The Pentagonal Number Theorem and Modular Forms

Dick Koch

March 28, 2010

1 Introduction

At first, the Pentagonal Number Theorem seems to be an isolated result on the fringes of mathematics, amusing but dispensable. But one hundred and fifty years after its discovery, it was connected to the central part of mathematics by an astonishing discovery of Dedekind. I'm going to sketch that development here, without proofs.

2 Math 112

Before studying calculus, students master two families of transcendental functions. The first contains the trigonometric functions: $\sin x, \cos x, \tan x, \arcsin x, \arctan x$. These functions satisfy powerful identities: $\sin^2 x + \cos^2 x = 1$ and $\sin(x+y) = \sin x \cos y + \cos x \sin y$. And they depend on a mysterious constant: π.

The second family contains $\ln x$ and e^x. Again there are powerful identities: $\ln xy = \ln x + \ln y$ and $e^{x+y} = e^x e^y$. And there is a mysterious constant: e.

It is a sort of accident that these families were discovered before calculus. The sine was introduced during the Roman empire by astronomers, although the modern form of the theory dates from textbooks written by Euler around 1750. But logarithms appeared only in 1614, shortly before Newton's work on calculus in 1664. If history had turned out differently, these families would have been discovered as a consequence of calculus, since they are required for integration of the simplest algebraic functions:

$$\int \frac{dx}{x} = \ln x$$

$$\int \frac{dx}{1+x^2} = \arctan x$$

The first of these formulas appears at the very start of Newton's first paper on calculus, *De Analysi per Aequationes Numero Terminorum Infinitas*, written in 1669 but published in 1711. Newton begins by stating that the area under the curve x^n is $\frac{1}{n+1}x^{n+1}$. He gives several examples, involving positive, negative, and fractional exponents. Then he faces the problem that the formula makes no sense when $n = -1$. To deal with this case, he shifts the axis and integrates $\frac{1}{1+x}$. To integrate, he expands in an infinite series, $1 - x + x^2 - x^3 + \ldots$, and integrates term by term to obtain $x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \frac{1}{4}x^4 + \ldots$. In unpublished papers, he computes many values using this series. For instance, he finds $\ln(1.1)$ to 79 decimal places.

In some modern calculus books, you can see how logarithms would have been developed if calculus had been invented first, because the book develops the theory again from scratch. A typical approach starts by defining

$$\ln x = \int_1^x \frac{dt}{t}$$

for $x > 0$. Next the author proves that $\ln xy = \ln x + \ln y$ directly from this definition. The fundamental theorem of calculus gives $\frac{d}{dx}\ln x = \frac{1}{x}$, which is positive, so the logarithm is increasing and thus one-to-one. It is easy to prove that $\ln x : (0, \infty) \to (-\infty, \infty)$ is also onto. In particular, there is a number x such that $\ln x = 1$. Call this number e and more generally call the inverse function e^x. Then the identity $\ln xy = \ln x + \ln y$ implies $e^{x+y} = e^x e^y$. It follows that e^x can be computed by raising the constant e to a power. There. That's the whole theory in a nutshell.

(I've always enjoyed this part of calculus. The students — not so much. They already know logs and want to get on to the new stuff.)

At the University of Oregon we use a book by Michael Spivak in a small honors calculus course. Spivak also develops trigonometry from scratch. He begins by defining

$$\arcsin x = \int_0^x \frac{dt}{\sqrt{1-t^2}} \quad \text{and} \quad \frac{\pi}{2} = \int_0^1 \frac{dt}{\sqrt{1-t^2}}$$

The rest of trigonometry falls out naturally.

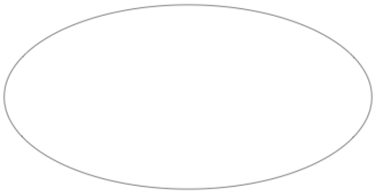
It is reasonable to guess that there are other undiscovered transcendental functions which arise from integrating simple expressions. But these simple expressions won't be rational functions because it is possible to integrate any rational function $f(x) = \frac{P(x)}{Q(x)}$ using only the elementary transcendental functions, as is shown in many calculus books.

These notes are about the simplest case when integration cannot be accomplished with the elementary transcendental functions. Instead, we must introduce new transcendental functions, even more beautiful than the elementary ones.

3 An Unsolved Problem from the Greeks

Shortly after Euclid's geometry was written, the Greeks began studying the conic sections, and in particular the ellipse

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1$$



It is natural to study such curves because circles observed from an angle appear elliptical. Two thousand years later, these conic sections also appeared as solutions of quadratic expressions $Ax^2+Bxy+Cy^2+Dx+Ey+F=0$, as planetary orbits, and elsewhere.

Up to rotation and translation, an ellipse is determined by the two numbers a and b, called the major and minor axes. We always rotate so a is the larger of the two. But it is much better to introduce a new number k defined by $k^2=\frac{a^2-b^2}{a^2}$ and describe the ellipse by giving a and k. The number k is independent of the size of the ellipse, for if we magnify by m, then a becomes ma and b becomes mb and m cancels out in the definition of k. Thus k, which is called the *eccentricity* of the ellipse, describes its shape, while a describes its size. The number k plays a major role in our story.

Perhaps the Greek's most famous formulas, aside from the Pythagorian theorem, are the formulas for the length $2\pi r$ and the area πr^2 of a circle of radius r. These formulas beg to be generalized to the ellipse. It is easy to generalize area. If an circle of radius 1 and area π is magnified in the x direction by a, it's area is also magnified by a. If the resulting figure is then magnified in the y direction by b, its area is magnified by b. The two operations convert the unit circle into an ellipse with major and minor axes a and b and area πab, neatly generalizing πr^2.

Sadly, this argument fails for lengths. Magnification in the x direction stretches more or less horizontal parts of the ellipse, but doesn't stretch more or less vertical parts; the average stretch isn't easy to calculate. So the Greeks never found a formula for the length of an ellipse.

Calculus provides more powerful techniques to compute curve length, as Newton noticed immediately. If a curve moves horizontally by an infinitesimal dx and vertically by an infinitesimal dy, the distance it moves is given by the Pythagorian theorem $\sqrt{(dx)^2+(dy)^2}$, so the total length of a curve is

$$\int \sqrt{(dx)^2+(dy)^2} = \int \sqrt{1+\left(\frac{dy}{dx}\right)^2}\,dx$$

This formula essentially appears in Newton's first calculus paper. So at last the calculation of the length of an ellipse appears imminent.

If we solve the equation of an ellipse for y, we get $y = b\sqrt{1-\frac{x^2}{a^2}}$. Substituting in the length formula, we find after a little algebra that the total length of an ellipse is

$$4a\int_0^1 \sqrt{\frac{1-k^2t^2}{1-t^2}}\;dt$$

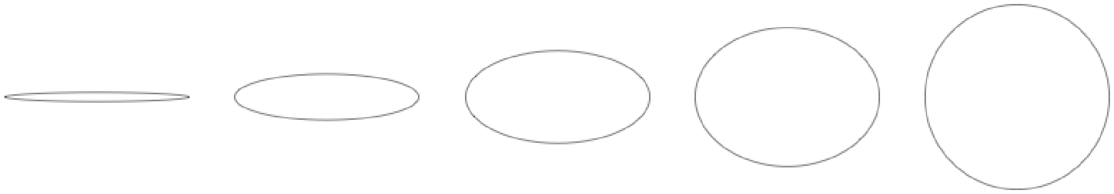
In the circle case when $k=0$, the integral is $\arcsin 1 = \frac{\pi}{2}$ and the above formula is $2\pi a$. In general, let us write

$$\pi(k) = 2\int_0^1 \sqrt{\frac{1-k^2t^2}{1-t^2}}$$

This $\pi(k)$ is a generalization of π, providing an infinite number of similar constants, one for each possible eccentricity. Then the length of an ellipse with major axis a and eccentricity k is

$$2\ \pi(k)\ a$$

generalizing the standard formula $2\pi r$. Below are ellipses of eccentricity $k = 1$, .94, .75, .44, and 0, and $\pi(k)$ values 2, 2.14, 2.42, 2.76, and 3.14.



It remains to actually calculate the above integral, but unhappily Liouville proved that it cannot be computed in terms of the elementary transcendental functions.

4 Elliptic Integrals

Already in the first calculus paper, Newton does not restrict himself to integrating rational functions of x. Since circles are fundamental in mathematics, Newton is forced to integrate expressions containing $\sqrt{1-x^2}$, and he has no trouble doing so.

Some calculus books push this development as far as possible, by considering the square root of an arbitrary quadratic expression $\sqrt{ax^2+bx+c}$, and then considering integrals of rational expressions of the form $R(x, \sqrt{ax^2+bx+c})$; these books show that any such expression can be integrated in terms of elementary transcendental functions. Here $R(x, \sqrt{ax^2+bx+c})$ is a fancy way to describe an arbitrary expression obtained from x and the square root by adding, subtracting, multiplying, and dividing. For instance, if the square root in question is $\sqrt{1-x^2}$, then the following expression is rational in x and the square root:

$$\frac{x^2+(x^5+5)\sqrt{1-x^2}}{5x^9+5\sqrt{1-x^2}}$$

Any such expression can be integrated by using what Spivak calls "the world's trickiest substitution." I'll let you look that up is in Spivak's book *Calculus*. Also, I don't claim that integrating this expression would be fun.

So if we are on the lookout for integrals which will give new transcendental functions, we must consider rational expressions of the above form where the polynomials under the square root have degree at least three. The integral giving the length of an ellipse is an example, because

$$\int \sqrt{\frac{1-k^2t^2}{1-t^2}}\,dt = \int \frac{1-k^2t^2}{\sqrt{(1-t^2)(1-k^2t^2)}}\,dt = \int R(t, \sqrt{(1-t^2)(1-k^2t^2)})\,dt$$

In this case we have the square root of a quartic polynomial.

After the invention of calculus, many different mathematicians ran into such integrals, and they gradually realized that something particularly interesting happens when the degree of the polynomial under the square root is three or four. We'll describe what is special about these cases later on. Such integrals are called *elliptic integrals* because the formula for the length of an ellipse is such an integral. Among the mathematicians involved were Euler, Gauss, Jacobi, and Legendre.

A simple transformation converts an elliptic integral based on a quartic to an elliptic integral based on a cubic, and conversely, so it suffices to restrict attention to one or the other. Between 1825 and 1828, Legendre published two volumes exhaustively studying elliptic integrals based on quartics. His central result was that any such expression can be reduced to a combination of rational functions, elementary transcendental functions, and three additional particular integrals called the elliptic integrals of the first, second, and third kinds. So in a sense he added three new transcendental functions to the mathematical toolkit. If we write the quartic as $(1-t^2)(1-k^2t^2)$, the three integrals are

$$\begin{aligned} \text{First Kind:} \quad & \int \frac{dt}{\sqrt{(1-t^2)(1-k^2t^2)}} \\ \text{Second Kind:} \quad & \int \sqrt{\frac{1-k^2t^2}{1-t^2}}\,dt \\ \text{Third Kind:} \quad & \int \frac{dt}{(1-nt^2)\sqrt{(1-t^2)(1-k^2t^2)}} \end{aligned}$$

You'll notice that I told a slight lie and there are infinitely many integrals of the third kind, depending on the parameter n.

The more important fact is that these integrals depend on the parameter k. So there are infinitely many new functions, determined by the particular quartic chosen or, in the case of the length of an ellipse, the eccentricity of that ellipse. This dependence on k is a major part of the story I'm telling.

In the old days, mathematicians owned a book called *The Handbook of Chemistry and Physics*, with extensive tables of logs, trig functions, etc. This book contained tables with values of elliptic integrals of the first, second, and third kinds. Nowadays, the tables have been replaced by computer programs. For instance, the three types of elliptic integrals are built-in functions provided by *Mathematica.*

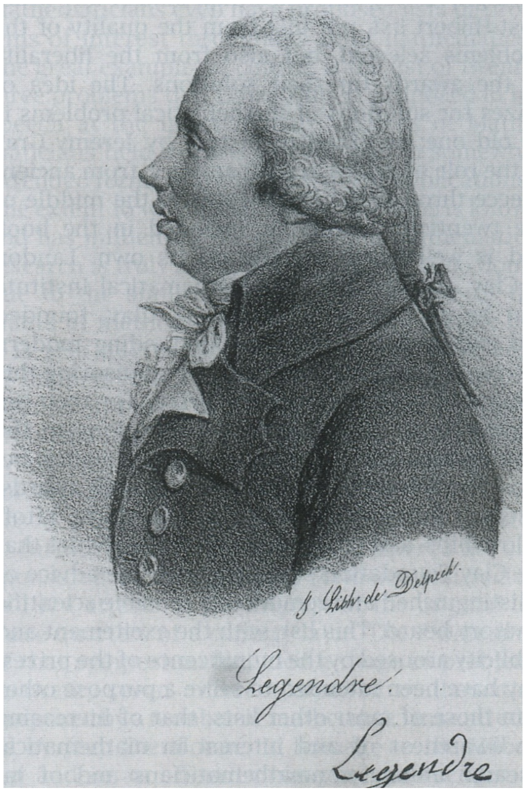
I'm not sure this is progress, because the Handbook used to cost \$3, and *Mathematica* is \$1000.

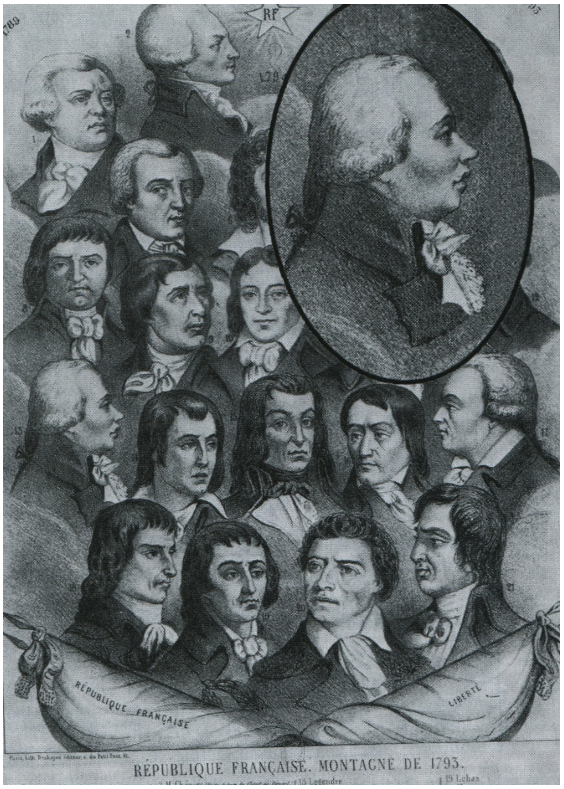
Legendre is one of my favorite mathematicians. He produced an enormous body of significant work. In 1794 he published *Elements de geometrie*, which remained a leading elementary textbook for one hundred years. In 1798 he wrote *Essai sur la theorie des nombres*, a book on number theory with many new theorems, including a proof of the law of quadratic reciprocity. He invented what we now call the Legendre polynomials, used for instance by physicists in the study of the hydrogen atom. And he discovered the method of least squares independently of Gauss.

Legendre also made mistakes, which makes his accomplishments a little more human. His geometry book included a proof of the parallel postulate, just a few years before the discovery of Non-Euclidean geometry showed that such a proof is impossible. His proof of quadratic reciprocity is incomplete, as carefully pointed out by Gauss in *Disquisitiones Arithmeticae* when Gauss gave the first acceptable proof. Moreover, at one point in Legendre's proof he used Dirichlet's theorem on primes in an arithmetic progression, about thirty years before Dirichlet actually provided that theorem. (Dirichlet wrote in the introduction to his paper: "a proof of this theorem is desirable, particularly because Legendre has already used it.")

Similarly, in his work on elliptic integrals, Legendre failed to discover the most important fact about those integrals: namely, that the resulting transcendental functions have very remarkable inverses. That was pointed out by Abel, whose work was generously quoted by Legendre in later editions of his books.

Legendre often said that he wanted to be known by his work, so only one picture survives. Here is that picture on the left, published in his biography and other books on the history of mathematics:

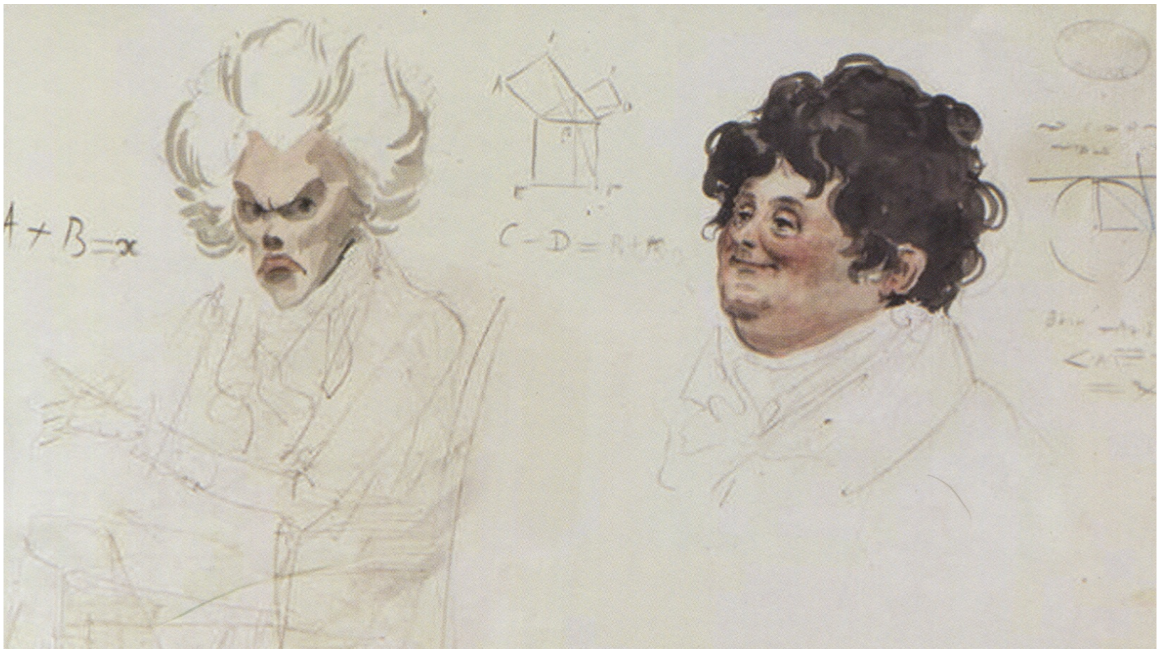




Legendre's first name was Adrien-Marie. There is another Legendre, Louis, who played a minor role in the French revolution; it has been known for some time that these two people were not related. In 2005, two students at the University of Strasbourg were astonished to discover that Google searches provide the same picture for both men. Louis Legendre was a member of the Montagne party, which also included such famous figures as Danton, Marat, and Robespierre. The picture on the right is a group portrait, with Louis Legendre's picture enlarged. You'll notice that it is the mirror reflection of the picture usually believed to be Adrien-Marie Legendre.

Further research confirms that the standard picture isn't Adrien-Marie at all, so no formal picture is known.

In 2007-2008, a caricature of the mathematical Legendre was located in an album in the *Bibliotheque de l'Institut de France*, and is now his only known likeness. Legendre is on the left; Fourier is on the right. The severe image of Legendre causes me to suspect that the artist had to endure a high school geometry course taught from his book.



5 Brief Intermission on Complex Differentiation

So far I've been implying that our functions are integrated over the real line, but to really understand the flavor of the theory we need to integrate in the complex plane. So let me say a few words about calculus using complex numbers. This subject was extensively developed by Euler, and thus a standard part of the mathematical toolkit in the nineteenth century.

Suppose $f(z)$ is defined for complex z, and takes complex values. There are many such functions, since any rational function $\frac{P(z)}{Q(z)}$ is an example. Define $\frac{df}{dz} = \lim \frac{f(z+h)-f(z)}{h}$ as usual. Then all of the standard differention rules hold, so you already know how to compute derivatives. A function which has a complex derivative at every point is called "holomorphic." From now on we assume all functions described are holomorphic.

You can think of $f(z)$ as a map which sends a piece of the plane near z_0 to another piece of the plane near $f(z_0)$, as if the plane were a sheet of rubber to be pulled and stretched as it is mapped. We only need one fact about differentiation in the following pages. If $f'(z_0) \neq 0$, then near z_0 this map is one-to-one and onto, so it does not fold the rubber or crush it.

Often we know a function $f(x)$ on the real numbers, and want to extend it to complex z. There is an amazing theorem about this called the *identity theorem*. It says that if there is any extension at all to a holomorphic function, it is unique. Said another way, if two holomorphic functions agree on the reals, they are the same everywhere. A surprising consequence of this theorem is that when a real function satisfies algebraic identities, the same identities are automatically true for the holomorphic extension. For example, if it is possible to extend e^x to all z, then $e^{z+w} = e^z e^w$ is automatically true for all complex z and w. And if it is possible to extend $\sin x$ and $\cos x$ to complex z, then automatically $\sin^2 z + \cos^2 z = 1$ for complex z.

An easy way to extend is to use power series. Suppose $f(x)$ is defined by a power series which converges for $|x| < R$ or possibly for all x. Then the same power series automatically converges to a holomorphic function of z in the disk of radius R about the origin; if the original series always converges, so does the new series. Since $e^x, \sin x$, and $\cos x$ have power series which always converge, they can automatically be extended to the entire complex plane.

It is useful to combine these theorems to understand extensions. Consider the case of e^z. Any complex z can be written $z = x + iy$ for real x and y, so $e^z = e^{x+iy} = e^x e^{iy}$. So to compute e^z, we don't need to plug z into the power series. Instead it is enough to compute e^{iy} for real y.

Let us compute that using power series. We have

$$e^{iy} = 1+iy+\frac{(iy)^2}{2!}+\frac{(iy)^3}{3!}+\frac{(iy)^4}{4!}+\ldots = \left(1-\frac{y^2}{2!}+\ldots\right)+i\left(y-\frac{y^3}{3!}+\ldots\right) = \cos y+i\sin y$$

and in particular

$$e^{i\pi} = \cos\pi + i\sin\pi = -1$$

So

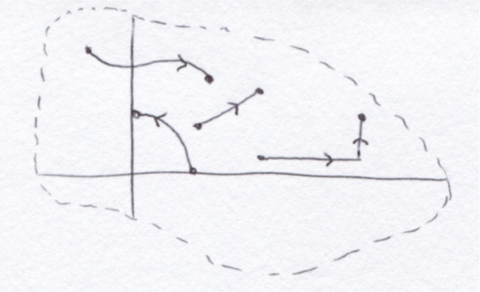
$$e^z = e^{x+iy} = e^x e^{iy} = e^x\cos y + ie^x\sin y$$

6 Brief Intermission on Complex Integration

The real power of complex variable theory comes from the complex integral. If $f(x)$ is a real function, we can compute $\int_a^b f(x)\, dx$ by integrating along the x axis from a to b.

When f is extended to a complex $f(z)$, we can integrate that function along any path γ in the plane from one point to another. We denote that integral $\int_\gamma f(z)\ dz$.

The picture below shows what we have in mind. Imagine a function $f(z)$ defined and holomorphic inside the dotted line. We can integrate the function to produce $\int_\gamma f(z)\ dz$ along any path. The picture shows four typical choices for γ.



Actually, you already know how to integrate because the fundamental theorem of calculus is still true in complex variables. So to integrate, guess a function $F(z)$ with $\frac{dF}{dz} = f(z)$ and then write

$$\int_\gamma f(z)\ dz = F(\text{end of path}) - F(\text{start of path})$$

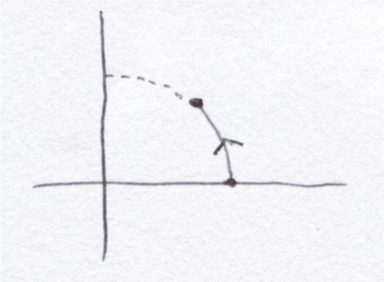
For example, the integral of $f(z) = \frac{1}{z^2}$ along the straight diagonal line from $1+i$ to $2+2i$ is just

$$-\frac{1}{z}\Big|_{1+i}^{2+2i} = -\frac{1}{2+2i} + \frac{1}{1+i} = \frac{1}{4}(1-i)$$

But sometimes we cannot guess F and instead our goal is to define F using the integral. That's what happens in ordinary calculus when we define $\ln x$ to be $\int_1^x \frac{dt}{t}$. Luckily there is a simple formula for complex integrals along a path which reduces integration to ordinary real integrals which can be computed in the usual way, or more often approximated numerically. Here's the rule. Parameterize the path by writing z as a function of a real t, which you can think of as time, so γ is defined by $z(t)$ for $a \le t \le b$. Then write $\int_\gamma f(z)\ dz = \int_a^b f(z(t))\frac{dz}{dt}\ dt$. This last expression gives a complex valued function of a real t. Separate the complex function into its real and imaginary parts and integrate each separately.

I'll give a simple example; the method used here always works. Suppose we want to integrate $f(z) = 2z$ along the piece of the unit circle illustrated on the next page. Then $z(t) = \cos t + i \sin t$ for $0 \le t \le \frac{\pi}{4}$. So the integral is

$$\int 2z\ dz = \int_0^{\frac{\pi}{4}} 2(\cos t + i\sin t)\frac{d}{dt}(\cos t + i \sin t)\ dt = 2\int_0^{\frac{\pi}{4}} (\cos t + i \sin t)(-\sin t + i\cos t)\ dt$$



Expanding, this is

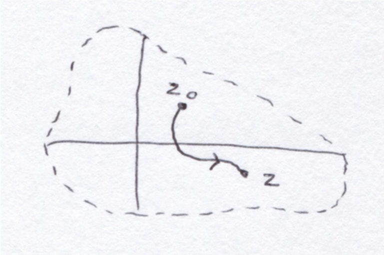
$$2\int_0^{\frac{\pi}{4}} -2\sin t\cos t + i(\cos^2 t - \sin^2 t)\ dt = -2\sin^2 t + i\sin 2t\Big|_0^{\frac{\pi}{4}} = -1 + i$$

We can check this answer because $2z$ is the derivative of z^2, so the answer should equal z^2 at the end of the path minus z^2 at the beginning of the path, and so

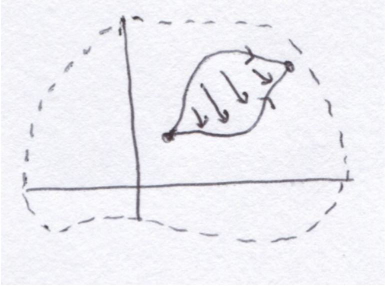
$$\left(\frac{1}{\sqrt{2}} + i\frac{1}{\sqrt{2}}\right)^2 - (1)^2 = i - 1$$

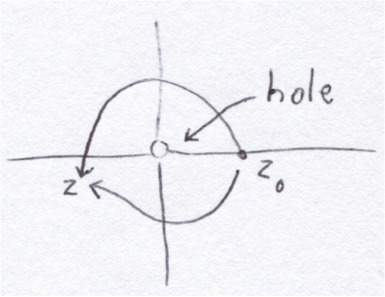
7 Two Important Complications

Summarizing the previous section, suppose we have $f(z)$ defined on a subset of the plane and fix a point z_0 in this subset. Suppose we want to find $F(z)$ on that subset with $\frac{dF}{dz} = f$. To compute $F(z)$, choose any path γ from z_0 to z and compute, probably numerically, $F(z) = \int_\gamma f(z)\ dz$. If we want a lot of values so we can get a picture of F, we have to pick lots of paths and compute lots of integrals.



But there is a complication. It may happen that different paths from z_0 to z yield different integrals; in that case, F doesn't exist at all. Luckily, there is a wonderful theorem due to Cauchy which rules out this complication in many cases. Suppose two paths start at z_0 and end at z, as in the picture on the next page, and suppose one path can be deformed to the other, as illustrated. Then Cauchy's theorem guarantees that they give the same value. If the domain of f has no holes, we can always deform and this complication does not arise. But if there are places where f is not defined, we may run into problems, as illustrated by the picture at the right.





Two examples illustrate this: $\frac{1}{z}$ and $\frac{1}{z^2}$. These functions are not defined at the origin, so the origin is a hole. Therefore these functions may not have antiderivatives. But holes don't always cause a problem, and actually $-\frac{1}{z}$ is the antiderivative of $\frac{1}{z^2}$.

However, there is trouble with $\frac{1}{z}$. The antiderivative wants to be $\ln z$, but there is a problem with complex logarithms. To see this, write points in the plane using polar coordinates: $(x, y) = (r\cos\theta, r\sin\theta)$. In complex analysis, r is called the absolute value of z and written $|z|$. So another way to write this is

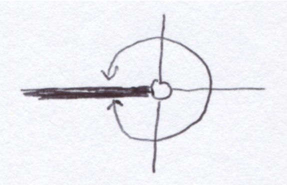
$$z = |z|\cos\theta + i|z|\sin\theta = |z|e^{i\theta} = e^{\ln|z| + i\theta}$$

Consequently, we should define

$$\ln z = \ln|z| + i\theta$$

The problem is that θ is only defined up to a multiple of 2π. If we set $\theta = 0$ on the positive real axis, then it increases to π as we rotate positively to the negative real axis, but decreases to $-\pi$ as we rotate negatively to the same negative real axis.

The standard solution to this problem is to take a pair of scissors and cut out the negative real axis, as shown below. Then the hole at the origin is no longer a hole and the remaining domain of $\frac{1}{z}$ contains no holes. So by Cauchy's result, we can find $F(z)$ with derivative $\frac{1}{z}$. That $F(z)$ is $\ln z = \ln|z| + i\theta$ for the unique θ satisfying $-\pi < \theta < \pi$.



In the remaining sections, we'll often cut domains with scissors to remove holes so we can guarantee the existence of antiderivatives.

The second complication concerns the square root. That's important for us because we are going to integrate the square root of a quartic. The problem is that each number except zero has two square roots. In real variable theory we solve this problem by always choosing the positive square root. But in complex analysis, square roots aren't real and there is no consistent way to choose one over another. For instance, consider $\sqrt{z}$. Let the square root be positive when z is a positive real number. If we rotate counterclockwise to the negative axis, the square root rotates counterclockwise half as fast, so $\sqrt{-1} = i$. But if we rotate clockwise to the negative axis, the square root rotates counterclockwise half as fast, so $\sqrt{-1} = -i$.

Here's the theorem which is used in complex analysis to solve this complication:

Theorem 1 *Suppose $f(z)$ is defined on a domain $\mathcal{U}$ which contains no holes, and suppose f is never zero on this domain. Then it is possible to consistently choose a continuous value for $\sqrt{f(z)}$.*

We'll just accept this result.

8 Gauss

I like to explain Abel's great discovery about elliptic integrals by showing you how Gauss privately discovered the same thing. Gauss didn't publish his results, probably because in the meantime Abel and then Jacobi independently worked out the details.

Recall again that

$$\arcsin x = \int_0^x \frac{dt}{\sqrt{1-t^2}}\, dt \quad \text{and} \quad \frac{\pi}{2} = \int_0^1 \frac{dt}{\sqrt{1-t^2}}\, dt$$

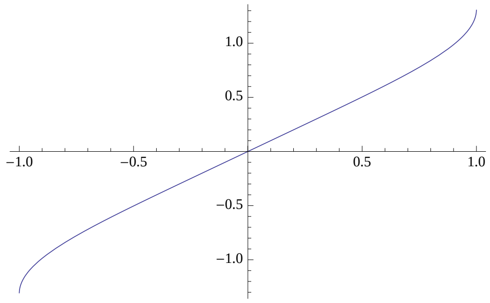
In his notebook, Gauss began experimenting with

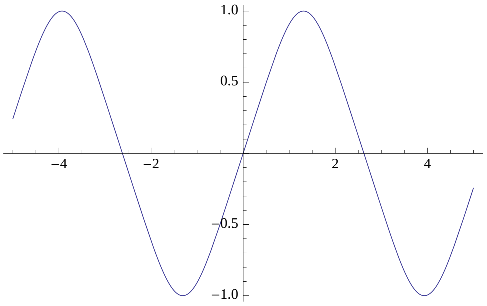
$$f(x) = \int_0^x \frac{dt}{\sqrt{1-t^n}}\, dt \quad \text{and} \quad \frac{K}{2} = \int_0^1 \frac{dt}{\sqrt{1-t^n}}\, dt$$

Here the second integral is the definition of a new constant K. Very rapidly Gauss zeroed in on the case $n = 4$ as especially interesting. Note that this is an elliptic integral:

$$f(x) = \int_0^x \frac{dt}{\sqrt{1-t^4}}\, dt \quad \text{and} \quad \frac{K}{2} = \int_0^1 \frac{dt}{\sqrt{1-t^4}}\, dt$$

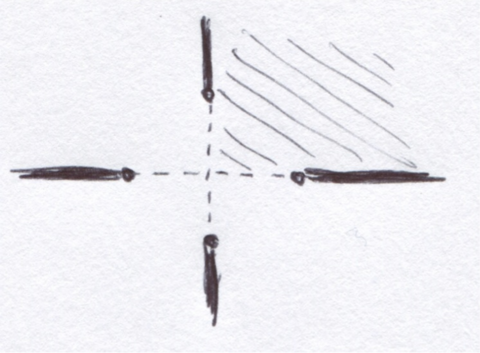
If we graph the function $f(x)$, we obtain the picture on the left. In the easier case of arcsin, the inverse function, $\sin x$, is even nicer, so Gauss inverted $f(x)$ by flipping the x and y axes, to obtain the central section of the picture on the right. This function has an obvious extension as a periodic function with period $2K$. Jacobi called this function $\mathrm{sn}(x)$.





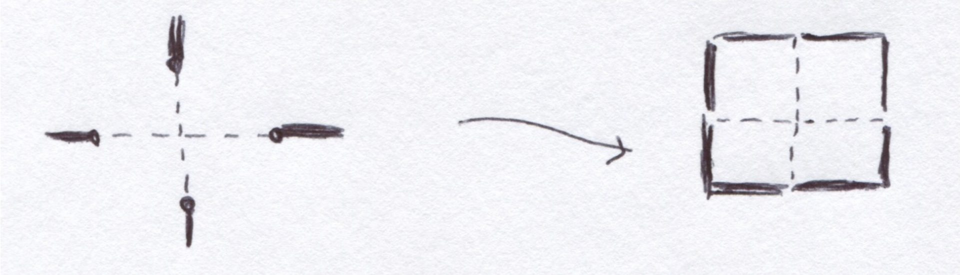
The last chapter of Gauss' famous book on number theory is about trigonometry. It is in that chapter that Gauss inscribes a 17-sided polygon in a circle with straightedge and compass. The introduction to the chapter contains a mysterious paragraph: "The principles of the theory which we are going to 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 $\int 1/\sqrt{1-t^4}\,dt$. Since we are preparing a substantial work on transcendental functions, we have decided to consider only circular functions here."

What concerns us here is the amazing thing that happens when we extend the integral of $\frac{1}{\sqrt{1-z^4}}$ into the complex plane. Note first that the denominator is zero at ± 1 and $\pm i$. So these are the holes of the sort we discussed earlier. We must remove them from the domain by cutting some lines, as indicated by the solid lines in the picture below. The ends of these lines are ± 1 and $\pm i$.

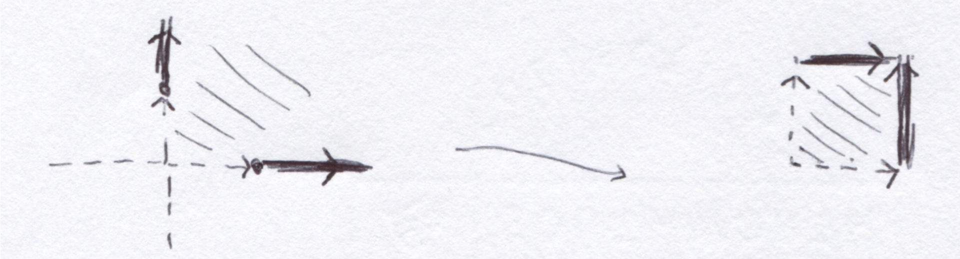


Once we remove these solid lines, the remaining set contains no holes. On this set, the function $1 - z^4$ has no zeros, so we can consistently choose a square root $\sqrt{1-z^4}$. We can choose the square root that is positive on the remaining portion of the x-axis. Since there are no holes, we can find a function $F(z)$ whose derivative is $\frac{1}{\sqrt{1-z^4}}$ by defining $F(z)$ to be the integral of $\frac{1}{\sqrt{1-z^4}}$ from the origin to z.

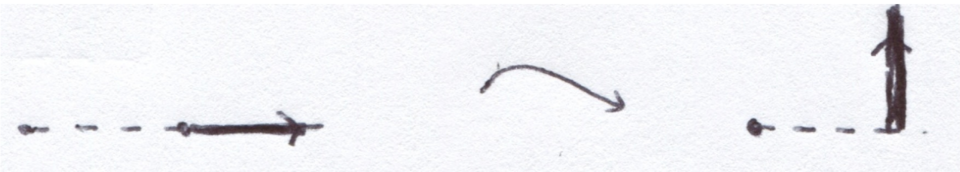
This F maps our set, with its missing lines, to some portion of the plane. We claim that its image is a square, as illustrated below, and the missing lines map to the boundary of this square. Amazing! The portions of the quadrants on the left which are far out near infinity map to the four corners of the square. We'll show that our map is one-to-one and onto, so it has an inverse. All pretty astonishing!



We'll concentrate on showing this for one quarter of the picture, shaded below. Let's pay attention to the behavior of F on the boundary of this shaded region.



First consider F on the x-axis at the bottom. This axis consists of two pieces, the set $0 \leq x \leq 1$ and the set $1 \leq x \leq \infty$. On the first piece, $F(x) = \int_0^x \frac{dx}{\sqrt{1-x^4}}$. When $x = 0$ the integral is zero, and as x increases, $F(x)$ increases through real numbers to the value $F(1) = \frac{K}{2}$. So F maps the dotted line on the left in the picture below to the dotted line on the right.



Next consider $1 \leq x \leq \infty$. Then

$$F(x) = \int_0^1 \frac{dt}{\sqrt{1-t^4}} + \int_1^x \frac{dt}{\sqrt{1-t^4}} = \int_0^1 \frac{dt}{\sqrt{1-t^4}} + i \int_1^x \frac{dt}{\sqrt{r^4-1}}$$

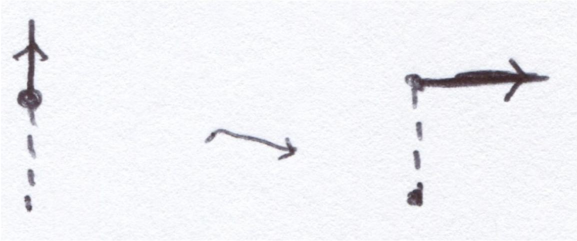
The first integral takes us to the end of the dotted line in the right in the previous picture, and the second integral takes us upward on the solid line. How far up do we go? Well, the final amount is $\int_1^\infty \frac{dt}{\sqrt{t^4-1}}$. Substitute $u = \frac{1}{t}$. A little algebra shows that

$$\int_1^\infty \frac{dt}{\sqrt{t^4-1}} = \int_1^0 \frac{u^2}{\sqrt{1-u^4}} \left(-\frac{1}{u^2}\right) du = \int_0^1 \frac{du}{\sqrt{1-u^4}} = \frac{K}{2}$$

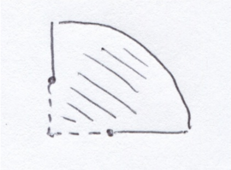
So we get a perfect square.

Next consider F on the left side of the shaded region, which also contains two pieces, a dotted portion and a solid portion. Our function $\sqrt{1-z^4}$ is the same on the x and y axes since $\sqrt{1-(iy)^4} = \sqrt{1-y^4}$ and the i cancels out. But when we integrate in the y direction, we write $z = it$ and the term $dz = idy$ has an extra i. So we get the same integrals as before, but with an extra i.

There is, however, another complication. It isn't clear which sign of the square root of $1-z^4$ is taken on the vertical axis. On the dotted portion, $1-z^4$ is real and close to the value 1 at the origin, where we picked the positive sign, so the square root must be positive. But on the solid portion we cut out, we might have one sign or the other, and the map might carry that portion left or right. We'll show it goes right by an indirect argument.



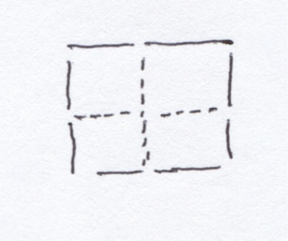
It is useful to connect the horizontal and vertical boundaries by a large circle as indicated below.



What happens to the circular piece? Well, integrating along it will be the integral of $\frac{1}{\sqrt{1-z^4}}$ over a portion of a circle of radius $|z|$. The size of the integrand is roughly $\frac{1}{|z|^2}$ and the length of the circle is at most $\frac{\pi}{2}|z|$ and thus the integral is at most $\frac{\pi}{2|z|}$. For large circles, this is almost zero. So in the limit, integrating the circular portion changes almost nothing and all the stuff on the large circle is scrunched into the top right corner of the square.

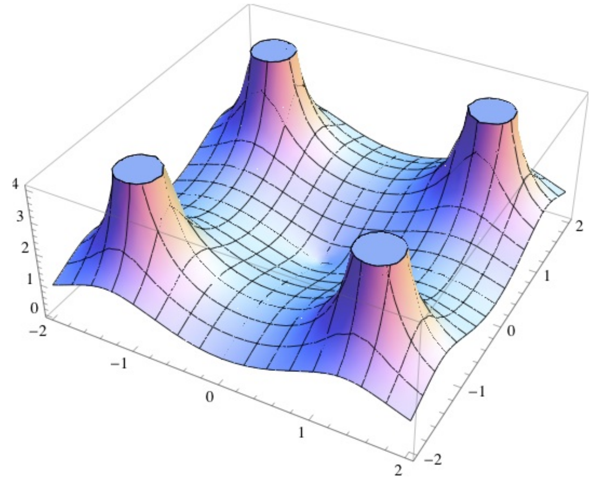
In particular, the solid boundary considered a paragraph earlier must go right, because the values are ultimately close to the these large values that map to the top right corner.

The remaining quadrants of our region are handled in the same way. So let's summarize. After we remove the four solid portions of the axes, the remaining region separates into four infinite quadrants, and each quadrant is mapped by F into one of the four portions of the illustrated $K \times K$ square.



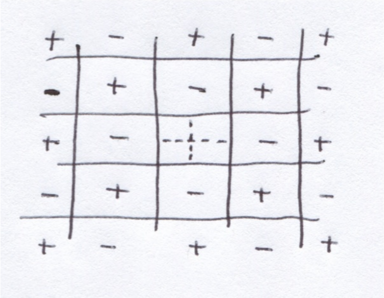
Now I claim that the map from our region to this square is one-to-one and onto. The key reason for this is that the derivative of $F(z)$ is $\frac{1}{\sqrt{1-z^4}}$ and this is not zero on the interior of our region. Consequently the map F cannot have folds. But if the interior of our region slops over the square under the mapping F, then this image must fold as it returns closer to the square as we reach the boundary and that cannot happen. Similarly if F takes the same value at two points inside, then there must be a fold between these points, which cannot happen. So F is one-to-one and onto on the interior, and thus we can form F^{-1}.

Below is a picture of F^{-1}, or rather, of its absolute value. The domain of this function is the square we just constructed, and it maps back to the original plane with portions of four axes removed. Since portions in the original quadrants close to infinity map to the corners of the square under F, the map F^{-1} gets very large near these four corners.



There is one further key fact. The values of F^{-1} on the left side of this square are exactly the negatives of the corresponding values on the right side. Similarly the values of F^{-1} on the bottom of this square are exactly the negatives of the corresponding values on the top. You can easily check this. For instance, F maps the removed interval $[1, \infty)$ to the right side of the square, and it maps the removed interval $(\infty, -1]$ to the left side of the square. If $x \in [1, \infty)$ is mapped to$\left(\frac{K}{2}, y\right)$ by F, then $-x$ is mapped to $\left(-\frac{K}{2}, y\right)$ by F. Similar remarks hold for the top and bottom maps.

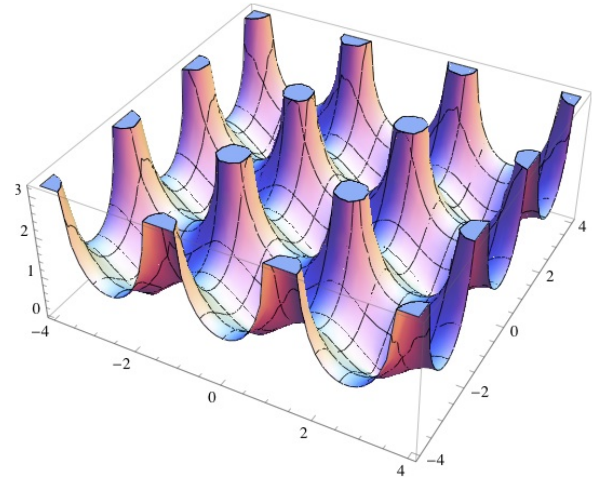
Because of these signs, we can glue our square into a grid of squares, as indicated below. Define a new function G from the entire plane to the complex numbers as follows. When a square has a plus sign, let $G = F^{-1}$, properly translated. When a square has a minus sign, let $G = -F^{-1}$. These pieces match on the boundary because of the remark in the previous paragraph.



This $G = F^{-1}$ is defined on the entire plane, and is holomorphic everywhere except on the grid lines, where it is continuous. There is one surprising theorem in complex analysis I forgot to tell you. It says that if a function is holomorphic on a region except on straight lines where it is merely continuous, then it is also holomorphic on the straight lines. That sure wouldn't be true in regular calculus, but it is true here. So G is actually a holomorphic function everywhere except the lattice points where it has poles.

The key point is that G is *doubly periodic*; it is periodic in two separate directions at the same time. The fact that F^{-1} has this property is the key discovery of Abel, and of Gauss. It turns out that the same thing happens if we start with *any elliptic integral of the first kind*, except that we often get a parallelogram rather than a perfect square. On the other hand, it only happens when integrating the square root of polynomials of degree 3 or 4, but not when integrating square roots of polynomials of higher degrees. That's why the elliptic case is so special.

Below is a picture of our $G = F^{-1}$ showing this double periodicity more clearly.

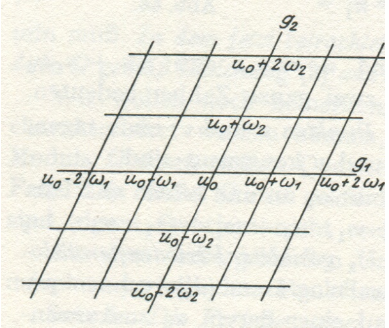


Whew. This paragraph was hard to write and I'm exhausted. I'm going to go to a movie and continue tomorrow!

9 The Weierstrass $\wp$ Function

I suspect you saw me sweating in the previous section, or at least desperately waving my hands. It is really hard to construct a doubly periodic function by inverting an elliptic integral! The great insight of Abel is that we shouldn't do it that way. It is far easier to construct doubly periodic functions from scratch, and then use them directly to calculate elliptic integrals. That is what Abel did, but I'm going to show you the easier construction due to Weierstrass about thirty years later.

Weierstrass begins by picking two linearly independent vectors in R^2, which we usually call ω_1 and ω_2. You can think of them are complex numbers. They will become the periods of our functions. The set of all $m\omega_1 + n\omega_2$ for integers m and n forms a lattice, which can be used to tile the plane with parallelograms. Below is a picture, where we usually set $u_0 = 0$.



We know from Gauss and Jacobi's $sn(z)$ that our functions will have singularities. To simplify matters, Weierstrass constructed a function whose singularities lie at the lattice points. The natural choice is $f(z) = \sum_{m,n} \frac{1}{z - m\omega_1 - n\omega_2}$, but this sum turns out not to converge. The next natural choice is $f(z) = \sum_{m,n} \frac{1}{(z - m\omega_1 - n\omega_2)^2}$, but this doesn't converge either, although it comes close. Weierstrass modified this sum by subtracting from each term its value at the origin; this makes the terms smaller close to the origin, and it turns out that the modified sum converges. Weierstrass' choice has been known ever since by the letter he used to describe it:

Definition 1 *The Weierstrass $\wp$ function is defined by*

$$\wp(z) = \frac{1}{z^2} + \sum_{(m,n)\neq(0,0)} \left(\frac{1}{(z - m\omega_1 - n\omega_2)^2} - \frac{1}{(m\omega_1 + n\omega_2)^2} \right)$$

Remark: It isn't clear that this function is doubly periodic, since for instance the singularity at the origin is treated differently than the remaining singularities. But when we differentiate, we get a function which is definitely doubly periodic:

$$\wp'(z) = (-2) \sum_{m,n)} \frac{1}{(z - m\omega_1 - n\omega_2)^3}$$

It follows that $\wp(z + \omega_1) - \wp(z)$ is constant, because its derivative is zero. Substituting $z = -\frac{\omega_1}{2}$ gives the value of this constant as $\wp\left(\frac{\omega_1}{2}\right) - \wp\left(-\frac{\omega_1}{2}\right)$. However, $\wp$ is clearly an even function taking the same value at z and $-z$, so this constant is zero and $\wp(z+\omega_1) = \wp(z)$. Similarly it is periodic in the ω_2 direction.

We can add, subtract, multiply, and divide doubly periodic functions to get more of them. Let us write $R(\wp)$ to denote an arbitrary rational expression of $\wp$. A typical such function is

$$\frac{\wp(z)^3 + 37\wp(z) - 5}{3wp(z)^2 + wp(z) + 1}$$

All of these functions are even, but the derivative of the $\wp$ function is odd: $\wp'(-z) = -\wp'(z)$. So the derivative of the $\wp$ function is not of the form $R(\wp)$.

Now we are ready for the big theorem:

Theorem 2 *Every doubly periodic function with periods ω_1 and ω_2 can be written uniquely in the form*

$$R_1(\wp(z)) + R_2(\wp(z))\wp'(z)$$

To me, this is amazing. There are so few doubly periodic functions that we can very explicitly describe them all. Every such function looks something like

$$\frac{\wp^3(z) + 3\wp(z) + 2}{\wp^2(z) + 7} + \frac{\wp(z) + 39}{\wp^5(z) + \wp(z)}\wp'(z)$$

Remark: If you add, subtract, multiply, or divide two doubly periodic functions, you get another one. The previous theorem almost, but not quite, describes this operation completely. Here is the missing ingredient:

Theorem 3

$$\left(\wp'(z)\right)^2 = 4\wp(z)^3 - g_2\ \wp(z) - g_3$$

where g_2 and g_3 are constants. Indeed

$$g_2 = 60 \sum_{(m,n)\neq(0,0)} \frac{1}{(m\omega_1 + n\omega_2)^4}$$

$$g_3 = 140 \sum_{(m,n)\neq(0,0)} \frac{1}{(m\omega_1 + n\omega_2)^6}$$

Remark: Once we have doubly periodic functions, we easily use them to compute elliptic integrals. Suppose we want to integrate

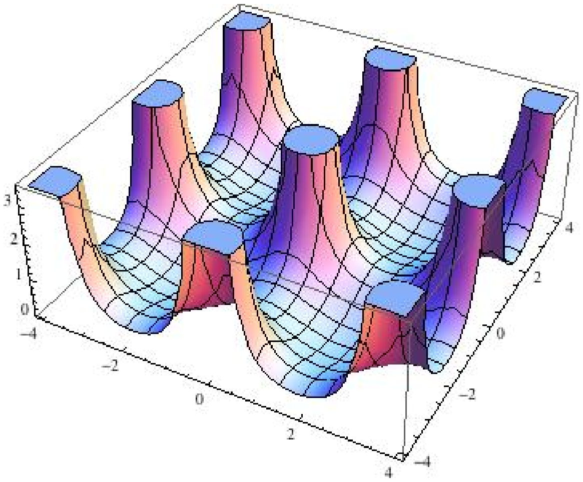
$$\int \frac{1}{\sqrt{x^3 - g_2 x - g_3}} dx$$

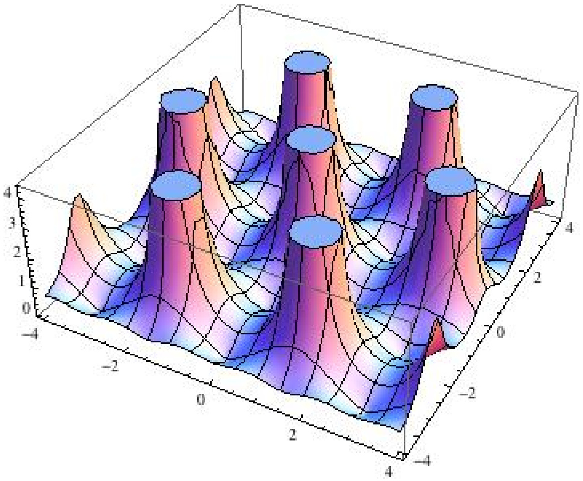
Make the substitution $x = \wp(z)$. Then the integral becomes

$$\int \frac{1}{\sqrt{\wp(z)^3 - g_2\wp(z) - g_3}}\wp'(z)\ dz = \int \frac{1}{\wp'(z)}\wp'(z)\ dz = \int 1\ dz = z + c = \wp^{-1}(x)$$

It is much easier this way than in the reverse direction starting with the elliptic integral.

Remark: Below are Weierstrass $\wp$ functions for the square lattice and the hexagonal lattice.





10 The Modular Picture

We now come to the climactic development. In this handout, we started with an ellipse and found a long chain of connections which ended with a doubly periodic function in the complex plane. There are infinitely many ellipses, depending on the constant k, so there are infinitely many chains of connections ending with infinitely many doubly periodic functions, each depending on a lattice of periodicity. Just as magnification and rotation of the ellipse are irrelevant and only eccentricity matters, so magnification and rotation of the lattice are irrelevant and only "lattice shape" matters. Our remaining goal is to explore this notion of "lattice shape."

To be more precise, the $\wp$ function and fundamental constants g_2 and g_3 change in easy ways if we magnify the lattice or rotate it. Multiplying by a complex number λ is the same thing as magnifying by $|\lambda|$ and then rotating by the argument of λ, so the claim that these operations are insignificant reduces to the following simple formulas:

$$\wp(\lambda z, \lambda\omega_1, \lambda\omega_2) = \frac{1}{(\lambda z)^2} + \sum \left(\frac{1}{(\lambda z - m\lambda\omega_1 - n\lambda\omega_2)^2} - \frac{1}{(m\lambda\omega_1 + n\lambda\omega_2)^2} \right) = \frac{1}{\lambda^2}\wp(z, \omega_1, \omega_2)$$

$$g_2(\lambda\omega_1, \lambda\omega_2) = 60 \sum \frac{1}{(m\lambda\omega_1 + n\lambda\omega_2)^4} = \frac{1}{\lambda^4} g_2(\omega_1, \omega_2)$$

$$g_3(\lambda\omega_1, \lambda\omega_2) = 140 \sum \frac{1}{(m\lambda\omega_1 + n\lambda\omega_2)^6} = \frac{1}{\lambda^6} g_3(\omega_1, \omega_2)$$

Therefore if we magnify and rotate, we don't have to recalculate $\wp$, g_2, and g_3; we can just multiply our previous calculations by appropriate powers of λ.

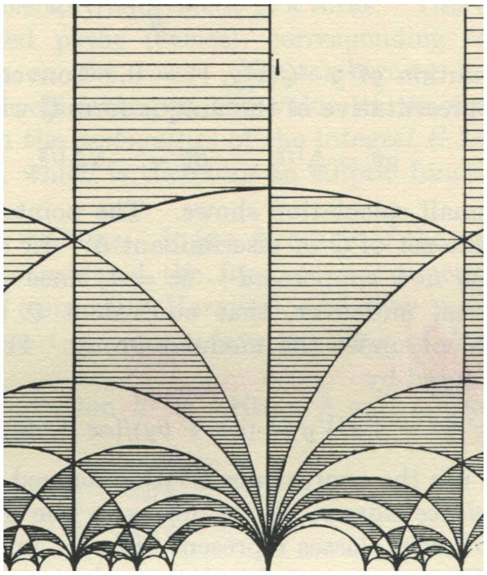
Moreover, the quantities $\wp, g_2$, and g_3 depend only on the lattice points, and not on the basis ω_1, ω_2. For instance, the lattice generated by 1 and i consists of all points with integer coefficients; this same set is generated by 1 and $i + 1$.

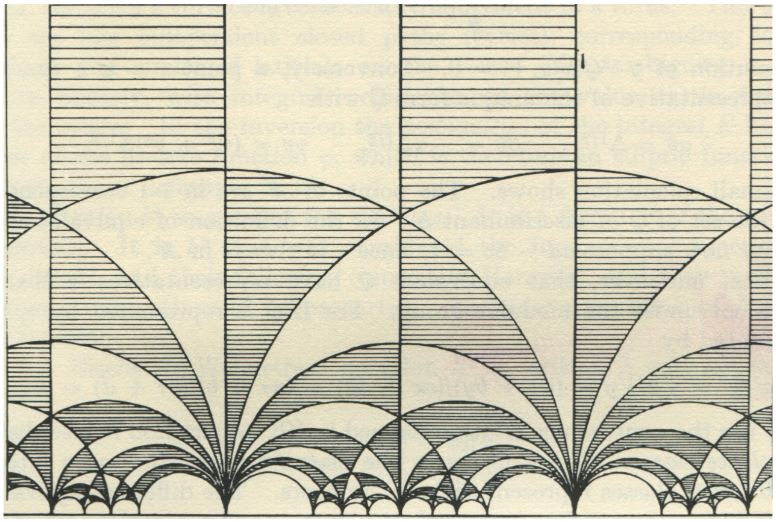
Our goal, then, is to classify lattices up to magnification, rotation, and change of basis. The ultimate theorem is very beautiful.

The theorem says that up these equivalences, every lattice can be generated by τ and 1 where τ is in the upper half plane. Moreover, this τ can be chosen in the vertical strip of numbers with real part between $-\frac{1}{2}$ and $\frac{1}{2}$. And finally, τ can be chosen to also satisfy $|\tau| >= 1$ and thus be outside the unit circle.

Two pictures of this region is shown on the next page.

The largest circle in the top picture has radius 1, and τ can be chosen outside it. The outside vertical lines are $x = -\frac{1}{2}$ and $x = \frac{1}{2}$ and τ can be chosen between them. So τ can be chosen in the union of the two vertical strips, one shaded and one not, which rise to infinity.





Actually, these strips can be replaced by the union of any two pieces provided one is shaded and one is not. So the small circular pictures near the x axis also suffice.

The idea of the proof will be important to us. Consider the two transformations $T(\tau) = \tau + 1$ and $S(\tau) = -\frac{1}{\tau}$ which map the upper half plane to itself. We claim that τ and $T(\tau)$ generate the same lattice, and τ and $S(\tau)$ generate equivalent lattices. The first assertion is easy since $m\tau + n$ can be rewritten $M(\tau + 1) + N$ by setting $M = m$ and $N = n - m$.

The second assertion is only slightly harder. If we magnify the lattice generated by τ and 1 by $\lambda = \frac{1}{\tau}$, we get an equivalent lattice generated by $\frac{1}{\tau}$ and 1. When τ is in the upper half plane, $\frac{1}{\tau}$ is in the lower half plane and $-\frac{1}{\tau}$ is in the upper half plane, and the new equivalent lattice can be generated by $S(\tau) = -\frac{1}{\tau}$ and 1.

We now sketch the proof of the main theorem. Suppose a lattice is generated by ω_1 and ω_2. Multiply by $\lambda = \frac{1}{\omega_2}$ to get an equivalent lattice generated by $\tau = \frac{\omega_1}{\omega_2}$ and 1. If τ is in the lower half plane, replace it by $-\tau$ in the upper half plane. The map T just translates right by one, so we can replace τ by an appropriate $T^k(\tau)$ whose real part is between $-\frac{1}{2}$ and $\frac{1}{2}$. If the resulting τ is inside the unit circle, replace it by $S(\tau)$ outside this circle.

Some details are missing from this proof. When τ is replaced by $S(\tau)$, the resulting number may no longer lie in the strip between $x = -\frac{1}{2}$ and $x = \frac{1}{2}$, so further translations may be necessary. These translations can bring the result back inside the circle of radius one, so further inversions may be needed. You can consult the literature to see why this process eventually ends.

A more serious objection is that there may be other transformations than S and T which produce equivalent lattices. Let's work this out. We know that every lattice is equivalent to a lattice generated by τ and 1 where τ is in the upper half plane. We could form an equivalent lattice by picking new generators $\omega_1 = a\tau + b$ and $\omega_2 = c\tau + d$ where a, b, c, d are integers. If these are really generators, we must be able to write $\tau = A\ \omega_1 + B\ \omega_2, 1 = C\ \omega_1 + D\ \omega_2$ for integers A, B, C, D. A little linear algebra shows that this will work only if $ad - bc = \pm 1$. Once we have a new basis, we can magnify by $\frac{1}{c\tau+d}$ to make the second generator 1, and then the first generator becomes $\frac{a\tau+b}{c\tau+d}$. If this new element is in the upper half plane, we need $ad - bc = 1$.

The set of all matrices $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with a, b, c, d integers and $ad - bc = 1$ forms a group named $SL(2, Z)$. Each element $A \in SL(2, Z)$ acts on the upper half plane by sending τ to $\frac{a\tau+b}{c\tau+d}$, and two elements define equivalent lattices if and only if there is an A mapping one to the other.

The essential fact about $SL(2, Z)$ is that S and T generate this group. Every element of the group is a product of powers of S and T. Consequently the study of lattice equivalence relations reduces to the study of S and T.

The proofs of all of these facts are interrelated; the full proof of the modular picture on page 24 is only slightly more difficult than this sketch and these results about $SL(2, Z)$ are corollaries.

11 Modular Forms

Long ago, we found that we could distinguish elliptical shapes by giving a single number k. Now we are going to show that we can define a similar number $j(\tau)$ which distinguishes lattice shapes. We are after a $j(\tau)$ that we can compute analytically without any group theory, which just happens to take the same values on points equivalent under the action of $SL(2,Z)$. Invariance requires only two equations:

$$j(T(\tau)) = j(\tau) \quad \text{and} \quad j(S(\tau) = j(\tau)$$

Temporarily we are going to call our unknown function $f(\tau)$ until the exciting final step. The first equation is just a fancy way to say that f must be periodic with period one. It is easy to produce a large number of candidates which satisfy this equation. If $g(z)$ is any holomorphic function on the unit disk, then $g(e^{2\pi i\tau})$ is defined on the upper half plane because

$$|e^{2\pi i(x+iy)}| = |e^{2\pi ix}|\;|e^{-2\pi y}| = e^{-2\pi y} < 1$$

and periodic because

$$e^{2\pi i(\tau+1)} = e^{2\pi i\tau}e^{2\pi i} = e^{2\pi i\tau}.$$

A holomorphic function on the unit disk can be expanded in a power series, so even more concretely, whenever we have a power series

$$g(z) = c_0 + c_1 z + c_2 z^2 + \dots$$

we get a candidate by writing $f(\tau) = g\left(e^{2\pi i\tau}\right)$. The rub is that it is extremely difficult to choose this power series so the corresponding $f(\tau)$ satisfies $f\left(-\frac{1}{\tau}\right) = f(\tau)$.

There are two candidates which come close, g_2 and g_3. These functions were studied by Eisenstein even before they came up naturally in Weierstrass' theory, so they are now called *Eisenstein series*:

$$g_2(\tau) = 60 \sum_{(m,n)\neq(0,0)} \frac{1}{(m\tau+n)^4} \qquad g_4(\tau) = 140 \sum_{(m,n)\neq(0,0)} \frac{1}{(m\tau+n)^6}$$

Since these functions depend only on lattice points, and not on the lattice basis, they satisfy $g_2(\tau+1) = g_2(\tau)$ and $g_3(\tau+1) = g_3(\tau)$. And since they behave under magnification as indicated on page 23, they satisfy $g_2\left(-\frac{1}{\tau}\right) = \tau^4\; g_2(\tau)$ and $g_3\left(-\frac{1}{\tau}\right) = \tau^6\; g_3(\tau)$

It isn't clear that these functions come from power series on the unit disk, but actually they do, and the calculation that shows this is one of the amusing exercises of complex analysis. The end result is that g_2 and g_3 come from following power series, where by definition $\sigma_k(n) = \sum_{d|n} d^k$:

$$g_2 : \frac{4\pi^4}{3}\left[1 + 240\sum_{n=1}^{\infty}\sigma_3(n)z^n\right]$$

$$g_3 : \frac{4\pi^6}{27}\left[1 - 504\sum_{n=1}^{\infty}\sigma_5(n)z^n\right]$$

It is useful to formalize what we have done so far before going on.

Definition 2 *Let $k \geq 0$ be an integer. A modular form of level k is a function $f(\tau)$ on the upper half plane associated with a power series $g(z) = c_0 + c_1 z + c_2 z^2 + \ldots$ by the formula $f(\tau) = g(e^{2\pi i\tau})$, which satisfies*

$$f\left(-\frac{1}{\tau}\right) = \tau^k f(\tau)$$

Let M_k be the set of all such modular forms.

We are interested in the special case when $k = 0$ because then f will be invariant under the transformation S. Sadly, we are going to be disappointed at first. Here is the amazing theorem

Theorem 4

1. *The M_k are finite dimensional vector spaces. When k is odd, they contain only the zero vector.*
2. *If k is even and $k \equiv 2 \pmod{12}, \dim M_k = \left[\frac{k}{12}\right]$*
3. *If k is even and $k \not\equiv 2 \pmod{12}, \dim M_k = \left[\frac{k}{12}\right] + 1$*
4. *Thus $M_0, M_2, M_4, M_6, M_8, M_{10},$ and M_{12} have dimensions $1, 0, 1, 1, 1, 1, 2$*
5. *The product of an element of M_k and an element of M_l is an element of M_{k+l}*
6. *$g_2 \in M_4$ and $g_3 \in M_6$*
7. *M_0 only contains constants*

Remark: This is a remarkable theorem! It shows dramatically just how restrictive the generalized equation $f\left(-\frac{1}{\tau}\right) = \tau^k f(\tau)$ is, since there are only a finite number of functions of a given level which satisfy it, up to linear combinations.

The dimension of M_0 is one, and since constants clearly satisfy the equations it follows that every element of M_0 is constant. So there are *no* functions arising from power series which are invariant under S and T. But don't despair; we'll fix this shortly.

Note that g_2 generates M_4 and g_4 generates M_6. Moreover, g_2^2 generates M_8 by 5), and $g_2 g_3$ generates M_{10}, again by 5). Finally, g_2^3 and g_3^2 belong to M_{12}.

Remark: And now the key remark. The space M_{12} is particularly interesting because its dimension is higher than one. Suppose $h_1(\tau)$ and $h_2(\tau)$ are linearly independent in M_{12}. Then

$$h_1\left(-\frac{1}{\tau}\right) = \tau^{12} h_1(\tau)$$

$$h_2\left(-\frac{1}{\tau}\right) = \tau^{12} h_2(\tau)$$

We could divide and define $j(\tau) = \frac{h_1(\tau)}{h_2(\tau)}$. Then the expression τ^{12} will cancel out and we have the long sought

$$j\left(-\frac{1}{\tau}\right) = j(\tau)$$

Of course this expression will be infinite when $h_2(\tau) = 0$, so we'd like h_2 to have as few zeros as possible in the upper half plane.

12 The Dedekind Eta Function

Do you remember the pentagonal number theorem? Really, after all this time?

The pentagonal number theorem gives a formal sum

$$g(z) = 1 - z - z^2 + z^5 + z^7 - z^{12} - z^{15} + z^{22} + z^{26} - \ldots$$

Earlier we didn't worry about convergence, but actually this series converges for $|z| < 1$ by comparison with the series $1 + z + z^2 + z^3 + \ldots$. So it induces a periodic function $f(\tau) = g\left(e^{2\pi i\tau}\right)$ as usual. But this is not very remarkable, since every convergent power series induces a periodic function. What would be remarkable would be a nice formula for $f\left(-\frac{1}{\tau}\right)$. There is absolutely no reason to expect such a formula.

Theorem 5 (Dedekind) *Consider the function* $\eta(\tau) = e^{\frac{2\pi i\tau}{24}} f(\tau)$. *Then*

$$\eta\left(-\frac{1}{\tau}\right) = \sqrt{-i\tau}\ \eta(\tau)$$

Remark: There are many proofs of this theorem. The proofs involve ingenious arguments in complex analysis, but are not really difficult.

The result can only be described as weird. The pentagonal number theorem has something to do with the modular theory, but what? There is that annoying twenty-fourth root, which ruins periodicity. And there is an unexpected square root of τ in a spot we usually see an integer power of τ.

However, we can fix all that. Instead of $\eta(\tau)$, consider $\eta^{24}(\tau) = e^{2\pi i\tau} f(\tau)$. This function satisfies

$$\eta^{24}(\tau+1) = \eta^{24}(\tau)$$

and

$$\eta^{24}\left(-\frac{1}{\tau}\right) = (-i\tau)^{12}\eta^{24}(\tau) = \tau^{12}\eta^{24}(\tau)$$

So $\eta^{24} \in M_{12}$. We could use it as a denominator to produce an invariant function, as we did at the end of the previous section. Indeed

$$j(\tau) = \frac{g_2^3(\tau)}{\eta^{24}(\tau)}$$

is completely invariant under $SL(2,Z)$ and defined except at zeros of η. So where are those zeros?

Remember again that $\eta^{24}(\tau) = e^{2\pi i\tau} g^{24}\left(e^{2\pi i\tau}\right)$ where g is given by the pentagonal number theorem. Thus η^{24} is obtained by plugging $z = e^{2\pi i\tau}$ in the power series

$$z\Big[1 - z - z^2 + z^5 + z^7 - z^{12} - z^{15} + z^{22} + \dots\Big]^{24}$$

$$= z\Big[(1-z)(1-z^2)(1-z^3)(1-z^4)\dots\Big]^{24}$$

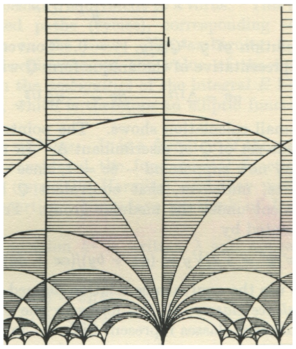
The zeros of this expression occur at $z = 0$ and at roots of unity on the unit circle. The mapping $\tau \to e^{2\pi i\tau}$ sends points in the upper half plane to non-zero points strictly inside the disk, so this function has no zeros at all in the upper half plane. It follows that $j(\tau)$, which is called by mathematicians the *Klein modular function*, has no singularities. Amazingly, the pentagonal number theorem has allowed us to produce an invariant function which is finite everywhere.

Even more is true. It is possible to prove that j takes every complex value exactly once in the fundamental region of the modular group. Consequently

Theorem 6 *Two lattices are equivalent under magnification, rotation, and basis change if and only if they have the same j-invariant.*

So j is the long-desired equivalent of k for lattices. Whew. It took longer for me to tell you that then I anticipated.

13 But There's One More Thing

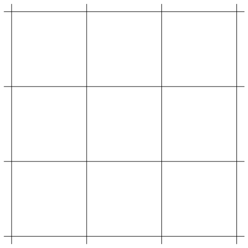


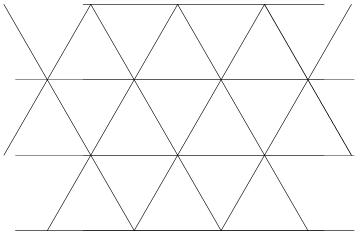
Consider the above picture. The map $e^{2\pi i\tau}$ carries the full portion of this picture between $x = -\frac{1}{2}$ and $x = \frac{1}{2}$ to the unit disk. In particular, it carries the boundary at the bottom where all of those circles terminate to the unit circle, and it takes the point at infinity beyond the top of the picture to the origin of the disk. Notice the extra z in G,

$$G(z) = z\Big[(1-z)(1-z^2)(1-z^3)(1-z^4)\dots\Big]^{24}$$

which makes this function vanish at the origin of the disk, or equivalently, at infinity beyond those shaded strips. By invariance, this function then needs to vanish at the ends of all equivalent shaded strips where they meet the x -axis. So invariance actually forces all those zeros at roots of unity.

There are two special points in the modular diagram. One is $\tau = i$, where the line $x = 0$ meets the circle in four separate regions. The other is $\tau = -\frac{1}{2} + \frac{\sqrt{3}}{2}i$, where six regions meet. There is something special about these values — they correspond to the square and hexagonal lattices, which are unusually symmetric.





Indeed, the square has 4-fold symmetric, and the hexagonal lattice has 6-fold symmetry. Compare this to the number 24 for the η function, and to the special roles of 2 and 3 in the theory of modular congruences.