

(February 27, 2006)

# Etymology, history, and artifacts

Paul Garrett garrett@math.umn.edu <http://www.math.umn.edu/~garrett/>

We will talk about *modular curves* or *modular forms* or *elliptic curves* or *elliptic modular something-or-others*, as examples of *automorphic* somethings. Without an explanation of the background, the reader who has not been completely beaten into submission will wonder what sense *modular* and *elliptic* have in this context. For example, ellipses and elliptic curves are not at all the same thing.

The etymology of *automorphic* is more reasonable, possibly suggesting the correct idea that an *automorphic* thing has *many symmetries*.

*Elliptic* really does have a connection with literal geometric *ellipses*, removed by a few steps. Integrals expressing arc length of ellipses are less elementary than integrals for inverse trigonometric functions, and were named *elliptic* because of this. Abel and Jacobi discovered in 1827 that the *inverse* functions are *doubly periodic* (explained below). This chain of events caused these functions to be called *elliptic functions*, which does not at all suggest that they are doubly periodic. Further, for a fixed periodicity lattice, the elliptic functions for that lattice satisfy *algebraic relations*. The basic relation is of the form (found by Weierstraß)

$$y^2 = 4x^3 - ax - b$$

Then such curves were called *elliptic curves*, because they relate the elliptic *functions* that invert the elliptic *integrals* that generalize the integral computing the arc length of an *ellipse*.

That story continues with *abelian integrals*, meaning more complicated multi-valued algebraic integrals, whose inverse functions are multiply periodic, and called *abelian functions*. These abelian functions satisfy *algebraic* relations which define *abelian varieties* as generalizations of elliptic curves. *Jacobian varieties* are a special case. These are so-named exactly because Abel and Jacobi studied them. <sup>[1]</sup>

The use of *modular* is *not* like the colloquial sense of admitting convenient rearrangement in pieces. Rather, the usage is more connected with *modul(e)* (the *e* was added only later), meaning something like a *lattice*  $\mathbb{Z}^n$  in  $\mathbb{R}^n$ , in a sense that later became *module* as in *module over a ring*. It seems likely that the contemporary use of *moduli space* as a way of saying *parameter space* (for some class of objects) is descended from the idea of looking at families of modules (lattices). One incarnation of *modular forms* is as functions on some space of lattices. Specifically (as explained below) *form* means *homogeneous function*.

Certain integrals arising in the discussion of *periodicity* thus earn the name *period integrals* or *periods*.

- Reconsideration of  $\sin x$
- Construction of singly-periodic functions
- Arc length of ellipses: elliptic integrals
- Doubly-periodic functions: elliptic functions
- Elliptic modular functions

---

## 1. Reconsideration of $\sin x$

By now we are well acquainted with the integral

$$g(x) = \int_0^x \frac{dt}{\sqrt{1-t^2}}$$

---

[1] A tiny irony is that abelian varieties *are* abelian groups, in the sense of being *commutative*.

and we know that this is  $\arcsin x$ . We also know that there is an issue about *multi-valued-ness* of  $\arcsin$ .<sup>[2]</sup> Instead of trying to *circumvent* the multi-valued-ness of  $\arcsin$ , we should follow Abel and Jacobi (1820's) and look at the *inverse* function,  $\sin$ , which will be *periodic* because  $\arcsin$  is multi-valued.

Ignoring the fact that we already knew something about the behavior of  $\arcsin$ , we can look at this integral directly, as practice for a more complicated example shortly. *And* we should think of it in the context of *complex* analysis, not merely real-number calculus. In that setting, there is great ambiguity in the choice of *path*  $\gamma$  from 0 to  $x$ , and, in general, the value of an integral depends greatly on the choice of path.<sup>[3]</sup> The arguably simplest choice is the obvious

$$\gamma : [0, 1] \longrightarrow \mathbb{C} \quad \text{by} \quad \gamma(t) = tx$$

which (for  $x \in (-1, +1)$ ) gives the usual choice of value. But promoting this choice doesn't explain the periodicity of  $\sin$ .

As a function of a complex variable, staying away from the bad points  $\pm 1$  where the argument to the square root vanishes, there are *two* square roots of  $1 - z^2$ . This does complicate matters, but, despite this, along every path in  $\Omega = \mathbb{C} - \{\pm 1\}$  the function  $\sqrt{1 - z^2}$  has an *analytic continuation*.<sup>[4]</sup> We make an initial choice of square roots near 0 by the convention, in effect at the *beginning* of a path from 0, that  $\sqrt{1} = +1$ .

Thus, for any path  $\gamma$  in  $\mathbb{C} - \{\pm 1\}$  from 0 to  $z \in \mathbb{C}$ , the integral

$$\int_{\gamma} \frac{d\zeta}{\sqrt{1 - \zeta^2}}$$

is *well-defined*. The basic ambiguity is that for a particular  $z$  we might choose a *different* path  $\delta$  which traverses extra *loops*<sup>[5]</sup> to get to the same point. Thus, *the ambiguities or multi-valuedness of the integral for arcsin are the integrals around closed paths (loops)*

$$\int_{\gamma} \frac{d\zeta}{\sqrt{1 - \zeta^2}} \quad (\gamma \text{ closed path})$$

[2] Typically, we impose some dumb rules on students to give an apparent resolution of the multi-valued-ness issue. There is little reason to heed or value these rules, of course.

[3] Properly, a *path*  $\gamma$  in an open set  $U \subset \mathbb{C}$  is a continuous map  $\gamma : [0, 1] \longrightarrow U$ . We want the ambient set to be *open* so that (to talk about analytic continuation) we can cover the path by disks inside the set  $U$ . The path  $\gamma$  is *closed* if  $\gamma(0) = \gamma(1)$ . In fact, we can only evaluate path *integrals* along *rectifiable* paths, which in practical terms means that the path is at least once-continuously-differentiable, apart from possibly finitely-many points of non-differentiability. Then the path integral of  $f$  along  $\gamma$  is  $\int_0^1 f(\gamma(t))\gamma'(t) dt$ .

[4] The notion of *analytic continuation* certainly needs explanation, especially since the usual context in which this appears offers little motivation. The procedure is roughly as follows, in an example. Let  $f$  be holomorphic on  $U = \mathbb{C} - \{z_1, \dots, z_n\}$ . To analytically continue a *square root*  $g$  of  $f$  at  $\gamma(0) = z_o$  along a path  $\gamma$  inside  $U$ , first cover (the image of)  $\gamma$  by *small* open disks  $D_1, D_2, \dots, D_n$  all lying inside  $U$ , such that  $D_i \cap D_{i+1} \neq \emptyset$  and  $D_1 \ni z_o$ . The disks must be small enough such that  $D_i \cap \gamma[0, 1]$  is *connected*. Choose square roots  $g_i$  of  $f$  on  $D_i$  such that  $g_i$  and  $g_{i+1}$  agree on  $D_i \cap D_{i+1}$  and such that  $g_1$  agrees with the given  $g$  at  $z_o$ . Then, the sequence of only locally defined functions is the analytic continuation. Note that the principal complication is that the path may intersect itself *and* it may be necessary to assign *different* fragments to the same point in  $\mathbb{C}$  corresponding to different parts of the path.

[5] *Loop* is synonym for *closed path*, probably *not* going directly through bad points. Here we are interested in closed paths around bad points. Changing these paths by continuous deformations (homotopies) that do not move paths across the points where the denominator has a 0 has no effect on the integrals' values, by Cauchy's theorem.

As we may remember, the value of any path integral depends only on its *homotopy class*<sup>[6]</sup> (fixing endpoints),<sup>[7]</sup> inside the region where the integrand is holomorphic.<sup>[8]</sup> In the present situation, the two-valued nature of the square root  $\sqrt{1-z^2}$  might suggest that we should only consider loops that go around *both* bad points  $\pm 1$ , or, perhaps, *twice* around a single bad point  $\pm 1$ , in order to return to the *same* square root.

**Claim:** Let  $\gamma$  be the path tracing a (large) circle of radius  $R$ , centered at 0. Then

$$\int_{\gamma} \frac{d\zeta}{\sqrt{1-\zeta^2}} = 2\pi$$

where the choice of square root in the integrand is made such that  $\sqrt{1-R^2} = i\sqrt{R^2-1}$ .

**Remark:** The fact that the indicated integral is  $2\pi$  proves that the sine function is periodic with period<sup>[9]</sup>  $2\pi$ , meaning exactly that

$$\sin(x+2\pi) = \sin x \quad (\text{for all } x)$$

That is, *arcsine* is *ambiguous* by multiples of  $2\pi$ , so its inverse function *sine* is *periodic* by multiples of  $2\pi$ .

*Proof:* For  $z$  a complex number with  $|z| < R$ , we can consider the function

$$F(z) = \int_{\gamma} \frac{d\zeta}{\sqrt{z-\zeta^2}}$$

For each  $\zeta$  with  $|\zeta| = R$ , there is a holomorphic (in  $z$ )  $\sqrt{z-\zeta^2}$  for  $|z| < R^2$ .<sup>[10]</sup> Then the *integral* with respect to  $\zeta$  of these holomorphic functions *should*<sup>[11]</sup> give a holomorphic function  $F(z)$ . Granting this, and

- 
- [6] The idea of *homotopy* is the equivalence relation on paths of being *continuous deformable* into each other without going outside the given region. *Fixing endpoints* means, as it sounds, that the endpoints should be kept fixed throughout the deformation. This still can be made (more) precise, but probably the reader completely unacquainted with the notion should try to *imagine* the *intent* rather than study the formal definition, for now.
- [7] Since path integrals are defined only for piecewise differentiable paths, one should worry that a merely *continuous* homotopy would involve intermediate paths that are not differentiable. This concern is justified, and one should prove, sometime, that existence of a *continuous* homotopy between piecewise differentiable paths implies existence of a suitably piecewise *differentiable* homotopy.
- [8] The *Seifert-vanKampen theorem* implies that the *homotopy group* of the plane with two points removed is free on two generators, the generators being two loops, one loop going around one point, the other loop going around the other. Such stuff provides reassurance that in fact the present heuristic discussion is fairly complete, though proof of completeness is not easy.
- [9] We are not making any claim that  $2\pi$  is the *smallest* period, or that there are not *other* symmetries.
- [10] The most direct argument for the existence of such a holomorphic square root is by using the logarithm  $\log z = \int_1^z d\zeta/\zeta$ , and taking advantage of the fact that the disk  $|z| < R$  is not only *simply-connected* (all paths are homotopic to points), but is *convex*, which greatly simplifies the verification that a holomorphic square root can be defined on it.
- [11] It really is true that an integral of holomorphic functions that depend *nicely* on a parameter is itself holomorphic. In the present context, one should probably accept this as at least a compelling heuristic, whether or not one sees how to prove it. A relatively elementary yet fairly conceptual proof can be given by using *Morera's theorem* together with *Fubini's theorem*, as follows. Morera's theorem asserts that a function is holomorphic if and only if its path *integrals* over small circles (or small squares, etc.) are all 0. This is a very useful conversion of a *derivative* condition to an *integral* condition. Fubini's theorem says that the order of integration can be interchanged if the double integral (here integral and a sum) is absolutely convergent. Since we are integrating continuous functions over compacta, this integrability holds, so the integral is holomorphic. This is an example of a happy ending which can be anticipated much more broadly, as we will see later.

granting that we can interchange differentiation and integration<sup>[12]</sup>

$$F'(z) = -\frac{1}{2} \int_{\gamma} \frac{d\zeta}{(z - \zeta^2)^{3/2}}$$

We estimate the absolute value of the latter integral, perhaps keeping  $|z| < R^2/2$  to make the estimate clearer, giving

$$|F'(z)| \leq \frac{1}{2} \cdot (2\pi R) \cdot \frac{1}{(R^2/2)^{3/2}} = \text{constant} \times \frac{1}{R^2}$$

Now notice that changing the path  $\gamma$  to have a larger radius is a homotopy that does not change the value of the integral. Thus, we may take  $R$  very large, and see that necessarily  $F'(z) = 0$  for  $|z| < R^2$ .

Thus,  $F(z)$  is *constant*, and we can evaluate the constant by taking  $z = 0$  (rather than the original  $z = 1$ ). Then the simplification

$$\sqrt{0 - \zeta^2} = \pm i \cdot \sqrt{\zeta^2} = \pm i \zeta$$

allows the integral to become

$$F(1) = F(0) = \int_{\gamma} \frac{d\zeta}{\pm i \zeta} = \frac{2\pi i}{\pm i} = \pm 2\pi$$

Since we would include any positive or negative integer multiples of this path as further possible ambiguities, we won't worry about figuring out the sign. ///

**Remark:** It must not be overlooked that the ambiguity of value arising from such integrals over closed paths is an ambiguity that is *the same* for all possible arguments to arcsin.

**Remark:** There is considerable charm in an argument which evaluates an integral by first making a constant be a variable parameter, showing that the whole depends rather innocently upon that parameter, then changing the parameter to a value where the whole is much easier to evaluate.

## 2. Construction of singly periodic functions

A **singly-periodic** function  $f$  on  $\mathbb{C}$  is a (probably holomorphic or meromorphic) function such that for some  $\omega \neq 0$

$$f(z + \omega) = f(z) \quad (\text{for all } z \in \mathbb{C})$$

(or at least for  $z$  away from poles of  $f$ ). Of course, since this holds for all  $z$ , for any integer  $n$  we have

$$f(z + n\omega) = f(z)$$

In other words,  $f$  is *invariant* under translation by the group  $\mathbb{Z} \cdot \omega$  inside  $\mathbb{C}$ .

Feigning ignorance of the trigonometric (and exponential) function, whether as inverse functions to integrals or not, as a warm-up to the later construction of *doubly*-periodic functions we should try to construct some singly-periodic functions.

For simplicity, let's try to make holomorphic or meromorphic functions  $f$  such that

$$f(z + 1) = f(z) \quad (\text{for all } z \in \mathbb{C})$$

---

<sup>[12]</sup> Differentiating an integral with respect to a parameter by differentiating inside the integral is non-trivial to justify, and it would be sad to give an ugly argument for this. The cleanest and most general argument for such invokes a yet more general, and useful, idea about characterization of vector-valued (for example, holomorphic-function-valued) integrals. Gelfand and Pettis (separately) developed the kernel of this idea starting about 1928.

That is, we might say that we want  $\mathbb{Z}$ -periodic functions on  $\mathbb{C}$ . A fundamental approach to manufacturing such things is *averaging*, also called *periodicization* or *automorphizing*, described as follows. Let  $\varphi$  be a given function, and consider

$$f(z) = \sum_{n \in \mathbb{Z}} \varphi(z + n)$$

If this converges reasonably<sup>[13]</sup> it is certainly periodic with period 1, since

$$f(z + 1) = \sum_{n \in \mathbb{Z}} \varphi(z + 1 + n) = \sum_{n \in \mathbb{Z}} \varphi(z + n)$$

by replacing  $n$  by  $n - 1$ , using the fact that we have summed over a group, and using the good convergence to be sure that rearrangements don't affect the sum.

An elementary function which makes the sum converge, apart from some poles, is  $\varphi(z) = 1/z^2$ . Thus, consider

$$f(z) = \sum_{n \in \mathbb{Z}} \frac{1}{(z + n)^2}$$

If we are lucky, this construction might manufacture a function related to the *sine* function, which arose here as the inverse function to the integral

$$\int_0^w \frac{d\zeta}{\sqrt{1 - \zeta^2}}$$

One traditional idea is to try to *guess* an expression for this function in terms of exponential functions by matching poles, subtract, prove that the difference has *no poles and* goes to 0 at infinity, so is 0.<sup>[14]</sup> To do this, we should first determine the Laurent expansion of  $f(z)$  near its poles. By periodicity, we'll understand all the poles if we understand the pole at  $z = 0$ . This is

$$f(z) = \frac{1}{z^2} + \sum_{n \neq 0} \frac{1}{(z + n)^2} = \frac{1}{z^2} + (\text{holomorphic near } z = 0)$$

A first reasonable guess for a function with double poles at integers expressible in terms of exponentials is something like

$$\left( \frac{1}{e^{2\pi iz} - 1} \right)^2$$

To understand the nature of each pole in detail, by periodicity it suffices to look near  $z = 0$ . There, we have

$$\begin{aligned} \left( \frac{1}{e^{2\pi iz} - 1} \right)^2 &= \left( \frac{1}{(1 + 2\pi iz + \dots) - 1} \right)^2 = \left( \frac{1}{2\pi iz + (2\pi iz)^2/2 + \dots} \right)^2 \\ &= \frac{1}{(2\pi i)^2} \left( \frac{1}{z + (\pi i)z^2 + \dots} \right)^2 \end{aligned}$$

by pulling the constant off the leading term of the power series in the denominator. But (using the geometric series expansion  $1/(1 + r) = 1 - r + \dots$ )

[13] *Reasonable* convergence is probably absolute convergence, and uniformly for the parameter on compacts not containing the obvious poles.

[14] Invoking Liouville's theorem: a bounded entire function (*entire* meaning that it is holomorphic in the entire plane) is *constant*. Thus, as an immediate corollary, an entire function which is bounded *and* goes to 0 as the imaginary part of  $z$  goes to infinity must be 0.

$$\begin{aligned} & \frac{1}{(z + (\pi i)z^2 + \dots)^2} = \frac{1}{z^2} \cdot \frac{1}{(1 + (2\pi iz + \dots))^2} \\ &= \frac{1}{z^2} \cdot (1 - (2\pi iz + \dots) + (2\pi iz + \dots)^2 - \dots)^2 = \frac{1}{z^2} \cdot (1 - 4\pi iz + \dots) \\ &= \frac{1}{z^2} - \frac{4\pi i}{z} + (\text{holomorphic near } z = 0) \end{aligned}$$

That is, even ignoring the constant  $(2\pi i)^2$ , we've not quite got the sort of pole we want, since we need the  $1/z$  term to be 0. Tweaking the thing slightly leads to

$$\left( \frac{2\pi i}{e^{\pi iz} - e^{-\pi iz}} \right)^2$$

which has the same Laurent expansion at poles, and goes to 0 when the imaginary part of  $z$  is large. Thus,

$$\sum_{n \in \mathbb{Z}} \frac{1}{(z+n)^2} - \left( \frac{2\pi i}{e^{\pi iz} - e^{-\pi iz}} \right)^2 = 0$$

since this difference has no poles, is bounded,<sup>[15]</sup> and goes to 0 when the imaginary part of  $z$  is large. That is, via complex analysis,

$$\sum_{n \in \mathbb{Z}} \frac{1}{(z+n)^2} = \frac{\pi^2}{\sin^2 \pi z}$$

**Remark:** That is, our *singly* periodic function  $\sum 1/(z+n)^2$  turns out to be identifiable in terms of already familiar items. Since this will *not* be the case for *doubly* periodic functions, it is profitable to experiment with another viewpoint that does *not* depend upon reduction to simpler things.

We reconsider  $f(z) = \sum 1/(z+n)^2$  by ascertaining a *differential equation* it satisfies. Without worrying about justifying<sup>[16]</sup> differentiation term-by-term, we have

$$f'(z) = -2 \sum \frac{1}{(z+n)^3}$$

Planning on invoking (again) Liouville's theorem, if we manage to arrange a polynomial in  $f$  and  $f'$  whose poles cancel, then this polynomial is necessarily *constant*, and we'll find a polynomial relation<sup>[17]</sup> between  $f'$  and  $f$ . It suffices to see the pole at  $z = 0$  cancel. Since  $f(z) = \frac{1}{z^2} + f_o(z)$  with

$$f_o(z) = \sum_{n \neq 0} \frac{1}{(z+n)^2}$$

[15] In the strip where  $|y| \leq 1$  (where  $y$  is the imaginary part of  $z$ ) the difference of  $f(z)$  and the exponential expression has no poles, so is continuous. It is therefore bounded on the compact set  $|z| \leq 2$ . Then, by periodicity, it is bounded on the whole strip  $|y| \leq 1$ . Easier estimates give the boundedness *off* this strip.

[16] Indeed, we are implicitly supposing that such sums give meromorphic functions, and can be differentiated term-by-term. That this is so can be proven by verifying that the series converges in a suitable sup-norm on compacta (away from the poles), and then invoking Morera's theorem (that functions with path integrals 0 over small triangles are holomorphic) to see that the limit of the partial sums is (locally) holomorphic.

[17] A non-linear polynomial relation between  $f$  and  $f'$  will be a *non-linear*, thus probably hard-to-solve, differential equation. The difficulty of solving non-linear differential equations in general is not the point here, however.

holomorphic near 0, the Laurent expansion of  $f(z)$  near  $z = 0$  is just the sum of  $1/z^2$  and the *power series* expansion of  $f_o(z)$  there. That is,

$$f(z) = \frac{1}{z^2} + f_o(0) + \frac{f_o''(z)}{2!}z^2 + \dots$$

The fact that  $f(-z) = f(z)$  assures that all the odd-order terms vanish. We have<sup>[18]</sup>

$$f_o(0) = \sum_{n \neq 0} \frac{1}{n^2} = 2\zeta(2) = \frac{\pi^2}{3}$$

and

$$f_o''(0)/2! = \frac{1}{2} \sum_{n \neq 0} \frac{(-2)(-3)}{n^4} = 6\zeta(4) = \frac{\pi^4}{15}$$

But we won't use the specific values for a moment. Thus, write<sup>[19]</sup>

$$f(z) = \frac{1}{z^2} + a + bz^2 + O(z^4)$$

Then

$$f'(z) = \frac{-2}{z^3} + 2bz + O(z^3)$$

And

$$\left(\frac{f'(z)}{-2}\right)^2 = \frac{1}{z^6} - \frac{2b}{z^2} + O(1)$$

We compute (carefully!?) that

$$f(z)^3 = \frac{1}{z^6} + \frac{3a}{z^4} + \frac{3a^2 + 3b}{z^2} + O(1)$$

so

$$\left(\frac{f'(z)}{-2}\right)^2 - f(z)^3 = -\frac{3a}{z^4} - \frac{3a^2 + 5b}{z^2} + O(1)$$

Using

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

we obtain

$$\left(\frac{f'(z)}{-2}\right)^2 - f(z)^3 + 3a \cdot f(z)^2 = \frac{-3a^2 - 5b + 6a^2}{z^2} + O(1) = \frac{3a^2 - 5b}{z^2} + O(1)$$

At this, point a small miracle occurs:

$$3a^2 - 5b = 3(\pi^2/3)^2 - 5\pi^4/15 = \pi^4 \cdot [1/3 - 1/3] = 0$$

Thus,

$$\left(\frac{f'(z)}{-2}\right)^2 - f(z)^3 + 3a \cdot f(z)^2 = O(1)$$

<sup>[18]</sup> Recall that the Euler-Riemann zeta function is  $\zeta(s) = \sum_{n \geq 1} 1/n^s$ . Evaluation of  $\zeta(s)$  had stymied the Bernouillis, and was one of Euler's first mathematical coups. For the present discussion, we'll take for granted that  $\zeta(2) = \pi^2/6$  and  $\zeta(4) = \pi^4/90$ . A bit later we will prove these as corollaries of *Fourier series* expansions of elementary functions.

<sup>[19]</sup> The *big-oh* notation  $O(z^4)$  means, as usual, that the rest is divisible by  $z^4$ .

and, by the same Liouville-theorem-based argument,

$$\left(\frac{f'(z)}{-2}\right)^2 - f(z)^3 + 3a \cdot f(z)^2 = 0$$

giving a polynomial relation<sup>[20]</sup> between  $f$  and  $f'$

$$\boxed{f'^2 = 4f^2(f - \pi^2)}$$

It is plausible that the double poles of  $f$  would suggest trying to take its *square root*, as well as taking its *reciprocal*, so let

$$u = \frac{1}{f^{1/2}}$$

Then

$$u' = -\frac{1}{2} \frac{f'}{f^{3/2}}$$

This suggests rearranging the boxed expression above into the form

$$\frac{f'^2}{4f^3} = 1 - \pi^2 \frac{1}{f}$$

which is

$$u'^2 = 1 - \pi^2 u^2$$

Of course, we can see in this the identity  $\cos^2 x = 1 - \sin^2 x$ , but let's continue to feign ignorance. Differentiating both sides will get rid of the constant 1, and give

$$2u''u' = -\pi^2 2u'u$$

Dividing through by  $2u'$  finally gives a *linear* differential equation, with constant coefficients:

$$u'' = -\pi^2 \cdot u$$

Thus, the presence of (complex) exponential functions can be detected in many different ways.

**Remark:** It is not at all clear that the function  $\sum 1/(z+n)^2$  should actually be of *exponential* decay as the imaginary part of  $z$  goes to infinity. But, from the re-expression in terms of exponential functions, it *is*.

### 3. Arc length of ellipses: elliptic integrals

One might naturally be interested in the integral for the length of a piece of arc of an ellipse. For example, the arc length of the piece the ellipse

$$kx^2 + y^2 = 1 \quad (\text{with } k > 0)$$

in the first quadrant is

$$\int_0^1 \sqrt{1 + \left(\frac{-\frac{1}{2} \cdot 2kx}{\sqrt{1-kx^2}}\right)^2} dx = \int_0^1 \sqrt{1 + \frac{k^2 x^2}{1-kx^2}} dx = \frac{1}{\sqrt{k}} \int_0^{\sqrt{k}} \sqrt{\frac{1-(1-k)x^2}{1-x^2}} dx$$

<sup>[20]</sup> This algebraic relation between a transcendental function and its derivative presages the *Weierstraß* equation for elliptic functions, below

by replacing  $x$  by  $x/\sqrt{k}$  in the last step. For the *circle*,  $k = 1$ , the numerator under the radical simplifies to 1, and we find a value of the integral discussed above, namely

$$\int_0^1 \frac{1}{\sqrt{1-x^2}} dx = \frac{\pi}{4}$$

Otherwise, though, there is no obvious reduction to truly elementary integrals. This somehow justifies calling an integral

$$\int_a^b \frac{\text{(rational expression in } z)}{\sqrt{\text{cubic or quartic in } z}} dz$$

an **elliptic integral**.<sup>[21]</sup> Many people studied the effect of changes of variables to transmute one form into another.<sup>[22]</sup>

As often happens in mathematics, the immediate problems of computing arc length or evaluating new integrals was eclipsed by a higher-level idea, in this case the discovery by Abel and Jacobi (independently) in 1827 of the *double periodicity*<sup>[23]</sup> of functions  $f(z)$  defined by

$$z = \int_0^{f(z)} \frac{d\zeta}{\sqrt{\text{quartic in } \zeta \text{ with distinct factors}}}$$

That is, there are **periods**  $\omega_1$  and  $\omega_2$  in  $\mathbb{C}$  such that

$$f(z + \omega_1) = f(z + \omega_2) = f(z) \quad (\text{for all } z \in \mathbb{C})$$

and  $\omega_1$  and  $\omega_2$  are linearly independent over  $\mathbb{R}$ .<sup>[24]</sup> These  $\omega_1$  and  $\omega_2$  will arise as *integrals* of  $1/\sqrt{\text{quartic}}$  over closed paths, which is why these integrals themselves have come to be called **period integrals**, or simply **periods**.

For example, write  $w = f(z)$  and consider

$$z = \int_0^w \frac{d\zeta}{\sqrt{1 + \zeta^4}}$$

The uniform *ambiguities* in the value of this integral viewed as a *path integral* from 0 to  $w$  are the values of the integrals along closed paths which circle an *even* number of the bad points  $e^{2\pi ik/8}$  with  $k = 1, 3, 5, 7$  (primitive 8<sup>th</sup> roots of 1).<sup>[25]</sup> With this denominator like  $1/|\zeta|^2$  for large  $|\zeta|$ , the integral over large circles

[21] It turns out, with hindsight, that when the expression inside the radical has more than 4 zeros, its behavior is similar but yet more complicated. Similarly, if the square root is replaced by a higher-order root, the behavior becomes more complicated. Thus, the case of square root of cubic or quartic is the simplest beyond more elementary integrals. Abel and Jacobi and others *did* subsequently consider the more complicated cases. This was a popular pastime throughout the 19<sup>th</sup> century.

[22] Already by 1757 Euler had studied relationships of the form  $dx/\sqrt{x^4+1} + dy/\sqrt{y^4+1} = 0$ , which lead to algebraic relations between  $x$  and  $y$ . Legendre (about 1811) had studied algebraic transformations of such integrals, giving special forms to which they can be reduced by various algebraic means.

[23] Gauss later claimed that he had found the double periodicity earlier, but had not published it. Abel and Jacobi published in 1827, and Legendre very civilly acknowledged their work in a new edition of his *Exercices de Calcul Intégral*. Later archival work did verify that Gauss had *privately* found the double periodicity in 1809.

[24] So  $\omega_1$  and  $\omega_2$  point in genuinely different directions in  $\mathbb{C}$ .

[25] As in the simpler case of the circle, to have the path return to the same square root at the end point as at the start requires that an even number of zeros be encircled. This becomes more intuitive if one speaks in terms of *sheets* of *Riemann surfaces* covering  $\mathbb{C}$ , but this viewpoint itself has a non-trivial cost.

goes to 0 as the radius goes to  $+\infty$ , unlike the earlier case. And the integral around two points added to the integral around the other two points is equal to<sup>[26]</sup> that outer circle integral, which is 0, so any two such integrals are merely negatives of each other.

Because of the decay for  $|\zeta|$  large, the integral of  $1/\sqrt{1+\zeta^4}$  along a path encircling  $e^{2\pi i/8}$  and  $e^{2\pi i 3/8}$  is equal (via a homotopy) to the integral along the real axis, namely

$$\lambda = \int_{-\infty}^{+\infty} \frac{dt}{\sqrt{1+t^4}}$$

Whatever else this may be, it is a positive real number. Similarly, the integral along a path encircling  $e^{2\pi i/8}$  and  $e^{2\pi i 7/8}$  is equal (via a homotopy) to the path integral along the imaginary axis, namely

$$\int_{-\infty}^{+\infty} \frac{d(it)}{\sqrt{1+(it)^4}} = i \cdot \int_{-\infty}^{+\infty} \frac{dt}{\sqrt{1+t^4}} = i\lambda$$

since  $i^4 = 1$ . Thus, defining

$$f(z) = w \quad \text{by} \quad z = \int_0^w \frac{d\zeta}{\sqrt{1+\zeta^4}}$$

we have

$$f(z + \lambda) = f(z) = f(z + i\lambda)$$

Since this holds for all complex  $z$ , for all integers  $m, n$

$$f(z + \lambda(m + in)) = f(z)$$

That is,  $f$  is **(doubly) periodic**, with respect to the **period lattice**

$$\Lambda = \mathbb{Z} \cdot \lambda + \mathbb{Z} \cdot i\lambda \subset \mathbb{C}$$

The modifier *doubly* emphasizes that the collection

$$\{\lambda \in \mathbb{C} : f(z + \lambda) = f(z) \text{ for all } z\}$$

is a  $\mathbb{Z}$ -module on *two* generators, rather than on a *single* generator.

**Remark:** The two integrals above are **periods** of the integrand. We made a choice of integral to make the linear independence (over  $\mathbb{R}$ ) of the two *periods* easy to verify.

**Remark:** In our example, the function  $f(z)$  has *poles*. Notice that the integral along the positive real axis

$$\int_0^{+\infty} \frac{dt}{\sqrt{1+t^4}}$$

is absolutely convergent, with value  $\lambda/2$ , where  $\lambda$  is the *whole* integral on the real line, as just above. That is,  $f(\lambda/2) = \infty$ , which is to say that  $f$  has a pole at  $\lambda/2$ . Likewise, the integral along the upper imaginary axis is absolutely convergent, to  $i\lambda/2$ , so another pole is at  $i\lambda/2$ . And then the periodicity implies that there are poles (at least) at all points

$$\frac{\lambda}{2} + (m + ni)\lambda \quad \frac{i\lambda}{2} + (m + ni)\lambda \quad (\text{for } m, n \in \mathbb{Z})$$

<sup>[26]</sup> This is easier to see in advance by invoking ideas about *homology*, rather than *homotopy*, though, again, there is some cost. We won't worry about the relations between the integrals, but, rather, just compute some examples to illustrate the double periodicity.

## 4. Doubly-periodic functions: elliptic functions

Once the existence of doubly periodic functions has been brought to our attention, we surely would want to construct some by other means than inversion of multi-valued integrals.

A **lattice** in  $\mathbb{C}$  is a subgroup of  $\mathbb{C}$  of the form<sup>[27]</sup>

$$\Lambda = \mathbb{Z} \cdot \omega_1 + \mathbb{Z} \cdot \omega_2 \quad (\text{with } \omega_1, \omega_2 \text{ linearly independent over } \mathbb{R})$$

We want to construct  $\Lambda$ -**periodic** functions, meaning meromorphic functions  $f$  on  $\mathbb{C}$  such that

$$f(z + \lambda) = f(z) \quad (\text{for all } z \in \mathbb{C} \text{ and } \lambda \in \Lambda)$$

These are **elliptic functions** with **period lattice**  $\Lambda$ . We might consider sums

$$\sum_{\lambda \in \Lambda} \frac{1}{(z + \lambda)^k}$$

which, when absolutely convergent,<sup>[28]</sup> are visibly invariant under  $z \rightarrow z + \lambda$  for  $\lambda \in \Lambda$ . The smallest exponent for which this sum converges (for  $z$  not in  $\Lambda$ ) is  $k = 3$ , but Weierstraß discovered<sup>[29]</sup> that it is best, with hindsight, to try to repair the convergence<sup>[30]</sup> in the  $k = 2$  case.<sup>[31]</sup> So define the **Weierstraß P-function**

$$\wp(z) = \wp_\Lambda(z) = \frac{1}{z^2} + \sum_{0 \neq \lambda \in \Lambda} \left( \frac{1}{(z + \lambda)^2} - \frac{1}{\lambda^2} \right)$$

This *does* converge absolutely, but now the easy argument for double periodicity fails.<sup>[32]</sup> Still, its derivative

$$\wp'(z) = \wp'_\Lambda(z) = -2 \sum_{\lambda \in \Lambda} \frac{1}{(z + \lambda)^3}$$

is nicely convergent *and* admits the easy argument for its periodicity.

**Claim:** The function  $\wp_\Lambda(z)$  is doubly periodic, with period lattice  $\Lambda$ .

[27] This usage *is* vaguely similar to a colloquial usage of *lattice*, if we connect nearby dots by straight lines.

[28] These series are absolutely convergent, uniformly on compacta, for  $k > 2$ .

[29] Here we follow Weierstraß's work on elliptic functions that came somewhat after Abel's and Jacobi's.

[30] One might wonder why we did not attempt a similar repair of badly-convergent sums in the construction of singly-periodic functions. *As it happens*, the analogous attempt in that situation is not productive, in that we do *not* produce  $1/\sin z$  as might be hoped. Thus, the fact that in the doubly periodic case this convergence repair *is* productive is all the more surprising.

[31] Since in general simple poles are preferable to double poles, the double-ness of the poles of  $\wp(z)$  is slightly disturbing. However, first, the subsequent discussion shows that there is *no* non-constant elliptic function with either no poles or a single pole modulo the period lattice. Second, to have all the poles at one point (modulo the period lattice) so that examination of the Laurent expansion at a single point can prove holomorphy is technically convenient. Thus, we should tolerate the double poles.

[32] Under  $z \rightarrow z + \mu$  for  $\mu \in \Lambda$ , the  $\lambda^{\text{th}}$  summand in  $\wp$  becomes  $1/(z + \mu + \lambda)^2 - 1/\lambda^2$ , and the fragility of the convergence does *not* allow us to break up these differences to rearrange and see the invariance.

*Proof:* The point is that for  $0 \neq \mu \in \Lambda$  the difference

$$\wp(z + \mu) - \wp(z) = \sum_{\lambda} \left( \frac{1}{(z + \mu + \lambda)^2} - \frac{1}{(z + \lambda)^2} \right)$$

has no poles, and is doubly periodic. By the following claim, whose usefulness we see at this point, this difference is *constant*  $C$ . Noting that  $\wp$  is in any case an *even* function,<sup>[33]</sup> let  $z = -\mu/2$  to see that

$$C = \wp(-\mu/2 + \mu) - \wp(\mu/2) = 0$$

This proves the periodicity of  $\wp$ . ///

**Claim:** An *entire* doubly periodic function is *constant*.

*Proof:* Let  $\omega_1, \omega_2$  be  $\mathbb{Z}$ -generators for  $\Lambda$ . Since the  $\omega_i$  are linearly independent over  $\mathbb{R}$ , every  $z \in \mathbb{C}$  is an  $\mathbb{R}$ -linear combination of them. Given  $z = a\omega_1 + b\omega_2$  with  $a, b \in \mathbb{R}$ , let  $m, n$  be integers such that  $0 \leq a - m < 1$  and  $0 \leq b - n < 1$ . Then

$$z = a\omega_1 + b\omega_2 = (a - m)\omega_1 + (b - n)\omega_2 + (m\omega_1 + n\omega_2)$$

Since  $m\omega_1 + n\omega_2$  is in the lattice  $\Lambda$ , this shows that every  $\Lambda$ -orbit on  $\mathbb{C}$  has a unique representative inside the so-called *fundamental domain*

$$F = \{r\omega_1 + s\omega_2 : 0 \leq r < 1, 0 \leq s < 1\}$$

for  $\Lambda$ . A  $\Lambda$ -periodic function's values on the whole plane are determined completely by its values on  $F$ . The set  $F$  has *compact* closure

$$\overline{F} = \{r\omega_1 + s\omega_2 : 0 \leq r \leq 1, 0 \leq s \leq 1\}$$

Thus, a continuous  $\Lambda$ -periodic function is *bounded* on  $\overline{F}$ , so bounded on  $\mathbb{C}$ . Thus, an entire  $\Lambda$ -periodic function is bounded. By Liouville's theorem, it is constant. ///

[... *iou* ...]pictures

**Claim:** Fix a lattice  $\Lambda$ . The Weierstraß P-function  $\wp(z)$  and its derivative  $\wp'(z)$  (attached to lattice  $\Lambda$ ) satisfy the algebraic relation<sup>[34]</sup>

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

where

$$g_2 = g_2(\Lambda) = 60 \sum_{0 \neq \lambda \in \Lambda} \frac{1}{\lambda^4} \quad g_3 = g_3(\Lambda) = 140 \sum_{0 \neq \lambda \in \Lambda} \frac{1}{\lambda^6}$$

**Remark:** We will *find* a relation satisfied by  $\wp$  and  $\wp'$ , not merely *verify* Weierstraß' relation. Determination of the relation strongly resembles the analogous computation for singly periodic functions earlier.

*Proof:* The poles of both  $\wp$  and  $\wp'$  are just on the lattice  $\Lambda$ , so if we can make a linear combination of powers of  $\wp$  and  $\wp'$  whose Laurent expansion at 0 has no negative terms or constant term, then<sup>[35]</sup> that

[33] To see that  $\wp(-z) = \wp(z)$ , observe that the term  $1/z^2$  is invariant under  $z \rightarrow -z$ , and the rest of the summands occur in pairs  $(z \pm \lambda)^2 - \lambda^2$  which are interchanged by  $z \rightarrow -z$ .

[34] We use the traditional notation. The 4 and the 60 and 140 are artifacts of the computation, as is the fact that there's no quadratic term on the right-hand side.

[35] From the previous claim, this linear combination is constant. Thus, if it vanishes at one point, it is identically 0.

linear combination of powers is identically 0. Let  $\wp_o(z)$  be  $\wp(z) - \frac{1}{z^2}$ . Then the Laurent expansion of  $\wp$  at 0 is the power series expansion of  $\wp_o$  plus the lone negative term  $\frac{1}{z^2}$ . Since  $\wp_o$  is even, at 0 its power series has no odd-degree terms. Thus, near 0,

$$\wp(z) = \frac{1}{z^2} + \wp_o(0) + \frac{\wp_o''(0)}{2!} z^2 + \frac{\wp_o''''(0)}{4!} z^4 + O(z^6) = \frac{1}{z^2} + az^2 + bz^4 + O(z^6)$$

for some constants  $a, b$  whose values we'll determine at the end. Note that  $\wp_o(0) = 0$ . Then

$$\wp'(z) = \frac{-2}{z^3} + 2az + 4bz^3 + O(z^5)$$

Then

$$\left(\frac{\wp'(z)}{-2}\right)^2 = \frac{1}{z^6} - \frac{2a}{z^2} - 4b + O(z)$$

Using

$$\wp(z)^3 = \frac{1}{z^6} + \frac{3a}{z^2} + 3b + O(z)$$

we have

$$\left(\frac{\wp'(z)}{-2}\right)^2 - \wp(z)^3 = \frac{-5a}{z^2} - 7b + O(z)$$

Then

$$\left(\frac{\wp'(z)}{-2}\right)^2 - \wp(z)^3 + 5a\wp(z) + 7b = O(z)$$

As remarked at the beginning, this linear combination of powers must be 0. That is

$$\wp'(z)^2 = 4\wp(z)^3 - 20a\wp(z) - 28b$$

Observing that

$$a = \frac{\wp_o''(0)}{2!} = \frac{1}{2!} \cdot \sum_{0 \neq \lambda \in \Lambda} \frac{(-2)(-3)}{\lambda^4} = 3 \cdot \sum_{0 \neq \lambda \in \Lambda} \frac{1}{\lambda^4}$$

and

$$b = \frac{\wp_o''''(0)}{2!} = \frac{1}{4!} \cdot \sum_{0 \neq \lambda \in \Lambda} \frac{(-2)(-3)(-4)(-5)}{\lambda^6} = 5 \cdot \sum_{0 \neq \lambda \in \Lambda} \frac{1}{\lambda^6}$$

we have Weierstraß'

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

as anticipated. ///

While we're here, it's impossible to resist proving that *every* elliptic function (with lattice  $\Lambda$ ) is expressible in terms of the corresponding  $\wp$  and  $\wp'$ :

**Theorem:** Given a lattice  $\Lambda$ , the field of meromorphic  $\Lambda$ -periodic functions is exactly the collection of rational expressions in  $\wp_\Lambda(z)$  and  $\wp'_\Lambda(z)$ . Indeed, all *even*  $\Lambda$ -periodic functions are rational expressions in  $\wp_\Lambda(z)$ .

Incidental to the proof of the theorem, we have

**Claim:** Let  $f$  be a  $\Lambda$ -periodic meromorphic function. For a fixed choice of basis  $\omega_1, \omega_2$  for  $\Lambda$ , let  $F$  be the corresponding *fundamental domain* as above. Let  $z_1, \dots, z_m$  be the zeros of  $f$  in  $F$ , and let  $p_1, \dots, p_n$  be the poles, both including multiplicities. <sup>[36]</sup> Then  $m = n$ . Further,

$$\sum_i z_i - \sum_j p_j = 0 \pmod{\Lambda}$$

<sup>[36]</sup> Usually *including multiplicities* means that for a zero  $z_o$  of order  $\ell$  the point  $z_o$  is included  $\ell$  times on the list of zeros. That is, this list is a *multiset*, not an ordinary set, since ordinary sets (by their nature) do not directly keep track of multiple occurrences of the same element.

*Proof:* Integrating  $f/f'$  around the boundary of  $F$  (make minor adaptations in case a zero or pole happens to be exactly on that path) computes  $2\pi i(m - n)$ , by Cauchy's residue theorem. On the other hand, by periodicity of  $f/f'$ , and since we integrate on opposite edges of the parallelogram  $F$  in opposite directions, this integral is 0. Thus,  $m = n$ .

Similarly, integrate  $z \cdot f'/f$  around the boundary of  $F$ . On one hand, by Cauchy's residue theorem this computes

$$2\pi i \cdot \left( \sum_i z_i - \sum_j p_j \right)$$

This time, since the function with the factor of  $z$  thrown in is *not* periodic, the integral is not 0. However, there is still some cancellation. The integral is

$$-\omega_2 \int_0^{\omega_1} \frac{f'}{f} + \omega_1 \int_0^{\omega_2} \frac{f'}{f}$$

One may easily overlook the fact that the two integrals are *integer* multiples of  $2\pi i$ , which follows from<sup>[37]</sup>

$$\int_0^{\omega_i} \frac{f'}{f} = \int_0^{\omega_i} \frac{d \log f}{d\zeta}$$

and the fact that  $f(0) = f(\omega_i)$ . That is, as  $\zeta$  goes from 0 to  $\omega_i$ , the function  $(\log f)(\zeta)$  traces out a closed path circling 0 some *integer* number of times, say  $k_i$ . Then the integral is

$$-\omega_2 \cdot 2\pi i k_1 + \omega_1 \cdot 2\pi i k_2 \in 2\pi i \cdot \Lambda$$

Cancelling the factor of  $2\pi i$ , equating the two outcomes gives

$$\sum_i z_i - \sum_j p_j \in \Lambda$$

as claimed. ///

*Proof:* Let  $f$  be a  $\Lambda$ -periodic meromorphic function on  $\mathbb{C}$ . We can break  $f$  into odd and even pieces by

$$f(z) = \frac{f(z) + f(-z)}{2} + \frac{f(z) - f(-z)}{2}$$

For  $f$  *odd*, the function  $\wp' \cdot f$  is *even*, so it suffices to prove that every *even* elliptic function is rational in  $\wp$ .

The previous claim has immediate implications for the values of  $\wp$ , which we use to form an expression in  $\wp$  that will duplicate the zeros and poles of the given *even*  $f$ . Generally, for *even*  $f$ , since  $f(-z) = f(z)$ , for  $2z_o \notin \Lambda$  and  $f(z_o) = 0$ , then  $f(-z_o) = 0$  and  $z_o$  and  $-z_o$  are distinct modulo  $\Lambda$ . For  $2z_o \in \Lambda$ , the *oddness* (and periodicity) of  $f'$  yields

$$f'(z_o) = -f'(-z_o) = -f'(-z_o + 2z_o) = -f'(z_o)$$

so  $f'(z_o) = 0$ , and the order of the zero  $z_o$  is at least 2.

In particular, by the previous claim, since  $\wp(z) - \wp(a)$  has the obvious double pole on  $\Lambda$ , it has exactly two zeros, whose sum is 0 modulo  $\Lambda$ . Obviously  $a$  itself is a 0, and for  $a \notin \frac{1}{2}\Lambda$  the unique (mod  $\Lambda$ ) other zero is  $-a$ . And for  $a \in \frac{1}{2}\Lambda$  it is a *double* zero of  $\wp(z) - \wp(a)$ .

<sup>[37]</sup> This is an instance of the *Argument Principle*.

Thus, for a zero  $z_o \notin \Lambda$  of  $f$ , the order of vanishing of  $\wp(z - \wp(z_o))$  at *all* its zeros is *at most* that of  $f$  at those zeros. Thus, by comparison to  $f(z)$ , the function

$$\frac{f(z)}{\wp(z) - \wp(z_o)}$$

has lost two zeros (either  $z_o$  and  $-z_o$  or a double zero at  $z_o$ ). The double pole of  $\wp(z) - \wp(z_o)$  at 0 makes  $f(z)/(\wp(z) - \wp(z_o))$  have order of vanishing at 0 two more than that of  $f(z)$ . No new poles are introduced by such an alteration, nor any zeros off  $\Lambda$ . Thus, since there are only finitely-many zeros (modulo  $\Lambda$ ), after finitely-many such modifications we have a function  $g(z)$  with *no* zeros off  $\Lambda$ .

Next, we get rid of *poles* of  $g(z)$  off  $\Lambda$  by a similar procedure, repeatedly *multiplying* by factors  $\wp(z) - \wp(z_o)$ . Thus, for some list of points  $z_i$  not in  $\Lambda$ , with positive and negative integer exponents  $e_i$ ,

$$f(z) \cdot \prod_i [\wp(z) - \wp(z_i)]^{e_i}$$

has no poles or zeros off  $\Lambda$ . From the previous discussion, this expression has *no* zeros or poles at all, and then is constant. ///

**Remark:** There is at least one other way to construct doubly periodic functions directions, due to Jacobi, who expressed doubly periodic functions as ratios of *entire* functions (*theta functions*) which are genuinely periodic with periods (for example)  $\mathbb{Z}$ , and nearly (but not quite) periodic in another direction. (Indeed, we saw just above that entire functions that are genuinely doubly periodic are constant!)

## 5. Elliptic modular forms

The functions traditionally denoted  $g_2 = g_2(\Lambda)$  and  $g_3 = g_3(\Lambda)$  in Weierstraß' equation relating  $\wp$  and  $\wp'$  certainly depend on the lattice, or *module*,  $\Lambda$ . It is in this sense that they are **modular forms**.<sup>[38]</sup> That is, as they arose historically, *modular forms* are functions on the set of lattices in  $\mathbb{C}$ .

The functions  $g_2$  and  $g_3$  have the further property of *homogeneity*, meaning that for any non-zero complex number  $\alpha$

$$\begin{aligned} g_2(\alpha \cdot \Lambda) &= \alpha^{-4} g_2(\Lambda) \\ g_3(\alpha \cdot \Lambda) &= \alpha^{-6} g_3(\Lambda) \end{aligned}$$

since

$$\sum_{0 \neq \lambda \in \Lambda} \frac{1}{(\alpha\lambda)^{2k}} = \alpha^{-2k} \sum_{0 \neq \lambda \in \Lambda} \frac{1}{\lambda^{2k}}$$

Of course we'd be happier if the inputs to these functions were something more familiar, rather than *lattices*, since initially we might see no better structure on the set of lattices than just that of *set*. Using the homogeneity, we can achieve this effect. Let  $F$  be a *homogeneous* function of degree  $-k$  on lattices, meaning that<sup>[39]</sup>

$$F(\alpha \cdot \Lambda) = \alpha^{-k} \cdot F(\Lambda)$$

[38] Why *form* rather than *function*? After all, these functions *are* literal functions (in our modern sense) on the set of lattices in  $\mathbb{C}$ . Certainly there was historical hesitancy to attempt to refer to functions on exotic spaces, since there was no completely abstract notion of *function* in the 19<sup>th</sup> century.

[39] Yes,  $g_2$  is homogeneous of degree  $-4$  and  $g_3$  is homogeneous of degree  $-6$ . Yes, it would have been better if their indices matched their degrees of homogeneity, at least up to sign, but the tradition developed otherwise.

Given an ordered  $\mathbb{Z}$ -basis  $\omega_1, \omega_2$  for a lattice  $\Lambda$ , we *normalize* the second basis element to 1, by multiplying  $\Lambda$  by  $\omega_2^{-1}$  and using basis  $z = \omega_1/\omega_2, 1$  for the dilated-and-rotated lattice  $\omega_2^{-1} \cdot \Lambda$ . Since

$$F(\omega_2^{-1} \cdot \Lambda) = \omega_2^k \cdot F(\Lambda)$$

we can recover the value of  $F$  on the original lattice from the value on the adjusted one.

This is not all that we know about functions of lattices. A function on lattices does not depend upon *choice of basis*. That is, for  $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$  in  $GL(2, \mathbb{Z})$ <sup>[40]</sup> the new basis

$$\begin{bmatrix} \omega'_1 \\ \omega'_2 \end{bmatrix} = \begin{bmatrix} a\omega_1 + b\omega_2 \\ c\omega_1 + d\omega_2 \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \end{bmatrix}$$

gives the same lattice, that is,

$$\mathbb{Z} \cdot \omega'_1 + \mathbb{Z} \cdot \omega'_2 = \mathbb{Z} \cdot \omega_1 + \mathbb{Z} \cdot \omega_2$$

We will combine the normalization and the change-of-basis. Given  $\mathbb{R}$ -linearly-independent  $\omega_1$  and  $\omega_2$ , let

$$z = \omega_1/\omega_2 \quad (\text{in } \mathbb{C}, \text{ not in } \mathbb{R})$$

Let

$$f(z) = F(\mathbb{Z} \cdot z + \mathbb{Z} \cdot 1)$$

For  $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$  in  $GL(2, \mathbb{Z})$ ,

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} z \\ 1 \end{bmatrix} = \begin{bmatrix} az + b \\ cz + d \end{bmatrix}$$

Thus, since change of basis does not alter the value of a function on lattices,

$$\begin{aligned} f(z) &= F(\mathbb{Z} \cdot z + \mathbb{Z} \cdot 1) = F(\mathbb{Z} \cdot (az + b) + \mathbb{Z} \cdot (cz + d)) \\ &= (cz + d)^{-k} F(\mathbb{Z} \cdot \frac{az + b}{cz + d} + \mathbb{Z} \cdot 1) = (cz + d)^{-k} f\left(\frac{az + b}{cz + d}\right) \end{aligned}$$

by the homogeneity. Thus, the action of  $GL(2, \mathbb{Z})$  or  $GL(2, \mathbb{C})$  by **linear fractional transformations**<sup>[41]</sup>

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} (z) = \frac{az + b}{cz + d}$$

arises here through renormalization of generators for lattices.

<sup>[40]</sup> As usual, for a commutative ring  $R$ , the group  $GL(2, R)$  is the group of invertible 2-by-2 matrices with entries in  $R$ . The invertibility implicitly demands that the inverse also have entries in  $R$ , which is equivalent to the determinant being in  $\mathbb{R}^\times$ .

<sup>[41]</sup> Also called *Möbius* transformations. The fact that this is a genuine group action, including associativity, is not obvious from the *ad hoc* presentation. A little later a presentation of this as being descended from a *linear* action on *projective space* will give a clear conceptual explanation for the good behavior of this action. Indeed, in general, linear fractional transformations truly act on the *Riemann sphere* complex projective one-space  $\mathbb{P}^1$ , which is  $\mathbb{C}$  with an additional point.

This brings us to the next version of **modular form**, more specifically **elliptic modular forms of weight  $k$** : these are *holomorphic* function  $f$  of a complex variable  $z$  on the upper complex half-plane<sup>[42]</sup>  $\mathfrak{H}$  or<sup>[43]</sup> on the union  $\mathfrak{H} \cup \overline{\mathfrak{H}}$  of the upper and lower half-planes, with the property that

$$f\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}(z)\right) = (cz + d)^k f(z)$$

where either<sup>[44]</sup>

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \in SL(2, \mathbb{Z}), \quad z \in \mathfrak{H} \quad (\text{for } f \text{ supposed to be defined on } \mathfrak{H})$$

or

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \in GL(2, \mathbb{Z}), \quad z \in \mathfrak{H} \cup \overline{\mathfrak{H}} \quad (\text{for } f \text{ defined on } \mathfrak{H} \cup \overline{\mathfrak{H}})$$

For example, the functions attached to the  $g_2$  and  $g_3$  above fit into a family of **Eisenstein series**<sup>[45]</sup>

$$E(z) = \sum_{c,d} \frac{1}{(cz + d)^k} \quad (\text{summed over } c, d \text{ not both } 0)$$

Since  $cz + d$  is complex, we must take  $k \in \mathbb{Z}$ . This series converges for  $k > 2$ . The series is identically 0, by obvious cancellation, for  $k$  odd.

Obvious variations suggest themselves: for fixed positive integer  $N$  and integers  $c_o, d_o$ , define Eisenstein series with **congruence conditions**

$$E(z) = \sum_{(c,d)=(c_o,d_o) \bmod N} \frac{1}{(cz + d)^k} \quad (c, d \text{ not both } 0)$$

This is an example of a modular form **of level  $N$**  (the first examples were tacitly of level 1), since we need a condition such as

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \bmod N \quad (\text{elementwise})$$

to be sure<sup>[46]</sup> that *this* Eisenstein series satisfies the *automorphy* relation

$$E\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}(z)\right) = (cz + d)^k E(z)$$

[42] The capital fraktur  $H$  for the upper half-plane is traditional, and at least saves other versions of  $H$  for other uses.

[43] The first place I saw modular forms defined as being functions on the union of the upper and lower half-planes was in writing of Deligne. At the time I mistakenly thought this was a needless or messy complication, despite whatever explanation was given there. In fact, it is much better in the long run to use the union of upper and lower half-spaces, since then the whole  $GL(2, \mathbb{R})$  and  $GL(2, \mathbb{Z})$  can act, rather than the awkward restriction to  $SL(2, \mathbb{R})$  and  $SL(2, \mathbb{Z})$ , the corresponding groups with elements required to have determinants 1.

[44] Most often these classical modular forms are assumed to live on the upper half-plane. And, indeed, in the short term, there's scant difference, since  $\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$  maps the upper to the lower.

[45] These are called Eisenstein series partly because Eisenstein studied them, and they needed a name. It is possible to give a general definition of what an *Eisenstein series* is, which we'll do later.

[46] Some choices of the data  $c_o, d_o$  modulo  $N$  may allow larger groups than  $\Gamma(N)$ . For example,  $c_o = d_o = 0$  does not require any congruence condition at all (and yields  $N^{-k}$  times the simplest Eisenstein series  $E(z) = \sum_{c,d} 1/(cz + d)^k$  summed over all  $c, d$  not both 0).

Such considerations motivate attention to natural subgroups of  $SL(2, \mathbb{Z})$ , with traditional notations: for a positive integer  $N$ ,

$$\begin{aligned}\Gamma(N) &= \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in SL(2, \mathbb{Z}) : a \equiv 1 \pmod{N}, b \equiv 0 \pmod{N}, c \equiv 0 \pmod{N}, d \equiv 1 \pmod{N} \right\} \\ &= \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in SL(2, \mathbb{Z}) : \begin{bmatrix} a & b \\ c & d \end{bmatrix} \equiv \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \pmod{N} \right\} \\ \Gamma_0(N) &= \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in SL(2, \mathbb{Z}) : \begin{bmatrix} a & b \\ c & d \end{bmatrix} \equiv \begin{bmatrix} * & * \\ 0 & * \end{bmatrix} \pmod{N} \right\} \\ \Gamma_1(N) &= \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in SL(2, \mathbb{Z}) : \begin{bmatrix} a & b \\ c & d \end{bmatrix} \equiv \begin{bmatrix} 1 & * \\ 0 & 1 \end{bmatrix} \pmod{N} \right\}\end{aligned}$$

In particular, the frequently occurring subgroup  $\Gamma(N)$  is also denoted  $\Gamma_N$  for reasons of economy:

$$\Gamma_N = \Gamma(N) = \textit{principal congruence subgroup of level } N$$

**Remark:** There is much more to be said about functions (modular forms) such as these Eisenstein series, even from an elementary viewpoint. We will indeed look at them further, in the sequel, from a technically stronger viewpoint, with an emphasis that does not necessarily adhere to historical origins. The goal of this little chapter was to sketch the early historical development, especially to shed some light on terminology, and perhaps we have said enough.

---