(f) \oplus C(f)g \oplus ... \oplus C(f)g^{n-1}$$

Therefore

$$g^n = a_0(f) + a_1(f)g + ... + a_{n-1}(f)g^{n-1}$$

where each a_i is rational. Clearing denominators gives a polynomial $P(X, Y)$ such that

$$P(f(z), g(z)) = 0$$

We proved several cases of this result. On the Riemann sphere, every meromorphic function is a rational function in z. On a torus, every meromorphic function equals

$$R_1(\wp(z)) + R_2(\wp(z))\wp'(z)$$

and

$$(\wp')^2 = \wp^3 - g_2\wp - g_3$$

We found a similar result on any hyperelliptic surface.

If we follow this idea further, the following theorem can eventually be proved.

Theorem 252. *Compact Riemann surfaces $\mathcal{S}$ satisfy these results:*

- *Two surfaces $\mathcal{S}_1$ and $\mathcal{S}_2$ are conformally equivalent if and only if their fields $\mathcal{M}_1$ and $\mathcal{M}_2$ are field isomorphic by an isomorphism which is the identity on C*
- *The field of meromorphic functions on any $\mathcal{S}$ is a finite algebraic extension of $C(X)$, the field of rational functions*

- *Conversely, every such field occurs as the field of meromorphic functions of some compact Riemann surface*

This theorem reduces compact Riemann surface theory to pure algebra; it becomes a chapter in the theory of fields. The first paper to carry out this program was by Dedekind and Weber, *The Theory of Algebraic Functions of One Variable*, published in 1882. Among other things, this paper contained a purely algebraic proof of the Riemann-Roch theorem, and since Riemann's work depended on results in analysis not yet cleanly proved, their's was the first complete proof of the theorem. However, Dedekind and Weber started with a field of the type described above, so their result only held for Riemann surfaces whose meromorphic function field has that form. Riemann's full theorem is required to show that *all* Riemann surfaces satisfy this hypothesis.

Dedekind was an expert in number theory, having edited Dirichlet's *Vorlesungen uber Zahlentheorie*. But "edited" doesn't truly describe what he did, because he added several supplements to this book in various editions describing his own work. Some of these supplements give the first modern treatment of algebraic number theory, that is, the theory of finite algebraic extensions of the field Q of rational numbers. The techniques Dedekind introduced in those supplements were extensively used in the paper with Weber.

If we start with an abstract field $\mathcal{M}$, how do we obtain the points of the Riemann surface? One possibility is to use the polynomial $P(X, Y)$ which field theory produces, and introduce the corresponding algebraic curve in projective space. The difficulty with that approach is that this curve usually has singularities, so points which ought to be distinct on the Riemann surface become the same point on the curve. This is not what Dedekind and Weber did.

Suppose p is a point on a compact Riemann surface. We then get a map

$$d_p : \mathcal{M}^\star \to Z$$

which assigns the degree of f at p to any non-zero meromorphic function. This map has the following properties.

Theorem 253. *If $p \in \mathcal{S}$,*

- $d_p(c) = 0$ *for any non-zero constant*
- $d_p(fg) = d_p(f) + d_p(g)$
- $d_p(f + g) \geq \min(d_p(f), d_p(g))$
- *the map d_p is onto Z*

The third condition has an inequality because when we add functions with poles, the poles can cancel, and when we add functions with zeros, additional zeros can be created. Note that the third condition does not make sense if $f + g = 0$. Some authors define $d_p(0) = \infty$ so d_p is defined on all of $\mathcal{M}$ and the four conditions always hold.

The final condition holds on a compact Riemann surfaces due to the Riemann-Roch theorem. Suppose we pick a point $q \neq p$ and form the divisor $D = mq$. For small m, $\dim L(D)$ may not increase as $deg(D)$ increases. Choose m large enough that this anomaly no longer occurs. Then consider $D = mq + np$ for $n = 0, 1, 2, 3, ...$ The dimension of $L(D)$ always increases as we increase n, so there are meromorphic functions with poles of order $0, 1, 2, 3, ...$ at p (and other poles at q). The inverses of these functions have zeros of all orders at p.

On a compact Riemann surface, if $p \neq q$ then $d_p \neq d_q$ by essentially the same argument.

A map $d : \mathcal{M} \to Z$ with all of these properties is called a *valuation*. Since the properties depend only on the field, we can talk about valuations when there is no underlying Riemann surface. We define the points of the *associated Riemann surface* to be all valuations of $\mathcal{M}$.

Suppose we have a valuation d. Let $\mathcal{R}$ be the set of all functions $f \in \mathcal{M}$ so $d(f) \geq 0$, together with the zero function. In the Riemann surface case, this would be all meromorphic functions holomorphic at p. Notice that constants belong to $\mathcal{R}$ and it is closed under addition and multiplication due to the axioms of a valuation. So $\mathcal{R} \subset \mathcal{M}$ is a subring. We call this the *valuation ring at* p. Let $\mathcal{R}_0$ be the subset of $\mathcal{R}$ containing 0 and all elements of $\mathcal{M}$ with $d(f) > 0$. In the Riemann surface case, this will be all meromorphic functions which vanish at p. In general it is an ideal in $\mathcal{R}$. So we can form the quotient ring $\mathcal{R}/\mathcal{R}_0$. In the Riemann surface case, this quotient space would C and a holomorphic function's equivalence class would be $f(p)$. It turns out that the quotient is canonically isomorphic to C in general. There is a natural map from C to this space, and since nonzero elements of C have value $d = 0$, this map is one-to-one. It can be proved to be onto as well.

If $f \in \mathcal{M}$, we can make it into a function on the associated Riemann surface as follows. If $d(f) < 0$, we define $f(p) = \infty$. Otherwise f is in the valuation ring and we define $f(p)$ to be the complex number it induces in the quotient ring.

In short, the valuations of our field form a set, and each element of the field induces a function from this set to $C \cup \{\infty\}$. We can give the set of valuations a topology making all of these functions continuous. Then it is not hard to believe that with further work this point set can be made into a Riemann surface, and the field $\mathcal{M}$ can be identified with the field of meromorphic functions on the surface.

It is illuminating to work out one step of this construction.

Theorem 254. *Every valuation of $R(X)$ has the form $order_c(f)$ or $order_\infty(f)$.*

Proof: Our field is the set of all $\frac{P(X)}{Q(X)}$ for polynomials P and Q. We make heavy use of the valuation ring $\mathcal{R}$ discussed earlier. Suppose first that $d(X) \geq 0$. Then the set of all polynomials is contained in $\mathcal{R}$, and the set of all polynomials with $d(P) > 0$ forms an ideal. If this ideal is trivial, then the value of any non-zero polynomial is zero, so the value of any quotient of two non-zero polynomials is zero, so d is trivial. Therefore the ideal is not zero and must

be generated by some polynomial $a(X)$. Factor $a(X) = (X - c_1)(X - c_2)...(X - c_k)$. Then $d(a) = \sum d(X - c_i)$, so one of the terms must have positive value. Say $d(X - c) > 0$. If $d(X - c_1) > 0$ for $c_1 \neq c$, then $d(c - c_1) \geq \min(d(c - X), d(X - c_1)) > 0$, contradicting the assumption that constants have value zero. So $d(X-c) > 0$ and $d(X-c_1) = 0$ for $c_1 \neq c$.

But any $\frac{P(X)}{Q(X)} = \prod(X - c_i)^{m_i}$ for integers m_i which can be positive or negative. It follows that d of this quotient equals $d(X - c)$ times the number of times $X - c$ appears in this product. This can only be onto if $d(X - c) = 1$ and then d of a rational function is the order of that function at c.

The remaining case is when $d(X) < 0$. Then $d\left(\frac{1}{X}\right) > 0$. It follows that $\left(\frac{1}{X} - c\right) \geq 0$. If $d\left(\frac{1}{X} - c\right) > 0$ for $c \neq 0$, then $d(c) \geq \min\left(\frac{1}{X}, c - \frac{1}{X}\right) > 0$, which cannot happen. So

$$0 = d\left(\frac{1}{X} - c\right) = d\left(\frac{1 - cX}{X}\right) = d(1 - cX) - d(X) = 0$$

and $(1-cX) = d(X)$. So $d(X-1/c) = d(X)$ and thus $d(X-\lambda) = d(X)$ for all λ. Therefore

$$d\left(\frac{P(X)}{Q(X)}\right) = d\left(\prod(X - c_i)^{m_i}\right) = d(X)\sum m_i$$

This d can only be onto Z if $d(X) = -1$, and then

$$d\left(\frac{P(X)}{Q(X)}\right) = -(\deg(P) - \deg(Q)) = order_\infty\left(\frac{P(X)}{Q(X)}\right)$$

QED.

Remark: It remains to find all valuations in the general case of a finite algebraic extension of the rational functions $R(X) \subset \mathcal{M}$. If d is a valuation of $\mathcal{M}$, we can restrict d to $R(X)$ and get one of the known valuations of the rational functions. So the problem is reduced to studying extensions of the standard valuations to the larger field. It turns out that there are at most k extensions when k is the dimension of $\mathcal{M}$ over $R(X)$. But sometimes there are fewer extensions, raised to powers. The picture we obtain is the exact analog of a branched covering of the Riemann sphere, where the various extensions of valuations correspond to various sheets of the covering, and powers of extensions corresponding to branch points. So the field theoretic point of view in the end returns to Riemann's original construction of Riemann surfaces.

Remark: This is not the end of the story, however. The definition of a valuation works for any field. The valuations on fields of meromorphic functions come from points on the Riemann surface. But what happens for other fields?

The natural starting point is the field Q of rational numbers. Our first axiom is about the subfield of complex numbers and no longer makes sense, so we ignore it. But all of the remaining axioms make sense.

It is easy to prove, using exactly the arguments of the previous theorem, that Q has one valuation for every prime number. Given p, factor a rational as

$$q = p_1^{m_1} p_2^{m_2} ... p_k^{m_k}$$

where the m_i are positive or negative integers. Define $d_p(q)$ to be m_p in this factorization. This is a valuation and all valuations have this form.

So in a sense the Riemann Surface of the rational numbers is the set of primes!

Remark: We can keep pushing this analogy. If Q corresponds to the meromorphic functions, then the holomorphic functions will be rational numbers whose valuation are non-negative. That describes the integers. So Q corresponds to meromorphic functions and Z corresponds to holomorphic functions.

Compact Riemann surfaces should correspond to finite algebraic extensions $Q \subset K$. These are exactly the algebraic number fields studied intensively in number theory.

For example, $Q \subset Q(i)$ is such an extension. The holomorphic functions on the extension are exactly the Gaussian integers. We obtain the points on the corresponding Riemann surface by extending ordinary valuations on Q. To do that , let $q_1, q_2, ...$ be the Gaussian primes. Recall that they are obtained by factoring ordinary primes. So up to units,

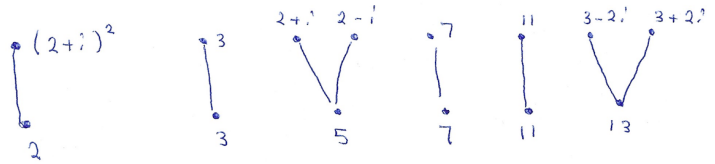
$$2 = (1+i)^2, 3, 5 = (2-i)(2+i), 7, 11, 13 = (3-2i)(3+2i), ...$$

Then any Gaussian number can be written as a unit times

$$q = (2-i)^{m_1} 3^{m_2} (2-i)^{m_3} (2+i)^{m_4} ...$$

Each of these terms gives a valuation, of the form $d_i(q) = m_i$. All valuations have this form.

Notice that these valuations arise from ordinary primes by factoring in the Gaussian primes. This picture corresponds to creating a Riemann surface using branch points and branch cuts. Think of the corresponding Riemann surface in number theory as given by the picture below:



This picture shows one branch point over 2, because 2 factors as a square of a prime. The remaining points are ordinary points, which corresponds to the sheets of the Riemann surface.

In this case, there are two sheets. But some primes factor and some don't, so the analogy seems to be breaking down. Actually it still holds; there is a new phenomenon in number theory that we don't see in Riemann surface theory. Recall that elements of the field become functions on the Riemann surface which take values in $\mathcal{R}/\mathcal{R}_0$, which is canonically C. In the rational number field, this quotient ring is Z_p and elements of the field take values in Z_p at the point p. When we extend to a branched cover, two things can happen. A point can split into two points, or else the value field Z_p can extend to become a field with p^2 elements. In the general case over a non-branch point, a prime p will split into primes $p_1, \ldots, p_k$ and the value quotient Z_p will become finite fields of order $f_1, \ldots, f_k$ and $\sum_{i=1}^{k} f_i = \dim_Q K$. So the Gaussian picture shows two sheets, but sometimes the value field increased size instead of splitting the prime.

When we generalize still more, we come to rings of algebraic integers which do not have unique factorization. Our valuations seem to depend on unique factorization. However, we can recover unique factorization in number fields by working with ideals rather than integers: every ideal can be uniquely written as a product of prime ideals. It turns out that in the general case, the valuations are associated with prime ideals rather than prime numbers. So the sheets in the branched covering do not correspond to factoring p into prime numbers, but rather correspond to factoring the ideal (p) generated by p into prime ideals. The invention of ideals was one of the great moments in algebraic number theory. It is astonishing that later on when valuations were invented by analogy with taking the order of a zero or a pole, the valuations already knew that we should be factoring ideals rather than primes!

This analogy continues to hold as the theories get deeper, but we are not in complex analysis land any more, so it is time to put down the pen.